International Conference

On

Statistical Models and Methods for Reliability and Survival Analysis and Their Validation

4-6 July 2012, Bordeaux, France

University of Bordeaux

Mathematical Institute of Bordeaux

In the honor

of

Professor Mikhail S. Nikulin

PREFACE

This volume is the proceedings' book of S2MRSA, an international conference in statistics organized in honor of Mikhail Nikulin for his twentieth anniversary as a Professor in the Bordeaux Segalen University, France. This event was aimed at gathering specialists from laboratories, industry sectors or academia sharing experiences and ideas on recent advanced themes that belong to M. Nikulin's fields of interest.

Mikhail Nikulin was born in Saint Petersbourg (Russia) in April 1944. After obtaining his M.D. graduation in Mathematics in 1966 at the St. Petersbourg State University, he taught for three years in college (specialty: Mathematics) in Brazzaville (République Populaire du Congo). In 1973 he obtained, under the steering of L. N. Bolshev, his Ph.D. in Theory of Probabilities and Mathematical Statistics in V. Steklov Institute of Mathematics (Moscow), and directly joined in 1974 the Statistical Methods Laboratory of this same institute. In 1991-1992 he was a visiting professor in the Queen's University, Kingston, Canada. In October of 1992 he joined the "*Sciences et modélisation"* Department, University Segalen of Bordeaux.

His research in mathematical statistics and its applications have produced an impressive literature. In 1973, he edited his 1st article: "Chi-squared test for continuous distributions with shift and scale parameters", *in Theory probability and its applications*. As a whole, he recorded not less than 170 articles published from 1973 through 2012, accrued by two dozen books for which he participated as an editor, and as well as an author. He gave his name to several statistics applicable to the fields of test theory and survival analysis. For instance, literature mentions the following items as *Dzhaparidze-Nikulin statistic, Rao-Robson-Nikulin statistic, Nikulin-Rao-Robson-Moore statistic and Bagdonavicius-Nikulin model*.

After 20 continuous working years developing mathematical statistics and their applications in Bordeaux, M. Nikulin highly deserves a tribute, which is the aim of the S2MRSA conference. S2MRSA is the endpoint of a long list of past events involving M. Nikulin as a co-organizer, where topics were mostly centered around Nikulin's favorite themes such as Statistical modeling in reliability & survival analysis, or Statistical testing, estimation and applications in biostatistics, medicine, industry and demography.

We all warmly wish that, surrounded by numerous friends from all over the world who are globally recognized in mathematics and statistics, Misha (the nickname he reserves for friends only) will appreciate this gathering in a place full of history and prestige.

Vincent Couallier & Léo Gerville réache

Acknowledgements

We would like to say thanks to all plenary, invited, and contributing speakers for making this conference a valuable learning platform for all participants. Also we are thankful to all colleagues and friends, even those also who didn't succeed to attend this event, for their support in organizing this conference. Especially we thank Catherine Huber, Nikolaos Limnios and Mounir Mesbah for their encouragements, and Ingrid Rochel, Melanie Toto and Ramzan Tahir for their contribution to the excellent preparation of this meeting. These three people are the dream team of the organizing committee.

Conference Committees

Chairmen

Vincent Couallier
Leo Gerville Réache

Organisation Committee

Ingrid Rochel
Mélanie Toto
Ramzan Tahir

Plenary Speakers

Vilijandas Bagdonavicius
On generalisations of Rao-Robson-Nikulin and Zhang tests to censored and regression data

Paul Deheuvels
Uniform in Bandwidth Functional Limit Theorems and Applications

In this talk, we discuss strong and weak uniform-in-bandwidth functional limit theorems for local empirical processes. A series of results of the kind have been obtained during the last decade, among others, by U. Einmahl, D. M. Mason, D. Varron and ourselves. Part of the present lecture is based on joint work on the subject with S. Ouadah. An example of a statistical application of this theory is as follows. Consider a sequence $X_1, X_2, \ldots$, of independent and identically distributed random variables, with a continuous density $f(\cdot)$, in a neighborhood of a non-degenerate bounded interval I. Let $K(\cdot)$ denote a compactly supported function (or kernel) with bounded variation, and such that $\int_R K(t)dt = 1$. Let, for $h > 0$ and $x \in R$

$$f_{n,h}(x) = (nh)^{-1} \sum_{i=1}^{n} K\big(h^{-1}(x - X_i)\big),$$

denote the kernel density estimator of $f(x)$, with bandwidth h. Now, set $\mathcal{H}_n = [a_n, b_n]$, where a_n and b_n denote sequences such that $0 < a_n \leq b_n < \infty$, and, as $n \to \infty$,

$$na_n / \log n \to \infty \text{ and } b_n \to 0.$$

Then, as $n \to \infty$,

$$\sup_{h \in \mathcal{H}_n} \left| \left\{ \frac{nh}{2\log(1/h)} \right\}^2 \sup_{x \in I} \pm \{f_{n,h}(x) - \mathbb{E}\, f_{n,h}(x)\} - \left\{ \sup_{x \in I} f(x) \int_R K^2(t)dt \right\}^{1/2} \right| = o_P(1).$$

This uniform-in-bandwidth limit theorem illustrates the statistical applications we have in mind. In this talk, we shall present a series of examples of the kind, together with a discussion of the methods which may be used in their proofs.

Ildar Ibragimov
Statistical estimation of analytic functions

The aim of the talk is to present some results about non-parametric estimation of analytic functions. It turns out that the a priori knowledge that the function which we would like to estimate belongs to a class of analytic functions helps to construct estimates almost such good as in the case of parametric estimation. We discuss a few problems of such type.

The second question we would like to consider is the following one The uniqueness theorem says that if two functions f and g are holomorphic in a region G and $f(z) = g(z)$ for all z in some sequence of distinct points with limit point in G, then $f(z) = g(z)$ everywhere in G. The theorem means in particular that if an entire analytic function is observed on an interval I, it can be restored immediately in the whole complex plane. Of course, this problem of restoration is an ill posed one and small perturbations of the observations may drastically change the solution on large distance from the region of observation. We are interesting how far from the region of observation a consistent restoration is possible under small stochastic perturbations.

Waltraud Kahle
Optimal preventive maintenance in degradation processes

We consider the Wiener process with drift as a model of damage and degradation. A failure occurs when the degradation reaches a given level h first time. In this case, the time to failure is inverse Gaussian distributed.
Estimators for the parameters of the degradation process and the resulting lifetime distribution are given. For preventive maintenance, inspections of the degradation are regularly carried out. If at inspection time the degradation is larger than a predefined level a, then the item will be replaced by a new one. For statistical modeling we develop the density of a process increment under the condition that the process has not yet exceeded the level h.
There are three kinds of costs:

- costs of inspection,
- costs of (preventive) maintenance,
- costs of a failure.

In the talk, we consider the problem of defining optimal time intervals between inspections, as well as an optimal replacement level level a.

Mei Ling Lee
First-hitting-time based threshold regression models and comparisons with proportional hazard models

The proportional hazards (PH) assumption required by PH regression is not appropriate in some applications. Moreover, PH regression focuses mainly on hazard ratios and thus does not offer many insights into underlying determinants of survival. Threshold regression (TR) is an alternative methodology that is not built on consideration of hazards.
Threshold regression methodology is based on the concept that the degradation of a machine or a patient's health status follows a stochastic process. For engineering applications, the degradation can often be observed. For medical research, a patient's health status is a complex unobservable process. The onset of disease, or death, occurs when the process first reaches a failure threshold (i.e., a first hitting time). Instead of calendar time, analytical time (also called operational time) can be included in TR regression. The TR model is intuitive and does not require the proportional hazards assumption. It thus provides an important alternative for analyzing time-to-event data.
In this talk, we discuss the connections between these two regression methodologies. A case demonstration is used to highlight the greater understanding of scientific foundations that TR can offer in comparison to PH regression. Applications will also be demonstrated.

Table of Contents

ESTIMATION OF CONDITIONAL SURVIVAL FUNCTION IN FIXED DESIGN REGRESSION MODEL UNDER RANDOM CENSORSHIP FROM BOTH SIDES

A.A.Abdikalikov
Dpt. Probability Theory and Mathematical Statistics
National University of Uzbekistan named after Mirzo Ulugbek
Tashkent, Uzbekistan
E-mail: a_abdushukurov@rambler.ru

Abstract— **We introduce an estimator for the conditional distribution function in fixed design regression model under random censorship from both sides. Such estimator generalizes the one proposed under independent censoring model. We demonstrate the asymptotic representation result by summs of random variables.**

Keywords- *Fixed design regression model, random censoring, relative-risk power estimator.*

I. INTRODUCTION

In survival data analysis, response random variable (r.v.) Z, the survival time of a patient, that usually can be inflienced by r.v. X, often called prognostic factor. In fact, in practical situations often occurs that not all the survival times $Z_1,...,Z_n$ corresponding to n individuals, are completely observed, they may be censored. In this article we consider the case, when lifetimes censored from both sides. So let $\left\{\left(Z_j, L_j, Y_j, X_j\right), j=\overline{1.n}\right\}$ are independent replicas of vector (Z, L, Y, X), where components of vector (Z, L, Y) are independent for diven covariate X. Our sample will be consist of n vectors $\left\{\left(\zeta_i, \chi_i^{(0)}, \chi_i^{(1)}, \chi_i^{(2)}, X_i\right), i=1,...,n\right\}=S^{(n)}$, where $\zeta_i=\max\left(L_i, \min\left(Z_i, Y_i\right)\right), \chi_i^{(0)}=I\left(\min\left(Z_i, Y_i\right)<L_i\right)$, $\chi_i^{(1)}=I\left(L_i \le Z_i \le Y_i\right), \chi_i^{(2)}=I\left(L_i \le Y_i < Z_i\right)$ with $I(A)$ - denoting the indicator of event A. In sample $S^{(n)}$ the r.v.-s of interest Y_j are abservable when $\chi_j^{(1)}=1$. We denote by F_x, G_x, K_x and H_x the conditional distribution functions (d.f-s) of r.v.-s Z_j, Y_j, L_j and ζ_j respectively, given that $X_j=x$ and suppose that they are continuous. Because of the assumed conditional independence we have that

$$H_x(t)=K_x(t)\left(1-\left(1-G_x(t)\right)\left(1-F_x(t)\right)\right),\ t\in R^+. \qquad (1.1)$$

We consider only fixed–design covariates.

Let $0\le x_1\le ...\le x_n\le 1$ denote n fixed design, points. For notational simplicity these design points x_i we denote as x. For some fixed point $\tau>0$ we consider estimation of conditional d.f. $F_{\tau x}(t)=P\left(Z_x\le t / Z_x\ge\tau\right), t\ge\tau$, given $X_j=x$ from sample $S^{(n)}$.

II. ESTIMATE OF CONDITIONAL D.F.

In order to constructing the estimator of $F_{\tau x}$ we introduce sub-d.f.-s for all $t\in R^+$:

$$\begin{aligned}
&T_x^{(0)}(t)=P\left(L_x\le t, \chi_x^{(0)}=1\right)=\\
&=\int_0^t\left(1-\left(1-G_x(s)\right)\left(1-F_x(s)\right)\right)dK_x(s),\\
&T_x^{(1)}(t)=P\left(Z_x\le t, \chi_x^{(1)}=1\right)=\\
&=\int_0^t K_x(s)\left(1-F_x(s)\right)dG_x(s),\\
&T_x^{(2)}(t)=P\left(Y_x\le t, \chi_x^{(2)}=1\right)=\\
&=\int_0^t K_x(s)\left(1-G_x(s)\right)dF_x(s),
\end{aligned} \qquad (2.1)$$

where $T_x^{(0)}(t)+T_x^{(1)}(t)+T_x^{(2)}(t)=H_x(t),\ t\in R^+$. In troduce the probability $q_x(t)=P\left(L_x\le t\le\min\left(Z_x, Y_x\right)\right)=$ $=K_x(t)-H_x(t)$. For the cumulative hazard function (c.h.f) of $F_{\tau x}$ we have representation

$$\begin{aligned}
&\Lambda_{\tau x}^{(1)}(t)=\int_\tau^t\frac{dF_{\tau x}(s)}{1-F_{\tau x}(s)}=\int_\tau^t\frac{dF_x(s)}{1-F_x(s)}=\\
&=\int_\tau^t\frac{dT_x^{(1)}(s)}{q_x(s)},\ t\ge\tau.
\end{aligned} \qquad (2.2)$$

For a left-side c.h.f. of d.f. K_x

$$\Lambda_x^{(0)}(t)=\int_t^{+\infty}\frac{dK_x(s)}{K_x(s)}=\int_t^{+\infty}\frac{dT_x^{(0)}(s)}{H_x(s)},\ t\ge\tau. \tag{2.3}$$

Let $\gamma_x(t)=1-(1-G_x(t))(1-F_x(t))$, $Sp(\gamma_x)=\{t:0<\gamma_x(t)<1\}$, $\Gamma_{\tau x}^{(m)}=\left\{t\ge\tau:0<\Lambda_x^{(m)}(t)<\infty\right\}, m=0,1$. Then a number $\tau=\tau(K_x,G_x,F_x)$ we choose from conditions:

$$\begin{cases}\inf\limits_{t\in S_p(\gamma_x)\cap[\tau,+\infty)}\left\{K_x(t)(1-\gamma_x(t))\right\}>0,\\ \Gamma_{\tau x}^{(0)}\cap\Gamma_{\tau x}^{(1)}\ne\varnothing\end{cases} \tag{2.4}$$

Let $\{\omega_{ni}(x;h_n), i=\overline{1,n}\}$ - are Gasser-Müller type weights, given by

$$\omega_{ni}(x;h_n)=\left(\int_0^{x_n}\frac{1}{h_n}\pi\left(\frac{x-y}{h_n}\right)dy\right)^{-1}\int_{x_{i-1}}^{x_i}\frac{1}{h_n}\pi\left(\frac{x-y}{h_n}\right)dy,$$

where $x_0=0,\pi$ is known density function and $\{h_n\downarrow 0\ as\ n\to\infty\}$ - sequence of bandwidths. Then the conditional d.f.-s (1.1) and (2.1) estimated by following Stone type kernel statistics for $t\in R^+$ [2]:

$$\begin{aligned}H_{xh}(t)&=\sum_{i=1}^n\omega_{ni}(x;h_n)I(\zeta_i\le t)\\ T_{xh}^{(m)}(t)&=\sum_{i=1}^n\omega_{ni}(x;h_n)I\left(\zeta_i\le t,\chi_i^{(m)}=1\right),\ m=0,1,2\end{aligned} \tag{2.5}$$

By solving the integral equation (2.3) with respect to d.f. K_x and using estimates (2.5) we obtain the estimator of K_x as

$$K_{xh}(t)=\exp\left\{\int_t^{+\infty}\frac{dT_{xh}^{(0)}(s)}{H_{xh}(s)}\right\}, t\ge\tau.$$

Then the probability $q_x(t)$ may be estimated by

$$q_{xh}(t)=K_{xh}(t)-H_{xh}(t). \tag{2.6}$$

Let $\Lambda_{\tau xh}^{(1)}(t)=\int_t^\tau\frac{dT_{xh}^{(1)}(s)}{q_{xh}(s)}$, $t\ge\tau$, is an estimate of c.h.f. (2.2). By using an ideas from [1], we introduce the following relative – risk power estimator of conditional survival function $1-F_{\tau x}(t)$:

$$1-F_{\tau xh}(t)=\left[\frac{q_{xh}(t)}{q_{xh}(\tau)}\right]^{R_{\tau xh}(t)},\quad t\ge\tau, \tag{2.7}$$

where $R_{\tau xh}(t)=\Lambda_{\tau xh}^{(1)}(t)\left[-\int_\tau^t\frac{dq_{xh}(s)}{q_{xh}(s)}\right]^{-1}$.

III. ASYMPTOTIC REPRESENTATION FOR ESTIMATOR

For investigating the estimator (2.7) we need in some notations and conditions. Let

$$\underline{\Delta_n}=\min_{1\le i\le n}(x_i-x_{i-1}),\quad \overline{\Delta_n}=\max_{1\le i\le n}(x_i-x_{i-1}),$$

and introduce the conditions

(C1) $As\ n\to\infty,\ x_n\to 1,\ \overline{\Delta_n}=O\left(\frac{1}{n}\right),\ \overline{\Delta_n}-\underline{\Delta_n}=o\left(\frac{1}{n}\right)$;

(C2) Kernel π have a compact support $[-M,M], M>0$,

$\int_{-\infty}^{\infty}u\,\pi(u)\,du=0$ and π is Lipschitz of order 1;

Let $N_x(t)$ is some d.f. Consider the following conditions for all $(x;t)\in[0,1]\times[0,T]$ for some $T<T_{N_x}=\inf\{t:N_x(t)=1\}$:

(C3) $\ddot{N}_x(t)=\frac{\partial^2}{\partial x^2}N_x(t)$, $N_x''(t)=\frac{\partial^2}{\partial t^2}N_x(t)$ and $\dot{N}_x'(t)=\frac{\partial^2}{\partial x\partial t}N_x(t)$ exist and continuous.

Let $\tau_{H_x}=\sup\{t:H_x(t)=0\}$ and $\tau_{H_x}<\tau<T<T_{H_x}$.

Theorem. Suppose that the conditions (2.4), (C1)-(C3) are hold and $h_n\to 0$, $\frac{\log n}{nh_n}\to 0$, $\frac{nh_n^5}{\log n}=O(1)$ as $n\to\infty$.

Then for all $t\in[\tau,T]$:

$$F_{\tau xh}(t)-F_{\tau x}(t)=\sum_{i=1}^n\omega_{ni}(x;h_n)\Psi_{tx}^{(n)}\left(\zeta_i,\chi_i^{(0)},\chi_i^{(1)},\chi_i^{(2)}\right)+R_n(t;x),$$

where $\sup\limits_{\tau\le t\le T}|R_n(t,x)|\overset{a.s.}{=}O\left(\left(\frac{\log n}{nh_n}\right)^{3/4}\right)$,

$$\Psi_{tx}^{(n)}\left(\zeta_i,\chi_i^{(0)},\chi_i^{(1)},\chi_i^{(2)}\right)=(1-F_{\tau x}(t))\times$$

$$\times\left\{\int_\tau^t\frac{\left(I(\zeta_i\le s)-H_x(s)\right)-\lambda_{xh}(s)\right)dT_x^{(1)}(s)}{\left(K_x(s)-H_x(s)\right)^2}+\right.$$

$$+\frac{\left(I\left(\zeta_i\le t,\chi_i^{(1)}=1\right)-T_x^{(1)}(t)\right)}{K_x(t)-H_x(t)}-\frac{I\left(\zeta_i\le\tau,\chi_i^{(1)}=1\right)-T_x^{(1)}(\tau)}{K_x(\tau)-H_x(\tau)}-$$

$$\left.-\int_\tau^t\frac{\left(I\left(\zeta_i\le s,\chi_i^{(1)}=1\right)-T_x^{(1)}(s)\right)d\left(K_x(s)-H_x(s)\right)}{\left(K_x(s)-H_x(s)\right)^2}\right\},$$

,

and

$$\lambda_{xh}(t)=-K_x(t)\sum_{i=1}^n\left\{\int_t^{+\infty}\frac{\left(I(\zeta_i\le s)-H_x(s)\right)dT_x^{(0)}(s)}{H_x^2(s)}+\right.$$

$$+\frac{\left(I\left(\zeta_i \le t, \chi_i^{(0)}=1\right)-T_x^{(0)}(t)\right)}{H_x(t)}-$$

$$-\int_t^{+\infty}\frac{\left(I\left(\zeta_i \le s, \chi_i^{(0)}=1\right)-T_x^{(0)}(s)\right)dH_x(s)}{H_x^2(s)}\Bigg\}.$$

REFERENCES

[1] Abdushukurov A.A. Estimation of unknown distributions by incomplete observations and its properties, LAMBERT Academic Publishing, 301 p. 2011.(In Russian).

[2] Stone C.J. Consistent nonparametric regression. //Ann. Statist.-1977/ -v.5.-p. 595-645.

Asymptotic representations for the likelihood ratio statistics in competing risks model under progressive censoring

Abdushukurov A.A
Dpt. Probability Theory and Mathematical Statistics
National University of Uzbekistan named after Mirzo Ulugbek
Tashkent, Uzbekistan
E-mail: a_abdushukurov@rambler.ru

Nurmukhamedova N.S.
Dpt. Probability Theory and Mathematical Statistics
National University of Uzbekistan named after Mirzo Ulugbek
Tashkent, Uzbekistan
E-mail: rasulova_nargiza@mail.ru

***Abstract*—One of the basic properties of likelihood ratio statistic is the local asymptotic normality(LAN). In this paper we present a theorem on local asymptotic normality in the model of competing risks in presence of random right censoring.**

Keywords-competing risks model, random censoring, likelihood ratio statistics, local asymptotic normality.

I. INTRODUCTION

Consider the following competing risks models [1]. Let X - a non-negative random variable (r.v.), meaning the lifetime of the test object is defined on a probability space $(\Omega, \mathcal{A}, P)$ with values in a measurable space $(\mathcal{X}, \mathcal{B})$, where $\mathcal{X} \subseteq \mathbb{R}^+ = [0,\infty)$ and $\mathcal{B} = \sigma(\mathcal{X})$. Let $\{A^{(1)},...,A^{(k)}\}$ - pairwise disjoint events (or at least $P(A^{(i)} \cap A^{(j)}) = 0,\ i \neq j,$ $i, j = \overline{1,k}$), such that $P(\bigcup_{i=1}^{k} A^{(i)}) = 1$. We are interesting in joint properties of the vector $Z = (X, \delta^{(1)},..., \delta^{(k)})$, where $\{\delta^{(i)} = I(A^{(i)}), i = \overline{1,k}\}$ - indicators of these events. Let the joint distribution of the random vector Z depends on unknown parameter $\theta \in \Theta$:

$$Q_\theta(x, y^{(1)},..., y^{(k)}) = P_\theta(X \le x, \delta^{(1)} = y^{(1)},..., \delta^{(k)} = y^{(k)}),$$

where $x \in \overline{R}^+ = [0,\infty]$, $y^{(i)} \in \{0,1\}$, $i = \overline{1,k}$; and Θ - open set in $\mathbb{R}^1 = (-\infty,\infty)$.

Let $H(x;\theta) = P_\theta(X \le x)$ и $H^{(i)}(x;\theta) = P_\theta(X < x, \delta^{(i)} = 1)$ - marginal distributions of r.v. X and pairs $\{(X, \delta^{(i)}), i = \overline{1,k}\}$. As $\delta^{(1)} + ... + \delta^{(k)} = 1$, therefore for all $(x;\theta) \in \overline{\mathbb{R}}^+ \times \Theta$: $H^{(1)}(x;\theta) + ... + H^{(k)}(x;\theta) = H(x;\theta)$. Suppose that sub distributions $H^{(i)}$ absolutely continuous and have densities $h^{(i)}$, then $h^{(1)}(x;\theta) + ... + h^{(k)}(x;\theta) = \frac{\partial H(x;\theta)}{\partial x} = h(x;\theta)$ for all $(x;\theta) \in \overline{\mathbb{R}}^+ \times \Theta$. We also introduce the hazards functions $\lambda^{(i)}(x;\theta) = h^{(i)}(x;\theta) / (1 - H(x;\theta))$, $\lambda(x;\theta) = h(x;\theta) / (1 - H(x;\theta)) = \lambda^{(1)}(x;\theta) + ... + \lambda^{(k)}(x;\theta)$, $(x;\theta) \in \overline{\mathbb{R}}^+ \times \Theta$. Let $\{(X_j, A_j^{(1)},..., A_j^{(k)}), j \ge 1\}$ - sequence of independent replicas of aggregate $(X, A^{(1)},..., A^{(k)})$ and on n-th step of the experiment observations available the sample of volume n : $\overline{\mathbb{Z}} = (Z_1,..., Z_n)$, where $Z_j = (X_j, \delta_j^{(1)},..., \delta_j^{(k)})$, $\delta_j^{(i)} = I(A_j^{(i)}),\quad i = \overline{1,k},\ j = \overline{1,n}.$ Let $X_{1n} < X_{2n} < ... < X_{nn}$ - ordered statistics of r.v.-s $\{X_1,..., X_n\}$, where by the continuity of distribution H equate the options are missing. Through $\{\delta_{jn}^{(1)},..., \delta_{jn}^{(k)}, j = \overline{1,n}\}$ denote the concomitant indicators corresponding the r.v.-s $\{X_{jn}, j = \overline{1,n}\}$. Let

$$\mathbb{Z}^{(m)} = (\mathcal{Z}_{1n},..., \mathcal{Z}_{mn}),\ m = 1,...,n;\ \mathbb{Z}^{(0)} = \mathcal{Z}_{0n} = 0, \qquad (1.1)$$

where $\mathcal{Z}_{jn} = (X_{jn}, \delta_{jn}^{(1)},..., \delta_{jn}^{(k)})$.

Let $(\mathcal{Y}^{(n)}, \mathcal{U}^{(n)}, Q_\theta^{(n)})$ denote the sequence of statistical experiments generated by observations (1.1). Therefore, $\mathcal{Y}^{(n)} = \{\mathcal{X} \times \{0,1\}^{(k)}\}^{(n)}$, $\mathcal{U}^{(n)} = \sigma(\mathcal{Y}^{(n)})$, $Q_\theta^{(n)}$ - distribution on $(\mathcal{Y}^{(n)}, \mathcal{U}^{(n)})$, being "$n$-fold product of one-dimensional" distributions Q_θ. The family of measures $\{Q_\theta^{(n)}, \theta \in \Theta, n \ge 1\}$ is absolutely continuous with respect to measure $\nu^{(n)} = \nu_1 \times ... \times \nu_n$, where $d\nu_m = d\gamma(x_m) \times \varepsilon_{y_m^{(1)}} \times ... \times \varepsilon_{y_m^{(k)}}$, γ - measure on $\mathcal{X}$ and $\varepsilon_{y_m^{(i)}}$ - counting measures concentrated at a point $y_m^{(i)} \in \{0,1\}, i = \overline{1,k}; m = \overline{1,n}$ and its density is given by

$$q_n(\mathbb{Z}^{(n)};\theta)=\frac{dQ_\theta^{(n)}(\mathbb{Z}^{(n)})}{d\nu^{(n)}(\mathbb{Z}^{(n)})}=\prod_{m=1}^{n}\prod_{i=1}^{k}[h^{(i)}(x_m;\theta)]^{y_m^{(i)}}.$$

Let $(\mathcal{Y}^*,\mathcal{U}^*)=\prod_{j=1}^{\infty}(\mathcal{Y}^{(j)},\mathcal{U}^{(j)})$ and $\mathcal{P}_\theta$ - probability distribution on $(\mathcal{Y}^*,\mathcal{U}^*)$, is a product measure of Q_θ and - E_θ the expectation operator with respect to $\mathcal{P}_\theta$. Let $U_{n,m}$ denote σ - algebra generated by the vectors (1.1). Let $\{\tau_n, n\ge 1\}$ - sequence of stopping times, measurable with respect to non-decreasing by m sequence of σ - algebras $\{U_{n,m}, 1\le m\le n\}$. Through $\mathcal{P}_{n,\theta}$ denote the restriction of $\mathcal{P}_\theta$ on σ - algebra U_{n,τ_n}. Next, we are interested in a LAN property of family of probability measures $\{\mathcal{P}_{n,\theta}, \theta\in\Theta\}$, corresponding progressive censored from the right sample $\mathbb{Z}^{(\tau_n)}$. Let $\mathcal{P}_{n,\theta}^{(m)}$ - restriction of the measure $\mathcal{P}_\theta$ on $U_{n,m}$. Then the joint density of the vector $\mathbb{Z}^{(m)}$ is given by

$$p_n(z^{(m)};\theta)=\frac{n!}{(n-m)!}\prod_{l=1}^{m}\prod_{i=1}^{k}\{[h^{(i)}(z_{ln};\theta)]^{y_{ln}^{(i)}}\}[1-H(z_{mn};\theta)]^{n-m}, \tag{1.2}$$

where $z^{(m)}=(z_{1n},...,z_{mn})$ - realization of the vector $\mathbb{Z}^{(m)}$. Support of density (1.2) is the set $N_{n,m}=\{z^{(m)}: 0<z_{1n}<...<z_{mn}<\infty\}$. Let $u_\theta(z_{mn}/U_{n,m-1})$ - the conditional density of Z_{mn} on σ - algebras $U_{n,m-1}$. Then it is easy to see that by (1.2) for $z_{mn}>Z_{m-1,n}$:

$$u_\theta(z_{mn}/U_{n,m-1})=(n-m+1)\prod_{i=1}^{k}\left\{\frac{[h^{(i)}(z_{mn};\theta)]^{y_{mn}^{(i)}}[1-H(z_{mn};\theta)]^{n-m}}{[1-H(z_{m-1,n};\theta)]^{n-m+1}}\right\}. \tag{1.3}$$

In view of (1.2) and (1.3) we have

$$p_n(z^{(m)};\theta)=p_n(z^{(m-1)};\theta)u_\theta(z_{mn}/U_{n,m-1})=\prod_{l=1}^{m}u_\theta(z_{ln}/U_{n,l-1}). \tag{1.4}$$

II. Local asymptotic normality

Let $\theta_0\in\Theta$ - fixed value of θ. For a given $u\in\mathbb{R}^1$, define the sequence $\theta_n=\theta_0+un^{-1/2}\in\Theta$. We introduce the Likelihood Ratio Statistics (LRS) for $m=\overline{1,n}$:

$$\frac{d\mathcal{P}_{n,\theta_n}^{(m)}}{d\mathcal{P}_{n,\theta_0}^{(m)}}=l_{n,m}(u)=\frac{p_n(z^{(m)};\theta_n)}{p_n(z^{(m)};\theta_0)}=\prod_{l=1}^{m}\frac{u_{\theta_n}(z_{ln}/U_{n,l-1})}{u_{\theta_0}(z_{ln}/U_{n,l-1})}. \tag{2.1}$$

To prove the LAN to the LRS (2.1) we introduce a regularity condition:

(C1) Support $\{Sp(h^{(i)})=\{x\ge 0: h^{(i)}(x;\theta)>0\}, i=\overline{1,k}\}$ - independent on θ;

(C2) Densities $\{h^{(i)} i=\overline{1,k}\}$ - continuously differentiable on θ and derivatives $\{\frac{\partial h^{(i)}}{\partial\theta}, i=\overline{1,k}\}$ - uniformly bounded for all $(x;\theta)\in\bar{\mathbb{R}}^+\times\Theta_0$, where Θ_0 - neighborhood θ_0;

(C3) For γ - almost all $x\in R^+$ functions $\{\frac{\partial\log\lambda^{(i)}(x;\theta_0)}{\partial\theta}, i=\overline{1,k}\}$ - differentiable;

(C4) There exists a number $\Delta>0$ such that for all $i=\overline{1,k}$:

$$E_\theta\left|\frac{\partial\log\lambda^{(i)}(X;\theta_0)}{\partial\theta}\right|^{2+\Delta}<\infty, E_\theta\left|\frac{\partial\log h^{(i)}(X;\theta_0)}{\partial\theta}\right|^{2+\Delta}<\infty;$$

(C5) There exists a number $\beta\in(0,1]$ such that $n\to\infty$, $\frac{\tau_n}{n}\xrightarrow{Q_{\theta_0}}\beta$ and for all $i=\overline{1,k}$:

$$I_\beta^{(i)}=I_\beta^{(i)}(\theta_0)=\int_0^{H^{-1}(\beta;\theta_0)}\left(\frac{\partial\log\lambda^{(i)}(x;\theta_0)}{\partial\theta}\right)^2 dH(x;\theta_0)>0,$$

where $H^{-1}(\beta;\theta_0)$ - quantile on level β of distributions $H(x;\theta_0)$;

(C6) For each $m=\overline{1,n}$ functions

$$\int_{z_{m-1,n}}^{\infty}u(z/U_{n,m-1})d\gamma(z)$$

- differentiable by θ in point θ_0 under the integral sign;

(C7) For each $u\in R^1$

$$\lim_{n\to\infty}E_\theta\left\{\sup_{|\theta-\theta_0|\le\frac{u}{\sqrt{n}}}\frac{1}{n}\sum_{m=1}^{n}\int_{z_{m-1,n}}^{+\infty}\left[\frac{\partial}{\partial\theta}\{u_\theta(z/U_{n,m-1})\}^{1/2}-\right.\right.$$

$$\left.\left.-\frac{\partial}{\partial\theta_0}\{u_{\theta_0}(z/U_{n,m-1})\}^{1/2}\right]^2 d\gamma(z)\right\}=0,$$

where $\frac{\partial}{\partial\theta_0}\{u_{\theta_0}(z/U_{n,m-1})\}^{1/2}=\left(\frac{\partial}{\partial\theta}\{u_\theta(z/U_{n,m-1})\}^{1/2}\right)_{\theta=\theta_0}$.

Note that the conditions (C1) - (C3) are the usual regularity conditions, (C4) provide for the establishment of the Lyapunov condition for LAN of LRS, (C5) gives the constraint on the growth of τ_n and requires a positive Fisher information $I_\beta^{(i)}$, and the conditions (C6) and (C7) requires continuity of the integrals conditional densities.

Later we use the notation in $m=\overline{1,n}$

$$\eta_{n,m}=\frac{\partial \log p_n(\mathbb{Z}^{(m)};\theta_0)}{\partial\theta}, \tag{2.2}$$

$$I_{n,m}=M_{\theta_0}\eta^2_{n,m},\quad I_{n,\tau_n}=M_{\theta_0}\eta^2_{n,\tau_n}. \tag{2.3}$$

We study asymptotic properties of statistics $\{\omega_{n,\tau_n}, n\geq 1\}$, where $\omega_{n,\tau_n}=\eta_{n,\tau_n}I^{-1/2}_{n,\tau_n}$. Let $I_\beta=I^{(1)}_\beta+...+I^{(k)}_\beta$.

Theorem. Under conditions (C1) - (C7) and for each $u\in R^1$ for LRS the following representation holds

$$l_{n,\tau_n}(u)=\exp\{uI^{1/2}_\beta\omega_{n,\tau_n}-\frac{u^2}{2}I_\beta+R_n(u)\}, \tag{2.4}$$

where at $n\to\infty$, $R_n(u)\overset{Q_{\theta_0}}{\to}0$, $\mathcal{L}(\omega_{n,\tau_n}/Q_{\theta_0})\to\mathcal{L}(\zeta)$, *and* $\zeta\overset{D}{=}N(0,1)$.

Remark. The representation (2.4) provides a LAN family of measures $\{Q^{(n)}_{\theta_0},\theta\in\Theta\}$ at point θ_0. In this case for each $u\in R^1$ and $n\to\infty$:

$$\mathcal{L}(l_{n,\tau_n}/Q_{\theta_0})\to\mathcal{L}(l^-(u)), \tag{2.5}$$

$$\mathcal{L}(l_{n,\tau_n}/Q_{\theta_n})\to\mathcal{L}(l^+(u)), \tag{2.6}$$

where $l^\pm(u)=\exp\{uI^{1/2}_\beta(\theta_0)\zeta\pm\frac{u^2}{2}I_\beta(\theta_0)\}$. The convergence (2.5) is a direct consequence of (2.4) and (2.6) follows from the first lemma of Le Cam and the fact that the family of probability measures $\{Q^{(n)}_{\theta_n},n\geq1\}$ contiguous with respect to the family $\{Q^{(n)}_{\theta_0},n\geq1\}$.

REFERENCES

[1] Abdushukurov A.A. Estimation of unknown distributions by incomplete observations and its properties, LAMBERT Academic Publishing, 301 p. 2011.(In Russian).

Estimation of survival and mean residual life functions from dependent right random censored data

A.A. Abdushukurov
Dpt. Probability theory and mathematical statistics
National University of Uzbekistan
Tashkent, Uzbekistan
a_abdushukurov@rambler.ru

N.T.Dushatov
Dpt. Probability theory and mathematical statistics
National University of Uzbekistan
Tashkent, Uzbekistan
n_dushatov@mail.ru

R.S. Muradov
Dpt. Probability theory and mathematical statistics
National University of Uzbekistan
Tashkent, Uzbekistan
r_muradov1985@rambler.ru

***Abstract*— In this article an estimators for survival and mean residual life functions with using Archimedean copulas under random censoring from the right are proposed. The properties of estimators are presented.**

Keywords- random censorship, survival function, mean residual life function, Sklar's theorem, Archimedean copulas.

I. Introduction

In survival analysis our interest focuses on a nonnegative random variables (r.v.-s) denoting death times of biological organisms or failure times of mechanical systems. A difficulty in the analysis of survival data is the possibility that the survival times can be subjected to random censoring by other nonnegative r.v.-s and therefore we observe incomplete data. There are various types of censoring mechanisms. In this article we consider only right censoring model and problem of estimation of survival and mean residual life functions when the survival times and censoring times are dependent and propose new estimates of survival functions assuming that the dependence structure is described by a known copula function.

II. THE MODEL AND ESTIMATOR OF SURVIVAL FUNCTION

On the probability space $(\Omega, \mathcal{A}, P)$ we consider $\{(X_k, Y_k), k \geq 1\}$ - a sequence of independent and identically distributed pairs of nonnegative r.v.-s with common joint distribution function (d.f.) $H(x,y) = P(X_1 \leq x, Y \leq y)$, $(x,y) \in \overline{R}^{+2} = [0,\infty]^2$. We suppose that the marginal d.f.-s $F(x) = P(X_1 \leq x) = H(x,\infty)$ and $G(y) = P(Y_1 \leq y) = H(+\infty, y)$, $x, y \in \overline{R}^{+}$, are continuous and $F(0) = G(0) = 0$. Assume that the sequence $\{X_k, k \geq 1\}$ is right censored by the sequence $\{Y_k, k \geq 1\}$ and at n-th stage of the experiment the observation is available the sample $\mathbb{V}^{(n)} = \{(Z_k, \delta_k), 1 \leq k \leq n\}$, where $Z_k = \min(X_k, Y_k)$, $\delta_k = I(Z_k = X_k)$ and $I(A)$ is the indicator of the event A. Should be noted that it does not require independence of sequences $\{X_k\}$ and $\{Y_k\}$. The problem is consist in estimating of the survival function $S^X(x) = \mathbb{P}(X_1 > x) = 1 - F(x)$, $x \in \overline{R}^{+}$, from the sample $\mathbb{V}^{(n)}$. Let $\overline{H}(x,y) = P(X_1 > x, Y_1 > y), (x,y) \in \overline{R}^{+2}$ - a joint survival function of the pairs (X_k, Y_k). According to Theorem of Sclar H and $\overline{H}$ can be submitted through the appropriate copula functions (see [4,5]):

$$\begin{aligned} H(x,y) &= C(F(x); G(y)), (x,y) \in \overline{R}^{+2}, \\ \overline{H}(x,y) &= C^*(S^X(x); S^Y(y)), (x,y) \in \overline{R}^{+2}, \end{aligned} \tag{1}$$

where copulas C and C^* are related as

$$C^*(u,v) = u + v - 1 + C(1-u; 1-v), (u,v) \in [0,1]^2. \tag{2}$$

In the sequel in order to construct estimates for the survival function S^X, assume that C^* is Archimedean copula, i.e. $C^*(u,v) = \varphi^{[-1]}[\varphi(u) + \varphi(v)], (u,v) \in [0,1]^2$, where $\varphi: [0,1] \to \overline{R}^{+}$ is some generator function with the pseudo inverse $\varphi^{[-1]}$. Thus, by (1) and (2)

$$\begin{aligned} \overline{H}(x,y) &= \varphi^{[-1]}[\varphi(S^X(x)) + \varphi(S^Y(y))], \ (x,y) \in \overline{R}^{+2}, \\ S^Z(x) &= \varphi^{[-1]}[\varphi(S^X(x)) + \varphi(S^Y(x))], \quad x \in \overline{R}^{+}. \end{aligned} \tag{3}$$

We introduce a usual λ^X, λ^Z and "crude" $\lambda-$ hazard functions

$$\begin{aligned} \lambda^X(x) &= \lim_{\Delta \downarrow 0} \frac{1}{\Delta} P(x < X_1 \leq x + \Delta / X_1 > x), \\ \lambda^Z(x) &= \lim_{\Delta \downarrow 0} \frac{1}{\Delta} P(x < Z_1 \leq x + \Delta / X_1 > x, Y_1 > x), \\ \lambda(x) &= \lim_{\Delta \downarrow 0} \frac{1}{\Delta} P(x < X_1 \leq x + \Delta / X_1 > x, Y_1 > x). \end{aligned} \tag{4}$$

In order to construct a copula estimates for S^X consider the following easily verifiable equality:

$$\lambda^X(x)S^X(x)\varphi'\left(S^X(x)\right)=\lambda(x)S^Z(x)\varphi'\left(S^Z(x)\right). \qquad (5)$$

Integrating (5) over the interval $[0,x]$ and denoting by $\Lambda(x)=\int_0^x\lambda(t)dt$ and $\Lambda^X(x)=\int_0^x\lambda^X(t)dt$ corresponding cumulative hazard functions we obtain the integral equation

$$\int_0^x S^X(t)\varphi'\left(S^X(t)\right)d\Lambda^X(t)=\int_0^x S^Z(t)\varphi'\left(S^Z(t)\right)d\Lambda(t),\quad x\in\overline{R}^+. \qquad (6)$$

Integral on the left side of (6) is equal to $-\varphi\left(S^X(x)\right)$ and then (6) takes the form

$$\varphi\left(S^X(x)\right)=-\int_0^x S^Z(t)\varphi'\left(S^Z(t)\right)d\Lambda(t),\quad x\in\overline{R}^+. \qquad (7)$$

Hence we find the expression for the survival function S^X:

$$S^X(x)=\varphi^{[-1]}\left[-\int_0^x S^Z(t)\varphi'\left(S^Z(t)\right)d\Lambda(t)\right],\quad x\in\overline{R}^+. \qquad (8)$$

Note that the survival function S^Z permit usual empirical estimation by the values Z_k observed in the sample $\mathbb{V}^{(n)}$:

$$S_n^Z(x)=\frac{1}{n}\sum_{k=1}^n I(Z_k>x),\quad x\in\overline{R}^+. \qquad (9)$$

Substituting (9) to the right of representation (8), we obtain a preliminary estimate of S^X as

$$\widetilde{S}_n^X(x)=\varphi^{[-1]}\left[-\int_0^x I\left(S_n^Z(t-)>0\right)S_n^Z(t-)\varphi'\left(S_n^Z(t)\right)d\Lambda_n(t)\right], \qquad (10)$$

where

$$\Lambda_n(t)=\frac{1}{n}\sum_{k=1}^n\frac{I(Z_k\le t,\delta_k=1)}{S_n^z(Z_k)-\frac{1}{n}}, \qquad (11)$$

-the corresponding estimate for $\Lambda(t)=\int_0^t\frac{dP(Z_1\le s,\delta_1=1)}{P(Z_1>s)}$. Estimate (10) plays a supporting role in the construction of the main estimates for S^X in the future. Let $N_k(t)=\ =I(Z_k\le t,\delta_k=1)$. Define the counting processes $\overline{\mathbb{N}}_n(t)=\sum_{k=1}^n N_k(t)$ and $\mathbb{J}_n(t)=nS_n^Z(t-)=$ $=\sum_{k=1}^n I(Z_k\ge t)$. Then the estimates (10) and (11) can be represented as

$$\widetilde{S}_n^X(x)=\varphi^{[-1]}\left[-\frac{1}{n}\int_0^x I\left(\mathbb{J}_n(t)>0\right)\varphi'\left(\frac{\mathbb{J}_n(t)}{n}\right)d\overline{\mathbb{N}}_n(t)\right], \qquad (12)$$

$$\Lambda_n(t)=\int_0^t\frac{I\left(\mathbb{J}_n(s)>0\right)}{\mathbb{J}_n(s)}\,d\overline{\mathbb{N}}_n(s).$$

Given the analog left side of (6), i.e.

$$\varphi\left(S^Z(x)\right)=-\int_0^x S^Z(t)\varphi'\left(S^Z(t)\right)d\Lambda^Z(t), \qquad (13)$$

where $\Lambda^Z(t)=\int_0^t\lambda^Z(s)ds$, together with (9) also obtain other estimate for S^Z as

$$\widetilde{S}_n^Z(x)=\varphi^{[-1]}\left[-\frac{1}{n}\int_0^x I\left(\mathbb{J}_n(t)>0\right)\varphi'\left(\frac{\mathbb{J}_n(t)}{n}\right)d\overline{N}_n^Z(t)\right], \qquad (14)$$

where $\Lambda_n^Z(t)=\int_0^t\frac{I\left(\mathbb{J}_n(s)>0\right)}{\mathbb{J}_n(s)}d\overline{N}_n^Z(s)$, is estimate for $\Lambda^Z(t)$ and $\overline{N}_n^Z(t)=n\left(1-S_n^Z(t)\right)=n-\mathbb{J}_n(t+)=\sum_{k=1}^n N_k^Z(t)=\sum_{k=1}^n I(Z_k\le t)$- the counting process. For S^X have the following obvious identity obtained from the representations (7) and (13):

$$S^X(x)=\varphi^{[-1]}\left[\varphi\left(S^Z(x)\right)\frac{\left(-\int_0^x S^Z(t)\varphi'\left(S^Z(t)\right)d\Lambda(t)\right)}{\left(-\int_0^x S^Z(t)\varphi'\left(S^Z(t)\right)d\Lambda^Z(t)\right)}\right]. \qquad (15)$$

Now substituting the empirical estimate of (9) under the first factor on the right of representation (15) and the corresponding estimates (12) and (14) instead of integrals we obtain the final estimate of S^X in the form

$$S_n^X(x)=\varphi^{[-1]}\left[\varphi\left(S_n^Z(x)\right)\frac{\left(-\int_0^x I\left(\mathbb{J}_n(t)>0\right)\varphi'\left(\frac{\mathbb{J}_n(t)}{n}\right)d\overline{N}_n(t)\right)}{\left(-\int_0^x I\left(\mathbb{J}_n(t)>0\right)\varphi'\left(\frac{\mathbb{J}_n(t)}{n}\right)d\overline{N}_n^Z(t)\right)}\right], \qquad (16)$$

where

$$\varphi\left(S_n^Z(x)\right)=-\int_0^x I\left(\mathbb{J}_n(s)>0\right)\left[\varphi\left(\frac{\mathbb{J}_n(s)}{n}\right)-\varphi\left(\frac{\mathbb{J}_n(s)}{n}-\frac{1}{n}\right)\right]d\overline{\mathbb{N}}_n^Z(s),$$

is estimator of $\varphi\left(S^Z(x)\right)$.

In fact, we suppose that in (15) the generator function φ is strong (that is $\varphi(0)=\infty$) and hence $\varphi^{[-1]}=\varphi^{-1}$ is usual inverse function.

Denote $Z^{(n)}=\sup\left\{x\ge 0:\mathbb{J}_n(x)>0\right\}$, $T_Z=\sup\left\{x\ge 0:\right.$ $\left.S^Z(x)>0\right\}$, $\Psi(x)=-x\varphi'(x)$. Introduce the regularity conditions with respect to S^X, S^Z and the copula generator φ. By Λ^* in conditions below denote both of Λ and Λ^Z:

(C1) The strong generator function $\varphi(\cdot)$ is strictly decreasing on (0,1] and is sufficiently smooth in the sense that the first two derivatives of the functions $\varphi(x)$ and $\Psi(x)$ are bounded for $x\in[\varepsilon,1]$, where $\varepsilon>0$ is arbitrary. Moreover,

the first derivative φ' is bounded away from zero on $[0,1]$;

(C2) $0 < \int_0^{T_Z} \left[\Psi\left(S^Z(x)\right)\right]^m d\Lambda^*(x) < \infty$ for $m = 0,1,2$;

(C3) $\int_0^{T_Z} \left|\Psi'\left(S^Z(x)\right)\right| d\Lambda^*(x) < \infty$;

(C4) $\limsup_{x \to T_Z} \int_x^{T_Z} \frac{\Psi\left(S^Z(t)\right)}{S^Z(t)} d\Lambda^*(t) = 0$;

(C5) $S^X(\cdot)-$ is continuous on $[0, T_Z]$ if $T_Z < \infty$. Otherwise, $S^X(\infty) = \lim_{x \to \infty} S^X(x)$.

At first we state the strong consistency of estimator (16).

Theorem 1. Let conditions (C1)-(C3) are hold. Then for $n \to \infty$

$$\sup_{0 \le x < \infty} \left|S_n^X(x) - S^X(x)\right| \xrightarrow{P} 0.$$

In the paper [4] we generate estimator (16) and theorem 1 for multivariate censoring case.

Now we demonstrate result on asymptotic normality of estimator (16). Introduce the stopped processes $q_n(x) = n^{1/2}\left(S_n^X\left(x \wedge Z^{(n)}\right) - S^X\left(x \wedge Z^{(n)}\right)\right)$, where $a \wedge b = \min(a,b)$. Let $q(x) = e(x)\left[\varphi'\left(S^X(x)\right)\right]^{-1} + \xi\left[\varphi'\left(S^X(T_Z)\right)\right]^{-1}$, where $e(x)$ is mean zero Gaussian process with covariance function

$$A(x_1,x_2) = \int_0^{x_1 \wedge x_2} S^Z(t)\left[\varphi'\left(S^Z(t)\right)\right]^2 d\Lambda(t) +$$
$$+2\int_0^{x_1 \wedge x_2}\int_0^t S^Z(t)\left(1-S^Z(s)\right)\Psi'\left(S^Z(t)\right)\Psi'\left(S^Z(s)\right)d\Lambda(s)d\Lambda(t) +$$
$$+2\int_0^{x_1 \wedge x_2}\int_0^t \varphi'\left(S^Z(s)\right)S^Z(t)\Psi'\left(S^Z(t)\right)d\Lambda(s)d\Lambda(t) +$$
$$+\int_{x_1 \wedge x_2}^{x_1 \vee x_2} S^Z(t)\Psi'\left(S^Z(t)\right)d\Lambda(t)\int_0^{x_1 \wedge x_2}\left[\left(1-S^Z(s)\right)\Psi'\left(S^Z(s)\right) + \varphi'\left(S^Z(s)\right)\right]d\Lambda(s),$$

$x_1 \vee x_2 = \max(x_1, x_2)$, $\xi \overset{D}{=} \mathbb{N}\left(0, \sigma_0^2\right)$ and $\sigma_0^2(x) = \lim_{x \to T_Z} A(x,x)$. Let $C(x) = \lim_{x \to T_Z} A(t,x)$.

Theorem 2. Let conditions (C1)-(C5) are hold, $\sigma_0^2 < \infty$, and for every $x \in [0, T_Z)$: $C(x) < \infty$. Then for $n \to \infty$:

$$q_n(x) \overset{D}{\Rightarrow} q(x) \text{ in } D[0,T_Z].$$

Remark. Consider independent censoring model (i.e. $\{X_k\}$ and $\{Y_k\}$ are mutually independent). In this case in (2) $C(u;v) = uv = C^*(u;v)$, $u,v \in [0,1]$ and hence $\varphi(u) = -\log u$, $u \in [0,1]$ and $\varphi^{[-1]}(t) = \varphi^{-1}(t) = \exp(-t)$, so that

$$S^Z(x) = S^X(x)S^Y(x),\ x \in \overline{R}^+. \tag{17}$$

It is easy to verify that from (12) and (16) respectively we obtain the exponential-hazard estimator

$$\tilde{S}_n^X(x) = \exp\left\{-\int_0^x \frac{I\left(\mathbb{J}_n(t) > 0\right)}{\mathbb{J}_n(t)} d\overline{\mathbb{N}}_n(t)\right\}, \tag{18}$$

and relative-risk power estimator of Abdushukurov (1998) (see[1]):

$$S_n^X(x) = \left[S_n^Z(x)\right]^{R_n(x)},\ R_n(x) = \frac{\Lambda_n(x)}{\Lambda_n^Z(x)}. \tag{19}$$

Note that the estimator (12) is investigated in [3]. Moreover the Zeng-Klein's (1994) copula-graphic estimator is (see [3,6]):

$$\hat{S}_n^X(x) = \varphi^{[-1]}\left[\int_0^x I(\mathbb{J}_n(t) > 0)\left(\varphi\left(\frac{\mathbb{J}_n(t)-1}{n}\right) - \varphi\left(\frac{\mathbb{J}_n(t)}{n}\right)\right)\right] d\overline{\mathbb{N}}_n(t), \tag{20}$$

which in independence model (17) is reduced to well - known Kaplan- Meier product - limit estimator

$$\hat{S}_n^X(x) = \prod_{t \le x}\left\{1 - \frac{d\overline{\mathbb{N}}_n(t)}{\mathbb{J}_n(t)}\right\}. \tag{21}$$

Let $\tilde{S}_n^Y$, S_n^Y and $\hat{S}_n^Y$ are respectively estimators of S^Y of exponential-hazard, relative-risk power and product-limit structures obtained from formulas (18), (19) and (21) by using events $\delta_k = 0$ instead of $\delta_k = 1$. Then we have:

(*a*) $\tilde{S}_n^X(x)\tilde{S}_n^Y(x) = \exp\left\{-\Lambda_n^Z(x)\right\} \ne S_n^Z(x)$ and for $x \ge Z_{(n)} = \max\{Z_k, 1 \le k \le n\}$, $\max\left\{\tilde{S}_n^X(x), \tilde{S}_n^Y(x)\right\} < 1$;

(*b*) $S_n^X(x)\,S_n^Y(x) = S_n^Z(x)$ for all $x \in \overline{R}^+$ and $S_n^X(x) = S_n^Y(x) = 0$, for $x \ge Z_{(n)}$;

(c) $\hat{S}_n^X(x)\hat{S}_n^Y(x) \ne S_n^Z(x)$ and for $x \ge Z_{(n)}$ the estimators $\hat{S}_n^X$ and $\hat{S}_n^Y$ are undefined. Moreover the estimators $\hat{S}_n^X$ and $\hat{S}_n^Y$ require also the condition $P(X_k = Y_k) = 0$, $k = 1,2,...$, which in many practical situations is not hold. Thus only the relative-risk power estimators have identifiability properties with independence censoring model satisfying empirical analogue of equality (17). Analogously a new estimator (16) is more suitable estimator for $\hat{S}_n^X$ than the estimators (12) and (20). In

picture below we demonstrate plots of estimators (12), (16) and (20) of $\hat{S}_n^X$ using well-known Channing House data of size n=97 (see [1]). Here, thin-solid line stands for $\tilde{S}_n^X$, medium-one for $\hat{S}_n^X$ and thick-solid line stands for a new estimator S_n^X. Note that estimate S_n^X is defined in whole line.

III. Estimation of mean residual life function

Let $E(x) = E(x; S^X) = E(X_1 - x / X_1 > x) = = \left(S^X(x)\right)^{-1} \cdot \int_x^{+\infty} S^X(t)\,dt,\ x \in [0, T_X)$, is mean residual life function of r.v. X_1. Consider estimate of $E(x)$:

$$E_n(x) = \begin{cases} E(x; S_n^X),\ x \in \left[0, \mathbb{Z}^{(n)}\right), \\ 0, \qquad x \geq \mathbb{Z}^{(n)}. \end{cases}$$

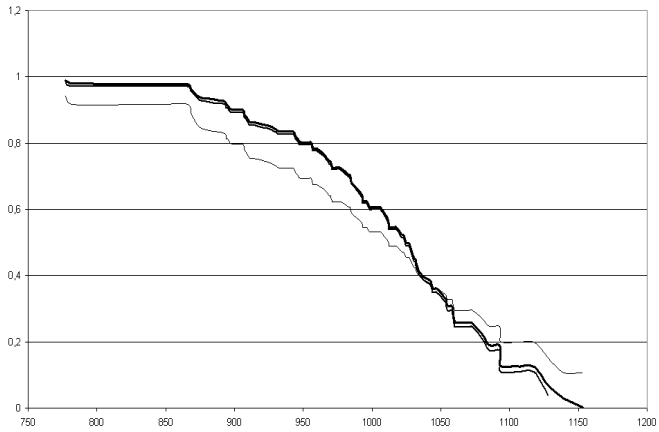
Now we state our result on consistency of E_n with weight function $\chi(\cdot)$. We assume following conditions for χ:

(C6) Function $\chi : [0,1] \to [0, \infty]$ is measurable and, for every $\eta > 0$: $\sup_{u \in [0, 1-\eta]} \{\chi(u)\} < \infty$;

(C7) Function $\chi(u)/(1-u)$ is nondecreasing in a neighborhood of 1;

(C8) $\int_0^{T_X} \left\{ \left(S^X(x)\right)^{-1} \int_x^{T_X} \chi(F(y))\,dy \right\} dF(x) < \infty$.

Theorem 3. Let $\mu = EX_1 < \infty$, conditions (C1)-(C3) and (C6)-(C8) are hold. Then for $n \to \infty$, $\varepsilon_n(F) \xrightarrow{P} 0$.



Picture 1. Plots of estimates $\tilde{S}_n^X$ (thin-solid), $\hat{S}_n^X$ (medium one) and S_n^X (thick-solid) for copula generator $\varphi(u) = \ln^2 u, u \in [0,1]$.

REFERENCES

[1] Abdushukurov A.A. Estimation of unknown distributions by incomplete observations and its properties, LAMBERT Academic Publishing, 301 p. 2011.(In Russian).

[2] Abdushukurov A.A.,Dushatov N.T., Muradov R.S. Estimating of functionals of a multidimensional distribution by censored observations with using copula functions. In: Statistical Methods of Estimation and Hyphoteses Testing. Perm State University. Issue 23, 2011, pp. 36-47. (In Russian).

[3] Li Y., Tiwari R.C., Guha S. Mixture cure survival models with dependent censoring. // J. Royal Statist. Soc. B. v. 69. Part 3. 2007, pp. 285-306.

[4] Muradov R.S., Abdushukurov A.A. () Estimation of multivariate distributions and its mixtures by incomplete data. LAMBERT Academic Publishing, 123 p. 2011.(In Russian).

[5] Nelsen R.B. An introduction to copulas. - Springer, New York. 269 p. 2006.

[6] Rivest L.-P.,Wells M.T.). A martingale approach to the copula-graphic estimator for the survival function under dependent censoring. // J. Multivar.Anal.v.79.2001,pp.138-155.

Flexible Modelling of Cumulative Effects in Survival Analyses

Michal Abrahamowicz
Department of Epidemiology & Biostatistics,
McGill University
Montreal, Quebec, Canada
Michal@epimgh.mcgill.ca

Marie-Pierre Sylvestre
Research Institute of the University of Montreal Hospital Centre
University of Montreal Hospital Centre (CRCHUM)
Montreal, Quebec, Canada
marie-pierre.sylvestre.chum@ssss.gouv.qc.ca

***Abstract*— Time-varying covariates are increasingly used in survival analysis to model the effects of prognostic factors and treatments whose values change during the follow-up. An accurate evaluation of time-varying covariate/exposure effects on survival requires defining an etiologically appropriate 'exposure metric'. This is complicated by the fact that potentially relevant information about the impact of a given time-varying prognostic factor on the current hazard is represented by the vector of its past values, rather than by a scalar. Indeed, most time-varying factors are likely to have cumulative effects on the hazard. Yet, modelling cumulative effects poses two important methodological challenges. Firstly, the impact of past values on the current hazard likely depends on time elapsed since those values were measured. While this calls for a differential weighting of the past values, such weights are generally unknown. Secondly, the form of the dose-response relationship between the prognostic factor and hazard is also typically unknown. We have recently proposed a flexible method for estimating the weight function assigning weights to past values of the prognostic factor, assuming linear dose-response [1]. We now extend this method and model the cumulative effect of past values of a time-varying prognostic factor in a flexible model allowing for simultaneous estimation of weight function and possibly non-linear dose response curve. In our new model, the weighed cumulative exposure (WCE) effect at time τ is a function of past exposure history, represented by the time-dependent vector of past values of a prognostic factor $x(t), at\ 0 < t < \tau$:**

$$(1)\quad WCE(|x(t), t <) \;=\; \sum w(-t) * s[x(t)]$$

where w(τ–t) assigns importance weights to past exposures as the function of time elapsed since exposure, while s[x(t)] represents a smooth dose-response curve describing the relationship between exposure intensity (dose) at a given point in time and the logarithm of the hazard. The estimated WCE is then included as a time-dependent covariate in the Cox's proportional hazards model. Both w(τ–t) and s[x(t)] are modelled using low-dimension cubic regression splines. Quasi-parametric likelihood ratio tests [1] are used to test the non-linearity of s[x(t)], and to compare w(τ–t) against the standard unweighted cumulative dose, as well as to test the Ho of no association between exposure and hazard. To assess the accuracy of the proposed estimates and tests, we present simulation results. To illustrate a real-life application, we use a large long-term cohort study with repeated measures of blood pressure to re-assess its effects on CVD mortality and morbidity.

Keywords: survival analysis; time-varying covariates; splines; cumulative effects.

REFERENCES

[1] Sylvestre, M.P. and Abrahamowicz M. Statistics in Medicine, 28, 27, 3437-3453, 2009.

Block and Basu Bivariate Lifetime Distribution in Presence of Cure Fraction

Jorge Alberto Achcar
Departamento de Medicina Social, FMRP
Universidade de São Paulo
Ribeirão Preto, SP, Brazil
Email: achcar@fmrp.usp.br

Emílio Augusto Coelho-Barros
Departamento de Estatística
Universidade Estadual de Maringá
Maringá, PR, Brazil
Email: eacbarros@uem.br

***Abstract*—In this paper, we introduce a Bayesian analysis for the Block and Basu bivariate lifetime distribution in the presence of covariates and cure fraction. Posterior summaries of interest are obtained using standard MCMC (Markov Chain Monte Carlo) methods and the OpenBugs software. An illustration of the proposed methodology is given for a marrow transplantation for leukemia data set.**

I. Introduction

In some areas of application, especially in medical or engineering studies, we could have two lifetimes T_1 and T_2 associated with each unit. Usually, these data are assumed to be independent, but in many cases, the lifetime of one component could affect the lifetime of the others component. This is the case as an example in the medical area of paired organs like kidneys, lungs, eyes, ears, dental implants among many others. In the literature we observe many papers related to bivarite lifetime parametric models [1]–[9]. One of these bivariate lifetime distributios is very popular: the bivariate exponential distribution introduced by Block and Basu (1974).

The bivariate exponential distribution of Block and Basu with parameters λ_1, λ_2 and λ_3 for the random variables T_1 and T_2 denoting the lifetimes of two components is given by the following joint density function,

$$f(t_1,t_2)=\begin{cases}\frac{\lambda\lambda_1\lambda_{23}}{\lambda_{12}}\exp\left(-\lambda_1 t_1-\lambda_{23}t_2\right) & \text{if } t_1<t_2\\ \frac{\lambda\lambda_2\lambda_{13}}{\lambda_{12}}\exp\left(-\lambda_{13}t_1-\lambda_2 t_2\right) & \text{if } t_1\geq t_2\end{cases} \quad (1)$$

where $\lambda_{12}=\lambda_1+\lambda_2$, $\lambda_{13}=\lambda_1+\lambda_3$, $\lambda_{23}=\lambda_2+\lambda_3$ and $\lambda=\lambda_1+\lambda_2+\lambda_3$.

The joint survival function of the Block and Basu distribution is given by,

$$S(t_1,t_2)=P(T_1>t_1,T_2>t_2)=\begin{cases}S_1(t_1,t_2) & \text{if } t_1<t_2\\ S_2(t_1,t_2) & \text{if } t_1\geq t_2\end{cases} \quad (2)$$

where,

$$S_1(t_1,t_2)=\frac{\lambda}{\lambda_{12}}\exp\left(-\lambda_1 t_1-\lambda_{23}t_2\right)-\frac{\lambda_3}{\lambda_{12}}\exp\left(-\lambda t_2\right),$$

and

$$S_2(t_1,t_2)=\frac{\lambda}{\lambda_{12}}\exp\left(-\lambda_{13} t_1-\lambda_{2}t_2\right)-\frac{\lambda_3}{\lambda_{12}}\exp\left(-\lambda t_1\right).$$

The joint generating function of the Block and Basu distribution is given by,

$$m(s_1,s_2)=\frac{\lambda}{\lambda_{12}(\lambda-s_1-s_2)}\left(\frac{\lambda_1\lambda_{23}}{\lambda_{23}-s_2}-\frac{\lambda_2\lambda_{13}}{\lambda_{13}-s_1}\right). \quad (3)$$

From equation (3), we get the means, variances and the covariance between T_1 and T_2. The correlation coefficient between T_1 and T_2 is given by,

$$\rho_{12}=\frac{\lambda_{12}\left[\left(\lambda_1^2+\lambda_2^2\right)\lambda_3\lambda+\lambda_1\lambda_2\lambda_3^2\right]}{\left[\lambda^2\lambda_{12}^2+\lambda_2\lambda_3\left(2\lambda_1\lambda+\lambda_2\lambda_3\right)\right]^{1/2}\left[\lambda^2\lambda_{12}^2+\lambda_1\lambda_3\left(2\lambda_2\lambda+\lambda_1\lambda_3\right)\right]^{1/2}}. \quad (4)$$

Observe that $0\leq\rho_{12}\leq 1$ and $\rho_{12}=0$ when $\lambda_3=0$ or $\lambda_1=\lambda_2=0$.

We also observe that the marginal distributions for T_1 and T_2 are exponential distributions with parameters λ_1 and λ_2, respectively

Usually we have the presence os censored observations, covariates and all individuals experience the event of interest.

A Bayesian analysis of the Block and Basu bivariate exponential distribution in the presence of covariates and censored data was introduced by Santos and Achcar (2011) using MCMC (Markov Chain Monte Carlo) methods to simulate samples of the joint posterior distribution of interest [11].

In some situations, we could have a fraction of individuals not expecting the occurrence of the event of interest, that is, these individuals are not at risk ("long term survivors or "cured individuals). Different approaches has been presented in the literature to model cure fraction, especially for univariate lifetime data [12]–[19]. Wienke et. al (2006) introduced a model for a cure fraction in bivariate time-to-event data.

In this paper, we develop a Bayesian analysis for bivariate lifetime data considering a generalization of the bivariate exponential distribution of Block and Basu in the presence of censored data, covariates and cure fraction.

Posterior summaries of interest are obtained using standard MCMC methods as the popular Gibbs Sampling algorithm [21] or the Metropolis-Hastings algorithm [22].

The paper is organized as follows: in Section II, we introduce a mixture cured fraction model; in Section III, we present the likelihood function; in Section IV, we present a Bayesian analysis for the model; finally, in Section V, we show

an application with a marrow transplantation for leukemia data set.

II. A Mixture Cured Fraction Model

In the literature we observe different cured fraction models based on mixture or non-mixture models [23]–[26].

The mixture cure rate model, also known as standard cure rate model, considering univariate lifetimes T, assumes that the studied populatin is a mixture of suscetible individuals, who experience the event of interest and non-suscetible individuals tha will never experiment this event, that is, these individuals are not at risk with respect to the event of interest and they are considered immune, non-suscetible or cured.

The standard cure fraction model [25] assumes that a given fraction p in the population is cured with respect to the specific cause of death (or failure) while the remainning $(1-p)$ fraction of the individuals is not cured, with a survival function $S(t) = P(T > t)$ for all population given by,

$$S(t) = p + (1-p) S_0(t) \tag{5}$$

where $0 \leq p \leq 1$ and $S_0(t)$ denotes a proper survival function for the uncured group.

Assuming the mixture model (1), the probability density function for $\mathbf{t}$ is given (from $-dS(t)/dt$) by,

$$f(t) = (1-p) f_0(t) \tag{6}$$

where $f_0(t)$ is the probability density function for the suscetible individuals.

A generalization of the mixture model (5) for bivariate lifetimes T_1 and T_2 is given by [20],

$$\begin{aligned} S(t) &= P(T_1 > t_1, T_2 > t_2) \\ &= \phi_{11} S_{110}(t_1, t_2) + \phi_{10} S_{10}(t_1) + \phi_{01} S_{20}(t_2) + \phi_{00} \end{aligned} \tag{7}$$

where $S_{110}(t_1, t_2)$ is the joint survival function for the suscetible individuals; $S_{10}(t_1) = P(T_1 > t_1)$ is the marginal survival function for T_1; $S_{20}(t_2) = P(T_1 > t_2)$ is the marginal survival function for T_2; $\phi_{11} = P(V_1 = 1, V_2 = 1)$, $\phi_{10} = P(V_1 = 1, V_2 = 0)$, $\phi_{01} = P(V_1 = 0, V_2 = 1)$, $\phi_{00} = P(V_1 = 0, V_2 = 0)$, $\phi_{11} + \phi_{10} + \phi_{01} + \phi_{00} = 1$; V_1 and V_2 are binary variables such that $V_1 = 1$ if the individual is suscetible for the lifetime T_1 and $V_1 = 0$ if the individual is immune; in the same way, $V_2 = 1$ if the individual is suscetible for the lifetime T_2 and $V_2 = 0$ if the individual is immune.

In this paper, we assume that the joint distribution function for the lfetimes T_1 and T_2 of suscetible individuals is given by Block and Basu distribution defined by the joint probability density function $f_{110}(t_1, t_2)$ given by (1) and joint survival function $S_{110}(t_1, t_2)$ given by (2). Observe that, in this case, $S_{10}(t_1)$ and $S_{20}(t_2)$ in (7) are survival functions of exponential distributions with parameters λ_1 and λ_2, respectively.

III. The Likelihood Function

Let us assume n pairs of lifetimes T_1 and T_2 that can be censored where censoring is independent of the lifetimes divided into four classes:

C_1 : both lifetimes t_{1i} and t_{2i} are observed, $i = 1, 2, \ldots, n$;
C_2 : t_{1i} is a lifetime and t_{2i} is a censored time (that is, we only know that $T_{2i} \geq t_{2i}$);
C_3 : t_{1i} is a censoring time and t_{2i} is a lifetime;
C_4 : both t_{1i} and t_{2i} are censoring times.

The likelihood function [27] is given by,

$$\begin{aligned} L &= \prod_{i \in C_1} f(t_{1i}, t_{2i}) \prod_{i \in C_2} \left[-\frac{\partial S(t_{1i}, t_{2i})}{\partial t_{1i}} \right] \times \\ &\quad \prod_{i \in C_3} \left[-\frac{\partial S(t_{1i}, t_{2i})}{\partial t_{2i}} \right] \prod_{i \in C_4} S(t_{1i}, t_{2i}), \end{aligned} \tag{8}$$

where $f(t_{1i}, t_{2i}) = \frac{\partial^2 S(t_{1i}, t_{2i})}{\partial t_{1i} \partial t_{2i}}$, $S(t_{1i}, t_{2i})$ is given by (7).

The contributions for the likelihood function (8) are given by,

(i) If $i \in C_1$,

$$f(t_{1i}, t_{2i}) = \phi_{11} f_{110}(t_{1i}, t_{2i}), \tag{9}$$

where

$$f_{110}(t_{1i}, t_{2i}) = \frac{\lambda \lambda_1 \lambda_{23}}{\lambda_{12}} \exp(-\lambda_1 t_{1i} - \lambda_{23} t_{2i}) \text{ if } t_{1i} < t_{2i},$$

and

$$f_{110}(t_{1i}, t_{2i}) = \frac{\lambda \lambda_2 \lambda_{13}}{\lambda_{12}} \exp(-\lambda_{13} t_{1i} - \lambda_2 t_{2i}) \text{ if } t_{1i} \geq t_{2i}.$$

(ii) If $i \in C_2$,

$$-\frac{\partial S(t_{1i}, t_{2i})}{\partial t_{1i}} = \phi_{11} S'_{t_1}(t_{1i}, t_{2i}) + \phi_{10} f_{10}(t_{1i}), \tag{10}$$

where

$$S'_{t_1}(t_{1i}, t_{2i}) = \frac{\lambda \lambda_1}{\lambda_{12}} \exp(-\lambda_1 t_{1i} - \lambda_{23} t_{2i}) \text{ if } t_{1i} < t_{2i},$$

and

$$\begin{aligned} S'_{t_1}(t_{1i}, t_{2i}) &= \frac{\lambda \lambda_{13}}{\lambda_{12}} \exp(-\lambda_{13} t_{1i} - \lambda_2 t_{2i}) - \\ &\quad \frac{\lambda \lambda_3}{\lambda_{12}} \exp(-\lambda t_{1i}) \qquad \text{if } t_{1i} \geq t_{2i}, \end{aligned}$$

and

$$f_{10}(t_{1i}) = \lambda_1 \exp(-\lambda_1 t_{1i}).$$

(iii) If $i \in C_3$,

$$-\frac{\partial S(t_{1i}, t_{2i})}{\partial t_{2i}} = \phi_{11} S'_{t_2}(t_{1i}, t_{2i}) + \phi_{01} f_{20}(t_{2i}), \tag{11}$$

where

$$\begin{aligned} S'_{t_2}(t_{1i}, t_{2i}) &= \frac{\lambda \lambda_{23}}{\lambda_{12}} \exp(-\lambda_1 t_{1i} - \lambda_{23} t_{2i}) - \\ &\quad \frac{\lambda \lambda_3}{\lambda_{12}} \exp(-\lambda t_{2i}) \qquad \text{if } t_{1i} < t_{2i}, \end{aligned}$$

and

$$S'_{t_2}(t_{1i}, t_{2i}) = \frac{\lambda\lambda_2}{\lambda_{12}} \exp(-\lambda_{13} t_{1i} - \lambda_2 t_{2i}) \text{ if } t_{1i} \geq t_{2i},$$

and

$$f_{20}(t_{2i}) = \lambda_2 \exp(-\lambda_2 t_{2i}).$$

(iv) If $i \in C_4$,

$$\begin{aligned} S(t_{1i}, t_{2i}) &= \phi_{11} S_{110}(t_{1i}, t_{2i}) + \phi_{10} S_{10}(t_{1i}) + \\ &\quad \phi_{01} S_{20}(t_{2i}) + \phi_{00}, \end{aligned} \tag{12}$$

where $S_{110}(t_{1i}, t_{2i}) = S_1(t_{1i}, t_{2i})$ if $t_{1i} < t_{2i}$, $S_{110}(t_{1i}, t_{2i}) = S_2(t_{1i}, t_{2i})$ if $t_{1i} \geq t_{2i}$ (see (2)), that is,

$$\begin{aligned} S_1(t_{1i}, t_{2i}) &= \frac{\lambda}{\lambda_{12}} \exp(-\lambda_1 t_{1i} - \lambda_{23} t_{2i}) - \\ &\quad \frac{\lambda_3}{\lambda_{12}} \exp(-\lambda t_{2i}), \end{aligned}$$

$$\begin{aligned} S_2(t_{1i}, t_{2i}) &= \frac{\lambda}{\lambda_{12}} \exp(-\lambda_{13} t_{1i} - \lambda_2 t_{2i}) - \\ &\quad \frac{\lambda_3}{\lambda_{12}} \exp(-\lambda t_{1i}), \end{aligned}$$

$S_{10}(t_{1i}) = \exp(-\lambda_1 t_{1i})$, and $S_{20}(t_{2i}) = \exp(-\lambda_2 t_{2i})$, $i = 1, \ldots, n$.

IV. A Bayesian Analysis For The Model

For a Bayesian analysis, we assume different cases of the model introduced in Section III.

Model 1: Let us assume the Block and Basu distribution (1) not considering the presence of covariates and cure fraction, that is, $\phi_{11} = 1$, $\phi_{10} = \phi_{01} = \phi_{00} = 0$. In this case we assume Gamma priors for λ_j, that is,

$$\lambda_j \sim Gamma(a_j, b_j) \tag{13}$$

for $j = 1, 2, 3$; $Gamma(a, b)$ denotes a gamma distribution with mean a/b and variance a/b^2. We assume known hyperparemeters a_j and b_j and prior independence among the random quantities λ_1, λ_2 and λ_3.

Model 2: Let us assume the Block and Basu distribution (1) in the presence of covariates but not considering the presence of cure fraction, that is, $\phi_{11} = 1$, $\phi_{10} = \phi_{01} = \phi_{00} = 0$. In this model, we assume the following regression model,

$$\begin{aligned} \lambda_{1i} &= \alpha_1 \exp(\beta'_{\mathbf{1}} \mathbf{x_i}) \\ \lambda_{2i} &= \alpha_2 \exp(\beta'_{\mathbf{2}} \mathbf{x_i}) \end{aligned} \tag{14}$$

where $\mathbf{x_i} = (x_{1i}, \ldots, x_{ki})'$ is a vector of covariates and $\beta_{\mathbf{l}} = (\beta_{l1}, \ldots, \beta_{lk})$, $l = 1, 2$ denotes a vector regression parameters. In this model, we assume the folowing prior distributions,

$$\begin{aligned} \theta_j &\sim U(c_j, d_j), \; j = 1, 2 \\ \theta_3 &\sim U(c_3, d_3) \\ \beta_{ls} &\sim U(e_{ls}, f_{ls}) \end{aligned} \tag{15}$$

where θ_j, $j = 1, 2$ is a reparametrization of α_j given by $\theta_j = \log(\alpha_j)$; $\theta_3 = \log(\lambda_3)$; $U(c, d)$ denotes an uniform distribution defined in the interval (c, d); c_j, d_j, e_{ls} and f_{ls} are known hyperparameters, $j = 1, 2, 3$; $l = 1, 2$; $s = 1, \ldots, k$.

Model 3: In this model we assume the Block Basu distribution (1) in the presence of cure fraction but not considering covariates. For this model, we assume the same gamma priors (13) for the random quantities λ_j, $j = 1, 2, 3$ and a Dirichlet prior distribution for the incidence parameters ϕ_{11}, ϕ_{10}, ϕ_{01}, and ϕ_{00}, where $\phi_{11} + \phi_{10} + \phi_{01} + \phi_{00} = 1$, with density,

$$\pi(\phi \mid \mathbf{a}) = \frac{1}{B(\mathbf{a})} \prod_{k=1}^{4} \phi_k^{a_k - 1} \tag{16}$$

where $\phi_1 = \phi_{11}$, $\phi_2 = \phi_{10}$, $\phi_3 = \phi_{01}$, $\phi_4 = \phi_{00}$ and $B(\mathbf{a})$ is a normalizing constant given by,

$$B(\mathbf{a}) = \frac{\prod_{k=1}^{4} \Gamma(a_k)}{\Gamma\left(\sum_{k=1}^{4} a_k\right)} \tag{17}$$

where $\mathbf{a} = (a_1, a_2, a_3, a_4)$ is a known vector of hyperparameters.

Model 4: In this model we assume the Block Basu distribution (1) in the presence of cure fraction and covariates. In this case, we assume the same regression models (14) and the same prior for θ_1, θ_2 and θ_3 ($\theta_j = \log(\alpha_j)$, $j = 1, 2$, $\theta_3 = \log(\lambda_3)$) given in (15) $l = 1, 2$, $s = 1, \ldots, k$. For the incidence parameters ϕ_{11}, ϕ_{10}, ϕ_{01}, and ϕ_{00}, we also assume the same Dirichlet prior distribution given by (16).

Posterior summaries of interest for each model are simulated using standard MCMC methods. We also have used the software Openbugs [28] available at the site *www.openbugs.info*.

V. An Application: Bone Marrow Transplantation For Leukemia

In this aplication, we consider a data set introduced by Klein and Moeschberger (1997, page 464). This data set is related to bone marrow transplants used as standard treatments for acute leukemia. Associated to the recovery of the patients, there are many risk factors known at the time of transplantation, such as patient and/or donor age and sex, the stage of initial disease, the time from prognosis to transplantation, amog many others.

The final prognosis may change depending on the patient post transplantation history and some events can be developed such as development of accute or chronic graft-versus-host disease (GVHD), return of the platelet count to normal levels, or development of infections. Transplantation is considered as a failure when leukemia returns (relapse) or when the patient dies in remission.

In this study, 137 patients with accute myeloctic leukemia (AML) and accute lymphoblastic leukemia (ALL) received a combination of 16 mg/Kg of oral Busulfan (BU) and 120 mg/Kg of intravenous cyclopho sphamide (Cy) (99 AML and 38 ALL patients).

In the analysis considered here, we assume as lifetimes, the times (in days) to accute-versus-host disease (TA) with 111 censored observations and 26 not-censured observations and the time (in days) to chronic graft-versus-host disease (TC) with 76 censored observations an 61 not-censored observations and some covariates as patient age (in years); donor age (in years); patient sex; donor sex; patient CMV (cytomegalovirus immune status); donor CMV and waiting time to transplantation.

Observe that among teh censored observations, some patients will never have the occurence ot the event of interest, that is, they are immunes.

In the analysis of this data set, we assume the four models introduced in Section IV. Firt os all, we assume the Block and Basu bivariate exponential distribution (1) not considering the presence of covariates and cure fraction ("Model 1" introduced in Section IV) to analyse the bone marrow transplantation data. For this model, we consider gamma prior distribution (13) for λ_j, $j = 1, 2, 3$ where $a_1 = a_2 = a_3 = 1$, $b_1 = b_2 = 100$ and $b_3 = 1000$.

Considering the Openbugs software, we simulate $5,010,00$ Gibbs samples, from the joint posterior distribution of interest, from which we discarded the first $10,000$ samples as a "burn-in-sample" to eliminate the effect ot the initial values in the simulation algorithm; after this "burn-in-sample" period, we selected every 5000^{th} sample, to have approximately uncorrelated samples, which totalizes a final sample of $1,000$ Gibbs samples to be used to obtain the posterior summaries of interest. Convergence of the Gibbs Sampling algorithm was monitored using standard graphical methods, as the traceplots of the simulated samples. In Table I, we have the posterior summaries os interest assuming "Model 1".

TABLE I
POSTERIOR SUMMARIES ("MODEL 1")

Parameter	Mean	S.D	95% Credible Interval
cov_{12}	76.11	58.3	$(2.642; 206.9)$
λ_1	0.000265	0.000054	$(0.000176; 0.0003814)$
λ_2	0.000937	0.000121	$(0.000727; 0.001169)$
λ_3	0.000037	0.000033	$(0.000002; 0.0001147)$
$mean_1$	3537.0	720.4	$(2415.0; 5118.0)$
$mean_2$	1049.0	128.6	$(840.2; 1320.0)$
ρ_{12}	0.000023	0.000019	$(0.000001; 0.00006817)$
sd_1	3481.0	725.8	$(2348.0; 5066.0)$
sd_2	1047.0	127.8	$(839.9; 1318.0)$

In Table I, cov_{12} denotes the covariance between TA and TC (or T_1 and T_2); $mean_1$ and $mean_2$ denote the means of T_1 and T_2; ρ_{12} denote the correlation between T_1 and T_2; sd_1 and sd_2 denote the standard deviation of T_1 and T_2.

Assuming the Block and Basu bivariate exponential distribution in presence of covariates but not considering cure fraction ("Model 2"), we assume de regression model 14 given by,

$$\begin{aligned}\lambda_{1i} &= \alpha_1 \exp\left[\beta_{11}\left(z_{1i}-\bar{z}_1\right)+\beta_{12}\left(z_{2i}-\bar{z}_2\right)+\beta_{13}z_{3i}+\right.\\ &\quad \left.\beta_{14}z_{4i}+\beta_{15}z_{5i}+\beta_{16}z_{6i}+\beta_{17}\left(z_{7i}-\bar{z}_7\right)\right]\end{aligned} \tag{18}$$

$$\begin{aligned}\lambda_{2i} &= \alpha_2 \exp\left[\beta_{21}\left(z_{1i}-\bar{z}_1\right)+\beta_{22}\left(z_{2i}-\bar{z}_2\right)+\beta_{23}z_{3i}+\right.\\ &\quad \left.\beta_{24}z_{4i}+\beta_{25}z_{5i}+\beta_{26}z_{6i}+\beta_{27}\left(z_{7i}-\bar{z}_7\right)\right]\end{aligned}$$

where $i = 1, \ldots, n$ $(n = 137)$ z_{1i} denotes the patient age; z_{2i} denotes the donor age; z_{3i} denotes the patient sex ($1 =$ male; $0 =$ female); z_{4i} denotes the donor sex ($1 =$ male; $0 =$ female); z_{5i} denotes the patient CMV status ($1 =$ CMV positive; $0 =$ CMV negative); z_{6i} denotes the donor CMV status ($1 =$ CMV positive; $0 =$ CMV negative) and z_{7i} denotes the waiting time to transplant in days; $\bar{z}_1$, $\bar{z}_2$ and $\bar{z}_7$ are, respectively, the samples averages of z_{1i}, z_{2i} and z_{7i}.

For a Bayesian analysis of "Model 2", we assume the prior distributions (15), with the following hyperparameters values: $c_1 = c_2 = -10$; $d_1 = d_2 = 0$; $c_3 = -20$ and $d_3 = 0$; $e_{11} = e_{12} = e_{13} = e_{14} = e_{16} = e_{21} = e_{22} = e_{23} = e_{24} = e_{25} = e_{26} = -0.1$; $f_{11} = f_{12} = f_{13} = f_{14} = f_{16} = f_{21} = f_{22} = f_{23} = f_{24} = f_{25} = f_{26} = 0.1$; $e_{15} = e_{17} = e_{27} = -0.002$ and $f_{15} = f_{17} = f_{27} = 0.002$.

In Table II, we have the posterior summaries of interest assuming $1,000$ simulated Gibbs samples ("burn-in-sample"=$2,000$ and choosing every 20^{th} simulated sample from $20,000$ Gibbs samples) using the Openbugs software.

TABLE II
POSTERIOR SUMMARIES ("MODEL 2")

Parameter	Mean	S.D	95% Credible Interval
α_1	0.0002358	0.0000551	$(0.0001429; 0.0003589)$
α_2	0.00101	0.0001457	$(0.0007531; 0.001316)$
β_{11}	0.04143	0.02736	$(-0.01586; 0.08936)$
β_{12}	0.04075	0.02778	$(-0.01503; 0.09016)$
β_{13}	-0.006977	0.05765	$(-0.09507; 0.09334)$
β_{14}	-0.008019	0.05752	$(-0.09495; 0.09123)$
β_{15}	0.0000567	0.001165	$(-0.001869; 0.001913)$
β_{16}	0.0131	0.05654	$(-0.09338; 0.09693)$
β_{17}	0.0002818	0.0004884	$(-0.0008465; 0.001124)$
β_{21}	-0.02514	0.02282	$(-0.06819; 0.01801)$
β_{22}	0.05112	0.02222	$(0.006327; 0.09387)$
β_{23}	-0.01138	0.05731	$(-0.09451; 0.09481)$
β_{24}	-0.01943	0.05637	$(-0.09799; 0.09056)$
β_{25}	0.01896	0.05558	$(-0.08954; 0.09745)$
β_{26}	-0.01012	0.05632	$(-0.09701; 0.09074)$
β_{27}	0.0005917	0.000343	$(-0.0001128; 0.001217)$
λ_3	0.0000038	0.0000107	$(0.000000002; 0.00003339)$

From the results of Table II, we observe that only the covariate donor age has some significative effect on the parameter λ_2, since the 95% credible interval for β_{22} does not contain the zero value. For all the other covariate, zero is included in the credible intervals for the associated regression parameters.

Assuming "Model 3" introduced in Section IV, that is, the Block and Basu distribution (1) in the presence of cure fraction but not considering the presence of covariates, and the same gamma prior distributions (13) for the parameters λ_j, $j = 1, 2, 3$ considered for "Model 1" and the Dirichlet

prior distribution (17) for ϕ_{11}, ϕ_{10}, ϕ_{01}, and ϕ_{00} with hyperparameter values $a_1 = a_2 = a_3 = a_4 = 1$ (non-informative priors for the incidence parameters), we have in Table III, the posterior summaries of interest.

In this simulation approach used for "Model 3" also considering the OpenBugs software, we simulated a "burn-in-sample" of size $10,000$; after this "burn-in-sample" period, we simulated another 1000 Gibbs samples taking every 500^{th} simulated sample.

TABLE III
POSTERIOR SUMMARIES ("MODEL 3")

Parameter	Mean	S.D	95% Credible Interval
λ_1	0.02153	0.00531	$(0.01204; 0.0324)$
λ_2	0.00546	0.00075	$(0.004136; 0.007006)$
λ_3	0.00070	0.00064	$(0.000026; 0.002466)$
ϕ_{00}	0.4252	0.04423	$(0.3392; 0.5127)$
ϕ_{01}	0.3641	0.04263	$(0.283; 0.4549)$
ϕ_{10}	0.05049	0.02369	$(0.01333; 0.1046)$
ϕ_{11}	0.1602	0.03504	$(0.09553; 0.2351)$

Assuming "Model 4" (Block and Basu distribution in the presence of covariates and cure fraction), the regression model (18) and priors (15) with hyperparameters values: $c_1 = c_2 = -10$; $c_3 = -20$; $d_1 = d_2 = d_3 = 0$; $e_{11} = -0.1$; $f_{11} = 0.15$; $e_{12} = e_{13} = e_{14} = e_{15} = e_{16} = e_{21} = e_{22} = e_{23} = e_{24} = e_{25} = e_{26} = -1$; $f_{12} = f_{13} = f_{14} = f_{15} = f_{16} = f_{21} = f_{22} = f_{23} = f_{24} = f_{25} = f_{26} = 1$; $e_{17} = -0.005$; $f_{17} = 0.005$; $e_{27} = -0.001$; $f_{27} = 0.001$, we have in Table IV, the posterior summaries of interest ("burn-in-sample"=2000; 1000 simulated Gibbs samples taking every 30^{th} sample).

TABLE IV
POSTERIOR SUMMARIES ("MODEL 4")

Parameter	Mean	S.D	95% Credible Interval
α_1	0.0072	0.00678	$(0.001439; 0.02667)$
α_2	0.005168	0.00155	$(0.002517; 0.008701)$
β_{11}	0.0364	0.04418	$(-0.05199; 0.121)$
β_{12}	0.1111	0.06422	$(-0.03762; 0.2202)$
β_{13}	-0.3092	0.4619	$(-0.9713; 0.6953)$
β_{14}	-0.135	0.5087	$(-0.9468; 0.8705)$
β_{15}	-0.09674	0.5285	$(-0.9364; 0.8786)$
β_{16}	0.4309	0.4414	$(-0.6561; 0.9829)$
β_{17}	0.000262	0.00066	$(-0.001233; 0.001448)$
β_{21}	-0.01254	0.02241	$(-0.05836; 0.03144)$
β_{22}	-0.00255	0.02176	$(-0.0448; 0.03978)$
β_{23}	0.08944	0.2926	$(-0.495; 0.6676)$
β_{24}	-0.00107	0.2891	$(-0.5404; 0.5365)$
β_{25}	0.003001	0.3074	$(-0.582; 0.6227)$
β_{26}	-0.08789	0.3053	$(-0.6612; 0.4964)$
β_{27}	0.000068	0.00035	$(-0.0006626; 0.0007316)$
λ_3	0.000150	0.00048	$(0.000000003; 0.00141)$
ϕ_{00}	0.3295	0.07036	$(0.1977; 0.4663)$
ϕ_{10}	0.05861	0.03069	$(0.01186; 0.128)$
ϕ_{01}	0.4152	0.05881	$(0.3024; 0.54)$
ϕ_{11}	0.1967	0.04769	$(0.112; 0.2976)$

From the results of Table IV, we conclude that the covariates associated in the model do not show significative resuts since zero is included in all 95% credible intervals for the regression parameters β_{ls}, $l = 1, 2$, $s = 1, \ldots, 7$.

For the discrimination among the different models, we could use a Bayesian criterion given by DIC (Deviance Information Criterium) introduced by Spiegelhalter et. al (2002), and given automatically by OpenBugs. In Table V, we have the Monte Carlo estimates for DIC assuming the different proposed models. Smaller values of DIC indicates better models.

From the results of Table V, we conclude that "Model 3" (Block and Basu distribution in presence of cure fraction) is better fitted by the data.

TABLE V
DIC FOR THE MODELS.

Model	DIC
Model 1	1457
Model 2	1445
Model 3	1290
Model 4	1305

REFERENCES

[1] B. C. Arnold and D. Strauss, "Bivariate distributions with exponential conditionals," *Journal of the American Statistical Association*, vol. 83, no. 402, pp. 522–527, 1988. [Online]. Available: http://links.jstor.org/sici?sici=0162-1459(198806)83:402¡522:BDWEC¿2.0.CO;2-T&origin=MSN

[2] H. W. Block and A. P. Basu, "A continuous bivariate exponential extension," *Journal of the American Statistical Association*, vol. 69, pp. 1031–1037, 1974.

[3] F. Downton, "Bivariate exponential distributions in reliability theory," *Journal of the Royal Statistical Society, Series B, Methodological*, vol. 32, pp. 408–417, 1970.

[4] J. E. Freund, "A bivariate extension of the exponential distribution," *Journal of the American Statistical Association*, vol. 56, pp. 971–977, 1961.

[5] E. J. Gumbel, "Bivariate exponential distributions," *Journal of the American Statistical Association*, vol. 55, pp. 698–707, 1960.

[6] A. G. Hawkes, "A bivariate exponential distribution with applications to reliability," *Journal of the Royal Statistical Society, Series B, Methodological*, vol. 34, pp. 129–131, 1972.

[7] P. Hougaard, "A class of multivariate failure time distributions," *Biometrika*, vol. 73, no. 3, pp. 671–678, 1986.

[8] A. W. Marshall and I. Olkin, "A multivariate exponential distribution," *J. Amer. Statist. Assoc.*, vol. 62, pp. 30–44, 1967.

[9] S. K. Sarkar, "A continuous bivariate exponential distribution," *Journal of the American Statistical Association*, vol. 82, no. 398, pp. 667–675, 1987. [Online]. Available: http://links.jstor.org/sici?sici=0162-1459(198706)82:398¡667:ACBED¿2.0.CO;2-E&origin=MSN

[10] C. A. dos Santos and J. A. Achcar, "A Bayesian analysis for the Block and Basu bivariate exponential distribution in the presence of covariates and censored data," *J. Appl. Stat.*, vol. 38, no. 10, pp. 2213–2223, 2011. [Online]. Available: http://dx.doi.org/10.1080/02664763.2010.545372

[11] J. A. Achcar and R. A. Leandro, "Use of Markov chain Monte Carlo methods in a Bayesian analysis of the block and Basu bivariate exponential distribution," *Ann. Inst. Statist. Math.*, vol. 50, no. 3, pp. 403–416, 1998. [Online]. Available: http://dx.doi.org/10.1023/A:1003582409664

[12] B. Yu, R. C. Tiwari, and K. Z. Cronin, "Cure fraction estimation from the mixture cure models for grouped survival times," *Statistics in Medicine*, vol. 23, pp. 1733–1747, 2004.

[13] V. T. Farewell, "The use of mixture models for the analysis of survival data with long-term survivors," *Biometrics*, vol. 38, pp. 1041–1046, 1982.

[14] J. W. Gamel, I. W. Mclean, and S. H. Rosenberg, "Proportion cured and mean log-survival time as functions of tumor size," *Statistical in Medicine*, vol. 9, pp. 999–1006, 1999.

[15] K. Yamaguchi, "Accelerated failure-time regression model with a regression model for the survining fraction: an application to the analysis of permanent employment in japan," *Journal of the American Statistical Association*, vol. 87, pp. 284–292, 1992.

[16] V. G. Cancho and H. Bolfarine, "Modeling the presence of immunes by using the exponentiated-Weibull model," *Journal of Applied Statistics*, vol. 28, no. 6, pp. 659–671, 2001. [Online]. Available: http://dx.doi.org/10.1080/02664760120059200

[17] W. Dunsmuir, R. Tweedie, L. Flack, and K. Mengersen, "Modeling the transitions between employement states for young australian," *Australians Journal of Statistics*, vol. 31, no. A, pp. 165–196, 1989.
[18] J. M. G. Taylor, "Semiparametric estimation in failure time mixture models," *Biometrics*, vol. 51, pp. 899–907, 1995.
[19] N. Kannan, D. Kundu, P. Nair, and R. C. Tripathi, "The generalized exponential cure rate model with covariates," *J. Appl. Stat.*, vol. 37, no. 9-10, pp. 1625–1636, 2010. [Online]. Available: http://dx.doi.org/10.1080/02664760903117739
[20] A. Wienke, I. Locatelli, and A. I. Yashin, "The modelling of a cure fraction in bivariate time-to-event data," *Austrian Journal of Statistics*, vol. 35, no. 1, pp. 67–76, 2006.
[21] A. E. Gelfand and A. F. M. Smith, "Sampling-based approaches to calculating marginal densities," *J. Amer. Statist. Assoc.*, vol. 85, no. 410, pp. 398–409, 1990. [Online]. Available: http://links.jstor.org/sici?sici=0162-1459(199006)85:410¡398:SATCMD¿2.0.CO;2-3&origin=MSN
[22] S. Chib and E. Greenberg, "Undestanding the metropolis-hastings algorithm," *The American Statistician*, vol. 49, no. 4, pp. 327–335, 1995.
[23] A. Y. Yakovlev, A. D. Tsodikov, and B. Asselain, *Stochastic models of tumor latency and their biostatistical applications*. Singapore: World Scientific, 1996.
[24] A. D. Tsodikov, J. G. Ibrahim, and A. Y. Yakovlev, "Estimating cure rates from survival data: an alternative to two-component mixture models," *J. Amer. Statist. Assoc.*, vol. 98, no. 464, pp. 1063–1078, 2003. [Online]. Available: http://dx.doi.org/10.1198/0162145030000001007
[25] R. A. Maller and X. Zhou, *Survival analysis with long-term survivors*, ser. Wiley Series in Probability and Statistics: Applied Probability and Statistics. Chichester: John Wiley & Sons Ltd., 1996.
[26] P. C. Lambert, J. R. Thompson, C. L. Weston, and P. W. Dickman, "Estimating and modeling the cure fraction in population-based cancer survival analysis," *Biostatistics*, vol. 8, no. 3, pp. 576–594, 2007.
[27] J. F. Lawless, *Statistical models and methods for lifetime data*. New York: John Wiley and Sons, 1982.
[28] D. Lunn, D. Spiegelhalter, A. Thomas, and N. Best, "The BUGS project: evolution, critique and future directions," *Stat. Med.*, vol. 28, no. 25, pp. 3049–3067, 2009. [Online]. Available: http://dx.doi.org/10.1002/sim.3680
[29] J. P. Klein and M. L. Moeschberger, *Survival Analysis: Techniques for Censored and Truncated Data*. Berlin: Springer-Verlag Telos, 1997.
[30] D. J. Spiegelhalter, N. G. Best, and A. Vander Linde, "A bayesian measure of model complexity and fit (with discussion)," *Journal of the Royal Statistical Society, B*, vol. 64, pp. 583–639, 2000.

Nonparametric prediction of system failure time using partially known signatures

Abdullah Al-nefaiee
Department of Mathematical Sciences
Durham University, UK
Email: a.h.al-nefaiee@durham.ac.uk

Frank P.A. Coolen
Department of Mathematical Sciences
Durham University, UK
Email: frank.coolen@durham.ac.uk

***Abstract*—System signatures provide a powerful framework for reliability assessment for systems consisting of exchangeable components. The use of signatures in nonparametric predictive inference has been presented and leads to lower and upper survival functions for the system failure time, given failure times of tested components. However, deriving the system signature is computationally complex. This paper presents how limited information about the signature can be used to derive bounds on such lower and upper survival functions and related inferences.**

I. Introduction

System signatures are a powerful tool for quantifying reliability of coherent systems consisting of exchangeable components [5]. Consider a system consisting of m exchangeable components, it could be said that such components are all 'of the same type'. Throughout this paper it is assumed that the system is coherent. Let the random failure time of the system be T_S, and let $T_{j:m}$ be the j-th order statistic of the m random component failure times for $j = 1, \ldots, m$, with $T_{1:m} \leq T_{2:m} \leq \ldots \leq T_{m:m}$. The system's signature is the m-vector q with j-th component $q_j = P(T_S = T_{j:m})$, the probability that the system fails at the moment of the j-th component failure. Assume that $\sum_{j=1}^{m} q_j = 1$, so the system functions if all components function, has failed if all components have failed, and system failure only occurs at times of component failures. The survival function of the system failure time is

$$P(T_S > t) = \sum_{j=1}^{m} q_j P(T_{j:m} > t) \qquad (1)$$

Recently, the use of signatures for nonparametric predictive inference (NPI) for system reliability has been presented [1]. In NPI for system reliability, lower and upper survival functions are derived for the system's failure time, these reflect the limited knowledge about reliability of the components, using only the information from component tests. A brief overview of the results in [1] is given in Section II.

Derivation of the signature is not straightforward, even for relatively basic systems. For specific inferences it may not be necessary to compute the exact signature. If computation of signatures is stopped before the exact signature is derived, one typically has bounds for the probabilities q_j. We explore the use of such bounds in NPI, leading to lower and upper bounds for the NPI lower and upper survival functions. For specific inferences, these bounds may already be conclusive, meaning that no further computation is needed. The basic results for the use of such bounds in NPI are presented in Section III, together with explanation of the possible use of information on signatures for subsystems and comparison of the failure times of two systems. Examples in Section IV illustrate the results in this paper.

II. Using signatures in NPI

Suppose that in a test of n components, exchangeable with those in the system considered, the observed failure times were $t_1 < t_2 < \ldots < t_n$. For ease of notation, define $t_0 = 0$ and $t_{n+1} = \infty$. These n observations partition the non-negative real-line into $n+1$ intervals $I_i = (t_{i-1}, t_i)$ for $i = 1, \ldots, n+1$. Consider reliability of a system with m components, so interest is in the m failure times of those components, say $T_1, \ldots, T_m$. The test data and the future observations $T_1, \ldots, T_m$ are linked via repeated use of the assumption $A_{(n)}$ [2]. The order statistics of the m future observations $T_1, \ldots, T_m$ are denoted by $T_{1:m} \leq T_{2:m} \leq \ldots \leq T_{m:m}$. The following probabilities hold for $T_{j:m}$, for $j = 1, \ldots, m$ and for $i = 1, \ldots, n+1$ [2]

$$P(T_{j:m} \in I_i) = \binom{i+j-2}{i-1} \binom{n-i+1+m-j}{n-i+1} \binom{n+m}{n}^{-1}$$

These probabilities lead to the following NPI lower and upper survival functions for $T_{j:m}$, which are the sharpest bounds for the probability of the event $T_{j:m} > t$ that can be justified without further assumptions. The NPI lower survival function for $T_{j:m}$ is, for $t \in (t_{i-1}, t_i]$

$$\underline{S}_{T_{j:m}}(t) = \underline{P}(T_{j:m} > t) = \sum_{l=i+1}^{n+1} P(T_{j:m} \in I_l)$$

and the NPI upper survival function is, for $t \in [t_{i-1}, t_i)$

$$\overline{S}_{T_{j:m}}(t) = \overline{P}(T_{j:m} > t) = \sum_{l=i}^{n+1} P(T_{j:m} \in I_l)$$

At observed failure times t_i there is no imprecision in the NPI lower and upper survival functions, that is $\underline{S}_{T_{j:m}}(t_i) = \overline{S}_{T_{j:m}}(t_i)$ for $i = 1, \ldots, n$, while $\underline{S}_{T_{j:m}}(0) = \overline{S}_{T_{j:m}}(0) = 1$. For $t > t_n$, $\underline{S}_{T_{j:m}}(t) = 0$ and $\overline{S}_{T_{j:m}}(t) = \prod_{l=j}^{m} \frac{l}{n+l} > 0$. This reflects that there is no evidence in favour of such components, and hence the system, surviving past time t_n (reflected by the

lower survival function being zero), but the evidence against this is limited as there are only n observations (reflected by the upper survival function being a positive decreasing function of n). The NPI lower and upper survival functions for the failure time T_S of a system with signature q are [1]

$$\underline{S}_{T_S}(t) = \underline{P}(T_S > t) = \sum_{j=1}^{m} q_j \underline{S}_{T_{j:m}}(t) \quad (2)$$

$$\overline{S}_{T_S}(t) = \overline{P}(T_S > t) = \sum_{j=1}^{m} q_j \overline{S}_{T_{j:m}}(t) \quad (3)$$

While this is a straightforward generalization of (1), the derivation involves m optimisation problems which take on the optima simultaneously [1].

III. Partially known signatures

Computation of the system signature is a complex problem due to the fact that $m!$ orderings in which the m components fail must be considered. Explicit expressions for the signature of some specific system structures are available [3], but general algorithms to compute signatures have not received much attention in the literature. As any computational method will have to deal with the very large number of orderings, it is interesting to consider if one really needs to know the exact signature for a specific inference on the system's reliability. It is likely that any method for computing the signature, if ended before the exact signature has been derived, will provide bounds for the probabilities q_j of the signature. We explore the use of bounds on q_j in NPI. The method presented can be applied throughout the process of computation of the signature and can indicate when further computation is not required.

Assume that bounds on the elements of signature $q = (q_1, \ldots, q_m)$ have been derived, with $0 \leq \underline{q}_j \leq q_j \leq \overline{q}_j \leq 1$. Assume $\sum_{j=1}^{m} \underline{q}_j \leq 1$ and $\sum_{j=1}^{m} \overline{q}_j \geq 1$, so at least one signature (with elements summing to one) exists between these bounds. We also assume, for all $j = 1, \ldots, m$

$$\underline{q}_j \geq 1 - \sum_{\substack{l=1 \\ l \neq j}}^{m} \overline{q}_l \text{ and } \overline{q}_j \leq 1 - \sum_{\substack{l=1 \\ l \neq j}}^{m} \underline{q}_l \quad (4)$$

If these inequalities are not satisfied then $\underline{q}_j$ can be increased or $\overline{q}_j$ decreased, to the value which gives equality in the corresponding inequality without any change to the signatures q whose elements are all within these bounds.

Suppose that we want to derive the NPI lower and upper survival functions (2) and (3) based on the observed failure times of n tested components which are exchangeable with those in the system. If the exact system signature is not known, but bounds $\underline{q}_j$ and $\overline{q}_j$ are available for each probability q_j, then these can be used to derive lower and upper bounds for these NPI lower and upper survival functions which are the tightest possible bounds corresponding to these bounds for the elements of the signature. Because $\underline{S}_{T_{j:m}}(t)$ and $\overline{S}_{T_{j:m}}(t)$ are increasing functions of j, for all $t > 0$, it is clear that we can derive two signatures with all their elements within the bounds and such that one of them provides the maximum lower bound for both $\underline{S}_{T_S}(t)$ and $\overline{S}_{T_S}(t)$ and the other provides the minimum upper bound for both $\underline{S}_{T_S}(t)$ and $\overline{S}_{T_S}(t)$, for all $t > 0$. This corresponds to the link between the stochastic ordering of random failure times of systems and the stochastic ordering of their signatures [5]. We call the signature within these bounds that provides the maximum lower bound for the NPI lower and upper survival functions the 'pessimistic signature', denoted by q^p, and the one that provides the minimum upper bound for the NPI lower and upper survival functions the 'optimistic signature', denoted by q^o. These terms follow the logical interpretation of 'pessimistic' and 'optimistic' in terms of survival of the system and the lack of knowledge of the actual NPI lower and upper survival functions as the exact signature is not known.

The pessimistic signature puts the probability mass that is flexible according to the given bounds $\underline{q}_j$ and $\overline{q}_j$ as far to the left as possible, so to elements with lower values of j, hence making earlier system failure more likely. The optimistic signature puts this probability mass as far to the right as possible, so to elements with higher values of j, hence making later system failure more likely. Algorithms to derive q^p and q^o are easy to implement, and lead to

$$q^p = (\overline{q}_1, \ldots, \overline{q}_{j_p-1}, 1 - \sum_{j=1}^{j_p-1} \overline{q}_j - \sum_{j=j_p+1}^{m} \underline{q}_j, \underline{q}_{j_p+1}, \ldots, \underline{q}_m)$$

$$q^o = (\underline{q}_1, \ldots, \underline{q}_{j_o-1}, 1 - \sum_{j=1}^{j_o-1} \underline{q}_j - \sum_{j=j_o+1}^{m} \overline{q}_j, \overline{q}_{j_o+1}, \ldots, \overline{q}_m)$$

for some $j_p, j_o \in \{1, \ldots, m\}$. The assumptions (4) ensure that the j_p, j_o are unique and $q^p_{j_p} \in [\underline{q}_{j_p}, \overline{q}_{j_p}]$ and $q^o_{j_o} \in [\underline{q}_{j_o}, \overline{q}_{j_o}]$.

The lower and upper bounds for the NPI lower and upper survival functions for T_S follow immediately from (2), (3) and the pessimistic and optimistic signatures q^p and q^0,

$$\underline{S}^p_{T_S}(t) = \sum_{j=1}^{m} q^p_j \underline{S}_{T_{j:m}}(t) \quad (5)$$

$$\underline{S}^o_{T_S}(t) = \sum_{j=1}^{m} q^o_j \underline{S}_{T_{j:m}}(t) \quad (6)$$

and the lower and upper bounds for the NPI upper survival function for T_S are

$$\overline{S}^p_{T_S}(t) = \sum_{j=1}^{m} q^p_j \overline{S}_{T_{j:m}}(t) \quad (7)$$

$$\overline{S}^o_{T_S}(t) = \sum_{j=1}^{m} q^o_j \overline{S}_{T_{j:m}}(t) \quad (8)$$

These are the sharpest bounds for the NPI lower and upper survival functions for T_S corresponding to the bounds $\underline{q}_j$ and $\overline{q}_j$ for q_j, for $j = 1, \ldots, m$. Due to the construction of these bounds, it is clear that they can actually be attained. So, when the real signature q is only known up to such bounds for its individual elements, it follows that the NPI lower and upper survival functions for T_S are between their

respective bounds, and nothing more can be deduced without additional assumptions or indeed without further computation of the signature. Further computation which falls short of deriving the exact signature will lead to new bounds for the NPI lower and upper survival functions which are within the corresponding earlier bounds. This may be useful for deciding if further computation is required for a specific inferential problem. For example, if one is interested in the system's reliability at time t^* and requires a minimum probability of p^* for the system to function at time t^*, then $\underline{S}^p_{T_S}(t^*) \geq p^*$ would imply that the reliability requirement is certainly met without need for further computation of the signature. Similarly, if $\overline{S}^o_{T_S}(t^*) \leq p^*$ then the reliability requirement is certainly not met. In the other situations one cannot draw a firm conclusion about whether or not the reliability requirement is met and one may want to continue computation of the system signature. Even with the exact signature it is possible that no firm conclusion can be drawn, namely if $\underline{S}_{T_S}(t^*) < p^* < \overline{S}_{T_S}(t^*)$. In such a case one would either require more test data or use additional information, insights or assumptions in order to reach a conclusion. We consider it an advantage of the use of lower and upper probabilities that such situations can occur, as they reflect the limits to the amount of information in test results. The use of these lower and upper bounds at different levels of computation of the system signature, so with increasingly accurate bounds, will be illustrated in Example 1 in Section IV. In all examples we will concentrate on the optimal lower bound for the NPI lower survival function and the optimal upper bound for the NPI upper survival function, which are likely to be of most relevance for inferences.

It may be possible to derive a system's signature by combining signatures of its subsystems. Gaofeng et al [4] present such algorithms for a system consisting of two subsystems in parallel or series configuration, with all components in the system exchangeable. For the NPI approach, bounds for the signatures of two subsystems in parallel or series configuration can be used to derive bounds for the full system's signature, using the same algorithms. The reason for this is the assumption that the system is coherent, which implies that a decrease (increase) in reliability of a component can never lead to increased (decreased) reliability of the system, therefore a decrease (increase) in reliability of a subsystem can never lead to increased (decreased) reliability of the system. The pessimistic signatures for the two subsystems can be combined to give the pessimistic signature for the full system, and combining the optimistic signatures for the two subsystems leads to the optimistic signature for the full system. Due to space restrictions we do not include the formulae for such combinations [4], Example 2 in Section IV illustrates this approach.

In addition the survival of a system consisting of exchangeable components, other inferences can be considered. Coolen and Al-nefaiee [1] considered the comparison of the failure times of two coherent systems, each consisting of exchangeable components. It is assumed that the failure times of the components in the different systems are fully independent, so any information about components' failure times of one system does not affect (lower and upper) probabilities involving only failure times of components of the other system. Due to the monotonicity of this comparison with regard to the systems' signatures, such comparison with exactly known signatures [1] can be generalized to partially known signatures. Let the signatures of systems A and B be q^a and q^b and their failure times T^a and T^b, and assume that these systems have m_a and m_b components and n_a and n_b components exchangeable with those in the respective system were tested with failure times $t^a_1 < t^a_2 < \ldots < t^a_{n_a}$ and $t^b_1 < t^b_2 < \ldots < t^b_{n_b}$. Let $t^a_0 = t^b_0 = 0$ and $t^a_{n_a+1} = t^b_{n_b+1} = \infty$. If the exact signatures are known, NPI lower and upper probabilities for the event $T^a \leq T^b + \delta$ are [1]

$$\underline{P}(T^a \leq T^b + \delta) = \sum_{i=1}^{m_a} \sum_{j=1}^{m_b} q^a_i q^b_j \underline{P}(T^a_{i:m_a} \leq T^b_{j:m_b} + \delta)$$

where

$$\underline{P}(T^a_{i:m_a} \leq T^b_{j:m_b} + \delta) = \sum_{l=1}^{n_a} P^{a,i}_l \underline{P}(T^b_{j:m_b} + \delta \geq t^a_l)$$

with $P^{a,i}_l = P(T^a_{i:m_a} \in (t^a_{l-1}, t^a_l))$. Let $v_{l,\delta} \in \{1, \ldots, n_b + 1\}$ be such that $t^b_{v_{l,\delta}-1} < t^a_l - \delta < t^b_{v_{l,\delta}}$, then

$$\underline{P}(T^b_{j:m_b} + \delta \geq t^a_l) = \sum_{v=v_{l,\delta}+1}^{n_b+1} P(T^b_{j:m_b} \in (t^b_{v-1}, t^b_v))$$

The corresponding NPI upper probability is

$$\overline{P}(T^a \leq T^b + \delta) = \sum_{i=1}^{m_a} \sum_{j=1}^{m_b} q^a_i q^b_j \overline{P}(T^a_{i:m_a} \leq T^b_{j:m_b} + \delta)$$

where

$$\overline{P}(T^a_{i:m_a} \leq T^b_{j:m_b} + \delta) = \sum_{l=1}^{n_a+1} P^{a,i}_l \overline{P}(T^b_{j:m_b} + \delta \geq t^a_{l-1})$$

and

$$\overline{P}(T^b_{j:m_b} + \delta \geq t^a_{l-1}) = \sum_{v=v_{l,\delta}}^{n_b+1} P(T^b_{j:m_b} \in (t^b_{v-1}, t^b_v))$$

If the exact signatures are not available but instead bounds $\underline{q}^a$ and $\overline{q}^a$ for q^a and $\underline{q}^b$ and $\overline{q}^b$ for q^b have been derived, which are assumed to satisfy (4), then the optimal lower bound for the NPI lower probability for the event $T^a \leq T^b + \delta$ is derived using the optimistic signature $q^{a,o}$ for System A and the pessimistic signature $q^{b,p}$ for System B

$$\underline{P}^l(T^a \leq T^b + \delta) = \sum_{i=1}^{m_a} \sum_{j=1}^{m_b} q^{a,o}_i q^{b,p}_j \underline{P}(T^a_{i:m_a} \leq T^b_{j:m_b} + \delta)$$

The optimal upper bound for the NPI upper probability for $T^a \leq T^b + \delta$ is derived using the pessimistic signature $q^{a,p}$ for System A and the optimistic signature $q^{b,o}$ for System B

$$\overline{P}^u(^a \leq T^b + \delta) = \sum_{i=1}^{m_a} \sum_{j=1}^{m_b} q^{a,p}_i q^{b,o}_j \underline{P}(T^a_{i:m_a} \leq T^b_{j:m_b} + \delta)$$

These bounds follow from the monotonicity of these NPI lower and upper probabilities with regard to the signatures. The lower bound for the NPI lower probability for this event corresponds to maximum optimism about the lifetime of System A and maximum pessimism about the lifetime of System B, which is fully in line with intuition, and of course the other way around for the upper bound for the NPI upper probability. The upper bound for the NPI lower probability and the lower bound for the NPI upper probability are of course derived by taking the alternative optimistic or pessimistic signatures, but these are less likely to be of interest. This is illustrated in Example 3 in Section IV.

IV. Examples

Example 1. For the system in Figure 1, computing the signature involves determining for all of the $7! = 5040$ orderings of the failure times of the components, at which of these ordered times the system fails. Of course, all $6! = 720$ orderings with failure of Component 1 occurring first lead to immediate failure, from which we can conclude the lower bound $\underline{q}_1 = 0.143$. It is easy to see that no other component's failure will lead to immediate system failure if it is the first to fail, hence also the upper bound $\overline{q} = 0.143$. In addition, it is easy to see that the system cannot function with at most two functioning components, this leads to the upper bounds $\overline{q}_6 = \overline{q}_7 = 0$. This information, using conditions (4) but without further computation, can be reflected by $\underline{q} = (0.143, 0, 0, 0, 0, 0, 0)$ and $\overline{q} = (0.143, 0.857, 0.857, 0.857, 0.857, 0, 0)$. The corresponding pessimistic and optimistic signatures are $q^p = (0.143, 0.857, 0, 0, 0, 0, 0)$ and $q^o = (0.143, 0, 0, 0, 0.857, 0, 0)$. Computation of signatures by counting orderings typically leads to information in the form of lower bounds $\underline{q}_j$ for individual elements of the signature. To illustrate the method presented in this paper further, Table I provides, in addition to the first case just mentioned, three more combinations of lower and upper bounds for this system's signature as occurred at different stages of its computation, with increasing amount of information in Cases 1 to 4. For each case the pessimistic and optimistic signatures are also presented in this table. Test component failure times were simulated for this example, with $n = 100$ observations taken from the Weibull distribution with shape parameter 3 and scale parameter 1. The corresponding lower bounds for the NPI lower survival function, $\underline{S}^p_{T_S}(t)$ as given in Equation (5), and the upper bounds for the NPI upper survival function, $\overline{S}^o_{T_S}(t)$ as given in Equation (8), are presented in the plots in Figure 2, where in each plot also the NPI lower and upper survival functions are presented based on the exact signature, which is $q = (1/5040) \times (720, 1200, 1392, 1440, 288, 0, 0) = (0.143, 0.238, 0.276, 0.286, 0.057, 0, 0)$. These plots illustrate the use of the bounds as presented in this paper, and also show that the lower bound of the NPI lower survival function moves up if more details about the signature become known, in which case the upper bound for the NPI upper survival function moves down. As possible use of these bounds in order to determine when no further computation for the signature is needed, suppose a reliability requirement that the system's failure time should exceed 0.5 with probability at least 0.8. With the bounds for the signature in Case 1, the upper bound for the NPI upper survival function at 0.5 is greater than 0.8 and the corresponding lower bound for the NPI lower survival function is less than 0.8, but for the bounds in Case 2, based on some additional computations, the upper bound for the NPI upper survival function at 0.5 is less than 0.8, so it is clear that the reliability requirement cannot be met and hence that no further computation of the signature is needed. Similarly, if one only requires that the system's failure time should exceed 0.5 with probability at least 0.3 then one needs no more computation once the bounds in Case 4 have been derived, as the corresponding lower bound for the NPI lower survival function at 0.5 exceeds 0.3 hence this reliability requirement is certainly met.

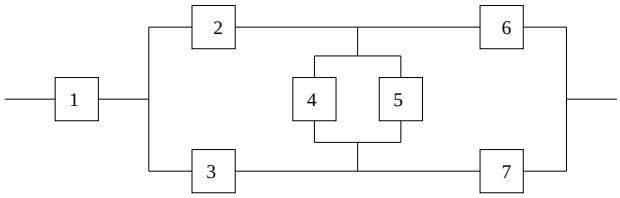


Fig. 1. A system with 7 components (Exs. 1,3)

TABLE I
Bounds, pessimistic and optimistic signatures (Ex. 1)

Case 1	$\underline{q}$	(0.143, 0, 0, 0, 0, 0, 0)
	$\overline{q}$	(0.143, 0.857, 0.857, 0.857, 0.857, 0, 0)
	q^p	(0.143, 0.857, 0, 0, 0, 0, 0)
	q^o	(0.143, 0, 0, 0, 0.857, 0, 0)
Case 2	$\underline{q}$	(0.143, 0.143, 0, 0, 0, 0, 0)
	$\overline{q}$	(0.143, 0.857, 0.714, 0.714, 0.714, 0, 0)
	q^p	(0.143, 0.857, 0, 0, 0, 0, 0)
	q^o	(0.143, 0.143, 0, 0, 0.714, 0, 0)
Case 3	$\underline{q}$	(0.143, 0.143, 0.076, 0, 0, 0, 0))
	$\overline{q}$	(0.143, 0.781, 0.714, 0.638, 0.638, 0, 0)
	q^p	(0.143, 0.781, 0.076, 0, 0, 0, 0)
	q^o	(0.143, 0.143, 0.076, 0, 0.638, 0, 0)
Case 4	$\underline{q}$	(0.143, 0.143, 0.152, 0.157, 0, 0, 0)
	$\overline{q}$	(0.143, 0.548, 0.557, 0.562, 0.405, 0, 0)
	q^p	(0.143, 0.548, 0.152, 0.157, 0, 0, 0)
	q^o	(0.143, 0.143, 0.152, 0.157, 0.405, 0, 0)

Example 2. Figure 3 shows a coherent system consisting of 17 exchangeable components, which consists of two subsystems in parallel configuration. Subsystem A is the same system, consisting of 7 components, as considered in Example 1. Subsystem B consists of 10 components. While the exact signature for this full system can be obtained by using the given signature for Subsystem A together with repeated use of the algorithm presented by Gaofeng et al [4] for Subsystem B and for the combination of the two subsystems, we assume, in order to illustrate the use of the bounds on signatures presented in this paper, that the signatures of subsystems A and B have

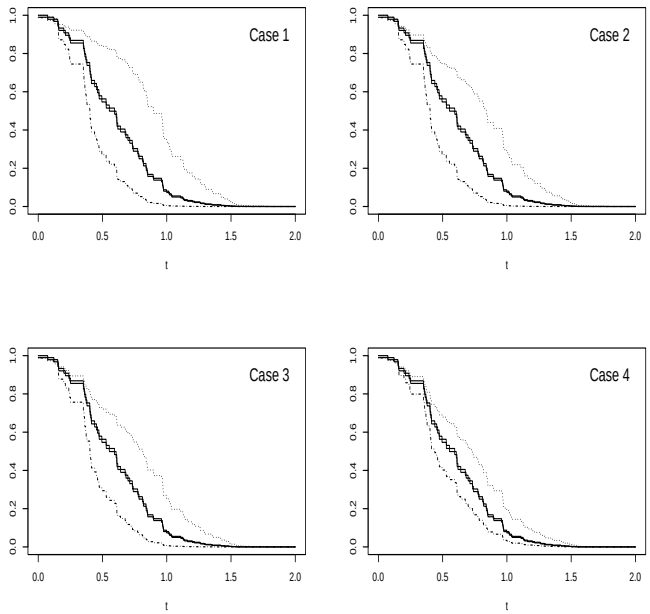


Fig. 2. NPI lower and upper survival functions (Ex. 1)

only been derived partially, with the bounds and corresponding pessimistic and optimistic signatures as presented in Table II.

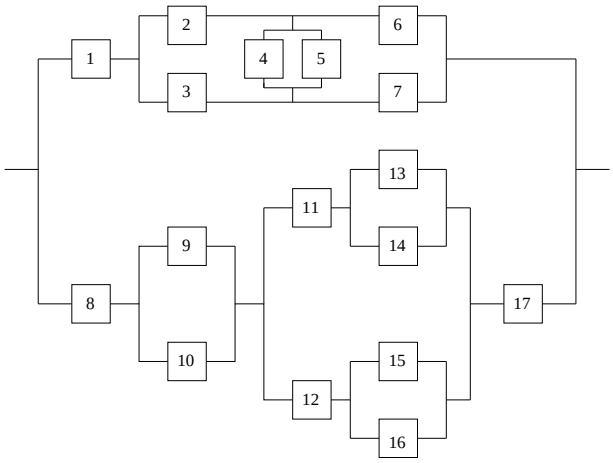


Fig. 3. Two subsystems in parallel (Ex. 2)

The pessimistic signature for the full 17-component system is derived by application of the algorithm presented by Gaofeng et al [4] with the use of the pessimistic signatures $q^{a,p}$ and $q^{b,p}$, which leads to

$$q^p = (0, 0.015, 0.050, 0.099, 0.161, 0.158, 0.136, 0.109, 0.084, 0.064, 0.048, 0.035, 0.023, 0.013, 0.005, 0, 0)$$

Applying the same algorithm with the optimistic signatures

TABLE II
BOUNDS, PESSIMISTIC AND OPTIMISTIC SIGNATURES FOR SUBSYSTEMS A AND B (EX. 2)

A	$\underline{q}^a$	(0.143, 0.143, 0.152, 0.157, 0.100, 0, 0)
	$\overline{q}^a$	(0.143, 0.448, 0.457, 0.462, 0.405, 0, 0)
	$q^{a,p}$	(0.143, 0.448, 0.152, 0.157, 0.100, 0, 0)
	$q^{a,o}$	(0.143, 0.143, 0.152, 0.157, 0.405, 0, 0)
B	$\underline{q}^b$	(0.200, 0.222, 0.072, 0.100, 0.046, 0.013, 0, 0, 0, 0)
	$\overline{q}^b$	(0.200, 0.222, 0.419, 0.447, 0.393, 0.360, 0, 0, 0, 0)
	$q^{b,p}$	(0.200, 0.222, 0.419, 0.100, 0.046, 0.013, 0, 0, 0, 0)
	$q^{b,o}$	(0.200, 0.222, 0.072, 0.100, 0.046, 0.360, 0, 0, 0, 0)

$q^{a,o}$ and $q^{b,o}$ leads to

$$q^o = (0, 0.015, 0.031, 0.040, 0.046, 0.051, 0.061, 0.078, 0.106, 0.128, 0.164, 0.128, 0.084, 0.047, 0.021, 0, 0)$$

In Figure 4, the left plot presents the lower bound for the NPI lower survival function and the upper bound for the NPI upper survival function, both for the failure time of the full system and based on $n = 10$ failure times of tested components which are exchangeable with those in the system (simulated from the Weibull distribution with shape parameter 2 and scale parameter 1). The right plot in Figure 4 is included for comparison with the following situation: Suppose that one would apply the NPI method presented in this paper directly to each subsystem individually, using the bounds given in Table II, but neglecting the fact that all components in both subsystems are exchangeable. Making this mistake, one could continue by calculating bounds for the full system's survival function following the standard way for simple parallel systems (effectively using '$1-(1-S_a)(1-S_b)$', with self-explanatory notation). The resulting lower and upper survival functions are greater than (or equal to) the correctly derived bounds for the NPI lower and upper survival function, because for the correct method the dependence of the components in both systems is taken into account. An intuitive explanation is as follows: The parallel system will only fail if both subsystems fail, and if one subsystem is known to fail this contains some information that suggests that the components are not very reliable, which as a consequence increases the (lower and upper) probability that the second subsystem also fails (when compared to the situation with the wrongly assumed independence between the two subsystems). This example shows the importance of taking the dependence of the exchangeable components, due to the limited information about their reliability from the test results, carefully into account, as is done by the NPI approach with the use of (bounds of) signatures.

Example 3. Consider the systems of Figures 5 and 1, called System A and System B, respectively. Assume that each system consists of exchangeable components but these are different for the two systems, assuming independence of the failure times of components in the different systems. Assume that $n_a = n_b = 30$ components exchangeable with those of each type in the respective system have been tested, leading to the failure times in Table IV. Assume that bounds $\underline{q}^a$

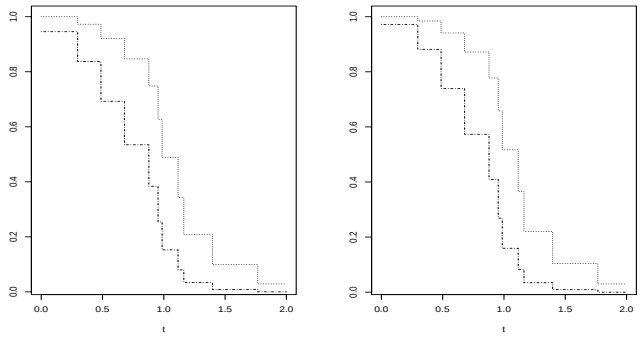


Fig. 4. Bounds on NPI lower and upper survival functions (Left); similar but resulting from wrongly assumed independence of subsystems (Right) (Ex. 2)

and $\overline{q}^a$ are available for the signature of System A, and bounds $\underline{q}^b$ and $\overline{q}^b$ for the signature of System B as given in Table III, which also presents the pessimistic and optimistic signatures corresponding to these bounds. The optimal lower bound for the NPI lower probability and the optimal upper bound for the NPI upper probability for the event $T_S^a \leq T_S^b + \delta$ are presented in Figure 6 as functions of δ. This figure also gives the NPI lower and upper probabilities for this event corresponding to the exact signatures [1], which for System B was given in Example 1 and for System A is equal to $q^a = (1/720) \times (0, 96, 192, 336, 96, 0) = (0, 0.133, 0.267, 0.467, 0.133, 0)$. Figure 6 gives a good impression of the actual difference between the failure times of these two systems, where it should be remarked that the bounds based on the partial information are still relatively wide compared to the NPI lower and upper probabilities based on the exact signatures as the vertical distances between the functions at specific values of δ must be considered.

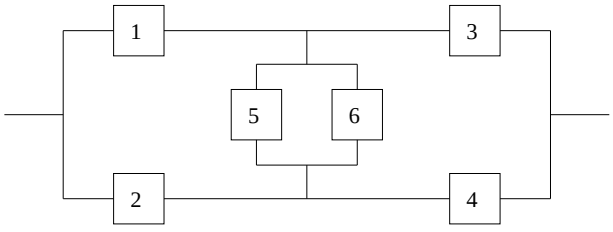


Fig. 5. System A (Ex. 3)

TABLE III
BOUNDS, PESSIMISTIC AND OPTIMISTIC SIGNATURES (EX. 3)

System A	$\underline{q}^a$	(0, 0.133, 0.267, 0.044, 0, 0)
	$\overline{q}^a$	(0, 0.133, 0.267, 0.600, 0.556, 0)
	$q^{a,p}$	(0, 0.133, 0.267, 0.600, 0, 0)
	$q^{a,o}$	(0, 0.133, 0.267, 0.044, 0.556, 0)
System B	$\underline{q}^b$	(0.143, 0.143, 0.152, 0.157, 0.100, 0, 0)
	$\overline{q}^b$	(0.143, 0.448, 0.457, 0.452, 0.405, 0, 0)
	$q^{b,p}$	(0.143, 0.448, 0.152, 0.157, 0.100, 0, 0)
	$q^{b,o}$	(0.143, 0.143, 0.152, 0.157, 0.405, 0, 0)

TABLE IV
COMPONENT FAILURE TIMES (EX. 3)

System A			System B		
0.223	0.747	0.994	0.154	0.585	1.076
0.265	0.798	1.008	0.155	0.598	1.169
0.372	0.807	1.073	0.347	0.642	1.239
0.419	0.824	1.115	0.402	0.692	1.248
0.564	0.850	1.167	0.483	0.738	1.327
0.630	0.887	1.182	0.512	0.822	1.421
0.675	0.914	1.275	0.513	0.843	1.569
0.685	0.921	1.397	0.548	0.848	1.643
0.709	0.981	1.400	0.563	0.863	1.735
0.727	0.987	1.425	0.574	0.938	2.565

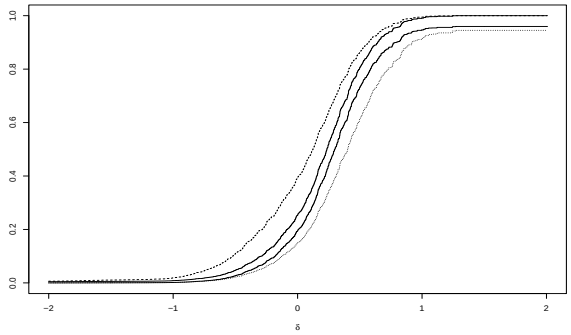


Fig. 6. (Bounds on) NPI lower and upper probabilities for $T_S^A < T_S^B + \delta$ (Ex. 3).

REFERENCES

[1] F.P.A. Coolen, A.N. Al-nefaiee, *Nonparametric predictive inference for failure times of systems with exchangeable components*. Journal of Risk and Reliability, to appear.

[2] F.P.A. Coolen, T.A. Maturi, *Nonparametric predictive inference for order statistics of future observations*. In: Combining Soft Computing and Statistical Methods in Data Analysis, C. Borgelt et al (Eds). Springer, pp. 97-104, 2010.

[3] S. Eryilmaz, *Review of recent advances in reliability of consecutive k-out-of-n and related systems*. Journal of Risk and Reliability 224, 225-237, 2010.

[4] D. Gaofeng, B. Zheng, H. Taizhong, *On computing signatures of coherent systems*. Journal of Multivariate Analysis 103, 142-150, 2012.

[5] F.J. Samaniego, *System Signatures and their Applications in Engineering Reliability*. Springer, 2007.

Markov-Modulated Linear Regression

Alexander M. Andronov
Dept. of Mathematical Methods and Modelling
Transport and Telecommunication Institute
Riga, Latvia
lora@mailbox.riga.lv

Nadezda Spiridovska
Dept. of Mathematical Methods and Modelling
Transport and Telecommunication Institute
Riga, Latvia
Spiridovska.N@tsi.lv

Abstract—**Classical linear regression is considered for a case when regression parameters depend on the external random environment. The last is described as a continuous time Markov chain with finite state space. Here the expected sojourn times in various states are additional regressors. Necessary formulas for an estimation of regression parameters have been derived. The numerical example illustrates the results obtained.**

Markov-Modulated processes; linear regression; external environment

I. MODEL DESCRIPTION

Classical linear regression [1 – 3] is of the form

$$Y_i = x_i\beta + Z_i,\ \ i = 1,...,n, \qquad (1)$$

where Y_i is scale response, $x_i = (x_{i,1}, x_{i,2},.., x_{i,k})$ is $1\times k$ vector, β is $k\times 1$ vector, Z_i is scale disturbance. The usual assumptions take place: the disturbances Z_i are independently, identically normally distributed with mean zero and variance σ^2, the $n\times k$ matrix $X = (x_{i,v}) = \left(x_i^{\ T}\right)^T$ has rank $r(X) = k$, so $\left(X^T X\right)^{-1}$ exists.

Now we suppose that model (1) corresponds to one unit of a continuous time, $Z_i(t)$ is Brown motion and responses $Y_i(t)$ are time-additive. Then for t > 0

$$Y_i(t) = x_i\beta\, t + Z_i\sqrt{t},\ \ i = 1,...,n, \qquad (2)$$

where Y_i (0) = 0 and disturbances Z_i are (as before) independently, identically normally distributed with mean zero and variance σ^2.

Let value Y_i of the *i*-th response be fixed after time t_i, so $Y_i = Y_i(t_i)$. In this case the generalized least square method's estimates of β and σ^2 are the following:

$$\widetilde{\beta} = \left(X^T W^2 X\right)^{-1} X^T Y,$$
$$\widetilde{\sigma}^2 = \frac{1}{n-k-1}\left(Y - W^2 X\widetilde{\beta}\right)^T W^{-2}\left(Y - W^2 X\widetilde{\beta}\right), \qquad (3)$$

where $W = diag\left(\sqrt{t_1},...,\sqrt{t_n}\right)$.

Additionally we suppose that model (2) operates in the so-called *external environment*, which has final state space *E* [4]. For the fixed state $s_j \in S$, j = 1,…, m, parameters β of model (2) are $\beta_j = \left(\beta_{1,j},...,\beta_{k,j}\right)^T$, but as before $Z_{i,j}$ are stochastically independent, normally distributed with mean zero and variances σ^2. Let $t_i = (t_{i,1},...,t_{i,m})$ be $1\times m$ vector, for that component $t_{i,j}$ means a sojourn time for response Y_i in the state $s_j \in S$. Note that $t_i = t_{i,1} + ... + t_{i,m}$. Then

$$Y_i(t_i) = x_i\sum_{j=1}^{m}\beta_j t_{i,j} + \sum_{j=1}^{m}\sqrt{t_{i,j}}Z_{i,j},\ \ i = 1,...,n.$$

Taking into account properties of the normal distribution, we can rewrite the last formula as

$$Y_i(t_i) = x_i\sum_{j=1}^{m}\beta_j t_{i,j} + Z_i\sqrt{t_i},\ \ i = 1,...,n. \qquad (4)$$

To write it in matrix notation we use Kronecker product $\otimes$ [2, 3, 5], the $k\times m$ matrix $\beta = (\beta_1,...,\beta_m) = (\beta_{v,j})$, the $n\times 1$ vector $Z = (Z_i)$, the $1\times m$ vector $\vec{t}_i = \left(t_{i,1},...,t_{i,m}\right)$, the *i*-th rows e_i of *n*-dimensional identity matrix, the *n*-dimensional diagonal matrix *diag*(*v*) with the vector *v* on the main diagonal, *vec* operator *vec A* of matrix *A*. Then

$$Y(T) = \left(Y_1(t_1),...,Y_n(t_n)\right)^T =$$
$$= \begin{pmatrix} \vec{t}_1 \otimes x_1 \\ \vec{t}_2 \otimes x_2 \\ \dots \\ \vec{t}_n \otimes x_n \end{pmatrix} vec\beta + diag(\sqrt{t_1}, \sqrt{t_2},..., \sqrt{t_n})Z. \qquad (5)$$

We see that the generalized linear regression model has place here. The expectation and the covariance matrix of *Y(T)* are the following:

The article is written with the financial assistance of European Social Fund. Project Nr.2009/0159/1DP/1.1.2.1.2/09/ IPIA/VIAA/006. The Support in Realisation of the Doctoral Programme "Telematics and Logistics" of the Transport and Telecommunication Institute.

$$E(Y(T)) = \begin{pmatrix} \vec{t}_1 \otimes x_1 \\ \vec{t}_2 \otimes x_{21} \\ \dots \\ \vec{t}_n \otimes x_n \end{pmatrix} vec\beta, \qquad (6)$$

$$Cov(Y(T)) = \sigma^2 diag(t_1, t_2, ..., t_n).$$

Now we are able to use the generalized least square method to estimate parameter matrix β, supposing that the matrix of regressors by *vec* β has full rank *mk*. If *Y* means an observed value of *Y(T)* then

$$vec\tilde{\beta} = \left(\begin{pmatrix} \vec{t}_1 \otimes x_1 \\ \vec{t}_2 \otimes x_2 \\ \dots \\ \vec{t}_n \otimes x_n \end{pmatrix}^T \begin{pmatrix} t_1 & 0 & \dots & 0 \\ 0 & t_2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & t_n \end{pmatrix}^{-1} \begin{pmatrix} \vec{t}_1 \otimes x_1 \\ \vec{t}_2 \otimes x_2 \\ \dots \\ \vec{t}_n \otimes x_n \end{pmatrix} \right)^{-1} \cdot$$

$$\cdot \begin{pmatrix} \vec{t}_1 \otimes x_1 \\ \vec{t}_2 \otimes x_2 \\ \dots \\ \vec{t}_n \otimes x_n \end{pmatrix}^T \begin{pmatrix} t_1 & 0 & \dots & 0 \\ 0 & t_2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & t_n \end{pmatrix}^{-1} Y =$$

$$= \left(\sum_{i=1}^{n} \frac{1}{t_i} \left(\vec{t}_i^{\,T} \vec{t}_i \right) \otimes \left(x_i^{\,T} x_i \right) \right)^{-1} \begin{pmatrix} t_1^{-1} \vec{t}_1 \otimes x_1 \\ t_2^{-1} \vec{t}_2 \otimes x_2 \\ \dots \\ t_n^{-1} \vec{t}_n \otimes x_n \end{pmatrix}^T Y. \qquad (7)$$

The variance σ^2 can be estimated as usually:

$$\tilde{\sigma}^2 = \frac{1}{n - mk - 1} \left(Y(T) - \begin{pmatrix} \vec{t}_1 \otimes x_1 \\ \vec{t}_2 \otimes x_2 \\ \dots \\ \vec{t}_n \otimes x_n \end{pmatrix} vec\tilde{\beta} \right)^T W^{-2} \cdot \qquad (8)$$

$$\cdot \left(Y(T) - \begin{pmatrix} \vec{t}_1 \otimes x_1 \\ \vec{t}_2 \otimes x_2 \\ \dots \\ \vec{t}_n \otimes x_n \end{pmatrix} vec\tilde{\beta} \right)$$

Note that the last estimate is biased because it doesn't take into account randomness of $\{\vec{t}_i\}$ values.

II. MARKOV-MODULATED CASE

Further we suppose that the external environment is a random one and is described by a continuous-time Markov chain $J(t)$, $t \geq 0$, with the finite state set $S = \{1, 2, \dots, m\}$ [4, 6]. Let $\lambda_{i,j}$ be the known transition rate from state s_i to state s_j, and $\Lambda_i = \sum_{j \neq i} \lambda_{i,j}$.

We have *n* independent realizations of this Markov chain. The *i*-th realization corresponds to the response Y_i. Let us use the following notation for the *i-th* realization: *N(i)* be the number of jumps of random environment *J*; $V_1, V_2, ..., V_{N(i)}$ be time moments of these jumps and $J_1, J_2, ..., J_{N(i)}$ be the corresponding states $J(V_1+), J(V_2+), ..., J(V_{N(t)}+)$; $I = J(0)$ be an initial state at time moment 0.

Therefore now sojourn time $T_{i,j}$ in the state s_j for the *i-th* realization is a sum of all terms $V_r - V_{r-1}$, for which $J_r = j$. Above, the case has been considered when the whole trajectory of environment *J(.)* is known, so $t_{i,j}$ is a fixed value of $T_{i,j}$. We suppose that total time t_i of the *i-th* observation is fixed, so $t_i = T_{i,1} + ... + T_{i,m}$. If $\vec{T}_i = (T_{i,1}, ..., T_{i,m})$ is the $1 \times m$ vector, then for the general case linear regression (5) is of the form

$$Y(T) = (Y_1(t_1), ..., Y_n(t_n))^T = \qquad (9)$$

$$= \begin{pmatrix} \vec{T}_1 \otimes x_1 \\ \vec{T}_2 \otimes x_2 \\ \dots \\ \vec{T}_n \otimes x_n \end{pmatrix} vec\beta + diag(\sqrt{t_1}, \sqrt{t_2} ... \sqrt{t_n}) Z$$

Note that $T = \{\vec{T}_1, ..., \vec{T}_n\}$ and *Z* are independent, so the expectation *E*(*Y* (*T*)) is the same as before in (6).

Now a special case of the given sample will be considered. It is supposed that for each realization *i* the following data are available: total observation time $t_i = T_{i,1} + ... + T_{i,m}$, initial $J_{i,0}$ and final $J_{i,\tau(i)}$ states of $J(.)$, and the response $Y_i = Y_i(t_i)$ from (9). On this basis we must estimate the unknown parameters: the $k \times m$ matrix $\beta = (\beta_1 \ \dots \ \beta_m) = (\beta_{v,j})$ and the variance σ^2. Additionally we use a knowledge on parameters of the modulated Markov chain *J*(.). One allows us to calculate the average sojourn time $E(T_{i,j} | t_i, J_{i,0}, J_{i,\tau(i)})$ in the state s_j during time t_i for the *i*-th realization, given fixed initial and final states $J_{i,0}$ and $J_{i,\tau(i)}$ of *J*(.), see below Section 3. This time will be used instead of $t_{i,j}$ in previous formulas (7) and (8) so we

set $t_{i,j} = E\left(T_{i,j} \middle| t_i, J_{i,0}, J_{i,\tau(i)}\right)$. Then $E\left(\vec{T}_i\right) = E\left(T_{i,1},...,T_{i,m}\right) = \vec{t}_i$ and as before

$$vec\tilde{\beta} = \left(\sum_{i=1}^{n} \frac{1}{t_i}\left(\vec{t}_i^{\,T}\vec{t}_i\right) \otimes \left(x_i^{\,T} x_i\right)\right)^{-1} \begin{pmatrix} t_1^{-1}\vec{t}_1 \otimes x_1 \\ t_2^{-1}\vec{t}_2 \otimes x_2 \\ \dots \\ t_n^{-1}\vec{t}_n \otimes x_n \end{pmatrix}^T Y. \quad (10)$$

Substitution *Y(T)* from (9), we get a form which is more convenient for statistical analysis

$$vec\tilde{\beta} = \left(\sum_{i=1}^{n} \frac{1}{t_i}\left(\vec{t}_i^{\,T}\vec{t}_i\right) \otimes \left(x_i^{\,T} x_i\right)\right)^{-1} \begin{pmatrix} t_1^{-1}\vec{t}_1 \otimes x_1 \\ t_2^{-1}\vec{t}_2 \otimes x_2 \\ \dots \\ t_n^{-1}\vec{t}_n \otimes x_n \end{pmatrix}^T \cdot$$

$$\cdot \left(\begin{pmatrix} t_1^{-1}\vec{T}_1 \otimes x_1 \\ t_2^{-1}\vec{T}_2 \otimes x_2 \\ \dots \\ t_n^{-1}\vec{T}_n \otimes x_n \end{pmatrix} vec\beta + diag(\sqrt{t_1} \ \dots \ \sqrt{t_n})Z \right). \quad (11)$$

As $E\left(\vec{t}_i^{\,T}\vec{T}_i \middle| t_i, J_{i,0}, J_{i,\tau(i)}\right) = \vec{t}_i^{\,T} E\left(\vec{T}_i \middle| t_i, J_{i,0}, J_{i,\tau(i)}\right) = = \vec{t}_i^{\,T}\vec{t}_i$, *T* and *Z* are independent, and *vecZ* has zero expectation, we can conclude that the estimate (10) of ***β* is unbiased**.

Further we give necessary formulas for a calculation of the conditional average sojourn time that allows us to get the estimates needed.

III. Modulating Markov chain

For transition probabilities $p_{i,j}(t) = P\{J(t) = j | J(0) = i\}$ of the above described Markov chain, a usual system of differential equations take place [4, 6]. If $P(t) = \left(p_{i,j}(t)\right)$ and $\lambda = \left(\lambda_{i,j}\right)$ are the $m \times m$ matrices, Λ is an *m*-dimensional diagonal matrix with a vector $\left(\Lambda_1,...,\Lambda_m\right)$ on the main diagonal then

$$\dot{P}(t) = -P(t)\Lambda + P(t)\lambda, \ \ t \ge 0.$$

The solution can be represented by the matrix exponent [7, 8]:

$$P(t) = \exp\left(t\left(\lambda - \Lambda\right)\right), \ \ t \ge 0, \quad (12)$$

where *P(0) = I*.

If all the eigenvalues of matrix $A = \lambda - \Lambda$ are different then the solution (12) can be represented more simply. Let v_η and Z_η, η = 1, …, *m*, be the eigenvalue and the corresponding eigenvector of *A*, $Z = \left(Z_1,...,Z_m\right)$ be the matrix of the eigenvectors and $\overline{Z} = Z^{-1} = \left(\overline{Z}_1^{\,T},...,\overline{Z}_k^{\,T}\right)^T$ be the corresponding inverse matrix (here $\overline{Z}_\eta$ is the η–th row of $\overline{Z}$). Then [7, 8]:

$$P(t) = \exp\left(tA\right) = Zdiag\left(\exp\left(v_1 t\right),...,\exp\left(v_m t\right)\right)Z^{-1} = \\ = \sum_{\eta=1}^{m} Z_\eta \exp(\gamma_\eta t)\overline{Z}_\eta. \quad (13)$$

For the conditional average sojourn time $t_{r,\nu}(\tau) = E\left(T_{r,\nu} \middle| t_r = \tau, J_{r,0} = i, J_{r,\tau} = j\right)$ in the state $\nu \in S$ on the interval $(0, \tau)$ we have

$$t_{r,\nu}(\tau) = \frac{1}{p_{i,j}(\tau)} \int_0^\tau p_{i,\nu}(u) p_{\nu,j}(\tau - u) du. \quad (14)$$

Further

$$p_{i,j}(\tau) = \sum_{\eta=1}^{m} Z_{i,\eta} \exp(\gamma_\eta \tau)\overline{Z}_{\eta,j}, \quad (15)$$

$$\int_0^\tau p_{i,\nu}(u) p_{\nu,j}(t-u) du = \\ = \int_0^\tau \sum_{\eta=1}^{m} Z_{i,\eta} \exp(\gamma_\eta u)\overline{Z}_{\eta,\nu} \sum_{\theta=1}^{m} Z_{\nu,\theta} \exp\left(\gamma_\theta(\tau - u)\right)\overline{Z}_{\theta,j} du = \\ = \sum_{\eta=1}^{m} Z_{i,\eta}\overline{Z}_{\eta,\nu} \sum_{\theta=1,\theta\ne\eta}^{m} Z_{\nu,\theta}\overline{Z}_{\theta,j} \exp\left(\gamma_\theta \tau\right) \frac{1}{\gamma_\theta - \gamma_\eta} \cdot \\ \cdot \left(1 - \exp\left(-\tau(\gamma_\theta - \gamma_\eta)\right)\right) + \\ + \tau \sum_{\eta=1}^{m} Z_{i,\eta}\overline{Z}_{\eta,\nu} Z_{\nu,\eta}\overline{Z}_{\eta,j} \exp\left(\gamma_\eta \tau\right) = \\ = \tau \sum_{\eta=1}^{m} Z_{i,\eta}\overline{Z}_{\eta,\nu} Z_{\nu,\eta}\overline{Z}_{\eta,j} \exp\left(\gamma_\eta \tau\right) + \\ + \sum_{\eta=1}^{m} Z_{i,\eta}\overline{Z}_{\eta,\nu} \sum_{\theta=1,\theta\ne\eta}^{m} Z_{\nu,\theta}\overline{Z}_{\theta,j} \frac{1}{\gamma_\theta - \gamma_\eta}\left(\exp\left(\gamma_\theta \tau\right) - \exp\left(\gamma_\eta \tau\right)\right). \quad (16)$$

Now we can make calculation by formula (14) and derive estimates (10) setting $t_{i,\nu} = t_{i,\nu}(t_i)$.

IV. Simulation study

Our example supposes three states of the environment (*m* = 3). The known transition rates $\{\lambda_{i,j}\}$ from state s_i to state s_j are set by matrix

$$\lambda = (\lambda_{i,j}) = \begin{pmatrix} 0 & 0.2 & 0.3 \\ 0.1 & 0 & 0.2 \\ 0.4 & 0 & 0 \end{pmatrix}.$$

Stationary state distribution is the following [6]: $\pi = (0.364 \quad 0.242 \quad 0.394)^T$.

The number of regressors equals three (k = 3). Firstly we consider a case of a small sample, when n = 15 observations take place. The regressors' values are the following:

$$X = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 4 & 5 & 7 & 3 & 8 & 2 & 3 & 9 & 5 & 4 & 6 & 3 & 2 & 5 & 7 \\ 1.1 & 2.5 & 4.9 & 0.9 & 3.4 & 2.4 & 1.9 & 4.1 & 4.9 & 2.6 & 3.6 & 2.9 & 1.6 & 3.5 & 1.9 \end{pmatrix}^T$$

The two vectors t and I contain values of the durations of the observations and initial states of the environment:

$$t = (5 \quad 8 \quad 3 \quad 6 \quad 9 \quad 6 \quad 4 \quad 6 \quad 9 \quad 8 \quad 5 \quad 7 \quad 8 \quad 10 \quad 5)^T,$$
$$I = (1 \quad 0 \quad 2 \quad 2 \quad 1 \quad 1 \quad 0 \quad 0 \quad 1 \quad 2 \quad 1 \quad 1 \quad 0 \quad 0 \quad 2)^T.$$

Note that the total observation time equals to 99.

A simulation has been used for our purpose. The following parameters of the regression model are supposed: σ = 1,

$$\beta = \begin{pmatrix} 0 & 2 & 4 \\ 1 & 3 & 6 \\ 2 & 5 & 8 \end{pmatrix}$$

so $vec(\beta) = (0 \; 1 \; 2 \; 2 \; 3 \; 5 \; 4 \; 6 \; 8)^T$.

The first run of the simulation model gives the following values of the final states of the environment (vector J) and responses (vector Y):

$$J = (0 \quad 2 \quad 0 \quad 2 \quad 2 \quad 0 \quad 1 \quad 1 \quad 1 \quad 2 \quad 2 \quad 1 \quad 0 \quad 1 \quad 0)^T,$$
$$Y = (67 \quad 230 \quad 231 \quad 174 \quad 296 \quad 126 \quad 43 \quad 226 \quad 427 \quad 187 \quad 284 \quad 179 \quad 146 \quad 346 \quad 262)^T.$$

The data obtained are used for estimating the parameters β, those being supposed as unknown. We begin with the estimates for a simple linear regression (2) with the three regressors from X. Formula (3) gives $\widetilde{\beta} = (2.060 \quad 3.454 \quad 5.090)^T$. Weighted residual square sum R = 9816 against R = 23200 for one with respect to the sample mean.

Now we exam the suggested approach. The values of conditional average sojourn times $\bar{t}_i = (t_{i,1},...,t_{i,m})$ have been calculated by formulas (14) – (16). They are given for all observations in Table 1.

TABLE I. Value of average sojourn time $\bar{t}_i^T = (t_{i,1}, t_{i,2}, t_{i,3})^T, i = 1,...,15$

$$E(T) = \begin{pmatrix} i & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 & 13 & 14 & 15 \\ j=1 & 1.8 & 3.2 & 1.4 & 1.5 & 1.9 & 2.1 & 1.7 & 2.6 & 1.7 & 2.2 & 0.7 & 1.1 & 4.5 & 4.0 & 2.3 \\ j=2 & 2.0 & 1.2 & 0.1 & 0.3 & 3.4 & 2.3 & 2.0 & 2.7 & 5.9 & 0.7 & 2.2 & 5.2 & 1.0 & 3.9 & 0.2 \\ j=3 & 1.2 & 3.6 & 1.5 & 4.2 & 3.7 & 1.6 & 0.3 & 0.7 & 1.4 & 5.1 & 2.1 & 0.7 & 2.5 & 2.1 & 2.5 \end{pmatrix}$$

Formula (3) gives the following estimates:

$$\widetilde{\beta} = \begin{pmatrix} -10.843 & -2.785 & 3.168 \\ 5.013 & -8.492 & 8.452 \\ 8.099 & 18.666 & 0.857 \end{pmatrix}.$$

They are very far from true β but weighted residual square sum is sufficiently less: R = 4674. Therefore in the considered example the suggested method gives inaccessible estimates of true β but predicts values of the responses well. What is the cause of such a fact? The given number of the observation (n = 15) is insufficient for the given number (9) of estimated parameters! After all, a random environment with exponential distributed sojourn times creates a big randomness. It can be seen from a value of estimated variance $\widetilde{\sigma}^2$, calculated by formula (8): one equals 934.8 although $\sigma^2 = 1$. This difference is explained by an essential randomness.

A possibility of using the approach described for a prediction can be seen from the following reasoning. For the considered regression model and given data on t, I, J, and $\bar{t}_i = (t_{i,1},...,t_{i,m})$ from Table 1, the expectation of the responses is the following:

$$E(Y) = (44.7 \quad 193.3 \quad 131.1 \quad 122.9 \quad 295.1 \quad 54.9 \quad 9.8$$
$$65.1 \quad 99.9 \quad 247.8 \quad 148.7 \quad 32.8 \quad 70.9 \quad 126.7 \quad 155.5)^T.$$

Let us use this expectation as observed responses Y. The considered approach gives the same result: the residual square sum equals 6.207×10^{-4} only! On other hand the estimates of β are insufficient ones:

$$vec(\widetilde{\beta}) = (0.024 \quad 8.091 \times 10^{-3} \quad -0.022 \quad 0.015 \quad -4.27 \times 10^{-4}$$
$$-3.658 \times 10^{-3} \quad 3.992 \quad 5.989 \quad 8.023)^T.$$

Note that the estimates $\widetilde{\beta}_3 = (3.992 \quad 5.989 \quad 8.023)^T$ are very close to the true parameter values for the third environment state: $\beta_3 = (4 \quad 6 \quad 8)^T$.

Formula (3) gives $\widetilde{\beta} = (7.503 \quad 3.485 \quad -2.149)^T$ and the weighted residual square sum 8.25×10^3 that is sufficiently worse.

Further we consider a case of a big sample. We organize the following simulation experiment. The last includes q independent blocks. Each block has the above described structure: the same random environment, observation number n = 15, regressors' number k = 3. The matrix of regressors, the initial state I of random environment J and the observations' times t are chosen at random according to the following distributions. The expectation of the matrix of regressors coincides with the above obtained matrix X, all elements of the second and third columns are independent and uniformly distributed on intervals (-2, 2) and (-1, 1) correspondingly. The expectation of observations' times coincides with previous values t, all times are independent and time of i-th observation t_i has uniform distribution on (0,

2 t_i). The initial states I for various observations are independent and are chosen with respect to stationary distribution of the states for the random environment J. Therefore the total number of the observation for one experiment equals $qn = 15\,q$.

The above described simulation procedure is realized for those conditions. Table 2 contains the obtained estimates of true parameters $vec(\beta) = \left(0\ 1\ 2\ 2\ 3\ 5\ 4\ 6\ 8\right)^T$ for increasing values of q.

TABLE II. RESULTS OF SIMULATION EXPERIMENT

q	500	1000	1500	2000	2500	3000	3500	4000	4500
$\beta_{1,1}$	0.112	0.777	0.058	-0.02	0.179	0.154	0.239	0.400	0.367
$\beta_{2,1}$	0.711	0.770	0.881	0.898	0.992	0.868	0.851	0.888	0.948
$\beta_{3,1}$	2.566	2.158	2.289	2.276	2.207	2.255	2.201	2.090	2.012
$\beta_{,1,2}$	2.746	1.572	2.170	2.306	2.105	2.092	2.026	2.047	2.061
$\beta_{2,2}$	3.030	3.164	3.035	3.014	3.025	3.050	3.063	3.034	3.011
$\beta_{3,2}$	4.707	4.932	4.935	4.913	4.948	4.928	4.934	4.960	4.978
$\beta_{1,3}$	3.496	3.568	3.634	3.572	3.556	3.776	3.841	3.789	3.738
$\beta_{2,3}$	6.152	6.079	6.058	6.061	6.113	6.105	6.099	6.071	6.079
$\beta_{,3,3}$	7.969	8.053	7.964	8.005	7.904	7.854	7.877	7.910	7.903

We see from Table 2 that a convergence to true values of parameters is very slow. One can be increased if an estimation procedure uses true values of variances for sojourn times. It will be a direction of our future investigations.

REFERENCES

[1] C. R. Rao. Linear statistical inference and its applications. New York – London – Sydney: Jihn Wiley & Sons, INC., 1965.

[2] M. S. Srivastava. Methods of Multivariate Statistics. New-York: Wiley-Interscience, 2002.

[3] D. A. Turkington. Matrix Calculus and Zero-One Matrices. Statistical and Econometric Applications. Cambridge: Cambridge University Press, 2002

[4] A. Pacheco, L. C. Tang. and N. U. Prabhu. Markov-Modulated Processes & Semiregenerative Phenomena. New Jersy: World Scientific, 2009.

[5] T. Kollo and D. von Rosen. Advanced Multivariate Statistics with Matrices. Dordrecht: Springer, 2005.

[6] M. Kijima. Markov Processes for Stochastic Modeling. London: Chapman & Hall, 1997.

[7] R. Bellman. Introduction to matrix analysis. New York–Toronto–-London: McGraw-Hill Book Company, INC., 1969.

[8] L. S. Pontryagin. Ordinary differential equations. Moscow: Nauka, 1965. (In Russian.).

.

ON THE CALCULATION OF THE REDUNDANT STRUCTURES RELIABILITY TO AGING ELEMENTS

A. Antonov, A. Plyaskin and Kh. Tataev
Institute for Nuclear Power Engineering
Obninsk, Russia
antonov@iate.obninsk.ru, plyaskin@iate.obninsk.ru and khizri@bk.ru

The paper discusses the calculation of reliability of redundant structures, taking into account the aging elements. The operation practice of the modern industrial plants is such that both the basic element and elements from the spare subject are affected by process. Failed objects have to be renewed. After repairing they supplement of spare elements. It should be noted when the repair is finished usually a partial renewal of capacity occurs. An object produces a part of the resource during the previous operation and full renewal is not happening. The problem of determine of spare elements composition taking into account renewal and working out of a certain part of their resource is solved by simulation.

Safety; Reliability; Spares Elements; Redundancy; Ageing; Geometric Process; Resource; Transition Graph; Incomplete Recovery

I. INTRODUCTION

Organizing the operation of industrial objects, especially high-risk objects such as nuclear power plants (NPP), high demands are made of safety and reliability of their operation. One way to improve reliability is to plan preventive maintenance, objects functional testing and the creation of sets of spare elements for the rapid replacement of faulty equipment.

In this paper we consider the calculation of reliability issues of restorable equipment taking into account the availability of spare elements and determination of the required number of spare elements that guarantee support of the specified reliability coefficients. The questions of calculation of the equipment reliability taking into account spare parts and of determination of its optimal composition were considered in both Russian and foreign specialists.

We note here one feature that is characteristic of a large amount of equipment in service in various industries. This feature is - the equipment has a big operating time. Often it exceeds a resource or life time that is established for it in the regulations. The objects operating in many industries were put into operation during the Soviet era. Replacement of the equipment is a slow process. Consequently, we can assume that aging process take place in this equipment due to wear and tear of materials and degradation processes occurring within the product. In this connection there is the problem of calculating of the equipment reliability with spare parts, there with account of aging.

II. THE OPERATION PROCESS OF THE RENEWAL AND REPAIRED PRODUCTS

Let's consider the method of calculation of the reliability characteristics of renewal and repaired object with n spare elements. The strategy of the element operating is such. The item is in good condition at the initial time.

The element fails with the intensity $\lambda(t)$. Element is replaced by a reserve one at the time of failure. The intensity of the replacement element is $\mu(t)$. The faulty element is sent in for repair. The element is consider to be renewed after repair and proceeds into reserve. The intensity of the repair is $\nu(t)$. If available elements in the reserve are left, then a failure comes. The above strategy of operation can be represented by the graph shown in Figure 1.

It is necessary to assess the reliability of the system (availability factor, the probability of system failure due to lack of spare elements) for the reviewed strategy and to determine the required number of spare elements that provide a given level of system availability.

We denote the state of the object on the graph by two symbols (k, i), where the first character indicates the number of the spare elements, k=0,…,n, the second character i the state of a basic element, i=1, the element is operational, i=0 element is not operational.

Consider the operation of the object with spare elements in more detail. An element is with probability 1 in the state (n, 1) in the beginning (there are n available spare elements, the object is operational). The element changes its state (n, 0) in a random moment in time with the intensity of failure $\lambda(t)$ (n spare elements, the object is in a failure state, replacing of the element is beginning).

The object passes into a (n-1, 1) state with the intensity of recovery $\mu(t)$ (n-1 spare elements, the object is operational). From this state, transitions to state (n, 1) with an intensity of recovery $\nu(t)$ (repair is finished, n elements are in reserve

again), or in the state of (n-1, 0) with intensity λ(t) (repairs are not completed until the next failure) are possible, and so on. Getting into a condition (0,0) the object stops working and it is in a condition of refusal till the moment of replenishment of spare parts.

TABLE I
SYMBOLS

Symbol	Description
λ(t)	failure rate
μ(t)	replacement element
ν(t)	intensity of the repair
n	number of spare elements

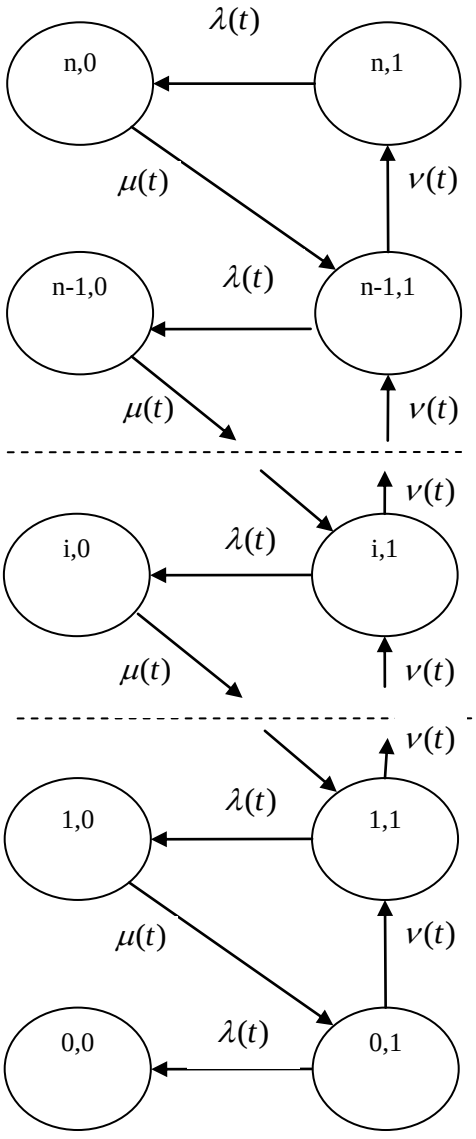


Fig. 1. Transition graph of renewal and repaired element.

The calculation of the probability of finding each of the intermediate states $P_{i,j}(t)$, where $i=1,n,\ j=0,1$ is of interest in this problem. The problem of determining of the probability of falling into a state $P_{0,0}(t)$ characterized by the fact that the basic element is failed and there are no spare elements is very important. The above problem will be solved by methods of simulation. To describe the dependence of the intensity of the elements failures we use the model of the geometric process.

III. THE MODEL OF THE GEOMETRIC PROCESS

Let us consider the model describing the variation of object reliability characteristics and taking into account incomplete repair of operability after failure [1-4]. The behavior of complex systems is well described by this model.

Let's consider the following strategy of the object operation. The object is functioning properly for random time. It is restored after the failure. It is understood that the restoration is not complete. Incompleteness of restoration is characterized by the coefficient γ. As a result of incomplete repair the operating time of the renewal object is reduced (by probability) by γ times in comparison with the previous operation phase:

$$\xi_2^{\ d} = \gamma\xi_1, \ldots \xi_n^{\ d} = \gamma^{n-1}\xi_1, 0 < \gamma \le 1.$$

Mathematical dependence between distribution functions of the operating time between failures of the restored object (taking into account incomplete repair) can be expressed as

$$F_{\xi_2}(t) = F_{\xi_1}\left(\frac{t}{\gamma}\right), \ldots, F_{\xi_1}\left(\frac{t}{\gamma^{n-1}}\right),$$

where $F_{\xi_i}(t)$ is the distribution function of the operating time between failures of $(i-1)$ times restored object and γ is the coefficient of incomplete restoration.). Then, distribution densities are related through the following equation

$$f_{\xi_n}(t) = \frac{1}{\gamma^{n-1}} f_{\xi_1}\left(\frac{t}{\gamma^{n-1}}\right).$$

We define the inverse value of the coefficient of incomplete restoration $\alpha = \frac{1}{\gamma}$ and call it the degradation coefficient. The degradation coefficient α is an average value that reflects the accumulated process of damages and defects and indirectly characterizes the process of gradual material weariness, physical ageing, wear ability, corrosion, etc.

In some cases, α can be understood as a factor, reflecting the increased load on the object due to variable operating conditions.

Definition. *Let* $\{\xi_i\}, i \geq 1$ *be the sequence of independent random variables. Each* ξ_i *corresponds to operating time between failures of the object with the distribution function* $F_{\xi_i}(t)$ *generated by the distribution* $F(t)$ *as follows*

$$F_{\xi_i}(t) = F\left(\frac{t}{\gamma^{i-1}}\right), i = 1,2,...,$$

where γ *is a positive constant. Then the sequence* $\{\xi_i\}, i \geq 1$ *is called geometric process.*

Let us define the expression that establishes the relationship between the failure rate at the initial stage of operation and the failure rate after the $(n-1)^{th}$ failure. By definition the failure rate can be defined by

$$\lambda(t) = \frac{f(t)}{1 - F(t)}.$$

Then the expression of the failure rate of $(n-1)$ times restored object can be written as

$$\lambda_{\xi_n}(t) = \frac{f_{\xi_n}(t)}{1 - F\xi_n(t)} = \frac{\frac{1}{\gamma^{n-1}} f_{\xi_1}\left(\frac{1}{\gamma^{n-1}}\right)}{1 - F_{\xi_1}\left(\frac{t}{\gamma^{n-1}}\right)} = \frac{1}{\gamma^{n-1}} \lambda_{\xi_1}\left(\frac{t}{\gamma^{n-1}}\right) (1)$$

Thus, after each restoration the failure rate becomes $\frac{1}{\gamma}$ times more than the failure rate during the previous time interval. The time scale of the process also changes.

IV. TASK SOLUTION

The simulation process is organized in accordance with the description provided in Section 1. The object is available at the initial time. Further, the failure of the object is occurred in the random time. Failed element is removed from operation, is sent being repaired, and the element from the spare set is put into its place. After repairing object is returned into the spare elements, and is put of the end of the queue to use. That is, the next time it will be installed in the system only after all the spare elements that are before it have worked out to failure. Similarly, operation process of the installation of a system, failure, repair and return of spare parts for all other elements are organized.

Note, that the object has the failure intensity $\lambda_{\xi_1}(t)$ in the first cycle, intensity varies according to the expression (1) after the first failure and is $\lambda_{\xi 2}(t)$, after the i-th failure, it will be equal $\lambda_{\xi_{i+1}}(t)$.

We represent the results of test calculations, which were carried out using the described method. The following values were used as the initial data for calculating:

- failure intensity of the element is λ (t) = 0.001*1/hr,
- the intensity of the element replacement is μ(t) = 1*1/hr,
- the intensity of repair is ν(t) = 0.1*1/hr.

The number of simulation is N = 10^6. Operating time of the redundant structure is equal to 16.000 hours.

At the first stage we will carry out calculations of changing of an average operating time of the element depending on the number of renewals. In this case we change the coefficient of degradation. The calculations considered a structure with a basic element and three spares. Figure 2 shows plots of the average time to failure of the element depending on the number of renewals of the elements for different values of degradation coefficient.

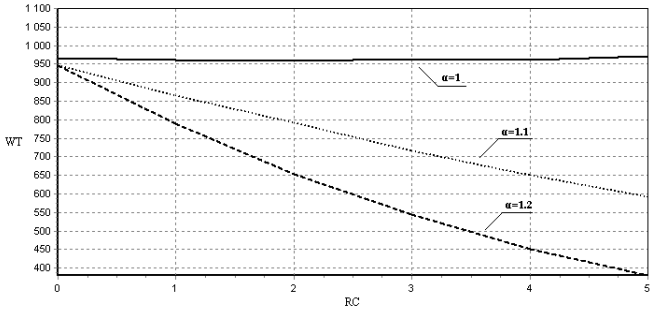


Fig. 2. Changing the average time to failure of the object depending on the number of renewals.

WT = work time, RC = recovery time, α= degradation coefficient.

Several lines of the average time to failure of the object for different α coefficients are presented on the figure.

We can see at the graphs average time to failure does not change for elements without taking into account the aging and it is equal to 1000 hours. A situation characterized by the fact that the greater the degradation coefficient is, the faster average time to failure is decrease for all other graphs. Availability coefficient to 16,000 hours of operation work was calculated for other parameter. The calculation results are presented in Table II.

TABLE II
PARAMETERS OF CALCULATION

Degradation coefficient α	1	1.05	1.1	1.15	1.2
Availability	0.999	0.9989	0.9988	0.9986	0.9979

The calculation results show that the availability coefficient decreases with increasing of degradation coefficient.

The proposed mathematical model can be used to solve the problem of calculating the required number of spare elements that guarantee the achievement of specified reliability parameters on the considered time interval of operation of the structure. We take availability coefficient as an indicator of the

reliability. Let it is necessary provide value of availability coefficient not less than 0.99 of the time of 16.000 hours The results of these calculations are shown in Table III.

TABLE III
PARAMETERS OF CALCULATION

Degradation coefficient α	1	1/1.05	1/1.1	1/1.15	1/1.2
The number of spare elements to ensure specified requirements	2	2	2	2	3
Availability coefficient	0.999	0.9989	0.9988	0.9986	0.9979

As can be seen from Table III the structure with one basic and two spare elements requirements specified by the value of the availability coefficient satisfies for degradation coefficients α = 1 – 1. 15, and three spare elements are for α = 1. 2.

CONCLUSION

Thus, we can say that a model for calculating the reliability redundant structure, taking into account the aging elements presents in this paper. It has been suggested that either a basic element or spare elements spend a part of there resource in the operation process. The result is a partial renewal of the system availability takes place. Investigations of the behavior of the system parameters, depending on the values of the degradation were fulfilled in the test examples. It is shown that the developed model can be used to calculate the required number of spare elements that guarantee the achievement of specified parameters of the reliability on the considered time interval of the structure operation.

REFERENCES

[1] A.V. Antonov, A. A. Poliakov. One Statistical Age-Dependent Reliability Model in Operating of Nuclear Power Plant Equipment. July 1 - 4, 2007. MMR' 2007. International Conference on Mathematical Methods in Reliability. "Methodology, Practice and Interference". Glasgow, GB, 2007.

[2] M. Finkelstein. A scale model of general repair. Microelectronics and Reliability, 1993, p. 41-46.

[3] Y. Lam. Some limit theorems in geometric processes. Acta Mathematicae Applicatae Sinica, 2003, p. 405-416.

[4] Y. Lam. Geometric process and the replacement problem. Acta Mathematicae Applicatae Sinica, 1988, p. 366-382.

Local generalizing ability in Manifold Learning Problem

Alexander Bernstein
Institute for System Analysis
Russian Academy of Sciences
Moscow, Russia
a.bernstein@mail.ru

Abstract*. This paper addresses the Manifold Learning Problem, which is stated as follows: given a finite number of points belonging to a q-dimensional data manifold $\mathbf{X}_\omega \subset R^p$, construct an embedding mapping h from $\mathbf{X}_\omega$ to a q-dimensional dataset $\mathbf{Y} = h(\mathbf{X}_\omega) \subset R^q$ and a reconstruction mapping g from the dataset Y to R^p which provide approximate equalities $g(h(X)) \approx X$ for all $X \in \mathbf{X}_\omega$. A local lower bound is obtained for the maximum reconstruction error $\max\{\|X - g(h(X))\|: X \in U_\varepsilon(X_0)\}$ in the ε-neighborhood $U_\varepsilon(X_0) = \{X \in \mathbf{X}_\omega: \|X - X_0\| \le \varepsilon\}$ of an arbitrary point $X_0 \in \mathbf{X}_\omega$. The lower bound is defined in terms of the distance between q-dimensional linear spaces in R^p that are tangent spaces to the data manifold $\mathbf{X}_\omega$ and to the empirical q-dimensional dataset-based manifold $\mathbf{X}_{emp}$ at the points $X_0 \in \mathbf{X}_\omega$ and $g(h(X)) \in \mathbf{X}_{emp}$, respectively.

Key words: Dimension Reduction, Manifold Learning Problem, Manifold Data Model, projection 2-norm in Grassmann manifold.

I. INTRODUCTION

A variety of problems grouped under the common title 'Dimension Reduction Problem (DR)' and formulated in various ways have been a subject of intensive research over the last decade. The problem considered in [1] – [9] consists in constructing a mapping h of a multidimensional dataset $\mathbf{X}_n = \{X_1, X_2, \dots, X_n\} \subset R^p$ to a space of a lower dimension $q < p$ such that the results $\mathbf{Y}_n = h(\mathbf{X}_n) = \{y_1, y_2, \dots, y_n\} \subset R^q$ of the mapping faithfully represent the multidimensional input data $\mathbf{X}_n$ while inheriting the specified subject-driven data properties. For example, the procedure h is required to preserve local data geometry, proximity relations, geodesic distances, etc. It is natural to refer to the problem thus defined, which amounts to the construction of a mapping h for the points from the dataset $\mathbf{X}_n$ only, as the Embedding Problem, or E-problem.

[10] - [12] examine an Extended Embedding Problem (EE-Problem), where the mapping

$$h: \mathbf{X} \subset R^p \to \mathbf{Y} = h(\mathbf{X}) \subset R^q, \tag{1}$$

must be constructed not only for points from the dataset $\mathbf{X}_n$ but also for new out-of-sample points $X_{new} \in \mathbf{X} / \mathbf{X}_n$, where $\mathbf{X} \subset R^p$ is a dataset the new points are picked from. The E-Problem could certainly be solved for each new point X_{new} with regard to the dataset $\{\mathbf{X}_n \cup X_{new}\}$; however, the embedding $\mathbf{Y}_n = h(\mathbf{X}_n)$ obtained previously for the initial sample $\mathbf{X}_n$ will not be preserved in the general case.

Some applied problems impose a requirement to prevent substantial data losses when using a reduced q-dimensional vector $y = h(X)$ instead of the initial p-dimensional vector X, which means that X can be approximately reconstructed given y. If put otherwise, the problem can be defined as building an embedding h (1) and a reconstruction mapping

$$g: \mathbf{Y} \subset R^q \to \mathbf{X} \subset R^p, \tag{2}$$

such that the pair of mappings $\theta = (h, g)$ ensures the approximate equality

$$r_\theta(X) \approx X \quad \text{for all } X \in \mathbf{X}, \tag{3}$$

where the mapping

$$r_\theta(X) = g(h(X)) \tag{4}$$

is a result of successively applying the embedding and reconstruction mappings to the vector $X \in \mathbf{X}$. The DR-problem thus defined will be referred to as the Full DR-problem, and the pair of mappings $\theta = (h, g)$, as its solution. In this work, we shall focus on the Full DR-problem.

Let $\delta_\theta(X) = \|X - r_\theta(X)\|$ be a reconstruction error in approximate equality (3) at the point $X \in \mathbf{X}$. The quantity $\delta_\theta(X)$ is checked for the sample points $\mathbf{X}_n$, and for $X \in \mathbf{X} \setminus \mathbf{X}_n$ it describes the generalization ability of the procedure θ at the point X. Assume that $X_0 \in \mathbf{X}$ is some selected point and

$$\delta_\theta(X_0, \varepsilon) = \max\{\delta_\theta(X): X \in U_\varepsilon(X_0)\} \tag{5}$$

is the maximum reconstruction error in the ε-neighborhood

$$U_\varepsilon(X_0) = \{X \in \mathbf{X}_\omega: \|X - X_0\| \le \varepsilon\}$$

of the point X_0. For $X_0 \in \mathbf{X}_n$, quantity (5) characterizes the local generalization ability of the procedure θ in the neighborhood of the point X_0 from the sample.

The author is partially supported by Laboratory for Structural Methods of Data Analysis in Predictive Modeling, MIPT, RF government grant, ag. 11.G34.31.0073.

The local estimate obtained in this work for quantity (5) helps establish further requirements to the Full DR-problem solution.

II. Manifold Learning Problem

The definitions of the EE- and Full DR-problems use values of the function h (1) for the out-of-sample points $X \in \mathbf{X} / \mathbf{X}_n$ and values of the function g (2) for the points $y = h(X) \notin \mathbf{Y}_n$ that are images of the out-of-sample points. Thus we have to define a Data Model ω describing the dataset $\mathbf{X}_\omega \subset R^p$ and a Sampling Model s(ω) offering a way for extracting both the sample $\mathbf{X}_n$ and the out-of-sample points $X \in \mathbf{X}_\omega / \mathbf{X}_n$ from the dataset $\mathbf{X}_\omega$. The most popular models in the DR problem ([12] - [20] et al.) are Manifold Data Models ω, in which the datasets

$$\mathbf{X}_\omega = \{X = f(b) \in R^p: b \in \mathbf{B} \subset R^q\} \subset R^p \quad (6)$$

are q-dimensional manifolds in R^p covered by a single coordinate chart that are called Data manifolds. The model ω is identified with either

- a pair (**B**, f) consisting of an open subset $\mathbf{B} \subset R^q$, called a coordinate space, and a smooth mapping f of the dataset **B** in R^p, or

- a parametrized manifold $\mathbf{M}_\omega = (\mathbf{X}_\omega, \tau_f)$, where the **parametrization function** $\tau_f = f^{-1}$ maps the data manifold $\mathbf{X}_\omega$ to the coordinate space $\mathbf{B} = \tau_f(\mathbf{X}_\omega)$.

In this work, we shall investigate the Full DR-problem based on Manifold Data Models, which is called the Manifold Learning Problem.

The Sampling Model s(ω) is typically defined as a probability space $(\mathbf{X}_\omega, \sigma(\mathbf{X}_\omega), \mu_\omega)$ with a probability measure μ_ω in the σ-algebra $\sigma(\mathbf{X}_\omega)$ of measurable subsets. In this space, the support $\mathrm{Supp}(\mu_\omega)$ of the measure μ_ω coincides with $\mathbf{X}_\omega$. In accordance with the model, the data $\mathbf{X}_n = \{X_1, X_2, \dots, X_n\} \subset \mathbf{X}_\omega$ and out-of-sample points $X \in \mathbf{X}_\omega / \mathbf{X}_n$ are selected from the manifold $\mathbf{X}_\omega$ independently of each other according to the probability measure μ_ω.

In R^p, reconstruction mapping g (2) defines an empirical manifold

$$\mathbf{X}_{emp} = \{X = g(y) \in R^p: y \in \mathbf{Y} \subset R^q\} \quad (7)$$

which is a q-dimensional manifold covered by a single q-dimensional coordinate system (chart) g with the domain of definition (coordinate space) $\mathbf{Y} = h(\mathbf{X}_\omega)$. Note that r_θ (4) is a mapping of the data manifold $\mathbf{X}_\omega$ (6) to the empirical manifold $\mathbf{X}_{emp}$ (7).

If the Data Manifold $\mathbf{X}_\omega$ lies in a tube $\mathrm{Tube}(\mathbf{X}_{emp})$ in the empirical manifold $\mathbf{X}_g$ composed of points in R^p such that their projections onto $\mathbf{X}_{emp}$ are unique, one can consider a new solution $\theta(g) = (h_g, g)$, where

$$h_g(X) = \arg\min\{||X - g(y)||, y \in \mathbf{Y}\} \text{ for } X \in \mathrm{Tube}(\mathbf{X}_{emp}) \quad (8)$$

is the projection function onto the empirical manifold $\mathbf{X}_{emp}$. By definition, we have the inequality

$$||X - r_{\theta(g)}(X)|| \le ||X - r_\theta(X)|| \text{ for } X \in \mathbf{X}_\omega, \quad (9)$$

where $r_{\theta(g)}(X) = g(h_g(X))$.

III. Local generalization ability theorem

Denote by $T_\omega(X)$ and $T_{emp}(X)$ the tangent subspaces to the manifolds $\mathbf{X}_\omega$ and $\mathbf{X}_{emp}$ at the points $X \in \mathbf{X}_\omega$ and $r_\theta(X) \in \mathbf{X}_{emp}$, respectively. The tangents, which are affine linear subspaces of dimension q in R^p, can be represented in the form of direct sums

$$T_\omega(X) = X \oplus L_\omega(X), \quad T_{emp}(X) = r_\theta(X) \oplus L_\theta(X), \quad (10)$$

where $L_\omega(X) = \mathrm{Span}(F(\tau_f(X)))$ and

$$L_\theta(X) = \mathrm{Span}(G(h(X))), \quad (11)$$

are linear subspaces of dimension q in R^p that are spanned by columns of Jacobians F(b) and G(y) of the mappings f(b) and g(y) at the points $b = \tau_f(X)$ and $y = h(X)$, respectively. In what follows, the linear spaces $L_\omega(X)$ and $L_\theta(X)$ will be treated as elements of the Grassmann manifold Grass(p, q) composed of q-dimensional linear subspaces in R^p [21].

For elements $L, L' \in \mathrm{Grass}(p, q)$ of the Grassmann manifold Grass(p, q), the quantity

$$d_{P,2}(L, L') = ||P_L - P_{L'}||_2 = \sin\zeta_{max}(L, L') \quad (12)$$

is a metric on the Grassmann manifold [22], [23]; here P_L and $P_{L'}$ are projectors onto the linear subspaces L and L′, and $\zeta_{max}(L, L')$ is the maximum principal angle [24], [25] between the subspaces L and L′. Metric (12) is also called the projection metric in the 2-norm [23] (projection 2-norm [26], [27]), or the Min Correlation Metric in statistics [24].

The following theorem and its corollary hold true.

Theorem. If the data manifold $\mathbf{X}_\omega$ lies in the tube $\mathrm{Tube}(\mathbf{X}_{emp})$ of the empirical manifold $\mathbf{X}_{emp}$, and if h and g are smooth full-rank mappings, then the following inequality holds for the local maximum reconstruction error as $\varepsilon \to 0$:

$$\delta^2_{\theta(g)}(X_0, \varepsilon) \ge \delta^2_{\theta(g)}(X_0) + \varepsilon^2 \times d^2_{P,2}(L_\omega(X_0), L_{\theta(g)}(X_0)) + o(\varepsilon^2); \quad (13)$$

where $L_{\theta(g)}(X) = \mathrm{Span}(G(h_g(X)))$; here and in what follows, the $o(\cdot)$ symbol is understood componentwise in the vector case. For $\delta_{\theta(g)}(X_0) = 0$, inequality (13) turns into equality.

Taking into account (9), the Theorem establishes a lower bound for the quantity $\delta_\theta(X_0, \varepsilon)$ (5).

IV. BRIEF PROOF OF THE THEOREM

Assume that $X \in U_\varepsilon(X_0)$ and $y_0 = h_g(X_0)$, $y = h_g(X)$. Then the Taylor formula yields

$$r_{\theta(g)}(X) = r_{\theta(g)}(X_0) + G(y_0) \times (y - y_0) + o(X - X_0),$$

and it follows from (8) that

$$y = y_0 + (G^T(y_0) \times G(y_0))^{-1} \times G^T(y_0) \times (X - r_{\theta(g)}(X_0)) + o(X - X_0).$$

Let

$$G(y) = Q_G(y) \times D_G(y) \times (V_G(y))^T$$

be a Singular Value Decomposition (SVD) [22] of the p×q matrix G(y), where $Q_G(y)$ is a p×q orthogonal matrix; then

$$r_{\theta(g)}(X) = r_{\theta(g)}(X_0) + \pi(y_0) \times (X - r_{\theta(g)}(X_0)) + o(X - X_0), \quad (14)$$

where $\pi(y_0) = Q_G(y_0) \times (Q_G(y_0))^T$ is the projector onto the linear space $L_{\theta(g)}(X_0)$ (11).

It follows from (8) that $(X_0 - r_{\theta(g)}(X_0)) \in (L_{\theta(g)}(X_0))^\perp$; hence,

$$\pi(y_0) \times (X - r_{\theta(g)}(X_0)) = \pi(y_0) \times (X - X_0),$$

and (14) takes the form

$$r_{\theta(g)}(X) = r_{\theta(g)}(X_0) + \pi(y_0) \times (X - X_0) + o(X - X_0),$$

whence comes the relation

$$X - r_{\theta(g)}(X) = (X_0 - r_{\theta(g)}(X_0)) + \pi^\perp(y_0) \times (X - X_0) + o(X - X_0), \quad (15)$$

where $\pi^\perp(y_0)$ is the projector onto the linear space $(L_{\theta(g)}(X_0))^\perp$.

Let $F(b) = Q_F(b) \times D_F(b) \times (V_F(b))^T$ be an SVD-decomposition of the p×q matrix F(b), where $Q_F(b)$ is a p×q orthogonal matrix. Consider a q×q matrix $(Q_G(y_0))^T \times Q_F(b_0)$ such that its SVD-decomposition has the form

$$(Q_G(y_0))^T \times Q_F(b_0) = O_1 \times \mathrm{Diag}(\cos(\zeta)) \times (O_2)^T,$$

where O_1 and O_2 are q×q orthogonal matrices and where the diagonal elements of the diagonal matrix

$$\mathrm{Diag}(\cos(\zeta)) = \mathrm{Diag}(\cos(\zeta_q), \cos(\zeta_{q-1}), \dots, \cos(\zeta_1))$$

are cosines of the principal angles [26], [27] between the subspaces $L_\omega(X_0)$ and $L_{\theta(g)}(X_0)$ arranged in ascending order:

$$0 \le \zeta_1 \le \zeta_2 \le \dots \le \zeta_q \le \pi/2.$$

The columns $\{t_{F,1}, t_{F,2}, \dots, t_{F,q}\}$ and $\{t_{G,1}, t_{G,2}, \dots, t_{G,q}\}$ of the p×q orthogonal matrices $Q_F(b_0) \times O_2$ and $Q_G(y_0) \times O_1$ are principal vectors [26], [27] of the subspaces $L_\omega(X_0)$ and $L_{\theta(g)}(X_0)$ and determine orthonormal bases for these subspaces, where

$$(t_{F,i}, t_{G,j}) = \delta_{ij} \times \cos(\zeta_j), \; i, j = 1, 2, \dots, p.$$

Taking into account the Taylor series expansion

$$X = X_0 + F(b_0) \times (\tau_f(X) - \tau_f(X_0)) + o(\varepsilon) \quad (16)$$

we obtain

$$\pi^\perp(y_0) \times (X - X_0) = \sum_{j=1}^{q} t_{F,G,j} \times \alpha_j(X),$$

where

$$t_{F,G,j} = \pi^\perp(y_0) \times t_{F,j} = t_{F,j} - t_{G,j} \times \cos\zeta_{q+1-j}, \; j = 1, 2, \dots, q,$$

are projections of the principal vectors $\{t_{F,1}, t_{F,2}, \dots, t_{F,q}\}$ onto the subspace $(L_{\theta(g)}(X_0))^\perp$ and

$$\alpha(X) = (O_2)^T \times D_F(b_0) \times (V_F(b_0))^T \times (\tau_f(X) - \tau_f(X_0)) \equiv$$

$$\equiv (\alpha_1(X), \alpha_2(X), \dots \alpha_q(X))^T.$$

Then it follows from (16) that the vector $\alpha(X)$ satisfies the condition

$$\max\{\|\alpha(X)\|, X \in U_\varepsilon(X_0)\} = \varepsilon + o(\varepsilon).$$

Taking into account the introduced notation and obtained relations, we get from (15) that

$$\delta^2_{\theta(g)}(X) = \sum_{j=1}^{q} \{A_j(X_0) + \alpha_j(X) \times \sin\zeta_{q+1-j}\}^2 + o(\|X - X_0\|^2),$$

where $A_1(X_0), A_2(X_0), \dots A_q(X_0)$ are components of the vector $\pi^\perp(y_0) \times (X_0 - r_{\theta(g)}(X_0))$.

It can be shown that the relation

$$\max\left\{\sum_{j=1}^{q}\left\{A_j(X_0)+\alpha_j(X)\times\sin\zeta_{q+1-j}\right\}^2, X\in U_\varepsilon(X_0)\right\}=$$

$$=\max\left\{\sum_{j=1}^{q}\left\{\left|A_j(X_0)\right|+\left|\alpha_j(X)\right|\times\sin\zeta_{q+1-j}\right\}^2, X\in U_\varepsilon(X_0)\right\}+o(\varepsilon^2)\geq$$

$$\geq\delta^2_{\theta(g)}(X_0)+\max\left\{\sum_{j=1}^{q}\alpha^2_j(X)\times\sin^2\zeta_{q+1-j}, |\alpha(X)|\leq\varepsilon\right\}+o(\varepsilon^2),$$

is valid, whence it follows that

$$\delta^2{}_{\theta(g)}(X_0,\varepsilon)\geq\delta^2{}_{\theta(g)}(X_0)+\varepsilon^2\times\sin^2\zeta_{\max}(L_\omega(X_0),L_{\theta(g)}(X_0))+o(\varepsilon^2),$$

which proves the Theorem

From the proof of the Theorem given above, one can derive the following corollary.

Corollary of the Theorem. For $X\in U_\varepsilon(X_0)$ and $\varepsilon\to 0$, we have the following asymptotic inequalities:

$$||(X-X_0)-(r_{\theta(g)}(X)-r_{\theta(g)}(X_0))||\leq$$

$$\leq||X-X_0||\times d_{P,2}(L_\omega(X_0),L_{\theta(g)}(X_0))+o(||X-X_0||), \qquad (17)$$

and

$$||X-X_0||\times(1-d^2{}_{P,2}(L_\omega(X_0),L_{\theta(g)}(X_0)))^{1/2}+o(||X-X_0||)\leq$$

$$\leq||r_{\theta(g)}(X)-r_{\theta(g)}(X_0)||\leq||X-X_0||+o(||X-X_0||). \qquad (18)$$

The left-hand side of the first inequality (17) indicates to what extent the mapping $r_{\theta(g)}$ preserves the local structure of the data manifold, while the second inequality (18) characterizes the local non-isometricity of the mapping.

V. Conclusion

It follows from the above formulas that the greater the distances between the linear subspaces $L_\omega(X_i)$ and $L_{\theta(g)}(X_i)$ at the sample points X_i, i = 1, 2, … , n, the lower the local generalization ability of the obtained solution becomes, the poorer the local structure of the data manifold is preserved, and the poorer the local non-isometricity properties are ensured. Thus, it is natural to require that the dimension reduction procedure θ = (h, g) ensures not only proximity (3) between the points $X\in\mathbf{X}_\omega$ and their images $r_{\theta(g)}(X)\in\mathbf{X}_{emp}$ but also proximity between the linear manifolds $L_\omega(X)$ and $L_{\theta(g)}(X)$ in the selected metric on the Grassmann manifold Grass(p, q), the latter proximity being naturally referred to as tangential proximity.

We note in conclusion that, taking into account (3) and (10), it follows from the tangential proximity that the tangent spaces $T_\omega(X)$ and $T_{emp}(X)$ are close too.

References

[1] S.T. Roweis and L.K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," Science, vol. 290(5500), pp. 2323 – 2326, December 2000.

[2] J.B. Tehenbaum, V. de Silva and J.C. Langford, "A global geometric framework for nonlinear dimensionality reduction," Science, 290(5500), pp. 2319 – 2323. December 2000.

[3] M. Belkin and P. Niyogi, "Laplacian eigenmaps and Spectral Techniques for Embedding and Clustering," Advances in Neural Information Processing Systems (NIPS 2001), MIT Press, Cambridge, MA, vol. 14, 2002.

[4] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," Neural Computation, vol. 15(6), pp. 1373 – 1396, June 2003.

[5] D.L. Donoho and C. Grimes, "Hessian eigenmaps: New locally linear embedding techniques for high-dimensional data", in Proceedings of the National Academy of Arts and Sciences, vol. 100, pp. 5591 – 5596, 2003.

[6] K. Q. Weinberger and L. K. Saul, "Maximum Variance Unfolding: Unsupervized Learning of Image Manifolds by Semidefinite Programming," International Journal of Computer Vision, vol. 70(1), pp. 77 – 90, 2006.

[7] L.K. Saul and S.T. Roweis, "Think globally, fit locally: unsupervised learning of low dimensional manifolds," Journal of Machine Learning Research, vol. 4, pp. 119 – 155, 2003.

[8] L. K. Saul, K. Q. Weinberger, J. H. Ham, F. Sha, and D. D. Lee, "Spectral methods for dimensionality reduction", in Semisupervised Learning, O. Chapelle, B. Schoelkopf, and A. Zien (eds.), MIT Press, Cambridge, MA, 2006, pp. 293-308.

[9] Jihun Ham, Daniel D. Lee, Sebastian Mika and Bernhard Schkolkopf, "A kernel view of the dimensionality reduction of manifolds," in Proceedings of the Twenty First International Conference on Machine Learning (ICML-04), Banff, Canada, 2004, pp. 369 - 376.

[10] Y. Bengio, O. Delalleau, N. Le Roux, J.-F. Paiement, P. Vincent and M. Ouimet, "Out-of-sample extension for LLE, Isomap, MDS, Eigenmaps, and spectral clustering," Advances in Neural Information Processing Systems (NIPS 2003), MIT Press, Cambridge, MA, vol. 16, 2004.

[11] Y. Bengio, O. Delalleau, N. Le Roux, J.-F. Paiement, P. Vincent and M. Ouimet, "Learning Eigenfunctions Link Spectral Embedding and Kernel PCA," Neural Computation, vol. 16(10), pp. 2197 – 2219, 2004.

[12] Tony Lin, Hougbin Zha and Sang Uk Lee, "Riemannian Manifold Learning for Nonlinear dimensionality reduction," in Proceedings of ECCV 2006, A. Leonardis, H. Bischof, and A. Prinz (Eds.), Springer-Verlag Berlin Heidelberg: Part I, LNCS 3951, 2006, pp. 44-55.

[13] Zhenyue Zhang and Hongyuan Zha, "Principal Manifolds and Nonlinear Dimension Reduction via Local Tangent Space Alignment," SIAM Journal on Scientific Computing, vol. 26 (1), pp. 313–338, 2005.

[14] De-Li Zhao, "Tangential Eigenmaps: A Unifying Geometric for Manifold Learning," Shanghai Jiao Tong University (Unpublished: see http://sites.google.com/site/zhaodeli/paper), Oct. 2005.

[15] Lawrence Cayton, "Algorithms for manifold learning," Univ of California at San Diego (UCSD)Tech Rep CS2008-0923, Publisher: Citeseer, pp. 541 – 555, June 2005.

[16] M. Brand, "Charting a manifold," in Advances in Neural Information Processing Systems (NIPS 2002), S. Becker, S. Thrun, and K. Obermayer, Eds, vol. 15, MIT Press, Cambridge, MA, 2003.

[17] V. Brand, "From subspace to submanifold methods," in Proceedings of British Machine Vision Conference (BMVC 2004), London, Kingston, England, 2004.

[18] Piotr Dollár, Vincent Rabaud, and Serge Belongie, "Learning to Traverse Image Manifolds," in Advances in Neural Information Processing Systems (NIPS 2006), vol. 19, MIT Press, Cambridge, MA, 2007, pp. 361-368.

[19] Anders Brun, Carl-Fredrik Westin, Magnus Herberthson, and Hans Knutsson, "Fast Manifold Learning Based on Riemannian Normal Coordinates," in Proceedings of the 14th Scandinavian conference on image analysis (SCIA'05), Joensuu, Finland, June, 2005, pp. 920 - 929.

[20] Zhenyue Zhang and Hongyuan Zha, "A Domain Decomposition Method for Fast Manifold Learning,' in Advances in Neural Information Processing Systems (NIPS 2005), vol. 18, MIT Press, Cambridge, MA, 2006, pp. 1625-1632.

[21] R.P. Woods, "Diffrrential geometry of Grassmann manifolds," Proc. Nqt. Acad. Sci. USA, vol. 57 (1967), pp. 589 – 594.

[22] G.H. Golub and C.F. Van Loan, Matrix Computation, 2nd ed., Johns Hopkins University Press, Baltimore, 1989; 3rd ed. Baltimore, MD: Johns Hopkins University Press, 1996.

[23] Liwei Wang, Xiao Wang, and Jufu Feng, "Subspace Distance Analysis with Application to Adaptive Bayesian Algorithm for Face Recognition, Pattern Recognition, vol. 39, Issue 3, pp. 456 – 464, 2006.

[24] H. Hotelling, "Relations between two sets of variables," Biometrika, vol. 28, pp. 321 – 377, 1936.

[25] A.T. James, "Normal multivariate analysis and the orthogonal group," Ann. Math. Statistics, v0l. 25, pp. 40 – 75, 1954.

[26] Jihun Hamm and Daniel D. Lee, "Grassmann Discriminant Analysis: a Unifying View on Subspace-Based Learning," in Proceedings of the 25th International Conference on Machine Learning (ICML 2008), July 2006.

[27] A. Edelman, T. A. Arias, and T. Smith, "The Geometry of Algorithms with Orthogonality Constraints," SIAM Journal on Matrix Analysis and Applications, vol. 20(2), pp. 303-353, 1999.

Estimating and comparing areas under time-dependent ROC curves in presence of censoring and competing risks

Paul Blanche
INSERM U897 and
Univ Bordeaux
Bordeaux, France
Email: Paul.Blanche@isped.u-bordeaux2.fr

Hélène Jacqmin-Gadda
INSERM U897 and
Univ Bordeaux
Bordeaux, France
Email: Helene.Jacqmin-Gadda@bordeaux.inserm.fr

***Abstract*—To quantify the ability of a marker to predict the onset of a clinical outcome in the future, area under the time-dependent ROC curve (AUC) is particularly relevant and is becoming more and more used. We first present non parametric Inverse Probability of Censoring Weighting (IPCW) estimators for the two AUC definitions according to how we define a case and a control with competing risk events. Then, we investigate the asymptotic property of the proposed estimators. Consequently, we derive confidence intervals and test statistics for the equality of the AUCs from two markers measured on the same subjects. A simulation study enlighten the finite sample behaviour of the tests and confidence intervals. The method is applied to the French cohort PAQUID to compare the abilities of two psychometric tests to predict dementia onset in the elderly accounting for death competing risk.**

I. Introduction

For many diseases, it would be relevant to identify a marker or a combination of markers that enables the identification of subjects at high and low risk of the disease in the future. In particular, Alzheimer's disease treatments given after the clinical diagnosis have been shown to have modest effects and research is currently focussing on preventive treatment given in the pre-diagnosis phase [1]. To ensure sufficient power to these preventive trials and then to apply the preventive treatment if its efficacy was demonstrated, validated markers for detecting subjects at high risk of Alzheimer's disease in next years would be required.

The diagnostic accuracy of a quantitative marker is often evaluated by the ROC curve that displays the sensitivity (probability that the marker M be above the cutpoint c for a diseased subject) versus 1-specificity (where the specificity is the probability that M be below c for a healthy subject) for all the possible cutpoints c [2]. The diagnostic accuracy is summarized by the Area Under the ROC Curve (AUC) that may be interpreted as the probability that the marker value of a randomly chosen case is above the marker value of a randomly chosen healthy subject [2]. In a diagnostic study, the marker and disease are measured at the same time and are known for all participants. In prognostic studies, the marker is measured at a given time (considered as time 0 in the following) while the disease may occur at any time thereafter. Thus sensitivity, specificity and ROC curve are time-dependent and may be computed for different time durations t (window of prediction).

Time dependent ROC methodology was first introduced by Heagerty et al. [3] to deal with censoring from survival data and then extended for competing risks setting by Saha and Heagerty [4] and Zheng et al. [5]. With competing risks setting, definition of cases is clear but two possibilities exist to define controls. Let $X(t)$ be the stochastic process that indicates the state of the subject at time t (e.g : $X(t) \in \{0, 1, 2\}$ with 0=health, 1=dementia, 2=death) and T the time to the first event defined by $T = \inf\{t \geqslant 0 : X(t) \neq 0\}$.

1) Saha and Heagerty [4] defined cases as ill subjects at time t ($T \leqslant t, X(T) = 1$) and controls as healthy subjects at time t ($T > t$). Consequently, the AUC can be interpreted as the probability :

$$\theta_t = \mathbb{P}\big(M_i > M_j | T_i \leq t, X_i(T_i) = 1, T_j > t\big)$$

with i and j the indices of two independent subjects.

2) Zheng et al. [5] proposed another definition of controls that include both healthy subjects at time t ($T > t$) and subjects who died without illness before time t ($T \leqslant t, X(T) \neq 1$). Consequently, the AUC can be interpreted as the probability :

$$\pi_t = \mathbb{P}\Big(M_i > M_j \Big| T_i \leq t, X_i(T_i) = 1, \{T_j > t\} \cup \{T_j \leq t, X_j(T_j) \neq 1\}\Big)$$

with i and j the indices of two independent subjects.

Saha and Heagerty [4] provided an estimator of θ_t but their methodology required a smoothing parameter that should be carefully chosen by the statistician. Zheng et al. [5] provided semi-parametric estimators of θ_t and π_t.

The first goal of this work is to propose simple inverse probability weighting estimators of θ_t and π_t, with asymptotic results to build confidence intervals. The second one is to provide a methodology to compare the areas under time-dependent ROC curves for two markers measured on the same sample. Indeed, we often have several competing markers to predict the onset of a disease and a relevant question is to know which one is the best. Delong et al. [6] proposed a non parametric test but only for uncensored data and Chiang and Hung [7] proposed a test for censored data without competing risks. The main goal of this work is therefore to propose a test for the difference between two AUCs by extending Chiang and Hung's asymptotic results [7], [8] to the competing risks setting.

II. Methodology

A. Additional notations

Let C denote a censoring time and $\Delta = \mathbb{1}_{(T \leqslant C)}$. Let $\tilde{T} = \min(T, C)$ denote the observed follow-up time and $\tilde{\delta} = \Delta.X(\tilde{T})$ the observed state at the end of follow-up. We observe the i.i.d sample of n subjects $\{(\tilde{T}_i, \tilde{\delta}_i, M_i), i = 1, \ldots, n\}$ of law $(T, \tilde{\delta}, M)$.
We denote $\mathcal{E}_{ij} = \mathbb{1}_{(M_i > M_j)}$, $S_C(t) = \mathbb{P}(C > t)$ and $\hat{S}_C(t)$ the Kaplan-Meier estimator of $S_C(t)$, and $\hat{S}_{\tilde{T}}(t)$ the empirical survival function of $\tilde{T}$.

B. Estimators and i.i.d representations

We define the IPCW estimator of θ_t :

$$\hat{\theta}_t = \frac{\sum_{i \neq j} \frac{\mathbb{1}_{(\tilde{\delta}_i = 1)}}{\hat{S}_C(\tilde{T}_i)} \mathbb{1}_{(\tilde{T}_i \leq t)} \mathbb{1}_{(\tilde{T}_j > t)} \mathcal{E}_{ij}}{n(n-1) \hat{S}_{\tilde{T}}(t) \left(\frac{1}{n} \sum_i \mathbb{1}_{(\tilde{T}_i \leq t)} \frac{\mathbb{1}_{(\tilde{\delta}_i = 1)}}{\hat{S}_C(\tilde{T}_i)} \right)}$$

and the IPCW estimator of π_t:

$$\hat{\pi}_t = \frac{\sum_{i \neq j} \frac{\mathbb{1}_{(\tilde{\delta}_i = 1)} \mathbb{1}_{(\tilde{T}_i \leq t)}}{\hat{S}_C(\tilde{T}_i)} \left(\frac{\mathbb{1}_{(\tilde{T}_j > t)}}{\hat{S}_C(t)} + \frac{\mathbb{1}_{(\tilde{T}_j \leq t)} \mathbb{1}_{(\tilde{\delta}_j \notin \{0,1\})}}{\hat{S}_C(\tilde{T}_j)} \right) \mathcal{E}_{ij}}{\left(\frac{1}{n} \sum_i \frac{\mathbb{1}_{(\tilde{T}_i \leq t)} \mathbb{1}_{(\tilde{\delta}_i = 1)}}{\hat{S}_C(\tilde{T}_i)} \right) \left(\frac{1}{n} \sum_j 1 - \frac{\mathbb{1}_{(\tilde{\delta}_j = 1)} \mathbb{1}_{(\tilde{T}_j \leq t)}}{\hat{S}_C(\tilde{T}_j)} \right)}$$

The inverse probability of censoring weighting technique was previously used by Scheike et al. [9] to estimate cumulative incidence probabilities which are key quantities when we deal with competing risk events.
Following similar arguments to those of Hung and Chiang [8], based on the martingale representation of the Kaplan-Meier estimator of the censoring distribution [10] and usual U-statistics theory [11], we show that :

$$\sqrt{n}(\hat{\theta}_t - \theta_t) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \eta_\theta(\tilde{T}_i, M_i, \tilde{\delta}_i, t) + o_p(1) \quad (1)$$

and

$$\sqrt{n}(\hat{\pi}_t - \pi_t) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \eta_\pi(\tilde{T}_i, M_i, \tilde{\delta}_i, t) + o_p(1) \quad (2)$$

where $\mathbb{E}\left(\eta_\theta(\tilde{T}_i, M_i, \tilde{\delta}_i, t)\right) = 0$ and $\mathbb{E}\left(\eta_\pi(\tilde{T}_i, M_i, \tilde{\delta}_i, t)\right) = 0$.
Therefore we obtain the usual $\sqrt{n}$-consistency and the asymptotic normality of the estimators :

$$\sqrt{n}\left(\hat{\theta}_t - \theta_t\right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \sigma_\theta^2\right)$$

and

$$\sqrt{n}\left(\hat{\pi}_t - \pi_t\right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \sigma_\pi^2\right)$$

with $\sigma_\theta^2 = \text{Var}\left(\eta_\theta(\tilde{T}_i, M_i, \tilde{\delta}_i, t)\right)$ and $\sigma_\pi^2 = \text{Var}\left(\eta_\pi(\tilde{T}_i, M_i, \tilde{\delta}_i, t)\right)$. As we are able to compute consistent estimators $\hat{\eta}_\theta(\tilde{T}_i, M_i, \tilde{\delta}_i, t)$ and $\hat{\eta}_\pi(\tilde{T}_i, M_i, \tilde{\delta}_i, t)$ of $\eta_\theta(\tilde{T}_i, M_i, \tilde{\delta}_i, t)$ and $\eta_\pi(\tilde{T}_i, M_i, \tilde{\delta}_i, t)$, σ_θ^2 and σ_π^2 can be consistently estimated by :

$$\hat{\sigma}_\theta^2 = \frac{1}{n} \sum_i \hat{\eta}_\theta(\tilde{T}_i, M_i, \tilde{\delta}_i, t)^2 \quad (3)$$

and

$$\hat{\sigma}_\pi^2 = \frac{1}{n} \sum_i \hat{\eta}_\pi(\tilde{T}_i, M_i, \tilde{\delta}_i, t)^2 \quad (4)$$

As a consequence we can easily compute confidence intervals.

C. Testing procedure for comparing two AUCs

Let M^1 and M^2 denote two competing markers and $\hat{\theta}_t^1$, $\hat{\pi}_t^1$ and $\hat{\theta}_t^2$, $\hat{\pi}_t^2$ their AUC estimates. From (1) and (2), under the null hypothesis $\mathcal{H}_0 : \theta_t^1 = \theta_t^2$:

$$\sqrt{n}\left(\hat{\theta}_t^1 - \theta_t^2\right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \sigma_{\theta^{12}}^2\right)$$

with $\sigma_{\theta^{12}}^2 = \text{Var}\left(\eta_\theta(\tilde{T}_i, M_i^1, \tilde{\delta}_i, t) - \eta_\theta(\tilde{T}_i, M_i^2, \tilde{\delta}_i, t)\right)$; and under the null hypothesis $\mathcal{H}_0 : \pi_t^1 = \pi_t^2$:

$$\sqrt{n}\left(\hat{\pi}_t^1 - \pi_t^2\right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \sigma_{\pi^{12}}^2\right)$$

with $\sigma_{\pi^{12}}^2 = \text{Var}\left(\eta_\pi(\tilde{T}_i, M_i^1, \tilde{\delta}_i, t) - \eta_\pi(\tilde{T}_i, M_i^2, \tilde{\pi}_i, t)\right)$.

Similarly to σ_θ^2 and σ_π^2 we are also able to consistently estimate $\sigma_{\theta^{12}}^2$ and $\sigma_{\pi^{12}}^2$ by :

$$\hat{\sigma}_{\theta^{12}}^2 = \frac{1}{n} \sum_i \left(\hat{\eta}_\theta(\tilde{T}_i, M_i^1, \tilde{\delta}_i, t) - \hat{\eta}_\theta(\tilde{T}_i, M_i^2, \tilde{\delta}_i, t)\right)^2 \quad (5)$$

and

$$\hat{\sigma}_{\pi^{12}}^2 = \frac{1}{n} \sum_i \left(\hat{\eta}_\pi(\tilde{T}_i, M_i^1, \tilde{\delta}_i, t) - \hat{\eta}_\pi(\tilde{T}_i, M_i^2, \tilde{\delta}_i, t)\right)^2 \quad (6)$$

Confidence intervals for the differences of two AUCs from two markers measured on the same subjects and testing procedure are therefore easily computable.

III. Simulation

A. Simulation scenarios

As described in [12], we simulated the survival time $T = \inf_{t>0}\{X(t) \neq 0\}$ through two cause specific hazard functions

$$\alpha_{0j}(t) = \lim_{dt\to 0} \mathbb{P}\big(T \in [t, t+dt), X(T) = j | T \geq t\big)/dt,$$

for $j = 1, 2$. We simulated two competing markers X_1, X_2, both $\mathcal{N}(0, 1)$ with correlation equals to 0.5. The cause-specific hazard functions $\alpha_{01}(t)$ and $\alpha_{02}(t)$ were defined constant equal to: $\alpha_{01} = \alpha_{001} \exp(\beta_{011}X_1 + \beta_{012}X_2)$ and $\alpha_{02} = \alpha_{002} \exp(\beta_{021}X_1 + \beta_{022}X_2)$. Let α denote $\alpha_{01} + \alpha_{02}$. We simulated T from an exponential distribution $\mathcal{E}(\alpha)$ and the cause of event $X(T)$ by a binomial distribution with probability α_{01}/α on cause 1. Then, we simulated censoring C from an independent exponential distribution.
We simulated several scenarios through several values of Δ_β with $\beta_{011} = 1 + \Delta_\beta$ and $\beta_{012} = 1 - \Delta_\beta$, and several sample sizes n.

B. Simulation results

We estimated the coverage probabilities of the 95% confidence intervals for θ_t and π_t, the type I error under the null hypotheses $\mathcal{H}_0 : \pi_t^1 = \pi_t^2$ and $\mathcal{H}_0 : \theta_t^1 = \theta_t^2$ and the power of the tests under several alternative hypotheses with increasing differences $\pi_t^1 - \pi_t^2$ and $\theta_t^1 - \theta_t^2$ from 0.05 to 0.15.
We compared finite sample behavior of the tests and confidence intervals when standard deviations σ_θ, σ_π, $\sigma_{\theta^{12}}$ and $\sigma_{\pi^{12}}$ were estimated using the formula (3) to (6) or by Bootstrap.
Simulation results show very good coverage probabilities of the confidence intervals and type I errors were close to the nominal value 0.05. Under the alternative hypotheses, simulations show reasonable power of the testing procedure that is similar to the popular Delong et al. [6] procedure for uncensored data. Bootstrap and proposed estimates of the asymptotic standard deviations lead to very close results.

IV. Application : comparing psychometric tests for predicting dementia in the elderly

A. The PAQUID cohort

The PAQUID cohort is a French prospective study on cognitive ageing including 3777 subjects aged 65 years and older and living at home at baseline [13]. Subjects were initially interviewed at home in 1988 and 1,3,5,8,10,13,15,17 and 20 years later. The cognition was evaluated at each visit using a battery of cognitive tests including the Digit Symbol Substitution Test (DSST) [14] and the Isaac Set Test (IST) [15].

B. Application and results

We estimated the areas under the time dependent ROC curve π_t and θ_t for the DSST and IST for different clinically relevant windows of time t such as $t = 5$ or 10 years. The DSST has a significantly better predictive accuracy than the IST.
Our results show a larger values for AUCs estimates for the definition θ_t compared with the definition π_t. These results are due to the fact that poor scores to cognitive test are also predictive of death. Indeed, θ_t measures discrimination between demented subjects before time t and subjects alive and free of dementia at time t, whereas π_t measures discrimination between demented subjects before time t and subjects alive and free of dementia at time t or dead without dementia before t. In practice, both definitions are interesting, but from our point of view the most relevant to define criteria for enrollment in preventive clinical trial should be π_t.

V. Discussion

We provide a new methodology for inference about predictive accuracy of biomarker under competing risks and presence of censoring through area under the time-dependent ROC curve.
Contrary to Zheng et al. [5], the strength of our methodology is to be fully non-parametric. Therefore no assumption about the link between marker and cause specific hazard functions such as proportional hazards assumption is required. However, contrary to them, our methodology does not allow adjustment on covariates.
By contrast to Saha and Heagerty [4] we propose two non-parametric estimators corresponding to two definition of AUCs depending on how we define cases and controls. We think this is important because in epidemiology both can be of interest as illustrated in our application. Moreover, we provide asymptotic results to build confidence intervals and make comparison tests.
We can note that the test we proposed is a direct extension of the popular Delong et al. [6] test that is frequently used for the statistical evaluation of biomarkers [2]. Indeed, as they are both based on the same U-statistic theory, without competing risk nor censoring, our estimators, confidence bands and testing procedures are equivalent to those of Delong et al. [6].

This method assumes censoring is independent from the marker. This assumption is sometimes too strong. Even if this kind of IPCW estimators are quite robust to marker-dependent censoring [7], [16], we could also compute the weights accounting for marker-dependent censoring. However, due to the curse of dimensionality only semi-parametric model such as Cox model could be used for these weights. Therefore assumption about the link between censoring and markers are required. This extension to marker-dependent censoring as well as extension to interval censoring due to intermittently observed data will be investigate in future work.

Acknowledgments

This work was partly funded by a grant from France Alzheimer awarded to Helene Jacqmin-Gadda in 2009. We also thank the PAQUID team for allowing us to use their data.

References

[1] B. Vellas, S. Andrieu, P. Ousset, M. Ouzid, and H. Mathiex-Fortunet, "The guidage study: Methodological issues. a 5-year double-blind randomized trial of the efficacy of egb 761 (r) for prevention of alzheimer disease in patients over 70 with a memory complaint," *Neurology*, vol. 67, no. 9, Supplement 3, p. S6, 2006.

[2] M. Pepe, *The statistical evaluation of medical tests for classification and prediction*. Oxford University Press, USA, 2004.
[3] P. Heagerty, T. Lumley, and M. Pepe, "Time-dependent ROC curves for censored survival data and a diagnostic marker," *Biometrics*, vol. 56, no. 2, pp. 337–344, 2000.
[4] P. Saha and P. Heagerty, "Time-dependent predictive accuracy in the presence of competing risks," *Biometrics*, vol. 66, no. 4, pp. 999–1011, 2010.
[5] Y. Zheng, T. Cai, Y. Jin, and Z. Feng, "Evaluating Prognostic Accuracy of Biomarkers under Competing Risk," *Biometrics*, 2011.
[6] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson, "Comparing the Areas Under Two or More correlated Receiver Operating characteristic Curves : A Nonparametric Approach," *Biometrics*, pp. 837–845, 1988.
[7] C. Chiang and H. Hung, "Non-parametric estimation for time-dependent AUC," *Journal of Statistical Planning and Inference*, vol. 140, no. 5, pp. 1162–1174, 2010.
[8] H. Hung and C. Chiang, "Estimation methods for time-dependent AUC models with survival data," *Canadian Journal of Statistics*, vol. 38, no. 1, pp. 8–26, 2010.
[9] T. Scheike, M. Zhang, and T. Gerds, "Predicting cumulative incidence probability by direct binomial regression," *Biometrika*, vol. 95, pp. 205–220, 2008.
[10] P. K. Andersen, Ø. Borgan, R. D. Gill, and N. Keiding, *Statistical Models Based on Counting Processes*. New York: Springer Verlag, 1992.
[11] R. Serfling, "Approximation theorems of mathematical statistics," *New York*, 1980.
[12] J. Beyersmann, A. Latouche, A. Buchholz, and M. Schumacher, "Simulating competing risks data in survival analysis," *Statistics in medicine*, vol. 28, no. 6, pp. 956–971, 2009.
[13] H. Amieva, H. Jacqmin-Gadda, J. Orgogozo, N. Le Carret, C. Helmer, L. Letenneur, P. Barberger-Gateau, C. Fabrigoule, and J. Dartigues, "The 9 year cognitive decline before dementia of the Alzheimer type: a prospective population-based study," *Brain*, vol. 128, no. 5, p. 1093, 2005.
[14] D. Wechsler, "Wechsler adult intelligence scale (rev. ed.)," *New York: Psychological Corporation*, 1981.
[15] B. Isaacs and A. Kennie, "The set test as an aid to the detection of dementia in old people," *The British Journal of Psychiatry*, vol. 123, no. 575, pp. 467–470, 1973.
[16] P. Blanche, J. Dartigues, and H. Jacqmin-Gadda, "ROC curve estimators for a time-dependent outcome with marker-dependent censoring," *submitted*.

ON GENERAL BOOTSTRAP OF EMPIRICAL ESTIMATOR OF A SEMI-MARKOV KERNEL

SALIM BOUZEBDA AND NIKOLAOS LIMNIOS

Extended abstract

Semi-Markov processes constitute an extension of jump Markov processes and renewal processes. They allow the use of any distributions for the sojourn times instead of the exponential (geometric) distributions of Markov processes (chains). Semi-Markov processes has seen to be a flexible tool and a powerful statistical modeling framework in a variety of applied and theoretical contexts, survival analysis [Andersen *et al.* (1993)], reliability [Limnios (2004)], queueing theory, finance and insurance [Janssen and Limnios (1999)]. The interested reader may refer to Pyke (1961a,b) and for recent sources of references to research literature in this area along with statistical applications consult Limnios and Oprişan (2001) and Barbu and Limnios (2008). In the literature, some nonparametric estimators for semi-Markov kernels are considered in several papers, refer for example to Limnios (2004). Unfortunately, the limiting distribution of these estimators, or their functionals, depend crucially on the unknown parameters, which is a serious problem in practice. To circumvent this matter, we shall propose, in this work, a general bootstrap of empirical semi-Markov kernels and of the conditional transition probabilities and study some of its properties by mean of martingale techniques.

The results obtained in this work are useful in many statistical problems, in the semi-Markov framework, as illustrated in the construction of confidence bands, the change point problem and the computation of the p-value of the test.

To define semi-Markov processes or equivalently Markov renewal processes, it is natural, first, to define semi-Markov kernels (see, for example, Limnios and Oprişan (2001)). Consider an infinite countable set, say E, and an E-valued càdlàg time-homogeneous semi-Markov process Z_t, $t \in \mathbb{R}_+$,with embedded Markov renewal process (J_k, S_k), for $k \in \mathbb{N}$, where (J_k) is the E-valued embedded Markov chain (EMC) of the successive visited states, and $0 = S_0 \leq S_1 \leq \cdots \leq S_k \leq S_{k+1} \leq \cdots$ are the jump times of (Z_t). Define also $X_k := S_k - S_{k-1}, k \geq 1$, the inter-jump times, and the process $N(t)$, $t \in \mathbb{R}_+$, which counts the number of jumps of (Z_t), in the time interval $(0, t]$, by $N(t) := \sup\{k \geq 0 : S_k \leq t\}$. Let us also define $N_i(t)$ as the number of visits of (Z_t) to state $i \in E$ up to time t, and $N_{ij}(t)$ the number of direct jumps of (Z_t) from state i to state j up to time t, [see, e.g., Limnios and Oprişan (2001)]. To be specific, $N_i(t) := \sum_{k=1}^{N(t)} \mathbb{1}_{\{J_{k-1}=i\}}$ and $N_{ij}(t) := \sum_{k=1}^{N(t)} \mathbb{1}_{\{J_k=i, J_{k-1}=j\}}$, here and elsewhere, $\mathbb{1}_A$ stands for the indicator function of the event A. In the case where we consider the renewal process $(S_k^i)_{k\geq 0}$ (eventually delayed, see, e.g., Limnios and Oprişan (2001)) of successive times of visits to state i, then $N_i(t)$ is the counting process of renewals. Denote by μ_{ii} the mean recurrence times of (S_n^i), i.e., $\mu_{ii} = \mathbf{E}[S_2^i - S_1^i]$ and by $\nu = (\nu_i)_{i\in E}$ the stationary distribution of the embedded Markov chain (J_n). Let us denote by $Q(t) = (Q_{ij}(t), i, j \in E)$, $t \geq 0$ the semi-Markov kernel wich is defined by

$$Q_{ij}(t) := \mathbf{P}(J_{k+1} = j, X_{k+1} \leq t | J_k = i) \quad = \quad P(i, j)F_{ij}(t), \ t \geq 0, \ i, j \in E, \tag{0.1}$$

where $P(i, j) := \mathbf{P}(J_{k+1} = j|J_k = i)$ is the transition kernel of the EMC (J_k), and $F_{ij}(t) := \mathbf{P}(X_{k+l} \leq t|J_k = i, J_{k+1} = j)$ is the conditional distribution function of the inter-jump times. Let us define also the distribution function $H_i(t) := \sum_{j\in E} Q_{ij}(t)$ and its mean value m_i, which is the mean sojourn time of (Z_t) in state i, i.e., $m_i = \int_0^\infty (1-H_i(t))dt$. The mean sojourn time of the semi-Markov process (Z_t) is defined to be $\overline{m} := \sum_{i\in E} \nu_i m_i$. In the sequel we need to recall the following useful property $m_{ii} = \overline{m}/\nu_i$. Let us define the following observation in the time interval $[0, t]$,

$$\mathcal{H}_t := \{Z_u, 0 \leq u \leq t\} = \begin{cases} \{J_0, X_1, \ldots, J_{N(t)}, U_t\} & \text{if} \quad N(t) > 0, \\ \{J_0, U_t = t\} & \text{if} \quad N(t) = 0, \end{cases}$$

Date: April 8, 2012.

where $U_t := t - S_{N(t)}$. Define

$$Q_{ij}^W(x,t) := \frac{1}{N_i(t)} \sum_{k=1}^{N(t)} W_{N(t)k} \mathbb{1}_{\{J_{k-1}=i, J_k=j, X_k \le x\}}, \ 0 \le x \le t, \ i,j \in E, \tag{0.2}$$

and the empirical estimator of the conditional transition distribution functions is defined by

$$F_{ij}^W(x,t) := \frac{1}{N_{ij}(t)} \sum_{k=1}^{N(t)} W_{N(t)k} \mathbb{1}_{\{J_{k-1}=i, J_k=j, X_k \le x\}}, \ 0 \le x \le t, \ i,j \in E, \tag{0.3}$$

where W_{ni}'s are the bootstrap weights defined on the probability space $(\mathcal{W}, \Omega, \mathbb{P}_W)$. The bootstrap weights W_{ni}'s are assumed to belong to the class of exchangeable bootstrap weights introduced in Præstgaard and Wellner (1993).

Theorem 0.1. *For any arbitrary but fixed $i,j \in E$, and fixed $x \in \mathbb{R}_+$, we have the following weak convergence, as $n \to \infty$, for $t > 0$,*

$$n^{1/2}(Q_{ij}^W(x,nt) - \widehat{Q}_{ij}(x,nt)) \rightsquigarrow (1+c^2)^{1/2} b_{ij}(x) W(t)/t, \tag{0.4}$$

provided that $b_{ij}^2(x) > 0$, where $b_{ij}^2(x) = \mu_{ii} Q_{ij}(x)(1 - Q_{ij}(x))$ and c depending on W_{in}.

Theorem 0.2. *For any arbitrary but fixed $i,j \in E$, and fixed $x \in \mathbb{R}_+$, we have the following weak convergence, as $n \to \infty$, for $t > 0$,*

$$n^{1/2}(F_{ij}^W(x,nt) - \widehat{F}_{ij}(x,nt)) \rightsquigarrow (1+c^2)^{1/2} c_{ij}(x) W(t)/t, \tag{0.5}$$

provided that $c_{ij}^2(x) > 0$, where $c_{ij}^2(x) = \frac{\mu_{ii}}{P(i,j)} F_{ij}(x)(1 - F_{ij}(x))$.

References

Andersen, P. K., Borgan, Ø., Gill, R. D., and Keiding, N. (1993). *Statistical models based on counting processes*. Springer Series in Statistics. Springer-Verlag, New York.

Barbu, V. S. and Limnios, N. (2008). *Semi-Markov chains and hidden semi-Markov models toward applications*, volume 191 of *Lecture Notes in Statistics*. Springer, New York. Their use in reliability and DNA analysis.

Janssen, J. and Limnios, N., editors (1999). *Semi-Markov models and applications*. Kluwer Academic Publishers, Dordrecht. Selected papers from the 2nd International Symposium on Semi-Markov Models: Theory and Applications held in Compiègne, December 1998.

Limnios, N. (2004). A functional central limit theorem for the empirical estimator of a semi-Markov kernel. *J. Nonparametr. Stat.*, **16**(1-2), 13–18. The International Conference on Recent Trends and Directions in Nonparametric Statistics.

Limnios, N. and Oprişan, G. (2001). *Semi-Markov processes and reliability*. Statistics for Industry and Technology. Birkhäuser Boston Inc., Boston, MA.

Præstgaard, J. and Wellner, J. A. (1993). Exchangeably weighted bootstraps of the general empirical process. *Ann. Probab.*, **21**(4), 2053–2086.

Pyke, R. (1961a). Markov renewal processes: definitions and preliminary properties. *Ann. Math. Statist.*, **32**, 1231–1242.

Pyke, R. (1961b). Markov renewal processes with finitely many states. *Ann. Math. Statist.*, **32**, 1243–1259.

Laboratoire de mathématiques appliquées-L.M.A.C., université de technologie de Compiègne, B.P. 529, 60205 Compiègne cedex, France, E-mail adresses : salim.bouzebda@utc.fr ; nikolaos.limnios@utc.fr

Estimation of survival probabilities from two-phases stratified samples

Norman Breslow, Thomas Lumley, and Jon A. Wellner
DEPARTMENT OF BIOSTATISTICS, UNIVERSITY OF WASHINGTON, SEATTLE
DEPARTMENT OF STATISTICS, UNIVERSITY OF AUCKLAND, NZ
DEPARTMENTS OF STATISTICS AND BIOSTATISTICS, UNIVERSITY OF WASHINGTON, SEATTLE
norm@uw.edu

Abstract

Epidemiologists employ stratified sampling from a defined cohort so that collection of costly covariate information, such as bioassays of stored tissue samples, may be limited to the most informative participants. From a survey perspective these designs involve two-phase stratified samples: a simple random sample (the main cohort) from an infinite super-population (model) at Phase I; and a finite population stratified (case-control) sample at Phase II. One approach to analysis involves inverse probability weighting (IPW) of general estimating equations.

In previous work we investigated IPW of infinite dimensional likelihood equations for both Euclidean and non-Euclidean parameters in semi-parametric models, of which the paradigm is the Cox model for survival data. The key idea was to separate the likelihood calculations, which are the same as those for simple random sampling, from weak convergence results for the IPW empirical process. For estimation of the Euclidean parameter (log hazard ratios), the problem was asymptotically equivalent to that of using the Phase II sample to estimate an unknown finite population total: the total of the unknown influence function contributions for subjects in the main cohort. Efficiency was improved, sometimes dramatically, through adjustment of the sampling weights by calibration to totals of auxiliary variables known for everyone or by estimation of the known weights using these same variables.

After briefly reviewing these results, this talk considers the extensions needed for joint estimation of hazard ratios and baseline hazard function in the Cox model, and hence for estimation of survival probabilities. The improvements in precision possible with calibrated or estimated weights are illustrated via simulations conducted using Lumley's R survey package to analyze data from the National Wilms Tumor Study.

References:

- Breslow NE, Wellner JA. Scand J Stat 34:86-102, 2007; 35:186-192, 2008.
- Breslow NE, Lumley T et al. Am J Epidemiol 35:1398-1405, 2009
- Breslow NE, Lumley T et al. Stat Biosci 1:32-49, 2009
- Lumley T. Complex Surveys, New York: Wiley, 2010

Conditional inference in parametric models

Michel Broniatowski
Email: Michel.broniatowski@upmc.fr

Abstract: This talk presents a new approach to conditional inference, based on the simulation of samples conditioned by a statistics of the data. Also an explicit expression for the approximation of the conditional likelihood of long runs of the sample given the observed statistics is provided. It is shown that when the conditioning statistics is sufficient for a given parameter, the approximating density is still invariant with respect to the parameter. A new Rao-Blackwellisation procedure is proposed and simulation shows that Lehmann Scheffé Theorem is valid for this approximation. Conditional inference for exponential families with nuisance parameter is also studied, leading to Monte carlo tests. Finally the estimation of the parameter of interest through conditional likelihood is considered. Comparison with the parametric bootstrap method is discussed.

Keywords: Conditional inference, Rao Blackwell Theorem, Lehmann Scheffé Theorem, Exponential families, Nuisance parameter, Simulation.

Nonparametric estimation for survival data with censoring indicators missing at random

Elodie Brunel
I3M, UMR 5149 CNRS,
Université Montpellier 2,
FRANCE,
ebrunel@math.univ-montp2.fr

Fabienne Comte
MAP5, UMR 8145 CNRS,
Université Paris Descartes,
Sorbonne Paris Cité
FRANCE,
fabienne.comte@parisdescartes.fr

Agathe Guilloux
LSTA, Université Paris 6,
Centre de Recherche Saint-Antoine (UMR S 938)
FRANCE,
agathe.guilloux@upmc.fr

***Abstract*—In this paper, we consider the problem of hazard rate estimation in presence of covariates, for survival data with censoring indicators missing at random. We propose in the context usually denoted by MAR (missing at random, in opposition to MCAR, missing completely at random, which requires an additional independence assumption), nonparametric adaptive strategies based on model selection methods for estimators admitting finite dimensional developments in functional orthonormal bases. Theoretical risks bounds are provided, they prove that the estimators behave well in term of Mean Square Integrated Error (MISE). Simulation experiments illustrate the statistical procedure.**

I. Introduction

We consider the problem of estimation from right-censored data in presence of covariates, when the censoring indicator is missing. Let T be a random variable representing the time to death from the cause of interest. Let C denote a right-censoring random time. Under usual random censorship, the observation is $Y = T \wedge C$ and $\delta = \mathbb{1}(T \leq C)$. Let X denote a real covariate. In what follows, it is assumed that T, C and X admit densities respectively denoted by f_T, g and f_X. In addition, C is assumed to be independent of T conditionally to X, see e.g. [Comte *et al.*(2011)] for comments on this assumption.

When the cause of death is not recorded, the censoring indicator is missing: this is the missing censoring indicator (MCI) model, see [Subramanian(2006)], which is defined as follows. Let ξ be the missingness indicator, that is $\xi = 1$ if δ is observed and $\xi = 0$ otherwise. The observed data are then given for individual $i \in \{1, \ldots, n\}$:

$$(Y_i, X_i, \delta_i, \xi_i = 1) \quad \text{or} \quad (Y_i, X_i, \xi_i = 0).$$

We shall say that the model is:

- MCAR under the assumption that the indicator are Missing Completely At Random, i.e. ξ is independent of T, C and X.
- MAR under the assumption that the indicator is Missing At Random i.e. ξ and δ are independent conditionally to Y, X.

In this paper, we mainly concentrate on the MAR model. The MCAR model will be consider in Section II-B.

This model has been considered by several authors in the last decade. Most papers are interested in survival function and cumulative hazard rate estimation. In particular, [van der Laan and McKeague(1998)] build a sieved nonparametric maximum likelihood estimator of the survival function in the MAR case and prove its efficiency. Their estimator is a generalization of the Kaplan-Meier estimator to this context and is the first proposal reaching the efficiency bound. [Subramanian(2004)] also proposes an efficient estimator of the survival function in the MAR case; he proves his estimate to be efficient as well.

Kernel methods have also been used to build different estimators in the MAR context. [Subramanian(2006)] estimates the cumulative hazard rate with a ratio of kernel estimators. He provides an almost sure representation, and a Central Limit Theorem (CLT). He deduces results of the same type for the survival function. A study in a similar context is also provided by [Wang and Ng(2008)]. Recently, [Wang *et al.*(2009)] proposed density estimator based on kernels and Kaplan Meier-type corrections of censoring. They prove a CLT and suggest a bandwidth selection strategy. Extensions of these works to conditional functions (both cumulative hazard and survival functions) in the presence of covariates is developed in [Wang and Shen(2008)].

Both our method and our aim are rather different. Indeed, we estimate the conditional hazard rate given a covariate. Moreover, we provide a nonparametric mean square strategy by considering approximations of the target function on finite dimensional linear spaces spanned by convenient and simple orthonormal (functional) bases. A collection of estimators is thus defined, indexed by the dimension of the multidimensional projection space, and a penalization device allows us to select a "good" space among all the proposals.

Our estimator has the advantage of being defined as a contrast minimizer and not a ratio of two estimators, as in standard kernel methodology. As a drawback, it depends on an unknown function, in its definition, which has to be replaced by an estimator, and its mean square risk has consequently the order of the anisotropic rate corresponding to the regularity of the function under estimation, plus the rate of the intermediate plug-in estimator, for which we propose a similar estimation

strategy.

The plan of the paper is the following. We first explain in Section II how the contrast is built, and how it allows us to compute a collection of estimators. We conclude the section by giving the penalization device that completes the definition of the data driven estimator, up to an estimator to be plugged in the procedure. In Section III, we state the theoretical results that ensure that the quadratic risk of our estimator behaves well provided that the intermediate estimator has small risk. Then, we show how similar tools can be used to build, compute and control the second estimator. The procedure is tested in a simulation for both hazard and conditional hazard rates (i.e. with or without covariate) and under different missing scheme: we give here only a summary of our conclusions. Both the simulation section and technical proofs can be found in the working paper MAP5 2012-08.

II. Definition of the conditional hazard rate estimator

A. Choice of the contrast

We consider the general MAR case as described in the introduction, the global assumption is denoted **(A0)** and has several parts that we specify below.

(A0-1) The random vectors (X_i, T_i, C_i) are independent copies, for $i = 1, \dots, n$, of (X, Y, C).

(A0-2) For $i = 1, \dots, n$, we observe X_i, $Y_i = T_i \wedge C_i$, ξ_i, and $\delta_i = \mathbb{1}(T_i \leq C_i)$ if $\xi_i = 1$, otherwise $\xi_i = 0$.

(A0-3) C is independent of T given X.

(A0-4) ξ and δ are independent given X, Y.

The unknown function λ to be estimated is the conditional hazard rate of the random variable T given $X = x$ defined, for all $z > 0$ by:

$$\lambda(x,t) = \lambda_{T|X}(x,t) = \frac{f_{T|X}(x,t)}{1 - F_{T|X}(x,t)},$$

where $f_{T|X}$ and $F_{T|X}$ are respectively the conditional probability density function (p.d.f.) and the conditional cumulative distribution function (c.d.f.) of T given X. We shall denote by $G_{C|X}$ the conditional c.d.f. of C given X. We define the conditional expectations of ξ and δ by:

$$\begin{aligned} \pi(x,y) &= \mathbb{E}(\xi | X = x, Y = y) \text{ and} \\ \zeta(x,y) &= \mathbb{E}(\delta | X = x, Y = y). \end{aligned}$$

The crucial property for the construction of an estimation procedure is the following: for any integrable function h, we have

$$\begin{aligned} &\mathbb{E}(\zeta(X,Y)h(X,Y)) = \mathbb{E}[\mathbb{E}(\delta|X,Y)h(X,Y)] \\ &= \mathbb{E}(\delta h(X,Y)) \\ = \;& \mathbb{E}[\mathbb{E}(\mathbb{1}(T \leq C)h(X,T)|X)] \\ = \;& \mathbb{E}[(1 - G_{C|X})(X,T)h(X,T)] \quad \text{with } \textbf{(A0-3)} \\ = \;& \iint h(x,t)(1 - G_{C|X})(x,t) f_{T|X}(x,t) f_X(x) dx dt. \end{aligned}$$

This yields the equality

$$\begin{aligned} \mathbb{E}(\zeta(X,Y)h(X,Y)) &= \mathbb{E}(\delta h(X,Y)) \\ &= \iint h(x,y)\lambda(x,y) d\mu(x,y) \quad (1) \end{aligned}$$

with

$$d\mu(x,y) = (1 - L_{Y|X}(y,x)) f_X(x) dx dy = f(x,y) dx dy,$$

where $f(x,y) = (1 - L_{Y|X}(y,x)) f_X(x)$, and

$$\begin{aligned} 1 - L_{Y|X}(y,x) &:= \mathbb{P}(Y \geq y | X = x) \\ &= (1 - F_{T|X}(x,y))(1 - G_{C|X}(x,y)). \end{aligned}$$

If ζ was known, we would consider the contrast:

$$\begin{aligned} \Gamma_n^{th}(h) &= \frac{1}{n} \sum_{i=1}^n \int_0^1 h^2(X_i, y) \mathbb{1}(Y_i \geq y) dy \\ &\quad - \frac{2}{n} \sum_{i=1}^n \left(\xi_i \delta_i + (1 - \xi_i)\zeta(X_i, Y_i)\right) h(X_i, Y_i), \end{aligned}$$

which is a natural extension to the MAR case of the contrast introduced in [Comte *et al.*(2011)]. We note that, with assumption **(A0-4)** and the definition of ζ, we have

$$\begin{aligned} &\mathbb{E}(\delta_i \xi_i + (1 - \xi_i)\zeta(X_i, Y_i) | X_i, Y_i) \\ = \;& \mathbb{E}(\delta_i | X_i, Y_i)\mathbb{E}(\xi_i | X_i, Y_i) + \mathbb{E}[(1 - \xi_i)\mathbb{E}(\delta_i | X_i, Y_i) | X_i, Y_i] \\ = \;& \mathbb{E}(\mathbb{E}(\delta_i | X_i, Y_i)(\xi_i + (1 - \xi_i)) | X_i, Y_i), \end{aligned}$$

that is

$$\mathbb{E}(\delta_i \xi_i + (1 - \xi_i)\zeta(X_i, Y_i) | X_i, Y_i) = \mathbb{E}(\delta_i | X_i, Y_i). \quad (2)$$

Thus, if we compute the expectation of this theoretical contrast, we obtain, under the MAR assumption and using (1) and (2),

$$\begin{aligned} \mathbb{E}\big(\Gamma_n^{th}(h)\big) &= \|h\|_\mu^2 - 2 \iint h(x,y)\lambda(x,y) d\mu(x,y) \\ &= \|h - \lambda\|_\mu^2 - \|\lambda\|_\mu^2. \end{aligned}$$

Clearly, the above quantity is small if h is near of λ, and the measure denoted by μ plays the role of a reference weighting norm. This explains why minimizing Γ_n^{th} over an appropriate set of functions would be a relevant strategy to estimate λ.

As ζ is unknown, we must replace it by an estimator $\tilde{\zeta}$. Consequently, we consider

$$\begin{aligned} \Gamma_n(h) &= \frac{1}{n} \sum_{i=1}^n \int_0^1 h^2(X_i, y) \mathbb{1}(Y_i \geq y) dy \\ &- \frac{2}{n} \sum_{i=1}^n \left(\xi_i \delta_i + (1 - \xi_i)\tilde{\zeta}(X_i, Y_i)\right) h(X_i, Y_i). \quad (3) \end{aligned}$$

An estimator of $\zeta(x,y)$ is constructed in Section III-C below. This strategy of estimation of the unknown hazard rate λ, via an estimation of ζ, is also considered in [Wang *et al.*(2009)].

The empirical reference norm associated with the contrast (3) is defined by

$$\|h\|_n^2 = \frac{1}{n}\sum_{i=1}^{n}\int_0^1 h^2(X_i,y)\mathbb{1}(Y_i \geq y)dy$$

and the natural resulting scalar product is denoted by $\langle h, h_2\rangle_n = (1/4)(\|h+h_2\|_n^2 - \|h-h_2\|_n^2)$, where

$$\mathbb{E}(\langle h, h_2\rangle_n) = \langle h, h_2\rangle_\mu.$$

Remark 1. We could consider another strategy for the construction of the contrast, namely

$$\Gamma_n(h) = \frac{1}{n}\sum_{i=1}^{n}\int_0^1 h^2(X_i,y)\tilde{\pi}(X_i,y)\mathbb{1}(Y_i \geq y)dy - \frac{2}{n}\sum_{i=1}^{n}\delta_i\xi_i h(X_i,Y_i). \quad (4)$$

where $\tilde{\pi}$ is an estimator of π. The second part in Equation (4) is weighted by $\xi_i\delta_i$ which means that fewer observations are used for the estimation. As a consequence, the contrast (3), that we consider, is not only more convenient (from algebraic point of view) but is also expected to be more relevant.

B. The MCAR case

In the MCAR case, the function π is constant, that is $\pi(x,y) = p = \mathbb{E}(\xi)$. The following contrast

$$\gamma_n^{(1)}(h) = \frac{1}{n}\sum_{i=1}^{n}\int_0^1 h^2(X_i,y)\xi_i\mathbb{1}(Y_i \geq y)dy - \frac{2}{n}\sum_{i=1}^{n}\delta_i\xi_i h(X_i,Y_i), \quad (5)$$

would be adequate for conditional hazard rate estimation with reference measure

$$d\mu(x,y) = p(1-L_{Y|X}(y,x))f_X(x)dxdy.$$

It has the advantage of not involving any estimator and the drawback that there is a ξ_i-factor in all terms, so that only observations with non missing indicator are taken into account.

Remark 2. The contrast $\gamma_n^{(1)}$, defined in (5), can be seen as a rewriting of the contrast (4) taking into account the MCAR assumption. Indeed, in that case, $\tilde{\pi}(X_i,y)$ can be replaced by ξ_i. Notice in addition that, in the simulation study, see Section IV, we used this contrast for the MCAR case because we experimented that it was giving much stabler and better results than the contrast $\gamma_n^{(0)}$.

Lastly, this contrast would be still valid for estimating λ for ξ independent of Y given X, with reference measure:

$$d\mu_2(x,y) = \pi(x)(1-L_{Y|X}(y,x))f_X(x)dxdy,$$

and

$$\pi(x) = \mathbb{E}(\xi|X=x) = \mathbb{E}(\xi|X=x, Y=y).$$

C. Computing the estimator

We consider that we estimate the hazard rate on a compact set

$$A = A_1 \times [0,\tau],$$

where A_1 is an interval such that all observations lie in the domain. We take $A_1 = [0,1]$ for simplicity and without loss of generality. Recall that $f(x,y) = (1-L_{Y|X}(y,x))f_X(x)$ and denote by $f_{(X,Y)}(x,y)$ the joint density of the random pair (X,Y). We set standard assumptions of boundednesses from above and below.

- **(A1)** $\forall (x,y)\in A,\ \ 0 < f_0 \leq f(x,y) \leq f$ for fixed positive constants f_0 and f.
- **(A2)** $\forall (x,y)\in A,\ \ \lambda(x,y) \leq \|\lambda\|_{A,\infty} < +\infty$.
- **(A3)** $\forall (x,y)\in A,\ \ 0 < f_0^* < f_{(X,Y)}(x,y) \leq f^* + \infty$.

First, we define an estimator $\hat{\lambda}_m$ on the space S_m by:

$$\hat{\lambda}_m = \arg\min_{h\in S_m}\Gamma_n(h)$$

where $S_m = S_{m_1}^{(1)} \otimes S_{m_2}^{(2)}$, with

$$S_{m_1}^{(1)} = \text{spanned}\{\varphi_j, j = 1,\dots,D_{m_1}^{(1)}\}$$

and

$$S_{m_2}^{(2)} = \text{spanned}\{\psi_k, k = 1,\dots,D_{m_2}^{(2)}\}$$

The φ_j's, as well as the ψ_k's, constitute an $\mathbb{L}^2$-orthonormal basis, and the function h is of the form $h = \sum_{j,k} a_{j,k}\varphi_j \otimes \psi_k$.

We consider in the following two specific and classical examples of bases:

1) **Trigonometric bases.** They are defined by $\varphi_0(x) = \mathbb{1}_{[0,1]}(x)$, $\varphi_{2j+1}(x) = \sqrt{2}\sin(2\pi jx)\mathbb{1}_{[0,1]}(x), \varphi_{2j}(x) = \sqrt{2}\cos(2\pi jx)\mathbb{1}_{[0,1]}(x)$ and $\psi_0(x) = (1/\sqrt{\tau})\mathbb{1}_{[0,\tau]}(x)$, $\psi_{2k+1}(x) = \sqrt{2/\tau}\sin(2\pi jx/\tau)\mathbb{1}_{[0,\tau]}(x), \psi_{2k}(x) = \sqrt{2/\tau}\cos(2\pi kx/\tau)\mathbb{1}_{[0,\tau]}(x)$. Considering $(\varphi_j)_{0\leq j\leq m_1-1}$ and $(\psi_k)_{0\leq k\leq m_2-1}$ yields spaces with odd dimensions m_1 and m_2. We denote by $\mathcal{S}_n$ the nesting space of the collection, i.e. the product space corresponding to maximal dimensions for $S_{m_1}^{(1)}$ and $S_{m_2}^{(2)}$.
2) **Histogram bases.** They are defined by $\varphi_j(x) = \sqrt{2^{m_1}}\mathbb{1}_{[(j-1)/2^{m_1}, j/2^{m_1}[}(x)$, for $j = 1,\dots,2^{m_1}$ and $\psi_k(x) = \sqrt{2^{m_2}/\tau}\mathbb{1}_{[(k-1)\tau/2^{m_2}, k\tau/2^{m_2}[}(x)$, for $k = 1,\dots,2^{m_2}$, so that $D_{m_1}^{(1)} = 2^{m_1}, D_{m_2}^{(2)} = 2^{m_2}$. We shall take $m \leq [\log_2(n)/2]$ and $m_2 \leq [\log_2(n)/2]$ where $[z]$ denotes the integer part of z and $\log_2(x) = \log(x)/\log(2)$. We denote by $\mathcal{S}_n$ the nesting space of the collection, that is $\mathcal{S}_n = S_{m_1(n)}^{(1)} \otimes S_{m_2(n)}^{(2)}$, where $2^{m_1(n)}2^{m_2(n)} \leq n$.

In both cases, we denote by $\mathcal{D}_n := \dim(\mathcal{S}_n) = \mathcal{D}_n^{(1)}\mathcal{D}_n^{(2)}$. These bases are representative examples of localized bases for the second one (as piecewise polynomials, wavelets) or bounded non localized bases for the first one.

Now, let us study the contrast minimization. Writing that $\partial\Gamma_n(h)/\partial a_{j_0,k_0} = 0$, we get that the coefficients $\hat{a}_{j,k}$ of the

estimate of $\hat{\lambda}_m$ verify

$$\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \varphi_{j_0}(X_i)\psi_{k_0}(y)(\sum_{j,k}\hat{a}_{j,k}\varphi_j(X_i)\psi_k(y))\mathbf{I}(Y_i\geq y)dy$$
$$=\frac{1}{n}\sum_{i=1}^{n}\left(\delta_i\xi_i+(1-\xi_i)\tilde{\zeta}(X_i,Y_i)\right)\varphi_{j_0}(X_i)\psi_{k_0}(Y_i).$$

In the histogram case, as $\varphi_j\varphi_{j'}\equiv 0$ for $j\neq j'$ and $\psi_k\psi_{k'}\equiv 0$ for $k\neq k'$, we get

$$\hat{a}_{j_0,k_0}=\frac{\sum_{i=1}^{n}\left(\delta_i\xi_i+(1-\xi_i)\tilde{\zeta}(X_i,Y_i)\right)\varphi_{j_0}(X_i)\psi_{k_0}(Y_i)}{\sum_{i=1}^{n}\int_0^\tau\varphi_{j_0}^2(X_i)\psi_{k_0}^2(y)\mathbf{I}(Y_i\geq y)dy}$$

if the denominator is non zero.

More generally, let us define the matrices

$$G_{m_1}^{\varphi}(X_i)=\left(\varphi_j(X_i)\varphi_{j'}(X_i)\right)_{1\leq j,j'\leq D_{m_1}^{(1)}},$$

and

$$H_{m_2}^{\psi}(y)=\left(\psi_k(y)\psi_{k'}(y)\right)_{1\leq k,k'\leq D_{m_2}^{(2)}},$$

so that their tensorial Kronecker product $G_{m_1}^{\varphi}(X_i)\otimes H_{m_2}^{\psi}(y)$ is of size $(D_{m_1}^{(1)}+D_{m_2}^{(2)})\times(D_{m_1}^{(1)}+D_{m_2}^{(2)})$. We set

$$\Theta_m:=\frac{1}{n}\sum_{i=1}^{n}\int G_{m_1}^{\varphi}(X_i)\otimes H_{m_2}^{\psi}(y)\mathbf{I}_{\{Y_i\geq y\}}dy.$$

Recall that the $\mathrm{ve}(.)$ operator stacks the columns of a matrix and let

$$\overrightarrow{\hat{a}}_m=\mathrm{ve}\left({}^t(\hat{a}_{j,k})_{1\leq j\leq D_{m_1}^{(1)},1\leq k\leq D_{m_2}^{(2)}}\right),$$
$$\Delta_m=\mathrm{ve}\left(\frac{1}{n}\sum_{i=1}^{n}(\delta_i\xi_i+(1-\xi_i)\tilde{\zeta}(X_i,Y_i)\varphi_j(X_i)\psi_k(Y_i))_{j,k}\right),$$

where $1\leq j\leq D_{m_1}^{(1)}, 1\leq k\leq D_{m_2}^{(2)}$, then the coefficients of the estimator must fulfill the matrix constraint:

$$\Theta_m\overrightarrow{\hat{a}}_m=\Delta_m.$$

It follows that the estimator is well defined if Θ_m is invertible. We define $\rho(M)$ as the spectral radius of a matrix M, i.e. the largest eigenvalue in modulus of M. We set

$$\overrightarrow{\hat{a}}_m=\Theta_m^{-1}\Delta_m \text{ if } \rho(\Theta_m)\geq\max(\hat{f}_0/3,n^{-1/2}) \quad (6)$$

and $\overrightarrow{\hat{a}}_m=0$ otherwise.

The quantity $\hat{f}_0$ is an estimator of $f_0=\min_{(x,y)\in A}f(x,y)$, where $f(x,y)=(1-L_{Y|X}(y,x))f_X(x)$. It is defined in [Comte *et al.*(2011)], and proved to satisfy, for n large enough:

(A4) For any integer $k\geq 1$, there exists a constant $C_k^{(f_0)}>0$ such that

$$\mathbb{P}(|\hat{f}_0-f_0|>f_0/2)\leq C_k^{(f_0)}/n^k$$

At this stage, we are in a position of defining an estimator of λ on S_m:

$$\hat{\lambda}_m(x,y)=\sum_{j,k}\hat{a}_{j,k}\varphi_j(x)\psi_k(y), \quad (7)$$

where the $\hat{a}_{j,k}$'s are defined in Equation (6).

D. Model selection by penalization

The model selection device is now based on the following criterion

$$\hat{m}=\arg\min_{m\in\mathcal{M}_n}(\Gamma_n(\hat{\lambda}_m)+\mathrm{pen}(m)) \quad (8)$$

where

$$\mathcal{M}_n = \{m=(m_1,m_2)\in\mathbb{N}\times\mathbb{N},\dim(S_{m_2}^{(2)})\geq\log(n),$$
$$\dim(S_{m_1}^{(1)}\otimes S_{m_2}^{(2)})\leq\mathcal{D}_n\},$$

and the penalty is defined by

$$\mathrm{pen}(m)=\kappa\|\hat{\lambda}\|_{\infty,A}\frac{\dim(S_m)}{n}, \quad (9)$$

where $\hat{\lambda}=\hat{\lambda}_{m_0}$ is an estimator in the collection, on a space S_{m_0} which is specified below. Note that the properties required on $(m_1,m_2)\in\mathcal{M}_n$ mean that all spaces of the collection are included in a nesting space with dimension $\mathcal{D}_n$. Moreover, the dimension $D_{m_2}^{(2)}$ of the space $S_{m_2}^{(2)}$ has to be larger than $\log(n)$, see the proof of Proposition **??**. Lastly, we define the theoretical counterpart of the penalty:

$$\mathrm{pen}^{th}(m)=\kappa\|\lambda\|_{\infty,A}\frac{\dim(S_m)}{n}.$$

III. Results

A. Main Theorem

In order to state our Theorem 1, we have to define the integral norm with respect to $d\varrho(x,y)=f_{(X,Y)}(x,y)dxdy$ where $f_{(X,Y)}$ is the density of the bivariate vector (X,Y), that is

$$\|\psi\|_\varrho^2=\iint\psi^2(x,y)d\varrho(x,y)=\iint\psi^2(x,y)f_{(X,Y)}(x,y)dxdy \quad (10)$$

and the associated empirical norm:

$$\|\psi\|_{\varrho,n}^2=\frac{1}{n}\sum_{i=1}^{n}\psi^2(X_i,Y_i) \quad (11)$$

Theorem 1: Let $\hat{\lambda}_{\hat{m}}$ be the estimator defined by (6)-(7)-(8)-(9). Under Assumptions **(A1)-(A4)**, and if $\mathcal{D}_n^2\leq n/\log^2(n)$ for basis (1) and $\mathcal{D}_n\leq n/\log^2(n)$ for basis (2), there exists a constant κ such that, for n large enough

$$\mathbb{E}(\|\lambda\mathbf{I}_A-\hat{\lambda}_{\hat{m}}\|_n^2) \leq C\inf_{m\in\mathcal{M}_n}(\|\lambda\mathbf{I}_A-\lambda_m\|_\mu^2+\mathrm{pen}^{th}(m))$$
$$+C'\mathbb{E}(\|\tilde{\zeta}-\zeta\|_\varrho^2)+\frac{C''}{n}, \quad (12)$$

where C is a numerical constant and C', C'' are constants depending on f_0, f, f_0^*, f^* and $\|\lambda\|_{\infty,A}$.

The result stated in (12) involves a first term: $\inf_{m\in\mathcal{M}_n}(\|\lambda\mathbf{I}_A-\lambda_m\|^2+\mathrm{pen}^{th}(m))$ which is the usual squared-bias ($\|\lambda\mathbf{I}_A-\lambda_m\|^2$)/variance ($\mathrm{pen}^{th}(m)$) compromise, and will lead to an optimal anisotropic rate for a given regularity $\alpha=(\alpha,\alpha_2)$ of λ. The second term in (12) is $\mathbb{E}(\|\tilde{\zeta}-\zeta\|_\varrho^2)$, that is the mean-square risk of the estimator of ζ on A. The last term C''/n is negligible.

B. Consequence on the rate

The next corollary shows that $\hat{\lambda}_{\hat{m}}$ adapts to the unknown anisotropic smoothness of λ, up to the performance of $\tilde{\zeta}$. Toward that end, assume that λ restricted to A belongs to the anisotropic Besov space $B^{\boldsymbol{\alpha}}_{2,\infty}(A)$ on A with regularity $\boldsymbol{\alpha} = (\alpha, \alpha_2)$. We mention that anisotropy is almost mandatory in this context, because the regularity in the covariate direction has no reason to be the same as the regularity in the y-direction.

Let us recall the definition of $B^{\boldsymbol{\alpha}}_{2,\infty}(A)$. Let $\{e, e_2\}$ the canonical basis of $\mathbb{R}^2$ and take $A^r_{h,i} := \{x \in \mathbb{R}^2; x, x + he_i, \ldots, x + rhe_i \in A\}$, for $i = 1, 2$. For $x \in A^r_{h,i}$, let

$$\Delta^r_{h,i} g(x) = \sum_{k=0}^{r} (-1)^{r-k} \binom{r}{k} g(x + khe_i)$$

be the rth difference operator with step h. For $t > 0$, the directional moduli of smoothness are given by

$$\omega_{r_i,i}(g, t) = \sup_{|h| \le t} \Big(\int_{A^{r_i}_{h,i}} |\Delta^{r_i}_{h,i} g(x)|^2 dx \Big)^{1/2}.$$

We say that g is in the Besov space $B^{\boldsymbol{\alpha}}_{2,\infty}(A)$ if $\sup_{t>0}(t^{-\alpha}\omega_{r,1}(g, t) + t^{-\alpha_2}\omega_{r_2,2}(g, t) < \infty$ for r, r_2, integers larger than α, α_2 respectively. More details concerning Besov spaces can be found in [Triebel(2006)].

Corollary 1: Assume that λ restricted to A belongs to the anisotropic Besov space $B^{\boldsymbol{\alpha}}_{2,\infty}(A)$ with regularity $\boldsymbol{\alpha} = (\alpha, \alpha_2)$ such that $\alpha > 1/2$ and $\alpha_2 > 1/2$. Consider the estimator in the histogram basis. Then, under the assumptions of Theorem 1, we have

$$\mathbb{E}\left(\|\lambda - \hat{\lambda}_{\hat{m}}\|^2_A \right) = O(n^{-\frac{2\bar{\alpha}}{2\bar{\alpha}+2}}) + \mathbb{E}(\|\tilde{\zeta} - \zeta \mathbb{1}_A\|^2_\varrho). \quad (13)$$

where $\bar{\alpha}$ is the harmonic mean of α and α_2 (i.e. $2/\bar{\alpha} = 1/\alpha + 1/\alpha_2$).

The proof follows the lines of Corollary 1 (p.1178) in Comte et al. (2011). At this point, to state our final result for the estimation of λ (stated in Corollary 2), we have to construct and study an estimator of ζ.

C. Estimation of $\zeta(x, y)$

Here we want to exhibit an estimator of ζ on A for which we can prove a bound for $\mathbb{E}(\|\tilde{\zeta} - \zeta\|^2_\varrho)$. We consider the mean-square regression estimator of ζ defined as the minimizer of

$$\tilde{\gamma}_n(T) = \frac{1}{n} \sum_{i=1}^{n} [\xi_i T^2(X_i, Y_i) - 2\xi_i \delta_i T(X_i, Y_i)],$$

for T in $S_m = S^{(1)}_{m_1} \otimes S^{(2)}_{m_2}$, with penalization

$$\widetilde{\text{pen}}(m) = \tilde{\kappa} \frac{\dim(S_m)}{n}.$$

Here the reference norm must be $\|.\|_\varrho$ defined by (10) but the empirical norm associated with the problem is

$$N^2_{\xi,n}(\psi) = \frac{1}{n} \sum_{i=1}^{n} \xi_i \psi^2(X_i, Y_i),$$

$$\mathbb{E}(N^2_{\xi,n}(\psi)) = \iint \psi^2(x, y)\pi(x, y) f_{(X,Y)}(x, y) dxdy := \|\psi\|^2_\xi.$$

We assume that there exists a constant π_0, such that:

(B1) $\forall (x, y) \in A, \ \ 0 < \pi_0 \le \pi(x, y) \le 1.$

If one is interested in a control of $\mathbb{E}(N^2_{\xi,n}(\hat{\zeta}_{\hat{m}} - \zeta))$, one may consider that only the vector $(\hat{\zeta}_m(X_i, Y_i))$ has to be correctly defined, and in this case, classical projection arguments can be used to prove that the definition is consistent without any additional tools.

But considering that our aim here is related to the estimation of conditional hazard rate of the previous section, we wish to provide a $\mathbb{L}^2$ control.

Let us consider the same bases as in Section II-C, and the matrices

$$G^{\psi}_{m_2}(Y_i) = (\psi_k(Y_i)\psi_{k'}(Y_i))_{1 \le k,k' \le D^{(2)}_{m_2}}.$$

Let us define

$$\Upsilon_m = \frac{1}{n} \sum_{i=1}^{n} \xi_i G^{\varphi}_{m_1}(X_i) \otimes G^{\psi}_{m_2}(Y_i).$$

If the estimate of ζ is denoted by $\hat{\zeta}_m(x, y) = \sum_{j,k} \hat{\zeta}_{j,k}\varphi_j(x)\psi_k(y)$ and $\hat{Z}_m = (\text{ve}(\,{}^t(\hat{\zeta}_{j,k})_{j,k}))$ and

$$\Xi_m = \text{ve}\left((\frac{1}{n} \sum_{i=1}^{n} \xi_i \delta_i \varphi_j(X_i)\psi_k(Y_i))_{1 \le j \le D^{(1)}_{m_1}, 1 \le k \le D^{(2)}_{m_2}} \right).$$

Then we get in the same way as previously that, if Υ_m is invertible, $\hat{Z}_m = \Upsilon^{-1}_m \Xi_m$ and we set more restrictively

$$\hat{Z}_m = \Upsilon^{-1}_m \Xi_m \text{ if } \rho(\Upsilon_m) \ge \max(\hat{\rho}_0/2, n^{-1/2}),$$

and $\hat{Z}_m = 0$ otherwise. Here $\hat{\rho}_0$ is an estimate of ρ_0, which can be defined as the minimum of a well-chosen estimator of $\pi f_{(X,Y)}$: for instance $\hat{\rho}_0 = \sqrt{\dim(S_{m^*})} \min_{j,k} |\hat{a}_{j,k}|$ where

$$\hat{a}_{j,k} = \frac{1}{n} \sum_{i=1}^{n} \xi_i \varphi_j(X_i)\psi_k(Y_i)$$

and S_{m^*} is associated to a large enough subdivision for histogram bases (φ_j) and (ψ_k). We consider the assumption

(B2) For any integer $k \ge 1$, there exists a constant $C^{(\rho_0)}_k > 0$ such that $\mathbb{P}(|\hat{\rho}_0 - \rho_0| > \rho_0/2) := \mathbb{P}(\Omega^c_{\rho_0}) \le C^{(\rho_0)}_k / n^k$.

Then we have the following result bounding the $\mathbb{L}^2_\varrho$-risk of the estimator.

Theorem 2: Under assumptions **(A1)**, **(B1)-(B2)**, and if $\mathcal{D}^2_n \le n/\log^2(n)$ for basis (1) and $\mathcal{D}_n \le n/\log^2(n)$ for basis (2), there exists a choice of $\tilde{\kappa}$ such that,

$$\mathbb{E}(\|\hat{\zeta}_{\hat{m}} - \zeta \mathbb{1}_A\|^2_\varrho) \le C \inf_{m \in M_n} \left(\|\zeta_m - \zeta \mathbb{1}_A\|^2_\varrho + \widetilde{\text{pen}}(m) \right) + \frac{C'}{n}.$$

The next corollary is an immediate consequence of Corollary 1 and Theorem 2.

Corollary 2: Under the assumptions of Corollary 1 and assuming that ζ restricted to A belongs to the anisotropic Besov space $B^{\boldsymbol{\beta}}_{2,\infty}(A)$ with regularity $\boldsymbol{\beta} = (\beta, \beta_2)$ such that $\beta > 1/2$ and $\beta_2 > 1/2$. We take the estimator in the

histogram basis. Then, under the assumptions of Corollary 1, the estimator $\hat{\lambda}_{\hat{m}}$ of λ verifies:

$$\mathbb{E}\left(\|\lambda - \hat{\lambda}_{\hat{m}}\|_A^2\right) = O(n^{-\frac{2\bar{\alpha}}{2\bar{\alpha}+2}}) + O(n^{-\frac{2\bar{\beta}}{2\bar{\beta}+2}}). \qquad (14)$$

where $\bar{\alpha}$ (resp. $\bar{\beta}$) is the harmonic mean of α and α_2 (resp. β and β_2).

IV. Conclusion

To evaluate the finite sample performances of our different proposals for hazard rate estimation, we provide in the paper Monte Carlo studies in different settings. We study the (possibly conditional) hazard rate estimators with or without covariate, and under both settings of dependence for the missing of censoring indicators. We can in this way check that, once the constant κ is calibrated, the method works well in the different contexts studied here.

In the setting without covariate, our results show that in the MCAR setting, the MAR estimator always behaves slightly better than the MCAR one. At first sight, this may seem surprising but this is related to Remark 2. Indeed, the MAR estimator is obtained via the contrast (3) which is based on imputation and, as a consequence, uses all the data. On the contrary, the MCAR estimator is obtained via the contrast (5) which uses only the non missing data. This is confirmed in the bivariate conditional case: again, the MAR estimator gives systematically better results than the MCAR one. In conclusion, we recommend the systematic use of the MAR estimator when censoring indicators are missing.

References

[Barron *et al.*(1999)] Barron, A.R., Birgé, L. and Massart, P. (1999). Risk bounds for model selection via penalization. *Probab. Theory Relat. Fields* **113**, 301-413.

[Baraud *et al.*(2001)] Baraud, Y., Comte, F. and Viennet, G. (2001). Adaptive estimation in an autoregression and a geometrical β-mixing regression framework. *The Annals of Statistics* **39**, 839-875.

[Birgé & Massart(1998)] Birgé, L. and Massart, P. (1998) Minimum contrast estimators on sieves: exponential bounds and rates of convergence. *Bernoulli* **4**, 329-375.

[Cohen *et al.* (1993)] Cohen, A., Daubechies, I. and Vial, P. (1993) Wavelets on the interval and fast wavelet transforms, Applied and *Computational Harmonic Analysis* **1**, 54–81.

[Comte *et al.*(2011)] Comte, F., Gaïffas, S. and Guilloux, A. (2011). Adaptive estimation of the conditional intensity of marker-dependent counting processes. To appear in *Ann. Inst. Henri Poincaré Probab. Stat.* **47**, 171-1196.

[Kaplan-Meier(1958)] Kaplan, E.L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *J. Amer. Statist. Assoc.*, **53**, 457-481.

[Lawless(2003)] Lawless, J.F. (2003). *Statistical models and methods for lifetime data.* Second edition. Wiley Series in Probability and Statistics. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ.

[Lo *et al.*(1989)] Lo, S.H., Mack, Y.P. and Wang, J.L. (1989). Density and hazard rate estimation for censored data via strong representation of the Kaplan-Meier estimator. *Probab. Theory Related Fields* **80**, 461-473.

[Nikol'skii(1975)] Nikol'skii, S. M. (1975). *Approximation of functions of several variables and imbedding theorems.* Springer-Verlag, New York. Translated from the Russian by John M. Danskin, Jr., Die Grundlehren der Mathematischen Wissenschaften, Band 205.

[Talagrand(1996)] Talagrand, M. (1996). New concentration inequalities in product spaces. *Invent. Math.* **126**, 505-563.

[van der Laan and McKeague(1998)] van der Laan, M. J. and McKeague, I. W. (1998) Efficient estimation from right-censored data when failure indicators are missing at random. *Ann. Statist.* **26**, 164-182.

[Subramanian(2006)] Subramanian, S. (2006) Survival analysis for the missing censoring indicator model using kernel density estimation techniques. *Stat. Methodol.* **3**, 125–136.

[Subramanian(2004)] Subramanian, S. (2004) Asymptotically efficient estimation of a survival function in the missing censoring indicator model. *J. Nonparametr. Stat.* **16**, 797-817.

[Triebel(2006)] Triebel, H. (2006). *Theory of function spaces. III.* Monographs in Mathematics, 100. Birkhäuser Verlag, Basel, 2006.

[Wang and Shen(2008)] Wang, Q. and Shen, J. (2008) Estimation and confidence bands of a conditional survival function with censoring indicators missing at random. *J. Multivariate Anal.* **99**, 928-948.

[Wang and Ng(2008)] Wang, Q. Ng, K. W. (2008) Asymptotically efficient product-limit estimators with censoring indicators missing at random. *Statist. Sinica* **18**, 749-768.

[Wang *et al.*(2009)] Wang, Q. , Liu, W. and Liu, C. (2009) Probability density estimation for survival data with censoring indicators missing at random. *J. Multivariate Anal.* **100**, 835-850.

[Zhou and Sun (2008)] Zhou, X. and Sun, L. (2003) Additive hazards regression with missing censoring information. *Statist. Sinica* **13**, 1237–1257.

Rank estimation methods for response biased sampling

Kani Chen
email : makchen@ust.hk

Abstract: Response-biased sampling, in which samples are drawn from a population according to the values of the response variable, is common in biomedical, epidemiological, economic and social studies. In case control studies, as a special but typical case of response biased sampling, prospective estimation approach for logistic regression model still works without any modification. We propose to use transformation models, known as the generalized accelerated failure time model in econometrics, for regression analysis of response-biased sampling with continuous response. We show that the prospective maximum rank correlation estimation is still valid for response-biased sampling. Unlike the inverse probability methods, the proposed method of estimation does not involve the sampling probabilities, which are often difficult to obtain in practice.

Comparative Analysis of Some Goodness-of-Fit Tests for Censored Data

E. Chimitova, A. Naumova, A. Tsivinskaya, M. Vedernikova
Department of Applied Mathematics
Novosibirsk State Technical University
Novosibirsk, Russia
ekaterina.chimitova@gmail.com

Abstract—**In this paper, we consider some goodness-of-fit tests for right censored samples: modified Kolmogorov, Cramer-von Mises-Smirnov, Anderson-Darling and χ^2 Nikulin-Rao-Robson tests. We also consider an approach based on the transformation of an original censored sample to a complete one and the consequent application of classical goodness-of-fit tests to the completed sample. We compare these tests by power in case of the II type censored samples and for comparison we give the power of the Neyman-Pearson test.**

Keywords-censored samples; goodness-of-fit; Kolmogorov test; Cramer-von Mises-Smirnov test; Anderson-Darling test; chi-square Nikulin-Rao-Robson test; Neyman-Pearson test

I. INTRODUCTION

In reliability or survival studies the observed data are usually presented as

$$\mathrm{X} = \left(X_1, \delta_1\right), \left(X_2, \delta_2\right), ..., \left(X_n, \delta_n\right),$$

where $X_i = \min\left(T_i, C_i\right)$ is the observation value, T_i is the lifetime and C_i is a censoring time. The indicator $\delta_i = 1$ if $X_i = T_i$ and $\delta_i = 0$ if $X_i = C_i$.

There are various types of right censoring schemes:

- If individuals are observed up to some predetermined time c, then censoring is called *type I censoring* and $\forall \delta_i = 0 : X_i = c$.
- If a life test is terminated whenever a specified number of failures r have occurred, it is called *type II censoring* and $\forall \delta_i = 0 : X_i = X_{(r)}$, where $X_{(r)}$ is the last observed failure time.
- Let the lifetime T and the censoring time C be independent random variables from the distribution functions $F(t)$ and $F^C(t)$, respectively. All lifetimes and censoring times are assumed mutually independent. Then censoring is called *independent random censoring*.

In this paper we consider the problem of testing simple goodness-of-fit hypotheses $H_0 : F = F_0$ and composite hypotheses, which can be presented as $H_0 : F \in \left\{F_0\left(\cdot;\theta\right), \theta \in \Theta\right\}$.

The modification of the classical Kolmogorov, Cramer-von Mises-Smirnov and Anderson-Darling tests for censoring types I and II are given in papers [2], [7] and [18]. In the case of randomly censored data these tests can be modified by using the Kaplan-Meier estimate instead of the empirical distribution function in the formulas of statistics (see, for example, [9], [12], [16] and [19]). In [6], [15] and [17] the distributions of the modified Kolmogorov, Cramer-von Mises-Smirnov and Anderson-Darling statistics were investigated by means of the Monte Carlo simulation method. In [15] the authors proposed approximations of the limiting distributions of the test statistics for type I and type II censored samples and carried out the power study.

Another group of goodness-of-fit tests for censored data contains the χ^2 type tests (see, for example, [1], [3], [8], [9], [10] and [11]). The idea of comparing observed and expected numbers of failures in time intervals was discussed in [2] and [9]. In [3] this direction was developed considering the choice of random grouping intervals as data functions and writing simple formulas useful for computing test statistics for mostly applied classes of survival distributions.

In this paper we discuss some advantages and disadvantages of the modified Kolmogorov, Cramer-von Mises-Smirnov, Anderson-Darling tests, as well as the tests for completed samples, the Nikulin-Rao-Robson and Neyman-Pearson tests and then compare these tests by power when testing simple and composite hypotheses. The investigation is carried out with the Monte Carlo method.

II. NONPARAMETRIC GOODNESS-OF-FIT TESTS FOR CENSORED DATA

A. Kolmogorov, Cramer-von Mises-Smirnov and Anderson-Darling Tests

One approach for testing goodness-of-fit hypotheses by complete samples is the application of the nonparametric tests:

the Kolmogorov, Cramer-von Mises-Smirnov and Anderson-Darling tests. The Kolmogorov test statistic is given by

$$D_n = \sup_{-\infty<t<\infty} \left|F_n(t) - F_0(t;\theta)\right|,$$

where $F_n(t)$ is the empirical distribution function. In practice, the statistic is usually used with the Bolshev correction [4] of the form

$$S_K = \frac{6nD_n + 1}{6\sqrt{n}}. \quad (1)$$

The Cramer-von Mises-Smirnov test statistic can be written by

$$S_\omega = \int_{-\infty}^{\infty} \left(F_n(t) - F_0(t;\theta)\right)^2 dF_0(t;\theta). \quad (2)$$

The Anderson-Darling test statistic can be presented as

$$S_\Omega = \int_{-\infty}^{\infty} \left(F_n(t) - F_0(t;\theta)\right)^2 \frac{dF_0(t;\theta)}{F_0(t;\theta)\left(1 - F_0(t;\theta)\right)}. \quad (3)$$

Let us denote the distribution of a test statistic under hypothesis H_0 as $G(S \mid H_0)$. In the case of testing simple hypotheses the distributions $G(S \mid H_0)$ of the considered statistics do not depend on the tested distribution. Statistic S_K belongs to the Kolmogorov distribution, S_ω and S_Ω belong to the $a1$ and the $a2$ distributions, respectively. For composite hypotheses the nonparametric test statistic distributions $G(S \mid H_0)$ are affected by a number of factors: the form of the tested lifetime distribution $F_0(t;\theta)$, the type and the number of estimated parameters, the method of parameter estimation and other factors. Approximations of limiting statistic distributions for testing various composite hypotheses have been proposed in [13] and [14].

The Kaplan-Meier estimate of the lifetime distribution is used instead of the empirical distribution $F_n(t)$ in modified Kolmogorov, Cramer-von Mises-Smirnov and Anderson-Darling tests for censored samples. In this case the statistics (1), (2) and (3) are calculated on the interval $[0,\tau]$, where $\tau = \max_{1\le i\le n}\{\delta_i X_i\}$ is the last observed failure time.

In [6], [15] and [17] we investigated the distributions of the modified statistics for type I, type II and randomly censored samples. The distributions of test statistics based on the Kaplan-Meier estimate considerably depend on the type of censoring and censoring degree. In the case of randomly censored samples distributions of considered test statistics also depend on the distribution of censoring times, which is usually unknown in practice.

So, for testing goodness-of-fit with the modified tests and calculating the *p*-value it is necessary to simulate the statistic distribution under a null hypothesis. But the process of simulation requires time and computational costs and it is not possible without special software. This is a significant disadvantage of the modified Kolmogorov, Cramer-von Mises-Smirnov and Anderson-Darling tests for censored samples.

B. Nikulin-Rao-Robson χ^2 test

Chi-square type tests require dividing an observed interval $[0,\tau]$ into k smaller intervals $I_j = (a_{j-1}, a_j]$, $a_0 = 0, a_k = \tau$.

Denote by $U_j = \sum_{X_i \in I_j}^{n} \delta_i$ the number of observed failures and by e_j the "expected" number of failures in the interval I_j, $j = 1,...,k$.

The Nikulin-Rao-Robson (NRR) χ^2 test statistic [3] can be written in the form

$$Y^2 = Z^T \hat{V}^{-} Z, \quad (4)$$

where $Z = (Z_1,...,Z_k)^T$, $Z_j = \frac{1}{\sqrt{n}}(U_j - e_j)$, $\hat{V}^{-}$ is the general inverse of the matrix $\hat{V}$, $\hat{V}^{-} = \hat{A}^{-1} + \hat{A}^{-1}\hat{C}^T\hat{G}^{-}\hat{C}\hat{A}^{-1}$, $\hat{A}$ is the diagonal $k \times k$ matrix with diagonal elements $\hat{A}_j = \frac{U_j}{n}, j = 1,...,k$, $\hat{G} = [\hat{g}_{ll'}]_{m\times m}$, $\hat{g}_{ll'} = \hat{i}_{ll'} - \sum_{j=1}^{k} \hat{C}_{lj}\hat{C}_{l'j}\hat{A}_j^{-1}$,

$$\hat{i}_{ll'} = \frac{1}{n}\sum_{\delta_i=1}^{n} \frac{\partial \ln \lambda(X_i;\hat{\theta})}{\partial \theta_l} \frac{\partial \ln \lambda(X_i;\hat{\theta})}{\partial \theta_{l'}}, \quad \hat{C}_{lj} = \frac{1}{n}\sum_{\substack{X_i \in I_j \\ \delta_i=1}}^{n} \frac{\partial \ln \lambda(X_i;\hat{\theta})}{\partial \theta_l},$$

where $\lambda(t)$ is the hazard rate function and $\hat{\theta}$ is the maximum likelihood estimate of unknown parameters.

In [3] it is recommended to take a_j as random data functions so that to divide the interval $[0,\tau]$ into k intervals with equal expected numbers of failures. So, a_j can be calculated as follows: define $E_k = \sum_{i=1}^{n} \Lambda(X_i;\hat{\theta})$, $b_i = (n-i)\Lambda(X_{(i)};\hat{\theta}) + \sum_{l=1}^{i} \Lambda(X_{(i)};\hat{\theta})$, $E_j = \frac{j}{k}E_k$, $j = 1,...,k$, $X_{(0)} = 0$. If i is the smallest natural number verifying $E_j \in [b_{i-1}, b_i]$, $j = 1,...,k-1$, then

$$\hat{a}_j = \Lambda^{-1}\left(\left(E_j - \sum_{l=1}^{i} \Lambda(X_{(l)};\hat{\theta})\right)/(n-i+1), \hat{\theta}\right), \quad \hat{a}_k = X_{(n)},$$

where Λ^{-1} is the inverse of the cumulative hazard function Λ.

The limiting distribution of the test statistic is χ^2_r, $r = rank(V^-)$. So the hypothesis is rejected with approximate significance level α if $Y^2 > \chi^2_\alpha(r)$.

If a simple hypothesis is tested, the test statistic has the following form

$$Y^2 = \sum_{j=1}^{k} \frac{\left(U_j - e_j\right)^2}{U_j}.$$

The choice of $\hat{a}_j$ is the same as in the composite hypothesis case, we need only to skip $\hat{\theta}$ in all formulas. The limiting distribution of the test statistic is χ^2_k. So, the hypothesis is rejected with approximate significance level α if $Y^2 > \chi^2_\alpha(k)$.

The NRR χ^2 test has a number of advantages as compared with the modified nonparametric tests mentioned above. In particular, there is no significant dependence of the NRR statistic distributions on the distribution of censoring times. As the sample size grows, NRR statistic distributions converge to the corresponding χ^2 distribution law. It is possible to choose the number of grouping intervals and boundary points for considered pair of competing hypotheses to increase the test power. Nevertheless, it is necessary to mention here that calculation of this statistic is rather complicated.

III. Application of Nonparametric Goodness-of-Fit Tests for Completed Samples

We propose to transform the original censored sample to a complete sample and to use the classical Kolmogorov, Cramer-von Mises-Smirnov and Anderson-Darling tests for the obtained sample.

A. Simple Hypothesis

Let us consider a simple hypothesis. That is, F_0 is completely known. For each censored observation $(X_i, \delta_i = 0)$ a value $\hat{T}_i$ is generated by $\hat{T}_i = F_0^{-1}(\xi_i)$, where $\xi_i \sim U\left[F_0(C_i), 1\right]$. The corresponding values of censored observations in the original sample are replaced by generated values. Hereby, we obtain a transformed complete sample $\hat{X}_1, \hat{X}_2, ..., \hat{X}_n$, in which

$$\hat{X}_i = \begin{cases} X_i, \text{ if } \delta_i = 1 \\ \hat{T}_i, \text{ if } \delta_i = 0 \end{cases}.$$

Under the simple null hypothesis the distribution of the completed sample is F_0. So, the algorithm of testing simple hypotheses can be written as follows

1) Specify the significance level α.

2) Transform the original censored sample X *to a complete sample $\hat{X}_1, \hat{X}_2, ..., \hat{X}_n$.*

3) Calculate the test statistic S^* ((1), (2) or (3)) by the obtained sample and the theoretical distribution F_0.

4) Compute the p-value: $p = 1 - G(S^* \mid H_0)$.

5) The hypothesis H_0 is rejected if the obtained p-value is less than α.

By computer simulation methods we have shown that test statistic distributions based on the transformed samples when testing simple hypotheses do not depend on the distribution $F^C(t)$ and coincide with the corresponding limiting distributions (for example, the Kolmogorov distribution for Kolmogorov's test statistic) [5].

B. Composite Hypothesis

Let $\hat{\theta}$ be a maximum likelihood estimate of unknown parameter calculated by the original censored sample. We replace all censored observations X_i with $\delta_i = 0$ by simulated times $\hat{T}_i = F_0^{-1}(\xi_i)$, where ξ_i is uniformly distributed at the interval $[F(C_i, \hat{\theta}), 1)$. So, we obtain a completed sample $\hat{X}_1, \hat{X}_2, ..., \hat{X}_n$.

After replacing censored observations by generated observations with values $\hat{T}_i$ it is necessary to estimate unknown parameters θ again. The algorithm of testing composite hypotheses by using transformation of a censored sample can be written as follows

1) Specify the significance level α.

2) Calculate the maximum likelihood estimates $\hat{\theta}$ by the original sample X.

3) Transform the original censored sample X *to complete sample $\hat{X}_1, \hat{X}_2, ..., \hat{X}_n$.*

4) Calculate the maximum likelihood estimates $\tilde{\theta}$ by the transformed sample $\hat{X}_1, \hat{X}_2, ..., \hat{X}_n$.

5) Calculate the test statistic S^* ((1), (2) or (3)) by the obtained sample and the theoretical distribution $F_0(\cdot; \tilde{\theta})$.

6) Compute the p-value: $p = 1 - G(S^* \mid H_0)$.

7) The hypothesis H_0 is rejected if the obtained p-value is less then α.

By computer simulation methods we have shown that test statistic distributions based on the transformed samples when testing composite goodness-of-fit hypotheses do not differ from corresponding approximations of the limiting distributions, if the censoring degree is small ($\leq 30\%$) [5].

So, in the case of testing a simple goodness-of-fit hypothesis by censored samples including randomly censored

samples it is possible to use the classical Kolmogorov, Cramer-von Mises-Smirnov and Anderson-Darling test, and to calculate the *p*-value based on the Kolmogorov, $a1$ and $a2$ limiting distributions. When testing a composite hypothesis it is necessary to take into account the number of censored observations, as if the censoring degree is large, then the application of the approximations obtained for complete samples to calculate the *p*-value can be incorrect.

IV. Neyman-Pearson Test

Carrying out a comparative analysis of different goodness-of-fit tests it is quite rational to compare considered tests with the most powerful test. According to the Neyman-Pearson lemma in the case of distinguishing between two models, one of which (the null model) is a special case of the other (the alternative model) and each of which has no unknown parameters, the likelihood ratio test has the highest power among all competitors. The likelihood ratio statistic is written as follows

$$\Lambda(\mathrm{X}) = \frac{L(\mathrm{X} \mid H_1)}{L(\mathrm{X} \mid H_0)}, \tag{5}$$

where $L(\mathrm{X}) = \prod_{i=1}^{n} f^{\delta_i}(X_i) \cdot (1 - F(X_i))^{1-\delta_i}$ is the likelihood function for $\mathrm{X} = (X_1, \delta_1), (X_2, \delta_2), ..., (X_n, \delta_n)$. Usually it is more convenient to consider the logarithm of the statistic (5). Let α be the specified significance level, then H_0 is rejected if $\Lambda(\mathrm{X}) > B_\alpha$ and B_α is such that $\mathrm{P}(\Lambda(\mathrm{X}) > B_\alpha \mid H_0) = \alpha$.

The distribution of the statistic $\Lambda(\mathrm{X})$ can be simulated by means of the Monte Carlo method. So, in the case of the simple hypothesis $H_0: F(x) = F_1(x; \theta = \theta_0)$ against the simple alternative $H_1: F(x) = F_1(x; \theta = \theta_1)$ the algorithm of simulating the distribution $G(\Lambda(\mathrm{X}) \mid H_0)$ can be written as follows

1) Generate a sample of observations X *according to the null hypothesis.*

2) Calculate the likelihood functions: $L(\mathrm{X} \mid H_0)$ *and* $L(\mathrm{X} \mid H_1)$ *and the value of the statistic (5).*

By repeating the above process N times, a random sample from the distribution of the test statistic is generated.

Now let us consider the composite hypothesis $H_0: F(x) = F_1(x; \theta \in \Omega_0)$ against the composite alternative $H_1: F(x) = F_1(x; \theta \in \Omega_1)$, where $\Omega_0 \subset \Omega_1, \Omega_1 \setminus \Omega_0 \neq \varnothing$. Then the algorithm of simulating the distribution $G(\Lambda(\mathrm{X}) \mid H_0)$ can be written as follows

1) Generate a sample of observations X *according to the null hypothesis.*

2) Estimate unknown parameters maximizing the likelihood function $L(\mathrm{X} \mid H_0)$ *by* $\theta \in \Omega_0$.

3) Estimate unknown parameters maximizing the likelihood function $L(\mathrm{X} \mid H_1)$ *by* $\theta \in \Omega_1$.

4) Calculate the statistic (5) in which $L(\mathrm{X} \mid H_0)$ *and* $L(\mathrm{X} \mid H_1)$ *have been obtained in steps 2 and 3 correspondingly.*

By repeating the above process N times, a random sample from the distribution of the test statistic is generated.

In some particular cases, the distribution of the Neyman-Pearson test statistic can be found analytically, but in most cases we need to apply Monte Carlo simulations.

V. Empirical Power Study

In [15] we investigated the power of the modified Kolmogorov, Cramer-von Mises-Smirnov and Anderson-Darling tests in the case of type II censored samples and in [6] these tests were compared by power with the NRR χ^2 test. The power of classical Kolmogorov, Cramer-von Mises-Smirnov and Anderson-Darling tests for samples completed from original randomly censored samples was studied in [5]. It was shown that the power of considered tests increases as the sample size grows. It is necessary to note that in the case of type I and random censoring we cannot simulate the same censoring scheme for both hypotheses H_0 and H_1, hereupon it is not correct to compare different tests by power in such situations.

In this paper we compare the considered goodness-of-fit tests by power on the example of two different pairs of competing hypotheses by means of computer simulation methods. The estimates of the power are calculated basing on the distributions of the test statistics $G(S \mid H_0)$ and $G(S \mid H_1)$, which are simulated for type II censored samples of the size $n = 200$. In the case of composite hypotheses the distribution parameters are estimated by the maximum likelihood method. The number of simulations is $N = 10^5$. The values of the power of the tests are calculated for the significance level $\alpha = 0.1$.

A. First Pair of Competing Hypotheses

The null hypothesis H_0 is the exponential distribution with the distribution function $F(x; \theta_1) = 1 - \exp\left(-\frac{x}{\theta_1}\right)$ and the parameter $\theta_1 = 2$. The competing hypothesis H_1 is the Weibull distribution with the distribution function $F(x; \theta_1, \theta_2) = 1 - \exp\left(-\left(\frac{x}{\theta_1}\right)^{\theta_2}\right)$ and parameters $\theta_1 = 2,\ \theta_2 = 0.9$.

In Table I, the powers of the considered tests are given for various censoring degrees in the case of testing the simple

hypothesis. The modified Kolmogorov, Cramer-von Mises-Smirnov and Anderson-Darling tests are denoted as S_K^M, S_ω^M and S_Ω^M, respectively, the classical Kolmogorov, Cramer-von Mises-Smirnov and Anderson-Darling tests for completed samples are denoted as S_K^R, S_ω^R and S_Ω^R, respectively, the NRR χ^2 test is denoted as Y^2 with the number of intervals k, and the Neyman-Pearson test is denoted by Λ.

TABLE I. TEST POWER FOR THE FIRST PAIR WHEN TESTING THE SIMPLE HYPOTHESIS

Test	Censoring Degrees								
	0%	*10%*	*20%*	*30%*	*40%*	*50%*	*60%*	*70%*	*80%*
S_K^M	0.21	0.21	0.21	0.21	0.20	0.22	0.22	0.23	0.22
S_ω^M	0.21	0.22	0.21	0.21	0.22	0.23	0.23	0.22	0.18
S_Ω^M	0.33	0.30	0.28	0.28	0.28	0.30	0.30	0.31	0.29
S_K^R	0.21	0.21	0.20	0.20	0.19	0.19	0.19	0.19	0.17
S_ω^R	0.21	0.21	0.21	0.20	0.19	0.20	0.21	0.20	0.18
S_Ω^R	0.33	0.31	0.27	0.25	0.24	0.25	0.25	0.24	0.23
$Y^2,3$	0.47	0.33	0.26	0.20	0.18	0.16	0.14	0.13	0.10
$Y^2,5$	0.42	0.31	0.25	0.20	0.17	0.13	0.11	0.08	0.05
$Y^2,7$	0.38	0.29	0.29	0.18	0.14	0.11	0.08	0.08	0.04
Λ	0.77	0.68	0.61	0.57	0.53	0.51	0.50	0.50	0.50

As you can see from Table I, the power estimates of considered tests decrease, as the censoring degree grows, with the exception of the modified nonparametric tests. The modified Anderson-Darling test has the higher power than the modified Kolmogorov and Cramer-von Mises-Smirnov tests. The same regularity is observed in the case of tests for samples completed from original censored samples. The power estimates of the Nikulin-Rao-Robson χ^2 test are higher for the lower number of intervals ($k = 3$). Moreover, this test is more powerful than the Kolmogorov, Cramer-von Mises-Smirnov and Anderson-Darling tests (both modified tests and tests for completed samples) when the censoring degree is lower than 30%. But in comparison with the Neyman-Pearson test, which is the most powerful test, the nonparametric tests considerably lose in power.

In Table II, the powers of the considered tests are given in the case of testing the composite hypothesis.

TABLE II. TEST POWER FOR THE FIRST PAIR WHEN TESTING THE COMPOSITE HYPOTHESIS

Test	Censoring Degrees			
	0%	*10%*	*20%*	*30%*
S_K^M	0.45	0.38	0.33	0.29
S_ω^M	0.51	0.43	0.37	0.33
S_Ω^M	0.56	0.49	0.43	0.39
S_K^R	0.45	0.35	0.29	0.24
S_ω^R	0.51	0.40	0.33	0.27
S_Ω^R	0.56	0.46	0.40	0.34
$Y^2,3$	0.39	0.32	0.27	0.23
$Y^2,5$	0.34	0.28	0.24	0.21
$Y^2,7$	0.31	0.26	0.22	0.19
Λ	0.59	0.49	0.44	0.38

As it is seen from Table II, the Neyman-Pearson test is more powerful than the other considered tests, just like in the case of testing simple hypothesis considered earlier. One can easily notice that the power of the Kolmogorov test is lower than the power of other nonparametric tests. And the Anderson-Darling test turned out to be the most powerful in comparison with the Cramer-von Mises-Smirnov, Kolmogorov and Nikulin-Rao-Robson tests.

B. Second Pair of Compiting Hypotheses

The null hypothesis H_0 is the Weibull distribution with parameters $\theta_1 = 2,\ \theta_2 = 2$. The competing hypothesis H_1 is the Generalized Weibull distribution with the distribution function $F(x;\theta_1,\theta_2,\theta_3) = 1 - \exp\left(1 - \left(1 + \left(\frac{x}{\theta_1}\right)^{\theta_2}\right)^{\frac{1}{\theta_3}}\right)$ and parameters $\theta_1 = 2,\ \theta_2 = 2,\ \theta_3 = 0.9$.

In Table III, the powers of the considered tests are given for various censoring degrees in the case of testing the simple hypothesis.

TABLE III. POWER FOR SECOND PAIR WHEN TESTING THE SIMPLE HYPOTHESIS

Test	Censoring Degrees								
	0%	*10%*	*20%*	*30%*	*40%*	*50%*	*60%*	*70%*	*80%*
S_K^M	0.50	0.49	0.46	0.41	0.34	0.28	0.22	0.16	0.11
S_ω^M	0.55	0.50	0.43	0.36	0.28	0.22	0.17	0.11	0.11
S_Ω^M	0.59	0.49	0.41	0.34	0.27	0.22	0.18	0.14	0.11
S_K^R	0.50	0.49	0.47	0.43	0.37	0.30	0.23	0.16	0.12
S_ω^R	0.55	0.54	0.50	0.45	0.38	0.31	0.24	0.17	0.13
S_Ω^R	0.59	0.55	0.49	0.43	0.36	0.30	0.23	0.17	0.14
$Y^2,3$	0.51	0.40	0.32	0.22	0.19	0.14	0.11	0.08	0.07
$Y^2,5$	0.35	0.25	0.19	0.15	0.11	0.09	0.07	0.06	0.06
$Y^2,7$	0.24	0.17	0.12	0.09	0.07	0.06	0.05	0.06	0.06
Λ	0.88	0.81	0.74	0.66	0.59	0.50	0.42	0.35	0.27

As you can see from Table III, the power of the NRR χ^2 test for $k = 3$ is almost equal to the power of the Kolmogorov, Cramer-von Mises-Smirnov and Anderson-Darling tests in the case of complete samples, but with increasing the censoring degree the NRR χ^2 test loses in power. The power of the classical Kolmogorov, Cramer-von Mises-Smirnov and Anderson-Darling tests for completed samples is higher than the power of the modified tests for all considered censoring degrees. The Anderson-Darling test for completed samples turned out to be the second in power among the considered tests, with the Neyman-Pearson test being the first, although the difference in power between them is huge.

The estimates of the power of all the considered tests in the case of the composite hypothesis turned out to be equal to the significance level ($\alpha = 0.1$), that is none of the considered tests can distinguish between the Weibull distribution and the Generalized Weibull distribution, which can be explained as follows: for the Generalized Weibull distribution with the chosen parameter values, it is possible to fit a very close Weibull distribution.

VI. CONCLUSIONS

So, what is the answer to the question: "Which goodness-of-fit test is most preferable for right censored samples?" There is no unique test, which could be the best in any situation. Nevertheless, the results of our investigation allow us to formulate some advantages and disadvantages of the considered tests. As to the modified Kolmogorov, Cramer-von Mises-Smirnov and Anderson-Darling tests: in the case of type I and type II censored data one can use the tables of percentage points for these tests in simple hypothesis testing [15]. While testing a composite hypothesis it is possible to simulate the distribution of test statistics for the distribution under test and censoring scheme given. But when we have a randomly censored sample the distributions of these statistics strongly depend on both the distribution of lifetimes and the distribution of censoring times, which is usually unknown in practice. We do not recommend using the modified Kolmogorov, Cramer-von Mises-Smirnov and Anderson-Darling tests for randomly censored data.

There is a good possibility to use the considered transformation of a censored sample to a complete one and then to apply the classical Kolmogorov, Cramer-von Mises-Smirnov and Anderson-Darling tests for a completed sample. In the case of small censoring degrees, when there is no significant bias of parameter estimates, it is possible to use the approximations of the limiting statistic distributions obtained in [11] for the calculation of a *p*-value while testing composite hypotheses. The loss of power is not significant if the censoring degree is not high.

The NRR test has a number of advantages compared with the considered nonparametric tests. In particular, there is no significant dependence of the NRR statistic distributions (in the case of limited sample sizes) on the distribution of censoring times. As the sample size grows, NRR statistic distributions converge to the corresponding χ^2 distribution law. In this paper we considered this test as a nonparametric test, but for some given pair of competing hypotheses it is possible to find the optimal number of grouping intervals and boundary points that maximize the test power.

REFERENCES

[1] M. G. Akritas, "Pearson-type goodness-of-fit tests: the univariate case," J. Amer. Statist. Assoc., Vol. 83, pp. 222-230, 1988.

[2] D. M. Barr and T. Davidson, "A Kolmogorov-Smirnov test for censored samples," Technometrics, 15, 4, 1973.

[3] V. Bagdonavicus, J. Kruopis and M. Nikulin, "Nonparametric tests for censored data," John Wiley and Sons, Inc., New York, 2010.

[4] L. N. Bolshev, N. V. Smirnov, "Tables of Mathematical Statistics," Moscow: Science. (in Russian), 1983.

[5] E. Chimitova, H. Liero and M. Vedernikova, "Application of classical Kolmogorov, Cramer-von Mises-Smirnov and Anderson-Darling tests for censored samples," Proccedings of the International Workshop "Applied Methods of Statistical Analysis". Novosibirsk: Publishing house of NSTU, pp. 176-185, 2011.

[6] E. V. Chimitova, A. O. Tsivinskaya, "Simulation study for the NRR chi-square test of goodness-of-fit for censored data," Proccedings of the International Workshop "Applied Methods of Statistical Analysis". Novosibirsk: Publishing house of NSTU, pp. 44-52, 2011.

[7] R. B. D'Agostino and M. A. Stephens, "Goodness of fit techniques," New York: Marcel Dekker, 1986.

[8] M. G. Habib, D. R. Thomas, "Chi-square goodness-of-_t tests for randomly censored data," Annals of Statistics, Vol. 14, pp. 759-765, 1986.

[9] N. L. Hjort, "On inference in parametric survival data," International Statistical Review, 60, 3, pp. 355-387, 1992.

[10] M. Hollander, E. A. Pena, "A chi-squared goodness-of-fit test for randomly censored data," J. Amer. Statist. Assoc., Vol. 87, pp. 458-463, 1992.

[11] J. H. Kim, "Chi-square goodness-of-fit tests for randomly censored data," The Annals of Mathematical Statistics, Vol. 21, № 3, pp. 1621-1639, 1993.

[12] J. A. Koziol and S. B. Green, "A Cramer-von Mises statistic for randomly censored data," Biometrika, 63, 3, pp. 465-474, 1976.

[13] B. Yu. Lemeshko, S. B. Lemeshko, "Distribution models for nonparametric tests for fit in verifying complicated hypotheses and maximum-likelihood estimators. Part 1," Measurement Techniques. Vol. 52, № 6, pp. 555-565, 2009.

[14] B. Yu. Lemeshko, S. B. Lemeshko, "Models for statistical distributions in nonparametric fitting tests on composite hypotheses based on maximumlikelihood estimators. Part II," Measurement Techniques. Vol. 52, № 8, pp. 799-812, 2009.

[15] B. Yu. Lemeshko, E. V. Chimitova, T. A. Pleshkova, "Testing simple and composite goodness-of-fit hypotheses by censored samples," Nauchniy vestnik NGTU, №4(41), pp. 13-28, 2010.

[16] V. Nair, "Plots and tests for goodness of fit with randomly censored data," Biometrika, 68, pp. 99-103, 1981.

[17] E. Chimitova, M. Nikulin, B. Lemeshko, A. Tsivinskaya, "Nonparametric goodness-of-fit tests for censored data," The 7th international Conference on "Mathematical methods in reliability. Theory. Methods. Applications", Beijing, China. June 20-24, pp.817-823, 2011.

[18] A. N. Pettitt and M. A. Stephens, "Modified Cramer von Mises statistics for censored data," Biometrika, 63, 2, 1976.

[19] D. Reineke and J. Crown, "Estimation of hazard, density and survival functions for randomly censored data," Journal of Applied Statistics, 31, 10, pp. 1211-1225, 2004.

A degradation model with stochastic process drift applied to coronary heart disease

Daniel Commenges
INSERM U897 and
Univ Bordeaux
Bordeaux, France
Email: daniel.commenges@isped.u-bordeaux2.fr

Boris Hejblum
INSERM U897 and
Univ Bordeaux
Bordeaux, France
Email: boris.hejblum@isped.u-bordeaux2.fr

***Abstract*—We envisage a degradation model for coronary heart disease (CHD). A degradation model make sense since CHD is the result of the progressive deposit of atheroma in the arteries. The model is that a CHD event occurs when a diffusion process representing the atheromatous process crosses a certain threshold. It is interesting to model the drift as a function of physiological conditions such as high blood pressure, obesity, inflammation and cholesterol level. If these factors are fixed, inference is easy because the hitting time distribution is inverse Gaussian. We may model their evolution in time as stochastic processes, leading to a drift which is a stochastic process. We discuss the way inference could be carried on in these cases, particularly when the processes describing physiological conditions are Ornstein-Ulhenbeck processes.**

I. Introduction

It is often the case that an event occurs when a degradation process reaches a certain threshold [1]–[5]. Gaussian processes have often been considered but also gamma processes have been proposed [6]; see [7] for recent advances in the topic. It has been shown that myocardial infarction (MI) is most commonly due to occlusion (blockage) of a coronary artery following the rupture of an atherosclerotic plaque [8], [9]. A Brownian motion with positive drift seems well adapted to describe the progressive growth of atheroma, with MI occurring when this processes reaches a certain threshold. One advantage of this approach is that we can link MI with broader coronary heart disease (CHD) events which happen before MI when occlusion is not complete but the heart already suffers from hypoxia. These CHD events may occur when the atheromatous process reaches a threshold below that required for MI. One of the use of such a model is to express the effect of risk factors. Several risk factors are already known for MI or CHD. Most analyzes focus on one particular risk factor rather than presenting a global model. Few works have attempted to develop more global dynamic analysis: Wilson et al. [10], using data from the Framingham study, developed prediction scores using indicators of lipid profile, diabetes, obesity, blood pressure and tobacco consumption as explanatory variables in a Cox model; Gamborg et al. [11] used the dynamic path analysis of Fosen et al. [12] to take into account the possible evolution of obesity and blood pressure.

A possibility with a degradation model is to express the drift of the degradation process as a function of risk factors. For CHD we may consider physiological conditions that have a direct effect on the atheromatous process: high blood pressure, obesity, inflammation and cholesterol level. If these factors are considered as fixed, inference is still relatively easy since conditionally on these factors, the hitting time distribution is inverse Gaussian. However in real life these factors may change with time. One may propose deterministic or stochastic functions for modeling these changes.

II. A degradation model for myocardial infarction and CHD

A. The degradation model; modeling the drift

As described in [13], the atheromatous process $A(t)$ can be described by a Brownian motion with drift: $dA(t) = \lambda \mathrm{d}t + \mathrm{d}B_{_A}(t)$; the time parameter can be taken as age (in year) minus 20. A basic degradation model is that a CHD event happens when the atheromatous process reaches a certain threshold η, that is, the counting process $\mathrm{CHD}(t)$ is defined as: $\mathrm{CHD}(t) = \mathbb{1}_{\{A(t)>\eta\}}$. Then the jump time T of $\mathrm{CHD}(t)$ (the hitting time of the Brownian motion with drift) has an inverse Gaussian ($\mathcal{I}G$) distribution with parameters $(\eta/\lambda, \eta^2)$; its density is:

$$f(t) = \left[\frac{\eta^2}{2\pi t^3}\right]^{1/2} \exp\left(\frac{-\lambda^2\left(t-\frac{\eta}{\lambda}\right)^2}{2t}\right) \quad \mathbb{1}_{\{t\geq 0\}}.$$

What is interesting from an epidemiological point of view is to model the drift as a function of physiological conditions suspected to play a role in the atheromatous process. Here we will take into account four of them: obesity, represented by body mass index (BMI), lipid profile represented by low density lipid concentration (LDL), inflammation process represented by C-reactive protein concentration (CRP) and blood pressure represented by systolic blood pressure (SBP). We assume a linear model for λ:

$$\lambda = \lambda_0 + \beta_{\text{BMI}}\text{BMI} + \beta_{\text{LDL}}\text{LDL} + \beta_{\text{CRP}}\text{CRP} + \beta_{\text{SBP}}\text{SBP} + \varepsilon,$$

where λ_0 is a baseline drift and ε has a normal distribution with zero expectation. Commenges and Hejblum [13] used Box-Cox transforms of BMI, LDL, CRP and SBP in order to get an approximately multinormal distribution. Conditionally

on the values of these factors, the time of occurrence of CHD has an $\mathcal{I}G$ distribution.

For inference there may be missing data. Thus we need all the marginal distributions of T, and for computing them, the marginal distributions of λ. Assuming the indicators have a multinormal distribution, all marginal distributions of λ are normal, so that the marginal distributions of T are inverse Gaussian Normal ($\mathcal{IGN}$) [14]. If the drift parameter λ has a $\mathcal{N}(m_\lambda, s_\lambda)$ distribution, then the hitting time T has the distribution $\mathcal{IGN}(m_\lambda, s_\lambda, \eta)$, with density:

$$f(t) = \left[\frac{\eta^2}{2\pi t^3\left(1+s_\lambda^2 t\right)}\right]^{1/2} \exp\left(\frac{-(\eta - m_\lambda t)^2}{2t(1+s_\lambda^2 t)}\right) \quad \mathbb{1}_{\{t\geq 0\}}$$

This is an improper distribution in that $P(T=\infty) > 0$; this is not a problem in our model since not everybody develops a MI.

The model can be extended for defining two hitting times. It often happens that the progression of the atheromatous process first produces symptoms related to hypoxia (like angina pectoris) before the completion of MI; CHD includes these symptoms as well as MI. Thus two thresholds that we will denote η_{CHD} and η_{MI} can be defined and determine the distribution of the time of occurrence of CHD, T_{CHD}, and of MI, T_{MI}.

B. Modeling drift and threshold

Another model arises if the threshold rather than the drift varies with the value of an explanatory variable. [5] (section 10.3.8) present a degradation model where both drift and starting point may depend on covariates (modeling the starting point or the threshold are mathematically equivalent, although they may reflect different interpretations). For instance it may be asked whether SBP has a cumulative effect or merely favors the occurrence of MI for a given state of the atheromatous process. The model for η could be: $\eta = \eta_0 + \beta'_{\text{SBP}}\text{SBP}$ while the model for λ involves BMI, LDL and CRP. Of course SBP could modify both η and λ.

Conditionally on the explanatory variables the distribution of T is still $\mathcal{I}G$. As before, in case of missing data we need marginal distributions for which both λ and η are normal. We call the resulting distribution inverse Gaussian normal-normal ($\mathcal{IGNN}$). It happens that the density of this distribution has an analytic form. [13] used these results for estimating the parameters of the model by making a synthesis analysis of several studies about CHD.

III. Physiological conditions as stochastic processes

If the factors involved in the model of the drift change with time, the hitting time distribution is no more inverse Gaussian. For instance, as proposed by [15], the physiological conditions could be modeled as Ornstein-Uhlenbeck processes. For instance, the LDL process could be modeled as:

$$d\text{LDL}_t = \theta_{\text{LDL}}(\text{LDL}_t - \mu_{\text{LDL}})dt + dB_{\text{LDL},t},$$

where B_{LDL} is a Brownian motion. The other factors involved in the drift of the atheromatous process could be modeled in a similar way. Thus the drift of the atheromatous process is a stochastic process, the law of which can be computed as a function of the parameters of the Ornstein-Uhlenbeck processes. It is itself a Gaussian processs.

This modeling raises a certain number of difficulties when it comes to estimating the parameters. We could have observations of the processes characterizing the physiological conditions. For instance we could observe $Y_j = \text{LDL}_{t_j} + \varepsilon_{\text{LDL},j}, j = 1, n_{\text{LDL}}$. We could also have observations of the atheromatous process itself. Finally we must have observations of CHD events. The joint distribution of the observations of Gaussian processes is a multinormal with expectation and variance which can be computed from the parameters of the model. The problem comes from writing the likelihood of the times of occurrence of CHD which are hitting times of a Gaussian process. In spite of recent progress in this field, [16], the hitting time distribution for a general Gaussian process is, to our knowledge, not available. We have to resort to simulation or approximation. There is a literature on computing the likelihood for SDE based on simulation, [17], that could be adapted to our problem. A possible simpler approach would be to approximate the likelihood by remarking that the observation T of the first hitting time of a level η is equivalent to observing $\{A(u) < \eta, 0 < u < T, A(T) = \eta\}$. We could replace the likelihood of this event by that of $\{A(u_k) < \eta, k = 1\ldots, K; u_K < T, A(T) = \eta\}$, which can be computed using numerical integration for a multinormal distribution. Good algorithms exist for this problem [18] so this is feasible if K is not too large.

IV. Conclusion

A degradation model is well adapted to describe the occurrence of CHD events. The drift and threshold can be modeled as a function of physiological conditions and [13] has proposed an approach for inference based on a synthesis of data from the literature. In a perspective of lifecourse emidemiology it is interesting to model the physiological conditions themselves as stochastic processes. This leads to consider that the drift of the degradation process is it self a stochastic process. If the physiological conditions are Gaussian processes, the degradation process is a Gaussian process. The likelihood for such a model has to be computed by simulation or approximated. In the case of a Gaussian process an approximation seems feasible and involves numerical integration.

References

[1] K. Doksum and S. Normand, "Gaussian models for degradation processes-part i: Methods for the analysis of biomarker data," *Lifetime Data Analysis*, vol. 1, no. 2, pp. 131–144, 1995.

[2] O. O. Aalen and H. k. K. Gjessing, "Understanding the shape of the hazard rate: a process point of view," *Statistical Science*, vol. 16, no. 1, pp. 1–22, 2001.

[3] R. Hashemi, H. Jacqmin-Gadda, and D. Commenges, "A latent process model for joint modeling of events and marker," *Lifetime Data Analysis*, vol. 9, no. 4, pp. 331–343, 2003.

[4] M.-L. T. Lee and G. A. Whitmore, "Threshold Regression for Survival Analysis: Modeling Event Times by a Stochastic Process Reaching a Boundary," *Statistical Science*, vol. 21, no. 4, pp. 501–513, Aug. 2006.
[5] O. Aalen, Ø. Borgan, and H. Gjessing, *Survival and event history analysis: a process point of view*. Springer Verlag, 2008.
[6] V. Bagdonavicius and M. Nikulin, "Estimation in degradation models with explanatory variables," *Lifetime Data Analysis*, vol. 7, no. 1, pp. 85–103, 2001.
[7] M. Nikulin, N. Limnios, and N. Balakrishnan, *Advances in Degradation Modeling: Applications to Reliability, Survival Analysis, and Finance*. Birkhauser, 2009.
[8] G. Hansson, "Inflammation, atherosclerosis, and coronary artery disease," *New England Journal of Medicine*, vol. 352, no. 16, pp. 1685–1695, 2005.
[9] S. J. Nicholls, "Relationship between LDL, HDL, blood pressure and atheroma progression in the coronaries," *Current Opinion in Lipidology*, vol. 20, no. 6, pp. 491–496, Dec. 2009.
[10] P. W. F. Wilson, R. B. D'Agostino, D. Levy, A. M. Belanger, H. Silbershatz, and W. B. Kannel, "Prediction of coronary heart disease using risk factor categories." *Circulation*, vol. 97, no. 18, pp. 1837–1847, May 1998.
[11] M. Gamborg, G. Jensen, T. Sørensen, and P. Andersen, "Dynamic path analysis in life-course epidemiology," *American Journal of Epidemiology*, vol. 173, no. 10, p. 1131, 2011.
[12] J. Fosen, E. Ferkingstad, Ø. Borgan, and O. Aalen, "Dynamic path analysisa new approach to analyzing time-dependent covariates," *Lifetime data analysis*, vol. 12, no. 2, pp. 143–167, 2006.
[13] D. Commenges and B. Hejblum, "Evidence synthesis through a degradation model applied to myocardial infarction," *submitted*, 2012.
[14] G. A. Whitmore, "Normal-gamma mixtures of inverse Gaussian distributions," *Scandinavian Journal of Statistics*, vol. 13, no. 3, pp. 211–220, 1986.
[15] D. Commenges, "The stochastic system approach to causality with a view toward lifecourse epidemiology," *submitted*, 2012.
[16] L. Decreusefond and D. Nualart, "Hitting times for gaussian processes," *The Annals of Probability*, vol. 36, no. 1, pp. 319–330, 2008.
[17] A. Beskos, O. Papaspiliopoulos, and G. Roberts, "Monte carlo maximum likelihood estimation for discretely observed diffusion processes," *The Annals of Statistics*, vol. 37, no. 1, pp. 223–245, 2009.
[18] A. Genz, "Numerical computation of multivariate normal probabilities," *Journal of computational and graphical statistics*, pp. 141–149, 1992.

Semiparametric Regression Analysis of Panel Count Data with Time-Dependent Covariates and Informative Observation and Censoring Times

Shirong Deng
Department of Applied Mathematics,
The Hong Kong Polytechnic University,
Hong Kong, P. R. China
Email: Deng.Shirong@connect.polyu.hk

***Abstract*—In this paper, we extend the joint frailty models proposed by Zhao and Tong (2011) to panel count data with the time-dependent covariates and informative observation and censoring times. A novel estimating equation approach that does not depend on the distribution of frailty variables and the link function is proposed for estimation of parameters, and the asymptotic properties of the proposed estimators are established. Simulation studies demonstrate that the proposed inference procedure performs well. The analysis of a bladder tumor data is presented to illustrate the method.**

I. Introduction

Recurrent events may occur frequently in a wide variety of settings. For their analysis, the important information including the observation times, the counts of recurrent events, the censoring or follow-up times and the covariates related to the study are recorded for each study subject. However, in some studies, it may not be possible to record the exact event times. For example, the examination may be too expensive or the events may occur too frequently for their exact times to be recorded. Each subject may be observed at several distinct times and only the numbers of events between two adjacent times are available. Moreover, the set of observation times may vary from subject to subject. Such data are called panel count data, which often occur in many fields, such as demographic studies, industrial reliability and clinical trials; see Kalbfleisch and Lawless (1985), Gaver and O'Muicheartaigh (1987), Thall and Lachin (1988) and Sun and Kalbfleisch (1995)

Recently, the nonparametric and semiparametric analysis of panel count data have attracted considerable attention. For example, Kalbfleisch and Lawless (1985) discussed the fitting of Markov model to panel count data. Sun and Kalbfleisch (1995), Wellner and Zhang (2000), Lu et al. (2007), and Hu et al. (2009a) considered different nonparametric estimation of the mean function of the underlying recurrent event process. Zhang (2002) and Wellner and Zhang (2007) proposed the semiparametric maximum pseudolikelihood and maximum likelihood estimation procedures. When the panel count data consist of independent samples randomly drawn from $k(k \geq 2)$ populations or groups, one important thing is to handle the treatment comparison. Thall and Lachin (1988), Sun and Fang (2003), Zhang (2006), Park et al. (2007), Balakrishnan and Zhao (2009, 2010, 2011), and Zhao and Sun (2011) presented nonparametric tests for the problem of nonparametric comparison of the mean function of counting process with different groups. Staniswalis et al. (1997), Sun and Wei (2000), Zhang (2002), Wellner and Zhang (2007), He et al. (2003), and Hu et al. (2009b) investigated regression analysis of panel count data. In addition, Zhao et al. (2011) provide a relatively complete discussion for the analysis of panel count data wherein more references can be found.

In many situations, the underlying recurrent process and the observation process are not independent even given covariates, such as example given by a set of panel count data arising from a bladder cancer follow-up study conducted by the Veterans Administration Cooperative Urological Research Group (Byar, 1980; Sun and Wei, 2000; Zhang, 2002). Many patients had multiple recurrences of new tumors during the study. One problem with the data set is that some patients in the study had significantly more clinical visits than others (Sun and Wei, 2000; Hu, et al., 2003; Zhao and Tong, 2011). This indicates that the number of clinical visits may contain some information about the tumor occurrence rate. For this, few references have studied the analysis of panel count data or recurrent event data with informative observation times. Huang et al. (2006) studied nonparametric and semiparametric models that allow the observation times to be correlated with the event process, where the correlation is induced by a frailty variable. Zhao and Tong (2011) proposed a joint modeling approach that used an unobserved frailty variable and a completely unspecified link function to characterize the correlation between the event process and the observation times with time-independent covariates considered in their models. However, in some applications, it would be desirable to develop estimation procedures for panel count data with informative observation times, and also with time-dependent covariates and informative censoring times. For this, we considered the same models for the underlying recurrent events and the observation times as given in Zhao and Tong (2011) except replacing the time-independent covariates with the time-dependent covariates and removing the assumption of noninformative censoring.

The remainder of this paper is organized as follows. We begin in Section 2 by introducing some notation and describing models for the underlying recurrent event process and the observation process. In Section 3, a novel estimation procedure that does not depend on the distribution of frailty variables is proposed for estimation of regression parameters and the consistency and asymptotic normality of the proposed estimators are established. In order to assess the finite-sample properties of the proposed inference procedure, we present some results obtained from simulation studies in Section 4. In Section 5, the proposed approaches are illustrated through the analysis of a data set from a bladder tumor study. Some concluding remarks are given in Section 6.

II. Statistical Models

Consider a recurrent event study that consists of n independent subjects, and let $N_i(t)$ denote the number of occurrences of the recurrent event of interest before or at time t for subjects i. Suppose that for each subject, there exist a p-dimensional possibly time-dependent covariates, denoted by $\mathbf{x}_i(t)$, and z_i is an unobserved positive random variable that is independent of the covariates with $E(z_i) = 1$. Then, for subject i, given $\mathbf{x}_i(t)$ and z_i, the mean function of $N_i(t)$ is assumed to have the form

$$E\{N_i(t)|\mathbf{x}_i(t), z_i\} = \mu_0(t)g(z_i)\exp\{\mathbf{x}_i'(t)\beta_0\}, \qquad (1)$$

where $\mu_0(\cdot)$ is a completely unknown continuous baseline mean function, β_0 is a vector of unknown regression parameters, and $g(\cdot)$ is a completely unspecified function. Since $N_i(t)$ is a counting process, the choice of time-dependent covariates should be constrained by the fact that $E\{N_i(t)|\mathbf{x}_i(t), z_i\}$ is a nondecreasing function of time. Also the covariate histories $\{\mathbf{x}_i(t) : 0 \le t \le C_i\}(i = 1, \cdots, n)$ are assumed to be observed.

For subject i, suppose that $N_i(\cdot)$ is observed only at finite time points $T_{i1} < \cdots < T_{iK_i}$, where K_i denotes the potential number of observation times, $i = 1, \cdots, n$. That is, only the values of $N_i(t)$ at these observation times are known and we have panel count data on the $N_i(t)$'s. Let C_i be the censoring time and thus $N_i(t)$ is observed only at these T_{ij}'s with $T_{ij} \le C_i$, $i = 1, \cdots, n$. Define $\tilde{H}_i(t) = H_i\{\min(t, C_i)\}$, where $H_i(t) = \sum_{j=1}^{K_i} I\{T_{ij} \le t\}, i = 1, \cdots, n$, and $I(\cdot)$ is a indicator function. Then $\tilde{H}_i(\cdot)$ is a point process characterizing the ith subject's observation process and jumps only at the observation times.

In the following, we assume that given $\mathbf{x}_i(t)$ and z_i, $H_i(\cdot)$ is a nonhomogeneous Poisson process with the intensity function

$$\lambda(t\ |\mathbf{x}_i(t), z_i) = \lambda_0(t)z_i\exp\{\mathbf{x}_i'(t)\gamma_0\}, \qquad (2)$$

where $\lambda_0(\cdot)$ is a completely unknown continuous baseline intensity function and γ_0 denotes a vector of regression parameters. Let $\Lambda_0(t) = \int_0^t \lambda_0(s)ds$. In addition, we assume that conditional on the covariates $\mathbf{x}_i(t)$'s and z_i's, N_i's, H_i's and C_i's are mutually independent, and $\{H_i(t), N_i(t), \mathbf{x}_i(t), C_i, z_i, 0 \le t \le \tau\}, i = 1, \cdots, n$, are independent and identically distributed, where τ is the length of study.

The special cases of models (1) and (2) have been studied individually by earlier researchers. For example, model (1) with $g(z_i) = 1$ and time-independent covariates was considered by Sun and Wei (2000), Zhang (2002), and Wellner and Zhang (2007) for regression analysis of panel count data; Huang et al. (2010) considered model (2) with time-dependent and time-independent covariates, and Wang et al. (2001) and Huang and Wang (2004) considered model (2) with time-independent covariates for recurrent event data; Furthermore, Zhao and Tong (2011) developed the joint analysis of the two models with time-independent covariates. In the following, we study the joint analysis of the two models together. The proposed models allow the underlying recurrent event process and the observation process to be correlated through their connections with the link function of the frailty; moreover, both the link function and the distribution of the frailty are considered as nuisance parameters. Our main goal here is to make inference about β. Toward this end, we develop a novel estimation procedure that depends neither on the form of the link function nor on the distribution of the frailty in the next section.

III. Estimation Procedure

For estimation of β_0 along with other parameters, we define $\tilde{N}_i(t) = \int_0^t N_i(s)d\tilde{H}_i(s)$, then this newly defined process only has possible jumps at the observation time points $\{T_{ij} \wedge C_i : j = 1, \cdots, K_i\}$ with respective jump sizes $N_i(T_{ij}), i = 1, \cdots, n$. Thus we have

$$E\{d\tilde{N}_i(t)|\mathbf{x}_i(t), C_i\} = \exp\{\mathbf{x}_i'(t)\theta_0\}\xi_i(t)d\phi_0(t).$$

where $\theta_0 = \beta_0 + \gamma_0$, $\xi_i(t) = I(C_i \ge t)$ and $\phi_0(t) = \int_0^t E[g(z)z]\mu_0(s)d\Lambda_0(s)$.

Similar to Hu et al. (2003), borrowing the structure of the Cox partial likelihood score function of the Andersen-Gill proportional intensity model (Anderson and Gill, 1982), we construct an estimating equation of θ_0 in the form of

$$U(\theta; \tilde{N}) = \sum_{i=1}^{n}\int_0^\tau W(t)\{\mathbf{x}_i(t) - \bar{\mathbf{X}}(t;\theta)\}d\tilde{N}_i(t) = 0$$

where $\bar{\mathbf{X}}(t;\theta) = S^{(1)}(t;\theta)/S^{(0)}(t;\theta)$, and

$$S^{(k)}(t;\theta) = n^{-1}\sum_{i=1}^{n}\xi_i(t)\mathbf{x}_i(t)^{\otimes k}\exp\{\mathbf{x}_i'(t)\theta\}, \quad k = 0, 1, 2,$$

where $a^{\otimes 0} = 1, a^{\otimes 1} = a, a^{\otimes 2} = aa'$ for a vector a .

It can be shown that this estimating equation $U(\theta; \tilde{N}) = 0$ is unbiased for θ (i.e., $E[U(\theta_0; \tilde{N})] = 0$). Solving the estimating equation provides us with an estimator of θ_0, denoted by $\hat{\theta}$, and thus, given γ_0, β_0 can be estimated by $\hat{\theta} - \gamma_0$. But γ_0 is unknown, we need to find an estimator for it.

Since

$$\begin{aligned} E\{dH_i(t)|\mathbf{x}_i(t)\} &= E\{E[dH_i(t)|\ \mathbf{x}_i(t), z_i]|\mathbf{x}_i(t)\} \\ &= \exp\{\mathbf{x}_i'(t)\gamma_0\}d\Lambda_0(t) \end{aligned}$$

and C_i's are independent of (N_i, H_i)'s conditional on covariate and the frailty, as in Liang et al. (2009), the methods

proposed by Lin et al. (2000) for the proportional rate model can be used to consistently estimate γ_0 and $\Lambda_0(\cdot)$. To be specific, γ_0 can be consistently estimated from the following estimating equation

$$U_2(\gamma; \tilde{H}) = \sum_{i=1}^{n} \int_0^{\tau} \{\mathbf{x}_i(t) - \bar{\mathbf{X}}(t;\gamma)\} d\tilde{H}_i(t) = 0,$$

where $\bar{\mathbf{X}}(t;\gamma) = S^{(1)}(t;\gamma)/S^{(0)}(t;\gamma)$, and

$$S^{(k)}(t;\gamma) = n^{-1} \sum_{i=1}^{n} \xi_i(t) \mathbf{x}_i(t)^{\otimes k} \exp\{\mathbf{x}_i'(t)\gamma\}, \quad k = 0,\ 1,\ 2.$$

The resulting estimator is denoted by $\hat{\gamma}$. In addition, $\Lambda_0(t)$ can be consistently estimated by the Aalen-Breslow-type estimator $\hat{\Lambda}_0(t) = \hat{\Lambda}_0(t;\hat{\gamma})$, where $\hat{\Lambda}_0(t;\gamma) = \sum_{i=1}^{n} \int_0^t \frac{d\tilde{H}_i(s)}{nS^{(0)}(s\ ;\gamma)}$.

Let

$$s^{(k)}(t;\mu) = \lim_{n\to\infty} S^{(k)}(t;\mu) = E[\xi_1(t) \exp\{\mathbf{x}_1'(t)\mu\} \mathbf{x}_1(t)^{\otimes k}],$$

for $k = 0,\ 1,\ 2$, and define $\bar{\mathbf{x}}(t;\mu) = s_1^{(1)}(t;\mu)/s_1^{(0)}(t;\mu)$.

To establish the asymptotic properties of $\hat{\theta}$, we need the following regularity conditions.

(C.1) $P(C \geq \tau) > 0$.

(C.2) $\mathbf{x}_i(t)$ have bounded total variations, i.e. $|x_{ji}(0)| + \int_0^{\tau} |x_{ji}(t)| \leq M_0$ for all $j = 1, \cdots, p$ and $i = 1, \cdots, n$, where x_{ji} is the jth component of $\mathbf{x}_i$ and M_0 is a constant.

(C.3) $\Lambda_0(\tau) \leq M_1, \mu_0(\tau) \leq M_2$, where M_1, M_2 are constants.

(C.4) $N_i(\tau)$ $(i = 1, \cdots, n)$ are bounded by a constant and the K_i's are bounded; $W(\cdot)$ is nonnegative and have bounded total variations with $W(\cdot) \to w(\cdot)$, as $n \to \infty$.

(C.5)

$$A_\theta(\theta_0) \equiv E \int_0^{\tau} w(t) \{\mathbf{x}_1(t) - \bar{\mathbf{x}}(t;\theta_0)\}^{\otimes 2} \xi_1(t) e^{\{\mathbf{x}_1'(t)\theta_0\}} d\phi_0(t),$$

and

$$A_\gamma(\gamma_0) \equiv E \int_0^{\tau} \{\mathbf{x}_1(t) - \bar{\mathbf{x}}(t;\gamma_0)\}^{\otimes 2} \xi_1(t) \exp\{\mathbf{x}_1'(t)\gamma_0\} d\Lambda_0(t)$$

are positive definite.

In practice, condition (C.1) can be enforced simple by not choosing τ to be greater than the maximum observation time. The boundedness conditions in (C.2), (C.3) and (C.4) simplify the derivation of the asymptotic results. Condition (C.5) can be interpreted that the sample covariance is asymptotically non-singular. The asymptotic properties are summarized as follows.

Theorem 3.1 (Consistency of $\hat{\theta}$). *Under conditions (C.1 – C.5), $\hat{\theta} \longrightarrow \theta_0, a.s.$*

Since $\hat{\gamma}$ is consistent as in Lin et al. (2000), then $\hat{\beta} = \hat{\theta} - \hat{\gamma}$ is a consistent estimator of β_0.

A consistent Aalen-Breslow-type estimator for $\phi_0(t)$ can be obtained as $\hat{\phi}_0(t) = \hat{\phi}_0(t;\hat{\theta}) = \int_0^t \frac{\sum_{i=1}^n d\tilde{N}_i(s)}{nS^{(0)}(s;\hat{\theta})}, \quad t \in [0,\tau]$.

To establish the asymptotic normality of $\hat{\beta}$, define

$$\hat{\tilde{M}}_i(t;\hat{\theta}) = \tilde{N}_i(t) - \int_0^t \xi_i(s) \exp\{\mathbf{x}_i'(s)\hat{\theta}\} d\hat{\phi}_0(s),$$

$$\hat{M}_i(t;\hat{\gamma}) = \tilde{H}_i(t) - \int_0^t \xi_i(s) \exp\{\mathbf{x}_i'(s)\hat{\gamma}\} d\hat{\Lambda}_0(s),$$

$$\hat{A}_\theta(\hat{\theta}) = n^{-1} \sum_{i=1}^{n} \int_0^{\tau} W(t) \xi_i(t) e^{\{\mathbf{x}_i'(t)\hat{\theta}\}} [\mathbf{x}_i(t) - \bar{\mathbf{X}}(t;\hat{\theta})]^{\otimes 2} d\hat{\phi}_0(t),$$

$$\hat{A}_\gamma(\hat{\gamma}) = n^{-1} \sum_{i=1}^{n} \int_0^{\tau} \{\mathbf{x}_i(t) - \bar{\mathbf{X}}(t;\hat{\gamma})\}^{\otimes 2} \xi_i(t) e^{\{\mathbf{x}_i'(t)\hat{\gamma}\}} d\hat{\Lambda}_0(t),$$

$$\hat{a}_i = \hat{A}_\theta(\hat{\theta})^{-1} \int_0^{\tau} W(t) [\mathbf{x}_i(t) - \bar{\mathbf{X}}(t;\hat{\theta})] d\hat{\tilde{M}}_i(t;\hat{\theta}),$$

$$\hat{b}_i = \hat{A}_\gamma(\hat{\gamma})^{-1} \int_0^{\tau} [\mathbf{x}_i(t) - \bar{\mathbf{X}}(t;\hat{\gamma})] d\hat{M}_i(t;\hat{\gamma}),$$

and $\hat{c}_i = \hat{a}_i - \hat{b}_i$.

Theorem 3.2 (Asymptotic normality of $\hat{\beta}$). *Under conditions (C.1 – C.5), $n^{1/2}(\hat{\beta} - \beta_0)$ is asymptotically zero-mean normal, with covariance matrix $\Sigma_\beta = E[c_1^{\otimes 2}]$, which can be consistently estimated by $\hat{\Sigma}_\beta = n^{-1} \sum_{i=1}^n \hat{c}_i^{\otimes 2}$, where c_1 is given in the proof of this theorem.*

Proof of Theorem 3.1.

By the strong law of large numbers, for each $t \in [0,\tau]$, $S^{(k)}(t;\theta)$ converges almost surely to $s^{(k)}(t;\theta)$, for every θ, $k = 0,\ 1,\ 2$. Define

$$Y_n(\theta) \equiv \frac{1}{n} \sum_{i=1}^{n} \int_0^{\tau} W(t) \left[(\theta - \theta_0)'\ \mathbf{x}_i(t) - \log\{\frac{S^{(0)}(t;\theta)}{S^{(0)}(t;\theta_0)}\} \right] d\tilde{N}_i(t)$$

and

$$\mathcal{Y}(\theta) \equiv E \int_0^{\tau} w(t) \left[(\theta - \theta_0)'\ \mathbf{x}_1(t) - \log\{\frac{s^{(0)}(t;\theta)}{s^{(0)}(t;\theta_0)}\} \right] d\tilde{N}_1(t).$$

We can see that $Y_n(\theta)$ converges almost surely to $\mathcal{Y}(\theta)$, for every θ and

$$\partial Y_n(\theta)/\partial\theta = n^{-1} U(\theta; \tilde{N}).$$

Clearly, $\partial^2 Y_n(\theta)/\partial\theta\partial\theta'$ is negative semidefinite. Thus, $Y_n(\theta)$ is concave, which implies that the convergence of $Y_n(\theta)$ to $\mathcal{Y}(\theta)$ is uniform on any compact set of θ (Rockafellar (1970), Th10.8). In particular, letting $\mathcal{A}_\epsilon(\theta_0) = \{\theta :\| \theta - \theta_0 \| \leq \varepsilon\}$, we have

$$\sup_{\theta \in\ \mathcal{A}_\epsilon(\theta_0)}\ \| Y_n(\theta) - \mathcal{Y}(\theta) \| \longrightarrow 0 \tag{3}$$

almost surely. It is easy to show that $\partial\mathcal{Y}(\theta_0)/\partial\theta = 0$ and $\partial^2\mathcal{Y}(\theta_0)/\partial\theta\partial\theta' = -A_\theta(\theta_0)$, where $A_\theta(\theta_0)$ is positive definite by (C.5). Thus $\mathcal{Y}(\theta)$ has a unique maximizer θ_0.

In particular, $\sup_{\theta \in \partial\mathcal{A}_\epsilon(\theta_0)} \mathcal{Y}(\theta) < \mathcal{Y}(\theta_0)$, where $\partial\mathcal{A}_\epsilon(\theta_0) = \{\theta :\| \theta - \theta_0 \| = \varepsilon\}$. This fact, together with (3) implies that $Y_n(\theta) < Y_n(\theta_0)$ for all $\theta \in \partial\mathcal{A}_\epsilon(\theta_0)$ and all large n. Therefore, there must be a maximizer of

$Y_n(\theta)$, i.e., a solution to $\partial Y_n(\theta)/\partial\theta = 0$, say $\hat{\theta}$, in the interior of $\mathcal{A}_\epsilon(\theta_0)$.

On the other hand, similar to the proof of consistency of their estimators in Hu et al. (2003), we can verify that $\partial^2 Y_n(\hat{\theta})/\partial\theta\partial\theta'$ is negative definitive, which implies that $\hat{\theta}$ is the unique global maximizer of $Y_n(\theta)$ in $\mathcal{A}_\epsilon(\theta_0)$, i.e., the unique solution to $U(\theta; \tilde{N}) = 0$.

Finally, since ϵ can be chosen arbitrarily small, $\hat{\theta}$ must converge to θ_0 almost surely, as $n \longrightarrow \infty$.

Proof of Theorem 3.2.

Notice that

$$U(\theta; \tilde{N}) = \sum_{i=1}^{n} \int_0^\tau W(t)\{\mathbf{x}_i(t) - \bar{\mathbf{X}}(t;\theta)\} d\tilde{M}_i(t;\theta),$$

where $\tilde{M}_i(t;\theta) = \tilde{N}_i(t) - \int_0^t \xi_i(s)\exp(\mathbf{x}_i'(t)\theta) d\phi_0(s)$.

Similar to the arguments of Lin and Wei (1989), we can show that

$$\begin{aligned} & n^{-1/2} U(\theta_0; \tilde{N}) \\ &= n^{-1/2} \sum_{i=1}^{n} \int_0^\tau W(t)\{\mathbf{x}_i(t) - \bar{\mathbf{x}}(t;\theta_0)\} d\tilde{M}_i(t;\theta_0) + o_p(1). \end{aligned}$$

By the Taylor expansion,

$$\begin{aligned} & n^{1/2}(\hat{\theta} - \theta_0) \\ &= \left[-n^{-1} \partial U(\theta; \tilde{N})/\partial\theta \Big|_{\theta=\theta_0} \right]^{-1} \left[n^{-1/2} U(\theta_0; \tilde{N}) \right] + o_p(1) \\ &\equiv n^{-1/2} \sum_{i=1}^{n} a_i + o_p(1), \end{aligned}$$

with $a_i = A_\theta(\theta_0)^{-1} \int_0^\tau W(t)\{\mathbf{x}_i(t) - \bar{\mathbf{x}}(t;\theta_0)\} d\tilde{M}_i(t;\theta_0)$.

By (A.5) of Lin et al. (2000),

$$\begin{aligned} & n^{1/2}(\hat{\gamma} - \gamma_0) \\ &= A_\gamma(\gamma_0)^{-1} n^{-\frac{1}{2}} \sum_{i=1}^{n} \int_0^\tau \{\mathbf{x}_i(t) - \bar{\mathbf{x}}(t;\gamma_0)\} dM_i(t;\gamma_0) + o_p(1) \\ &\equiv n^{-1/2} \sum_{i=1}^{n} b_i + o_p(1), \end{aligned}$$

where $M_i(t;\gamma) = \tilde{H}_i(t) - \int_0^t \xi_i(t)\exp\{\mathbf{x}_i'(t)\gamma\} d\Lambda_0(t)$.

Thus, $n^{1/2}(\hat{\beta} - \beta_0) = n^{-1/2}\sum_{i=1}^n c_i + o_p(1)$, where $c_i = a_i - b_i, i = 1, \cdots, n$. Then, by the multivariate central limit theorem, we conclude that $n^{1/2}(\hat{\beta} - \beta_0)$ is asymptotically zero-mean normal with covariance matrix $\Sigma_\beta = E[c_1^{\otimes 2}]$.

Next, similar to the argument in the A.3 of Lin et al. (2000), we can verify that Σ_β can be consistently estimated by $\hat{\Sigma}_\beta$ as defined in Theorem 3.2.

IV. Simulation Study

We conducted Monte Carlo simulation studies to evaluate the finite-sample properties of the proposed estimators. To generate the simulated data, we first generated z_i from the gamma distribution with mean 1 and variance σ^2, and let $g(z_i) = z_i^\alpha$. We assume that the time-dependent covariate $x_i(t)$ takes the form $u_i \log(t)$, where u_i has a uniform distribution over $[0, 0.5]$, and the follow-up times C_i's were generated from the uniform distribution over $(\tau/2, \tau)$ with $\tau = 18$. Here the symbol of α characterizes the relationship between the observation process and the recurrent event process. When $\alpha > 0$, a subject with more frequent observations would have a higher occurrence rate of the recurrent event and the two processes are positively correlated; when $\alpha = 0$, the two processes have no correlation given the covariates; when $\alpha < 0$, a subject with more frequent observations would have a lower occurrence rate of the recurrent event and the two processes are negatively correlated.

For observation process, we assume that H_i is a homogeneous Poisson process with $\lambda_0(t) = 1$. Then, given $x_i, C_i,\ z_i,\ K_i^* = \xi_i(C_i)H_i(C_i)$, the total number of real observation times for subjects i, follows the Poisson distribution with mean

$$\Lambda_0(C_i \mid x_i, z_i) = \int_0^{C_i} z_i \exp\{x_i(t)\gamma_0\}\lambda_0(t)dt = z_i \frac{C_i^{u_i\gamma_0+1}}{u_i\gamma_0 + 1}.$$

In this case, the observation times $(T_{i1}, \cdots, T_{iK_i^*})$ are the order statistics of a random sample of size K_i^* from the uniform distribution over $(0, C_i)$. Finally, given K_i^* and $(T_{i1}, \cdots, T_{iK_i^*})$, we generate $N_i(T_{ij})'s$ by taking $N_i(T_{ij}) = N_i(T_{i1}) + N_i(T_{i2}) - N_i(T_{i1}) + \cdots + N_i(T_{ij}) - N_i(T_{ij-1})$, where $N_i(t) - N_i(s) \sim \text{Poisson}(0.5t^2 g(z_i)\exp\{x_i(t)\beta_0\} - 0.5s^2 g(z_i)\exp\{x_i(s)\beta_0\})$, for $j = 1, \cdots, K_i^*, i = 1, \cdots, n$.

Set $\gamma_0 = 1$ and $\beta_0 = -1,\ 0,\ 1$, representing the different effect of the covariate $x(t)$ on the panel counts. On one hand, in order to check the effect of the estimators with time-independent covariates, we performed Monte Carlo studies when the time-independent covariate x_i follows a Bernoulli distribution with success probability 0.5. On the other hand, we also considered the situation that the observation process H_i follows a nonhomogeneous Poisson process with $\lambda_0(t) = (t+1)/(\tau/2+1)$ to verify that whether the different forms of the observation process H_i will affect the estimation of β or not. For each setting, we consider the sample size $n = 100$. All the results reported here are based on 500 Monte Carlo replications using R software.

Tables 1 presents the simulation results on estimation of β with time-independent and time-dependent covariates respectively under the homogeneous poisson observation process with $n = 100$, while Table 2 presents those under the nonhomogeneous poisson observation process. The tables include the bias (Bias) given by the sample means of the point estimates $\hat{\beta}$ minus the true values, the sample standard deviations of the estimates (SSD), the means of the estimated standard deviations (ESD), and the empirical 95% coverage probabilities (CP) for β. These results indicate that the estimate $\hat{\beta}$ seems to be unbiased and the proposed variance estimation procedure provides reasonable estimates. Also the results on the empirical coverage probabilities indicate that the normal approximation seems to be appropriate.

In addition, one can see from Tables 1 and 2 that the biases of the estimators of β, the SSD and ESD of the estimators

of β with time-independent covariates are smaller than those with time-dependent covariates, which means that estimators with time-independent covariates are more precise and more stable than those with time-dependent covariates since there are more nondeterminacy with the time-varying covariates. Furthermore, one can see that the effect of the estimators with time-dependent covariates worsens rapidly as the variance of the frailty increases as discussed in Lin et al. (2000).

Table 3 shows the results of the estimators of β under the homogeneous and nonhomogeneous poisson observation process respectively with $n = 200$ and time-independent covariates. Compared with the corresponding results in Tables 1 and 2, we can see that the SSD and ESD of the estimators decreases when the sample size increases. As shown in Tables 1 and 2, the variance seems underestimated; a possible reason is that the simulated data were generated from the joint model including random effects, and the estimating equation only involves the means of random effects. The results in Table 3 indicate that this does not seem to be a problem for large sample size.

Table 1: Estimation of β with time-independent and time-dependent covariates respectively and $n = 100$ under the homogeneous poisson observation process

$\alpha = -0.5$: H and N are negatively correlated						
β_0	1	0	-1	1	0	-1
	Time-indep covariates			Time-dep covariates		
Bias$_{\hat{\beta}}$	0.0022	-0.0040	0.0021	-0.0339	-0.0222	-0.0293
SSD	0.0746	0.0775	0.1031	0.1226	0.1079	0.1262
ESD	0.0738	0.0744	0.0963	0.1148	0.1044	0.1247
CP	0.9380	0.9300	0.9260	0.9100	0.9440	0.9480
$\alpha = 0$: H and N have no correlation						
β_0	1	0	-1	1	0	-1
	Time-indep covariates			Time-dep covariates		
Bias$_{\hat{\beta}}$	0.0002	0.0048	0.0016	-0.0292	-0.0239	-0.0207
SSD	0.0614	0.0665	0.0714	0.1117	0.0843	0.1103
ESD	0.0605	0.0622	0.0668	0.1008	0.0801	0.1012
CP	0.9380	0.9160	0.9220	0.9000	0.9280	0.9360
$\alpha = 0.5$: H and N are positively correlated						
β_0	1	0	-1	1	0	-1
	Time-indep covariates			Time-dep covariates		
Bias$_{\hat{\beta}}$	-0.0039	0.0035	-0.0051	-0.0249	-0.0321	- 0.0239
SSD	0.1006	0.0991	0.0800	0.1791	0.1354	0.1473
ESD	0.0927	0.0932	0.0793	0.1649	0.1234	0.1379
CP	0.9280	0.9280	0.9340	0.9160	0.9100	0.9360

V. An Application

This section presents an analysis of the bladder cancer data by applying our proposed methods. There were 121 subjects with superficial bladder tumors randomized into one of three treatment groups: placebo, thiotepa, and pyridoxine. In the following, we restrict our attention to the placebo and thiotepa groups with respective sizes of 47 and 38. For each patient, the observed information includes times when he or she made clinical visits and the numbers of recurrent tumors between clinical visits. Two baseline covariates were observed and they are the number of initial tumors and the size of the largest initial tumor.

To analysis the data, for patient i, define x_{i1} to be equal to 1 if the ith patient was given the thiotepa treatment and 0 otherwise, x_{i2} to be the number of initial tumors and x_{i3} to be the size of the largest initial tumor, $i = 1, \cdots, 85$. Assume that the occurrence process of the bladder tumors and the clinical visit process can be described by joint models (1) and (2). Let $N_i(\cdot)$ represent the accumulated new tumor numbers of patient i over study period. We took the last visit time of the subject to approximate C_i in the analysis.

The application of the estimation procedure proposed in the previous sections gave $\hat{\gamma}_1 = 0.5071, \hat{\gamma}_2 = -0.0049, \hat{\gamma}_3 = 0.0321, \hat{\beta}_1 = -1.4905, \hat{\beta}_2 = 0.2867, \hat{\beta}_3 = -0.0821$ with the estimated standard errors being 0.1175, 0.0343, 0.0359, 0.3287, 0.0615 and 0.1056, which correspond to p-values of 1.5905e-05, 0.8864, 0.3712, 5.7732e-06, 3.1347e-06 and 0.4369, respectively based on the asymptotic results of the estimators. Here γ_1 and β_1, γ_2 and β_2, and γ_3 and β_3 represent regression coefficients corresponding to the treatment indicator, the number of initial tumors, and the size of the largest initial tumor, respectively. These results indicate that

Table 2: Estimation of β with time-independent and time-dependent covariates respectively and $n = 100$ under the nonhomogeneous poisson observation process

$\alpha = -0.5$: H and N are negatively correlated						
β_0	1	0	-1	1	0	-1
	Time-indep covariates			Time-dep covariates		
Bias$_{\hat{\beta}}$	0.0036	0.0004	0.0062	-0.0305	-0.0411	-0.0262
SSD	0.0811	0.0798	0.0844	0.1232	0.1253	0.1395
ESD	0.0777	0.0786	0.0833	0.1202	0.1176	0.1324
CP	0.9500	0.9260	0.9240	0.9240	0.9200	0.9320
$\alpha = 0$: H and N have no correlation						
β_0	1	0	-1	1	0	-1
	time-indep covariates			time-dep covariates		
Bias$_{\hat{\beta}}$	0.0060	0.0025	0.0049	-0.0324	-0.0337	-0.0315
SSD	0.0669	0.0714	0.0746	0.1138	0.0905	0.1113
ESD	0.0638	0.0657	0.0706	0.1041	0.0814	0.1065
CP	0.9280	0.9240	0.9280	0.9160	0.9080	0.9120
$\alpha = 0.5$: H and N are positively correlated						
β_0	1	0	-1	1	0	-1
	Time-indep covariates			Time-dep covariates		
Bias$_{\hat{\beta}}$	0.0072	-0.0021	-0.0047	-0.0349	-0.0330	- 0.0290
SSD	0.1011	0.0995	0.1161	0.1847	0.1452	0.1611
ESD	0.0974	0.0953	0.1022	0.1642	0.1281	0.1412
CP	0.9380	0.9280	0.9160	0.8740	0.8900	0.9220

Table 3: Estimation of β under the homogeneous and nonhomogeneous poisson observation process respectively with n=200 and time-independent covariates

$\alpha = -0.5$: H and N are negatively correlated						
β_0	1	0	-1	1	0	-1
	Homogeneous			Nonhomogeneous		
Bias$_{\hat{\beta}}$	0.0001	0.00391	0.0042	0.0080	0.0048	0.0028
SSD	0.0553	0.0527	0.0558	0.0555	0.0564	0.0585
ESD	0.0525	0.0534	0.0562	0.0554	0.0562	0.0591
CP	0.9460	0.9440	0.9420	0.9420	0.9500	0.9540
$\alpha = 0$: H and N have no correlation						
β_0	1	0	-1	1	0	-1
	Homogeneous			Nonhomogeneous		
Bias$_{\hat{\beta}}$	0.0009	-0.0004	0.0002	0.0043	0.0008	-0.0005
SSD	0.0449	0.0471	0.0502	0.0468	0.0483	0.0507
ESD	0.0436	0.0451	0.0486	0.0459	0.0472	0.0513
CP	0.9440	0.9360	0.9420	0.9380	0.9440	0.9440
$\alpha = 0.5$: H and N are positively correlated						
β_0	1	0	-1	1	0	-1
	Homogeneous			Nonhomogeneous		
Bias$_{\hat{\beta}}$	-0.0023	0.0078	-0.0007	0.0057	0.0049	0.0056
SSD	0.0703	0.0766	0.0755	0.0747	0.0744	0.0775
ESD	0.0667	0.0684	0.0711	0.0722	0.0722	0.0749
CP	0.9420	0.9120	0.9380	0.9320	0.9440	0.9420

the thiotepa treatment significantly reduces the occurrence rate of the bladder tumors and the number of initial tumors has a significant positive effect on the tumor recurrence rate but no significant effect on the visit process. However, both the occurrence rate of the bladder tumors and the visit times do not seem to be significantly related to the size of the largest initial tumor. These conclusions are consistent with the analysis results presented in Sun and Wei (2000), Hu et al. (2003) and Zhao and Tong (2011). Furthermore, one can see that our proposed approach yields the smallest standard deviations except that the standard deviation of $\hat{\beta}_3$ is slightly higher than that of Zhao and Tong (2011), which suggests that our approach works well in applications.

VI. Conclusion

In this article, we have generalized Zhao and Tong (2011)'s joint modeling approach for the analysis of panel count data to the situations where the covariates are time-dependent and the observation and censoring times are informative. For estimation of the covariate effect on the underlying recurrent process, we have developed a novel estimating equation-based procedure, which depends on neither the form of the link function of the frailty nor the distribution of the frailty, and established the consistency and asymptotic normality of the resulting estimates.

By using the approach proposed by Huang et al. (2010), one can obtain the estimators of the parameter γ and $\Lambda_0(\cdot)$ in model (2), which are different from the approach proposed in Lin et al. (2000). Then, by replacing $\hat{\gamma}$ and $\hat{\lambda}_0(\cdot)$ with those given in Huang et al. (2010), one can get another estimator for β_0, which is different from our proposed estimator. Thus, it is desirable to compute the efficiency of these two different estimators.

In practice, it is important to predict the mean of panel counts. However, it is hard to estimate the baseline mean function $\mu_0(t)$ in the current setting. Further research is needed to address this issue.

Just as Zhao and Tong (2011) mentioned, the time-dependent frailty, the non-poisson observation process are also important issues to be studied.

Acknowledgment

The author would like to thank Professor Liu Li in Wuhan University for the computing support.

References

[1] Andersen, P. K. and Gill, R. D. (1982). Cox's regression model for counting processes: a large sample study. *Ann. Statist.* **10:** 1100 - 1120.

[2] Balakrishnan, N. and Zhao, X. (2009). New multi-sample nonparametric tests for panel count data. *Ann. Statist.* **37:** 1112 - 1149.

[3] Balakrishnan, N. and Zhao, X. (2010). A nonparametric test for the equality of counting processes with panel count data. *Comput. Statist. Data. Anal.* **54:** 135 - 142.

[4] Balakrishnan, N. and Zhao, X. (2011). A class of multi-sample nonparametric tests for panel count data. *Ann. Inst. Statist. Math.* **63:** 135-156.

[5] Byar, D. P. (1980). The veterans administration study of chemoprophylaxis for recurrent stage I bladder tumors: comparisons of placebo, pyridoxine and topical thiotepa. In: Pavane-Macaluso, M., Smith, P. H., Edsmyr, F. (Eds), Bladder Tumors and Other Topics in Urological Oncology. Plenum, New York, pp. 363 - 370.

[6] Gaver, D. P. and O'Muircheartaigh, I. G. (1987). Robust empirical Bayes analysis of event rates. *Technometics* **29:** 1 - 15.

[7] He, X., Tong, X., Sun, J. and Cook, R. (2008). Regression analysis of multivariate panel Count data. *Biostatistics* **9:** 234-248.

[8] Hu, X. J., Lagakos, S. W. and Lockhart, R. A. (2009a). Generalized least squares estimation of the mean function of a counting process based on panel counts. *Statist. Sinica* **19:** 561- 580.

[9] Hu, X. J., Lagakos, S. W. and Lockhart, R. A. (2009b). Marginal analysis of panel counts through estimating functions . *Biometrika* **96:** 445 - 456.

[10] Hu, X. J., Sun, J. and Wei, L. J. (2003). Regression parameter estimation from panel counts . *Scand. J. Statist.* **30** : 25 - 43.

[11] Huang, C. Y. and Wang, M. C. (2004). Joint modeling and estimation for recurrent event processes and failure time data. *J . Am. Statist. Assoc.* **47:** 1153 - 1165.

[12] Huang, C. Y., Wang, M. C. and Zhang, Y. (2006). Analysing panel count data with informative observation times. *Biometrika* **93:** 763-775.

[13] Huang, C. Y., Qin, J. and Wang, M. C. (2010). Semiparametric analysis for recurrent event data with time-dependent covariates and informative censoring. *Biometrics* **66:** 39 - 49.

[14] Kalbfleisch, J. D. and Lawless, J. F. (1985). The analysis of panel count data under a Markov assumption. *J . Am. Statist. Assoc.* **80:** 863 - 871.

[15] Lawless, J. F. and Nadeau, J. C. (1995). Some simple robust methods for the analysis of recurrent events. *Technometrics* **37**: 158 - 168.

[16] Liang, Y., Lu, W. and Ying, Z. L. (2009). Joint modeling and analysis of longitudinal data with informative observation times. *Biometrics* **65:** 377 - 384.

[17] Lin, D. Y. and Wei, L. J. (1989). The robust inference for the cox proportional hazards model. *J . Am. Stat. Assoc.*, Vol. **84,** No. **408**, pp. 1074 - 1078.

[18] Lin, D. Y., Wei, L. J., Yang, I. and Ying, Z. (2000). Semiparametric regression for the mean and rate functions of recurrent events. *J. R. Statist. Soc. B* **62:** 711-730.

[19] Lu, M., Zhang, Y. and Huang, J. (2007). Estimation of the mean function with panel count data usinmg monotone polynomial splines. *Biometrika* **94:** 705 - 718.

[20] Park, D. H., Sun, J. and Zhao, X. (2007). A class of two-sample nonparametric tests for panel count Data. *Commun. Stat. Theor. M.* **36:** 1611-1625.

[21] Pollard, D. (1990). *Empirical Processes: Theory and Applications*. Hayward: Institute of Mathematical Statistics.

[22] Rockaffellar, R. T. (1970). *Convex Analysis*. Princeton University Press.

[23] Staniswalis, J. G., Thall, P. F. and Salch, J. (1997). Semiaprametric regression analysis for recurrent event interval counts. *Biometrics* **53:** 1334 - 1353.

[24] Sun, J. and Kalbleisch, J. D. (1995). Estimation of the mean function of point processes based on panel count data. *Statist. Sinica* **5:** 279-290.

[25] Sun, J. and Wei, L. J. (2000). Regression analysis of panel count data with covariate-dependent observation and censoring times. *J. R. Statist. Soc. B* **62:** 293-302.

[26] Sun, J. and Fang, H. B. (2003). A nonparametric test for panel count data. *Biometrika* **90:** 199 - 208.

[27] Thall, P. F. and Lachin J. M. (1988). Analysis of recurrent events: nonparametric methods for random-interval count data. *J. Am. Statist. Assoc.* **83:** 339 - 347.

[28] Wang, M. C., Qin, J. and Chiang, C. T. (2001). Analyzing recurrent event data with informative censoring. *J. Am. Statist. Assoc.* **96:** 1057 - 1065.

[29] Wellner, J. A. and Zhang, Y. (2000). Two estimators of the mean of a counting process with panel count data. *Ann. Statist.* **28:** 779 - 814.

[30] Wellner, J. A. and Zhang, Y. (2007). Two likelihood based semiparametric estimation methods for panel count data with covariates. *Ann. Statist.* **35:** 2106- 2142.

[31] Zhang, Y. (2002). A semipartametric pseudolikehood estimation method for panel count data. *Biometrika* **89:** 39-48.

[32] Zhang, Y. (2006). Nonparametric k-sample tests with panel count data. *Biometrika* **93:** 777-790.

[33] Zhao, X., Balakrishnan, N. and Sun, J. (2011). Nonparametric inference based on panel count data. *Test* **20:** 1-42.

[34] Zhao, X. and Sun, J. (2011). Nonparametric comparison for panel count data with unequal observation processes. *Biometrics*, in press.

[35] Zhao, X., and Tong, X. (2011). Semiparametric regression analysis of panel count data with informative observation times. *Comput. Statist. Data. Anal.* **55:** 291 - 300.

RELSYS® methodology for calculating the reliability parameters by the scientific calculation

Jérôme de Reffye, *PI-RAMSES Consulting, 10, rue Carnot, 78000 Versailles*
e-mail : dereffye.pi-ramses@sfr.

Abstract : A model of feasibility of the RELSYS methodology is developed. Its main goal is to prove that this way for solving all calculations of the reliability parameters is the most complete it is possible to realize. We obtain a synthesis of previous methods and the problem of the dynamical reliability simulation is solved. An effort will be made about the quality of physical data and it remains to develop the definitive software.

Keywords : dynamic reliability scientific calculation electronic-mechanical systems

1 Introduction and position of the problem

For any years the author is working on an unification of the different methods existing to calculate the reliability of components and systems. These methods can be so separated between these ones dedicated to the estimation of the reliability of components and these ones dedicated to the reliability of systems. The first ones can be divided on the methods based on a calculation and the other ones based on the data bases or the results of trials while the second ones can be divided on the methods based on experiment data from the Return of Experiment and the methods based on trails and comparison with norms. A last subdivision remains between the electronic components and the mechanical ones. Without to forget the problem put by the reliability of software based systems.
Consequently the existing software is equally divided with the same separation. The goal of this study is to propose an unified approach with the same methodology for every component and a synthesis between the components and the systems. Consequently a software methodology will be developed for the reliability of every system. The mathematical calculation of the reliability of components and systems is only related to the mechanics. It is based on the strength-stress method. The remaining of all methods is based on the use of experiment data. By this approach it is impossible to obtain time-dependent parameters. To make dynamical reliability and mathematical simulation with real working conditions it is necessary to use mathematical modeling. This methodology exposed below is the synthesis between :
The mechanical and electronic components whom the reliability is calculated with stress-strength method,
The components and the system linked through the fault trees fed by the reliability of the components and feeding the Bayesian net of the system
The merge of the theoretical results previously obtained and data of trials and REX by data fusion methods is the last operation to finish the RELSYS® methodology.

2 Description of the methodology RELSYS®

This proposed methodology is mainly based on the mathematical formulation of every problem of reliability:

- Equations of mechanics for the mechanical components
- Equations of electronics for the electronic components

These equations are used into a simplified version. The reason why we so proceed is the precision of calculation what can be less than this one obtained in the numerical simulation of the partial differential equations of physics. Every time it is important to upgrade the precision of the calculation the software can be coupled with a FEM method.

A other reason is the huge number of calculation for all components of a real system.

The main goal is to obtain mainly overestimating values for safety reasons.

The different phases of this methodology are the following ones:

- Decomposition of the system on its components

- Decomposition of each component with its different materials and its different working modes
- FMEA of each component with regards to its materials and functional analysis
- Calculation of the failure probability of each mode of each component by the strength-stress method
- Calculation of the probability of function loss via the fault tree
- Calculation of the probability of the state realization of the system
- Calculation of the transition probabilities and fill the Bayesian net of the system
- Introduction of the experiment data by a fusion method or a Bayesian procedure

The mathematical framework is the finite energy functions Hilbert Space for the deterministic ones and the second order random functions Hilbert Space for the random ones.
As the theoretical explanation could find too much time and too much paper sheets one prefers demonstrate the methodology with a didactic example. More explanations are available in [1], [2], [3]

3 A canonical example

We want to determine the reliability parameters of a hoist with a regulation in speed while the descent or the climb of the cable.

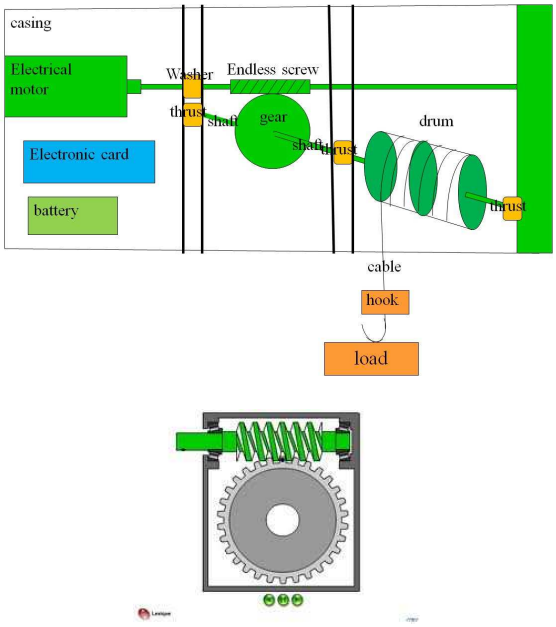


Figure 1: Synoptic Scheme of an hoist

We take on account a simple version of a real hoist to facilitate the calculation and the understood of the methodology.

4 Reliability of the electronic components

4.1 Modeling the engine and the control-command system. Failure modes and failure probabilities of the electronic card (Component 1)

The whole circuit is compound with the circuit (modeling the engine) and the loop of the control using a tachometer generator for the speed measure as shown in the figure 2 [1]:

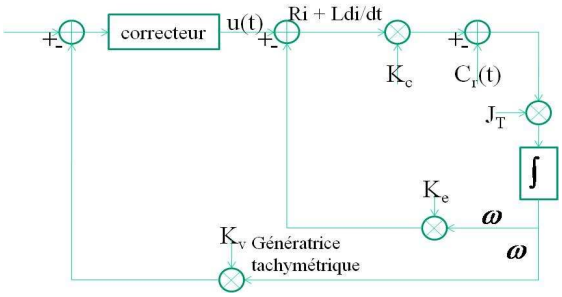


Figure 2: scheme of the work of a speed control-command system of an electrical engine

Where the equation of the proportional corrector is :

$$u(t) = K_v(\omega_0 - \omega(t)) \quad (1)$$

And the instantaneous current is given by a differential equation excited by a random function:

$$(K_v K_c + \frac{K_e}{J_T} K_c) i(t) + R\frac{di}{dt} + L\frac{d^2 i}{dt^2} = (K_v + \frac{K_e}{J_T}) C_r(t) \quad (2)$$

This equation is a linear second order differential equation with the excitation defined by the opposite torque of the load which is a time random function. This equation is very important to calculate the reliability of the semi-conductor because it links the local conditions (the current in the semi-conductor) to the external conditions (the load suffered by the electronic circuit).

Mode 1 : Over intensity or over temperature of the power semi-conductor

$$proba_f(t) = proba(T(t) > T_{\max} \cup i(t) > i_{\max}) \quad (3)$$

Mode 2 : Excess of vibrations on the electronic card:

$$P_f(t) = proba(\sup(\sigma_{xx}, \sigma_{yy} \geq \sigma_D) \quad (4)$$

Mode 3 : Excess of thermo-mechanical stresses:

$$P_f(t) = proba(E\alpha\Delta T \geq \sigma_D) \quad (5)$$

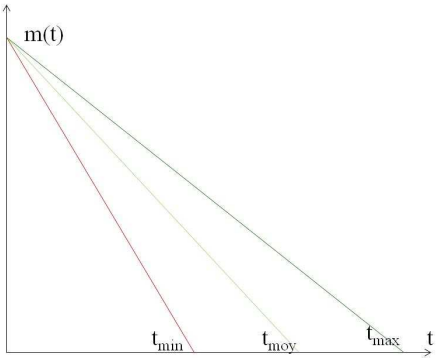
4.2 Modeling the reliability of the brushes of the Cc motor (component 2)

The single failure mode took on account. We suppose an uniform distribution surrounding the mean value of the wearing and we approximate it by a Weibull's law [4].
Tearing law of the brushsses Archard's law :

$$W(lost\ of\ material) \cong K.s.\frac{P_N}{p_m}$$

The life duration is quit equal to 1000 hours. The stress-strength relationship is degenerated . While the damage of brushes keeps material nothing appears. But at the end of the rubbing material the failure appears immediately.

We put a uncertainty = 20 % and the following law of failure :



We have to define the m(t) stochastic process from m(0)

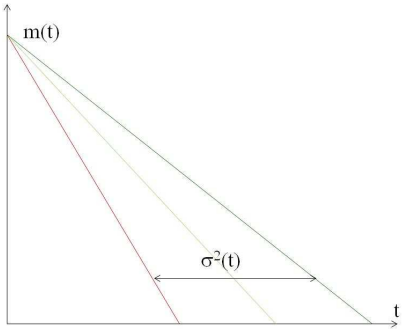


Figure 3 : stochastic process of tearing of brushes

We adopt the following probability law :

$m(t) = m(0)(T - \alpha(t))$ avec : $\alpha(t) = t + \sigma(t)$
and : $-0{,}2t \le \sigma(t) \le 0{,}2t$ (uniform law)

The failure is realizing to zero.
$\mathbb{P}(m(t) = 0) = \mathbb{P}(T - \alpha(t) = 0)$

Hence :

$$\mathbb{P}(m(t) = 0) = \mathbb{P}(T - t - \sigma(t) = 0) = P(\sigma(t) = T - t)$$

While : T-t > 0,2t, P(m(t)=0) = 0

If T/1,2 < t < T $P(m(t) = 0) = \int_{T-t}^{0,2t} \frac{dt}{0,4} = \frac{1,2t - T}{0,4}$

As soon as : t ≥ 1,2T P(m(t)=0) = 1

In this strongly unsteady state the failure probability is the first failure at t probability.
This law is approximated by a Weibull's law with eta = T and beta ~ 15.

Failure probability of the Cc electrical motor :

$$P_f(t) = 1 - e^{-\left(\frac{t}{\eta}\right)^{\beta}} \quad P_f(t) = 1 - e^{-\left(\frac{t}{1000}\right)^{15}} \qquad \textbf{(6)}$$

5 Reliability of the mechanical component

The reliability of all mechanical components is calculated by a strength-stress method applied in mean. The modeling is classical:

- Endless screw and gear (component 3): we take only the failure mode on wearing. The wearing law which is used is the Archard's law given in [5]
- Line shaft (component 4) : failure mode on torsion.
- Drum (component 5): The considered failure mode is the torsion.

6 Reliability of the hoist

The numerical application is the climbing of 1000 kg on 30 m with a speed 1m/s

6.1 Physical links between components and system : functional analysis

For building the links between components and system it is compulsory to define the internal functions to be defined for the realization of the external functions which are these ones seen by the users. The loose of functions is related to the degraded states of the system and its damages.

6.1.1 Dysfunctional analysis and fault trees

From the failure probabilities to the probabilities of the realization of the states of the system the dysfunctional analysis is the study of the causes of the loss of

external functions. These ones are lost when one or any components used to realize this function are damaged. So the dysfunctional analysis links the loss of external function from the damages of the components. Through the fault tree which is obtained by a logical inversion of the logical component-function tree it is obtained the probabilities of the functional loss.

The FMEA analysis gives the board below :

Components / Failure modes	Over intensity	Over temperature	Vibrations	dilatation	brushes	wearing	deformation
Electronic card	x	x	x	x			
Electrical motor					x		
Reduction gear						x	
Line shaft							x
Drum							x

Board 1 : FMEA analysis

6.1.2 Definition of the different working states of the hoist

This part is defined by the engineering : what kind of degradation damage and failures are admissible and when do they realize ?
In this example we decompose the work of the system in four states:
We represent the links between components functions and states by the following scheme:

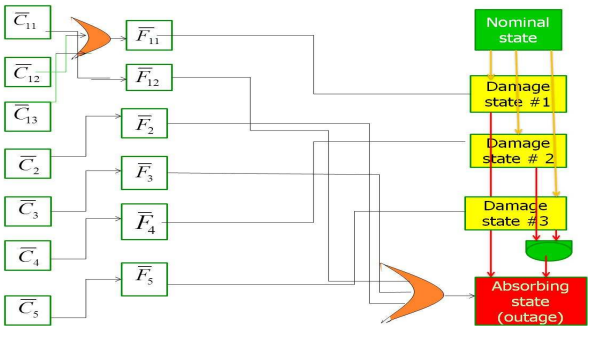


Figure 4: logical scheme of the reliability of the system

It is rather complex because we have the merging of two concepts :

- A net of chained states

- A sum of independent functions whom the loss involves the occurrence of the outage (realization of the absorbing state). The net is analyzed as a bayesian one :

$$P(A) = P(A/\overline{F}_{11})P(\overline{F}_{11}) + P(A/\overline{F}_4)P(\overline{F}_4/\overline{F}_5) + P(A/\overline{F}_5)P(\overline{F}_5/\overline{F}_4) \quad (7)$$

The sum of independent events involves the outage as soon as an event is realized :

$$P_{ind}(A) = P(\inf_t events) = 1 - \prod_i (1 - P_{fi}(t)) \quad (8)$$

6.1.3 State probabilities and Bayesian net

We achieve the method to calculate the probabilities of realization of states the transition probabilities between states and the completion of the Bayesian net.
We suppose that the realization of the $\overline{C}_{13}$ after $\overline{C}_{12}$ or $\overline{C}_{12}$ after $\overline{C}_{13}$ involves the absorbing state. So we obtain the following state graph :

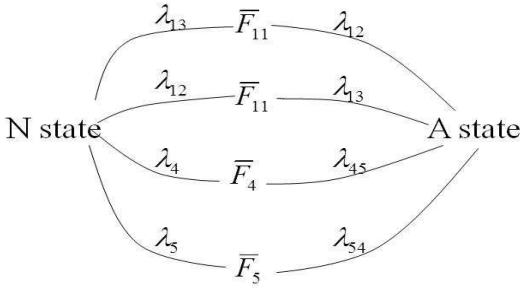


Figure 5 : transitions of the graph

6.2 Failure probabilities

- Components :

$\overline{C}_{11}$: $P_f(t) = 2{,}26 \cdot 10^{-4}$;
$\overline{C}_{12}$: $P_f(t) = 5.10^{-4}$;
$\overline{C}_{13}$: $P_f(t) = 1{,}75 \cdot 10^{-6}$;
$\overline{C}_2$: $P_f(t) = 1 - e^{-\left(\frac{t}{1000}\right)^{15}}$;
$\overline{C}_3$: $P_f(t) = 8{,}28 \cdot 10^{-2}$;
$\overline{C}_4$: $P_f(t) = 4{,}32.10^{-2}$;
$\overline{C}_5$: $P_f(t) = 4{,}12.10^{-7}$

- Functions :

$\overline{F}_{11}$: $P_{fF11} = P_{C12} + P_{C13} = 5.10^{-4} + 1{,}75 \cdot 10^{-6}$;
$\overline{F}_{12}$: $P_{fF12} = 2{,}26 \cdot 10^{-4}$;

$\overline{F_2}$: $P_{fF2} = P_f(t) = 1 - e^{-\left(\frac{t}{1000}\right)^{15}}$;

$\overline{F_3}$: P_{fF3} $P_f(t) = 8{,}28 \cdot 10^{-2}$;

$\overline{F_4}$: P_{fF4} $Pf(t) = 4{,}32.10^{-2}$;

$\overline{F_5}$: $P_{fF5} = P_f(t) = 4{,}12.10^{-7}$

For obtaining the time-dependent formulations it is necessary to introduce the damage laws in the strength – stress method. A simplifier procedure is usable. For that we understand that the previous ones are asymptotic formulations of time dependent probabilities.

$P_{fFj} = \lim_{t \to \infty} P_{fFj}(t)$, where in first approximation:

$$P_{fFj}(t) \approx P_{fFj}\left(1 - e^{-\frac{t}{T_{F_j}}}\right) \qquad (9)$$

Where T_{Fj} is the life duration of the F_j function which is depending on the life duration of its components. One supposes that its one is much minor than the life duration of the hoist. This one is admitted equal to approximately 5000 hours.

We suppose that the interval of confidence of T is equal to 5000(1 ±0,1). Approximately the interval of confidence of λ is equal to 20%

6.2.1 Calculation of the state and transition probabilities of the Bayesian net

The transition probabilities of the Bayesian network defined by the 5-state graph and the rates of transition between the corresponding states are given by:

$\lambda_{12}(t) = 1 \cdot 10^{-7}$ / hour ;

$\lambda_{13}(t) = 1 \cdot 10^{-7}$ / hour

$$\lambda_2(t) = 1{,}5.10^{-2}\left(\frac{t}{1000}\right)^{14} ;$$

$\lambda_3(t) = 1{,}66 \cdot 10^{-5}$ / hour ;

$\lambda_4(t) = 8{,}64 \cdot 10^{-6}$ / hour ;

$\lambda_5(t) = 8{,}24 \cdot 10^{-11}$ / hour ;

$\lambda_{45}(t) = \lambda_4(t)$; $\lambda_{54}(t) = \lambda_5(t)$ (10)

The probabilities of realization of the different states are given by the Kolmogorov's equations supplying the probabilities of the states from 0 to 4 with : $P_0(t) = P_N(t) : P_4(t) = P_A(t)$:

$$\begin{bmatrix} \frac{dP_0(t)}{dt} \\ \frac{dP_1(t)}{dt} \\ \frac{dP_2(t)}{dt} \\ \frac{dP_3(t)}{dt} \\ \frac{dP_4(t)}{dt} \end{bmatrix} = \begin{bmatrix} -(\lambda_{12}+\lambda_{13}) & 0 & 0 & 0 & 0 \\ (\lambda_{12}+\lambda_{13}) & -(\lambda_{12}+\lambda_{13}) & 0 & 0 & 0 \\ \lambda_4 & 0 & -\lambda_4 & 0 & 0 \\ \lambda_5 & 0 & 0 & -\lambda_5 & 0 \\ 0 & \lambda_{12}+\lambda_{13} & \lambda_{45} & \lambda_{54} & -(\lambda_{12}+\lambda_{13}+\lambda_{45}+\lambda_{54}) \end{bmatrix} \begin{bmatrix} P_0(t) \\ P_1(t) \\ P_2(t) \\ P_3(t) \\ P_4(t) \end{bmatrix}$$

$$+ \begin{bmatrix} (\lambda_{12} + \lambda_{13})P_0(0) \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \qquad (11)$$

We find after an iterative resolution the solution of these equations. The transition probabilities are defined by the probability of function loss. And the Bayesian network is completely defined by: The state graph, the transition probabilities, the state probabilities.

6.2.2 Calculation of the realization of the failures of independent functions

For independent functions we apply the minimum time law :

$$P_{ind}(A) = 1 - (1 - 2{,}26.10^{-4}).(1 - 8{,}28.10^{-2}).$$

$$(1 - (1 - e^{\left(\frac{t}{1000}\right)^{15}}))$$

The failure probability of the hoist is the sum of the P_4 and P_{ind} probabilities.

6.3 Merging the physical forecast of the reliability with the data of reliability given by the return of experiment (REX)

We suppose that the observed failures are classified according to the lost functions. We suppose that the brushes of the electrical motor are overhauled every 500 hours. No failure is observed about the motor.
We observe only failures on the reduction gear and the line shaft. The two components are simultaneously overhauled because the screw is integrated in the shaft.

About a fleet of 10 helicopters we note a failure a year on their hoists. The failure rate observed is equal to : **1/87600 = 1,14.10^{-5} h^{-1}** and we note that

70% of the outages are due to the reduction gear while the remaining is due to the line shaft. The observed ratio between the two components is equal to 0,42.

The estimated failure probability of the shaft and the gear is :

P(shaft) = $1{,}14.10^{-5}$ / 1,42 = $8{,}05.10^{-6}$ / h ;
P(gear) = $8{,}05.10^{-6}$ / 0,42 = $1{,}92.10^{-5}$ / h

The physical estimation and the statistical estimation are noisy: The uncertainties in the physical parameters on the one and the errors of measurements in life or in trials on the other one. The general method to merge the two estimations is the data fusion which the general principle is as below :
Suppose that the experimental relative interval of confidence is equal to 30% of the mean value.
Taking on account the interval of confidence of the theoretical value one obtains:

$$\lambda_{th,\exp} = \frac{\sigma_{theoratical}\lambda_{theoretical} + \sigma_{\exp erimental}\lambda_{\exp erimental}}{\sigma_{theoretical} + \sigma_{\exp erimental}} \quad (8)$$

For the gear: $\lambda_{th,\exp} = 1{,}82.10^{-5}/h$

For the shaft: $\lambda_{th,\exp} = 8{,}29.10^{-6}/h$ (12)

This relationship can be applied to the previous formulations. The aftermaths of the calculations are the same. It exists also other technical ways of data fusions what are more fine but it is not the goal of this study to make a lesson about their. Some papers about are available in the signal processing area.

7 Conclusion

The RELSYS® methodology is the synthesis between the methods for components and the methods for systems. It realizes the whole chain of calculation from the strength-stress method for components to the Bayesian net of the system. Moreover it is the first time the reliability of the electronic components is calculated by a strength-stress method and one obtains a unified general formulation of the reliability of the components. All tools existing presently each independently one another are placed in the chain of calculation according to a precise order. The experiment data obtained by trials or Return of Experiment are merged with the theoretical results by means of a data fusion method or a Bayesian procedure. The technical gap with regards to the present methods is very important and the supplied results are calculated in a rigorous mathematical framework. This approach is very useful for the new systems to estimate their reliability and fit the maintenance operations at the beginning of using. This approach is very analytic and it is rigorous in the chain of calculation. The parameters of strength in the strength-stress method are well known and given by the manufacturer and the stress parameters are deduced from the conditions of use and environment. As soon as data of REX are obtained, a data fusion between the reliability estimated by the physical approach and the statistical approach is the best approach to merge the two sets of data the first one by the physical parameters with this methodology and the second one by the return of experiment and a statistical processing. The maintenance operations are adjusted with the experimental knowledge increasing with the time. This very interesting result is given by means of the scientific calculation which is the most powerful tool of analysis.

8 Bibliography

[1] J. de Reffye : Reliability of the power electronic components by their dynamical simulation in real working conditions, ESREL 2011, Troyes, France

[2] J. de Reffye : A new Methodology to forecast Reliability and to determine appropriate maintenance, 41° ESREDA, Session 3 : Cost-effective Maintenance Policies and Model, 8, La Rochelle, France, 2011

[3] William Feller : An Introduction to Probability theory and its application, 2 Volumes, J. Wiley

[4] Johson & Kotz : distributions in statistics, continuous univariate distributions – 1, J; Wiley

[5] Archard's law : Code-Aster, Loi d'Achard

Bounds on Reliability of Coherent Systems Using Signatures

Isha Dewan
Indian Statistical Institute
New Delhi, India 110016
Email: isha@isid.ac.in

J.V. Deshpande
Indian Institute of Technology
Mumbai, India
Email: jayant921@yahoo.com

***Abstract*—The reliability of a coherent system is defined as the probability that it survives time t. Assume that the life times of the components of this system are independent and identically distributed continuous positive valued random variables. One needs the knowledge of the structure function of the system to calculate its reliability function . We assume that the component lifetimes are either IFRA or NBUFRA. In this paper we use the concept of signature of a system and the closure property of coherent systems with IFRA or NBUFRA components to find bounds for the reliability function . These bounds are in terms of the quantile of a specified order or the mean of the common component life distribution and the signature of the coherent system. The reliability bounds have been calculated for several 3, 4 and 5 component systems and we find that they are very close to the true reliability in certain instances.We are reporting partial results here.**

I. Introduction

Consider a coherent system consisting of n components with structure function $\phi(x_1, x_2, \ldots, x_n)$. That is to say,

$$\phi(x_1, x_2, \ldots, x_n) = \begin{array}{l} 1, \text{ if the system is working} \\ 0, \text{ if the system has failed,} \end{array} \quad (1)$$

where

$$x_i = \begin{array}{l} 1, \text{ if the ith component is working} \\ 0, \text{ if it has failed.} \end{array}$$

A system is a coherent system if the structure function ϕ satisfies the following two conditions

(i) ϕ is monotonic increasing in each of its arguments
(ii) $\phi(x_1, \ldots, x_{i-1}, 0, x_{i+1}, \ldots, x_n) = 0$

and $\phi(x_1, \ldots, x_{i-1}, 1, x_{i+1}, \ldots, x_n) = 0$
for some value of $(x_1, \ldots, x_{i-1}, 0, x_{i+1}, \ldots, x_n)$.

The second condition specifies the requirement that every component in the coherent system is relevant.

The reliability of any system at time t is the probability that the system does not fail upto time t. Let the random variable T denote the life time of the system, then the reliability at time t is $R(t) = P(T > t)$. It is well known that

$$R(t) = P(T > t) = \phi(\bar{F}_1(t), \bar{F}_2(t), \ldots, \bar{F}_n(t)), \quad (2)$$

where ϕ is the structure function defined in (I) and $\bar{F}_i(t)$ is the continuous survival function of the ith component with random lifetime $X_i,\ i = 1, 2, \ldots, n$. If component lifetimes are independent and identically distributed with common survival function $\bar{F}(t)$, then the reliability of the system is given as

$$R(t) = P(T > t) = \phi(\bar{F}(t), \bar{F}(t), \ldots, \bar{F}(t)). \quad (3)$$

It is usually very difficult to calculate the exact reliability of the coherent system as the structure function ϕ may be difficult to find out and the life distribution of the components need not be known . The next best thing one can do is to find bounds for the reliability function of the system .

The earliest bounds were given by Barlow and Proschan (1975 Ch 4). Bodin (1970), Beichelt and Spross (1989) found useful bounds based on cut and path sets and module-decompositions of the coherent system and Chaudhuri, Deshpande and Dharmadhikari (1991) found bounds for coherent systems where components were independent and had IFRA life distributions. However, in order to work out these bounds one would need the knowlede of the structure function of the system, which, as mentioned earlier may not be known for most complex systems. For other papers on bounds of reliability function see Deshpande and Karia (1997), Koutras, Papastavridis and Patakos (1996) and Hsieh (2003), From (2011) .

Recently Samaniego (1985), Kochar, Mukerjee and Samaniego (1999) and Samaniego (2007) have introduced and utilized the concept of 'signature' of a coherent system. They observed that the system fails when one of the component fails. For example, a series system fails when the component with the smallest lifelength fails. And a parallel system fails when the component with the longest lifetime fails. Thus, lifelength of a series system is the $min(X_1, X_2, \ldots, X_n)$ and that of a parallel system is $max(X_1, X_2, \ldots, X_n)$. In general, the system lifelength T is one of the component lifetimes. Let $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ be the ordered component life times, and

$$P[T = X_{(i)}] = s_i, \quad i = 1, 2, \ldots, n,$$

then $\underline{s} = (s_1, s_2, \ldots, s_n)$ is a probability distribution, that is, $s_i \geq 0$ and $\sum_{i=1}^{n} s_i = 1$. The vector $\underline{s}$ is defined as the

signature of the coherent system. It is seen that

$$R(t) = P(T > t) = \sum_{i=1}^{n} s_i \bar{F}_{(i)}(t), \tag{4}$$

where

$$\bar{F}_{(i)}(t) = P(X_{(i)} > t) = \sum_{j=0}^{i-1} \binom{n}{j} (1-\bar{F}(t))^j (\bar{F}(t))^{n-j} \tag{5}$$

is the survival function of the ith order statistics of the random sample $X_1, X_2, \ldots, X_n$, the lifetimes of the n components which have i.i.d. life times. The survival function represented in (5) is the reliability function of the $(n-i+1)$-out-of-n system.

Next, for sake of completeness, we state some common notions of positive ageing. If X is a continuous random variable denoting the system lifelength with distribution function H, survival function $\bar{H}$ and failure rate $r_H(x)$, then several positive ageing criteria are defined in terms of the failure rate.

Definition 1. H is said to be IFR iff $r_H(x)$ is non-decreasing in x.

Definition 2. H is said to be IFRA iff $\frac{\int_0^x r_H(u)du}{x}$ is non-decreasing in x.

Definition 3. H is said to be New Better Than Used in Failure Rate (NBUFR) distribution iff $r_H(0) \leq r_H(x)$ for all $x \geq 0$.

Definition 4. (Loh (1984)) H is said to be New Better Than Used in Failure Rate Average (NBUFRA) distribution iff $r_H(0) \leq \frac{\int_0^x r_H(u)du}{x}$ for all $x \geq 0$.

It is known from Barlow and Proschan (1975) and Deshpande Kochar and Singh (1986) that

$$IFR implies IFRA implies NBUFR implies NBUFRA$$

A. Reliability Bounds - IFRA components

In this section we look at bounds of the reliability function of a coherent system with n independent IFRA components. It is well known that if the life distributions of the components are IFRA, then the life distribution of the system is IFRA. Another result that is used is that any IFRA distribution function crosses every exponential distribution function at most once from below, if it does, and exactly once if the IFRA distribution and the exponential distribution have a common quantile ζ_p of any order $p, (0 < p < 1)$.

Barlow and Proschan (1975) proved the following result.

Lemma 2.1: Let F be an IFRA distribution and $0 < a < \infty$. Then

$$\begin{aligned} \bar{F}(t) &\geq [\bar{F}(a)]^{t/a}, \quad 0 < t \leq a, \\ \bar{F}(t) &\leq [\bar{F}(a)]^{t/a}, \quad a \leq t \leq \infty. \end{aligned} \tag{6}$$

Using the above Lemma and (2) Chaudhuri et al (1991) proved the following result.

Theorem 2.2: Let $F_i(t)$ be IFRA life distributions and $0 < a_i < \infty, i = 1, 2, \ldots, n$. Then

$$\begin{aligned} & \phi(\bar{F}_1(t), \bar{F}_2(t), \ldots, \bar{F}_n(t)) \\ \geq\ & \phi([\bar{F}_1(a_1)]^{t/a_1}, [\bar{F}_2(a_2)]^{t/a_2}, \ldots, [\bar{F}_n(a_n)]^{t/a_n}) \\ & \qquad t \leq min(a_1, a_2, \ldots, a_n), \\ \leq\ & \phi([\bar{F}_1(a_1)]^{t/a_1}, [\bar{F}_2(a_2)]^{t/a_2}, \ldots, [\bar{F}_n(a_n)]^{t/a_n}) \\ & \qquad t \geq max(a_1, a_2, \ldots, a_n). \end{aligned} \tag{7}$$

As a special case Chaudhuri et al (1991) proved that

Corollary 2.3:

$$\begin{aligned} & \phi(\bar{F}_1(t), \bar{F}_2(t), \ldots, \bar{F}_n(t)) \\ \geq\ & \phi([\bar{F}_1(a)]^{t/a}, [\bar{F}_2(a)]^{t/a}, \ldots, [\bar{F}_n(a)]^{t/a}) \\ & \qquad t \leq a, \\ \leq\ & \phi([\bar{F}_1(a)]^{t/a}, [\bar{F}_2(a)]^{t/a}, \ldots, [\bar{F}_n(a)]^{t/a}) \\ & \qquad t \geq a. \end{aligned} \tag{8}$$

Note that one needs to know the structure function of the system to work out these bounds to the reliability function. However, the following theorem gives bounds based on the signature of the system.

Theorem 2.4: Let $(s_1, s_2, \ldots, s_n)$ be the signature of a coherent system composed of n components with independent and identically distributed lifetimes with common distribution F being an IFRA distribution. Suppose that the survival function $\bar{F}$ has the quantile ζ_p of order p. Then, the reliability of the system is bounded as below:

$$\begin{aligned} R(t) &\geq \sum_{i=1}^{n} s_i \sum_{j=0}^{i-1} \binom{n}{j} (1 - e^{t\frac{log(1-p)}{\xi_p}})^j (e^{t\frac{log(1-p)}{\xi_p}})^{n-j} \\ & \qquad t \leq \xi_p, \\ R(t) &\leq \sum_{i=1}^{n} s_i \sum_{j=0}^{i-1} \binom{n}{j} (1 - e^{t\frac{log(1-p)}{\xi_p}})^j (e^{t\frac{log(1-p)}{\xi_p}})^{n-j} \\ & \qquad t \geq \xi_p. \end{aligned} \tag{9}$$

The bounds given in the above theorem are based on the signature of the coherent system and the quantile of order p of the common component survival function. Besides it will be easier to find bounds based on (9) instead of the structure function. These bounds are sharp as lower and upper bounds together is the exact reliability function of the coherent system under consideration where components are exponentially distributed with specified parameter.

If $\bar{F}_i$ is IFRA then it intersects each exponential reliabilty function atmost once. If the two distributions have the same mean (say μ_i), then the corresponding reliability functions must intersect. Let t_{i0} be the unique point of interesection of $\bar{F}_i$ and $exp(-t/\mu_i)$. Then Chaudhuri et al (1991) proved the following result.

Corollary 2.5:

$$\begin{aligned} & \phi(\bar{F}_1(t), \bar{F}_2(t), \ldots, \bar{F}_n(t)) \\ \geq \quad & \phi(e^{-t/\mu_1}, e^{-t/\mu_2}, \ldots, e^{-t/\mu_n}) \\ & \qquad\qquad t \leq min(t_{10}, t_{20}, \ldots, t_{n0}), \\ \leq \quad & \phi(e^{-t/\mu_1}, e^{-t/\mu_2}, \ldots, e^{-t/\mu_n}) \\ & \qquad\qquad t \geq max(t_{10}, t_{20}, \ldots, t_{n0}). \end{aligned} \tag{10}$$

The following bounds are in terms of the signature and the mean of the distribution.

Theorem 2.6: Let $(s_1, s_2, \ldots, s_n)$ be the signature of a coherent system composed of n components with independent and identically distributed lifetimes with common survival function $\bar{F}$ having mean μ. Then the reliability of the system is bounded as below:

$$\begin{aligned} R(t) \quad \geq \quad & \sum_{i=1}^{n} s_i \sum_{j=0}^{i-1} \binom{n}{j} (1 - e^{t\frac{log\bar{F}(\mu)}{\mu}})^j (e^{t\frac{log\bar{F}(\mu)}{\mu}})^{n-j} \\ & \qquad\qquad t \leq \mu, \\ R(t) \quad \leq \quad & \sum_{i=1}^{n} s_i \sum_{j=0}^{i-1} \binom{n}{j} (1 - e^{t\frac{log\bar{F}(\mu)}{\mu}})^j (e^{t\frac{log\bar{F}(\mu)}{\mu}})^{n-j} \\ & \qquad\qquad t \geq \mu. \end{aligned} \tag{11}$$

Remark 2.7: Note that the bound in (11) can also be expressed as

$$\begin{aligned} & \sum_{i=1}^{n} s_i \sum_{j=0}^{i-1} \binom{n}{j} (1 - e^{-t\frac{log\bar{F}(\mu)}{\mu}})^j (e^{-t\frac{log\bar{F}(\mu)}{\mu}})^{n-j} \\ = \quad & \sum_{i=1}^{n-1} \binom{n}{i} (1 - e^{-t\frac{log\bar{F}(\mu)}{\mu}})^i (e^{-t\frac{log\bar{F}(\mu)}{\mu}})^{n-i} \sum_{j=i}^{n} s_j. \end{aligned}$$

1) Numerical Results : We carried out numerical results to compare the exact reliability of the system with the reliability bounds based on signatures.

Figures 3.1-3.5 give the expressions for the same for a series system consisting of 4 independent IFRA componnets. The red curve gives the exact reliability function and the blue gives the reliability bounds. We have upper bounds before the point of intersection and lower bounds after that.

Similarly figures 3.6-3.10 give the expressions for the exact reliability and its bounds for a parallel system consisting of 4 independent IFRA componnets.

Figure 3.1 (Series system (4) $p = .1$)

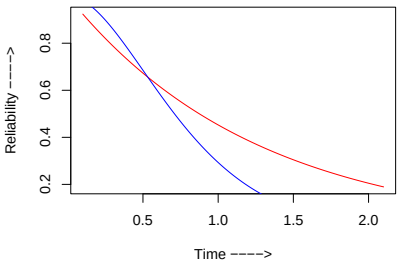
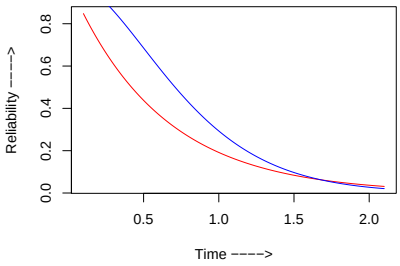


Figure 3.5 (Series system (4) $p = .5$)



Similar numerical results have been obtained for a 5 component bridge structure. Comparisons have been done for bounds based on structure functions and bounds based on signatures.

Figure 3.8 (Parallel system (4) $p = .3$)

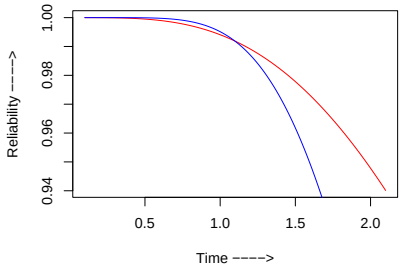
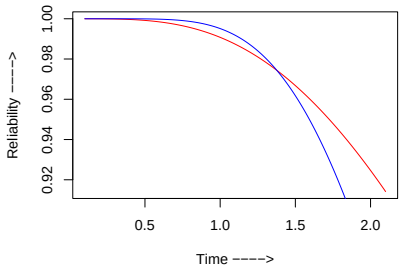


Figure 3.9 (Parallel system (4) $p = .4$)



II. Conclusion

Notice that both the lower bound part and the upper bound part are close to the exact reliability for the systems considered above. Once we know that component lifetimes have IFRA distribution we can use the bound values as the actual values of the relaibility.

We have also found bounds for the reliability function based on structure functions and signatures when the components are independent and have NBUFRA life distributions.

The work is being extended to systems with exchangeable components.

References

[1] Barlow, R.E. and Proschan, F. *Statistical Theory of Reliability and Life Testing*, Holt, Rinehart and Winston, New York, 1975.

[2] Beichelt, F. A. and Spross, L. (1989). Bounds on reliability of binary coherent systems. *IEEE Transactions on Reliability Theory*, **38**, 425-427.

[3] Bodin, L.D. (1970). Approximations to system reliability using a modular decomposition. *Technometrics*, **12**, 335-344.

[4] Chaudhuri, G., Deshpande, J.V. and Dharmadhikari, A.D. (1991). Some bounds on reliability of coherent systems of IFRA components. *J. Appl. Probab.*, **28**, 709-714.

[5] Deshpande, Jayant V.; Kochar, Subhash C.and Singh, H. (1996). Aspects of positive ageing. J. Appl. Probab., **23** , 748758.

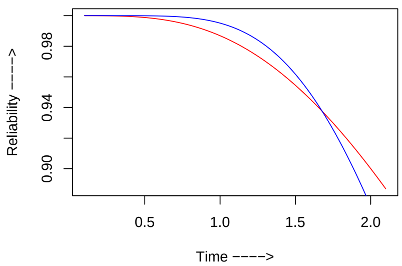
[6] From, S.G. (2011).Some new reliability bounds for sums of NBUE random variables. *Probab. Engg. Information Sciences*, **25**, 83-102.

[7] Hsieh, Y.C. (2003), New reliability bounds for coherent systems. *J. Oper. Res*, **54**, 995-1001

[8] Deshpande, J. V. and Karia, S. R.(1997), Bounds for the joint survival and incidence functions through coherent system data. *Adv. in Appl. Probab.*, **29**, 478497.

[9] Kochar, S.C., Mukerjee, H. and Samaniego, F.J. (1999). The signature of a coherent system and its application to comparisons among systems. *Naval Research Logistics*, **46**, 507-523.

Figure 3.10 (Parallel system (4) $p = .5$)



[10] Koutras, M.V., Papastavridis S.G. and Patakos K., (1996). Bounds for coherent reliability structures, *Statist. and Probab. Letters*, **26**, 285-292.

[11] Loh, W.V.(1984), A new generalization of the class of NBU distributions. *IEEE Transactions of Reliability*, **33**, 419-422.

[12] Navarro, J. and Rychlik, T. (2007). Reliability and expectation bounds for coherent systems with exchangeable components. *J. Multivariate Analysis*, **98**, 102-113.

[13] Navarro, J. and Rubio R., (2009), Computations of signatures of coherent systems of five components. *Comm. Statist.. Sim and Comp*, **39**, 68-84.

[14] Navarro, J. and Rychlik, T. (2010). Comparisons and bounds for expected lifetimes of reliability systems. *European J. Oper. Res.*, **207**, 309317.

[15] Samaniego, F. J. (1985). On closure of the IFR class under formation of coherent systems. *IEEE Transactions on Reliability Theory*, **34**, 69-72.

[16] Samaniego, F. J. . *System Signatures and their Applications in Engineering Reliability*. Springer, New York. 2007

On A Generalized Shot-noise Type Failure Model

Maxim Finkelstein
Department of Mathematical Statistics, University of the Free State, 339 Bloemfontein 9300, South Africa
e-mail : FinkelM@ufs.ac.za
Max Planck Institute for Demographic Research, Germany

Ji Hwan Cha
Department of Statistics, Ewha Womans University, Seoul, 120-750, Korea
e-mail : jhcha@ewha.ac.kr

Abstract— **Standard assumptions in shock models are that failures of items are related either to the cumulative effect of shocks (cumulative models) or that they are caused by shocks that exceed a certain critical level (extreme shocks models). In this paper, we present a useful generalization of this setting to the case when an item is deteriorating itself, e.g., when the boundary for the fatal shock magnitude is decreasing with time. A generalized shot-noise type shock model will be studied. Explicit formula for the corresponding survival function is derived and several simple examples are considered.**

Keywords-reliability; shot-noise process; cumulative shock model; extreme shock model; intensity process

I. INTRODUCTION

Many of the currently used failure models are developed under the premise that the operating environment is static. In these cases, the basic assumption is that the prevailing environmental conditions either do not change in time or, in case they do, have no effect on the deterioration and failure process of the device. Therefore, in these cases, models not depending on the external environmental conditions are proposed and studied.

However, devices often work in varying environments and so their performance is significantly affected by these varying environmental conditions. In this paper, we consider external (environmental) shocks as a cause for system's failure and deterioration. For instance, numerous electronic devices are frequently subject to random shocks caused by fluctuations of unstable electric power. In these cases, the changes in external conditions result either in immediate failure or deterioration of equipment.

Shock models usually consider systems that are subject to shocks of random magnitudes at random times. Traditionally, one distinguishes between two major types: cumulative shock models (systems break down because of a cumulative effect) and extreme shock models (systems break down because of one single large shock). Some references (to name a few) are: Shanthikumar and Sumita [1], Sumita and Shanthikumar [2], Gut [3], Finkelstein [4], Cha and Finkelstein [5], Finkelstein and Marais [6]. A combination of these models was investigated by Gut and Hüsler [7], where the failures were due either to a cumulative effect, or to a single, fatal shock. In this paper, we are somehow in the framework of the latter setting generalizing it to the case when a system itself (apart from the shock process) is deteriorating with time. However, mathematically, our approach is closer to the paper by Lemoine and Wenocur [8] (see also Lemoine and Wenocur [9]) and is based on considering the shot noise process-type stochastic intensity as a model for shocks accumulation.

Assume that a system is subject to the NHPP (Nonhomogeneous Poisson Process) of shocks $N(t), t \geq 0$ with rate $\lambda(t)$, which is the only possible cause of its failure. The consequences of shocks are accumulated in accordance with the 'standard' shot-noise process $X(t)$, $X(0)=0$ (see e.g., Rice [10], Ross [11]) and therefore, define the level of the cumulative *stress* (from all prior shocks) at time t as the following stochastic process:

$$X(t)=\sum_{j=1}^{N(t)} D_j h(t-T_j), \tag{1}$$

where T_n is the n-th arrival time of the shock process, $D_j, j=1,2,...$ (i.i.d.) are the magnitudes of shocks and $h(t)$ is a nonnegative, non-increasing for $t \geq 0$, deterministic function and $h(t)=0$ for $t<0$. The usual assumption for considering asymptotic properties of $X(t)$ is that $h(t)$ vanishes as $t \to \infty$ and its integral in $[0,\infty)$ is finite (see, e.g., Lund *et al.* [12]), but here, we formally do not need this rather restrictive assumption. The shock process $\{N(t), t \geq 0\}$ and the sequence $\{D_1, D_2, ...\}$ are supposed to be independent.

The cumulative stress eventually results in failures, which T can be probabilistically described in different ways. Denote by the failure time of our system. Lemoine and Wenocur [8], for example, model the distribution of T by assuming that the corresponding conditional failure (intensity) rate process (on condition that $\{N(t), T_1, T_2, ..., T_{N(t)}\}$ and $\{D_1, D_2, ..., D_{N(t)}\}$ are given) is proportional to $X(t)$. This is a reasonable assumption that describes the proportional dependence of the probability of failure in the infinitesimal interval of time on the level of stress:

$$\lambda_t \equiv kX(t)=k\sum_{j=1}^{N(t)} D_j h(t-T_j), \tag{2}$$

where $k>0$ is the constant of proportionality. Then

$$P(T>t \mid N(s),0\le s\le t, D_1, D_2,..., D_{N(t)})$$
$$= \exp\{-k\int_0^t \sum_{j=1}^{N(x)} D_j h(x-T_j)dx\}, \quad (3)$$

This probability should be understood conditionally on the corresponding realizations of $N(s),0\le s\le t$ and $D_1, D_2,..., D_{N(t)}$. Therefore,

$$P(T>t) = E\left[\exp\left\{-k\int_0^t X(u)du\right\}\right].$$

Lemoine and Wenocur [8] had finally derived the following formula for this probability:

$$P(T>t) = \exp\{-\Lambda(t)\}\exp\left\{\int_0^t L(kH(u))\lambda(t-u)du\right\}, \quad (4)$$

where $\Lambda(t)=\int_0^t \lambda(u)du$, $H(t)=\int_0^t h(u)du$ and L is the operator of the Laplace transform with respect to the distribution of the shock's magnitude.

The main goal of our paper is to generalize this approach to the case when a system can also fail due to a fatal shock with the magnitude exceeding the time-dependent bound, which is more realistic in practice.

The structure of the paper is as follows. In Section 2, a shot noise-type failure model, which generalizes the initial model of Lemoine and Wenocur [8] is suggested by considering deterioration of our system in time (apart from the shock process). In this model, we also assume that the conditional failure rate function is proportional to the accumulated stress caused by a stochastic shock process. The survival function and the corresponding failure rate function are derived and relevant discussions are presented. Finally in Section 3, the concluding remarks are given.

II. Generalized Shot-noise Type Failure Model

In addition to the general assumptions of Lemoine and Wenocur [8] stated in Section 1 (see also equations (2)-(4)), let on each shock, depending on its magnitude $D_j, j=1,2...$, the following mutually exclusive events occur,

(i) If $D_j > g_U(T_j)$, then the shock results in an immediate system's failure

(ii) If $D_j \le g_L(T_j)$, then the shock does not cause any change in the system (harmless)

(iii) If $g_L(T_j) < D_j \le g_U(T_j)$, then the shock increases the stress by $D_j h(0)$,

where $g_U(t)$, $g_L(t)$ are decreasing, deterministic functions. Note that, in accordance with (1), the 'remaining' stress after s units of time from the occurrence of a single non-fatal shock is $D_j h(s)$.

The functions $g_U(t)$, $g_L(t)$ are the upper and the lower bounds, which are the functions of operating time. Because they are decreasing, this means that the probability that the shock arriving at time t results in the system's failure is increasing in time, whereas the probability that the shock is harmless is decreasing with time. Therefore, obviously, a deterioration of our system is described in this way. The function $g_U(t)$ can also be interpreted as the strength of our system with respect to shocks, whereas the function $g_L(t)$, as the 'sensitivity' to shocks.

Define the following 'membership function':

$$\xi(T_j, D_j) = \begin{cases} 1, & g_L(T_j) < D_j \le g_U(T_j) \\ 0, & D_j \le g_L(T_j) \end{cases}. \quad (5)$$

Using this notation, the cumulative stress, similar to (1), can be written as

$$X(t) \equiv \sum_{j=1}^{N(t)} \xi(T_j, D_j) D_j h(t-T_j), \quad (6)$$

provided that the system is operating at time t (i.e., the event $D_j > g_U(T_j), j=1,2,...$ did not happen in $[0,t)$).

Generalizing (2), assume that the conditional failure rate process $\hat{\lambda}_t$ (on condition that the event $D_j > g_U(T_j), j=1,2,...$ did not happen in $[0,t)$ and $\{N(t), T_1, T_2,..., T_{N(t)}\}$ and $\{D_1, D_2,..., D_{N(t)}\}$ are given) is proportional to $X(t)$:

$$\hat{\lambda}_t \equiv kX(t) = k\sum_{n=1}^{N(t)} \xi(T_j, D_j) D_j h(t-T_j), k>0. \quad (7)$$

It is clear that conditionally on the corresponding history:

(i) If $D_j > g_U(T_j)$, for at least one j, then

$$P(T>t \mid N(s),0\le s\le t, D_1, D_2,..., D_{N(t)}) = 0;$$

(ii) If $D_j \le g_U(t)$, for all j, then

$$P(T>t \mid N(s),0\le s\le t, D_1, D_2,..., D_{N(t)})$$
$$= \exp\{-k\int_0^t \sum_{j=1}^{N(x)} \xi(T_j, D_j) D_j h(x-T_j)dx\}.$$

Therefore,

$$P(T>t \mid N(s),0\le s\le t, D_1, D_2,..., D_{N(t)})$$
$$= \prod_{j=1}^{N(t)} \gamma(T_j, D_j) \cdot \exp\{-k\int_0^t \sum_{j=1}^{N(x)} \xi(T_j, D_j) D_j h(x-T_j)dx\}, \quad (8)$$

where

$$\gamma(T_j, D_j) = \begin{cases} 0, & D_j > g_U(T_j) \\ 1, & D_j \le g_U(T_j) \end{cases}. \quad (9)$$

Thus, we have described a rather general model that extends (3) to the defined deterioration pattern. Indeed, if $g_U(t)=\infty; g_L(t)=0$, then $\xi(T_j, D_j)\equiv 1$ and (8) reduces to (3) with the corresponding survival probability (4). On the

other hand, let $g_U(t) = g_L(t) = g(t)$. Then, defining $p(t) = P(D_j > g(t))$ as the probability of failure under a shock at time t ($q(t) = P(D_j \le g(t))$, we obviously arrive at the, so called, $p(t) \Leftrightarrow q(t)$ model (Cha and Finkelstein [5]) for which the unconditional probability of survival is

$$P(T > t) = \exp\left\{-\int_0^t p(u)\lambda(u)du\right\}, \quad (10)$$

with the corresponding failure rate

$$r(t) = p(t)\lambda(t), \quad (11)$$

which is also widely known in the literature as the extreme shock model (e.g., Gut and Hüsler [7]).

Based on the above described model, we will derive now the (unconditional) survival function and the corresponding failure rate function. We will need the following general lemma (See, Cha and Mi [13] for the proof):

Lemma 1. *Let $X_1, X_2, ..., X_n$ be i.i.d. random variables and $Z_1, Z_2, ..., Z_n$ be i.i.d. continuous random variables with the corresponding common pdf. Furthermore, let $\mathrm{X} = (X_1, X_2, ..., X_n)$ and $\mathrm{Z} = (Z_1, Z_2, ..., Z_n)$ be independent. Suppose that the function $\varphi(x, z): R^n \times R^n \to R$ satisfies $\varphi(\mathrm{X}, t) =^d \varphi(\mathrm{X}, \pi(t))$, for any vector $t \in R^n$ and for any n-dimensional permutation function $\pi(\cdot)$. Then*

$$\varphi(\mathrm{X}, \mathrm{Z}) =^d \varphi(\mathrm{X}, \mathrm{Z}^*),$$

where $\mathrm{Z}^* = (Z_{(1)}, Z_{(2)}, ..., Z_{(n)})$ is the vector of the order statistics of Z.

We are ready now to prove the main theorem.

Theorem 1. *Let $H(t) = \int_0^t h(v)dv$, $\Lambda(t) \equiv E(N(t)) = \int_0^t \lambda(x)dx$ and $f_D(u)$, $F_D(u)$ be the pdf and the Cdf of $D =^d D_j$, $j = 1, 2, ...$. Assume that the inverse function $\Lambda^{-1}(t)$ exists. Then the survival function that corresponds to the lifetime T is*

$$P(T > t) = \exp\left\{-\int_0^t \bar{F}_D(g_L(u))\lambda(u)du\right\} \times \exp\left\{\int_0^t \int_{g_L(s)}^{g_U(s)} \exp\{-kuH(t-s)\} f_D(u)du\lambda(s)ds\right\}, \quad (12)$$

and the corresponding failure rate is

$$r(t) = P(D > g_U(t))\lambda(t) + \int_0^t \int_{g_L(s)}^{g_U(s)} kuh(t-s)\exp\{-kuH(t-s)\} f_D(u)du\lambda(s)ds. \quad (13)$$

Proof.

Observe that

$$P(T > t \mid N(s), 0 \le s \le t, D_1, D_2, ..., D_{N(t)}) = \prod_{j=1}^{N(t)} \gamma(T_j, D_j)\exp\{-k\sum_{j=1}^{N(t)} \xi(T_j, D_j)D_j H(t-T_j)\}$$

$$= \exp\{\sum_{j=1}^{N(t)} (\ln\gamma(T_j, D_j) - k\xi(T_j, D_j)D_j H(t-T_j))\}.$$

Therefore,

$$P(T > t) = E\left[\exp\{\sum_{j=1}^{N(t)} (\ln\gamma(T_j, D_j) - k\xi(T_j, D_j)D_j H(t-T_j))\}\right]$$

$$= E\left[E\left(\exp\{\sum_{j=1}^{N(t)} (\ln\gamma(T_j, D_j) - k\xi(T_j, D_j)D_j H(t-T_j))\} \middle| N(t)\right)\right].$$

Observe that, if $\Lambda^{-1}(t)$ exists, then the joint distribution of $T_1, T_2, ..., T_n$, given $N(t) = n$, is the same as the joint distribution of the order statistics $T_{(1)}' \le T_{(2)}' \le ... \le T_{(n)}'$ of i.i.d. random variables $T_1', T_2', ..., T_n'$, where the pdf of the common distribution of T_j''s is given by $\lambda(x)/\Lambda(t)$. Thus,

$$E\left(\exp\{\sum_{j=1}^{N(t)} (\ln\gamma(T_j, D_j) - k\xi(T_j, D_j)D_j H(t-T_j))\} \middle| N(t) = n\right)$$

$$= E\left(\exp\{\sum_{j=1}^{n} (\ln\gamma(T_{(j)}', D_j) - k\xi(T_{(j)}', D_j)D_j H(t-T_{(j)}'))\}\right).$$

Let $\mathrm{X} = (D_1, D_2, ..., D_n)$, $\mathrm{Z} = (T_1', T_2', ..., T_n')$ and

$$\varphi(\mathrm{X}, \mathrm{Z}) \equiv \sum_{j=1}^{n} (\ln\gamma(T_j', D_j) - k\xi(T_j', D_j)D_j H(t-T_j')). \quad (14)$$

Note that, as it was mentioned, if $g_U(t) = \infty; g_L(t) = 0$, then $\xi(T_j, D_j) \equiv 1$ and our model reduces to the original model of Lemoine and Wenocur [8], where each term in $\varphi(\mathrm{X}, \mathrm{Z})$ is just a simple product of D_j and $H(t-T_j')$. Due to this simplicity, the rest was straightforward. Now we have a much more complex form of $\varphi(\mathrm{X}, \mathrm{Z})$, as given in (14), where the terms in the sum cannot be factorized.

Observe that the function $\varphi(x, z)$ satisfies

$$\varphi(\mathrm{X}, t) =^d \varphi(\mathrm{X}, \pi(t)),$$

for any vector $t \in R^n$ and for any n-dimensional permutation function $\pi(\cdot)$. Thus applying Lemma 1,

$$\sum_{j=1}^{n} (\ln\gamma(T_j', D_j) - k\xi(T_j', D_j)D_j H(t-T_j')) =^d \sum_{j=1}^{n} (\ln\gamma(T_{(j)}', D_j) - k\xi(T_{(j)}', D_j)D_j H(t-T_{(j)}')),$$

and therefore,

$$E\left(\exp\{\sum_{j=1}^{n} (\ln\gamma(T_{(j)}', D_j) - k\xi(T_{(j)}', D_j)D_j H(t-T_{(j)}'))\}\right)$$

$$= E\left(\exp\{\sum_{j=1}^{n} (\ln\gamma(T_j', D_j) - k\xi(T_j', D_j)D_j H(t-T_j'))\}\right)$$

$$= \left(E\left(\exp\{\ln\gamma(T_1', D_1) - k\xi(T_1', D_1)D_1H(t-T_1')\}\right)\right)^n.$$

As

$$E\left[\exp\{\ln\gamma(T_1', D_1) - k\xi(T_1', D_1)D_1H(t-T_1')\} \mid T_1' = s\right]$$

$$= E\left[\exp\{\ln\gamma(s, D_1) - k\xi(s, D_1)D_1H(t-s)\}\right]$$

$$= \int_{g_L(s)}^{g_U(s)} \exp\{-kuH(t-s)\} f_D(u)du + P(D_1 \le g_L(s)), \qquad (15)$$

where for $D_1 > g_U(s)$,

$\exp\{\ln\gamma(s, D_1) - k\xi(s, D_1)D_1H(t-s)\} = 0$, for all $s > 0$, the unconditional expectation is

$$E\left[\exp\{\ln\gamma(T_1', D_1) - k\xi(T_1', D_1)D_1H(t-T_1')\}\right]$$
$$= \int_0^t \int_{g_L(s)}^{g_U(s)} \exp\{-kuH(t-s)\} f_D(u)du \frac{\lambda(s)}{\Lambda(t)} ds + \int_0^t P(D_1 \le g_L(s)) \frac{\lambda(s)}{\Lambda(t)} ds \cdot$$

Let

$$\alpha(t) \equiv \int_0^t \int_{g_L(s)}^{g_U(s)} \exp\{-kuH(t-s)\} f_D(u)du\lambda(s)ds + \int_0^t P(D_1 \le g_L(s))\lambda(s)ds,$$

and we finally arrive at

$$P(T > t)$$
$$= \sum_{n=0}^{\infty} \left(\frac{\alpha(t)}{\Lambda(t)}\right)^n \cdot \frac{\Lambda(t)^n}{n!} \exp\{-\int_0^t \lambda(u)du\}$$
$$= \exp\{-\int_0^t \lambda(u)du$$
$$+ \int_0^t \int_{g_L(s)}^{g_U(s)} \exp\{-kuH(t-s)\} f_D(u)du\lambda(s)ds + \int_0^t P(D_1 \le g_L(u))\lambda(u)du\},$$

which is obviously equal to (12).

The corresponding failure rate can be obtained as

$$r(t) = -\frac{d}{dt} \ln P(T > t)$$
$$= \lambda(t) - P(g_L(t) \le D_1 \le g_u(t))\lambda(t) + \int_0^t \int_{g_L(s)}^{g_U(s)} kuh(t-s) \exp\{-kuH(t-s)\} f_D(u)du\lambda(s)ds - P(D_1 \le g_L(t))\lambda(t)$$
$$= P(D_1 > g_U(t))\lambda(t) + \int_0^t \int_{g_L(s)}^{g_U(s)} kuh(t-s) \exp\{-kuH(t-s)\} f_D(u)du\lambda(s)ds,$$

where the Leibnitz rule was used for differentiation of the double integral. ■

Remark 1. Relationship (13) suggests that (12) can be equivalently written as

$$P(T > t) = \exp\left\{-\int_0^t \overline{F}_D(g_U(u))\lambda(u)du\right\} \times \exp\left\{-\int_0^t \int_{g_L(s)}^{g_U(s)} kuh(t-s) \exp\{-kuH(t-s)\} f_D(u)du\lambda(s)ds\right\}$$

and therefore, we can interpret our system as a series one with two independent components: one that fails only due to fatal (critical) shocks and the other that fails only due to non-fatal shocks.

Remark 2. This result can be generalized in a straightforward way to the case when the shock's magnitude (at time t) is multiplied by the increasing function $\phi(t)$, which also models additional deterioration of our system. For this generalized model, (6) can be written as:

$$X(t) \equiv \sum_{j=1}^{N(t)} \xi(T_j, D_j) \cdot D_j\phi(T_j) \cdot h(t-T_j) \cdot$$

Equation (13), e.g., in this case is modified to

$$r(t) = P(D > g_U(t))\lambda(t) + \int_0^t \int_{g_L(s)}^{g_U(s)} ku\phi(s)h(t-s) \exp\{-ku\phi(s)H(t-s)\} f_D(u)du\lambda(s)ds \cdot$$

Example 1. Consider the special case when $g_U(t) = \infty$ and $g_L(t) = 0$. Then the survival function in (12) is

$$P(T > t) = \exp\left\{-\int_0^t \overline{F}_D(g_L(u))\lambda(u)du\right\} \times \exp\left\{\int_0^t \int_{g_L(s)}^{g_U(s)} \exp\{-kuH(t-s)\} f_D(u)du\lambda(s)ds\right\}$$
$$= \exp\{-\Lambda(t)\} \exp\left\{\int_0^t L(kH(t-s))\lambda(s)ds\right\}$$
$$= \exp\{-\Lambda(t)\} \exp\left\{\int_0^t L(kH(u))\lambda(t-u)du\right\},$$

where L is the operator of the Laplace transform with respect to $f_D(u)$. Therefore, we arrive at the relationship (4) obtained by Lemoine and Wenocur [8].

Example 2. Suppose that $\lambda(t) = \lambda$, $t \ge 0$, $D_j \equiv d$, $j = 1,2,\ldots$, and there exist $t_2 > t_1 > 0$ such that $g_U(t) > g_L(t) > d$, for $0 \le t < t_1$ (shocks are harmless); $d > g_U(t) > g_L(t)$, for $t_2 < t$ (shocks are fatal), and $g_U(t) > d > g_L(t)$, for $t_1 < t < t_2$; $g_L(t_1) = g_U(t_2) = d$.

Let for the sake of further integration, $h(t) = 1/(1+t)$, $t \ge 0$, and $k = 1/d$ (for simplicity of notation). The intermediate equation in (15) can be written as

$$E\left[\exp\{\ln\gamma(T_1', D_1) - k\xi(T_1', D_1)D_1H(t-T_1')\} \mid T_1' = s\right]$$

$$= \exp\{\ln\gamma(s,d) - k\xi(s,d)dH(t-s)\}$$

$$= \begin{cases} 0, \text{ if } g_U(s) > d \ (s > t_2) \\ \exp\{-H(t-s)\}, \text{ if } g_L(s) < d \le g_U(s) \ (t_1 < s \le t_2) \\ 1, \text{ if } d \le g_L(s) (s \le t_1) \end{cases}$$

$$= \exp\{-H(t-s)\} I(g_L(s) < d \le g_U(s)) + I(d \le g_L(s))$$

$$= \exp\{-H(t-s)\} I(t_1 < s \le t_2) + I(s \le t_1).$$

Thus, 'integrating $T_1' = s$ out':

$$E\left[\exp\{\ln\gamma(T_1', D_1) - k\xi(T_1', D_1) D_1 H(t - T_1')\}\right]$$

$$= \frac{1}{\Lambda(t)}\left[\int_0^t \exp\{-H(t-s)\} I(t_1 < s \le t_2)\lambda(s)ds + \int_0^t I(s \le t_1)\lambda(s)ds\right].$$

Then,

$$P(T > t) = \exp\{-\int_0^t \lambda(u)du$$

$$+ \int_0^t \exp\{-H(t-s)\} I(t_1 < s \le t_2)\lambda(s)ds + \int_0^t I(s \le t_1)\lambda(s)ds\}$$

$$= \exp\{-\int_0^t I(s > t_1)\lambda(s)ds + \int_0^t \exp\{-H(t-s)\} I(t_1 < s \le t_2)\lambda(s)ds\}$$

Thus:

(i) For $0 \le t \le t_1$, $P(T > t) = 1$;

(ii) For $t_1 \le t \le t_2$,

$$P(T > t) = \exp\left\{-\int_{t_1}^t \lambda du\right\} \exp\left\{\lambda \int_{t_1}^t \exp\{-H(t-s)\}ds\right\}$$

$$= \exp\{-\lambda(t - t_1)\} \exp\{\lambda \ln(1 + t - t_1)\}$$

$$= \exp\{-\lambda(t - t_1)\}(1 + t - t_1)^{\lambda};$$

(iii) For $t_2 \le t$,

$$P(T > t) = \exp\left\{-\int_{t_1}^t \lambda du\right\} \exp\left\{\lambda \int_{t_1}^{t_2} \exp\{-H(t-s)\}ds\right\}$$

$$= \exp\{-\lambda(t - t_1)\}(1 + t_2 - t_1)^{\lambda},$$

which shows (compared with case (ii)) that if the system has survived in $0 \le t \le t_2$, then the next shock with probability 1 will 'kill it'.

III. Concluding Remarks

In this paper, we consider the combined shock model when the failure of a system can occur either due to the fatal (critical) shock or due to accumulation in the corresponding intensity process (hazard rate process). This accumulation is modeled via the shot noise-type processes that take into account the consequences of all shocks that occurred previously.

We generalize the previous results in the literature to the case when a system is deteriorating with time, i.e., when the boundaries that define the corresponding operable region are decreasing with time. In the absence of boundaries, our results reduce to the findings of Lemoine and Wenocur [8]. On the other hand, when both boundaries coincide, the well-known $p(t) \Leftrightarrow q(t)$ model emerges as the specific case.

The important assumption, which allows for tractable formulas (for the corresponding survival functions and failure rates), is that the process of shocks is NHPP.

Acknowledgements

The work of the first author was supported by the NRF (National Research Foundation of South Africa) grant FA2006040700002. The work of the second author was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) (No. 2011-0017338).

References

[1] J. G. Shanthikumar, U. Sumita, "Distribution properties of the system failure time in a general shock model", Advances in Applied Probability, vol. 16**,** pp. 363-377, 1984.

[2] U. Sumita U, J. G. Shanthikumar, "A class of correlated cumulative shocks models", Advances in Applied Probability, vol. 17, pp. 347-366, 1985.

[3] A. Gut, "Cumulated shock models". Advances in Applied Probability; vol. 22, pp. 504-507, 1990.

[4] M. Finkelstein, Failure rate Modelling for Reliability and Risk. London: Springer, 2008.

[5] J. H. Cha, M. Finkelstein, "On terminating shock process with independent wear increments", Journal of Applied Probability, vol. 46, pp. 353-362, 2009.

[6] M. Finkelstein, F. Marais, "On terminating Poisson processes in some shock models", Reliability Engineering and System Safety, vol. 95, pp. 874-879, 2010.

[7] A. Gut, J. Hüsler, "Realistic variation of shock models", Statistics & Probability Letters, vol. 74**,** pp. 187-204, 2005.

[8] A. J. Lemoine, M. L. Wenocur, "A note on shot-noise and reliability modeling", Operations Research, vol. 34, pp. 320-323. 1986.

[9] A. J. Lemoine, M. L. Wenocur, "On failure modeling", Naval Research Logistics, vol. 32, pp. 497-508, 1985.

[10] J. Rice, "On generalized shot noise", Advances in Applied Probability, vol. 9, pp. 553-565, 1977

[11] S. M. Ross, Stochastic Processes. New York: John Wiley, 1996.

R. Lund, W. McCormic, U. Xiao, "Limiting properties of Poisson shot noise processes", Journal of Applied Probability, vol. 41, pp. 911-918, 2004.

[12] J. H. Cha, J. Mi, "On a stochastic survival model for a system under randomly variable environment", Methodology and Computing in Applied Probability, vol. 13**,** pp. 549-561, 2011.

Probabilistic modeling of SN curve

Remy Fouchereau
SNECMA
SAFRAN Group Villaroche France
and
Département de Mathématiques
Université de Paris Sud France
Email: remy.fouchereau@math.u-psud.fr

Patrick Pamphile
Département de Mathématiques
Université de Paris Sud France
Email: patrick.pamphile@math.u-psud.fr

Gilles Celeux
Département de Mathématiques
Université de Paris Sud France
Email: gilles.celeux@math.u-psud.fr

***Abstract*—A probabilistic model for the construction of SN-curve has been proposed. Our probabilistic model is based on a fracture mechanic approach: few parameters are required and they are easily interpreted by mechanic or material engineers. In general, fatigue test results are widely scattered for High Cycle Fatigue region and "duplex" SN-curves appear for Very High Cycle region. That's why classic models from mechanic of rupture theory on the one hand, probability theory on the other hand, do not fit SN-curve on the whole range of cycles. Our proposed model has been applied to both simulated and real fatigue test data sets. The SN-curves have been well fitted on the whole range of cycles. The parameters have been estimated using EM algorithm, combining Newton-Raphson's optimisation method and Monte-Carlo integral estimations.**

I. Introduction

A Fatigue failure occurs, when a component is subject to a repetitive stress over a long period of time. If the failed component is critical for a structure safety, then a fatal accident can result. Fatigue failures are all the more dangerous since they can occur, even, for stress not higher than the service loading. In aeronautic industry, fatigue is the most common causes of breaking for mechanic parts. It is worth mentioning that fatigue failure analysis is, also, an important point for reliability analysis and structures design in the power generation industry, the automotive industry and transportation, the construction industry, civil engineering or biomedical engineering.

Fatigue test is the main basic tool for analyzing fatigue lifetime of a given material, component, or structure. An sample of the material is subjected to cyclic loading S (stress, force, strain, etc), by a testing machine which counts N, the number of cycles to failure. Fatigue test results are plotted on a S-N curve cf. figure 1.

Fatigue tests usually take a long time, and require a large budget. The S-N curve is then obtained from a database collected over several years, with different material batch. Furthermore the fatigue phenomenon is complex and subjected to effects of various mechanical, microstructural, and environmental factors. The resulting lifetime data are then invariably scattered. Based on this, many stochastic models for fatigue lifetime prediction have been developed. In 1870, Wöhler suggests that N, the fatigue lifetime, can be expressed

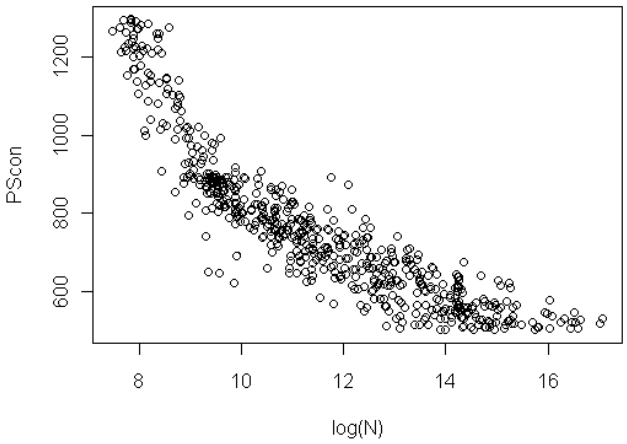


Fig. 1. SN Curve

as follow:

$$\ln(N) = (a\,S\,+\,b) + \epsilon.$$

where $a\,S\,+\,b$ is the trend fatigue lifetime relationship to the strain S, and ϵ represents the randomness of the phenomenon. Since, using regression models, a lot of works have been made to provide better fit to test results cf.[1]. A particular attention is paid to the High Cycle Fatigue region ($10^4 < N \leq 10^7$) where the results are highly scattered cf. [2],[3].

For Very High Cycle Fatigue region ($N > 10^7$), with the improve of testing means, more recent researches report that there is a "duplex S-N curve" : SN curve exhibits two different trends for VHCF region cf. [4]. So, to predict fatigue lifetime there is two different points of view: determinist models based on the fracture mechanic theory, and probabilistic models. Models from the fracture mechanic theory involve microstructure parameters which explain the SN curve duality: crack nucleation in surface/subsurface, crack growth rate for small-crack/large-crack, cf. [5] for example. Unfortunately mechanic theory based models are not very robust: in the first place, the SN-curve duality is not well understood as yet, several uncertainty factors more than material microstructure have a close relation to it cf. [6]. In the second place, material microstructure parameters involved in the model are hardly

predictable. Probabilistic methods provides mixing models cf. [7], [8] or competing risk models cf. [9],[10]. Unfortunately competing risk model does not fit the SN curve over all strain range, and mixing model gives a too conservative prediction for VHCF region.

In this paper, we propose a new probabilistic model based on a fracture mechanic approach: fatigue lifetime is usually considered as the sum of the initiation and propagation lifetimes. The initiation lifetime, N_i, may be defined as the number of cycles required to form a crack of a certain small size, that is on the order of a material grain size. The propagation lifetime, N_p, is the number of cycles required to extend the crack from that small size to the critical size at which fracture occurs. The proportion of the total fatigue lifetime taken by each of these depends mainly on material and the stress level cf. [12]. Let N be a fatigue test lifetime : either $N = N_p$, or $N = N_i + N_p$, that we have modeled by a mixing model :

$$f_N = \pi\, f_{N_p} \;+\; (1-\pi) f_{N_i+N_p},$$

where π is the probability of having a crack initiation at the first load. Thus we have is a probabilistic model for collected database of fatigue test lifetimes.

To fit the S-N curve, $(N; S)$, we have to estimate parameters involved in the model especially the probability π. Nevertheless, most of the times we don't known if a lifetime test has got a crack initiation at the first load or not. This required a fractography analysis by a Scanning Electron Microscope which is exceptionally done. Since data are incomplete, we have to use Expectation-Maximisation algorithm. The maximization part of the EM algorithm is hardly complicated in this situation, as the expected complete log-likelihood involves a complicated non-linear function of the parameters, with a non closed form integrate. We have carried out the required maximization part through the Newton-Raphson's method and the integral computation through Monte-Carlo simulations.

II. S-N curve modeling

For a fatigue test, the initiation lifetime N_i is defined as the number of cycles required to form a crack of a certain small size. The propagation lifetime, N_p, is defined as the number of cycles required to extend the crack from that small size to the critical size at which fracture occurs. Those two lifetime variables follow a log-normal distribution with a linear model for the mean relationship to strain S:

$$f_{N_i}(n,s) = \frac{1}{n\,\sigma_i\,\sqrt{2\pi}}\,\exp\left(-\frac{[\ln(n)-(\alpha_i\,s+\beta_i)]^2}{2\sigma_i^2}\right); \tag{1}$$

$$f_{N_p}(n,s) = \frac{1}{n\,\sigma_p\,\sqrt{2\pi}}\,\exp\left(-\frac{[\ln(n)-(\alpha_p\,s+\beta_p)]^2}{2\sigma_p^2}\right). \tag{2}$$

Then fatigue test lifetime N is a mixture of the propagation lifetime, N_i, and the total lifetime $N_i + N_p$. Its probability density fonction is as followed:

$$f_N = \pi\, f_{N_p} \;+\; (1-\pi) f_{N_i+N_p}, \tag{3}$$

where π is the probability of having a crack initiation at the first load.

The total lifetime, $N_{total} = N_i + N_p$, represents the "usual" fatigue comportement; whereas the first term N_p represents an "unusual" fatigue comportement which depends mainly on material microstructure comportement.

III. S-N curve fitting

To fit the S-N curve, $(N; S)$, we have had to estimate the parameters π, $\theta = (\alpha_i,\, \beta_i,\, \sigma_i,\, \alpha_p,\, \beta_p,\, \sigma_p)$. Let's consider the density

$$f_{(N,S,Z)}(n,s,z\,;\,\theta) = [f_{N_i}(n,s)]^z + [f_{N_i+N_p}(n,s)]^{(1-z)}$$

and the likelihood

$$L(N,S,Z\,;\,\theta) = \prod_k^m f_{(N,S,Z)}(n_k,s_k,z_k\,;\,\theta). \tag{4}$$

Most of the time, the indicator variable Z was not observed, therefore the complete likelihood $L(N,S,Z\,;\,\theta)$ cannot be maximized directly. Let's define the "observed" density

$$\begin{aligned} &f_{(N,S)}(n,s\,;\,\theta) = \mathbf{E}_Z[f((N,Z);\theta)] \\ &= \pi\, f_{N_p}(n,s\,;\,\alpha_p\,;\,\beta_p) \\ &+(1-\pi)\, f_{N_p+N_i}(n,s\,;\,\alpha_i\,;\,\beta_i\,;\,\alpha_p\,;\,\beta_p). \end{aligned}$$

Then the "observed" likelihood

$$L(N,S\,;\,\theta) = \prod_k^m f_{(N,S)}(n_k,s_k\,;\,\theta). \tag{5}$$

can be maximized if the marginal distribution of Z is estimated. This is the idea of the Expectation-Maximisation method cf. [13].

A. Estimation-Maximisation method

EM method is a iterative algorithm that try to maximize the log-likelihood in two step :

1) Expectation :
   Given provisional estimations $\theta^{(j)}$ and $\pi^{(j)}$, using the Bayes, the conditional distribution of Z is as followed :

$$\begin{aligned} \hat{t}_k &= \mathbf{E}[Z|N,S;\theta^{(j)}] \\ &= \mathbf{P}(Z=1|N=n_k,S=s_k\,;\,\theta=\theta^{(j)}) \\ &= \frac{\pi^{(j)} f_{N_p}(n_k,s_k\,;\,\alpha_p^{(j)};\beta_p^{(j)})}{f_{(N,S)}(n_k,s_k\,;\,\theta^{(j)})} \end{aligned}$$

   Thus, in the E-step, the expected value of the complete loglikelihood has been calculated with respect to the marginal distribution of Z, according to the current parameter $\theta^{(j)}$:

$$\begin{aligned} Q(\theta|\theta^{(j)}) &= \mathbf{E}\left[\ln L(N,S,Z\,;\,\theta)\,|\,\theta^{(j)}\right] \\ &= \sum_k^m \big[\hat{t}_k \times \ln(f_{N_p}(n_k,s_k\,;\alpha_k;\beta_k)) \\ &+ (1-\hat{t}_k) \times \ln(f_{N_i+N_p}(n_k,s_k\,;\,\theta))\big]. \end{aligned} \tag{6}$$

2) Maximisation :
The M-step consists in the maximisation of $Q(\theta|\theta^{(j)})$

$$\hat{\theta}^{(j+1)} = \arg\max_{\theta} Q(\theta|\theta^{(j)}).$$

and

$$\hat{\pi}^{(j+1)} = \frac{1}{m}\sum_{k}^{m} \hat{t}_k.$$

The EM algorithm is then continued iteratively.

The sequence $\theta^{(1)}$, $\theta^{(2)}$,... converges to a local maximum of the observed-data likelihood $L(N, S\,;\theta)$ under fairly general conditions cf. [13].

In fact we have used a Stochastic EM cf. [14] which has improved the rate of convergence.

The algorithm has been initialized with

- $\pi^{(1)} = 0.5$;
- $\theta^{(1)} = (\alpha_p^{(1)}; \beta_p^{(1)}; \alpha_i^{(1)}; \beta_i^{(1)})$ has been obtained by a clusterwise regression cf. [15]. Indeed, for Low Cycle Fatigue region(N¡10^4), the fatigue testing lifetime $N \simeq N_p$ whereas for Very High Cycle region $N \simeq N_i$.

Moreover, the maximization of $Q(\theta|\theta^{(j)})$ required a Newton-Raphson algorithm mixed with a Monte-Carlo algorithm for the evaluation of $f_{N_i+N_p}$.

B. Monte-Carlo algorithm

The convolution product

$$\begin{aligned} f_{N_i+N_p}(n) &= \int_0^n f_{N_i}(n-x) f_{N_p}(x)dx &(7)\\ &= \mathsf{E}_{N_p}[f_{N_i}(n-N_p)^+] &(8)\\ &= \mathsf{E}_{N_i}[f_{N_p}(n-N_i)^+] &(9)\end{aligned}$$

used in (6) doesn't lead to closed form. Using a Gauss-Legendre Quadrature here wouldn't be efficient because for some x values, the function under the integral is highly picked. That's why we have used a Monte-Carlo approximation in order to complete the calculation.

Monte-Carlo Algorithm : The integral (7) has been estimated using either (8), either (9), according to $\hat{\sigma}(N_p) < \hat{\sigma}(N_i)$ or not. Without loss of generality assuming that $\hat{\sigma}(N_p) < \hat{\sigma}(N_i)$ then

1) N_p sampling $(n_{p_1}, \ldots, n_{p_b})$ has been simulated ;
2) using (8), $f_{N_i+N_p}(n)$ has been estimated by :

$$\hat{f}_{N_i+N_p}(n) = \frac{1}{b}\sum_{j=1}^{b} f_{n_i}(n-n_{p_j})^+.$$

IV. Results

We have simulated $m = 600$ lifetimes from (3).

parameters	π
% error	70%

parameters	α_i	β_i	σ_i
% error	$0,3\%$	$1,7\%$	$2,5\%$

parameters	α_p	β_p	σ_p
% error	$1,7\%$	$1,8\%$	$20,0\%$

TABLE I
Parameters estimation

A. Parameters estimation

Parameters involved in (3) have been estimated by a EM-algorithm. The percent errors are given in the table I:

The π estimation has been quite hard. In first place π is the probability of a "unusual" fatigue comportement during fatigue test, then it is a small probability (less than 5%). Next, for high strain levels N_i vanishes, then N_p and $N_p + N_i$ are quite similar. The classification "unusual" comportement ($Z = 1$) or "usual" comportement ($Z = 0$), has been then difficult on this strain region (but it has been easily done for lower strain).

Things have been looking better for the other parameters cf. table I.

B. SN fitting

The fitting results are showed in figure 2.

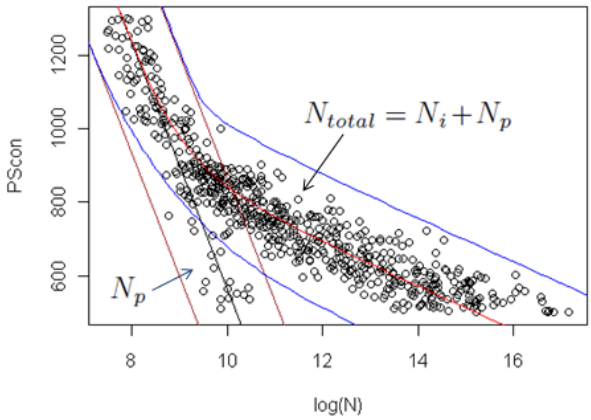


Fig. 2. SN curve fit

For each component of the mixture model (3) we have represented the trend and the 99 th and 1st percentiles of the lifetime. Thus simulated data have been well fitting. Snecma data are unavailable for publication but they have been also well fitted be our model.

V. Discussion

Fatigue tests are used to study physical and mechanical properties of the material, to design the structure or to asses the reliability. We have modeled the fatigue test lifetime by a mixture of the total lifetime $N_{total} = N_i + N_p$ representing the "usual" fatigue comportement and the propagation lifetime N_p representing an "unusual" fatigue comportement. Thus

engineer must pay attention to fatigue lifetimes under the 1st percentile of N_{total}.

Either, from fractography or production process analysis, causes of this "unusual" fatigue comportement can be identified then those data are withdrawn. Thus the total lifetime N_{total} can be used to asses the reliability or to design structure, cf. figure (3).

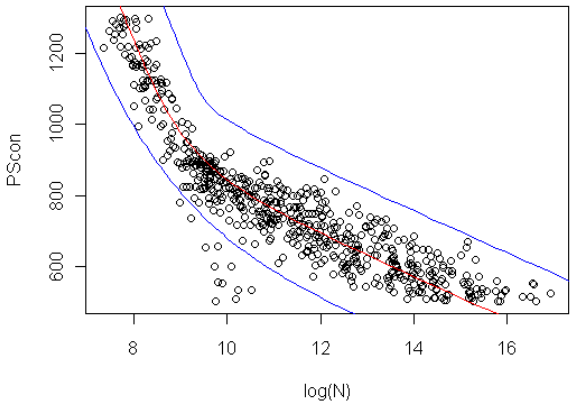


Fig. 3. SN-curve fit with $\pi = 0$, the "usual" fatigue comportement

Either the propagation lifetime N_p must be used to asses the reliability, cf. figure (4).

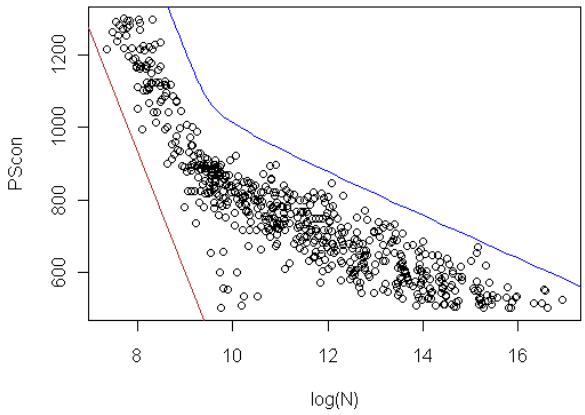


Fig. 4. SN-curve fit with $\pi > 0$, mixture with "usual and "unusual" fatigue comportement

VI. Conclusion

Fatigue test is the main basic tool for analyzing fatigue lifetime of a given material, component, or structure. In general, fatigue test results are widely scattered for High Cycle Fatigue region and "duplex" SN-curves appear for Very High Cycle region. That's why classic models from mechanic of rupture theory on the one hand, probability theory on the other hand, do not fit SN-curve on the whole range of cycles. In the paper we have proposed a probabilistic model for collected database of fatigue test lifetimes. This model is based on a fracture mechanic approach, and therefore is easily interpreted by material or mechanic engineers. Furthermore our model requires only few parameters which are estimated using EM algorithm.

Both simulated and real fatigue test data sets has been fitted with this model : not only the classic wide scattering for High Cycle Fatigue region of the SN-curve has been well fitting, but also the "duplex" phenomenon recently observed for High Cycle Fatigue region.

Then we have provided engineers with a probabilistic tool for reliability analysis of mechanic component or for structure design. But also a diagnostic tool for material elaboration.

Acknowledgment

The research was supported by SNECMA, subsidiary of the SAFRAN Group.

References

[1] F.G. Pascual and W.Q. Meeker *Estimating Fatigue Curves with the Random Fatigue-Limit Model.* 1em plus 0.5em minus 0.4emTechnometrics,1999; 41(4);277-290.

[2] T. Sakai, H. Nakayasu and I. Nishikawa *Establishment of JSMS standard regression method of SN curves for metallic materials*. Safety Reliab. Eng. Syst. Struct., 2005:380815

[3] S. Hanaki, M. Yamashita, H. Uchida and M. Zako *On stochastic evaluation of SN data based on fatigue strength distribution.* Int. J. of Fatigue, 2010;32:605609.

[4] S.K. Jha and K.S. Ravi Chandran *An unusual fatigue phenomenon: duality of the SN fatigue curve in the b-titanium alloy Ti10V2Fe3Al.* Scripta Materialia 2003;48:12071212.

[5] K. Shiozawa, M. Murai and Y. Shimatani, T. Yoshimoto *Transition of fatigue failure mode of NiCrMo low-alloy steel in very high cycle regime.* Int. J. of Fatigue, 2010; 32:541550.

[6] K.W. Chan *Role of microsture in fatigue crack initiation.* Int. J. of Fatigue, 2010;32:14281447.

[7] S.K. Jha, M.J. Caton and J.M. Larsen *A new paradigm of fatigue variability behavior and implications for life prediction.* Mater. Sci. Eng. A, 2007;468470:2332.

[8] G. McLachlan and D. Peel *Finite mixture models* Wiley 2000.

[9] K.S. Ravi Chandran, P. Chang and G.T. Cashman *Competing failure modes and complex SN curves in fatigue of structural materials* Int. J. of Fatigue, 2010;32:482491.

[10] M. Crowder *Classical Competing Risks.*, 3rd ed. Harlow, England: Addison-Wesley, 1999.

[11] W. Nelson *Applied life data analysis.* Wiley 1982.

[12] F. Alexandre *Aspects probabilistes et microstructuraux de l'amorçage des fissures de fatigue dans l'alliage INCO 718.* Thèse de doctorat. Ecole Normale Supérieure des Mines de Paris, 2004.

[13] A.P. Dempster *Maximum Likelihood from Incomplete Data via the EM Algorithm.* Journal of the Royal Statistical Society, 1977.

[14] G. Celeux, D. Chauveau, and J. Diebolt. *On Stochastic Versions of the EM Algorithm.* Technical Report 2514, INRIA, Mar, 1995.

[15] W.S. DeSarbo and W.L. Cron *A maximum likelihood methodology for clusterwise linear regression.* Int. Journal of Classification 1988;5(1):249-282.

On the use of Fleming and Harrington's test to detect late effects in clinical trials.

Valérie Garès
INSERM Unity 1027
Team Aging and Alzheimer's disease
Toulouse, France.
valerie.gares@inserm.fr

Jean-François Dupuy
INSA de Rennes
Centre de Mathématiques
Rennes, France.
jean-francois.dupuy@insa-rennes.fr

Nicolas Savy
Université Paul Sabatier
Institut de Mathématiques de Toulouse
Toulouse, France.
nicolas.savy@math.univ-toulouse.fr

***Abstract*—In this work, we deal with the question of detection of late effects in the setting of clinical trials. The most natural test for detecting this kind of effects was introduced by Fleming and Harrington [1], [2]. However, this test depends on a parameter that, in the context of clinical trials, must be chosen *a priori*. We examine the reasons why this test is adapted to the detection of late effects by studying its optimality in terms of Pitman Asymptotic Relative Efficiency. We give an explicit form of the function describing alternatives for which the test is optimal. Moreover, we will observe, by means of a simulation study, that this test is not very sensitive to the value of the parameter, which is very reassuring for its use in clinical trials.**

I. Introduction.

Neurodegenerative dementias, such as Alzheimer disease, are a growing public health concern. The global prevalence of Alzheimer disease is estimated at 115.4 million in 2050 [3]. There are currently no effective treatment for these pathologies, making its prevention a priority. Prevention is feasible due to the long asymptomatic latent period of the disease. Even an intervention that delayed disease onset by just a few years could reduce the burden of this disease on society and public health-care systems [4], [5].
To date, the rare published articles in the field of clinical trials, whose criterion of judgment is the appearance of the event "develop a dementia" are negative [6], [7], [8], [9].

The statistical design of these trials proposes to analyse data using Logrank test [10] and the conclusions are the non-significance of the treatment. Logrank test permits to test the equality of survival functions of censored failure time data. It is known to be the most powerful under the proportional hazards model [11] and thus it is not appropriate to prevention clinical trials. Indeed, we know that preventive treatments suppose an impregnation more or less long until noticing a effect. The proportional hazards assumption is thus not realistic.

In order to overpass this problem, we propose to use weighted Logrank tests. These families of tests are constructed from the Logrank test by plugging in the statistic a weight $(W_n(s),\ s \in \mathbb{R}^+)_{n\in\mathbb{N}}$. The choice of this weight is motivated by the deviations from equality that we are interested in detecting. These tests have been widely studied. Some examples are Gehan's test [12] where $W_n(s) = Y_n(s)$, Peto-Peto's test [13] where $W_n(s) = \tilde{S}(s)$ (with $\tilde{S}$ denoting a modified Kaplan-Meier estimator of S obtained from pooled samples), Tarone and Ware's test [14] where $W_n(s) = \sqrt{Y_n(s)}$, Peto-Prentice's test [15] where $W_n(s) = \widehat{S}(s)$ and Fleming-Harrington's test [2] where $W_n(s) = \widehat{S}(s)^p$. Most of these tests are constructed in order to detect early differences. Late effects have attracted little attention, until the extension of Fleming-Harrington's test [1] where

$$W_n^{p,q}(s) = \widehat{S}(s)^p[1-\widehat{S}(s)]^q,$$

with $p, q \geqslant 0$ and $\widehat{S}$ the Kaplan-Meier estimator of S.

The application of Fleming-Harrington's test to clinical trials is not immediate. In fact, if we focus on late effects, we do not handle a test but a family of tests indexed by the parameter q. In the framework of clinical trials, this parameter has to be fixed *a priori* in the statistical analysis design. The link between q and the characteristics of the trial is not obvious. We are only able to give a rough estimation of this value. But as we will see in this paper, a rough estimation is enough because the test is little sensitive to the choice of this parameter.

The paper is organized as follows. First, we introduce some notations and we review how we can extend the Logrank test to the so-called $\mathcal{K}$-class of weighted tests. Then, using the notion of Asymptotic Relative Efficiency (ARE) introduced by Pitman, we compare the tests of the class $\mathcal{K}$. The ARE enables us to make links between alternative hypotheses and weighted tests of the class $\mathcal{K}$. Finally, we apply these results to the Fleming-Harrington's test, and we exhibit a function for which this test is optimal under Shift Alternative Assumptions. This function allows us to make simulation studies, where we investigate the performance of the test, and to study the sensitivity of the test to the choice of the parameter.

II. Weighted Logrank test.

A. Notations.

Let T be a nonnegative random variable. T denotes the duration between the origin date and the time of occurrence of some specific event. Its cumulative distribution function is

denoted by F, its survival function by S, its hazard function by λ, and its cumulative hazard function by $\Lambda(t) = \int_0^t \lambda(s)ds$. T is assumed to be right-censored that is, we only observe events which happen before a certain date C: $(T < C)$. The i-th subject has latent survival time and censoring times T^i and C^i, respectively. The distribution function of the censoring times $(C^i)_{i=1,\dots,n}$ is G. We assume that C^i and T^i are independent. The observations consist of $(X^i, \delta^i)_{i=1\dots n}$ where $X^i = T^i \wedge C^i$ and $\delta^i = \mathbb{I}_{\{T^i \leqslant C^i\}}$. τ denotes the length of the study, from the origin. Let $\tau^{'} = \inf_{t \geqslant 0} \{(1 - F(t))\,(1 - G(t)) = 0\}$. We assume $\tau < \tau^{'}$. Let us define the random variables

$$N_n(t) = \sum_{i=1}^{n} \mathbb{I}_{\{X^i \leqslant t, \delta^i = 1\}}, \qquad Y_n(t) = \sum_{i=1}^{n} \mathbb{I}_{\{X^i \geqslant t\}}.$$

$N_n(t)$ is the number of failures at t and $Y_n(t)$ the number of subjects at risk at t^-. Finally, define

$$J_n(t) = \mathbb{I}_{\{Y_n(t) > 0\}}.$$

Consider a clinical trial with two arms, where n_T patients receive a drug (or treatment) and n_P patients receive a placebo. The duration of the follow-up is fixed. We note $n = n_P + n_T$ (resp. $N_n = N^P_{n_P} + N^T_{n_T}$, resp. $Y_n = Y^P_{n_P} + Y^T_{n_T}$) the variables corresponding to the pooled samples.

B. From the Logrank test to the tests of class $\mathcal{K}$.

In order to test the following hypotheses:

$$\begin{cases} \mathcal{H}_0 & : \ F^P = F^T = F, \\ \mathcal{H}_1 & : \ F^P \neq F^T, \end{cases}$$

we usually use the Logrank test, whose statistic at time t can be written as:

$$LR(t) = \int_0^t \left(\frac{n_P + n_T}{n_P n_T}\right)^{1/2} \frac{Y^P_{n_P}(s) Y^T_{n_T}(s) J_n(s)}{Y_n(s)} \left[\frac{dN^P_{n_P}(s)}{Y^P_{n_P}(s)} - \frac{dN^T_{n_T}(s)}{Y^T_{n_T}(s)}\right].$$

The asymptotic behavior (as n tends to infinity) of this test can be investigated in a functional point of view.

Assumptions 2.1: There exists a^P in $]0,1[$ and $a^T = 1 - a^P$ such that :

$$\frac{n_P}{n} \xrightarrow[n\to\infty]{} a^P \qquad \text{and} \qquad \frac{n_T}{n} \xrightarrow[n\to\infty]{} a^T.$$

In the following, we assume this assumption fulfilled.

Remark 2.1: When n tends to $+\infty$, n_T tends to $+\infty$ and n_P tends to $+\infty$.

Remark 2.2: In a clinical trial with two regular arms, Assumption 2.1 is fulfilled for $a^P = \frac{1}{2}$ and there exists a constant c such that:

$$\left(\frac{n_P + n_T}{n_P n_T}\right)^{1/2} \underset{n\to\infty}{\sim} \frac{c}{\sqrt{n}}.$$

This point is crucial in the proofs of the theorems below.

Lemma 2.1: For $i = P, T$, given $\pi^i = (1 - F^i)(1 - G^i)$, we have the convergence:

$$\sup_{t \in \mathbb{R}^+} \left| \frac{Y^i_{n_i}}{n_i}(t) - \pi^i(t) \right| \xrightarrow[n\to\infty]{p.s.} 0.$$

Theorem 2.1: Under $\mathcal{H}_0$,

$$LR \xrightarrow[n\to\infty]{\mathcal{L}(\mathbb{D})} \mathbb{G},$$

where $\mathbb{G}$ is a Gaussian process with mean 0 and variance function:

$$t \to \sigma^2_{\mathbb{G}}(t) = \int_0^t \frac{\pi^P(s)\pi^T(s)}{a^P \pi^P(s) + a^T \pi^T(s)} (1 - \Delta\Lambda(s)) d\Lambda(s).$$

Note that this convergence is functional that is, it should be understood as a convergence in distribution in the space $\mathbb{D}$ of càdlàg functions.

As discussed in the Introduction, the Logrank test is known to be the most powerful for constant effects (proportional hazards assumption). One may thus wonder what the properties of this test are, under late effects. In order to answer this question, weighted Logrank tests are introduced, by plugging a weight W_n in the Logrank statistic, so as to give more importance to certain values. The resulting statistic writes, at time t:

$$LR_{W_n}(t) = \int_0^t W_n(s) \left(\frac{n_P + n_T}{n_P n_T}\right)^{1/2} \frac{Y^P_{n_P}(s) Y^T_{n_T}(s) J_n(s)}{Y_n(s)} \left[\frac{dN^P_{n_P}(s)}{Y^P_{n_P}(s)} - \frac{dN^T_{n_T}(s)}{Y^T_{n_T}(s)}\right].$$

The key point we have to investigate is the asymptotic behavior of this statistic. The so-called $\mathcal{K}$ class of tests is defined as follows:

Definition 2.1 ([16]): A weighted Logrank statistic LR_{W_n} is in class $\mathcal{K}$ if W_n is an adapted bounded nonnegative predictable process.

C. Asymptotic behavior of class-$\mathcal{K}$ tests.

Let $\{F_\theta : \theta \in \Theta\}$ be a family of continuous cumulative distribution functions on $[0, \infty)$ indexed by a parameter $\theta \in \Theta$, and consider the following hypotheses:

$$\begin{cases} \mathcal{H}_0 & : \ F^T = F^P = F_{\theta_0}, \\ \mathcal{H}_1 & : \ F^T = F_{\theta^T_{n_T}} \quad \text{and} \quad F^P = F_{\theta^P_{n_P}}, \end{cases} \tag{1}$$

where for $i = T, P$, $(\theta^i_{n_i})$ is a sequence of Θ such that

$$\theta^i_{n_i} \xrightarrow[n\to\infty]{} \theta_0.$$

In order to derive the asymptotic behavior of the weighted Logrank tests, we make the following assumptions:

Assumptions 2.2: There exists a function $w \in \mathbb{D}$ such that:

$$W_n(s) \xrightarrow[n\to\infty]{p.s.} w(s).$$

Assumptions 2.3: For $i = P, T$, there exists a function $\gamma^i \in \mathbb{D}$ such that:

$$\sqrt{\frac{n_T n_P}{n}} \left(\frac{d\Lambda_{\theta^i_{n_i}}(s)}{d\Lambda_{\theta_0}(s)} - 1\right) \xrightarrow[n\to\infty]{p.s.} \gamma^i(s).$$

The main result for the weighted statistics of class $\mathcal{K}$ is:

Theorem 2.2: Suppose that λ is continuous in θ_0. Let LR_{W_n} be a statistic in the class $\mathcal{K}$, satisfying the Assumptions 2.2 and 2.3. Then, under $\mathcal{H}_1$,

$$LR_{W_n} \xrightarrow[n\to\infty]{\mathcal{L}(\mathbb{D})} \mathbb{G},$$

where $\mathbb{G}$ is a Gaussian process with mean function

$$\mu_{\mathbb{G}} : t \to \int_0^t k(s)(\gamma^P - \gamma^T)(s) d\Lambda_{\theta_0}(s),$$

with

$$k(s) = w(s) \frac{\pi^P(s)\pi^T(s)}{a^P\pi^P(s) + a^T\pi^T(s)}, \tag{2}$$

and variance function

$$\sigma^2_{\mathbb{G}} : t \to \int_0^t \frac{a^P\pi^P(s) + a^T\pi^T(s)}{\pi^P(s)\pi^T(s)} k^2(s)(1-\Delta\Lambda_{\theta_0}(s)) d\Lambda_{\theta_0}(s).$$

Corollary 2.1: Suppose that λ is continuous in θ_0. Let LR_{W_n} be a statistic in the class $\mathcal{K}$, satisfying the Assumption 2.2. Then, under $\mathcal{H}_0$,

$$LR_{W_n} \xrightarrow[n\to\infty]{\mathcal{L}(\mathbb{D})} \mathbb{G}',$$

where $\mathbb{G}'$ is a Gaussian process with mean 0 and variance function $\sigma^2_{\mathbb{G}}$.

III. Comparison of class-$\mathcal{K}$ tests.

A. Consistency and Pitman Asymptotic Relative Efficiency.

Under some conditions, weighted logrank tests can be consistent against the ordered hazard alternative or the alternative of stochastic ordering [1], [16]. Thus we cannot compare these tests by the limits of their statistics, nor by the limit of their power since, in any case, it converges to 1 as n tends to infinity. In this framework, a good procedure to compare tests is to investigate the behavior of the corresponding tests under a sequence of alternatives converging to the null hypothesis as $n \to \infty$. This is the so-called Asymptotic Relative Efficiency [17] (ARE for short) procedure. Here, we use the Pitman's ARE, which is defined as follows:

Definition 3.1: Consider two tests denoted by $\mathcal{T}$ and $\mathcal{T}'$. Consider a sequence of testing problems consisting in testing a null hypothesis $\mathcal{H}_0 : \theta = \theta_0$, against the alternative $\mathcal{H}_1 : \theta = \theta_*$, where θ_* tends to θ_0. Given a level α and a power $1-\beta$, let $N(\alpha, \beta, \theta_*)$ (resp. $N'(\alpha, \beta, \theta_*)$) be the minimal number of observations needed to reach the level α and the power $1-\beta$, for test $\mathcal{T}$ (resp. $\mathcal{T}$') . Then, if it exists,

$$\lim_{\theta_* \to \theta_0} \frac{N'(\alpha, \beta, \theta_*)}{N(\alpha, \beta, \theta_*)}$$

is called the Pitman's Asymptotic Relative Efficiency of $\mathcal{T}$ with respect to $\mathcal{T}'$.

B. ARE of class-$\mathcal{K}$ tests.

In order to calculate the ARE of class-$\mathcal{K}$ tests for the hypotheses (1), we assume that, for a constant c,

$$\theta^P_{n_P} = \theta_0 + c\left(\frac{n_T}{n_P(n_P+n_T)}\right)^{1/2}, \tag{3}$$
$$\theta^T_{n_T} = \theta_0 - c\left(\frac{n_P}{n_T(n_P+n_T)}\right)^{1/2}.$$

Theorem 3.1 ([16]): Let $LR_{W^1_n}$ and $LR_{W^2_n}$ be two statistics in the class $\mathcal{K}$, satisfying Assumptions 2.2 and 2.3. Moreover, assume that λ is differentiable in θ_0. Given sequences of alternatives defined by (1) with $\theta^P_{n_P}$ and $\theta^T_{n_T}$ defined by (3), we call Asymptotic Efficiency of $LR_{W^j_n}$ the quantity

$$\frac{|\int_0^\infty k^j(s)(\gamma^P - \gamma^T)(s) d\Lambda_{\theta_0}(s)|}{\left(\int_0^\infty (k^j)^2(s) \frac{a^P\pi^P(s)+a^T\pi^T(s)}{\pi^P(s)\pi^T(s)}(1-\Delta\Lambda_{\theta_0}(s)) d\Lambda_{\theta_0}(s)\right)^{1/2}},$$

which will be denoted by $AE(LR_{W^j_n})$ in the sequel. Then the Pitman Asymptotic Relative Efficiency of $LR_{W^1_n}$ with respect to $LR_{W^2_n}$ is given by:

$$ARE(LR_{W^1_n}, LR_{W^2_n}) = \frac{AE(LR_{W^2_n})}{AE(LR_{W^1_n})}.$$

Theorem 3.2 ([16]): If λ is differentiable in θ_0, the statistic in class $\mathcal{K}$ with maximal efficiency has a limit weight function w such that k defined by (2) is expressed by:

$$s \to \kappa \frac{\lambda'(\theta_0)}{\lambda(\theta_0)}(s) \left(\frac{\pi^P(s)\pi^T(s)}{a^P\pi^P(s) + a^T\pi^T(s)}\right) \frac{1}{1-\Delta\Lambda_{\theta_0}(s)}$$

where κ is a constant. Moreover, we have

$$\gamma^P - \gamma^T = c\frac{\lambda'(\theta_0)}{\lambda(\theta_0)}.$$

C. Particular case of Shift Assumptions.

In the framework of the Logrank test, a useful strategy is to consider a particular pattern of the alternative hypotheses, called: Shift Assumptions up to a change of time. This can be defined through the following family of distribution functions:

$$F_\theta(t) = \Psi(g(t) + \theta), \qquad \theta \in \Theta, \tag{4}$$

where g is a differentiable nondecreasing function from $[0, \infty[$ to $]-\infty, u^+[$, with $u^+ \in \overline{\mathbb{R}}$, and Ψ is a continuous cumulative distribution function with value in $[0,1]$, with positive density Ψ', having an almost everywhere continuous second derivative Ψ''. By using the Shift Assumptions, we prove that the Logrank test is the most powerful test for an alternative proportional hazards hypothesis.

Remark 3.1: According to the definition of F_θ, F is continuous so λ is differentiable in θ_0. We suppose that the Assumptions 2.2 and 2.3 are fulfilled.

Theorem 3.3: Consider the parametric family (4). The limit weights of the statistic in the class $\mathcal{K}$ for which the asymptotic

efficiency is maximal to test the hypothesis given by (1) are proportional and verify, for all $t \in \mathbb{R}^+$:

$$w(t) = L'[\Psi] \circ \Psi^{-1} \circ F_{\theta_0}(t), \qquad \text{where } L(\Psi) = \ln\left(\frac{\Psi'}{1-\Psi}\right).$$

Remark 3.2: We can use this theorem in two ways: given a weight, we can find alternatives for which the test is optimal, and given an hypotheses pattern, we can find the weight for which the test is optimal.

Corollary 3.1: Consider $W_n(s) = W(\widehat{S}(s))$ for all $s \geqslant 0$, where W is a continuous nonnegative function on $[0,1]$. The statistic LR_{W_n} has a maximal efficiency to test

$$\begin{cases} \mathcal{H}_0 & : \ F^T = F^P = F_{\theta_0}, \\ \mathcal{H}_1 & : \ F^T = \Psi(g + \theta^T_{n_T}) \text{ and } F^P = \Psi(g + \theta^P_{n_P}) \end{cases} \tag{5}$$

with $(\theta^i_{n_i})$ for $i = T, P$ given by (3) and

$$\Psi(u) = 1 - (\mathcal{L})^{-1}(u + c)$$

where c is a constant, and $\mathcal{L}$ is a one-to-one map from $]0,1[$ to $\mathbb{R}^-$, defined as the primitive of the function defined from $]0,1[$ to $\mathbb{R}^-$ by:

$$x \to \frac{1}{xL(x)} \qquad \text{with} \quad L(x) = -\int_x^1 \frac{W(s)}{s} ds.$$

IV. Application to Fleming and Harrington's test.

To simplify, we denote $LR^{p,q}_n$ instead of $LR_{W^{p,q}_n}$.

A. Optimal hypotheses under Shift Assumptions.

An application of Corollary 3.1 with $W(s) = s^p(1-s)^q$ allows us to determine the function Ψ for which Fleming Harrington's test is optimal in the ARE sense. The main result is thus:

Theorem 4.1: Fleming and Harrington's statistic $LR^{p,q}_n$ has maximal efficiency in the ARE's sense to test hypotheses (5) with

$$\Psi(u) = \Psi^{p,q}(u) = \begin{cases} 1 - \exp(-e^{u+c}) & \text{if } p = 0, \ q = 0, \\ 1 - (1 + pe^{u+c})^{-\frac{1}{p}} & \text{if } p > 0, \ q = 0, \\ 1 - (\mathcal{L}^{p,q})^{-1}(u + c) & \text{if } p > 0, \ q > 0, \end{cases}$$

where $u = g(t) + \theta_0$, c is some constant, $\mathcal{L}^{p,q}$ is constructed as in Theorem 3.1 from $L^{p,q}(x) = -B_{inc}(x-1, q+1, p)$, and B_{inc} is the incomplete beta function given by:

$$B_{inc}(x, a, b) = \int_0^x s^{a-1}(1-s)^{b-1} ds.$$

In the sequel, the function $\Psi^{0,q}$ will be noted by Ψ^q to ighten the notations.

Remark 4.1: It is important to notice that:

- $\mathcal{L}^q$ tends to minus infinity when x tends to 1, moreover we have:
$$\mathcal{L}^q(x) \underset{x \to 1}{\sim} -\frac{1}{(1-x)^q}.$$
- $\mathcal{L}^q$ tends to infinity when x tends to 0, moreover we have:
$$\mathcal{L}^q(x) \underset{x \to 0}{\sim} \ln(-\ln(x)).$$

By a relevant choice of g, it is possible to write the link between the risk of each group up to the shift $\Delta = \theta^P - \theta^T$. This leads us to the following Proposition.

Proposition 4.1: Fix $\Delta = \theta^P - \theta^T$.

- For $p = 0$, testing the hypotheses (5) is equivalent to test
$$\begin{cases} \mathcal{H}_0 & : \ \lambda^T = \lambda^P, \\ \mathcal{H}_1 & : \ \lambda^T = \lambda^P \ e^{\Delta} \end{cases}.$$
- For $p > 0$, testing the hypotheses (5) is equivalent to test
$$\begin{cases} \mathcal{H}_0 & : \ \lambda^T = \lambda^P, \\ \mathcal{H}_1 & : \ \lambda^T = \lambda^P \ e^{\Delta}[(S^P)^p + [1 - (S^P)^p]e^{\Delta}]^{-1} \end{cases}.$$
- For $p > 0, q > 0$, testing the hypotheses (5) is equivalent to test
$$\begin{cases} \mathcal{H}_0 & : \ \lambda^T = \lambda^P, \\ \mathcal{H}_1 & : \ \lambda^T = \lambda^P \ \Gamma^{p,q}(., \Delta), \end{cases}$$
where for any $t \in \mathbb{R}^+$,
$$\Gamma^{p,q}(t, \Delta) = \frac{L^{p,q}((\mathcal{L}^{p,q})^{-1}(\mathcal{L}^{p,q}(S^P(t)) + \Delta))}{L^{p,q}(S^P(t))}.$$

Remark 4.2: The first point of the Proposition is nothing but the proof of optimality of the Logrank test under the proportionality of the risks.

B. Performances and Sensitivity of the test.

Consider the case where $p \geqslant 0$ and $q = 0$. As pointed in [2], it is easier to compare a risk function to a constant rather than two risk functions. So, we choose the placebo survival function to be an exponential with parameter a, while the risk function of the treatment group takes the form:

$$\lambda^T(t) = e^{\Delta}(e^{-pat} + (1 - e^{-pat})e^{\Delta})^{-1}.$$

On Figure 1, we plot the risk (on the left) and survival functions (on the right) for the placebo group and for the treatment group, for different values of p (0, 1, 2, 3 and 4). In this example, we choose τ equal to 5 years, a proportion of censorship equal to 0.5, and a rate, defined by $r = \left(\frac{S^T(\tau) - S^P(\tau)}{1 - S^P(\tau)}\right)$ equal to 0.05% after τ years of follow up.

Figure 1 shows us the pattern of the survival functions of placebo and treatment groups for which the Fleming-Harrington's test is optimal in the sense of ARE. We clearly see that it detects proportional hazards for $p = 0$ and early effects for $p > 0$. Moreover, we can notice that as expected, the larger p is, the earlier the treatment effect is detected.

Consider now the case where $q \geqslant 0$ and $p = 0$. This is the setting we are interested in. Choosing the placebo survival function to be an exponential with parameter a, the risk function of the treatment group becomes:

$$\lambda^T(t) = a \ \frac{L^q((\mathcal{L}^q)^{-1}(\mathcal{L}^q(e^{-at}) + \Delta))}{L^q(e^{-at})}. \tag{6}$$

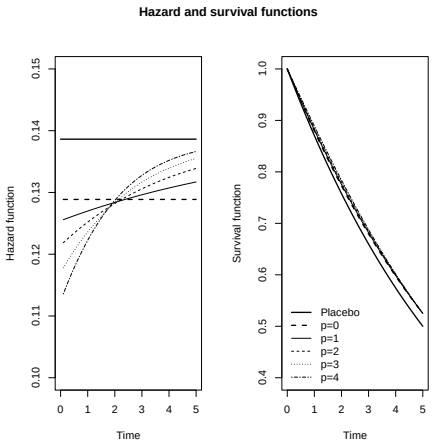


Fig. 1: Hazard and survival functions for $q = 0$, $p \geqslant 0$.

The situation is harder here because of the functions L^q and $\mathcal{L}^q$, which are not easy to handle and necessitate the use of numerical integrations. On Figure 2, we plot the risk (on the left) and survival functions (on the right) for the placebo group and for the treatment group, for different values of q (0, 1, 2, 3 and 4). In this example, we choose τ equal to 5 years, a proportion of censorship equal to 0.5, and a rate equal to 20% after τ years of follow up.

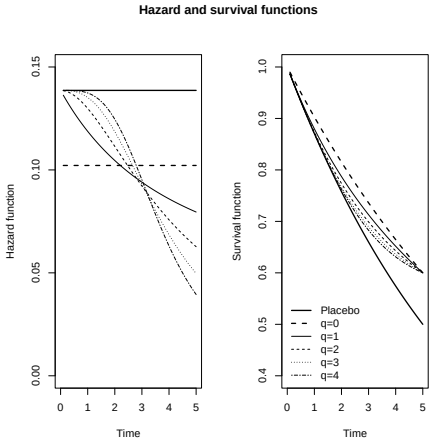


Fig. 2: Hazard and survival functions for $p = 0$, $q \geqslant 0$.

Figure 2 shows the alternatives when Fleming and Harrington's test with $q = 0$ is optimal. We clearly see that it detects proportional hazards for $q = 0$ and late effects for $q > 0$. Moreover, we can notice that, as expected, the more q is large, later the treatment effects is detected.

We now focus on the setting $q \geqslant 0$ and $p = 0$. The knowledge of the function Ψ^q allows us to perform Monte Carlo simulations in order to test the performance of the test and its sensitivity to the choice of q.
To do so, we perform 2000 Monte Carlo simulations for each scenario. Data for the placebo group are simulated from an exponential distribution whose parameter is fixed by the probability of censoring, and data for the treatment group are simulated by the use of (6) with $q = 3$. In order to investigate the importance of the sample size, we consider scenarios with $n^P = n^T = 100,\ 500,\ 1000\ or\ 2000$. To investigate the role of the censorship, we perform scenarios with a proportion of censored data equal to 20%, 50% and 80%. We choose a rate equals to 5%, 10%, 20% or 30%. Δ in the expression of $\Gamma^{p,q}(t, \Delta)$ is given by $\Delta = \theta^P - \theta^T = \mathcal{L}(\Delta S(1 - S^P(\tau)) + S^P(\tau)) - \mathcal{L}(S^P(\tau))$.
In each scenario, we evaluate the empirical level and power for both the Logrank and Fleming-Harrington's tests, for different parameters q around 3 (2, 3, 4). Table I gives an example of the results for $n^P = 2000$.

n	c	ΔS	Logrank	$q = 2$	$q = 3$	$q = 4$
2000	0.2	0.05	0.3500	0.6470	**0.6620**	0.6605
		0.1	0.8775	0.9990	**0.9990**	0.9990
		0.2	1.0000	1.0000	**1.0000**	1.0000
		0.3	1.0000	1.0000	**1.0000**	1.0000
	0.5	0.05	0.1595	0.2925	**0.3010**	0.2950
		0.1	0.4545	0.7915	**0.8025**	0.7955
		0.2	0.9700	1.0000	**1.0000**	1.0000
		0.3	0.9995	1.0000	**1.0000**	1.0000
	0.8	0.05	0.0890	0.1470	**0.1490**	0.1460
		0.1	0.1875	0.3815	**0.3960**	0.3900
		0.2	0.6225	0.9200	**0.9260**	0.9240
		0.3	0.9340	1.0000	**1.0000**	1.0000

TABLE I: Empirical powers of Fleming and Harrington's tests for different parameters q from data generated.

As expected, we observe that the power increases with n and r, and decreases when censoring increases. In each scenario, the power of the Logrank test is clearly lower than those of Fleming-Harrington's test, which confirms what we have proved: Fleming-Harrington's test is better than the Logrank for detecting late effects. It is worth noting that the Fleming-Harrington's test has a maximal power when q is chosen equal to 3 (recall that $q = 3$ was used to simulate the data), but we observe a very low variation of the power for a variation of this parameter. This means that the sensitivity of this test to the value of the parameter q is very low. Moreover, let us notice that this sensitivity decreases with n. Thus, an error on the choice of q will have a limited impact on the result of the test, which is very reassuring for its application in clinical trials.

C. Recommendation for the choice of q: use of inflexion point.

As we have already pointed out, the choice of q in the Fleming-Harrington's test is mandatory in clinical trials. Unfortunately, there are no simple links between the setting of the trials and this value. The result of the previous section allows us to give a rough estimation of q without risks on the quality of the results. To do so, we make use of the following proposition:

Proposition 4.2: For $p > 0$ and $q = 0$,

$$t^* = \frac{\ln(e^{-\Delta} - 1)}{ap} \tag{7}$$

is an inflexion point of the hazard function of the treatment group.

This point is of interest because before time t^*, the differences between treatment and placebo are emphasized (perhaps too much for small values of p), and after time t^*, these differences are diminished (perhaps too much for large values of p). Thus the values around t^* are the values which collect the best information. In most clinical trials, we are able (from previous studies for instance) to determine a rough estimation of t^*, and the formula (7) gives a rough estimation of the value of p.

In the case of interest where $p = 0$ and $q > 0$, the same arguments hold (up to a swap of "emphasized" and "diminished"). But although the second derivative of the hazard function of the treatment group is calculable, it is not possible to explicitly determine the value of t^*. However, the formula is closed and this value can be calculated numerically.

REFERENCES

[1] T. R. Fleming and D. P. Harrington, *Counting processes and survival analysis*, ser. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. New York: John Wiley & Sons Inc., 1991.

[2] D. P. Harrington and T. R. Fleming, "A class of rank test procedures for censored survival data," *Biometrika*, vol. 69, no. 3, pp. 553–566, 1982. [Online]. Available: http://dx.doi.org/10.1093/biomet/69.3.553

[3] A. Wimo and M. Prince, "World Alzheimer report 2010," Tech. Rep., 2010.

[4] R. Brookmeyer, "Forecasting the global burden of alzheimer's disease," *Alzheimers Dement*, vol. 3(3), pp. 186–91, 2007.

[5] S. G. Brookmeyer, R. and C. Kawas, "Projections of alzheimer's disease in the united states and the public health impact of delaying disease onset." *Am J Public Health*, vol. 88(9), pp. 1337–42, 1998.

[6] S. DeKosky, "Ginkgo biloba for prevention of dementia: a randomized controlled trial," *JAMA*, vol. 300(19), pp. 2253–2262, 2008.

[7] C. Lyketsos, "Naproxen and celecoxib do not prevent ad in early results from a randomized controlled trial," *Neurology*, vol. 68(21), pp. 1800–1808, 2007.

[8] S. Shumaker, "Estrogen plus progestin and the incidence of dementia and mild cognitive impairment in postmenopausal women: the women's health initiative memory study: a randomized controlled trial," *JAMA*, vol. 289(20), pp. 2651–2662, 2003.

[9] ——, "Conjugated equine estrogens and incidence of probable dementia and mild cognitive impairment in postmenopausal women: Women's health initiative memory study," *JAMA*, vol. 291(24), pp. 2947–2958, 2004.

[10] Mantel and Haenszel, "Statistical aspects of the analysis of data from retrospective studies of disease," *J. Nat. Cancer Inst*, vol. 22, pp. 719–748, 1959.

[11] D. R. Cox, "Regression models and life-tables," *J. Roy. Statist. Soc. Ser. B*, vol. 34, pp. 187–220, 1972, with discussion by F. Downton, Richard Peto, D. J. Bartholomew, D. V. Lindley, P. W. Glassborow, D. E. Barton, Susannah Howard, B. Benjamin, John J. Gart, L. D. Meshalkin, A. R. Kagan, M. Zelen, R. E. Barlow, Jack Kalbfleisch, R. L. Prentice and Norman Breslow, and a reply by D. R. Cox.

[12] E. A. Gehan, "A generalized Wilcoxon test for comparing arbitrarily singly-censored samples," *Biometrika*, vol. 52, pp. 203–223, 1965.

[13] R. Peto and J. Peto, "Asymptotically efficient rank invariant test procedures," *J. Roy. Statist. Soc.*, vol. A(135), pp. 185–206, 1972.

[14] R. E. Tarone and J. Ware, "On distribution-free tests for equality of survival distributions," *Biometrika*, vol. 64, no. 1, pp. 156–160, 1977.

[15] R. L. Prentice, "Linear rank tests with right censored data," *Biometrika*, vol. 65, no. 1, pp. 167–179, 1978. [Online]. Available: http://dx.doi.org/10.1093/biomet/65.1.167

[16] R. D. Gill, "Censoring and stochastic integrals," vol. 124, pp. v+178, 1980.

[17] A. W. van der Vaart, *Asymptotic statistics*, ser. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge: Cambridge University Press, 1998, vol. 3.

A Universal Goodness of Fit Test based on Regression Techniques

Florence George and Sneh Gulati
Department of Mathematics and Statistics
Florida International University
Miami, FL 33199, USA

Abstract— **Model fitting to describe real world phenomena is an important part of statistical research. However, model fitting cannot occur without assessing the fit of the model. Gulati and Shapiro [1] and Gulati [2] developed goodness of fit tests for the Pareto, the Rayleigh and the Laplace distribution based on the regression test of Brain and Shapiro [3]. This paper extends the results of [1] and [2] to develop a universal goodness of fit test for all distributions by transforming the data to the exponential distribution and applying the techniques of Brain and Shapiro [3]. The power of the procedure is tested via simulations for the Uniform, Gumbel and the Weibull distributions.**

Keywords and Phrases; Goodness of Fit; exponential Distribution; regression Tests; probability integral transform

I. INTRODUCTION

Model fitting is an important part of all statistical research. However, model fitting must be accompanied by an assessment of the fit of the model. The development of the first numerical techniques to test goodness of fit can probably be attributed to Karl Pearson who developed the well-known chi-squared test. The chi-squared test groups a given data set into intervals and then compares the observed cell counts to the expected cell counts. Chi-squared goodness-of-fit tests were followed by test of the EDF type; the well-known Kolmogrov-Smirnov test and the Cramér-von Mises statistic being part of this family. EDF tests are based on measures of distance between the empirical cumulative distribution function of the data and the hypothesized cumulative distribution function. Lately though, there has been an explosion in the number of procedures available to test distributional assumptions; for example tests based on regression and correlation, tests based on the characterizing properties of specific distributions, test based on empirical Laplace transforms etc. (see D'Agastino and Stephens [4] for more details).

This paper presents an omnibus test to test any distributional assumption and is based on the regression test of Brain and Shapiro [3] to test for an underlying exponential distribution. The techniques of Brain and Shapiro [3] were used by Gulati and Shapiro [1] and Gulati [2] to develop goodness of fit procedures for the Pareto, Rayleigh and the Laplace distributions. The data were transformed to an exponential random sample using simple transformations and then tested for exponentiality using the Brain and Shapiro test. This paper attempts to extend the results [1] and [2] to develop a universal test for all univariate (belonging to the scale family of distributions) or bivariate distributions (characterized by a scale and a shape parameter) based on the Brain and Shapiro test. The data are transformed to an exponential distribution either directly or indirectly through the probability integral transform and then tested for exponentiality.

Certainly, both the chi-square type tests and EDF tests are universal tests and can be used to test any for underlying family. The main drawback of the chi-squared test is its dependency on the grouping intervals and the fact that it tends to be conservative. Tests based on the EDF statistics are more powerful but when the null hypothesis is composite the exact distribution of the test statistic becomes very difficult to find. For the tests of the Anderson darling type, asymptotic theory is available however; the computation of the null percentiles still needs to be done separately for each family being tested. For the K-S type statistics, even asymptotic theory is not available and the percentiles need to be computed (see [4]). In comparison, the test statistic proposed in this paper is extremely simple to compute and has an asymptotic chi square distribution under the null. Simulation studies indicate that the null distribution of the proposed test statistic for hypothesized distributions with a single parameter is a chi squared two degree distribution (exponential) for sample sizes as small as 10, while for distributions with both a scale and a shape parameter, the test has a chi-square distribution with one degree of freedom. Thus a p value can be obtained directly from a chi-squared table and there is no need for extensive computations for the calculation of the null percentiles.

The rest of the paper is organized as follows. Section two discusses the procedure proposed by Brain and Shapiro [3], while the third section describes the application of this test to any underlying distribution. Sections four and five are devoted to simulation studies to assess the null distribution of the statistic and corresponding power studies for the uniform, Weibull and the Gumbel distribution. Section 6 contains some concluding remarks.

II. BRAIN AND SHAPIRO PROCEDURE FOR THE EXPONENTIAL DISTRIBUTION

Brain and Shapiro [1] developed a test for exponentiality based on the Laplace Test, suggested by Cox and Lewis (Cox [5], page 53). The procedure determines exponentiality of the given data by testing whether the underlying failure rate function is constant and is based on the fact that the ordered weighted spacings of i.i.d. exponential random variables are i.i.d. exponential. The test statistic is a combination of two test statistics: one tests for a monotone hazard rate and is a function of the linear regression of the ordered weighted spacings vs. the order number. The second test statistic tests for a non-monotone and a non-constant hazard rate and is a function of a quadratic regression of the order number vs. the weighted ordered spacings. Under the null hypothesis the two statistics are distributed as independent normal variates. Since the final test statistic is obtained by adding the squares of the two statistics its null distribution is that of a chi-squared random variable with two degrees of freedom. The test is summarized below:

1. Let n be the number of observations and denote the ordered observations as $X_{(1)} \leq X_{(2)} \leq \ldots\ldots \leq X_{(n)}$.

2. The null hypothesis of interest is :

H_0: The data come from an exponential distribution.

3. Let the ith weighted spacing be given by $Y_i = (n - i + 1)*(X_i - X_{(i-1)})$ $(i=1,2,\ldots n)$ and $X_0 = 0$. The n observations will generate n-1 spacings.

4. Let $t_i = \sum_{j=1}^{i} Y_j$ and let $u_i = t_i/t_n$, $i=1,2,\ldots,n-1$; then the test statistic for the linear regression is given by:

$$Z_1 = \sqrt{12(n-1)}\,(\bar{u} - 0.5) \qquad (1)$$

5. The test statistic for the quadratic component is given by:

$$Z_2 = \sqrt{\frac{5(n-1)}{4(n+1)(n-2)}}\left(n - 2 + 6n\bar{u} - 12\sum_{i=1}^{n-1}\frac{iu_i}{n-1}\right) \qquad (2)$$

6. The final test statistic then is

$$Z = Z_1^2 + Z_2^2 \qquad (3)$$

As mentioned earlier the null distribution of Z is in the limit a chi-squared distribution with 2 degrees of freedom or an exponential distribution with a mean of 2.0. The test is an upper tail test and thus the p-value is given by

$$p = e^{-Z/2} \qquad (4)$$

A Monte Carlo analysis was conducted of the null distribution of Z in Brain and Shapiro [3] showed that for small to moderate samples, a sample size correction factor resulted in an exact percentile but the correction factor was small and approached zero fairly quickly.

III. APPLICATION OF THE BRAIN AND SHAPIRO PROCEDURE

The Brain and Shapiro Test can be used to test for any underlying distribution since every distribution can be transformed to an exponential distribution either directly or indirectly. For example, the Weibull, Pareto and the Gumbel distributions are easily transformed to the exponential via straightforward transformations. For distributions, where the transformation is not straightforward, the probability integral transform leads to a uniform distribution which can then be transformed to an exponential distribution. As mentioned in the introduction, Gulati and Shapiro [1] and Gulati [2] used the Brain and Shapiro test to develop testing procedures for the Pareto, Rayleigh and the Laplace distributions by transforming them and then testing the transformed data for exponentiality. Here we extend this concept to include all distributions and test it for some specific distributions.

The details of the testing procedure are similar to those outlined in Section 2 with the exception of the fact that we work with transformed data rather than the actual observations. Thus if the ordered data are given by $X_{(1)} \leq X_{(2)} \leq \ldots\ldots \leq X_{(n)}$ (with the underlying assumption that they come from some continuous distribution, F(x)), we carry out the following procedure:

1. Estimate the underlying parameters (if they are unknown) and transform the data so that the transformed observations are approximately (this will be the case when the underlying parameters are unknown) exponential under the null hypothesis.

2. Denote the ordered transformed observations by $W(1) < W(2) < \ldots\ldots < W(n)$. If the data cannot be transformed directly, then transform the data to the uniform using the probability integral transform and then to the exponential. Thus the transformation will be (here represents F with estimated parameters).

3. Compute the ith weighted spacing as before as $Y_i = (n - i + 1)*(W_i - W_{(i-1)})$ $(i=1,2,\ldots n)$ and $X_0 = 0$.

4. Compute t_i , Z_1, Z_2 as in Section 2.

5. As in Section 2, the final test statistic will be given by

$$Z = Z_1^2 + Z_2^2 \qquad (5)$$

The null distribution of Z should be in the limit a chi-squared distribution with 2 degrees of freedom or an exponential distribution with a mean of 2.0. As before, the p-value of the test will be given by $p = e^{-Z/2}$.

IV. NULL DISTRIBUTION OF THE TEST STATISTIC FOR SPECIFIC DISTRIBUTIONS

As a preliminary study, we conducted Monte Carlo simulations to determine the null distribution of Z for distributions characterized by a single parameter only (usually a scale parameter) or two parameters (both a scale and a shape parameter.) The purpose of the simulations was two-fold. We wanted to determine how much (if any) information was lost by using parameter estimates instead of the true parameter values and if the transformation still preserved exponentiality. The simulations were conducted for the uniform distribution on (0,1), the Weibull with known shape, the Weibull with unknown shape, and the Gumbel distribution. The sample size varied from 10 – 100 (10) and the number of simulation runs was fixed at 10,000. We found that for the uniform distribution and the Weibull distribution with known shape, the empirical percentiles were fairly close to the percentiles of the chi-squared distribution with 2 degrees of freedom. The approximation improved with increasing sample size and was generally independent of the true parameter value. However for two-parameter distributions, we found that the null percentiles of Z were closely approximated by a chi-square distribution with one degree of freedom. This was not surprising since we had observed the same phenomenon in the case of the two parameter Pareto distribution (see [1]) and it adheres to the principle of "lose one degree of freedom for every estimated parameter"

Tables 1, II and III give the results of the simulations for the Uniform distribution, Weibull distribution with shape 0.5 but assumed to be unknown and shape 0.5, but assumed known respectively (for both Tables II and III, the scale parameter is set to 1 without loss of generality.)

TABLE I. PERCENTILES OF THE UNIFORM(0,1) DISTRIBTION

Sample Size		
	CV 90	**CV 95**
10	4.433	6.050
20	4.483	5.877
30	4.472	5.958
40	4.517	5.933
50	4.503	5.908
60	4.467	5.836
70	4.513	5.879
80	4.583	5.879
90	4.469	5.928
100	4.488	5.757
CV 90 – 90th percentile, CV 95 – 95th percentile percentile		

TABLE II. PERCENTILES OF THE WEIBULL DISTRIBUTION WITH SHAPE = 0.5 (ASSUMED UNKNOWN) AND SCALE = 1

Sample Size		
	CV 90	**CV 95**
10	2.924	3.758
20	2.839	3.797
30	2.819	3.826
40	2.834	3.854
50	2.809	3.932
60	2.775	3.833
70	2.838	3.934
80	2.751	3.854
90	2.772	3.959
100	2.816	3.914
CV 90 – 90th percentile, CV 95 – 95th percentile		

TABLE III. PERCENTILES OF THE WEIBULL DISTRIBUTION WITH KNOWN SHAPE =0.5 AND SCALE = 1

Sample Size		
	CV 90	**CV 95**
10	4.433	6.050
20	4.483	5.877
30	4.472	5.958
40	4.517	5.933
50	4.503	5.908
60	4.467	5.836
70	4.513	5.879
80	4.583	5.879
90	4.469	5.928
100	4.488	5.757
CV 90 – 90th percentile, CV 95 – 95th percentile		

V. POWER STUDIES

Monte Carlo studies were conducted to determine the power of the procedure. Previous studies have shown that this procedure generally has high power (see Brain and Shapiro [3], Gulati and Shapiro [1] and Gulati, [2]). Here we studied the power of the procedure for the Uniform distribution, the Gumbel distribution and for the Weibull distribution with known and unknown shape. All these studies were based on 10,000 simulation runs with sample sizes ranging between 10 and 100 (in increments of 10). In all cases, we chose alternate distributions that were skewed as well as symmetric. We found that the power of the procedure was extremely high for the uniform distribution for left skewed and symmetric alternatives and for the Weibull distribution with known shape. For the Gumbel distribution, the power depends on the alternate distribution but tends to be high in general. However, for the Weibull distribution, the procedure tends to have low power. Some of these results are displayed in Tables IV, V and VI for n = 20 and 50 and $\alpha = 0.1$.

TABLE IV. POWER OF TEST FOR THE UNIFORM DISTRIBUTION, TYPE I ERROR = 0.1

Alternate Distribution	n=20	n=50
Beta (2, 1) Skewed Right	0.0908	0..0954
Beta (2,2) Symmetric	0.5563	0.9292
Beta (1,2) Skewed Left	0.5030	0.904

TABLE V. POWER OF THE TEST FOR THE GUMBEL DISTRIBUTION, TYPE I ERROR = 0.1

Alternate Distribution	n=20	n=50
Gamma (2, 0.5)	0.2377	0.4597
U(0,1)	0.1526	0.1719
LN(0,1)	0.7307	0.9823
Weibull (0.5, 1)	0.9956	1.000
N(0,1)	0.3177	0.6685
T(1)	0.8674	0.9981
T(4)	0.5097	0.8672
HN	0.2952	0.5910
Cauchy	0.8707	0.996
Gamma (a, b) = Gamma with shape a and scale b, U(0,1) = standard uniform distribution, LN(0,1) = standard lognormal, T(k) = t distribution with k degrees of freedom, HN= Half Normal distribution		

TABLE VI. POWER OF THE TEST FOR THE WEIBULL DISTRIBUTION (BOTH PARAMETERS UNKNOWN), TYPE I ERROR = 0.1

Alternate Distribution	n=20	n=50
Gamma (1, 1)	0.1105	0.0988
Gamma (1,2)	0.1177	0.1078
Chisq(1)	0.1546	0.2252
Chisq(4)	0.1291	0.1774
LN	0.3163	0.6726
HN	0.1516	0.2115
Gamma (a, b) = Gamma with shape a and scale b, Chisq(k) = chi squared distribution with k degrees of freedom, LN= standard lognormal, HN= Half Normal distribution		

VI. CONCLUDING REMARKS

This paper extends the results of Gulati and Shapiro [1] and Gulati [2] to develop a universal test. The procedure transforms the underlying data and tests the transformed data for exponentiality using the regression test of Brain and Shapiro [3]. The test statistic is simple to compute and the p-values of the test can be obtained easily from the chi-squared distribution. The power of the procedure is generally high (with the exception for the Weibull distribution). The authors plan to include more distributions in the final version of the paper as well as study the effect of the probability integral transformation on the null distribution of the test statistic.

REFERENCES

[1] Sneh Gulati and Samuel Shapiro. "Goodness of Fit Tests for the Pareto Distribution". Statistical Models and Methods for Biomedical and Technical Systems published by Birkhauser, Boston (Vonta, F., Nikulin, M., Limnios, N. and Huber, C. editors), pp. 263-277, 2006.

[2] Sneh Gulati "Goodness of Fit for the Rayleigh and the Laplace Distributions" International journal of Applied Mathematics and Statistics, 24, pp. 74-85, 2011.

[3] Brain, C. W. and Shapiro, S. S. "A Regression Test for Exponentiality: Censored and Complete Samples," Technometrics, 25, pp. 69-76. 1983.

[4] D'Agastino, R.B. and Stephens, M. Goodness of Fit Techniques, Marcel Dekker, New York, 1986

[5] Cox, D. R. and Lewis, P. A. The Statistical Analysis of a Series of Events, London: Methune, 1966.

Comparison of two nonparametric estimators for reliability of discrete-time semi-Markov systems based on multiple independent observations

Stylianos Georgiadis and Nikolaos Limnios

Université de Technologie de Compiègne,
Laboratoire de Mathématiques Appliquées de Compiègne,
Centre de Recherches de Royallieu, BP 20529,
60205 Compiègne, Cedex, France
stylianos.georgiadis@utc.fr nikolaos.limnios@utc.fr

***Abstract*—We consider a discrete-time semi-Markov system, with a finite state space. The empirical and the exact maximum likelihood estimator for the semi-Markov kernel are given in the case of multiple parallel observations of the same process. Afterwards, we describe a reliability model described by a discrete-time semi-Markov process and we derive basic reliability measures, such as reliability, availability, failure rates and mean hitting times. Finally, we present a comparison between empirical and exact maximum likelihood estimators for these measures through a numerical application.**

***Index Terms*—Discrete-time semi-Markov system, nonparametric estimation, exact maximum likelihood estimation, reliability, mean hitting times.**

I. Discrete-time semi-Markov system

In recent literature, discrete-time semi-Markov models have achieved significant importance in probabilistic and statistical modeling especially the ones with a finite space state. System reliability and relative dependability measures consist, amongst others, an important application field. The term chain will be used for a discrete-time semi-Markov process. A general study on the semi-Markov chains is given by Barbu and Limnios [1] toward applications. Some statistical inference problems, such as the proposition of a computation procedure for solving the corresponding Markov renewal equation and the study of an empirical estimator of the semi-Markov kernel and other measurements in the case of one observed trajectory, are presented.

We consider a semi-Markov chain with finite state space and the sequence of the backward recurrence times, which form a coupled Markov chain. The basic properties of this Markov chain have been studied in Chryssaphinou et al. [2]. Trevezas and Limnios (2011) [3] present the exact maximum likelihood (EML) estimation of the semi-Markov kernel for a single trajectory of a semi-Markov system up to an arbitrary fixed time, when the length of the observation tends to infinity, and, next, when multiple independent observed trajectories generated by the same semi-Markov kernel, censored at a fixed time, when the number of trajectories tends to infinity, and study its asymptotic properties. In the present work, we focus on the latter case for a nonparametric semi-Markov model, which, from a practical point of view, corresponds to the evolution of multiple identical components of a repairable system or systems. Based on the maximum likelihood estimation of the coupled Markov chain, we examine the estimation of several reliability measures of a discrete-time semi-Markov system.

We give now all the necessary preliminaries concerned a semi-Markov chain. From now on we will use the following notation for the non-zero natural numbers $\mathbb{N}^* = \mathbb{N} \setminus \{0\}$ and take by convention that $0/0 := 0$.

Consider the finite set $E = \{1, \ldots, s\}$, $s \in \mathbb{N}^*$, and an E-valued stochastic chain $\boldsymbol{Z} := (Z_k)_{k\in\mathbb{N}}$. Let $\boldsymbol{J} := (J_n)_{n\in\mathbb{N}}$ be the successive visited states of $\boldsymbol{Z}$ with state space E and $\boldsymbol{S} := (S_n)_{n\in\mathbb{N}}$ are the jump times of $\boldsymbol{Z}$ with values in $\mathbb{N}$ with $0 = S_0 \leq S_1 \leq \cdots \leq S_n \leq S_{n+1} \leq \cdots$. Also, let us denote $X_n := S_n - S_{n-1}$, $n \in \mathbb{N}^*$, as the sojourn times in these states with values in $\mathbb{N}$.

Definition 1. *The stochastic process $(\boldsymbol{J}, \boldsymbol{S}) := (J_n, S_n)_{n\in\mathbb{N}}$, with state space E, is said to be a Markov renewal chain (MRC), if, for all $j \in E$, $k \in \mathbb{N}$ and $n \in \mathbb{N}$, it satisfies a.s. the following equality*

$$\begin{aligned}&\mathbb{P}(J_{n+1} = j, X_{n+1} = k | J_0, \ldots, J_n; S_1, \ldots, S_n)\\ &= \mathbb{P}(J_{n+1} = j, X_{n+1} = k | J_n).\end{aligned}$$

In this case, $\boldsymbol{Z}$ is called a semi-Markov chain (SMC).

Actually, $\boldsymbol{Z}$ gives the state of the process at time k. We assume that the MRC $(\boldsymbol{J}, \boldsymbol{S})$ is time homogeneous, that is, the above probability is independent of n and S_n. The process $\boldsymbol{J}$ is a Markov chain (MC) with state space E and transition kernel $\boldsymbol{p} := (p_{ij};\ i, j \in E)$, where

$$p_{ij} := \mathbb{P}(J_{n+1} = j | J_n = i),$$

called the embedded Markov chain (EMC) of $\boldsymbol{Z}$. We denote by $N(k)$, $k \in \mathbb{N}$, the process which counts the number of jumps of $\boldsymbol{Z}$ in the interval $(0, k]$, defined by $N(k) := \max\{n \geq 0 :$

$S_n \leq k\}$. The SMC $\boldsymbol{Z}$ is associated with the MRC $(\boldsymbol{J}, \boldsymbol{S})$ by

$$Z_k := J_{N(k)}, \quad k \in \mathbb{N}.$$

Let $N_i(k)$ be the number of visits of $\boldsymbol{Z}$ to state $i \in E$ up to time k, and $N_{ij}(k)$ the number of direct jumps of $\boldsymbol{Z}$ from state i to state j up to time k. To be specific,

$$N_i(k) := \sum_{m=1}^{N(k)} \mathbf{1}_{\{J_{m-1}=i\}}$$
$$N_{ij}(k) := \sum_{m=1}^{N(k)} \mathbf{1}_{\{J_{m-1}=i, J_m=j\}},$$

where $\mathbf{1}_A$ is the indicator function of the set A.

Definition 2. *The transition kernel $\boldsymbol{q}(k) := (q_{ij}(k);\ i,j \in E)$, $k \in \mathbb{N}$, is called the discrete-time semi-Markov kernel (DTSMK) of the SMC $\boldsymbol{Z}$ and it is defined by*

$$q_{ij}(k) := \mathbb{P}(J_{n+1} = j, X_{n+1} = k | J_n = i). \tag{1}$$

For all $i,j \in E$, let $\boldsymbol{f}(k) := (f_{ij}(k); i,j \in E)$ be the conditional distribution function of the sojourn time in any state i, given that the next visited state is j, $j \neq i$, defined as follows

$$f_{ij}(k) := \mathbb{P}(X_{n+1} = k | J_n = i, J_{n+1} = j) = \begin{cases} \frac{q_{ij}(k)}{p_{ij}}, & \text{if } p_{ij} \neq 0, \\ \mathbf{1}_{\{k=\infty\}}, & \text{if } p_{ij} = 0. \end{cases}$$

Definition 3. *For all $i,j \in E$, let us denote by $\boldsymbol{H}(k) := diag(H_i(k);\ i \in E)^\top$, $k \in \mathbb{N}$, the sojourn time cumulative distribution function in any state i*

$$H_i(k) := \mathbb{P}(X_{n+1} \leq k | J_n = i) = \sum_{j \in E} \sum_{l=0}^{k} q_{ij}(l).$$

and by $\bar{\boldsymbol{H}}(k) := (\bar{H}_i(k);\ i \in E)^\top$, $k \in \mathbb{N}$, the survival function in any state i.

Let us denote by μ_{ii} the mean recurrence time of state i for the SMC $\boldsymbol{Z}$, by $\boldsymbol{\pi} = (\pi_i; i \in E)$ and $\boldsymbol{\nu} = (\nu_i; i \in E)$, the stationary distribution of the SMC $\boldsymbol{Z}$ and the EMC $\boldsymbol{J}$, respectively. Let $\boldsymbol{m} := (m_i; i \in E)^\top$ be the vector with m_i to be the mean sojourn time of $\boldsymbol{Z}$ in state $i \in E$, i.e. $m_i := \sum_{n \in \mathbb{N}} [1 - H_i(n)]$, and $\bar{m}$ the mean sojourn time of $\boldsymbol{Z}$ defined as $\bar{m} := \sum_{k \in E} \nu_k m_k$.

Definition 4. *The matrix function $\boldsymbol{\psi}(k) := (\psi_{ij}(k); i,j \in E)$, $k \in \mathbb{N}$, is called Markov renewal function and it is defined by*

$$\psi_{ij}(k) := \mathbb{P}(\bigcup_{n=0}^{k} \{J_n = j, S_n = k\} | J_0 = i) := \sum_{n=0}^{k} q_{ij}^{(n)}(k),$$

where $\boldsymbol{q}^{(n)}(k) := (q_{ij}^{(n)}(k);\ i,j \in E)$, $n,k \in \mathbb{N}$, is the n-fold discrete-time convolution (see [1]), given as

$$q_{ij}^{(n)}(k) := \mathbb{P}(J_n = j, S_n = k | J_0 = i).$$

Let $\boldsymbol{I} := (I(k);\ k \in \mathbb{N})$, where $I(k) := (\mathbf{1}_{\{i=j\}}(k);\ i,j \in E)$ and

$$\mathbf{1}_{\{i=j\}}(k) := \begin{cases} 1, & \text{if } i = j,\ k \geq 0, \\ 0, & \text{otherwise.} \end{cases}$$

We denote by $*$, the convolution between two (matrix-valued) functions.

Definition 5. *The transition function $\boldsymbol{P}(k) := (P_{ij}(k);\ i,j \in E)$, $k \in \mathbb{N}$, of the SMC $\boldsymbol{Z}$ is defined by $P_{ij}(k) := \mathbb{P}(Z_k = j | Z_0 = i)$ and, in matrix form, is written as*

$$\boldsymbol{P}(k) = \boldsymbol{\psi} * (\boldsymbol{I} - \boldsymbol{H})(k).$$

The definition of the sequence of the backward recurrence times is now given.

Definition 6. *For all $k \in \mathbb{N}$, we define $\boldsymbol{U} := (U_k)_{k \in \mathbb{N}}$ as the sequence of the backward recurrence times for the SMC $\boldsymbol{Z}$ given by*

$$U_k := \begin{cases} k, & \text{if } k < S_1, \\ k - S_{N(k)}, & \text{if } k \geq S_1. \end{cases}$$

We note that, for all $k \in \mathbb{N}$, $U_k \leq k$. The stochastic process $(\boldsymbol{Z}, \boldsymbol{U}) := (Z_k, U_k)_{k \in \mathbb{N}}$ is a MC with values in $E \times \mathbb{N}$. In our case, where $S_0 = 0$, we get that $U_0 = 0$.

Definition 7. *The transition matrix $\boldsymbol{P}^B := (p_{i,u;j};\ i,j \in E, u \in \mathbb{N})$ of the MC $(\boldsymbol{Z}, \boldsymbol{U})$ is defined as*

$$p_{i,u;j} := \begin{cases} \mathbb{P}(Z_{k+1} = j, U_{k+1} = 0 | Z_k = i, U_k = u), & j \neq i, \\ \mathbb{P}(Z_{k+1} = i, U_{k+1} = u+1 | Z_k = i, U_k = u), & j = i. \end{cases}$$

The value of U_{k+1} is fully determined by the value of Z_{k+1}. So, for all $(i,u) \in E \times \mathbb{N}$ and all $k \in \mathbb{N}$ such that $\mathbb{P}(Z_k = i, U_k = u) > 0$, the transition probabilities of the MC $(\boldsymbol{Z}, \boldsymbol{U})$ are written as

$$p_{i,u;j} = \begin{cases} \frac{q_{ij}(u+1)}{\bar{H}_i(u)}, & j \neq i, \\ \frac{\bar{H}_i(u+1)}{\bar{H}_i(u)}, & j = i. \end{cases}$$

We assume that the MRC $(\boldsymbol{J}, \boldsymbol{S})$ is irreducible and aperiodic, with finite mean sojourn time. The MC $(\boldsymbol{Z}, \boldsymbol{U})$ is therefore irreducible.

II. Empirical and Exact Maximum Likelihood Estimation

In this section, we present the nonparametric estimation of semi-Markov chains by two different aspects; the empirical and EML estimation. We observe a SMC in the interval $[0, M]$, where $M \in \mathbb{N}^*$ a fixed censoring time.

Definition 8. *Let us define the observation of the SMC $\boldsymbol{Z}$ censored at time $M \in \mathbb{N}^*$*

$$\mathcal{H}_M := \{Z_u;\ 0 \leq u \leq M\} := \{J_0, X_1, J_1, \ldots, X_{N(M)}, J_{N(M)}, U_M\},$$

where $U_M := M - S_{N(M)}$.

Now, we suppose the realization of L, $L \geq 2$ independent observed trajectories observations censored at a time $M \in \mathbb{N}^*$, fixed for all, when the number of the observations tends to infinity. We collect the total information in the interval $[0, M]$ and we exclude the results without separating the different trajectories.

Let $N^l(M)$, $N_i^l(M)$ and $N_{ij}^l(M)$ be the l-th realizations of the counting processes $N(M)$, $N_i(M)$ and $N_{ij}(M)$, respectively, as defined in the previous section.

For all $i, j \in E$, $k \in \{0, 1, \ldots, M\}$, $M \in \mathbb{N}^*$, and $l = \{1, \ldots, L\}$, we define the following discrete-time counting process $N_{ij}^l(k, M)$ that gives the number of visits from state i to state j, up to time M, with sojourn time in state i equal to k, for the l-th trajectory, defined as

$$N_{ij}^l(k, M) := \sum_{n=1}^{N^l(M)} \mathbf{1}_{\{J_{n-1}^l = i, J_n^l = j, X_{n+1}^l = k\}}.$$

Definition 9. *Let L be the number of independent observed trajectories up to fixed time $M \in \mathbb{N}^*$. For any $i, j \in E$ and any $k \in \{0, 1, \ldots, M\}$, we define the following counting processes*

$$N_i(M, L) := \sum_{l=1}^{L} N_i^l(M),$$
$$N_{ij}(M, L) := \sum_{l=1}^{L} N_{ij}^l(M),$$
$$N_{ij}(k, M, L) := \sum_{l=1}^{L} N_{ij}^l(k, M).$$

For both empirical and EML estimation, the estimated initial law $\hat{\boldsymbol{\alpha}}(M, L) := (\hat{\alpha}_i(M, L);\ i \in E)$ and the estimated transition matrix $\hat{\boldsymbol{p}}(M, L) := (\hat{p}_{ij}(M, L);\ i, j \in E)$ of L trajectories, for any $M \in \mathbb{N}^*$, are given by

$$\hat{\alpha}_i(M, L) := \frac{N_i^{\alpha}(L)}{L} := \frac{1}{L} \sum_{l=1}^{L} \mathbf{1}_{\{Z_0^l = i\}},$$
$$\hat{p}_{ij}(M, L) = \frac{N_{ij}(M, L)}{N_i(M, L)}.$$

Definition 10. *Let L independent observations of a SMC $\boldsymbol{Z}$ up to a fixed censoring time $M \in \mathbb{N}^*$. For any $i, j \in E$ and $k \in \{0, 1, \ldots, M\}$, the empirical estimator $\tilde{\boldsymbol{q}}(k, M) := (\tilde{q}_{ij}(k, M);\ i, j \in E)$, for the DTSMK (2) is given as follows*

$$\hat{q}_{ij}(k, M, L) = \frac{N_{ij}(k, M, L)}{N_i(M, L)}. \tag{2}$$

The EML estimator is based on the time from the last jump of an observation up the time k.

Definition 11. *For all $i, j \in \mathbb{N}$, $k \in \{0, 1, \ldots, M\}$, $M \in \mathbb{N}^*$, and $l = \{1, \ldots, L\}$, we define the following discrete-time counting processes*

1) $N_{i,u}^{B,l}(M) := \sum_{n=1}^{M} \mathbf{1}_{\{Z_{n-1}^l = i, U_{n-1}^l = u\}}$ *: the number of visits in the state $(i, u) \in E \times \{0, 1, \ldots, M-1\}$, up to time $M \in \mathbb{N}^*$, neglecting the last visited state (J_M, U_M).*
2) $N_{i,u}^{B,l}(j, M) := \sum_{n=1}^{M} \mathbf{1}_{\{Z_{n-1}^l = i, Z_n^l = j, U_{n-1}^l = u\}}$ *: the number of visits of $\boldsymbol{Z}$ from state i to state j, with backward recurrence time u, up to time $M \in \mathbb{N}^*$.*

Definition 12. *For all $i, j \in E$ and $u \in \{0, 1, \ldots, M-1\}$, $M \in \mathbb{N}^*$, we define the counting processes*

$$N_{i,u}^B(M, L) := \sum_{l=1}^{L} N_{i,u}^{B,l}(M),$$
$$N_{i,u}^B(j, M, L) := \sum_{l=1}^{L} N_{i,u}^{B,l}(j, M).$$

Proposition 1 ([3])**.** *For any fixed time $M \in \mathbb{N}^*$, the EML estimator $\tilde{\boldsymbol{q}}(k, M, L) := (\tilde{q}_{ij}(k, M, L);\ i, j \in E)$, $i, j \in E$, $i \neq j$, $k \in \{1, \ldots, M\}$, for the DTSMK (2) in case of L trajectories is given as follows*

$$\tilde{q}_{ij}(k, M, L) = \begin{cases} \tilde{p}_{i,0}(j, M, L), & k = 1, \\ \tilde{p}_{i,k-1}(j, M, L) \prod_{u=0}^{k-2} \tilde{p}_{i,u}(i, M, L), & 2 \leq k \leq M, \end{cases} \tag{3}$$

where

$$\tilde{p}_{i,u}(j, M, L) = \frac{N_{i,u}^B(j, M, L)}{N_{i,u}^B(M, L)}, \quad u \in \{0, 1, \ldots, M-1\}.$$

III. Reliability model

A scientific field that semi-Markov models have been applied is, among others, reliability theory. We present the main measures of reliability, availability, failure rates and mean hitting times and how the theory of semi-Markov chains contribute to their study.

For a stochastic system with state space E, described by a SMC, we distinguish the up and down states of the system, denoted by U and D accordingly, i.e. $E = U \cup D$, with $U \cap D = \emptyset$ and $U, D \neq \emptyset$. For a finite state space $E = \{1, \ldots s\}$, we enumerate first the up states, $U = \{1, \ldots, r\}$, and next the down states, $D = \{r+1, \ldots, s\}$. For $m, n \in \mathbb{N}^*$, with $m > n$, let $\mathbf{1}_{m,n}$ denote the m-column vector whose the n first elements are 1 and the last $m-n$ ones are 0. For $m \in \mathbb{N}^*$, let $\mathbf{1}_m$ denote the m-column vector with all elements equal to one.

Now, let us denote by $\boldsymbol{\alpha_1}$ and $\boldsymbol{\alpha_2}$, the vectors of the initial law on U and D respectively (in the same manner, we consider the partitions of the sojourn time cumulative distribution function $\boldsymbol{H}(k)$ and the mean sojourn times $\boldsymbol{m}$). Considering the transition kernel $\boldsymbol{p}$, the submatrices $\boldsymbol{p_{11}}$, $\boldsymbol{p_{12}}$, $\boldsymbol{p_{21}}$ and $\boldsymbol{p_{22}}$ are the restrictions of $\boldsymbol{p}$ on $E \times E$, $E \times U$, $U \times E$ and $U \times U$ respectively (similarly, we act for the DTSMK $\boldsymbol{q}(k)$, the Markov renewal function $\boldsymbol{\psi}(k)$ and the transition function $\boldsymbol{P}(k)$).

Also, let us denote by T_D the first passage time in subset D, called *the lifetime of the system*, and by T_U the first hitting time of subset U given that $\boldsymbol{\alpha_1} = \mathbf{0}$, called *the repair time*. That is,

$$T_D := \min\{n \in \mathbb{N} :\ Z_n \in D\},$$
$$T_U := \min\{n \in \mathbb{N} :\ Z_n \in U\},$$

with $\min \emptyset := \infty$.

A. Reliability

Definition 13. *The reliability R of a system at time $k \in \mathbb{N}$, starting to function at time $k = 0$, is defined as the probability that the system has functioned without failure in the interval $[0, k]$, i.e.*

$$R(k) := \mathbb{P}(Z_n \in U;\ \forall n \in [0, k]).$$

In the framework of a semi-Markov model, for all $k \in \mathbb{N}$, the reliability is defined by the following equation

$$R(k) = \boldsymbol{\alpha_1 P_{11}}(k)\mathbf{1}_r.$$

B. Availability

Definition 14. *The pointwise availability A of a system at time $k \in \mathbb{N}$ is the probability that the system is operational at time k (independently of the fact that the system has failed or not in $[0, k)$), i.e.*

$$A(k) := \mathbb{P}(Z_k \in U).$$

That means that the system functions at the time k, ignoring its history. For all $k \in \mathbb{N}$, the pointwise availability is given by

$$A(k) = \boldsymbol{\alpha P}(k)\mathbf{1}_{s,r}.$$

Definition 15. *The steady-state availability A_∞ of a system is defined as the limit of the pointwise availability (when the limit exists), as the time tends to infinity, i.e.*

$$A_\infty := \lim_{k\to\infty} A(k).$$

For a semi-Markov system, the steady-state availability is given by

$$A_\infty = \frac{1}{\boldsymbol{\nu m}} \boldsymbol{m}^\top diag(\boldsymbol{\nu})\mathbf{1}_{s,r}.$$

C. Failure rate functions

1) BMP-failure rate function:

Definition 16. *The BMP-failure rate function λ of a system at time $k \in \mathbb{N}$, starting working at time $k = 0$, is the conditional probability that the failure of the system occurs at time k, given that the system has worked until time $k - 1$, i.e.*

$$\lambda(k) := \mathbb{P}(T_D = k | T_D \geq k).$$

The BMP-failure rate at time $k \geq 1$ is given by

$$\lambda(k) = \begin{cases} 1 - \frac{\boldsymbol{\alpha_1 P_{11}}(k)\mathbf{1}_r}{\boldsymbol{\alpha_1 P_{11}}(k-1)\mathbf{1}_r}, & R(k-1) \neq 0, \\ 0, & otherwise, \end{cases}$$
$$= \begin{cases} 1 - \frac{R(k)}{R(k-1)}, & R(k-1) \neq 0, \\ 0, & otherwise, \end{cases}$$

with $\lambda(0) = 1 - R(0)$.

2) RG-failure rate function:

Due to some difficulties in applying the BMP-failure rate function on some discrete-time systems, an alternative failure rate function r has been proposed

$$r(k) = \begin{cases} -\ln \frac{R(k)}{R(k-1)}, & k \geq 1, \\ -\ln R(0), & otherwise, \end{cases}$$

called *the RG-failure rate function* at time $k \in \mathbb{N}$.

D. Mean Hitting Times

1) Mean time to failure:

Definition 17. *The mean time to failure (MTTF) is defined as the mean lifetime, i.e. the expectation of the hitting time to the down set D, $MTTF := \mathbb{E}[T_D]$.*

The mean time to failure in a semi-Markov model follows

$$MTTF = \boldsymbol{\alpha_1}(I - \boldsymbol{p_{11}})^{-1}\boldsymbol{m_1}.$$

2) Mean time to repair:

Definition 18. *The mean time to repair (MTTR) is defined as the mean of the repair duration, i.e. the expectation of the hitting time to the up set U, $MTTR := \mathbb{E}[T_U]$.*

The mean time to repair is given as

$$MTTR = \boldsymbol{\alpha_2}(I - \boldsymbol{p_{22}})^{-1}\boldsymbol{m_2}.$$

IV. Numerical Application

In this section, we apply the previous results to a three-state semi-Markov system described as follows. The state space of the system $E = \{1, 2, 3\}$ is partitioned into the up-state set $U = \{1, 2\}$ and the down-state set $D = \{3\}$. To define it completely, we need the initial law $\boldsymbol{\alpha} = (0.9 \quad 0.1 \quad 0)$ and the transition kernel $\boldsymbol{p}$ of the EMC $\boldsymbol{J}$, given by

$$\boldsymbol{p} = \begin{pmatrix} 0 & 1 & 0 \\ 0.6 & 0 & 0.4 \\ 1 & 0 & 0 \end{pmatrix}.$$

The conditional distributions of the sojourn times are

$$\boldsymbol{f}(k) = \begin{pmatrix} 0 & f_{12}(k) & 0 \\ f_{21}(k) & 0 & f_{23}(k) \\ f_{31}(k) & 0 & 0 \end{pmatrix},$$

where the conditional distributions for the sojourn times $f_{12}(k)$ and $f_{31}(k)$ are geometric distributions with parameters $p = 0.15$ and $p = 0.20$ respectively, and the distributions $f_{21}(k)$ and $f_{23}(k)$ follow the discrete Weibull distribution with parameters $(q, b) = (0.9, 1.2)$ for the transition $2 \to 1$ and $(q, b) = (0.8, 1.2)$ for the transition $2 \to 3$. The realization of a trajectory of fixed length $M \in \mathbb{N}^*$ of the SMC $\boldsymbol{Z}$ with state space E, transition matrix $\boldsymbol{p}$ and initial law $\boldsymbol{\alpha}$ is simulated through a Monte Carlo method.

We observe 20000 independent trajectories of the SMC $\boldsymbol{Z}$ up to the censoring time $M = 100$. The empirical and EML estimations for all the measures are based on the estimators (2) and (3) of the DTSMK. We present now the plots for the

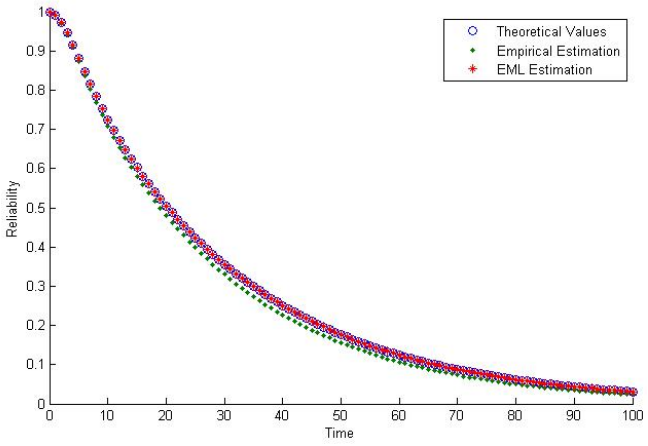


Fig. 1. Reliability plot.

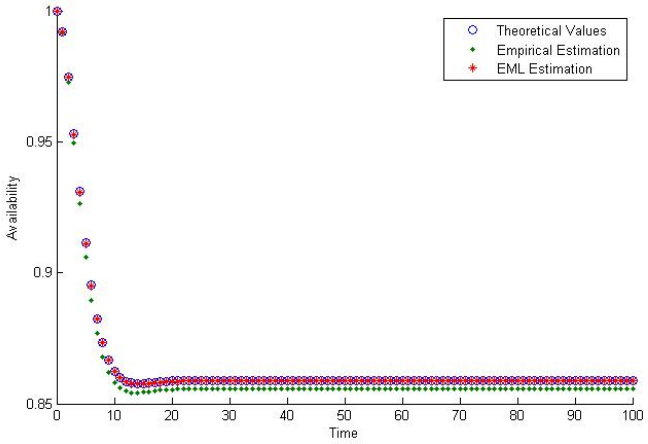


Fig. 2. Availability plot.

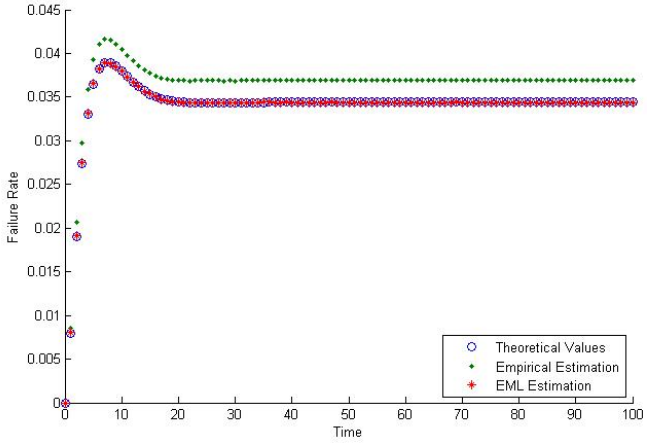


Fig. 3. BMP-failure rate plot.

reliability, availability and failure rates, presented in Section III and their estimations.

In Table I, the empirical and EML estimation are given for the steady-state availability and mean hitting times

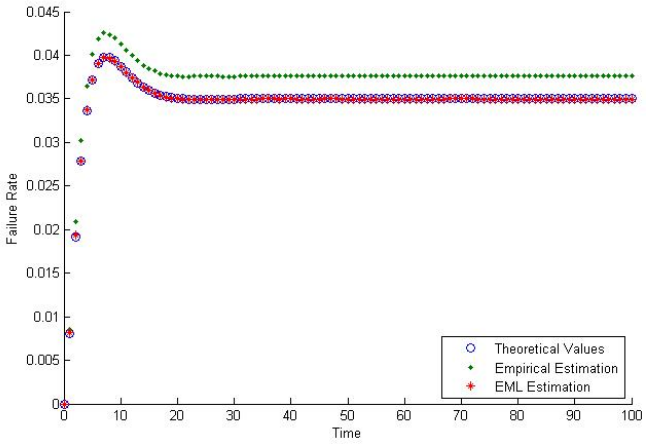


Fig. 4. RG-failure rate plot.

	True Value	Empirical Estimation	EML Estimation
A_∞	0.8589	0.8556	0.8566
$MTTF$	29.7578	27.8198	29.2270
$MTTR$	5.0000	4.8017	5.0067

TABLE I
ESTIMATION OF STEADY-STATE AVAILABILITY AND MEAN HITTING TIMES.

V. CONCLUSION

In the case of a single observed trajectory, the backward recurrence time at time M, is neglected from empirical estimators. In contrast, in the case of multiple observations, significant difference between the two estimations is observed. Even in the case of a large number of trajectories, when the time M is small, the estimated values of all the reliability measures differ, making the empirical estimation to seem less accurate than the EML estimation, which appear to be almost identical to the theoretical values.

The time interval until the system reaches the steady state is of great significance for real data applications and provide important information for the evolution of a system. For this interval, the differences of the two estimators in Figures 1 and 2 seem even wider.

REFERENCES

[1] V. S. Barbu and N. Limnios, *Semi-Markov Chains and Hidden Semi-Markov Models Toward Applications*. New York: Springer, 2008.

[2] O. Chryssaphinou, M. Karaliopoulou, and N. Limnios, "On discrete-time semi-Markov chains and applications in words occurrences," *Communications in Statistics - Theory and Methods*, vol. **37**, pp. 1306–1322, 2008.

[3] S. Trevezas and N. Limnios, "Exact MLE and asymptotic properties for nonparametric semi-Markov models," *Journal of Nonparametric Statistics*, pp. 1–21, 2011.

Analysis of different estimation methods from an accelerated test plan

Fabrice Guerin, Pauline Beaumont, Matteo Luca, Facchinetti, Guy Martin Borret, Pascal Lantieri

In an innovative industrial framework, where every prototype is supposed to be validated more and more quickly at the lowest cost, the reliability has to be assessed effectively. Dealing with fatigue phenomena, which represent the main failure mode for metallic chassis, the reliability is associated to validation tests involving cyclic stresses, applied during a given number of cycles or up to failure. In order to reduce the specimen number (i.e. expensive prototypes) and to shorten testing procedures (i.e. time-to-result delay), Accelerated Fatigue Tests (AFT) are developed under the constraint of the needed high-reliability assessment. This paper reviews a selected fatigue test procedure currently used in the automotive industry (i.e. StairCase), then addresses a comparison of several reliability estimation methods applied to the results data of this test procedure : the Dixon & Mood model , the Maximum Likelihood method and the Accelerated Life Testing models (ALT).

A partially accelerated life test planning with competing risks and linear degradation path model

Firoozeh Haghighi
Department of Mathematics, Statistics and Computer Sciences
University of Tehran
Tehran, Iran
Email: haghighi@khayam.ut.ac.ir

***Abstract*—In this work, we proposed a partially accelerated life test planning in the presence of competing risks and based on the assumption that the underlying degradation path is linear. Dependence of competing risk intensities on the degradation level is included into the plan and a tampered failure rate model is hold. Parametric estimation method is used and estimators of competing risks intensities and reliability function are given. A simulation study is conducted to evaluate the performance of the method.**

I. Introduction

In reliability studies, accelerated life testing is commonly used to shorten unit lifetime faster. In such testing, units run at higher than normal conditions to collect more failure times in a limited time. The stress loading in accelerated life testing can be applied various ways. In this work, we considered a simple step-stress plan with the normal stress at first level of loading. Such testing is called partially accelerated life testing. We assumed that the test unit has a linear degradation path (without measurement errors) and fails due to one of several risks, called competing risks. In this plan for each unit, the failure time, corresponding competing risk mode and the level of degradation at the moment of failure are recorded. These information are extrapolated to estimate the lifetime characteristics at normal conditions through a tampered failure rate model. Tampered failure rate model proposed by Battacharyya and Zanzawi (1989), relates the hazard rate of a unit at one stress level to the hazard rate of that unit at the next stress level. The direction in this modeling is based on the hazard acceleration. This approach differs from one in the cumulative exposure model which commonly used in ALT. Although, the cumulative exposure model proposed by Nelson (1990) has a clear mathematical formulation "but its physical motivation is by no means transparent except for the special case of a scale family".

Accelerated planning in presence of competing risks is studied by Pascual (2007, 2008). He considered that the failure times due to competing risks have a Weibull distribution. This assumption may be not be practical in many applications. Hence, we try to plan the more realistic situation of accelerated test wherein no assumption are made about the form of underlying distribution. However, we do require the assumption that it is possible to measure the value of degradation of unit at the moment of failure and the intensity functions corresponding to competing risks depend only on degradation level. The literature on simultaneous analysis of degradation and failure times is scare because of complexity of subject. Bagdonavicius *et al.* (2004) considered statistical analysis of linear degradation and failure time data with multiple failure modes. Semi-parametric estimations for reliability characteristics under general degradation path model and failure time data in the presence of several failure modes were studied by Bagdonavicius *et al.* (2005). In this work, we try to plan a partially accelerated life test for obtaining data in shorter time and based of this plan a statistical analysis of linear degradation and accelerated failure times in presence of competing risks is proposed.

II. Model

We make the following assumptions:

1) Two stress levels S_0 and S_1 $(S_0 < S_1)$ are used. S_0 is stress at normal use conditions.

2) Intensity function corresponding to kth competing risk λ^k at each level of stress depends only on degradation of unit and as a rule it is an increasing function.

3) A linear degradation path model and a tampered failure rate model are hold.

The test is conducted as follows. All test units are initially placed on normal stress S_0, and run until time τ. Then, the stress is changed to high stress S_1, and the test continues until all remaining units fail. Failure time of a unit at lth conditions denotes by $T_l = min(T_l^1, ..., T_l^s)$, where $l \in \{S_0, S_1\}$ and $T_l^k, k = 1, ..., s$ is the failure time corresponding to kth competing risk. Let V denotes the indicator of the failure due to competing risks:

$$V = \begin{cases} 1, & T_l = T_l^1, \\ 2, & T_l = T_l^2, \\ . & . \\ . & . \\ s, & T_l = T_l^s. \end{cases} \quad (1)$$

Consider the degradation process Z, follows a linear degradation path model as follow, (Meeker and Escobar (1998)).

$$Z(t) = \frac{t}{A},$$

where A is a random vector with distribution function π.
Denote n_1, n_2 the number of failures under conditions S_0, S_1. Let $\lambda_l^k(z)$ and $S_l^k(.|A = a)$ to be respectively, the intensity

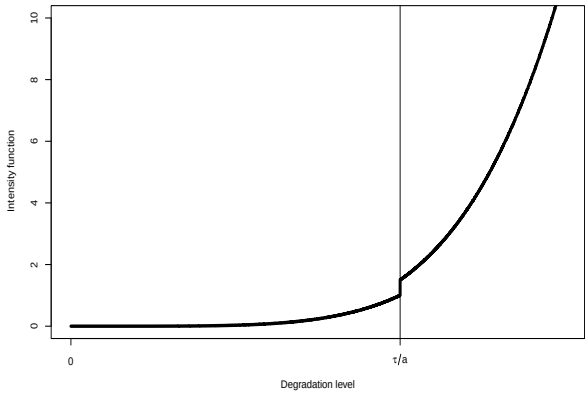


Fig. 1. Tampered failure ate model

and conditional survival functions corresponding to kth risk at lth level of stress.

$$\begin{aligned} S_0^k(t|A=a) &= \mathbf{P}\{T_0^k > t|A=a\} \\ &= \exp\{-\int_0^t \lambda_0^k(\frac{s}{a})ds\} \end{aligned} \tag{2}$$

$$\begin{aligned} S_1^k(t|A=a) &= S_0^k(\tau|A=a)\mathbf{P}\{T_1^k > t|A=a\} \\ &= S_0^k(\tau|A=a)\exp\{-\int_\tau^t \lambda_1^k(\frac{s}{a})ds\} \end{aligned} \tag{3}$$

Substituting $z = \frac{s}{a}$, we have

$$S_0^k(t|A=a) = \exp\{-\int_0^{\frac{t}{a}} a\lambda_0^k(z)dz\} \tag{4}$$

$$S_1^k(t|A=a) = S_0^k(\tau|A=a)\exp\{-\int_{\frac{\tau}{a}}^{\frac{t}{a}} a\lambda_1^k(z)dz\} \tag{5}$$

From tampered failure rate model, we have

$$\lambda^k(z) = \begin{cases} \lambda_0^k(z), & z \le \frac{\tau}{a}, \\ \alpha_k\lambda_0^k(z), & z > \frac{\tau}{a}. \end{cases} \tag{6}$$

The factor α_k corresponding to kth competing risk will depend on stress levels. Figure 1, depicts the two intensity functions for stress S_0 and S_1. From (3) and (6), we have

$$\begin{aligned} S_1^k(t|A=a) &= S_0^k(\tau|A=a)\exp\{-\int_{\frac{\tau}{a}}^{\frac{t}{a}} a\alpha_k\lambda_0^k(z)dz\} \\ &= S_0^k(\tau|A=a)\left[\frac{S_0^k(t|A=a)}{S_0^k(\tau|A=a)}\right]^{\alpha_k} \\ &= \left[S_0^k(\tau|A=a)\right]^{1-\alpha_k}\left[S_0^k(t|A=a)\right]^{\alpha_k} \end{aligned} \tag{7}$$

From the assumptions that the model is a tampered failure rate the conditional survival function of a test unit in the presence of competing risks and under simple step-stress test is

$$S^k(t|A=a) = \begin{cases} S_0^k(t|A=a) \\ \text{where } \ 0 \le t < \tau; \\ \\ \left[S_0^k(\tau|A=a)\right]^{1-\alpha_k}\left[S_0^k(t|A=a)\right]^{\alpha_k} \\ \text{where } \ \tau \le t < \infty. \end{cases} \tag{8}$$

III. Estimation

Consider n units are on test and n_1 units fail under S_0 and n_2 units fail under S_1. Suppose that the failure times T_i, the type of failure modes V_i and the degradation values Z_i at the failure moment under each level of stress are observed. So, the data are as follow:

Under normal conditions: $(T_1, V_1, Z_1), ..., (T_{n_1}, V_{n_1}, Z_{n_1})$.
Under high conditions: $(T_1, V_1, Z_1), ..., (T_{n_2}, V_{n_2}, Z_{n_2})$.

Let

$$\delta_i = \begin{cases} k, & \text{if ith unit fails by kth competing risk,} \\ 0, & \text{o.w.} \end{cases} \tag{9}$$

The likelihood function resulted from observed data under S_0 is

$$L^1 = \prod_{i=1}^{n_1} \lambda_0^{\delta_i}(Z_i, \gamma_k) \prod_{l=1}^{s} S_0^l(T_i|A_i, \gamma_k)\pi_0'(A_i)$$

where γ_k is a multi-dimensional parameter.
The likelihood function resulted from observations under S_1 is

$$L^2 = \prod_{i=1}^{n_2} \alpha_{\delta_i}\lambda_0^{\delta_i}(Z_i, \gamma_i) \prod_{l=1}^{s} S_1^l(T_i|A_i, \gamma_k)\pi_1'(A_i)$$

Then we can write

$$L = L^1 L^2$$

Hence, the log-likelihood function can be expressed as as:

$$\begin{aligned} l &\simeq \sum_{i=1}^{n_1} \log\left[\lambda_0^k(Z_i, \gamma_k)\right]\mathbf{I}(V_i = k) \\ &+ \sum_{i=1}^{n_1}\sum_{k=1}^{s} \log\left[S_0^k(T_i|A_i, \gamma_k)\right] \\ &+ \sum_{i=1}^{n_2} \log\left[\alpha_k\lambda_0^{(k)}(Z_i, \gamma_k)\right]\mathbf{I}(V_i = k) \\ &+ \sum_{i=1}^{n_2}\sum_{k=1}^{s} (1-\alpha_k)\log\left[S_0^k(\tau|A_i, \gamma_k)\right] \\ &+ \sum_{i=1}^{n_2}\sum_{k=1}^{s} \alpha_k \log\left[S_0^k(T_i|A_i, \gamma_k)\right] \end{aligned} \tag{11}$$

IV. Example

Here, we demonstrate the proposed method using tire wear data.

A. Tire Wear

It is known that the degradation path of tire wear is linear and there are several causes that could be redounded to failure for a tire. So, it is seemed that accelerated failure test data for tires may be adopted for our study. We considered a special case that in which there are two competing risks and corresponding intensity functions are as follow:

$$\lambda^k(z,\gamma_k) = (\frac{z}{\theta_k})^{\nu_k}, \quad \gamma_k = (\theta_k, \nu_k), \quad k = 1, 2. \tag{12}$$

This intensity function was proposed by Bagdonavicius *et al.* (2004) for tire wear study.

B. Estimation

Let n_{1k}, n_{2k}, $n_{.k}$ denote respectively, the number of units that fail in normal and high conditions and the total units that fail because of k th competing risk. Under assumption (12), the log-likelihood function is obtained as (15). The MLE's can be obtained by maximizing the log-likelihood through the numerical methods. The likelihood equations as well as some useful relations between the parameters are given in (15), (16) and (17). In normal conditions, the survival function is expressed as follow:

$$\begin{aligned} S(t) &= \mathbf{P}(T > t) = \mathbf{P}\{min(T_0^1, T_0^2) > t\} \\ &= \int_0^\infty \mathbf{P}\{min(T_0^1, T_0^2) > t | A = a\} d\pi(a) \\ &= \int_0^\infty \prod_{k=1}^{2} S_0^k(t|A = a) d\pi(a) \end{aligned}$$

Finally, the estimated survival function is obtained as follow:

$$\begin{aligned} \hat{S}(t) &= \int_0^\infty \prod_{k=1}^{2} \hat{S}_0^{(k)}(t|A = a) d\pi(a) \\ &= \int_0^\infty \exp\left\{-a\hat{\Lambda}_{\cdot}(\frac{t}{a})\right\} d\pi(a) \end{aligned} \tag{13}$$

where $\hat{\Lambda}_{\cdot}(z) = \sum_{k=1}^{2} \hat{\Lambda}^k(z) = \sum_{k=1}^{2} \int_0^z \lambda^k(y, \hat{\gamma}_2) dy$.

C. Data Analysis

Bagdonavicius *et al.* $(2004, 2005)$ studied a real data set based on lifetimes and wears of 101 bus tires. In their study, a failure occurs either by one of two competing risks or by reaching to a critical level of degradation. Time to failure, type of failure and level of degradation at moment of failure for each tire were recorded in this data set. In normal conditions, according to data, $n_{11} = 31$ tires fail because of first competing risk and $n_{12} = 22$ tires fail due to second competing risk. In our study, this part of data set is considered as information which resulted from first level of stress (normal conditions). For the rest of tires $n_{2\cdot} = 48$ the failure times due to competing risks were censored by reaching to critical level of degradation. We assumed that these tires are placed in high conditions and the test continues until remaining tires fail by competing risks. This part of information were simulated. From the previous study by Haghighi and Nikulin (2010) and Levuliene (2002), we know that in normal conditions π belongs to Weibull family with parameters $(\hat{\alpha} = 10.6, \hat{\beta} = 4.77)$ and $\hat{\nu}_1 = 6.883$, $\hat{\nu}_2 = 10.116$, $\hat{\theta}_1 = 23.256$ and $\hat{\theta}_2 = 20$. The stress changing point was set at $\tau = 77$. Using (7), for $k = 1, 2$, the accelerated failure times under high conditions are obtained as follow:

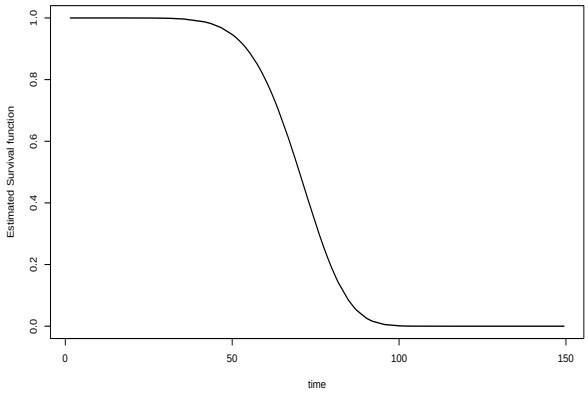


Fig. 2. Estimated Survival function

$$T_2^k = \left[\frac{(A_i\theta_k)^{\nu_k}(\nu_k + 1)(-\log U) - (1 - \alpha_k)\tau^{\nu_k+1}}{\alpha_k}\right]^{\frac{1}{\nu_k+1}} \tag{14}$$

where $U \sim$ uniform$(0, 1)$. Considering medianT_2^1 = median$T_2^2 = 85$, we obtained $\hat{\alpha}_1 = 3.88, \hat{\alpha}_2 = 2.26$. For fixed value of $(\hat{\nu}_1, \hat{\nu}_2, \hat{\theta}_1, \hat{\theta}_2, \hat{\alpha}_1, \hat{\alpha}_2)$ we calculated T_2^1 and T_2^2 and accelerated failure times follows as $T_2 = min(T_2^1, T_2^2)$. The data under normal and high conditions are given in Table I. The maximum likelihood estimates are resulted by maximizing (11) as follow:

$$(10.73, 16.85, 18.90, 17.86, 1.18, 1.44)$$

The estimated survival function are obtained by substituting the MLE's in (13).

$$\hat{S}(t) = \int_0^\infty exp\{-a\sum_{k=1}^{2}(\frac{t^{(\hat{\nu}_k+1)}}{(a\hat{\theta}_k)^{\hat{\nu}_k}(\hat{\nu}_k + 1)})\} d(exp\{-(\frac{a}{\hat{\beta}})^{\hat{\alpha}}\})$$

It is seen that the estimated survival function doesn't have a closed form. Hence, we used Monte-Carlo method for computing it . The Figure 2 shows the estimated survival function versus time for simulated data.

ACKNOWLEDGMENT

Part of this work is supported by a grant from university of Tehran.

$$
\begin{aligned}
l &= \sum_{i=1}^{n1} \nu_k(\log Z_i - \log\theta_k)\mathbf{I}(V_i = k) - \sum_{i=1}^{n1}\sum_{k=1}^{2} \frac{T_i^{\nu_k+1}}{(A_i\theta_k)^{\nu_k}(\nu_k+1)} + \sum_{i=1}^{n2} \{\log\alpha_k + \nu_k(\log Z_i - \log\theta_k)\}\,\mathbf{I}(V_i = k) \\
&- \sum_{i=1}^{n2}\sum_{k=1}^{2} \left\{(1-\alpha_k)\frac{\tau^{\nu_k+1}}{(A_i\theta_k)^{\nu_k}(\nu_k+1)} + \alpha_k \frac{T_i^{\nu_k+1}}{(A_i\theta_k)^{\nu_k}(\nu_k+1)}\right\}, \qquad (15)
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial l}{\partial\theta_k} = 0 &\implies n_{\cdot k} - \sum_{i=1}^{n1} \frac{Z_i^{\nu_1} T_i}{\theta_k^{\nu_k}(\nu_k+1)} - \alpha_k \sum_{i=1}^{n2} \frac{Z_{ik}^{\nu} T_i}{\theta_k^{\nu_k}(\nu_k+1)} - (1-\alpha_k)\tau \sum_{i=1}^{n2} \frac{(\frac{\tau}{A_i})^{\nu_k}}{\theta_k^{\nu_k}(\nu_k+1)} = 0, \\
&\implies \theta_k^{\nu_k} = \frac{\sum_{i=1}^{n1} {Z_i}^{\nu_k} T_i + \alpha_k \sum_{i=1}^{n2} {Z_i}^{\nu_k} T_i + (1-\alpha_k)\tau \sum_{i=1}^{n2} (\frac{\tau}{A_i})^{\nu_k}}{n_{\cdot k}(\nu_k+1)}. \qquad (16)
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial l}{\partial\alpha_k} = 0 &\implies \frac{n_{2k}}{\alpha_k} - \sum_{i=1}^{n2} \left\{\frac{{Z_i}^{\nu_k} T_i}{\theta_k^{\nu_k}(\nu_k+1)} + \frac{\tau^{\nu_k+1}}{(A_i\theta_k)^{\nu_k}(\nu_k+1)}\right\} = 0, \\
&\implies \alpha_k = \frac{\theta_k^{\nu_k}(\nu_k+1) n_{2k}}{\sum_{i=1}^{n2} Z_i^{\nu_k} T_i + \tau^{\nu_k+1} \sum_{i=1}^{n2} \frac{1}{A_i}}. \qquad (17)
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial l}{\partial\nu_k} = 0 &\implies \theta_k^{\nu_k}(\nu_k+1) \sum_{i=1}^{n} (\log Z_i - \log\theta_k)\mathbf{I}(V_i = k) - \sum_{i=1}^{n1} {Z_i}^{\nu_k} T_i (\log Z_i - \log\theta_k - \frac{1}{\nu_k+1}) \\
&- (1-\alpha_k)\tau^{\nu_k+1} \sum_{i=1}^{n2} \frac{1}{A_i^{\nu_k}} \left\{\log\tau - \log(A_i\theta_k) - \frac{1}{\nu_k+1}\right\} - \alpha_k \sum_{i=1}^{n2} {Z_i}^{\nu_k} T_i (\log Z_i - \log\theta_k - \frac{1}{\nu_k+1}) = 0.
\end{aligned}
$$

TABLE I
SIMULATED FAILURE TIMES DATA

Step stress	(T_i, V_i)							
Normal conditions	(58.3,1)	(55.0,2)	(57.1,1)	(56.0,2)	(51.0,1)	(36.8,1)	(55.2,1)	(56.6,2)
	(56.2,1)	(55.0,2)	(51.2,1)	(66.0,1)	(68.9,2)	(57.9,1)	(60.1,1)	(62.8,1)
	(72.0,1)	(72.7,2)	(57.4,1)	(59.3,1)	(65.1,1)	(76.2,1)	(76.1,2)	(75.6,2)
	(75.0,2)	(72.7,1)	(58.9,2)	(67.5,2)	(66.9,1)	(52.1,1)	(56.9,1)	(51.0,2)
	(47.3,1)	(61.3,1)	(65.6,2)	(68.8,1)	(55.0,1)	(53.8,1)	(58.6,1)	(59.2,2)
	(56.8,1)	(69.9,2)	(54.3,1)	(51.7,2)	(59.0,2)	(50.8,1)	(53.0,1)	(59.6,1)
	(61.3,2)	(51.7,2)	(50.0,2)	(68.3,2)	(58.1,2)			
High conditions	(85.54,2)	(85.35,2)	(84.11,1)	(77.75,1)	(83.70,2)	(84.30,2)	(85.30,2)	(81.85,1)
	(81.51,1)	(83.45,2)	(84.04,1)	(79.77,2)	(84.67,2)	(78.26,1)	(77.29,1)	(84.03,1)
	(82.30,1)	(85.23,1)	(84.88,2)	(84.92,2)	(85.54,2)	(82.66,1)	(78.46,1)	(84.08,2)
	(85.22,2)	(79.11,1)	(82.66,1)	(81.01,1)	(84.29,1)	(83.40,2)	(84.49,1)	(82.06,2)
	(85.54,2)	(84.65,1)	(83.11,1)	(85.23,2)	(78.56,2)	(79.99,1)	(77.53,2)	(84.60,2)
	(84.62,1)	(74.16,2)	(79.82,2)	(84.42,1)	(85.51,2)	(85.17,2)	(80.57,1)	(85.33,1)

REFERENCES

[1] V. Bagdonavicius, A. Bikelis and V. Kazakevicius, " Statistical analysis of linear degradation and failure time data with multiple failure modes," *Lifetime data analysis,* no. 10, pp. 65-81, 2004.

[2] V. Bagdonavicius, F. Haghighi and M. Nikulin, " Statistical analysis of general degradation path model and failure time data with multiple failure modes," *Communications in statistics- Theory and methods,* no. 34, pp. 1771-1791, 2005.

[3] G. K.Bhattacharyya and S. Zanzawi, " A tampered failure rate model for step-stress accelerated life test," *Communications in statistics- Theory and methods,* vol. 5, no. 18, pp. 1627-1643, 1989.

[4] F. Haghighi and M. Nikulin, " On the linear degradation model with multiple failure modes," *Journal of applied statistics,* no. 36, pp. 1499-1507, 2010.

[5] R. Levuliene, " Semiparametric estimates and goodness-of-fit tests for tire wear and failure time data," *Nonlinear Analysis: Modelling and Contorl,* no. 1, pp. 61-95, 2002.

[6] W. Q. Meeker and L. A. Escobar, *Statistical methods for reliability data* . New York: John Wiley and Sons, 1998.

[7] W. B. Nelson, *Accelerated Testing Statistical Models, Test Plans, and data Analysis* . New York: John Wiley and Sons, 1990.

[8] F. Pascual, " Accelerated life testing with independent Weibull competing risks with known shape parameter, " *IEEE Trans. Reliability,* no. 56, pp. 85-93, 2007.

[9] F. Pascual, " Accelerated life testing with independent Weibull competing risks, " *IEEE Trans. Reliability,* no. 56, pp. 435-444, 2008.

Procedure for Incorporating a Deterioration in Field Reliability into Test Cost Optimization Decisions

Bahman Honari
National University of Ireland Maynooth
Maynooth, Ireland
bahman.honari@eeng.nuim.ie

John Donovan
School of Engineering
Institute of Technology Sligo
Sligo, Ireland
donovan.john@itsligo.ie

Eamonn Murphy
University of Limerick,
Limerick, Ireland
eamonn.murphy@ul.ie

Abstract—**The optimum duration of an Environmental Stress Test should take into account the cost of the conducting this test versus the cost of warranty failures in the field. The optimum duration is the point at which the total integrated test-field costs are minimized. A cost model is proposed that evaluates the trade-off between time on an environmental stress test and additional warranty costs associated with field failures. This cost model incorporates production test information with field reliability performance. This paper introduces three techniques that can achieve this integration and optimization, Bayesian Networks, Weibull regression and a linear classifier model. It is imperative that ongoing systematic monitoring of early field failures ensures the optimization decision remains appropriate. This ensures that any adverse changes in the production process are identified quickly in order to avoid facing serious reliability and warranty problems. This early detection field failure system also allows one to regularly update the initial test optimization decision that was made based on test-field integrated cost model. This paper develops an integrated test-field cost model that incorporates feedback from the early field failure detection system and allows one to regularly optimize the test duration thereby continuing to minimize the test-field total cost.**

Keywords- Environmental Stress Testing; Field Reliability; Cost optimization

I. INTRODUCTION

Environmental Stress Testing (EST) is the most common approach to precipitating latent defects during the manufacturing of electronic products. This testing consists of applying environmental stresses to the product. Typically these environmental stresses consist of cycling temperature between a high and low extreme, while dwelling at these extremes for specific periods of time. While temperature cycling is the most common approach, it is not the only one and other stresses such as vibration, humidity and power cycling can also be used. The product under evaluation is generally monitored during the test so that any deterioration in performance is detected. The test may be terminated at a point where the latent defects have largely been removed. Determination of the optimum test duration has been traditionally based on the identification of the change point at which infant mortality has largely been removed from the units. The time-on-test is typically the only factor that influences this decision.

On the other hand, monitoring of changes in the pattern of failure field data is an important requirement of any reliability management program. It is imperative that a method be available for the reliability engineers to effectively monitor and detect reliability changes and to understand if the reliability of the product has changed significantly from what was expected. Systematic monitoring of field failures also ensures that adverse changes in the production process are identified quickly in order to avoid facing serious reliability and warranty problems. These kinds of problems are mainly caused by unanticipated failure modes, unknown changes in raw material, changes in operating environmental conditions, etc.

This paper proposes a model that integrates the influence of the production test failures and the field performance including their respective costs into a single unified model. A field reliability deterioration early detection system is then integrated into the total cost model to optimize the test duration for the minimum total cost.

II. TEMPERATURE CYCLING TEST

A temperature cycling test is the most commonly applied environmental stress test. Units are placed in an environmental chamber where the temperature is periodically and gradually changed between extremes of low and high temperature. The high and low temperature extremes are referred to a dwells while the rate at which the temperature changes is referred to as ramps. The temperature and duration of the dwells and the rate of temperature ramp are decided by the production test personnel and will depend on such factors as the time available, the product characteristics and the chamber capability. Each successive portion of the test that includes a decreasing ramp, a cold dwell, an increasing ramp and a hot dwell is called a cycle. The EST will typically contain a number of these cycles and may last for a few hours. Fig. 1 shows the actual temperature cycling profile that was applied to the units studied in this paper.

As the temperature cycling test is an energy intensive and expensive process, it is desirable to reduce the test duration and optimize the test regime to reduce the test cost. This is typically performed after a certain period of time, when confidence has been gained in the product and its test regime. Although a lengthy production test may initially be recommended for a new

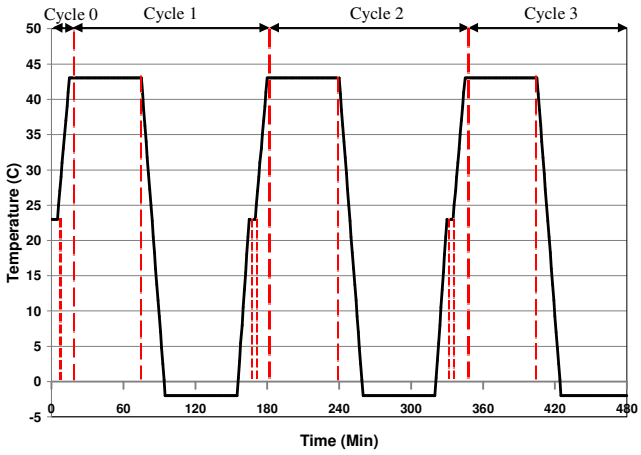


Figure 1. Environmental Stress Test Profile

product, invariably it is the intention to reduce this time based on the product's performance in production and in the field. Readers are referred to [1]-[8] for a useful list of related studies.

III. Optimum Test Duration For Minimizing Test-Field Total Cost

Reliability goals, engineering restrictions and production environments are the factors that affect the EST duration. By reducing the EST duration, the total production cycle time and associated cost can be reduced. However, once latent defects have been removed, any further testing will reduce the useful lifetime in the field. Therefore, there has been strong research motivation to determine the optimal EST duration. In many applications it is desirable to determine a reliability target that minimizes the total test and field cost. If both test and field costs are quantified, as presented in Fig. 2, the optimal total cost can be obtained. In this paper, the optimal EST duration decision is made based on their related costs, i.e. EST fixed and variable costs (C_f, C_v), repair and replacement costs for the EST failure (C_r) and warranty cost (C_w), the lifetime distribution, and repair/replacement models. The optimal EST duration minimizes the total EST-field cost.

A graphical representation of the EST repair and Field process are shown in Fig. 3. The is a specific cost to the manufacturer of performing EST. Should a unit fails during EST, the manufacturer also has to pay to investigate, repair and retest the unit once again. As the EST test is not necessarily perfect, it is possible that some latent defects cannot be detected during the test. These defects may then cause a failure during the warranty period in the field. The immediate result of such a warranty failure is to impose a further cost on the manufacturer. Therefore the test cost is presented as

$$T = C_f + C_v \frac{\tau}{\tau_0} + C_r \qquad (1)$$

where τ_0 and τ represent the initial and reduced test duration respectively. The initial EST duration is 8 hours. On the other hand, a unit that fails during the warranty period in the field imposes the warranty cost C_w on the manufacturer.

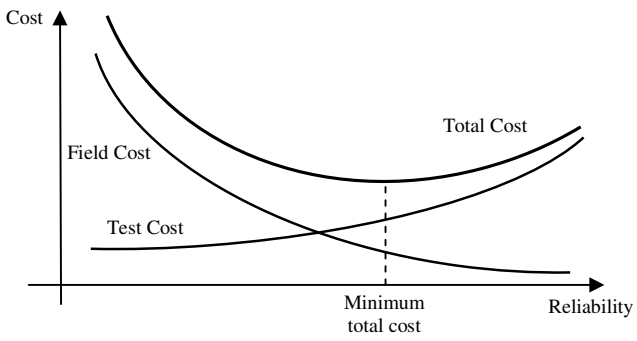


Figure 2. Total Cost Optimization

Furthermore, any unit that fails during the warranty period is returned and will go through the production repair process and is retested once again. This procedure is also shown in Fig. 3. If it passes the test it will be released, and if it fails the test it will be investigated, repaired and put back on test once again. Therefore the cost of such a field failure is presented as:

$$T = C_f + C_v \frac{\tau}{\tau_0} + C_w \qquad (2)$$

Various methodologies have been developed by authors to incorporate both test and field cost into the total cost optimization model. The main purpose of all those methodologies is to relate the units' performance in the field to the EST results. A Weibull regression model can be developed for this purpose [9]. This model can be presented as:

$$\ln(t_p) = \alpha_0 + \sum_i \alpha_i x_i + \frac{1}{\beta} \ln(-\ln(1-p)) \qquad (3)$$

where t_p is the p-quantile of the field failure time distribution x_i [10]. The model shows the dependence of the natural log of the failure time on the predictor x_i which represents the EST parameters. α and β can be estimated using maximum likelihood techniques [11]. The units' EST and field performance can also be related using a Bayesian Network (BN) where the state of variables are defined as the units' EST and field performance and the conditional distributions of the BN variable can be determined from the empirical data [12]. The authors have also proposed a linear classifier model which allows one to predict and discriminate the units' field performance based on their behavior in the EST [13].

The EST duration is reduced in a step manner in order to determine the optimum EST termination time. Suppose the EST duration was reduced to a time τ less than eight hours, then the proportion of units that failed in EST between (τ, τ_0) will no longer fail in EST because of the reduction in the test duration. We assume that the proportion of units that have typically failed during EST after time τ will now fail in the field during the warranty period. If we assume that the EST now terminates at time τ, then we should add the portion of the units that fail after time τ, to the field failures. This is shown in Fig. 4.

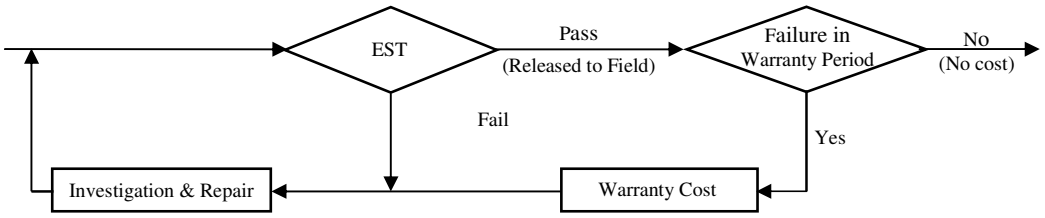


Figure 3. Graphical Representation of the EST Repair and Field Process

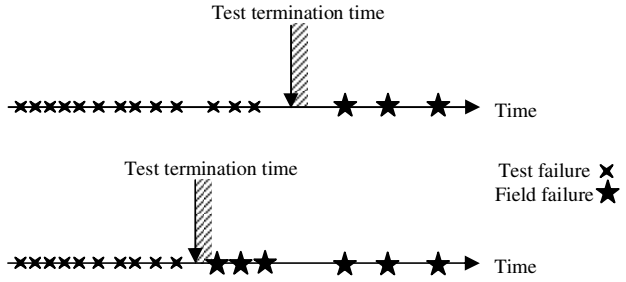


Figure 4. Late Test Failures Appear as Early Field Failures when Test Duration Is Reduced

IV. DETECTING CHANGES IN FIELD RELIABILITY

The total cost minimization analysis discussed in the last section is based on the information about the test and field performance for the units that were produced, installed and returned in the period of study. Although this piece of information can be periodically updated, the obtained results do not allow one to regularly monitor and study the effect of a possible deterioration in field reliability. Therefore the total cost optimization of test and field needs to be combined with a early detection field failure system. This combined model incorporates the field reliability variations into the total cost optimization decision.

The product return rate represents a typical measure for reliability performance in the field. This is generally defined as the total number of products received for servicing or repair divided by the total installed base of the product [14]. Occasionally this may be modified to take into account the hours that the product was actually operating, if this were known. These return rates are generally calculated on a monthly and yearly basis. However, as most products do not have constant failure rates and the number of units installed each month rarely remains constant, a change in reliability may go unnoticed for a long time. For any particular month, the number of products returned is a function of the number of units installed and portion of the reliability bathtub curve they represent [15].

To detect reliability deterioration, the monitoring procedure starts as soon as the field returns data become available. To detect the reliability changes, the dataset is stratified into different periods (months were chosen in this research). The monitoring procedure is periodically repeated at the end of each period as more data are accumulated.

A. Notation and Problem Formulation

The following notation is used to explain the early detection procedure. Let n_i denotes the number of units produced in the i^{th} period. We assume n_{ij} represents the number of units that were produced in month i and installed in month j. Finally, r_{ijk} denotes the observed value for the random variable R_{ijk} that represents the number of units produced in month i, installed in month j and returned after k months service. For instance, r_{231} denotes number of units that produced, installed and returned in months 2, 3 and 4, respectively. Table 1 shows the time for each i, j and k, that the values of r_{ijk} that are available.

In this paper, the warranty period for the units is assumed to be 6 months. R_{ijk}s are assumed to independently follow the Poisson distribution $P(n_{ij}, \lambda_k)$, where λ_k denotes the rate of field returns after k months in service. The reference value for is denoted by λ_k^0 which is actually the expected return rate after k months in service. This can be obtained from the historical data or from similar products.

B. Hypothesis Testing and False Alarms

The hypothesis test of

$$H_0: \lambda_0 \le \lambda_0^0,\ \lambda_1 \le \lambda_1^0, \dots, \lambda_p \le \lambda_p^0 \quad (4)$$

$$H_1: \exists\, k \mid \lambda_k \le \lambda_k^0\ ,\ k = 0,1,\dots,p \quad (5)$$

presents the statistical rule for the early detection procedure where p denotes the number of periods for which the field returns will be monitored [14]. This number of periods may vary based on the reliability targets. In this paper the warranty period is assumed to be 6 months. However, the period for monitoring the field returns could be extended for longer.

In month 1, among the n_1 manufactured units, n_{11} were installed. However r_{110} of units failed during the same month (i.e. 0 months in service) and were returned to the manufacturer. r_{110} is the observed value for the random variable R_{110} that follows the distribution $P(n_{11}, \lambda_0)$. In this period the available data are only sufficient to test $H_0: \lambda_0 \le \lambda_0^0$ vs. $H_1: \lambda_0 > \lambda_0^0$. This test is based on comparing r_{110} to the critical value c_{110} that is determined base on α, the test level of significance.

TABLE I. AVAILABILITY OF INFORMATION FOR FIELD RETURNS

	Data Available in Month						
Installed in Month	1	2	3	4	5	6	Produced in Month
1	r_{110}	r_{111}	r_{112}	r_{113}	r_{114}	r_{115}	1
2		r_{120}	r_{121}	r_{122}	r_{123}	r_{124}	
3			r_{130}	r_{131}	r_{132}	r_{133}	
4				r_{140}	r_{141}	r_{142}	
5					r_{150}	r_{151}	
6						r_{160}	
2		r_{220}	r_{221}	r_{222}	r_{223}	r_{224}	2
3			r_{230}	r_{231}	r_{232}	r_{233}	
4				r_{240}	r_{241}	r_{242}	
5					r_{250}	r_{251}	
6						r_{260}	
3			r_{330}	r_{331}	r_{332}	r_{333}	3
4				r_{340}	r_{341}	r_{342}	
5					r_{350}	r_{351}	
6						r_{360}	
4				r_{440}	r_{441}	r_{442}	4
5					r_{450}	r_{451}	
6						r_{460}	
5					r_{550}	r_{551}	5
6						r_{560}	
6						r_{660}	6

At the end of month 2, sufficient data for 3 tests are available. The first test is related to λ_0 for the units produced in month 1, regardless of installation time. For this test, the accumulated number of returns $r_{110} + r_{120}$ is compared to the critical value $c_{110} + c_{120}$. The next test corresponds to λ_0 for the units produced in month 2. For this test, r_{220} is compared to the critical value c_{220}. Finally, the last test at the end of month 2, is performed λ_1 for by comparing r_{111} to critical value c_{111}. Using the same procedure, hypothesis testing for other λ_is can be conducted based on the accumulated data from the following months.

One should note that as R_{ijk}s are independent, the tests in (4) can be performed individually. Furthermore, the overall probability of false alarm for the test in (4) can be presented as

$$\alpha = 1 - \prod_{k=0}^{p}(1 - \alpha_k) \qquad (6)$$

where α_k is the probability of false alarm in testing λ_k.

V. INCORPORATING THE EARLY DETECTION PROCEDURE INTO THE TOTAL COST OPTIMIZATION DECISION

Section 3 explained the motivation and methodologies for test-field total cost optimization. The philosophy and procedure for the early detection of field reliability was discussed in section 4. Section 5 provides an algorithm to combine the early detection technique with the total cost optimization decision. The proposed algorithm allows one to incorporate the updated information from the early detection procedure into the total cost minimization methodology.

Fig. 5 shows the proposed algorithm that starts at the end of month i. The total test-field cost is initially optimized based on the techniques described in section 3. At the end of month i+1 the early detection technique is applied to the available set of data. If any deterioration in field reliability is detected, the EST duration will be reviewed and the units produced in the next month are put through the revised optimized test. Otherwise, the same EST is applied to the units and the early detection test will be conducted for the units produced in month i+2, and so on.

The most important point in using the algorithm is determining the required values of α_k. Allocating proper values to α_k depends on different factors of the reliability plan. For example it may be assumed that detecting the early failures is critically important. However, in this paper we assume that a false alarm for the later months of monitoring, cause more costly changes to the EST optimization decision. This is because it is assumed that detecting the latent defects that are observed during the later months of warranty period requires a longer, intensive and so more expensive EST. Therefore, a false alarm for the later months of warranty period is assumed to be more critical. As a result one can simply assume

$$\alpha_k \propto \left(\sum_{i=1}^{p} i\right) - k \qquad (7)$$

Obviously, different weighting or allocation plans for α_k s can be developed. Critical values for rejecting the null hypothesis can then be determined.

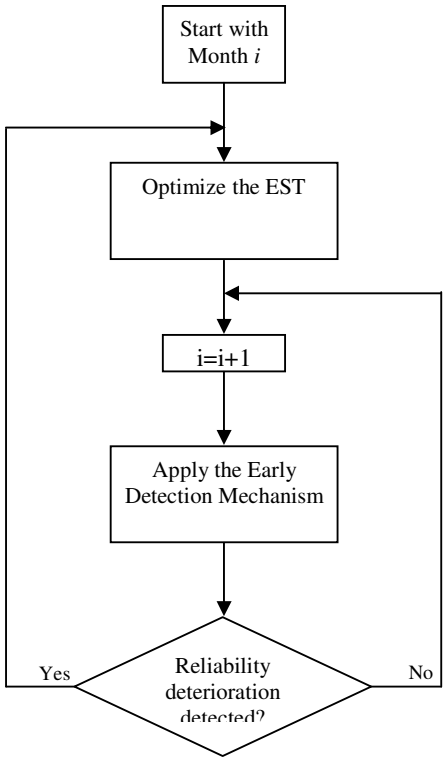


Figure 5. An Algorithm for Incorporating the Field Reliability Variations in Test-Field Total Cost Optimization

VI. NUMERICAL EXAMPLE

Real-life production and field data was used in the official verification of this proposed algorithm. However, due to confidentiality reasons only simulated data could be used for the purposes of illustrating the technique in this paper. Significant work was conducted on the simulated data to ensure that it represented a legitimate structure, similar to that of the real data in all major respects. Each record in the database represented an individual unit and included its EST and functional test results and field performance.

The EST duration was initially 480 minutes. This duration was then optimized based on a dataset of 20 months test and field failure data. For typical values of $C_f = \$6$, $C_v = \$2$, $C_r = \$6$ and $C_w = \$90$ the optimum EST duration was determined to be 360 minutes. The expected failure rate in each month is also determined based on this dataset. The maximum number of failure in each month is then obtained based on the expected failure rate, the number of units installed and the probability of false alarm determined by (6). These values are shown in Table 2. The total probability of false alarm is assumed to be 0.05.

The new set of units was manufactured at the beginning of the month that in early detection procedure is called month 1. All units that manufactured in each month are installed in the same month. The number of returned units for the following months is shown in Table 3. At the end of month 1, the only available information is r_{110} and this is used to perform the hypothesis test for λ_0. As the null hypothesis is not rejected, the same EST duration (360 minutes) is applied to the units produced in month 1. This continued until month 3 when the early detection test signals an alarm as there is a significant increase in the number of returned units. The EST is then optimized by applying the methodologies described in Section 3. The result is presented in Fig. 6. As the result shows the optimum EST duration may now be set at 390 minutes with the total test-warranty cost of $11.36 per unit.

TABLE II. THE MAXIMUM NUMBER OF EXPECTED FAILURES

Month (j)	Number of units installed	Number of months in service (k)					
		0	1	2	3	4	5
1	4053	1	5	4	5	9	6
2	2151	1	6	7	11	3	
3	1796	1	2	0	1		
4	3622	3	1	3			
5	1493	2	1				
6	2680	4					

TABLE III. THE NUMBER OF INSTALLED AND RETURNED UNITS

Month (j)	Number of units installed	Number of months in service (k)					
		0	1	2	3	4	5
1	4053	13	12	11	11	11	11
2	2151	7	7	7	7	7	
3	1796	6	5	5	5		
4	3622	11	10	10			
5	1493	5	5				
6	2680	7					

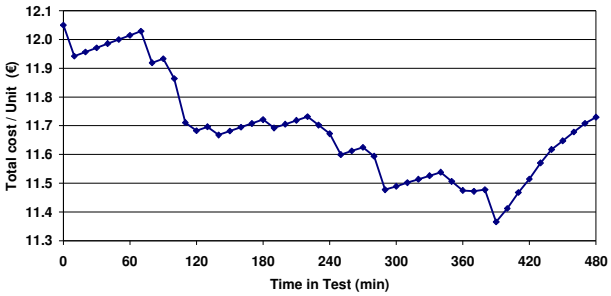


Figure 6. Optimum EST Duration Determined at the end of Month 3

One should note that any decision for changing the EST duration has to be treated carefully. Applying the revised optimum EST duration to the production process may not be recommended. In fact, different factors such as cost and technical aspects of this change, the amount of information that used in determining deterioration, and other factors have also to be considered.

VII. CONCLUSION AND FUTURE WORKS

In this paper a model was presented to determine and optimize the test-field total cost. An early detection technique for field reliability was presented that enables one to detect a deterioration in the field reliability. Those two models then were combined in an algorithm to incorporate the field reliability changes into the EST optimization decision. This algorithm provides a platform for combining the test-field total cost optimization decision and an early detection mechanism. It was shown that negative changes in the field reliability may necessitate a longer (or more intensive) EST in order to detect more latent defects and avoid them occurring in the field during the warranty period. However, applying the EST with the revised optimum duration, in practice, depends on many other factors in the manufacturing environment.

While the proposed technique was applied to the temperature cycling test, a similar algorithm can be applied to other environmental stress tests.

Furthermore, based on the product reliability target, various assumptions for false alarm allocation to the different periods of monitoring can be considered.

ACKNOWLEDGMENT

This work was supported by Science Foundation Ireland under SFI grant 03/CE3/1405.

REFERENCES

[1] F. Jensen and N.E. Petersen, Burn-In: an engineering approach to the design and analysis of burn-in procedures, John Wiley & Sons, 1983

[2] S. R. Dalal and C.L. Mallows, "Some graphical aids for deciding when to stop testing software," IEEE Journal on Selected Areas in Communications, vol. 8, No.2; pp. 169-175, 1990

[3] W.Q. Meeker and L. Escobar, "A review of accelerated test models," Statistical Science, 21(4); pp. 552-577, 2006.

[4] H. W. Block and T. H. Savits, "Burn-In," Statistical Science, 12(1), pp. 1-19, 1997.

[5] W. Kuo and Y. Kuo, “Facing the headaches early failures: A state of-the art review of burn-in decisions,” Proc. of IEEE, vol. 71(II), pp.1257–1266, 1983.

[6] L. M. Leemis and M. Beneke, “Burn-in models and methods: A review,” IIE Transactions, pp. 172-180, 1990.

[7] G. S. Watson and W. T. Wells, “On the possibility of improving mean Useful life of items by eliminating those with short lives,” Technometrics, vol 3, pp. 281-298, 1961.

[8] L. Stewart and J. D. Johnson, “Determining optimum burn-in and replacement time using Bayesian decision theory,” IEEE Transactions on Reliability, 21(3), pp. 170–175, 1972.

[9] B. Honari, J. Donovan, T. Joyce and S.P. Wilson, “Application of generalized linear models for optimizing production stress testing,” Proc. Annual Reliability and Maintainability Symp., pp. 267-271, 2008.

[10] P. McCullagh, and J.A. Nelder, Generalized Linear Models, Chapman & Hall, 1999.

[11] D.N. Prabhakar Murthy, M. Xie and R. Jiang, Weibull Models, Wiley Interscience, 2004.

[12] B. Honari, J. Donovan, E. Murphy, “Applying Bayesian Networks to test-field cost optimization decisions”, Proc. ISSAT, 15th Conference on Reliability and Quality in Design, (Aug.), pp 170-175, 2009.

[13] B. Honari, J. Donovan, E. Murphy, “Application of mixed integer linear classifier in test-field cost optimization decisions”, Proc. the 26th International Manufacturing Conference, Sept. 2009, pp 321-328.

[14] H. Wu, W. Q. Meeker, “Early detection of reliability problems using information from warranty databases”, Technometrics, vol. 44, no. 2, 2002, pp 120-133.

[15] B. Honari, J. Donovan, T. Joyce, S. Wilson, “Early Detection of Reliability Changes for A Non-Poisson Life Model Using Field Failure Data”, Proc. Ann. Reliability and Maintainability Symp., (Jan.) 2007, pp 346-349.

Within the sample Comparison of prediction performances of models and sub-models. Application to Alzheimer disease.

Catherine Huber-Carol
Université Paris Descartes
45 rue des Saints-Pères
75006 Paris, FRANCE
and INSERM U 1018, 94804 Villejuif
16 bis av. P.V. Couturier
catherine.huber@parisdescartes.fr

Shulamith T. Gross
Bernard M. Baruch College
Statistical Consulting Laboratory
City University of New York
One Baruch way
NY 10010, USA
shulamith.gross@baruch.cuny.edu

Annick Alpérovitch
Université Pierre et Marie Curie
Unté INSERM 708
CHU Pitié-Salpetrière
91 bd de l'Hôpital
75013 Paris, FRANCE
annick.alperovitch@upmc.fr

***Résumé*—Our objective is to compare the predictive ability of several nested models. It stems from the following problem in epidemiology : the occurrence of a certain disease is to be predicted to happen within a fixed period of time thanks to the values of a number of items measured on the observed patients. It may happen that one or several items, proved to be relevant for the best fitting model, have a non significant contribution to the prediction of who is at risk of developing the disease. The indices we use to compare the respective predictive ability of two models are the Integrated Discrimination Improvement (IDI) and the BRier's score Improvement (BRI). Estimation of the models and their relative IDI and BRI are conducted on the same sample, and their respective asymptotic properties are proved. We apply the results to Alzheimer disease on a French cohort.**

I. Introduction

When the objective of modeling a data set is explanatory, it is most appropriate to choose the best fitting model using the usual model selection procedures. But if the objective is to predict and not to explain the facts, and some of the factors selected by the "best fitting model" are not available for all subjects, as is sometimes the case for certain genetic markers, one can compare models by their prediction qualities rather than by their goodness of fit to the data. In this paper, we consider the case of two competing, nested, probability predicting models consistent with Pencina et al. (2008) setting. The nested model contains traditional factors, and the larger model contains in addition some expensive, or generally hard to obtain, often genetic, relevant markers.

II. Framework

A. General description of the data set and the models

The data set is as follows. $X = (Y, \mathbf{Z})$ is a random variable such that the response variable Y is binary with values in $\{0, 1\}$, and $\mathbf{Z}$ is a k-dimensional real variable. Observed is $\mathbf{X} = (X_1, \cdots, X_n)$, n i.i.d. observations of X, and two models for predicting Y on the basis of $\mathbf{Z}$ are to be compared :

$$\begin{array}{lrcl} \text{Model 1} & P(Y=1|\mathbf{Z}=\mathbf{z}) & = & p_1(\mathbf{z}) \\ \text{Model 2} & P(Y=1|\mathbf{Z}=\mathbf{z}) & = & p_2(\mathbf{z}) \end{array}$$

while the true distribution of Y given $Z = z$, which remains unknown all along, is given by

$$P(Y=1|\mathbf{Z}=\mathbf{z}) \quad = \quad p(\mathbf{z})$$

This setting originates from a problem in epidemiology : Y_i is the indicator of the occurrence of a specific disease for subject i within a given period of time. The prediction of occurrence of this event is based on the value $\mathbf{z_i}$ of $\mathbf{Z}$ observed on subject i. In the special case of linear logistic models, p_1 and p_2 are denoted g_1 and g_2. While g_1 is including all k components of $\mathbf{Z}$, g_2 is obtained by dropping k" components of $\mathbf{Z}$, keeping thus only $k' = k - k" < k$ of them. Without restriction of the generality, we treat in detail the case $k" = 1$. The theoretical aim of this work is to derive the asymptotic properties of the estimators of the IDI and the BRI in order to obtain for them confidence intervals within the sample used to estimate the two models.

B. Definition of the performance prediction criteria : IDI and BRI

From Pencina et al (op. cit.), the IDI of model 2 with respect to model 1 is

$$\begin{aligned} IDI_{2/1} &= E[p_2(\mathbf{Z}) - p_1(\mathbf{Z})|Y=1] \\ &\quad -E[p_2(\mathbf{Z}) - p_1(\mathbf{Z})|\mathbf{Y}=\mathbf{0}] \end{aligned}$$

where E denotes the expectation with respect to the distribution of X. We denote

$$\pi = P(Y=1) := E[p(\mathbf{Z})] ,$$

the population prevalence of the event under study. We derive a simpler expressions for $IDI_{2/1}$ to be used later :

$$IDI_{2/1} = E\left[(p_2(\mathbf{Z}) - p_1(\mathbf{Z}))\left(\frac{Y-\pi}{\pi(1-\pi)}\right)\right] \quad (1)$$

Gu and Pepe define the PEV (proportion of explained variance) of a model :

$$PEV = \frac{var(P(Y=1|Z))}{\pi(1-\pi)} \quad (2)$$

from which it is clear that :

$$IDI_{2/1} = PEV_2 - PEV_1 \quad (3)$$

For a single model p the Brier score is defined as

$$BR(p) = E[(Y - p(\mathbf{Z}))^2] \quad (4)$$

As the bigger the Brier's score the worst is the model, we define the BRier's score improvement for model 2 with respect to model 1 as

$$\begin{aligned} BRI_{2/1} &= BR(p_1) - BR(p_2) \\ &= E[(p_1(\mathbf{Z}) - p_2(\mathbf{Z}))\times \\ &\quad (p_1(\mathbf{Z}) + p_2(\mathbf{Z}) - 2Y)] \end{aligned} \quad (5)$$

A negative $IDI_{2/1}$ as well as a negative $BRI_{2/1}$ means that the predictive properties of model 2 are not as good as those of model 1. The ranges of IDI and BRI are respectively -2 , $+2$ and -1 , $+1$.

III. Estimation of IDI and BRI

Now we assume that we have a sample of size n of X, and the two prediction models $p_j(\theta_j, \mathbf{z})$, defined for $j = 1, 2$ through the respective parameters $\theta_\mathbf{1}$, $\theta_\mathbf{2}$, as :

$$\begin{aligned} P(Y=1|\text{ model 1}) &= p_1(\theta_\mathbf{1}, \mathbf{z}) \\ P(Y=1|\text{ model 2}) &= p_2(\theta_\mathbf{2}, \mathbf{z}) \end{aligned}$$

A. General estimating equations for IDI and BRI

Using (1) and (5), and denoting for simplicity

$$\widehat{p_{ji}} := p_j(\widehat{\theta_j}, \mathbf{z_i}) \quad \text{for } \mathbf{j} = \mathbf{1}, \mathbf{2}, \mathbf{i} = \mathbf{1}, \cdots, \mathbf{n}. \quad (6)$$

where $\widehat{\theta_j}, j = 1, 2$ are maximum likelihood estimates for the parameters of models 1 and 2, natural estimates of $IDI_{2/1}$ and $BRI_{2/1}$ are respectively

$$\widehat{IDI_{2/1}} = \frac{1}{n}\sum_{i=1}^{n}(\widehat{p_{2i}} - \widehat{p_{1i}})\frac{y_i - \overline{y}}{\overline{y}(1-\overline{y})} \quad (7)$$

$$\widehat{BRI_{2/1}} = \frac{2}{n}\sum_{i=1}^{n}[(\widehat{p_{1i}} - \widehat{p_{2i}})(\frac{(\widehat{p_{1i}} + \widehat{p_{2i}})}{2} - Y)] \quad (8)$$

Under usual regularity conditions on models 1 and 2, their parameters estimates are asymptotically consistent and normal.

B. Estimation of IDI and BRI in the logistic case

Let $u =< \theta, \mathbf{z} >$ be the scalar product of two $k+1$ dimensional real vectors $\theta = (\theta_0, \cdots, \theta_k)$, $\mathbf{z} = (1, z_1, \cdots, z_k)$) and g the function

$$g(u) = \frac{e^u}{1+e^u} \quad (9)$$

Models 1 and 2 are logistic : $p_1 \equiv g_1$ and $p_2 \equiv g_2$:

$$\begin{aligned} g_1(\mathbf{z}) &= g(< \theta_\mathbf{1}, \mathbf{z} >) \\ g_2(\mathbf{z}) &= g(< \theta_\mathbf{2}, \mathbf{z} >) \end{aligned}$$

where some components of $\theta_\mathbf{1}$ and some (possibly different) components of $\theta_\mathbf{2}$ are predefined. If we refer to the motivating example of the introduction :

$$\begin{aligned} \theta_\mathbf{1} &= (\theta_0, \theta_1, \cdots, \theta_{k-1}, \theta_k) \\ \theta_\mathbf{2} &= (\theta'_0, \theta'_1, \cdots, \theta'_{k-1}, 0) \end{aligned}$$

Dropping the index j for simplicity of notation, we get the log-likelihood L_n :

$$\begin{aligned} L_n(\theta) &= \frac{1}{n}\sum_{i=1}^{n} y_i \log(g(< \theta, \mathbf{z_i} >)) \\ &\quad +(1-y_i)\log(1 - g(< \theta, \mathbf{z_i} >)) \\ &= \frac{1}{n}\sum_{i=1}^{n}[y_i < \theta, \mathbf{z_i} > - \log(1 + e^{<\theta, \mathbf{z_i}>})] \end{aligned} \quad (10)$$

Then we estimate the parameters θ_1 and θ_2 of the two logistic models. Those two estimators, obtained through classical maximum likelihood equations, are proved to be consistent and asymptotically normal.

1) Asymptotics of $\widehat{IDI_{2/1}}$ for logistic predictors: For simplicity, and since there is no ambiguity as we always consider how model 2 behaves with respect to model 1, we now drop the index $2/1$.
From equation (7) we get

$$\begin{aligned}\widehat{IDI} &= \frac{1}{n}\sum_{i=1}^{n}\left[(\hat{g}_{2i}-\hat{g}_{1i})-(g_{2i}-g_{1i})\right]\frac{y_i-\overline{y}}{\overline{y}(1-\overline{y})}\\ &+ \frac{1}{n}\sum_{i=1}^{n}[g_{2i}-g_{1i}]\,\frac{y_i-\overline{y}}{\overline{y}(1-\overline{y})}\end{aligned}$$

By analyzing the two terms of this decomposition of $\widehat{IDI}$, we get the two following theorems (S.T. Gross et al, 2012)

Theorem 1 (Consistency of $\widehat{IDI}$):

$$\widehat{IDI} \xrightarrow[n\to\infty]{a.s.} IDI.$$

Theorem 2 (CLT of $\widehat{IDI}$):

$$\sqrt{n}(\widehat{IDI}-IDI) \xrightarrow[n\to\infty]{\mathscr{L}} \mathscr{N}(0,\sigma^2). \qquad (11)$$

where $\sigma^2=(\frac{1}{(1-\pi)\,\pi})^2\,var(V)$ with V defined as

$$\begin{aligned}V &= (g(\theta_2,\mathbf{Z_2})-g(\theta_1,\mathbf{Z_1})-E_\Delta)(Y-\pi)\\ &+ (Y-g(\theta_2,\mathbf{Z_2}))\;{}^t(\mathbf{Z_2})(I^{-1}(\theta_2))E_2\\ &- (Y-g(\theta_1,\mathbf{Z_1}))^t(\mathbf{Z_1})(I^{-1}(\theta_1))E_1\\ &+ IDI\,(2\pi-1)\,(Y-\pi)\end{aligned}$$

E_Δ and $E_j,\ \ ,\ \ j=1,2$ are expectations involving the functions g of the models, and $\sigma^2(V)$ can be consistently estimated by its empirical analog. $\widehat{\sigma^2}$.

The estimated $V's$ are obtained by replacing all parameters by their maximum likelihood estimates and expectations by sample averages.

2) Asymptotics of $\widehat{BRI_{2/1}}$ for logistic predictors: Dropping again the index 2/1, we consider the estimated BRI (III-B2).

$$\begin{aligned}\widehat{BRI} &= \frac{2}{n}\sum_{i=1}^{n}[(g(<\hat{\theta}_1,Z_i>)-g(<\hat{\theta}_2,Z_i>))\times\\ &\times\ (\frac{(g(<\hat{\theta}_1,Z_i>)+g(<\hat{\theta}_2,Z_i>))}{2}-Y_i)]\end{aligned}$$

Using the same method as for IDI, we get

$$\sqrt{n}(\widehat{BRI}-BRI) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(W_i-BRI)+o_P(1)$$

where the random variables W_i are defined as

$$\begin{aligned}W_i &= 2[(g(<\theta_{\mathbf{1}},\mathbf{Z}_i>)-g(<\theta_{\mathbf{2}},\mathbf{Z}_i>))\times\\ &\times\ (\frac{(g(<\theta_{\mathbf{2}},\mathbf{Z}_i>)+g(<\theta_{\mathbf{1}},\mathbf{Z}_i>))}{2}-Y_i)]\\ &- (Y_i-g(<\theta_{\mathbf{2}},\mathbf{Z}_{2i}>))^tZ_{2i}I^{-1}(\theta_2)(2E_2-E_4)\\ &+ (Y_i-g(<\theta_{\mathbf{1}},\mathbf{Z}_{1i}>))^tZ_{1i}I^{-1}(\theta_1)(2E_1-E_3)\end{aligned}$$

where the $E_j, j=1,\cdots,4$ are expectations involving the functions g defining the two models.
Actually, we have,

Theorem 3 (Consistency of $\widehat{BRI}$):

$$\widehat{BRI} \xrightarrow[n\to\infty]{a.s.} BRI \qquad (12)$$

Theorem 4 (CLT for $\widehat{BRI}$):

$$\sqrt{n}(\widehat{BRI}-BRI) \xrightarrow[n\to\infty]{\mathscr{L}} \mathscr{N}(0,\sigma_B^2). \qquad (13)$$

A consistent estimate of $\sigma_B^2=var(W)$ is given by

$$\widehat{\sigma_B^2}=\frac{1}{n-1}\sum_{i=1}^{n}(\widehat{W_i}-\hat{\overline{W}})^2.$$

The estimated W's are obtained by replacing all parameters by their maximum likelihood estimates and expectations by sample averages.

IV. Simulation studies

Our basic model is a logistic one with two independent covariates : $Z_1\in\{-1,0,1\}$ with respective probabilities $(.2,.4,.4)$, and $Z_2\in\{0,1\}$ with respective probabilities $(.2,.8)$, with $\theta=(0,2,1)$ so that

$$\mathscr{L}(Y_i|Z_i)=\text{Bernoulli}(\frac{\exp(2Z_{1i}+Z_{2i})}{1+\exp(2Z_{1i}+Z_{2i})}).$$

We generate 1000 samples of 1000 triplets (Y_i,Z_{1i},Z_{2i}). On each sample, three logistic models are considered : model 1 including both covariates, $\theta_{\mathbf{1}}=(\theta_{10},\theta_{11},\theta_{12})\equiv(0\ ,\ 2\ ,\ 1)$, model 2 including only Z_1, $\theta_{\mathbf{2}}=(\theta_{20},\theta_{21},0):=(0.096\ ,\ 1.969\ ,\ 0)$ and model 2' including only Z_2, $\theta_{\mathbf{2'}}=(\theta_{2'0},0,\theta_{2'2}):=(0.307\ ,\ 0\ ,\ 0.674)$. The true values of $\theta_{\mathbf{2}}$ and $\theta'_{\mathbf{2}}$ are obtained by minimizing the Kullback distance between the true law and the logistic based on covariate Z_1 alone and Z_2 alone respectively. The true values of $IDI_{2/1}$ and $IDI_{2'/1}$ can thus be computed :

The true values of the Brier's score of the three models lead to the true values of the BRI of models 2 and 2' with respect to model 1 :

TABLE I
TRUE VALUES OF $IDI_{2/1}$ AND $IDI_{2'/1}$

	True values
$IDI_{2/1}$	- 0.01037
$IDI_{2'/1}$	- 0.3282

TABLE II
TRUE VALUES OF BRIER'S SCORES AND BRI'S.

	Brier's score	BRI with respect to model 1
model 1	0.1603	0
model 2	0.16281	- 0.002511
model 2'	0.23962	- 0.079316

TABLE III
95% CONFIDENCE INTERVALS FOR $IDI_{2/1}$ AND $IDI_{2'/1}$.

	95% confidence	interval	True value
$IDI_{2/1}$	-0.02140714	-0.00076843	- 0.01037
$IDI_{2'/1}$	-0.38141	-0.27829	- 0.3282

TABLE IV
95% CONFIDENCE INTERVALS FOR $BRI_{2/1}$ AND $BRI_{2'/1}$.

	95% confidence	interval	True value
$BRI_{2/1}$	-0.00535781	-0.00067996	-0.002511
$BRI_{2'/1}$	-0.090107	-0.069466	-0.079316

For each sample, we estimated the three models, the two IDI's, $IDI_{2/1}$ and $IDI_{2'/1}$, as well as their standard errors. The same for the two BRI's. A comparison of the mean estimated asymptotic, as well as the empirical standard errors of the estimated IDI's and BRI's proved to be fairly close.

V. THE THREE CITY STUDY OF ALZHEIMER DISEASE.

The Three-City (3C) study is a cohort study conducted in three cities in France (Bordeaux, Dijon, and Montpellier), aiming to estimate the risk of dementia and cognitive impairment attributable to vascular factors. We were given a sub-sample of n = n=4214 men and women among which 162 developed Alzheimer within four years. The best fitting model included the genetic marker APOE4 :

Model 1 that includes APOE4 is significantly a better fit to the data than model 2. But we show that excluding APOE4 does not decrease significantly the prediction ability of model 2, by estimating $IDI_{2/1}$ and $BRI_{2/1}$ directly from the data and from a bootstrap with 1000 repetitions to obtain a competing estimator of their standard errors, as well as 95% confidence intervals.

TABLE V
MODEL 1 :LOGISTIC MODEL INCLUDING APOE4.

	Estimate	Std. Error	Pr(>|z|)
(Intercept)	-2.944	0.176	< 2e-16
age.fac.31	-2.089	0.330	2.3e-10
age.fac.32	-0.984	0.191	2.5e-07
nivetudes.spec	-0.430	0.180	0.0167
card	0.616	0.233	0.0081
depress	0.786	0.201	9.5e-05
incap	1.180	0.206	1.1e-08
APOE4	0.634	0.195	0.0012
AIC	=	1094	

TABLE VI
MODEL 2 : LOGISTIC MODEL WITHOUT APOE4.

	Estimate	Standard Error	p-value
(Intercept)	-2.797	0.168	< 2e-16
age.fac.31	-2.060	0.330	4.0e-10
age.fac.32	-0.963	0.190	4.2e-07
nivetudes.spec	-0.434	0.179	0.0155
card	0.677	0.231	0.0034
depress	0.805	0.201	6.2e-05
incap	1.124	0.206	4.6e-08
AIC	=	1102	

VI. CONCLUSION.

Those results enable researchers to do inference on two indices for measuring the relative effectiveness of two models in predicting the probabilities of future events. We allow for model estimation prior to index computation on the same data by providing new standard errors for both the IDI and the BRI when the indices are computed on the same data that provided model parameter estimates. We consider now additional indices, in particular the difference in the area under the ROC curve of two models and Pencina et al's (2008) Net Reclassification Improvement NRI, and are currently working on the asymptotic theory for these indices when parameters are estimated from the same data as the indices.

RÉFÉRENCES

[1] Shulamith T. Gross, Catherine Huber-Carol and Annick Alpérovitch, *Within the Sample comparison of Statistical Models Prediction, with Applcation to Alzheimer.*, submitted, 2012.

[2] Wen Gu and Margaret Pepe, *Measures to summarize and compare the predictive capacity of markers.*, The International Journal of Biostatistics, vol. 5, issue 1,The Berkeley Electronic Press., 2009.

[3] M. J. Pencina , R. B. D'Agostino Sr and R. B. D'Agostino Jr, *Evaluating the added predictive ability of a new marker. From area under the ROC curve to reclassification and beyond.*, Statistics in Medicine, vol. 27, 157–172, 2008.

Reliability Growth Models for Series Systems

Patricia A. Jacobs & Donald P. Gaver
Operations Research Department
Naval Postgraduate School, Monterey, CA 93943 USA
pajacobs@nps.edu & dgaver@nps.edu

***Abstract*—A series system of $S \geq 1$ stages initially has $m_s(0)$ failure modes (FMs) in stage s, $s=1,2,\ldots,S$. A stage, s, operates on test $t \geq 1$ if the system operates at all previous stages on test t, *and* on stage s. If at least one FM activates in stage s, then the test stops and stages $s+1,\ldots,S$ are not accessed. Under stagewise Poisson failure mode assumptions and Negative Binomial approximations, we calculate the probability of system survival after a fixed number of tests. Results are compared with simulations.**

Keywords-series system; failure mode masking; Negative Binomial distribution; Poisson distribution

I. PROBLEM FORMULATION

Consider a system that functions serially in stages $s=1,2,\ldots,S$. All stages must operate for a test or mission to succeed. However, inevitably design faults or *failure modes* (FMs) or software bugs intervene at each stage. Tests are conducted to locate and remove these. We shall characterize the initial number of FMs in each stage s as independently Poisson distributed with mean $m_s(0)$. On a given test, $t \geq 1$, an FM at stage s can only be accessed and activated and removed if all previous stages: $1,2,\ldots,s\text{-}1$ are accessed without activating any FM therein. We assume that any FM in accessed stage s is independently activated with probability p_s. When stage s is accessed, the conditional distribution of the number of activated FMs given the number of FMs remaining in that stage is Binomial with number of trials equal to the remaining number of FMs and probability of success p_s. Initial tests, conducted until the first stoppage at either S, the last stage, or more likely an intermediate stage, will be re-started after correction of the FMs found. The entire sequence is conducted for a fixed number of tests. This paper provides analytical insight concerning the total number of tests needed to demonstrate future total system fault-free operation with (approximate) quantitative certainty.

Note that questions of statistical inferences will be addressed later, using likelihood or Bayesian methods.

The distribution of the number of FMs remaining in each stage after a fixed number of tests is a mixture of Poisson distributions; c.f. Gaver et al. (2003). In this paper we present (Generalized) Negative Binomial approximations to the mixture distribution. In the next section we present equations to compute the mean and variance of the number of FMs remaining in each stage after t tests. Negative Binomial approximations to the distribution of the number of FMs remaining and the probability 0 FMs activate on the next test are presented in Section 3. Section 4 presents a numerical example. The paper ends with conclusions and extensions.

II. EQUATIONS FOR THE MEAN AND VARIANCE OF THE NUMBER OF FMS REMAINING IN EACH STAGE

Let $m_s(t)$ be the mean number of FMs in stage s after test t; $m_s(0)$ is the mean initial number of FMs in stage s

$$m_1(t) = m_1(0)\underbrace{(1-p_1)^t}_{\substack{\text{Probability FM in Stage 1 Survives } t \text{ Tests}}} \tag{1}$$

$$\begin{aligned} & m_2(t+1) \\ & = \underbrace{e^{-m_1(t)p_1}}_{\substack{\text{0 FMs activate in stage 1}}} m_2(t)(1-p_2) \\ & \quad + \left[1 - e^{-m_1(t)p_1}\right] m_2(t) \end{aligned} \tag{2}$$

$$\begin{aligned} & m_s(t+1) \\ & = \underbrace{e^{-\sum_{k=1}^{s-1} m_k(t)p_k}}_{\substack{\text{0 FMs activate in stages } 1,\ldots,s\text{-}1}} m_s(t)(1-p_s) \\ & \quad + \left[1 - e^{-\sum_{k=1}^{s-1} m_k(t)p_k}\right] m_s(t) \end{aligned} \tag{3}$$

The second moment of the number of FMs in stage *s* after test *t* is

$$c_s(t) = g_s(t) + m_s(t) \tag{4}$$

where $g_s(t)$ is the second factorial moment and satisfies the system of equations

$$g_1(t) = \left[m_1(0)(1-p_1)^t \right]^2 \tag{5}$$

$$g_s(0) = \left[m_s(0) \right]^2 \tag{6}$$

$$\begin{aligned} & g_2(t+1) \\ & = e^{-m_1(t)p_1} g_2(t)\left[(1-p_2)\right]^2 \\ & + \left[1 - e^{-m_1(t)p_1} \right] g_2(t) \end{aligned} \tag{7}$$

$$\begin{aligned} & g_s(t+1) \\ & = e^{-\sum_{k=1}^{s-1} m_k(t)p_k} g_s(t)\left[(1-p_s)\right]^2 \\ & + \left[1 - e^{-\sum_{k=1}^{s-1} m_k(t)p_k} \right] g_s(t) \end{aligned} \tag{8}$$

The variance of the number of FMs remaining in stage *s* after test *t* is

$$v_s(t) = g_s(t) + m_s(t) - \left[m_s(t) \right]^2. \tag{9}$$

The number of FMs remaining in stage *s* after *t* tests will have a mixed Poisson distribution.

III. Negative Binomial Approximations

Let $\boldsymbol{\Lambda}$ be a random variable having a Gamma distribution with

$$E\left[\boldsymbol{\Lambda}\right] = \frac{\beta}{\alpha}, Var\left[\boldsymbol{\Lambda}\right] = \frac{\beta}{\alpha^2}. \tag{10}$$

Let $\boldsymbol{N}$ be a random variable whose conditional distribution given $\boldsymbol{\Lambda}$ is Poisson with mean $\boldsymbol{\Lambda}$. Then

$$E\left[\boldsymbol{N}\right] = E\left[\boldsymbol{\Lambda}\right] = \frac{\beta}{\alpha}; \tag{11}$$

$$\begin{aligned} Var\left[\boldsymbol{N}\right] &= Var\left[E\left[\boldsymbol{N} \mid \boldsymbol{\Lambda}\right]\right] + E\left[Var\left[\boldsymbol{N} \mid \boldsymbol{\Lambda}\right]\right] \\ &= Var\left[\boldsymbol{\Lambda}\right] + E\left[\boldsymbol{\Lambda}\right] \\ &= \frac{\beta}{\alpha^2} + \frac{\beta}{\alpha} \end{aligned} \tag{12}$$

$\boldsymbol{N}$ has a (Generalized) Negative Binomial distribution; cf. Feller (1968).

A. An S-Stage Approximation

If $v_s(t) > m_s(t)$ then approximate the mixed distribution of the number of FMs remaining in stage *s* after *t* tests by a Negative Binomial distribution with scale and shape parameters chosen so that the mean (11) (respectively variance (12)) equals (3) (respectively (9)).

For $s = 2, \ldots, S$, let $\alpha_s(t)$ (respectively $\beta_s(t)$) be the Gamma scale parameter (respectively shape parameter) so that the resulting Negative Binomial mean and variance are equal to the calculated mean and variance for stage *s* after *t* tests. Let $\boldsymbol{\Lambda}_s(t)$ be a Gamma random variable having scale $\alpha_s(t)$ and shape $\beta_s(t)$. The probability that after *t* tests 0 FMs activate in stage *s* during additional tests $t+1, \ldots, t+k$ is approximated by

$$\begin{aligned} & \int_0^\infty e^{-yp_s\left[1+(1-p_s)+\ldots+(1-p_s)^{k-1}\right]} P\left\{\boldsymbol{\Lambda}_s(t) \in dy\right\} \\ & = \left[\frac{\alpha_s(t)}{\alpha_s(t) + \left[1-(1-p_s)^k\right]} \right]^{\beta_s(t)} \end{aligned} \tag{13}$$

A Negative Binomial approximation to the probability that after *t* tests 0 FMs activate in all stages during additional tests $t+1, \ldots, t+k$ is

$$\begin{aligned} q(t) = & e^{-m_1(0)(1-p_1)^t\left[1-(1-p_1)^k\right]} \\ & \times \prod_{s=2}^{S} \left[\frac{\alpha_s(t)}{\alpha_s(t) + \left[1-(1-p_s)^k\right]} \right]^{\beta_s(t)} \end{aligned} \tag{14}$$

recalling that the number of FMs remaining in stage 1 after *t* tests has a Poisson distribution with mean $m_1(0)(1-p_1)^t$; cf. Feller (1968).

B. A One-Stage Approximation

Let $M(t)=\sum_{s=1}^{S} m_s(t)$, the total mean number of FMs remaining in all stages after t tests. The one-stage approximation assumes that the numbers of FMs remaining in each stage are independent random variables; thus the approximate variance of the total number of FMs remaining in all stages is $V(t)=\sum_{s=1}^{S} v_s(t)$. If $V(t)>M(t)$, then approximate the distribution of the number of FMs remaining in all stages by a Negative Binomial distribution; otherwise approximate it by a Poisson distribution. Let $\overline{\alpha(t)}$ (respectively $\overline{\beta(t)}$) be the Gamma scale parameter (respectively shape parameter) so that the resulting Negative Binomial mean and variance are equal to the calculated mean and variance of the total number of FMs remaining after t tests. Approximate the probability a FM activates in the approximating one-stage system by

$$\overline{p(t)}=\frac{\sum_{s=1}^{S} p_s m_s(t)}{\sum_{s=1}^{S} m_s(t)}. \qquad (15)$$

The one-stage Negative Binomial approximation for the probability that after t tests 0 FMs are activated in all stages during additional tests $t+1,...,t+k$ is

$$\overline{q(t)}=\left[\frac{\overline{\alpha(t)}}{\overline{\alpha(t)}+\left[1-\left[1-\overline{p(t)}\right]^{k}\right]}\right]^{\overline{\beta(t)}}. \qquad (16)$$

The one-stage exponential approximation for the probability that after t tests 0 FMs are activated in all stages during additional tests $t+1,...,t+k$ is

$$\overline{q_E(t)}=\exp\left\{-M(t)\left[1-\left(1-\overline{p(t)}\right)^{k}\right]\right\}. \qquad (17)$$

IV. Example

The system consists of 4 stages. Before testing begins, the numbers of FMs in each stage are independent having a Poisson distribution with mean 6. The probability that a FM activates when a stage is accessed is 0.5 independently of other FMs for all stages. Tables I and II report the results of a simulation of the number of FMs remaining after t tests. Table I displays the results of 200 simulation replications and the 4-stage and 1-stage Negative Binomial approximations; displayed are the simulation's mean probability the next test will have 0 FMs activating in all stages and its standard error and the two Negative Binomial approximations for the probability: (14) and (16). Table II displays the results for the probability 0 FMs will activate during the next 3 tests.

For the probability 0 FMs activate during the next test, the two approximations give about the same values; the approximation is smaller than the simulation mean but seems adequate. For the probability 0 FMs activate during the next 3 tests, the 1-stage approximation tends to be smaller than the 4-stage approximation; the approximations are closer to the simulation values for larger numbers of tests. For the cases considered, the results suggest that at least 15 tests should be conducted.

TABLE I. Probability 0 FMs Activate in All Stages During Next Test

Number of Tests	Simulation	Approximations
	Mean (Standard Deviation)	***4-Stage [1-Stage]***
10	0.35 (0.02)	0.25 [0.25]
12	0.66 (0.02)	0.57 [0.57]
14	0.87 (0.01)	0.86 [0.86]
15	0.92 (0.008)	0.92 [0.92]
16	0.96 (0.005)	0.96 [0.96]
18	0.99 (0.003)	0.99 [0.99]

TABLE II. Probability 0 FMs Activate in All Stages During Next Three Tests

Number of Tests	Simulation	Approximations
	Mean (Standard Deviation)	***4-Stage [1-Stage]***
10	0.21 (0.01)	0.11 [0.11]
12	0.52 (0.02)	0.44 [0.37]
14	0.80 (0.01)	0.78 [0.76]
15	0.88 (0.01)	0.88 [0.87]
16	0.93 (0.008)	0.93 [0.93]
18	0.98 (0.001)	0.98 [0.98]

V. Conclusions and Extensions

We have presented Negative Binomial approximations for the number of failure modes (FM) remaining in a series system with FM masking. The approximations make use of known mean number of FMs in each stage before testing and known probabilities the FMs in each stage activate when the stage is accessed. Maximum likelihood can be used to estimate these parameters from data obtained. The approximations may be adequate to recursively estimate the probability of 0 FMs activating in all stages on the next test.

References

[1] D. P. Gaver, P. A. Jacobs, K. D. Glazebrook, and E. A. Seglie, "Probability models for sequential-stage system reliability growth via failure mode removal," International Journal of Reliability, Quality and Safety Engineering, vol. 10, pp. 15-40, 2003.

[2] W. Feller, An Introduction to Probability Theory and its Applications, 3rd ed., vol. 1. New York: John Wiley & Sons, Inc., 1968.

Survivability and Life Expectancy Modeling for Items Subjected to Complex Life Profile

Mordechai Jaeger and Ze'ev Porat,
Rafael, P. Box 2250, Haifa, Israel

***Abstract*—this article presents probabilistic models addressing the problem of assessing the survivability and life expectancy of items subjected to repetitive multi phases usage cycles. This life profile characterizes the operational pattern of many daily used technologies, such as air and ground vehicles, home and services appliances, medical instruments, industrial machineries, etc. The topic was first studied by us in reference [2], where simple approximated solutions were given for the Weibull life distribution. Here we show that the model can be generalized to any family of life distributions having zero location parameter, where their scale parameter varies with alternating environmental conditions.**
An exception is the exponential distribution for which we show that exact calculations can be performed with low computational effort.

Keywords: Survivability, Life Expectency, MTTF, Reliability, Weibull, Exponential, Gamma, Lognormal, CEM (Cumulative Exposure Model).

I. INTRODUCTION

Let us assume that an item's the life profile comprises cycles of usage as described in Figure 1.

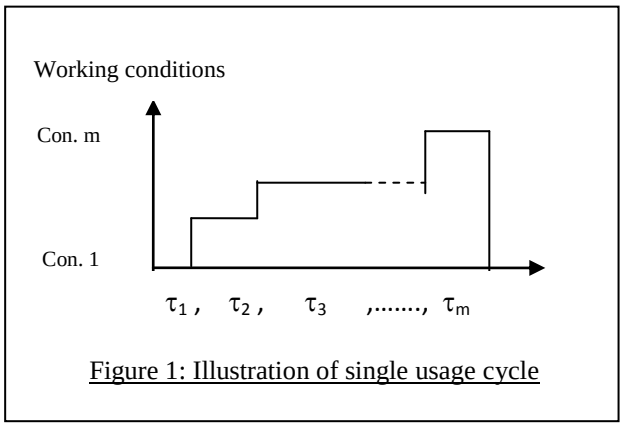


Figure 1: Illustration of single usage cycle

We introduce models to assess the item survival probability over specified time duration (Survivability) and the item expected life time – MTTF (Mean Time To Failure). The model outline has been formulated in [2] for the Weibull life distribution case. The models in [2] gave approximated assessments to the above characteristics. The degree of accuracy in the MTTF was also addressed and quantified there.

Here we show that the models can be generalized to any family of life distributions having zero location parameter, where their scale parameter varies with alternating environmental conditions.

For the exponential case we present exact derivations of the survivability function and the MTTF.

II. MODEL DESCRIPTION

A. *Notation*

Con. i - working and environmental conditions in usage phase i, i=1, m

m - number of phases in a usage cycle

τ_i - time in phase i, i=1, m

t_i - $\sum_{j=1}^{i} \tau_j$, time spent in the usage profile until the end of phase i

T - $\sum_{i=1}^{m} \tau_i$, lengh of a usage cycle, i=1, m

$F_X(t)$ - the Cumulative Distribution Function (CDF)

$R_X(t)$ - =1- $F_X(t)$ the reliability function associated with that distribution

$S(t)$ - the survivability function – probability of surviving till time t.

B. *Model Assumptions*

We assume that the life distributions in each phase are of the same type, with location parameter equals zero, while the scale parameter varies at each phase.

Thus, let X_i be the random variable that represents the failure time distribution at each phase and let a_i be the scale parameter associated with X_i. In this case a "standard" variable Y exists so that

$$R_{X_i}(t) = R_Y(\frac{t}{a_i}) \quad (1)$$

The total damage to the component, accumulated over the time is calculated according to the Cumulative Exposure Model (CEM) described in [1], [3] and [4]. The model is allowing computation of equivalent working time at the start of phase # i in terms of X_i as follows: Let $S(t_{i-1})$ be the value of the survivability function at the end of the i-1 phase, the equivalent working time at this point in terms of X_i is the value $t_{i/eq}$ for which

$$R_{X_i}(t_{i/eq}) = S(t_{i-1}) \quad (2)$$

and the survivability function at any time point t in the i'th phase is equal

$$S(t) = R_{X_i}[t_{i/eq} + (t - t_{i-1})] \quad (3)$$

C. *Results for the Weibull Case[2]*

The Weibull reliability formula is

$$R_X(t) = e^{-(t/\alpha)^\beta} \quad (4)$$

In this case we assume that the scale parameter α varies at each phase, while the shape parameter β is invariant.
In [2] we presented the following results for the Weibull case:

$$\textbf{Assign } \alpha_{eq} = \frac{\mathrm{T}}{\sum_{i=1}^{m} \frac{\tau_i}{\alpha_i}} \quad (5)$$

The survivability function after n cycles of length T is:

$$S(nT) = \exp[-(\frac{nT}{\alpha_{eq}})^\beta] \quad (6)$$

The system MTTF can be approximated as

$$MTTF \approx \alpha_{eq}\Gamma(1 + 1/\beta) \quad (7)$$

where the approximation error of the MTTF never exceeds the value of the cycle length T.

III . Results for The general Model

Let X_i be the random variable that represents the failure time distribution at each phase. As mentioned before, given the scale parameters a_i, a "standard" variable Y exists so that

$$R_{X_i}(t) = R_Y(\frac{t}{a_i}) \quad (8)$$

The Weibull model results are generalized as follows:

$$\text{Assign: } a_{eq} = \frac{\mathrm{T}}{\sum_{i=1}^{m} \frac{\tau_i}{a_i}} \text{ , } \quad X_{eq} = a_{eq} Y \quad (9)$$

Proposition 1: The survivability function after n cycles is equal

$$S(nT) = R_Y(\frac{nT}{a_{eq}}) = R_{X_{eq}}(nT) \quad (10)$$

Proposition 2: The system MTTF can be approximated as,

$$MTTF \approx E(X_{eq}) = a_{eq} E(Y) \text{ , } \quad (11)$$

where the approximation error never exceeds the value of the cycle length T.

Proof of the propositions is given in Appendix A.

Examples

Gamma Distribution

Let $Xi \sim Gamma(\lambda_i, \mathrm{r})$ *defining* $a_i = \frac{1}{\lambda i}$, we get

$$\lambda_{eq} = \frac{1}{T}\sum_{i=1}^{m} \lambda_i \tau_i, \ MTTF \cong \frac{\mathrm{r}}{\lambda_{eq}} \quad (12)$$

Lognormal Distribution

Let: $Xi \sim Lognotmal(\mu_i, \sigma)$ and $a_i = e^{\mu_i}$ yeilding:

$$\mu_{eq} = \ln\{\frac{T}{\sum_{i=1}^{m}[\tau_i / e^{\mu_i}]}\} \quad (13)$$

$$MTTF \cong e^{(\mu_{eq} + 0.5\sigma^2)} \quad (14)$$

IV. The Exponential Case

In this case $X_i \sim exp(\lambda_i)$. It should be noticed that since the exponential distribution has the memory less property, the assumption of CEM is not necessary for the development of the MTTF model.

For any time t, the survivability function (S) at time (t+T) is equal:

$$S(t+T) = S(t) \cdot e^{-\sum_{i=1}^{m} \lambda_i \tau_i} \quad (15)$$

Assign: $Q = e^{-\sum_{i=1}^{m} \lambda_i \tau_i}$, based on the above, the survivability at any time point can be easily calculated as

$$S(t) = S(t')Q^k \quad (16)$$

where:

- k = INT(t/T), is the number of cycles completed before t
- t' = t – kT, is the "compatible" time point for t in the first cycle.

The **exact** formula for the MTTF is

$$MTTF = \int_0^\infty S(t)dt = \sum_{j=1}^{\infty} S(t_{j-1}) \cdot \int_{t_{j-1}}^{t_j} e^{-\lambda_j (t - t_{j-1})} dt \quad (17)$$

$$= \sum_{k=0}^{\infty} Q^k \left(\int_0^{t_1} e^{-\lambda_1 t} dt + \ldots S(t_{m-1}) \cdot \int_{t_{m-1}}^{t_m} e^{-\lambda_m (t - t_{m-1})} dt \right) \quad (18)$$

$$= \frac{1}{1-Q} \sum_{j=1}^{m} S(t_{j-1}) \frac{1 - e^{-\lambda_j (t_j - t_{j-1})}}{\lambda_j} \quad (19)$$

Example: For a 3 phases cycle:

$S(t_1) = e^{-\lambda_1 t_1}$

$S(t_2) = S(t_1) e^{-\lambda_2 (t_2 - t_1)}$

$S(T) = S(t_2) e^{-\lambda_3 (T - t_2)}$, $Q = S(T)$

- The survivability function S(t) is calculated as follows:

For $t \le t_1$ $S(t) = e^{-\lambda_1 t}$,

for $t_1 \le t \le t_2$ $S(t) = S(t_1) e^{-\lambda_2 (t - t_1)}$,

for $t_2 \le t \le T$ $S(t) = S(t_2) e^{-\lambda_3 (t - t_2)}$,

and for any $t > T$ $S(t) = S(t') Q^k$, where

$k = \mathrm{INT}(t/T)$ and $t' = t - kT$.

- The exact value of the MTTF is:

$$\frac{1}{1-Q} \left(\frac{1 - e^{-\lambda_1 t_1}}{\lambda_1} + S(t_1) \frac{1 - e^{-\lambda_2 (t_2 - t_1)}}{\lambda_2} + S(t_2) \frac{1 - e^{-\lambda_3 (T - t_2)}}{\lambda_2} \right) \quad (20)$$

V. SUMMARY

In this paper we presented a simple fairly accurate method to evaluate the survivability function and the MTTF of an item under a complex life profile. It was shown that the method is suitable for any lifetime distribution with a zero location parameter. Exact derivations have been presented for the exponential distribution case.

VI. APPENDIX A: PROOF OF PROPOSITIONS 1 AND 2

Proof of proposition 1: Let $S(t_{i-1})$ be the survivability function at the end of the i-1'st phase. According to the Cumulative Exposure Model the equivalent working time at this point in terms of Y is the value t_{eq} for which $R_Y(t_{eq}) = S(t_{i-1})$

The value of the survivability function at the end of the i'th phase is equal

$$S(t_i) = R_{X_i}(a_i t_{eq} + \tau_i) = R_Y(t_{eq} + \frac{\tau_i}{a_i}) \quad (A.1)$$

Thus by induction, for any k

$$S(t_k) = R_Y(\sum_{i=1}^{k} \frac{\tau_i}{a_i}) \quad (A.2)$$

Assign $a_{eq} = \frac{T}{\sum_{i=1}^{m} \frac{\tau_i}{\alpha_i}}$. Based on the above, the survivability function after n cycles of length T is equal

$$S(nT) = R_Y(\frac{nT}{a_{eq}}) = R_{X_{eq}}(nT) \quad (A.3)$$

Proof of proposition 2: Let $R_{X_{eq}}(t)$ be the reliability function of X_{eq}, $S(t)$ the true system survivability function.

The true component's MTTF equals $\int_0^{\infty} S(t) dt$ while the approximated value is $\int_0^{\infty} R_{X_{eq}}(t) dt = E(X_{eq})$

For convenience, assume that the cycle length T is equal one time unit. Both $R_{X_{eq}}(t)$ and $S(t)$ are monotone decreasing functions, and due to our methodology they share equal values at the start of each time unit (t=0, 1, 2..). Therefore they exhibit the following property:

$$\sum_{n=1}^{\infty} R(n) < \int_0^{\infty} R(t) dt < \sum_{n=0}^{\infty} R(n) \quad (A.4)$$

Namely, both MTTF values are bounded between the same two above sums.
Thus, the difference between them is less than

$$\sum_{n=0}^{\infty} R(n) - \sum_{n=1}^{\infty} R(n) = R(0) = 1 \text{ unit of t.} \quad (A.5)$$

Meaning that the difference is less than T (i.e. one cycle length), so the approximation becomes more accurate as the MTTF is longer in terms of T.

VII. REFERENCES

[1] N. Balakrishnan, Q. Xie, D. Kunduy,"Exact Inference For A Simple Step-Stress Model From The Exponential Distribution Under Time Constraint", Annals of the Inst. of Statistical Statistics Mathematics (2009), Vol. 61, pp: 251-274.

[2] A. Filis, N. Pundak, M. Barak, Z. Porat, M. Jaeger."Ricor's New Development of Highly Reliable Integral Rotary Cooler, Engineering and Reliability Aspects", Proceedings of SPIA, Vol. 8012, May 2011.

[3] R. Miller, W. B. Nelson. "Optimum Step-Stress Plan For Accelerated Life testing", IEEE Trans. Reliability, vol. R-32, April 1983, pp 59-65.

[4] W. B. Nelson. "Accelerated Life testing Step-Stress models and Data analysis ", IEEE Trans. Reliability, vol. R-29, June 1980, pp 103-108.

Concordance Estimation and its application for censored data

ZhehZen Jin

Email: zj7@columbia.edu

Abstract: The concordance can be used to assess discriminative and predictive accuracy in statistical models. Recently, there has been significant development .In this talk, we will present methods for the estimation of concordance probability based on nonparametric or semiparametric modeling approach for right censored data. The asymptotic properties of the proposed estimators will be presented and the methods will be illustrated with real examples.

Characterizations and Inference for Survival and Reliability Models with Generalized Divergence Measures

Alex Karagrigoriou
Department of Mathematics and Statistics
University of Cyprus
CY-1678 Nicosia, Cyprus
Email: alex@ucy.ac.cy

Ilia Vonta
Department of Mathematics
National Technical University of Athens
Athens, Greece
Email: vonta@math.ntua.gr

***Abstract*—Measures of divergence or discrepancy are used either to measure mutual information concerning two variables or to construct model selection criteria. In this paper we are focusing on divergence measures that are based on a class of measures known as Csiszar's divergence measures. In particular, we propose a measure of divergence between residual lives of two items that have both survived up to some time t as well as a measure of divergence between past lives, both based on Csiszar's class of measures. Furthermore, we derive properties of these measures and provide examples based on the Cox model and frailty or transformation models.**

I. Introduction

A measure of divergence is used as a way to evaluate the distance (divergence) between any two populations or functions. In the present work, we concentrate on divergence measures that are based on a class of measures known as Csiszar's family of divergence measures or Csiszar's φ-divergence (Csiszar, (1963); Ali and Silvey, (1966)).

An issue of fundamental importance in Statistics is the investigation of Information Measures. These measures are classified in different categories and measure the quantity of information contained in the data with respect to a parameter θ, the divergence between two populations or functions, the information we get after the execution of an experiment and other important information according to the application they are used for. Traditionally, the measures of information are classified in four main categories namely divergence - type, entropy - type, Fisher - type and Bayesian - type.

Measures of divergence between two probability distributions have a very long history initiated by the pioneer work of Pearson, Mahalanobis, Lévy and Kolmogorov. Among the most popular measures of divergence are the Kullback-Leibler measure of divergence and the Csiszar's φ-divergence family of measures. Recently, the BHHJ divergence measure was proposed by Basu et *al.* (1998) and generalized to the BHHJ family of measures by Mattheou et *al.* (2009).

Ebrahimi and Kirmani (1996a) introduced a measure of discrepancy between the lifetimes X and Y of two items at time t. In survival analysis or in reliability we might know the current age t of a biomedical or technical system. We need to take this information into consideration when we compare two systems or populations. Ebrahimi and Kirmani (1996a) achieved this by replacing the distribution functions of the random variables X and Y in the Kullback-Leibler divergence of X and Y, by the distributions of their residual lifetimes. Di Crescenzo and Longobardi (2004) define a dual measure of divergence which constitutes a distance between past life distributions.

II. Generalized Divergence Measures

Let X and Y be absolutely continuous, non-negative random variables that describe the lifetimes of two items. Let $f(x)$, $F(x)$ and $\overline{F}(x)$ be the density function, the cumulative distribution function and the survival function of X respectively. Let also $g(x)$, $G(x)$ and $\overline{G}(x)$ be the density function, the cumulative distribution function and the survival function of Y respectively. Let

$$h_X(x) = f(x)/\overline{F}(x) \quad \text{and} \quad h_Y(x) = g(x)/\overline{G}(x)$$

be the hazard rate functions of X and Y while

$$\tau_X(x) = f(x)/F(x) \quad \text{and} \quad \tau_Y(x) = g(x)/G(x)$$

are the reversed hazard rate functions of X and Y. Without loss of generality we assume throughout the paper that the support of f and g is $(0, +\infty)$.

The Kullback-Leibler distance between F and G is defined by

$$I_{X,Y} = \int_0^\infty f(x) \log\left(\frac{f(x)}{g(x)}\right) dx \tag{1}$$

where $\log$ denotes the natural logarithm. A generalization of this distance is defined as

$$I^{\varphi}_{X,Y} = \int_0^\infty g(x)\varphi\left(\frac{f(x)}{g(x)}\right) dx \tag{2}$$

and is known as Csiszar's family of measures of divergence.

When the function φ is defined as

$$\varphi(u) = u\log u \quad \text{or} \quad \varphi(u) = u\log u + 1 - u$$

then the above measure reduces to the Kullback-Leibler measure. If $\varphi(u) = (1-u)^2$, Csiszar's measure yields the Pearson's chi-square divergence. If

$$\varphi(u) = \big(u^{a+1} - u - a(u-1)\big)/(a(a+1))$$

we obtain the Cressie and Read power divergence (Cressie and Read, (1984)), $a \neq 0, -1$. If $\varphi(u) = (1-\sqrt{u})^2$, we obtain the Matusita's divergence (Matusita, (1967)).

We define also the function

$$\varphi(u) = 1 - (1+\frac{1}{a})u + \frac{u^{1+a}}{a}, \quad a > 0$$

which is related to a recently proposed measure of divergence (BHHJ power divergence, Basu et *al.* (1998)). Another function that we consider is

$$\varphi(u) = u^{1+a} - (1+\frac{1}{a})u^a + \frac{1}{a}, \quad a > 0.$$

These last two functions are special cases of the BHHJ family of measures of divergence proposed by Mattheou et *al.* (2009)

$$\begin{aligned} I_X^a(g,f) = & \quad E_g\left(g^a(X)\varphi\left(\tfrac{f(X)}{g(X)}\right)\right) \\ = & \quad \int g^{1+a}(z)\,\varphi\left(\tfrac{f(z)}{g(z)}\right)d\mu, \quad a > 0, \end{aligned} \tag{3}$$

where μ represents the Lebesgue measure. Appropriately chosen functions $\varphi(\cdot)$ give rise to special measures mentioned above.

Ebrahimi and Kirmani (1996a) introduced a measure of discrepancy between X and Y at time t as follows

$$I_{X,Y}(t) = \int_t^\infty \frac{f(x)}{\overline{F}(t)} \log\left(\frac{f(x)/\overline{F}(t)}{g(x)/\overline{G}(t)}\right) dx, \qquad t > 0. \tag{4}$$

A dual measure is defined in Di Crescenzo and Longobardi (2004) which constitutes a distance between past lifetimes

$$\overline{I}_{X,Y}(t) = \int_0^t \frac{f(x)}{F(t)} \log\left(\frac{f(x)/F(t)}{g(x)/G(t)}\right) dx, \qquad t > 0. \tag{5}$$

In this paper we propose two measures of discrepancy which are based on the Csiszar's φ-divergence family, namely, the *φ-distance between residual lifetimes*

$$I_{X,Y}^\varphi(t) = \int_t^\infty \frac{g(x)}{\overline{G}(t)}\,\varphi\left(\frac{f(x)/\overline{F}(t)}{g(x)/\overline{G}(t)}\right) dx, \qquad t > 0 \tag{6}$$

and the *φ-distance between past lifetimes*

$$\overline{I}_{X,Y}^\varphi(t) = \int_0^t \frac{g(x)}{G(t)}\varphi\left(\frac{f(x)/F(t)}{g(x)/G(t)}\right) dx, \qquad t > 0 \tag{7}$$

where the function φ belongs to a class of convex functions Φ that satisfy some regularity conditions.

III. Measures in Survival and Reliability Models

A. Proportional hazards and Proportional reverse hazards models

In this section we examine properties of the proposed measures of divergence and find various discrimination measures in cases like the proportional hazards model (PH), the proportional reverse hazards model (PRH) and the frailty or transformation models. For the latter case, we provide the φ-distance between the respective residual and past lifetimes associated with the Cox and frailty models respectively.

For the case of proportional hazards let X and Y be random variables with distribution functions F and G respectively for which it holds that

$$\overline{G}(x) = (\overline{F}(x))^\theta \text{ for all } x > 0 \text{ and } \theta > 0. \tag{8}$$

Theorem. The discrimination measure $I_{X,Y}^\varphi(t)$ between the random variables X and Y which satisfy the proportional hazards assumption (8) is independent of t and is given as

$$I_{X,Y}^\varphi(t) = \int_0^1 \varphi\left(\frac{1}{\theta y^{\theta-1}}\right) dy^\theta. \tag{9}$$

(ii) If $I_{X,Y}^\varphi(t)$ is independent of t, then there exists a constant $\theta > 0$ such that (8) holds.

Let now X and Y be random variables with distribution functions F and G respectively which satisfy the proportional hazards assumption but with reverse proportionality, that is,

$$\overline{F}(x) = (\overline{G}(x))^\theta.$$

In this case, the discrimination measure takes the form

$$I_{X,Y}^\varphi(t) = \int_0^1 \varphi\left(\theta y^{\theta-1}\right) dy. \tag{10}$$

For the proportional reverse hazards model which is defined as

$$G(x) = (F(x))^\theta \ \textit{for all}\ x > 0 \ \textit{and}\ \theta > 0 \tag{11}$$

the following result holds.

Theorem. (i) The discrimination measure $\overline{I}_{X,Y}^\varphi(t)$ between the random variables X and Y which satisfy the proportional reverse hazards assumption (11) is independent of t and is given as

$$\overline{I}_{X,Y}^\varphi(t) = \int_0^1 \varphi\left(\frac{1}{\theta y^{\theta-1}}\right) dy^\theta. \tag{12}$$

(ii) If $\overline{I}_{X,Y}^\varphi(t)$ is independent of t, then there exists a constant $\theta > 0$ such that (11) holds.

Let now X and Y be random variables with distribution functions F and G respectively which satisfy the proportional reverse hazards assumption but with reverse proportionality. In that case, the discrimination measure $\overline{I}_{X,Y}^\varphi(t)$ is given by (10).

B. Frailty/transformation vs. Cox regression

Let now X and Y be random variables with distribution functions F_1 and F_2, probability density functions f_1 and f_2 and survival functions S_1 and S_2 respectively. Let H be the baseline cumulative hazard function and h the baseline intensity hazard function. Let X follow a Cox model (Cox 1972) under which

$$S_1(x) = e^{-\theta H(x)}, \quad \theta > 0. \tag{13}$$

Let also Y follow a frailty model under which (Vonta, 1996)

$$S_2(x) = e^{-G(\theta H(x))}, \quad \theta > 0 \tag{14}$$

where the function G is assumed to be concave, increasing with $G(0) = 0$ and $G(\infty) = \infty$.

We have derived the usual divergence between the distributions of X and Y using the Kullback-Leibler and the Csiszar's divergence as well as the φ-distance between the respective past and residual lifetimes of X and Y. We provide here for the latter cases the relative results which are stated below.

Theorem. The discrimination measure $I^{\varphi}_{X,Y}(t)$ between the random variables X and Y which follow the Cox proportional hazards model and the frailty or transformation model (14) respectively, is given as

$$I^{\varphi}_{X,Y}(t) = \int_{\theta H(t)}^{\infty} \frac{e^{-G(y)}G'(y)}{e^{-G(\theta H(t))}} \varphi\left(\frac{e^{-y}/e^{-\theta H(t)}}{e^{-G(y)}G'(y)/e^{-G(\theta H(t))}}\right) dy \tag{15}$$

for $t > 0$.

Theorem. The discrimination measure $\overline{I}^{\varphi}_{X,Y}(t)$ between the random variables X and Y which follow the Cox proportional hazards model and the frailty or transformation model (14) respectively, is given as

$$\overline{I}^{\varphi}_{X,Y}(t) = \int_{0}^{\theta H(t)} \frac{e^{-G(y)}G'(y)}{1-e^{-G(\theta H(t))}} \times \times \varphi\left(\frac{e^{-y}/1-e^{-\theta H(t)}}{e^{-G(y)}G'(y)/1-e^{-G(\theta H(t))}}\right) dy \tag{16}$$

for $t > 0$.

IV. Applications

It should be pointed out that measures of divergence play a significant role in statistical inference and have several applications. Measures of divergence can be used in statistical inference for estimating purposes (Toma, 2008; 2009), in the construction of test statistics for tests of fit (e.g. Zografos et. al, 1990, Huber-Carol et. al, 2002 and Zhang, 2002) or in statistical modeling for the construction of model selection criteria like the Kullback-Leibler measure which has been used for the development of various criteria (e.g. Akaike, 1973 and Cavanaugh, 2004). Special cases of the generalized family of divergence measures have been recently used for the construction of the Divergence Information Criterion (DIC, Mattheou et. al, 2009, Mantalos et. al, 2009)) and the development of a new test statistic for goodness of fit (gof) tests for multinomial distributions (Mattheou and Karagrigoriou, 2010).

The problem of goodness-of-fit to a distribution, $H_0 : F = F_0$, is frequently treated by partitioning the data-range in disjoint intervals and testing the hypothesis

$$H_0 : \mathbf{p} = \mathbf{p}_0$$

about the parameter-vector of the resulting multinomial distribution.

Let $P = \{E_i\}_{i=1,\ldots,m}$ be a partition of the real line R into m intervals. Let

$$\mathbf{p} = (p_1, \ldots, p_m)'$$

and

$$\mathbf{p}_0 = (p_{10}, \ldots, p_{m0})'$$

be the true and the hypothesized probabilities of the intervals $E_i,\ i = 1, \ldots, m$, respectively, in such a way that

$$p_i = P_F(E_i),\ i = 1, \ldots, m$$

and

$$p_{i0} = P_{F_0}(E_i) = \int_{E_i} dF_0,\ i = 1, \ldots, m.$$

Let $Y_1, \ldots, Y_N$ be a random sample from F, let $n_i = \sum_{j=1}^{N} I_{E_i}(Y_j)$, where and 0 otherwise, $\widehat{\mathbf{p}} = (\widehat{p}_1, \widehat{p}_2, ..., \widehat{p}_m)'$ with $\widehat{p}_i = n_i/N$, $i = 1, \ldots, m$ be the maximum likelihood estimator (MLE) of p_i, the true probability of the E_i interval, and $\sum_{i=1}^{m} n_i = N$. For testing the simple null hypothesis that $(n_1, n_2, ..., n_m)$ come from a known multinomial distribution $F_0 = M(N, P_0)$, where $P_0 = (p_{10}, p_{20}, ..., p_{m0})'$, namely $H_0 : \mathbf{p} = \mathbf{p}_0$ or equivalently

$$H_0 : F = M(N, P_0),\ P_0 = (p_{10}, p_{20}, ..., p_{m0})'$$

the most commonly used test statistics are Pearson's or chi-squared test statistic, given by

$$X^2 = \sum_{i=1}^{m} \frac{(n_i - Np_{i0})^2}{Np_{i0}}$$

and the likelihood ratio test statistic given by

$$G^2 = 2\sum_{i=1}^{m} n_i \log\left(\frac{n_i}{Np_{i0}}\right).$$

Both of these test statistics are special cases of the family of power-divergence test statistics (Cressie and Read, 1984) which is based on the $I^{\varphi}_{X,Y}$ measure of divergence with $\varphi(\cdot)$ given by

$$\varphi(u) = \big(u^{a+1} - u - a(u-1)\big)/(a(a+1)).$$

In this work we explore the applicability of the measures of divergence by focusing on goodness of fit tests for composite hypotheses when the underlying distribution depends on a $d-$dimensional parameter $\theta \in R^d$. More specifically, we propose to test the hypothesis

$$H_0 : F = F = M(N, P_{0,\theta}),\ P_0 = (p_{10,\theta}, p_{20,\theta}, ..., p_{m0,\theta})$$

using the test statistic (Vonta and Chouchoumis 2010, Vonta and Tsanousa, 2010)

$$T_{n,t}^{\varphi}(\hat{\theta}_{\varphi}) = T_{n,t}^{\varphi}(\hat{p}_t, p_t(\hat{\theta}_{\varphi})) = \frac{2n}{\varphi''(1)} \sum_{i=1}^{M} p_{i,t}(\hat{\theta}_{\varphi}) \varphi\left(\frac{\hat{p}_{i,t}}{p_{i,t}(\hat{\theta}_{\varphi})}\right),$$

where the estimation of the parameter θ is being done from

$$\hat{\theta}_{\varphi} = \arg \min_{\theta \in \Theta \subseteq R^d} \left(\sum_{i=1}^{M} p_{i,t}(\theta) \varphi\left(\frac{\hat{p}_{i,t}}{p_{i,t}(\theta)}\right)\right)$$

and $t > 0$ is the time of interest.

REFERENCES

[1] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. *Proc. of the* 2^{nd} *International Symposium on Information Theory*, Petrov B. N. and Csaki F., eds., Akademia Kaido, Budapest, 267-281.

[2] Ali, S. M. and Silvey, S. D. (1966). A general class of coefficients of divergence of one distribution from another, *J. Roy. Statist. Soc. B*, *28*, 131-142.

[3] Basu, A., Harris, I. R., Hjort, N. L. and Jones, M. C. (1998). Robust and efficient estimation by minimising a density power divergence, *Biometrika*, *85*, 549-559.

[4] Cavanaugh, J. E. (2004). Criteria for linear model selection based on Kullback's symmetric divergence. *Australian and New Zealand Journal of Statistics* 46, 257-274.

[5] Cox, D. R. (1972). Regression models and life tables, *J. Roy. Statist. Soc.*, *B 34*, 187-202.

[6] Cressie, N. and Read, T. R. C. (1984). Multinomial goodness-of-fit tests, *J. R. Statist. Soc*, *5*, 440-454.

[7] Csiszar, I. (1963). Eine Informationstheoretische Ungleichung und ihre Anwendung auf den Bewis der Ergodizitat on Markhoffschen Ketten, *Publ. of the Math. Inst. of the Hungarian Academy of Sc.*, *8*, 84–108.

[8] Di Crescenzo, A. and Longobardi, M. (2002) Entropy-based measure of uncertainty in past lifetime distributions, *J. Appl. Prob.*, *39*, 434-440.

[9] Di Crescenzo, A. and Longobardi, M. (2004) A measure of discrimination between past lifetime distributions, *Statist. Probab. Lett.*, *67*, 173-182.

[10] Ebrahimi, N. and Kirmani, S.N.U.A. (1996a) A measure of discrimination between two residual life-time distributions and its applications, *Ann, Inst. Statist. Math.*, *48*, 257-265.

[11] Ebrahimi, N. and Kirmani, S.N.U.A. (1996b) A characterization of the proportional hazards model through a measure of discrimination between two residual life distributions, *Biometrika*, *83*, 233-235.

[12] Huber-Carol, C., Balakrishnan, N., Nikulin, M. S., and Mesbah, M. (2002). *Goodness-of-fit Tests and Model Validity*, Birkhäuser, Boston.

[13] Mantalos, P., Mattheou, K., Karagrigoriou, A. (2010). Forecasting ARMA Models: A comparative study of information criteria focusing on MDIC, *J. Statist. Comput. and Simul*, 80 (1), 61-73.

[14] Mattheou, K. and Karagrigoriou, A. (2010). A new family of divergence measures for tests of fit. *Australian and New Zealand Journal of Statistics*, 52 (2), 187-200.

[15] Mattheou, K., Lee, S., and Karagrigoriou, A. (2009). A model selection criterion based on the BHHJ measure of divergence, *J. of Statist. Plan. and Infer.*, *139*, 128-135.

[16] Matusita, K. (1967). On the notion of affinity of several distributions and some of its applications, *Ann. Inst. Statist. Math.*, *19*, 181-192.

[17] Toma, A. (2009). Optimal robust M-estimators using divergences, *Statistics and Prob. Letters*, 79, 1-5.

[18] Toma, A. (2008). Minimum Hellinger distance estimators for multivariate distributions from the Johnson system, *J. Statist. Plan. and Infer.*, 138, 803-816.

[19] Vonta, F. (1996) Efficient estimation in a nonproportional hazards model in survival analysis, *Scand. Jour. Statist.*, *23*, 49-62.

[20] Vonta, F. and Chouchoumis I. (2010) *φ-measures of divergence and their applications to survivalk analysis and reliability*, Master thesis, National Technical University of Athens.

[21] Vonta F. and Tsanousa A. (2010). *Hypothesis testing via φ-measures of divergence*, Master thesis, National Technical University of Athens.

[22] Zhang, J. (2002). Powerful goddness-of-fit tests based on likelihood ratio. *J. R. Stat. Soc. Ser. B* 64 (2), 281-294.

[23] Zografos, K., Ferentinos, K. and Papaioannou, T. (1990). Φ-divergence statistics: Sampling properties, multinomial goodness of fit and divergence tests, *Comm. in Statist. Theor. Meth.*, 19 (5), 1785–1802.

Powering stochastic reliability models by discrete event simulation

Igor Kozine
Department of Management Engineering
Technical University of Denmark
Kgs. Lyngby, Denmark
igko@dtu.dk

Xiaoyun Wang
Department of Mathematics and Physics
Tsinghua University
Beijing, P.R.China

***Abstract*-Markov reliability models are widely practiced tools for the analysis of repairable systems. Nevertheless, the assumptions of the Markov model may appear too restrictive to adequately model a real system and the explosion in the number of states as the size of the system increases may make it difficult to find a solution to the problem. The power of modern computers and recent developments in discrete-event simulation (DES) software enable to diminish some of the drawbacks of stochastic models. In this paper we describe the insights we have gained based on using both Markov and DES models for simple systems. By contrasting the results of the two models we illuminate their advantages and disadvantages as well as we conclude that it is a good way of model validation.**

Keywords-Markov reliability model, discrete event simulation

I. INTRODUCTION

Assessing the reliability of systems which are subject to certain inspection-repair-replacement policies is a complex task. At any instant in time the system can be in one of many possible states. The number of distinguished states depends on the number and function of the system equipment. To be able to model such a system and predict its reliability, the deterioration law of the system is usually assumed to be Markovian; that is, the future course of the system depends only on its state at present time and not on its past history. Given this assumption and that each component has approximately an exponential failure law and the reliability of it is fully restored after repair, the complete system can be described approximately by a Markov process. The models of this type are widely practiced by reliability analysts and, generally, they are regarded powerful tools in reliability, maintainability and safety engineering and commonly used to study the dependability of complex systems. The advantage of the Markov process is that it describes both the failure of an item and its subsequent repair and/or preventive maintenance and periodic testing. The Markov process can easily describe degraded states of operation, where the item has either partially failed or is in a degraded state where some functions are performed while other are not. (Markov models are extensively described in the literature, see, for example, [1])

All in all, a Markov model is a well-established and widely used method to solve stochastic event problems and is perhaps the most practiced analytical model of repairable systems and its appeal is in being able to derive reliability measures through analytical calculations. Nevertheless, as it often is, analytical solutions for complex systems are often based on assumptions the influence of which on the results may be underestimated and not well understood. For example, the Markovian property, which is the memoryless property of a stochastic process, cannot be regarded adequate in many reliability applications and should be employed consciously. Or one more point to think over: for the sake of mathematical convenience one often has to accept the governance of time between failures and repair times by exponential distributions, while available failure data may not strongly support such choice. It is clear that in this case the computed values may deviate significantly from the true probabilistic measures.

The usually stressed major drawback of the Markov method is the explosion in the number of states as the size of the system increases. The resulting diagrams for relatively complex systems are generally extremely large and complicated, difficult to construct and computationally extensive [2].

However, the rapid increase in computer power and the associated development of easy-to-use modelling tools promote the use of computer modelling and simulation as a standard tool for reliability and risk practitioner. Discrete Event Simulation (DES) models appear a competitive alternative to the conventional reliability analysis models and systems analysis methods [3]-[5]. Systems subjected to certain inspection-repair-replacement policies can also be modelled in DES environments. This way, the analyst is not confined to any specific assumptions that she is not confident to. For example, the assumptions of the Markov model can easily be discarded and a more adequate solution can be implemented. DES models give a great deal of flexibility when striving for adequate system presentation and, if properly developed, they become an effective system reliability analysis tool, in particular, for systems operated under certain inspection-repair-replacement policies.

In this paper we compare the Markov model and DES modeling approach in terms of the results they produce for different state probabilities. To do so, we have chosen a rather simple set of examples.

II. EXAMPLE: POWER SUPPLY SYSTEM

Consider a simple power supply system the layout of which is shown in Figure 1 [7] to see what kind of insights one can get by having a DES model. In case of normal operation all busbars are fed from the grid. If power supply from the grid fails, the main busbar and the emergency busbar are disconnected. The diesel generator starts and is switched to the emergency busbar. The following system states can be defined:

State 1: Grid is in operation (A) - Diesel generator is available (B)

State 2: Grid has failed and is under repair (A) - Diesel generator is in operation (B)

State 3: Grid has failed and is under repair (A) - Diesel generator has failed and is under repair (B)

State 4: Grid has been restored (A) - Diesel generator is under repair (B)

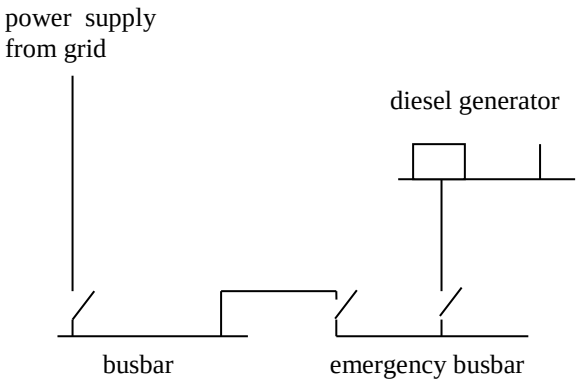


Figure 1. System layout of the power supply system

The corresponding state diagram is depicted in Figure 2. Special attention is called for the failure-to-start probability, Q_B, of the emergency diesel generator.

In state 1 both busbars are fed from the grid. The diesel generator is available but does not run. In state 2 power supply from grid is not available but the diesel generator has started and is in operation. In state 3 power supply from both the grid and the diesel generator is not available. The unavailability of the diesel generator can be a result of either a failure to start or a failure to run after a successful start. Two repair teams are at disposal to restore the subsystems concurrently in case they both fail. In state 4 power supply from the grid is restored, however the diesel generator is still under repair.

III. MARKOV MODEL OF THE POWER SUPPLY SYSTEM

Assume that the failure and repair rates of the grid and diesel generator are constant, so the transition rates between the different systems states are homogeneous, which means that the failure process as well as the repair process is exponentially distributed. Then by constructing a transition matrix, we can get for each state steady-state solutions which are attained after having the system run for a long enough period of time.

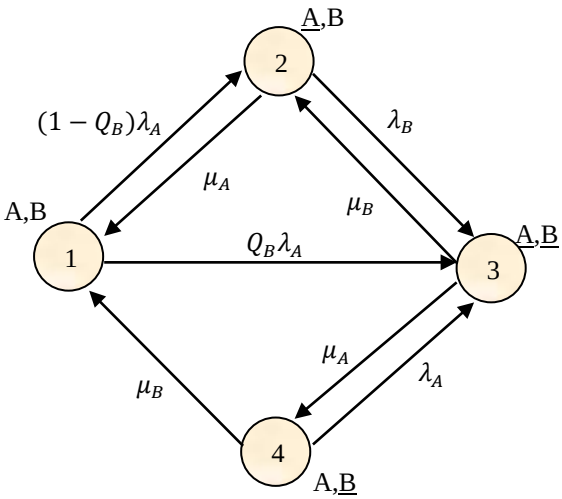


Figure 2. State diagram of the power supply system

Let λ_A and λ_B be the failure rates for the grid and diesel generator, respectively, while μ_A and μ_B be their repair rates. The probability of failure to start the diesel generator is denoted by Q_B. The transitions from state to state are shown in Figure 2, from which we can construct the transition matrix of the system and write Kolmogorov equations for computing the state probabilities P_1, P_2, P_3, and P_4:

$$\begin{bmatrix} \frac{dP_1}{dt} \\ \frac{dP_2}{dt} \\ \frac{dP_3}{dt} \\ \frac{dP_4}{dt} \end{bmatrix} = \begin{bmatrix} -Q_B\lambda_A-(1-Q_B)\lambda_A & \mu_A & 0 & \mu_B \\ (1-Q_B)\lambda_A & -\lambda_B-\mu_A & \mu_B & 0 \\ Q_B\lambda_A & \lambda_B & -\mu_A-\mu_B & \lambda_A \\ 0 & 0 & \mu_A & -\lambda_A-\mu_B \end{bmatrix} \cdot \begin{bmatrix} P_1 \\ P_2 \\ P_3 \\ P_4 \end{bmatrix}$$

It must also hold that

$$P_1+P_2+P_3+P_4=1.$$

For the steady-states, all derivatives $\frac{dP_i}{dt}$ are equalized to zero and, by doing this, the above system of equations becomes the system of algebraic equations that are rather easy to solve. With the help of MATLAB the following formulas were derived:

$$P_1 = \frac{\mu_A{\mu_B}^2 + {\mu_A}^2\mu_B + \lambda_A\mu_A\mu_B + \lambda_B\mu_A\mu_B}{A+B}, \quad (1)$$

$$P_2 = \frac{\lambda_A{\mu_B}^2 + {\lambda_A}^2\mu_B + \lambda_A\mu_A\mu_B - \lambda_A\mu_A\mu_BQ_B}{A+B}, \quad (2)$$

$$P_3 = \frac{{\lambda_A}^2\lambda_B + \lambda_A\lambda_B\mu_B + {\lambda_A}^2\mu_AQ_B + \lambda_A\mu_A\mu_BQ_B}{A+B}, \quad (3)$$

$$P_4 = \frac{\lambda_A Q_B {\mu_A}^2 + \lambda_A \lambda_B \mu_A}{A + B} \qquad (4)$$

where

$$A = Q_B {\lambda_A}^2 \mu_A + {\lambda_A}^2 \mu_B + \lambda_B {\lambda_A}^2 + Q_B \lambda_A {\mu_A}^2 + 2\lambda_A \mu_A \mu_B$$

$$B = \lambda_B \lambda_A \mu_A + \lambda_A {\mu_B}^2 + \lambda_B \lambda_A \mu_B + {\mu_A}^2 \mu_B + \mu_A {\mu_B}^2 + \lambda_B \mu_A \mu_B$$

It should be noted that formula (3) is different from that given in [7], as we did not neglect any terms, while the other formulas for state probabilities P_2, P_3, and P_4 were not provided at all in any form in [7].

IV. DES Model of the Power Supply

The appeal of models developed in a DES environment is that their logic follows the natural way the modelled system behaves. For example, the model of the power supply system works as follows. (See the logic of the model on the diagram given in Figure 3. This figure is a screen shot of the model buiolt in Arena.)

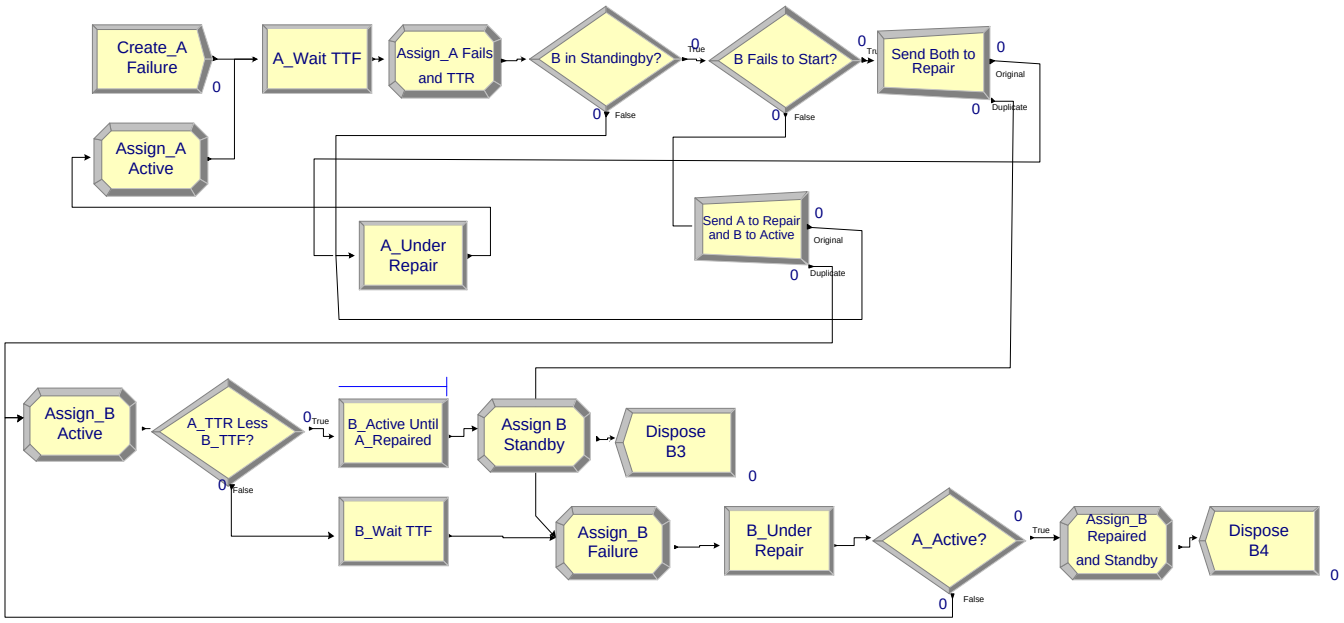


Figure 3. The diagram of the DES model of the power supply system as it appears in the Arena model window

At the instatnt of start running the model an entity is created in the "Create_A Failure" modul and sent further to the model. In the following modul, a time to failure governed by a specified probability distribution is generated and the entity is held in the "A_Wait TTF" modul for the the generated time. Then it moves to the Assign module, where the state of A is changed from "active" to "failure" and time to repair is generated according to a predefined probability distribution. Next, in the IF module it is checked whether the diesel generator, B, is in standby. If it is, then in the following IF module it is verified whether B fails to start. If it is, then from the following Separate module two entities are sent to the corresponding modules mimicking the repair of A and B. This is done in the modules "A_Under Repair" and "B_Under Repair" that simply delay the move of the entities for the specified times to repair of the both subsystems. The reader can further follow the logic and behaviour of the model on the diagram Figure 3. The model is run for a predefined period of time mimicking the system operation for thouzands of years. While the model is running, the times spent in the different states are accumulated and when the model run ends, the accumulated times are divided by the simulation time providing the state probabilities as an output.

The DES model of the power supply system was validated by comparing the values of the state probabilities obtained by simulation with those computed by formulas (1) – (4). A very high precision agreement was observed for the simulated and computed results. In this way, confidence to formulas (1) – (4) becomes higher as well. The two-way validation can be of utmost importance for the reliability analyst even for a simple system like that shown in Figure 2.

For example, let us take state probability P_3 as it is given in [7]:

$$P_3 \approx \frac{\lambda_A \times \lambda_B + \lambda_A \times \mu_A \times Q_B}{\mu_A(\mu_A + \mu_B)} \qquad (5)$$

It may be not straightforward and easy to come to the conclusion that formula (5) is an approximation of (3). As well as it may be not obvious that (3) is correct. By running the

simulation model exemplified in Figure 3 and comparing the results with those analytically computed the confidence to the both becomes definitely higher.

Another important point in support to mutual benefit for both Markov and DES model is a sensitivity or robustness analysis that can be rather easily conducted by DES. As soon as a DES model has been validated, one can drop the assumption of exponentially distributed time between failures and time to repair. The results obtained based on other distribution laws can be easily generated and by doing this the sensitivity of the model to the change of distributions can be numerically analysed.

Collecting representative samples of failure observations is a real problem reliability analysts face. The assessment of mean times between failures (MTBFs) and the variances based on a limited number of observations can rarely support the definitive conclusion that the times are exponentially distributed. Nonetheless, not having another tool except for the Markov model, the analyst is compelled to use the exponential distributions. How differently the results would be if TBFs were governed by other alternative distributions? Answering this question is often a requirement of reliability and risk analyses.

Besides the exponential distribution we have chosen three other (Rayleigh, log-normal and truncated normal) to see how the modeling results are sensitive to the change of the distribution. The parameters of the probability distributions were calculated based on a chosen fixed MTBFs and mean time to repair (MTTR). Surprisingly, the results of this exercise have demonstrated good model robustness.The highest spread in probability was for state 4. Although, even for a rather unrealistic ratio between MTBF and MTTR for the grid and generator (20 time units for MTBF: 1 time unit for MTTR) the range was not noticeably broad. More realistic ratios like 100:1 or 200:1 show more narrow uncertainty intervals for the state probabilities. A plot of the outputted results for the worst modeled case is presented in Figure 4. The differences in the probability are observed only in the third decimal.

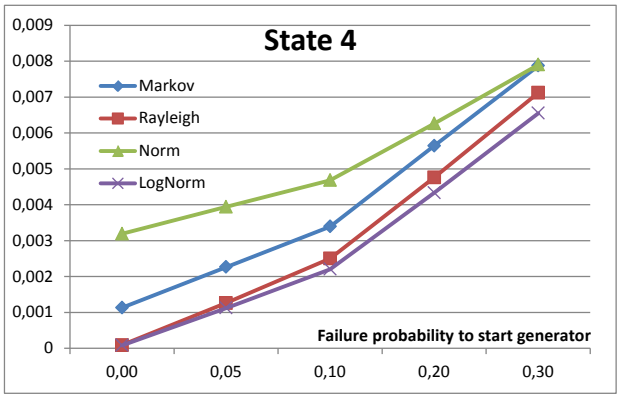


Figure 4. The probability of residing in State 4 obtained by simulation for the ration between MTBF and MTTR as 20:1

V. DES IN DEFENSE OF MARKOV MODEL

In this section we provide an example of fallacious in our opinion statements about some flaws of Markov models. To check the validity of the seemingly unquestionable statements, we computed the state probabilities by both the analytical approach and DES. By comparing the results, the conclusion was made in support to adequate modelling by the Markov process.

In [8] Markov models of multi-disk fault tolerant systems are briefly discussed. It is stated that the accuracy of Markov models and their utility decreases as the redundancy in the system increases. “For multi-disk fault tolerant systems, both rebuild models, the serial and concurrent (Figure 5), are incorrect. The rebuild transitions for states 2 through m are incorrect: they model the rebuild of the disk that failed most recently, whereas reliability is dominated by the rebuild of the disk that failed earliest. In essence, traditional Markov models reset the rebuild time for all disks being rebuilt whenever another disk fails. The traditional serial rebuild Markov model thus models a rebuild policy in which each subsequent disk failure changes which disk is being rebuilt, and “re-fails” the disk currently being rebuilt. The traditional concurrent rebuild Markov model thus models a rebuild policy in which each subsequent disk failure restarts the rebuild of all failed disks.” These assertions appeared to us logical until DES modeling proven their invalidity.

We have modelled a system consisting of five disks running in parallel and assumed that the system is operational until all five disks fail. Failure and repair rates were taken the same for the all disks. The both rebuild policies were subject to the modeling by the both methods.

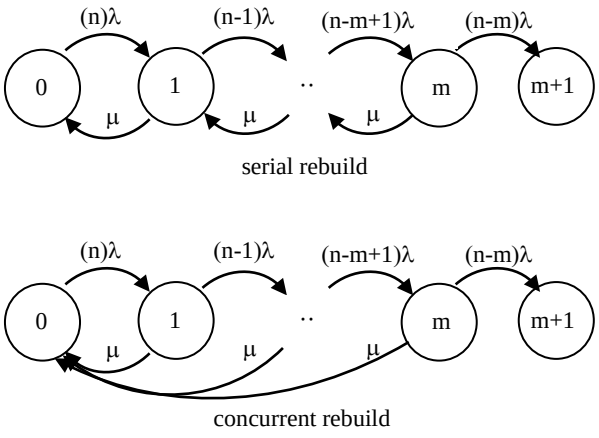


Figure 5. Traditional Markov models for rebuild policies

The analytical solutions for the state probabilities are given by the following formulas (see for example [9])

$$p_k = p_0 \prod_{i=1}^{k} \frac{\lambda_i}{\mu_i}, \quad p_0 = \frac{1}{1 + \sum_{k=1}^{n} \prod_{i=1}^{k} \frac{\lambda_i}{\mu_i}} \qquad (6)$$

The logics of the DES models are very simple. Figure 6 exemplifies the one for the concurrent rebuild policy.

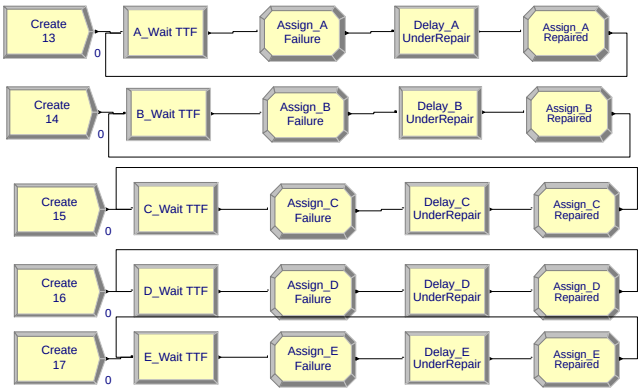


Figure 6. The outlook of the DES model made in Arena for the concurrent rebuild policy

The computed probabilities (6) of being in each of the six states and the results of the simulation have demonstrated very high agreement for all states. The expectation of having conservative reliability measures has not been supported by the modelling results.

VI. An Example of Explicit Advantage of DES

As stated in the introduction, the usually stressed major drawback of the Markov method is the explosion in the number of states as the size of the system increases. In fact, this drawback can be observed even for very simple systems like the ones depicted in Figure 5 with the only difference that each component (disk) has a distinctive failure rate. If this is the case, we have to enumerate all possible combinations of the components' failure each of which will represent a distinctive state. For a system of five components, the number of states will amount to 2^5 with numerous transition intersections on a state diagram. Assuming on top of that different repair rates for the all components results in a non-overviewable state diagram and a very-difficult-to-solve problem.

On contrary, the complexity of the DES model does not change at all when assigning different values to MTBFs and MTTRs, which enables to conduct nuanced analyses of the system

VII. Conclusion

The continuing increase of computer power and growing functionality of software tools supporting mathematical and reliability computations change the way the analysts tackle the problems they face. Analytical reliability models have undisputable advantages over numerical computations given they adequately count for important features of the system. Numerical models, including advanced Monte-Carlo simulation, can in turn be easily detached from restrictive assumptions of the analytical models, which gives the analyst a greater flexibility in building adequate models. In some cases, the two modeling approaches can be used to complement and validate each other.

Our experience in applying DES models shows one more their positive feature. As they simply mimic the behavior of the systems in time, they are easily understandable by domain experts that are not experienced in abstract mathematical modeling. This way the domain experts become collaborators in model development and contribute to model validation and greater confidence to the outputted results. That is to say, a frequently existing gap between the "black boxes" of complex mathematical models and a lack of confidence to them from the practitioners' side can be bridged by employing the alternative modeling approach.

The advantageous use of the DES models compared to Markov models has been stressed in the medical domain. As stated in [6], the DES model predicts the course of the HIV disease naturally, with few restrictions. This may give the model superior face validity with decision makers. Furthermore, this model automatically provides a probabilistic sensitivity analysis, which is cumbersome to perform with a Markov model. DES models allow inclusion of more variables without aggregation, which may improve model precision. The capacity of DES for additional data capture helps explain why this model consistently predicts better survival and thus greater savings than the Markov model. The DES model is better than the Markov model in isolating long-term implications of small but important differences in crucial input data.

A shortcoming of the use of DES consists in dependence on a specific simulation environment (software) in which the model is built and run. High costs of DES software and inability to run DES models in other environments, except for the one where they have been built, are the limitations against the analytical results obtained on Markov modelling.

References

[1] N.J. McCornick, Reliability and risk analysis. Academic Press, 1981.

[2] K.S. Trivedi and D. Selvamuthu, "Markov Modelling in Reliability" in Encyclopedia of Quantitative Risk Analysis and Assessment, vol. 3, E.L. Melnick and B.S. Everitt, Eds. Wiley, 2088, pp. 1045-1049.

[3] I. Kozine, F. Markert, and A. Alapetite; Discrete event simulation in support to hydrogen supply reliability, Paper 159, International Conference on Hydrogen Safety ICHS-3, Ajaccio, Corsica, France; September, 17th 2009

[4] I. Kozine, Discrete event simulation versus conventional system reliability analysis approaches, Reliability, Risk and Safety (2010), Eds. Ale, Papazoglou & Zio, Taylor & Francis Group London, ISBN 978-0-415-60427-7.

[5] I. Kozine, Simulation of human performance in time-pressured scenarios, Proc. IMechE Vol.221 Part O: J. Risk and Reliability (2007) 141-151

[6] K. N. Simpson, A. Strassburger, W.J. Jones, B. Dietz and R. Rajagopalan, Comparison of Markov Model and Discrete-Event Simulation Techniques for HIV. Pharmacoeconomics 27 (2), 2099, pp. 159-165.

[7] Shuller, J.C.H., et al. Methods for determining and processing probabilities. 'Red Book'. Second edition. CPR 12E. VROM. The Hague, 1997

[8] K. Greenan and J. J.Wylie. Reliability Markov models are becoming unreliable (WIP submission). http://www.usenix.org/events/fast08/wips_posters/greenan-wip.pdf

[9] J. L. Hafner and K. Rao. Notes on reliability models for non-MDS erasure codes. Technical Report RJ– 10391, IBM, October 2006.

Goodness of Fit Tests for Diffusion Processes

Yury A. Kutoyants
Université du Maine
Laboratoire de Statistique et Processus,
72085 Le Mans, France

***Abstract*—We present a review of several problems of goodness of fit testing by the observations of ergodic diffusion processes in continuous time. The basic hypothesis is supposed to be simple and composite. Our goal is to find the asymptotically distribution free and asymptotically parameter free tests. All tests are based on empirical distribution function and local time estimator of the invariant density.**

I. Introduction

If we observe n i.i.d. r. v.'s $(X_1,\dots,X_n)=X^n$ with distribution function $F(x)$ and the basic hypothesis is simple

$$\mathscr{H}_0,\qquad F(x)\equiv F_*(x),\qquad x\in\boldsymbol{R},$$

then the Cramér-von Mises and Kolmogorov-Smirnov statistics are

$$W_n^2=n\int\left[\hat{F}_n(x)-F_*(x)\right]^2\mathrm{d}F_*(x),$$
$$D_n=\sup_x\sqrt{n}\left|\hat{F}_n(x)-F_*(x)\right|$$

respectively. Here

$$\hat{F}_n(x)=\frac{1}{n}\sum_{j=1}^n\mathbb{1}_{\{X_j<x\}}$$

is the empirical distribution function. Let us denote by $\mathcal{K}_\alpha$ the class of tests of asymptotic size α, i.e.;

$$\mathcal{K}_\alpha=\left\{\bar{\psi}\quad:\quad\mathbf{E}_0\,\bar{\psi}=\alpha+o(1)\right\}.$$

We have the convergence

$$W_n^2\Longrightarrow\int_0^1W_0(s)^2\,\mathrm{d}s,\quad D_n\Longrightarrow\sup_{0\le s\le1}|W_0(s)|,$$

where $W_0(\cdot)$ is Brownian bridge. Introduce the constants c_α,d_α:

$$\mathbf{P}\left\{\|W_0(\cdot)\|^2>c_\alpha\right\}=\alpha,\quad\mathbf{P}\left\{\|W_0(\cdot)\|_\infty>d_\alpha\right\}=\alpha.$$

Then the *Cramér-von Mises* and *Kolmogorov-Smirnov* tests

$$\psi_n(X^n)=\mathbb{1}_{\{W_n^2>c_\alpha\}}\in\mathcal{K}_\alpha,\qquad\phi_n(X^n)=\mathbb{1}_{\{D_n>d_\alpha\}}\in\mathcal{K}_\alpha.$$

These tests are

- *asymptotically distribution-free* , belong to $\mathcal{K}_\alpha$,
- uniformly consistent against any alternative

$$\mathscr{H}_\rho=\{F(\cdot):\|F(\cdot)-F_*(\cdot)\|\ge\rho\},\qquad\rho>0.$$

Our goal is to present a similar theory in the case of diffusion processes observed in continuous time.

Diffusion Processes

Let $X=\{X_t,0\le t\le T\}$ be an observation of solution of some SDE

$$\mathrm{d}X_t=S(X_t)\,\mathrm{d}t+\sigma(X_t)\,\mathrm{d}W_t,\quad X_0,\quad0\le t\le T,$$

and we would like to know if this SDE is of the following form

$$\mathrm{d}X_t=S_*(X_t)\,\mathrm{d}t+\sigma(X_t)\,\mathrm{d}W_t,\quad X_0,\quad0\le t\le T,$$

where the trend $S_*(\cdot)$ and diffusion coefficient $\sigma(\cdot)^2$ are known functions.

II. Test

Hence we have to test the following two hypotheses

$$\mathscr{H}_0,\qquad S(x)=S_*(x),$$
$$\mathscr{H}_1,\qquad S(x)\ne S_*(x).$$

Our goal is to construct a test (called *Goodness-of-Fit*) which can solve this hypotheses testing problem and are

- asyptotically distribution free, belong to $\mathcal{K}_\alpha$,
- uniformly consistent against a wide classe of alternatives

The tests studied are of the form

$$\psi(X)=\mathbb{1}_{\{\Delta(X)>c_\alpha\}},\qquad\lim\mathbf{E}_0\psi(X)=\alpha$$

We study such tests in two types of asymptotics: *small noise* ($\sigma\to0$) and *large samples* ($T\to\infty$).

Small Noise Asymptotics

Suppose that the observed process $X^\varepsilon=\{X_t,0\le t\le T\}$ is

$$\mathrm{d}X_t=S(X_t)\,\mathrm{d}t+\varepsilon\sigma(X_t)\,\mathrm{d}W_t,\qquad X_0=x_0,\quad0\le t\le T.$$

If $\varepsilon\to0$ then the stochastic process X^ε converges to the deterministic function $\{x_t,0\le t\le T\}$, solution of the ordinary DE

$$\frac{\mathrm{d}x_t}{\mathrm{d}t}=S(x_t),\qquad x_0,\quad0\le t\le T.$$

Our goal is to construct the GoF tests (C-vM and K-S type) for this model. We suppose that $S_*(x)>0,x\in[x_0,x_T]$.

The basic hypothesis is simple:

$$\mathscr{H}_0:\qquad S(x)=S_*(x),\quad x\in[x_0,x_T^*],$$
$$\mathscr{H}_\rho:\qquad S(x)\in\mathcal{F}_\rho$$

Here x_t^* is solution x_t under hypothesis $\mathscr{H}_0$.

$$\frac{\mathrm{d}x_t^*}{\mathrm{d}t} = S_*\left(x_t^*\right), \quad 0 \le t \le T,$$
$$\mathcal{F}_\rho = \left\{S\left(\cdot\right) : \left\|x_\cdot - x_\cdot^*\right\| \ge \rho\right\}, \qquad \rho > 0$$

Introduce two statistics

$$W_\varepsilon^2 = \int_0^T \left(\frac{X_t - x_t^*}{\tau\,\varepsilon\,S_*\left(x_t^*\right)^2}\right)^2 \sigma\left(x_t^*\right)^2\,\mathrm{d}t,$$
$$D_\varepsilon = \sup_{0\le t\le T}\left|\frac{X_t - x_t^*}{\sqrt{\tau}\,\varepsilon\,S_*\left(x_t^*\right)}\right|,$$
$$\tau\left(s\right) = \int_0^s \left(\frac{\sigma\left(x_t^*\right)}{S_*\left(x_t^*\right)}\right)^2 \mathrm{d}t, \quad \tau = \tau\left(T\right)$$

We have the convergence (under $\mathscr{H}_0$)

$$W_\varepsilon^2 \Longrightarrow \int_0^1 W\left(s\right)^2\,\mathrm{d}s, \qquad D_\varepsilon \Longrightarrow \sup_{0\le s\le 1}\left|W\left(s\right)\right|.$$

Hence if we choose the constants c_α, d_α satisfying

$$\mathbf{P}\left\{\int_0^1 W\left(s\right)^2\,\mathrm{d}s > c_\alpha\right\} = \alpha,$$
$$\mathbf{P}\left\{\sup_{0\le s\le 1}\left|W\left(s\right)\right| > d_\alpha\right\} = \alpha,$$

then the *C-vM and K-S type tests*

$$\psi_\varepsilon\left(X^\varepsilon\right) = \mathbb{1}_{\left\{W_\varepsilon^2 > c_\alpha\right\}} \in \mathcal{K}_\alpha, \quad \phi_\varepsilon\left(X^\varepsilon\right) = \mathbb{1}_{\left\{D_\varepsilon > d_\alpha\right\}} \in \mathcal{K}_\alpha$$

The both tests are *asymptotically distribution free* and uniformly consistent against any alternative $\mathscr{H}_\rho$.

On local time and GoF testing

The local time of the diffusion process is defined as

$$\Lambda_T\left(x\right) = \lim_{\nu\downarrow 0}\frac{\varepsilon^2}{2\nu}\int_0^T \mathbb{1}_{\left\{\left|X_t - x\right|\le\nu\right\}}\,\sigma\left(X_t\right)^2\,\mathrm{d}t$$

and admits the Tanaka-Meyer representation

$$\Lambda_T\left(x\right) = \left|X_T - x\right| - \left|x_0 - x\right| - \int_0^T \mathrm{sgn}\left(X_t - x\right)\,\mathrm{d}X_t.$$

Note that in ergodic case ($\varepsilon \equiv 1$ and $T \to \infty$), the local time is asymptotically normal :

$$\frac{\Lambda_T\left(x\right)}{T\sigma\left(x\right)^2} \to f\left(x\right), \quad \sqrt{T}\left(\frac{\Lambda_T\left(x\right)}{T\sigma\left(x\right)^2} - f\left(x\right)\right) \Longrightarrow \mathcal{N},$$

In small noise case

$$\lim_{\varepsilon\to 0}\frac{\Lambda_T\left(x\right)}{\varepsilon^2} = \frac{\sigma\left(x\right)^2}{S_*\left(x\right)}, \quad \text{for} \quad x \in \left[x_0, x_T^*\right],$$

but

$$\frac{1}{\varepsilon}\left(\frac{\Lambda_T\left(x\right)}{\varepsilon^2} - \frac{\sigma\left(x\right)^2}{S_*\left(x\right)}\right) \quad \text{has no limit}$$

It is shown that

$$\frac{1}{\varepsilon}\int_{x_0}^x \left(\frac{1}{S_0\left(y\right)} - \frac{\Lambda_T\left(y\right)}{\varepsilon^2\sigma\left(y\right)^2}\right)\mathrm{d}y \Longrightarrow W\left(\int_{x_0}^x \frac{\sigma\left(y\right)^2}{S_0\left(y\right)^3}\,\mathrm{d}y\right).$$

Hence if we put

$$\delta_\varepsilon = \frac{\displaystyle\int_{x_0}^{x_T}\frac{\sigma\left(x\right)^2}{S_0\left(x\right)^3}\left(\int_{x_0}^x\left(\frac{1}{S_0\left(y\right)} - \frac{\Lambda_T\left(y\right)}{\varepsilon^2\sigma\left(x\right)^2}\right)\mathrm{d}y\right)^2\mathrm{d}x}{\left(\varepsilon\displaystyle\int_{x_0}^{x_T}\frac{\sigma\left(x\right)^2}{S_0\left(x\right)^3}\mathrm{d}x\right)^2}.$$

Then

$$\delta_\varepsilon \Longrightarrow \int_0^1 W\left(s\right)^2\,\mathrm{d}s$$

and the test $\psi_\varepsilon = \mathbb{1}_{\left\{\delta_\varepsilon > c_\alpha\right\}}$ with the corresponding c_α belongs to $\mathcal{K}_\alpha$.

For the proofs see [5]

Large samples asymptotics

III. Simple Basic Hypothesis

The observed process under hypothesis is

$$\mathrm{d}X_t = S_*\left(X_t\right)\mathrm{d}t + \sigma\left(X_t\right)\mathrm{d}W_t, \quad 0 \le t \le T.$$

The empirical distribution function is

$$\hat{F}_T(x) = \frac{1}{T}\int_0^T \mathbb{1}_{\{X_t<x\}}\mathrm{t}.$$

and the local time estimator of the invariant density

$$\hat{f}_T\left(x\right) = \frac{\Lambda_T\left(x\right)}{T\sigma\left(x\right)^2}.$$

The goodness of fit tests are based on these two estimators.

Local time estimator. Let us denote μ the median of invariant law and consider the HT problem with *one-sided alternatives*: changes are for $x \ge \mu$. Put

$$V_T^2 = T\int_\mu^\infty h\left(x\right)\left(\hat{f}_T\left(x\right) - f_*\left(x\right)\right)^2\mathrm{d}F_*\left(x\right)$$

with

$$h\left(x\right) = \frac{2F_*\left(x\right) - 1}{4\Phi\left(\mu\right)^2\sigma\left(x\right)^2 f_*\left(x\right)^4}\;e^{-\Phi(x)/\Phi(\mu)}\;1_{\{x\ge\mu\}},$$

and

$$\Phi\left(x\right) = \int_{-\infty}^\infty \frac{\left(1_{\{y>x\}} - F_*\left(y\right)\right)^2}{\sigma\left(y\right)^2 f_*\left(y\right)}\,\mathrm{d}y.$$

We show that

$$V_T^2 \Longrightarrow V^2 = \int_1^\infty W\left(s\right)^2 e^{-s}\mathrm{d}s$$

Hence the test $\psi_T = \mathbb{1}_{\left\{V_T^2 > c_\alpha\right\}}$ is ADF and consistent [4]. To calculate the threshold we can use the K-L expansion [2]

$$W(s) = \sum_{n=1}^\infty \frac{\sqrt{2}}{\delta_n}\,\xi_n\;\frac{J_0\left(\delta_n\,e^{-(s-1)}\right)}{\sqrt{\left(\delta_n^2+1\right)}\;J_1(\delta_n)}, \quad s \ge 1$$

where $\{\xi_n, n \ge 1\}$, are i.i.d. $\mathcal{N}(0,1)$ random variables and $\{\delta_n, n \ge 1\}$ are the positive zeros of the equation $J_0(\delta_n) - \delta_n J_1(\delta_n) = 0$. We have

$$V^2 = \sum_{n=1}^\infty \frac{\xi_n^2}{\delta_n^2}$$

Empirical Distribution function. Let us put

$$W_T^2 = T\int_{-\infty}^{\infty} H\left(x\right)\left(\hat{F}_T\left(x\right)-F_*\left(x\right)\right)^2 \mathrm{d}F_*\left(x\right)$$

where

$$H\left(x\right)=\frac{\Psi'\left(x\right)}{f_*\left(x\right)\left[F_*\left(x\right)-1\right]^2}\ e^{-\Psi(x)} \tag{1}$$

and

$$\Psi\left(x\right)=\int_{-\infty}^{x}\frac{F_*\left(y\right)^2}{\sigma\left(y\right)^2 f_*\left(y\right)}\mathrm{d}y + F_*\left(x\right)^2\int_x^{\infty}\left(\frac{F_*\left(y\right)-1}{F_*\left(x\right)-1}\right)^2\frac{\mathrm{d}y}{\sigma\left(y\right)^2 f_*\left(y\right)}.$$

We show that

$$W_T^2 \Longrightarrow \int_0^{\infty} W\left(s\right)^2 e^{-s}\mathrm{d}s \equiv W^2$$

Hence the test

$$\hat{\psi}_T = \mathbb{I}_{\left\{W_T^2>c_\alpha\right\}},\qquad \mathbf{P}\left\{W^2>c_\alpha\right\}=\alpha$$

belongs to $\mathcal{K}_\alpha$ [4]. To calculate the threshold we can use the K-L expansion [2]

$$W(s)=\sum_{n=1}^{\infty}\frac{\sqrt{2}}{z_{0,n}}\xi_n\ \frac{J_0\left(z_{0,n}\,e^{-s)}\right)}{J_1(z_{0,n})},\quad s\geq 0$$

where $\{\xi_n, n\geq 1\}$, are i.i.d. $\mathcal{N}(0,1)$ random variables and $\{z_{0,n}, n\geq 1\}$ are the positive zeros of the Bessel function $J_0\left(\cdot\right)$. We have

$$W^2=\sum_{n=1}^{\infty}\frac{\xi_n^2}{z_{0,n}^2}$$

IV. Composite Basic Hypothesis

The observed process under hypothesis is

$$\mathrm{d}X_t = S_*\left(\vartheta, X_t\right)\mathrm{d}t+\sigma\left(X_t\right)\mathrm{d}W_t,\quad 0\leq t\leq T$$

where $\vartheta\in\Theta$ is unknown parameter.

Remind that in the case of i.i.d. observations with distribution $F\left(\vartheta,x\right)$ depending on unknown parameter ϑ the Cramer-von Mises test is based on the statistics

$$\Delta_n = n\int_{-\infty}^{\infty}\left[\hat{F}_n\left(x\right)-F\left(\hat{\vartheta}_n,x\right)\right]^2\mathrm{d}F\left(\hat{\vartheta}_n,x\right)$$

and its limit distribution depends of the value ϑ. There are at least two particular cases when this limit distribution does not depend on ϑ : in the case of shift and scale parameters. The distributions of the Cramér-von Mises type statistics with estimated parameters were studied by many authors.

In our case we propose two Cramer-von Mises type statistics similar to Δ_n. At particularly, we study the statistics like

$$\Delta_T = T\int_{-\infty}^{\infty}\left[\hat{F}_T\left(x\right)-F\left(\hat{\vartheta}_T,x\right)\right]^2\mathrm{d}F\left(\hat{\vartheta}_T,x\right),$$

where $\hat{F}_T\left(x\right)$ is empirical distribution function and $\hat{\vartheta}_T$ is the MLE. Remind that under regularity conditions the MLE is asymptotically normal [3]

$$\sqrt{T}\left(\hat{\vartheta}_T-\vartheta\right)\Longrightarrow\xi\left(\vartheta\right)\ \sim\ \mathcal{N}\left(0,\mathrm{I}\left(\vartheta\right)^{-1}\right).$$

Here $\mathrm{I}\left(\vartheta\right)$ is the information matrix. It can be shown that under regularity conditions the statistic Δ_T has the following behaviour (as $T\to\infty$)

$$\begin{aligned}\Delta_T &= \int_{-\infty}^{\infty}\left[\sqrt{T}\left(\hat{F}_T\left(x\right)-F\left(\vartheta,x\right)\right)\right.\\ &\qquad\left.+\sqrt{T}\left(F\left(\vartheta,x\right)-F\left(\hat{\vartheta}_T,x\right)\right)\right]^2\mathrm{d}F\left(\hat{\vartheta}_T,x\right)\\ &= \int_{-\infty}^{\infty}\left[\eta_T\left(x\right)-\sqrt{T}\left(\left(\hat{\vartheta}_T-\vartheta\right)\dot{F}\left(\tilde{\vartheta},x\right)\right)\right]^2\mathrm{d}F\left(\hat{\vartheta}_T,x\right)\\ \Longrightarrow &\int_{-\infty}^{\infty}\left[\eta\left(\vartheta,x\right)-\left(\xi\left(\vartheta\right),\dot{F}\left(\vartheta,x\right)\right)\right]^2 f\left(\vartheta,x\right)\mathrm{d}x\equiv\Delta\end{aligned}$$

where $\eta\left(\vartheta,x\right)$ is a Gaussian process which admits the representation

$$\eta\left(\vartheta,x\right)=2\int_{-\infty}^{\infty}\frac{F\left(\vartheta,x\wedge y\right)-F\left(\vartheta,y\right)F\left(\vartheta,x\right)}{\sqrt{f\left(\vartheta,y\right)}}\ \mathrm{d}W\left(y\right).$$

Therefore, the limit distribution of the statistic Δ_T depends on the parameter ϑ and the choice of the threshold c_α can be a difficult problem.

It is interesting to find models for which this limit distribution does not depend on ϑ. Then the choice of the threshold c_α can be done much more easier.

Let us consider the model of ergodic diffusion process with shift parameter (under hypothesis)

$$\mathrm{d}X_t = S_*\left(X_t-\vartheta\right)\mathrm{d}t+\mathrm{d}W_t,\quad X_0, 0\leq t\leq T.$$

As before, we consider the GoF tests based on the local time empirical density estimators.

Local time estimator. Let us introduce the statistic

$$\delta_T = T\int_{-\infty}^{\infty}(\hat{f}_T(x)-f(x-\hat{\vartheta}_T))^2\mathrm{d}x,$$

where $f\left(x\right)$ is the invariant density of the observed process with $\vartheta=0$. We have the converges in distribution to

$$\delta=\int_{-\infty}^{\infty}\left(\int_{-\infty}^{\infty}\left(2f(x)\frac{\mathbb{I}_{\{y>x\}}-F(y)}{\sqrt{f(y)}}\right.\right.\\ \left.\left.-\frac{1}{I}S_*'(y)\sqrt{f(y)}f'(x)\right)\mathrm{d}W(y)\right)^2\mathrm{d}x,$$

with $W(y)=W_1(y),\ y\in\mathbb{R}^+$, $W(y)=W_2(-y),\ y\in\mathbb{R}^-$, where W_1 and W_2 are independent Wiener processes.

Therefore the test $\psi_T=\mathbb{I}_{\{\delta_T>c_\alpha\}}$ belongs to $\mathcal{K}_\alpha$ [8].

Empirical distribution function.

The second test is based on the same MLE and the empirical distribution function (EDF).

The corresponding statistic is

$$\Delta_T = T\int_{-\infty}^{\infty}\left(\hat{F}_T(x)-F(x-\hat{\vartheta}_T)\right)^2\mathrm{d}x,$$

which converges in distribution to

$$\Delta=\int_{-\infty}^{\infty}\left(\int_{-\infty}^{\infty}\left(2\frac{F(y\wedge x)-F(y)F(x)}{\sqrt{f(y)}}\right.\right.\\ \left.\left.-\frac{1}{I}S_*'(y)\sqrt{f(y)}f(x)\right)\mathrm{d}W(y)\right)^2\mathrm{d}x.$$

hence the test $\psi_T = \mathbb{1}_{\{\delta_T > c_\alpha\}}$ belongs to $\mathcal{K}_\alpha$ [8].

V. GoF for O-U process

We suppose that under basic hypothesis we observe an Ornstein-Uhlenbeck (O-U) process

$$\mathrm{d}X_t = -\beta\left(X_t - \mu\right)\,\mathrm{d}t + \sigma\,\mathrm{d}W_t, \qquad X_0, \quad 0 \le t \le T$$

where $\vartheta = (\mu, \beta) \in \Theta$ is unknown parameter. The set $\Theta = (a_1, a_2) \times (b_1, b_2)$ and $b_1 > 0$. The process X_t has ergodic properties with invariant density $f(\vartheta, x) \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{2\beta}\right)$. The corresponding Cramer-von Mises type statistics are

$$\delta_T\left(X^T\right) = \sigma^2 T \int_{-\infty}^{\infty} \left|\hat{f}_T(x) - f\left(\hat{\vartheta}_T, x\right)\right|^2 \mathrm{d}F\left(\hat{\vartheta}_T, x\right)$$

and

$$\Delta_T\left(X^T\right) = \hat{\beta}_T T \int_{-\infty}^{\infty} \left|\hat{F}_T(x) - F\left(\hat{\vartheta}_T, x\right)\right|^2 \mathrm{d}F\left(\hat{\vartheta}_T, x\right).$$

Here $\hat{\vartheta}_T = \left(\hat{\mu}_T, \hat{\beta}_T\right)$ is the maximum likelihood estimator (MLE) of ϑ.

Introduce Gaussian processes

$$\zeta_f(x) = \int \left[2\frac{F_0(y) - \mathbb{1}_{\{y>x\}}}{f_0(y)} - 2x + \left(1 - 2x^2\right) y\right] \mathrm{d}W(y)$$

and

$$\zeta_F(x) = \int \left[2\frac{F_0(y)\,F_0(x) - F_0(y \wedge x)}{f_0(y)\,f_0(x)} + 1 + xy\right] \mathrm{d}W(y).$$

Here $W(y), y \in R$ is two-sided Wiener process. Define two constants c_α and d_α by the equations

$$\mathbf{P}\left\{\int_{-\infty}^{\infty} \zeta_f(x)^2 f_0(x)^3\,\mathrm{d}x > c_\alpha\right\} = \alpha,$$
$$\mathbf{P}\left\{\int_{-\infty}^{\infty} \zeta_F(x)^2 f_0(x)^3\,\mathrm{d}x > d_\alpha\right\} = \alpha.$$

The tests

$$\psi_T\left(X^T\right) = \mathbb{1}_{\{\delta_T > c_\varepsilon\}}, \qquad \phi_T\left(X^T\right) = \mathbb{1}_{\{\Delta_T > d_\varepsilon\}}$$

belong to $\mathcal{K}_\alpha$. [7].

Introduce the random process

$$Z_t = \frac{\sqrt{\beta}\left(X_t - \alpha\right)}{\sigma}, \qquad 0 \le t \le T.$$

The equation O-U we can re-write as

$$\mathrm{d}\frac{\sqrt{\beta}\left(X_t - \alpha\right)}{\sigma} = -\frac{\sqrt{\beta}\left(X_t - \alpha\right)}{\sigma}\,\mathrm{d}\left(\beta t\right) + \sqrt{\beta}\,\mathrm{d}W_t,$$

and so we obtain the equation for Z_t:

$$\mathrm{d}Z_t = -Z_t\,\mathrm{d}\left(\beta t\right) + \sqrt{\beta}\mathrm{d}W_t, \quad Z_0, \quad 0 \le t \le T.$$

If we put $Y_s = Z_{\frac{s}{\beta}}, 0 \le s \le T\beta$, then the equation for Y_s is

$$\mathrm{d}Y_s = -Y_s\,\mathrm{d}s + \,\mathrm{d}w_s, \qquad Y_0, \qquad 0 \le s \le T\beta,$$

where $w_s = \sqrt{\beta}W_{s/\beta}, s > 0$ is a Wiener process.

Then we represent the statistics Δ_T and δ_T as functionals of Y.

VI. Threshold models

Suppose that under hypothesis the observed diffusion process satisfies the equation

$$\mathrm{d}X_t = \sum_{j=1}^{k+1} S_j\left(X_t\right)\,\mathbb{1}_{\{\vartheta_{j-1} < X_t \le \vartheta_j\}}\mathrm{d}t + \sigma\left(X_t\right)\mathrm{d}W_t, \quad X_0,$$

where $\vartheta_0 = -\infty$, $\vartheta_j \in \Theta_j = (\alpha_j, \beta_j), j = 1, \ldots, k$, $\vartheta_{k+1} = \infty$, $\beta_j < \alpha_{j+1}$. The unknown parameter is $\boldsymbol{\vartheta} = (\vartheta_1, \ldots, \vartheta_k) \in \boldsymbol{\Theta} = \Theta_1 \times \ldots \times \Theta_k$. Introduce the statistic

$$\Delta_T\left(X^T\right) = T\int_{-\infty}^{\infty} H\left(\hat{\boldsymbol{\vartheta}}_T, x\right)\left|\hat{F}_T(x)\right.$$
$$\left. - F\left(\hat{\boldsymbol{\vartheta}}_T, x\right)\right|^2 \mathrm{d}F\left(\hat{\boldsymbol{\vartheta}}_T, x\right)$$

where the function $H(\boldsymbol{\vartheta}, x)$ is the same as in (1). It can be shown that $\Delta_T\left(X^T\right)$ (under hypothesis) converges to the limit

$$\Delta = \int_{-\infty}^{\infty} W(s)^2\,e^{-s}\mathrm{d}s.$$

Therefore the test $\psi_T = \mathbb{1}_{\{\Delta_T(X^T) > c_\alpha\}}$ belongs to $\mathcal{K}_\alpha$ [6].

References

[1] Dachian, S. and Kutoyants, Yu.A. (2007) On the goodness-of-fit tests for some continuous time processes, in *Statistical Models and Methods for Biomedical and Technical Systems*, F.Vonta *et al.* (Eds), Birkhäuser, Boston, 395-413.

[2] Gassem, A. (2010) Test d'ajustement d'un processus de diffusion ergodique à changement de règime. Thèse PhD, Université du Maine, Le Mans.

[3] Kutoyants, Yu.A. (2004) Statistical Inference for Ergodic Diffusion Processes, Springer, London.

[4] Kutoyants Yu. A., (2010) On the goodness-of-fit testing for ergodic diffusion processes. Journal of Nonparametric Statistics, 22, 4, 529-543.

[5] Kutoyants, Yu. A., (2011) On goodness-of-fit tests for perturbed dynamical systems. Journal of Statistical Planning and Inference, 141,1655-1666.

[6] Kutoyants Yu. A., (2012) On identification of the threshold diffusion processes, Annals of the Institute of Statistical Mathematics, 64, 2, 383-413.

[7] Kutoyants, Yu. A., (2012) Goodness-of-fit test for Ornstein-Uhlenbeck process. Submitted.

[8] Negri, I. and Zhou, L., (2012) On goodness-of-fit testing for ergodic diffusion process with shift parameter. Sumitted.

Conditional Distributions and Multivariate Statistical Scaling

Henning Läuter

Institute of Mathematics, University of Potsdam (Germany)

Abstract

We consider multivariate random variables Z_1, Z_2 where the variable Z_1 is considered as the endogenous variable and Z_2 is the response. We are looking at the conditional variable $\widetilde{U} := Z_2 \,|Z_1 = t$ and we will find the influence of t in this variable. We admit for Z_1 both metric and categorical variables. We model these conditional variables, find assumptions for estimability of parameters. For discrete conditional variables the χ^2-statistic is used for checking the statistical model. Here connections to results of M. Nikulin are discussed.

These results yield the basis to the statistical scaling of categorical variables. With this approach it is possible to find optimal scalings or also maximum likelihood scalings. As a consequence the asymptotic properties of the scalings can be found.

We point out that these results for scaling are the basis for some other statistical analyses of categorical data. We can apply our results for modeling contingency tables and in multivariate analysis of variance problems. Furthermore we find methods for discrimination and classification if the exploratory variables are categorically.

References:

[1] Agresti A.(2002). Categorical Data Analysis. Wiley, New York.

[2] Everett B.S., Dunn G. (2001). Applied Multivariate Data Analysis. Hodder Education, London.

[3] Greenwood P.E., Nikulin, M. (1996) A Guide to Chi-Squared Testing. John Wiley, New York

[4] Läuter H., Ramadan A.M. (2010). Modeling and scaling of categorical data. Preprint. 03/2010 Univ. Potsdam.

[5] Srivastava M.S. (2002). Methods of Multivariate Statistics. Wiley, New York.

Computer methods for «real-time» investigation of statistical regularities as means for ensuring correctness of statistical inferences in testing composite hypotheses of goodness-of-fit

Boris Yu. Lemeshko, Stanislav B. Lemeshko, Andrey P. Rogozhnikov
Department of applied mathematics
Novosibirsk State Technical University
Novosibirsk, Russia
lemeshko@fpm.ami.nstu.ru, rogozhnikov.andrey@gmail.com

***Abstract* — In present work, a "real-time" ability to simulate and research distributions of test statistics in a course of testing a composite goodness-of-fit hypothesis (for distributions with estimated parameters) is implemented by use of parallel computing. It makes it possible to make correct statistical inferences even in those situations when a distribution of a test statistic is unknown (before the testing procedure starts).**

Keywords — goodness-of-fit test; composite hypotheses testing; Kolmogorov test; Cramer-Mises-Smirnov test; Anderson-Darling test; methods of statistical simulation.

I. INTRODUCTION

In composite hypotheses testing in the form H_0: $F(x)\in\{F(x,\theta),\ \theta\in\Theta\}$, when the estimate $\hat{\theta}$ of a scalar or vector parameter of distribution and a test statistic are calculated by the same sample, the nonparametric goodness-of-fit tests of Kolmogorov (K), ω^2 Cramer-Mises-Smirnov (CMS), and Ω^2 Anderson-Darling (AD) lose their distribution-free property.

The value

$$D_n = \sup_{|x|<\infty}\left|F_n(x)-F(x,\theta)\right|,$$

where $F_n(x)$ is an empirical distribution function, n is a sample size, is used in Kolmogorov test as a distance between empirical and theoretical laws. When testing hypotheses, this statistic is usually used with Bolshev's correction [3] in the form [4]

$$S_K = \frac{6nD_n+1}{6\sqrt{n}}, \qquad (1)$$

where $D_n = \max(D_n^+, D_n^-)$, $D_n^+ = \max\limits_{1\le i\le n}\left\{\frac{i}{n}-F(x_i,\theta)\right\}$, $D_n^- = \max\limits_{1\le i\le n}\left\{F(x_i,\theta)-\frac{i-1}{n}\right\}$, $x_1, x_2, \ldots, x_n$ are sample values in an increasing order. The distribution of statistic (1) in testing simple hypotheses obeys Kolmogorov distribution law $K(S)=\sum\limits_{k=-\infty}^{\infty}(-1)^k e^{-2k^2s^2}$.

In ω^2 Cramer-Mises-Smirnov test, one uses the statistic in the form

$$S_\omega = n\omega_n^2 = \frac{1}{12n}+\sum_{i=1}^{n}\left\{F(x_i,\theta)-\frac{2i-1}{2n}\right\}^2, \qquad (2)$$

and in the test of Ω^2 Anderson-Darling type [1], [2], the statistic in the form

$$S_\Omega = -n-2\sum_{i=1}^{n}\left\{\frac{2i-1}{2n}\ln F(x_i,\theta)+\left(1-\frac{2i-1}{2n}\right)\ln(1-F(x_i,\theta))\right\}. \qquad (3)$$

In testing a simple hypothesis, the statistic (2) obeys the distribution [4] with the CDF

$$a1(S)=\frac{1}{\sqrt{2s}}\sum_{j=0}^{\infty}\frac{\Gamma(j+1/2)\sqrt{4j+1}}{\Gamma(1/2)\Gamma(j+1)}\exp\left\{-\frac{(4j+1)^2}{16S}\right\}\times$$
$$\times\left\{I_{-\frac{1}{4}}\left[\frac{(4j+1)^2}{16S}\right]-I_{\frac{1}{4}}\left[\frac{(4j+1)^2}{16S}\right]\right\},$$

where $I_{-\frac{1}{4}}(\cdot)$ and $I_{\frac{1}{4}}(\cdot)$ are modified Bessel functions,

$$I_\nu(z)=\sum_{k=0}^{\infty}\frac{(z/2)^{\nu+2k}}{\Gamma(k+1)\Gamma(k+\nu+1)},\quad |z|<\infty,\quad |\arg z|<\pi,$$

and the statistic (3) obeys the distribution [4] with the CDF

$$a2(S)=\frac{\sqrt{2\pi}}{S}\sum_{j=0}^{\infty}(-1)^j\frac{\Gamma(j+1/2)(4j+1)}{\Gamma(1/2)\Gamma(j+1)}\exp\left\{-\frac{(4j+1)^2\pi^2}{8S}\right\}\times$$
$$\times\int_0^{\infty}\exp\left\{\frac{S}{8(y^2+1)}-\frac{(4j+1)^2\pi^2y^2}{8S}\right\}dy.$$

II. Distributions of the Test Statistics in Testing Composite Hypotheses

In composite hypotheses testing, the conditional distribution law of a statistic $G(S|H_0)$ is affected by number of factors: a form of an observed law $F(x,\theta)$ that corresponds to the true hypothesis H_0; types and number of parameters to be estimated; a method of parameter estimation; sometimes, it is a specific value of a parameter (e.g., in case of gamma-distribution, inverse Gaussian law, generalized Weibull distribution, beta-distribution families).

The paper Kac *et al.* [13] was a pioneer in investigating distributions of statistics of the nonparametric goodness-of-fit tests for composite hypotheses. Then, various approaches to the solution to this problem where used [5-12, 31-32, 34-39].

In our research [14-25, 30] statistic distributions of the nonparametric goodness-of-fit tests are investigated by means of methods of statistical simulation, and approximate models are found for constructed empirical distributions. The most complete list of the constructed models of the statistic distributions and the tables of percentage points for the nonparametric goodness-of-fit tests is provided in [18, 20-21, 24-25, 29]. These models and tables are usable when testing composite hypotheses and maximum likelihood estimators are applied.

For a number of distributions often used in applications for description of random variates, distributions of statistics of the nonparametric goodness-of-fit tests only have a limited set of dependences: the form of the observed law $F(x,\theta)$ that corresponds to the true hypothesis H_0; types and number of parameters to be estimated; the method of estimation of parameter. In these cases, there are no impediments for studying test statistic distributions by means of statistical simulation and for further construction of approximate models for them when testing composite hypothesis [18, 22-23].

Complications arise in a case when the statistic distributions $G(S|H_0)$ of the nonparametric goodness-of-fit tests depend on a value of specific parameter (or values of several parameters) of the distribution $F(x,\theta)$ (for gamma distribution, two-sided exponential law, inverse Gaussian law, generalized Weibull distribution, and beta-distribution families).

The existing dependence on parameters values should not be neglected, e.g., in composite hypotheses testing subject to gamma-distribution with the density function

$$f(x,\theta)=\frac{x^{\theta_0-1}}{\theta_1^{\theta_0}\Gamma(\theta_0)}\exp\left(-\frac{x}{\theta_1}\right),$$

the limit distributions of the statistics of the nonparametric goodness-of-fit tests depend on the value of the form parameter θ_0. Fig. 1 illustrates the dependence of the distribution of Kolmogorov statistic upon the value of θ_0 when testing a composite hypothesis in the case of calculating MLE for the scale parameter only of gamma-distribution.

This dependence on the values of parameters of the observed law is the most serious impediment to a complete solution of the problem of testing composite hypotheses with the use of non-parametric goodness-of-fit tests. In [18-25] models of distributions of the statistics were obtained for a limited set of combinations of (integer) values of shape parameters (for gamma distribution, two-sided exponential law, inverse Gaussian law, generalized Weibull distribution, and beta-distribution families). It is unrealistic to build such models for an infinite set of combinations of real values of parameters.

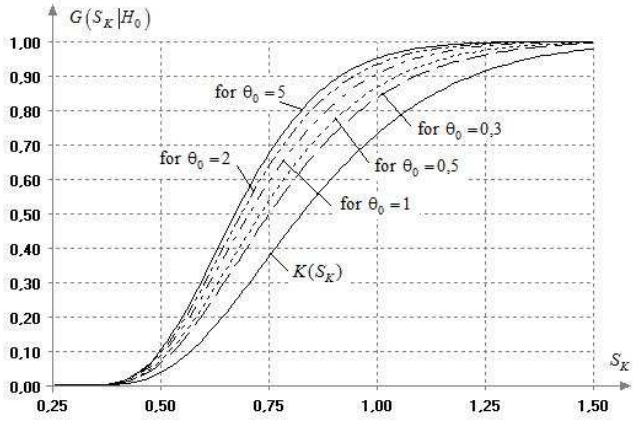


Figure 1. The Kolmogorov statistic (1) distributions for testing composite hypotheses with calculating MLE of scale parameter depend on the form parameter value of gamma-distribution

In present work, a "real-time" ability to simulate and research the distributions of tests statistics in the course of testing composite goodness-of-fit hypotheses is implemented by the use of parallel computing. It makes it possible to make correct statistical inferences even in those situations when the distribution of the test statistic is unknown before the testing procedure starts.

III. Testing Composite Hypotheses in "Real-Time"

In present work, an approach is proposed and implemented that is based upon authors' evolving software and the use of simulation [27, 29]. Computational processes in the simulation of statistics of various tests can be parallelized rather easily by the use of available resources of a nearby computer network. This makes it possible to dramatically reduce the time required for simulation (studying) an unknown distribution of the statistic $G(S|H_0)$. Statistical analysis is carried out by the

following scheme (Fig. 2) in case of use of nonparametric goodness-of-fit tests for testing composite hypotheses in regard to laws with the characteristic dependence of statistic distribution on values of parameters. Such an approach was used in [22, 26]. Here, the studying of $G(S|H_0)$ is carried out in "real-time" testing of the hypothesis [28].

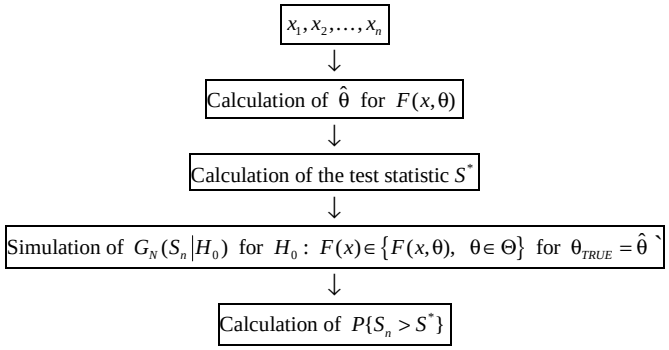


Figure 2. Testing the composite hypothesis H_0: $F(x)\in\{F(x,\theta),\ \theta\in\Theta\}$

When testing the composite hypothesis H_0: $F(x)\in\{F(x,\theta),\ \theta\in\Theta\}$ for an existing sample $x_1, x_2, \ldots, x_n$, the estimate $\hat{\theta}$ of the parameter vector for the law $F(x,\theta)$ is found in accordance with the selected method, e.g. MLE. Then, the value S^* of the statistic of the used goodness-of-fit test is calculated in accordance with the estimate $\hat{\theta}$ found. For making an inference on whether to reject or to accept H_0, it is necessary to know the distribution $G(S|H_0)$ of the test statistic that corresponds to the parameter value $\hat{\theta}$.

After that, statistical simulation procedure is started that results in obtaining an empirical conditional distribution $G_N(S_n|H_0)$ of the test statistic for the corresponding sample volume n and the given number of simulations N under $\theta=\hat{\theta}$. One can find an estimate of an achieved significance level (p-value) $P\{S_n > S^*\}$ or estimates of percentage points by the empirical distribution $G_N(S_n|H_0)$. The hypothesis is not rejected if $P\{S_n > S^*\} > \alpha$, where α is the given type I error probability.

The value of N defines the accuracy of the simulation of $G(S_n|H_0)$: the greater N the better. However, time spent for simulation increases along with growth of N, therefore, one can determine N during parallelization of simulation basing upon available computer resources (number of processors and cores) that could be used for the problem under solution.

The probability that elements of $\hat{\theta}$ are integer is zero. Thus, one should cautiously use the models and percentage points of test statistic distributions for values of parameters close to integer ones provided in [18-23] as, with interpolation applied, results obtained can be far from the true distribution $G(S|H_0)$ with the given $\hat{\theta}$.

Let us consider an example where a composite hypothesis is tested in regard to inverse Gaussian law (IGD) with the density function

$$f(x)=\left(\frac{\theta_1}{2\pi x^3}\right)^{1/2}\exp\left(-\frac{\theta_1(x-\theta_0)^2}{2\theta_0^2 x}\right).$$

In this case, distributions $G(S|H_0)$ of the nonparametric tests depend on specific values of θ_0 and θ_1 [22, 29]. This dependence is shown in Fig. 2 for distribution of Anderson-Darling statistic in case when MLE is applied to two parameters.

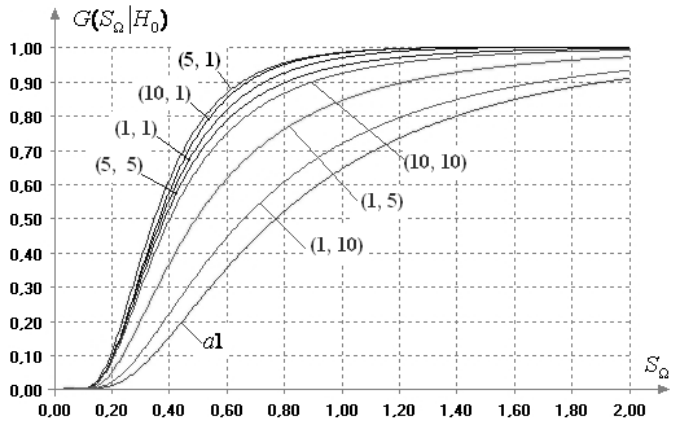


Figure 3. The dependence of Anderson-Darling statistic (3) on values of parameters (θ_0,θ_1) of IGD when MLE is used

The sample under analysis is following:

0.278	0.633	0.928	1.078	1.334	1.937	2.297	2.630	3.554	5.674
0.312	0.686	0.933	1.080	1.497	1.965	2.362	2.919	3.593	5.989
0.358	0.716	0.936	1.089	1.612	1.991	2.364	2.995	3.948	6.284
0.361	0.776	0.938	1.113	1.671	2.012	2.417	3.002	3.996	6.863
0.362	0.777	0.956	1.119	1.680	2.026	2.467	3.120	4.053	7.580
0.374	0.789	0.996	1.159	1.687	2.027	2.566	3.149	4.141	7.644
0.403	0.796	1.038	1.165	1.731	2.069	2.577	3.166	4.363	7.874
0.590	0.805	1.053	1.166	1.735	2.146	2.599	3.224	4.597	9.236
0.597	0.822	1.060	1.192	1.763	2.210	2.621	3.278	5.022	11.704
0.599	0.849	1.066	1.245	1.898	2.213	2.628	3.528	5.201	20.069

Maximum likelihood estimates of parameters calculated are $\hat{\theta}_0$ =2.4706 and $\hat{\theta}_1$ =2.5769. Values of the test statistics and the achieved significance levels obtained from simulated (in "real time") test statistic distributions under different values of N are given in Table I.

TABLE I. ACHIEVED SIGNIFICANCE LEVELS FOR IGD OBTAINED BY SIMULATION

Test	S^*	$P\{S_n > S^*\}$				
		$N=10^3$	$N=5\cdot10^3$	$N=10^4$	$N=10^5$	$N=10^6$
K	0.5936	0.638	0.653	0.654	0.655	0.656
CMS	0.0538	0.539	0.555	0.556	0.558	0.558
AD	0.3502	0.528	0.550	0.550	0.549	0.548

It should be noted, that distributions of the statistics (1) – (3) under $\hat{\theta}_0$ =2.4706 and $\hat{\theta}_1$ =2.5769 differ substantially from the corresponding distributions under different combinations of integer values of θ_0 and θ_1 [22, 29].

Let us see how well the sample can be described by the generalized Weibull distribution (GWD). The distribution function of GWD is

$$F(x;\theta_0,...,\theta_3)=1-\exp\left\{1-\left(1+\left((x-\theta_3)/\theta_2\right)^{\theta_0}\right)^{1/\theta_1}\right\}.$$

In the case of testing composite hypothesis, the distributions of the statistics (1) – (3) under true null hypothesis corresponding to GWD depend on value of shape parameter θ_1 (Fig. 4). Percentage points of the distributions for integer values of θ_1 are obtained in [23, 29]. In addition to them, percentage points and distribution models are given in Tables III-V for θ_1 =6, 7, 8.

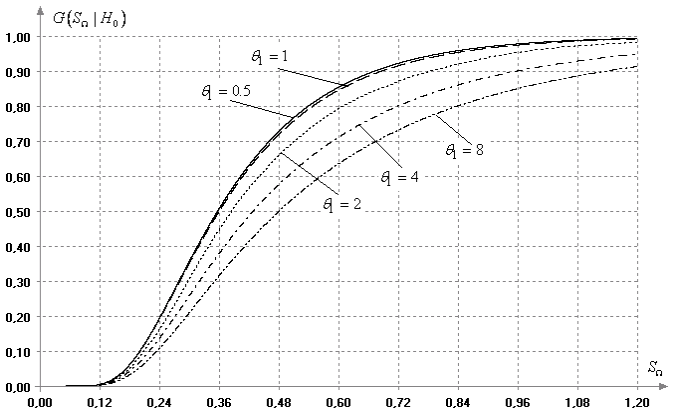


Figure 4. Distributions of the statistic (3) under true null hypothesis corresponding to GWD. MLE is applied to θ_0 and θ_1.

Distributions $G(S|H_0)$ of the statistics (1)-(3) are best approximated by the family of type III beta-distributions with the density function

$$B_3(x;\theta_0,...,\theta_4)=\frac{\theta_2^{\theta_0}}{\theta_3 \mathrm{B}(\theta_0,\theta_1)}\frac{\left(\frac{x-\theta_4}{\theta_3}\right)^{\theta_0-1}\left(1-\frac{x-\theta_4}{\theta_3}\right)^{\theta_1-1}}{\left[1+(\theta_2-1)\frac{x-\theta_4}{\theta_3}\right]^{\theta_0+\theta_1}},$$

or by the family of the *Sb*-Johnson distributions

$$Sb(x;\theta_0,...,\theta_3)=$$

$$=\frac{\theta_1\theta_2}{(x-\theta_3)(\theta_2+\theta_3-x)}\exp\left\{-\frac{1}{2}\left(\theta_0-\theta_1\ln\frac{x-\theta_3}{\theta_2+\theta_3-x}\right)^2\right\}.$$

In our case, MLE gives the closest fit to GWD with $\hat{\theta}_0$ =0.9542, $\hat{\theta}_1$ =0.1706, $\hat{\theta}_2$ =19.791, and $\hat{\theta}_3$ =0.2582. The obtained values of the statistics are given in the Table II with the simulated *p*-values under different numbers of simulations.

TABLE II. ACHIEVED SIGNIFICANCE LEVELS FOR GWD OBTAINED BY SIMULATION

Test	S^*	$P\{S_n > S^*\}$				
		$N=10^3$	$N=5\cdot10^3$	$N=10^4$	$N=10^5$	$N=10^6$
K	0.8724	0.286	0.279	0.276	0.274	0.273
CMS	0.1029	0.394	0.392	0.387	0.383	0.383
AD	2.4987	0.011	0.009	0.008	0.009	0.009

TABLE III. PERCENTAGE POINTS AND MODELS OF LIMIT DISTRIBUTIONS OF STATISTICS OF THE NONPARAMETRIC GOODNESS-OF-FIT TESTS WHEN MLE IS USED FOR PARAMETER ESTIMATION (θ_1 =6)

Parameter estimated	Percentage points			Model
	0.9	0.95	0.99	
for Kolmogorov test				
θ_0	1.110	1.230	1.471	*Sb*(2.3074;1.7536;2.3680;0.2609)
θ_1	1.084	1.199	1.427	B_3 (4.2825;5.6444;3 .1666;2.1430;0.3350)
θ_2	0.931	1.018	1.193	B_3 (6.9118;6.0758;3.3941;1.6000;0.2700)
θ_0 , θ_1	1.057	1.170	1.400	*Sb* (2.3818;1.7319;2.2919;0.2652)
θ_0 , θ_2	0.877	0.959	1.126	B_3 (4.4099;7.2683;2.3131;1.6000;0.3100)
θ_1 , θ_2	0.882	0.967	1.140	B_3 (5.2085;2.3642;4.3216;0.9100;0.3100)
θ_0 , θ_1 , θ_2	0.818	0.893	1.052	B_3 (4.8898;13.5936;2.0728;2.2000;0.2840)
for Cramer-von Mises-Smirnov test				
θ_0	0.251	0.331	0.530	*Sb*(3.4475;1.1730;1.7364;0.0066)
θ_1	0.227	0.296	0.466	B_3 (7.1391;2.5851;39.6158;1.2500;0.0000)
θ_2	0.148	0.184	0.270	B_3 (5.8171;4.3928;16.4865;0.8900;0.0000)
θ_0 , θ_1	0.207	0.271	0.430	*Sb* (3.8191;1.2084;1.7811;0.0073)
θ_0 , θ_2	0.116	0.143	0.208	B_3 (9.1508;3.9676;25.5001;0.6300;0.000)
θ_1 , θ_2	0.117	0.146	0.218	B_3 (11.2722;4.0296;63.6319;1.2000;0.000)
θ_0 , θ_1 , θ_2	0.093	0.114	0.167	B_3 (3.8686;3.8037;11.8789;0.4544;0.0092)
for Anderson-Darling test				
θ_0	1.404	1.812	2.837	B_3 (4.2257;2.6227;14.8469;5.0000;0.0800)
θ_1	1.279	1.625	2.478	B_3 (6.7753;2.9441;31.0263;6.8675;0.0535)
θ_2	0.940	1.150	1.652	B_3 (14.9774;4.0274;44.0425;5.5800;0.000)
θ_0 , θ_1	1.073	1.371	2.116	B_3 (5.9148;3.2311;37.2579;8.270;0.0569)
θ_0 , θ_2	0.671	0.806	1.129	B_3 (3.9744;4.9635;8.1400;3.1000;0.0800)
θ_1 , θ_2	0.679	0.823	1.173	B_3 (25.5888;4.1738;49.2114;3.0000;0.000)
θ_0 , θ_1 , θ_2	0.539	0.644	0.908	B_3 (6.6461;4.0515;13.9839;2.1420;0.0591)

IV. CONCLUSIONS

It required 4 hours 50 minutes to obtain the *p*-values in Table III under 10^6 simulations on a single core of Intel Core i7 processor and gave an error in only the third digit after the decimal period. In an analogous computation by means of a parallel algorithm in our software, performance increases linearly with more cores added. On this 8-threaded processor a simulation up to an error in the second digit takes about 20 seconds.

Computers with processors similar to that are not exclusively rare now and the further development of computations moves towards a growth of number of computing

units in a single device [40], thereby making a simulation of *p*-values transform from a long-lasting dedicated process into a routine procedure of statistical analysis with relatively small time cost. Such a computing performance eliminates the difficulties that arise when testing composite hypotheses with the use of nonparametric goodness-of-fit tests in cases when statistic distributions depend on specific values of parameters of observed distributions.

TABLE IV. PERCENTAGE POINTS AND MODELS OF LIMIT DISTRIBUTIONS OF STATISTICS OF THE NONPARAMETRIC GOODNESS-OF-FIT TESTS WHEN MLE IS USED FOR PARAMETER ESTIMATION (FOR θ_1 =8)

Parameter estimated	Percentage points			Model
	0.9	0.95	0.99	
for Kolmogorov test				
θ_0	1.100	1.218	1.454	B_3 (8.0781;4.8128;5.8094;2.0960;0.2735)
θ_1	1.084	1.199	1.428	*Sb*(2.4326;1.7778;2.3797;0.2673)
θ_2	0.978	1.074	1.266	B_3 (8.4485;5.1812;5.5890;1.8364;0.2700)
θ_0 , θ_1	1.072	1.186	1.417	B_3 (5.7833;6.1641;3.2903;2.1269;0.2699)
θ_0 , θ_2	0.911	0.999	1.179	*Sb*(2.6863;1.8734;2.0545;0.2559)
θ_1 , θ_2	0.929	1.012	1.198	*Sb*(2.6357;1.8244;2.0497;0.2612)
θ_0 , θ_1 , θ_2	0.863	0.948	1.117	B_3 (11.1281;6.1031;6.0962;1.7021;0.2200)
for Cramer-Von Mises-Smirnov test				
θ_0	0.242	0.319	0.507	B_3 (4.5895;2.5584;15.2153;0.8500;0.0000)
θ_1	0.228	0.296	0.467	B_3 (6.0112;2.5379;23.0339;0.8900;0.0000)
θ_2	0.166	0.209	0.314	B_3 (5.8877;4.0329;23.3907;1.2150;0.0000)
θ_0 , θ_1	0.217	0.284	0.450	*Sb*(3.4552;1.1997;1.4606;0.0061)
θ_0 , θ_2	0.128	0.160	0.239	*Sb*(4.6035;1.4434;1.3182;0.0060)
θ_1 , θ_2	0.131	0.165	0.252	*Sb*(4.4612;1.4003;1.3183;0.0059)
θ_0 , θ_1 , θ_2	0.107	0.134	0.201	B_3 (6.9845;2.7596;2.6920;0.4000;0.0060)
for Anderson-Darling test				
θ_0	1.363	1.752	2.722	B_3 (5.6824;4.0065;18.9636;8.4000;0.0000)
θ_1	1.279	1.624	2.477	*Sb*(3.4000;1.3163;7.4752;0.0535)
θ_2	1.005	1.240	1.799	B_3 (3.4843;5.3032;9.1592;6.2767;0.0800)
θ_0 , θ_1	1.139	1.457	2.243	B_3 (6.3736;2.8599;35.0312;6.7458;0.0538)
θ_0 , θ_2	0.713	0.864	1.227	B_3 (4.6820;5.7296;7.8880;3.4597;0.0523)
θ_1 , θ_2	0.722	0.881	1.275	B_3 (4.3613;6.0352;7.6499;3.8116;0.0520)
θ_0 , θ_1 , θ_2	0.589	0.714	1.009	B_3 (7.9147;4.0088;21.3294;2.9120;0.0500)

ACKNOWLEDGMENT

This research was partially supported by the Russian Foundation for Basic Research (Project no. 09-01-00056-a), the Analytical Departmental Targeted Program "Development of the Potential of Institutes of Higher Education" (Project No. 2.1.2/11855), and the Federal Targeted Program of the Ministry of Education and Science of the Russian Federation "Academic and Teaching Staff of an Innovative Russia".

TABLE V. PERCENTAGE POINTS AND MODELS OF LIMIT DISTRIBUTIONS OF STATISTICS OF THE NONPARAMETRIC GOODNESS-OF-FIT TESTS WHEN MLE IS USED FOR PARAMETER ESTIMATION (FOR θ_1 =7)

Parameter estimated	Percentage points			Model
	0.9	0.95	0.99	
for Kolmogorov test				
θ_0	1.104	1.223	1.461	B_3 (5.0453;5.6018;3.3300;2.1145;0.3100)
θ_1	1.084	1.199	1.427	B_3 (5.3655;6.0543;3.3092;2.1402;0.3000)
θ_2	0.955	1.047	1.231	B_3 (8.8643;20.9468;7.9001;9.1000;0.2300)
θ_0 , θ_1	1.066	1.180	1.409	*Sb*(2.4625;1.7390;2.3814;0.2668)
θ_0 , θ_2	0.895	0.980	1.153	B_3 (4.2520;7.5684;2.1829;1.6786;0.3100)
θ_1 , θ_2	0.902	0.991	1.169	B_3 (4.5096;5.6482;3.0218;1.6000;0.3100)
θ_0 , θ_1 , θ_2	0.839	0.918	1.079	B_3 (8.5291;6.5470;4.4062;1.6000;0.2400)
for Cramer-von Mises-Smirnov test				
θ_0	0.246	0.323	0.516	B_3 (7.5042;2.4317;48.3146;1.4000;0.0000)
θ_1	0.227	0.296	0.466	B_3 (6.2641;2.8729;33.7742;1.3750;0.0000)
θ_2	0.156	0.196	0.290	B_3 (4.1621;3.9072;14.0226;0.8986;0.0059)
θ_0 , θ_1	0.213	0.278	0.441	*Sb*(3.4488;1.2020;1.4196;0.0061)
θ_0 , θ_2	0.122	0.151	0.223	B_3 (7.9405;3.8743;23.4697;0.6700;0.0000)
θ_1 , θ_2	0.124	0.155	0.234	B_3 (7.5192;4.0675;25.1497;0.7945;0.0000)
θ_0 , θ_1 , θ_2	0.010	0.123	0.181	B_3 (5.5784;3.2913;17.0579;0.4290;0.0067)
for Anderson-Darling test				
θ_0	1.380	1.778	2.770	*Sb*(3.7593;1.3295;9.6362;0.0552)
θ_1	1.279	1.625	2.478	B_3 (4.8031;3.4732;17.6302;6.8675;0.0535)
θ_2	0.972	1.193	1.724	B_3 (3.4890;4.7102;10.2828;5.9597;0.0800)
θ_0 , θ_1	1.111	1.420	2.184	B_3 (5.4232;3.1894;26.3229;6.7000;0.0539)
θ_0 , θ_2	0.692	0.835	1.178	B_3 (11.9769;4.7144;19.2233;3.0000;0.000)
θ_1 , θ_2	0.701	0.852	1.225	B_3 (22.3537;4.1744;51.2639;3.6000;0.000)
θ_0 , θ_1 , θ_2	0.563	0.677	0.955	B_3 (8.0353;3.9949;18.1724;2.4000;0.0500)

REFERENCES

[1] T.W. Anderson and D.A. Darling, "Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes," Ann. Math. Statist., vol. 23, pp. 193-212, 1952.

[2] T.W. Anderson and D.A. Darling, "A test of goodness of fit," J. Amer. Statist. Assoc., vol. 29, pp. 765-769, 1954.

[3] L.N. Bolshev, "On the question on testing some composite statistical hypotheses," in Theory of Probability and Mathematical Statistics. Selected Works, Moscow: Nauka, 1987, pp. 5-63.

[4] L.N. Bolshev and N.V. Smirnov, Tables of Mathematical Statistics. Moscow: Nauka, 1983. (in Russian)

[5] M. Chandra, N.D. Singpurwalla, and M.A. Stephens, "Kolmogorov statistics for tests of fit for the extreme–value and Weibull distribution," J. Am. Statist. Assoc., vol. 76, no. 375, pp. 729-731, 1981.

[6] D.A. Darling, "The Cramer-Smirnov test in the parametric case," Ann. Math. Statist., vol. 26, pp. 1-20, 1955.

[7] D.A. Darling, "The Cramer-Smirnov test in the parametric case," Ann. Math. Statist., vol. 28, pp. 823-838, 1957.

[8] J. Durbin, "Weak convergence of the sample distribution function when parameters are estimated," Ann. Statist., vol. 1, pp. 279-290, 1973.

[9] J. Durbin, "Kolmogorov-Smirnov tests when parameters are estimated with applications to tests of exponentiality and tests of spacings," Biometrika, vol. 62, pp. 5-22, 1975.

[10] J. Durbin, “Kolmogorov–Smirnov Test when Parameters are Estimated,” Lect. Notes Math., vol. 566, pp. 33–44, 1976.

[11] K.O. Dzhaparidze and M.S. Nikulin, “Probability distribution of the Kolmogorov and omega-square statistics for continuous distributions with shift and scale parameters,” J. Soviet Math., vol. 20, pp. 2147-2163, 1982.

[12] I.I. Gihman, “Some remarks on the consistency criterion of A.N. Kolmogorov,” Dokl. Akad. Nauk SSSR, vol. 91(4), pp. 715-718, 1953.

[13] M. Kac, J. Kiefer, and J. Wolfowitz, “On tests of normality and other tests of goodness of fit based on distance methods,” Ann. Math. Stat., vol. 26, pp. 189–211, 1955.

[14] B.Yu. Lemeshko and S.N. Postovalov, “Statistical distributions of nonparametric goodness-of-fit tests as estimated by the sample parameters of experimentally observed laws,” Industrial laboratory (Ind. lab.), vol. 64, no. 3, pp. 197-208, 1998. (Consultants Bureau, New York)

[15] B.Yu. Lemeshko and S.N. Postovalov, “Application of the nonparametric goodness-of-fit tests in testing composite hypotheses,” Optoelectronics, Instrumentation and Data Processing, vol. 37, no. 2, pp. 76-88, 2001.

[16] B.Yu. Lemeshko and S.N. Postovalov, “The nonparametric goodness-of-fit tests about fit with Johnson distributions in testing composite hypotheses,” News of the SB AS HS, no. 1(5), pp. 65-74, 2002. (in Russian)

[17] B.Yu. Lemeshko and A.A. Maklakov, “Nonparametric test in testing composite hypotheses on goodness of fit exponential family distributions,” Optoelectronics, Instrumentation and Data Processing, vol. 40, no. 3, pp. 3-18, 2004.

[18] B.Yu. Lemeshko, S.B. Lemeshko, and S.N. Postovalov. “Statistic distribution models for some nonparametric goodness-of-fit tests in testing composite hypotheses,” Communications in Statistics - Theory and Methods, vol. 39, no. 3, pp. 460-471, 2010.

[19] B.Yu. Lemeshko and S.B. Lemeshko, “Statistic distributions of the nonparametric goodness-of-fit tests in testing hypotheses relative to beta-distributions,” News of the SB AS HS, no. 2(9), pp. 6-16, 2007. (in Russian)

[20] B.Yu. Lemeshko and S.B. Lemeshko, “Distribution models for nonparametric tests for fit in verifying complicated hypotheses and maximum-likelihood estimators. Part I,” Measurement Techniques, vol. 52, no. 6, pp. 555-565, 2009.

[21] B.Yu. Lemeshko and S.B. Lemeshko, “Models for statistical distributions in nonparametric fitting tests on composite hypotheses based on maximum-likelihood estimators. Part II,” Measurement Techniques, vol. 52, no. 8, pp. 799-812, 2009.

[22] B.Yu. Lemeshko, S.B. Lemeshko, M.S. Nikulin, and N. Saaidia, “Modeling statistic distributions for nonparametric goodness-of-fit criteria for testing complex hypotheses with respect to the inverse Gaussian law,” Automation and Remote Control, vol. 71, no. 7, pp. 1358-1373, 2010.

[23] B.Yu. Lemeshko, S.B. Lemeshko, and K.A. Akushkina, “Models of statistical distributions of nonparametric goodness-of-fit tests in testing composite hypotheses of the generalized Weibull distribution,” Proceedings of the Third International Conference on Accelerated Life Testing, Reliability-based Analysis and Design. Clermont-Ferrand, France, 2010, pp. 125-132.

[24] B.Yu. Lemeshko and S.B. Lemeshko, “Models of statistic distributions of nonparametric goodness-of-fit tests in composite hypotheses testing for double exponential law cases,” Communications in Statistics - Theory and Methods, vol. 40, no. 16, pp. 2879-2892, 2011.

[25] B.Yu. Lemeshko and S.B. Lemeshko, “Construction of statistic distribution models for nonparametric goodness-of-fit tests in testing composite hypotheses: the computer approach,” Quality Technology & Quantitative Management, vol. 8,no. 4, pp. 359-373, 2011.

[26] B.Yu. Lemeshko, S.B. Lemeshko, K.A. Akushkina, M.S. Nikulin, and N. Saaidia, “Inverse Gaussian model and its applications in reliability and survival analysis,” in Mathematical and Statistical Models and Methods in Reliability. Applications to Medicine, Finance, and Quality Control, V. Rykov, N. Balakrishnan, and M. Nikulin, Eds. Boston: Birkhäuser, 2011, pp. 433-453.

[27] B.Yu. Lemeshko, S.B. Lemeshko, E.V. Chimitova, S.N. Postovalov, and A.P. Rogozhnikov, “Software system for simulation and research of probabilistic regularities and statistical data analysis in reliability and quality control,” in Mathematical and Statistical Models and Methods in Reliability. Applications to Medicine, Finance, and Quality Control, V. Rykov, N. Balakrishnan, and M. Nikulin, Eds. Boston: Birkhäuser, 2011, pp. 417-432.

[28] B.Yu. Lemeshko, S.B. Lemeshko, and A.P. Rogozhnikov, “Real-time studying of statistic distributions of non-parametric goodness-of-fit tests when testing complex hypotheses,” in Proceedings of the International Workshop “Applied Methods of Statistical Analysis. Simulations and Statistical Inference” – AMSA’2011. Novosibirsk, Russia, 20-22 September, 2011, pp. 19-27.

[29] B.Yu. Lemeshko, S.B. Lemeshko, S.N. Postovalov, and E.V. Chimitova, Statistical Data Analysis, Simulation and Study of Probability Regularities. Computer Approach : monograph. Novosibirsk : NSTU Publishing House, 2011. (“NSTU Monographs” series, in Russian)

[30] S.B. Lemeshko, Expansion of applied opportunities of some classical methods of mathematical statistics. The dissertation on competition of a scientific degree of Cand. Tech. Sci. Novosibirsk State Technical University, 2007. (in Russian)

[31] G.V. Martynov, Omega-Square Tests. Moscow: Science, 1978. (in Russian)

[32] G. Martynov, “Weighted Cramer-von Mises test with estimated parameters,” Communications in Statistics – Theory and Methods, vol. 40, no. 19-20, pp. 3569-3586, 2011.

[33] M.S. Nikulin, “Gihman and goodness-of-fit tests for grouped data,” Mathematical Reports of the academy of Science of the Royal Society of Canada, vol. 14, no. 4, pp. 151-156, 1992.

[34] M.S. Nikulin, “A variant of the generalized omega-square statistic,” J. Soviet Math., vol. 61, no. 4, pp. 1896-1900, 1992.

[35] E.S. Pearson and H.O. Hartley, Biometrica Tables for Statistics, vol. 2. Cambridge: University Press, 1972.

[36] M.A. Stephens. “Use of Kolmogorov–Smirnov, Cramer – von Mises and related statistics – without extensive tables,” J. R. Stat. Soc., vol. 32, pp. 115-122, 1970.

[37] M.A. Stephens, “EDF statistics for goodness of fit and some comparisons,” J. Am. Statist. Assoc., vol. 69, pp. 730-737, 1974.

[38] Yu.N. Tyurin, “On the limiting Kolmogorov-Smirnov statistic distribution for composite hypothesis,” News of the AS USSR. Ser. Math., vol. 48, no. 6, pp. 1314-1343, 1984. (in Russian)

[39] Yu.N. Tyurin and N.E. Savvushkina, “Goodness-of-fit test for Weibull-Gnedenko distribution,” News of the AS USSR. Ser. Techn. Cybernetics, vol. 3, pp. 109-112, 1984. (in Russian)

[40] H. Sutter, “Welcome to the Jungle,” http://herbsutter.com/welcome-to-the-jungle/, 2011. (Internet publication)

Estimating hospital expected costs with censored data and outliers

Isabella Locatelli
Institute of social and preventive medicine (IUMSP),
Lausanne University Hospital
Route de la Corniche 10
1010 Lausanne, Switzerland
Email: Isabella.Locatelli@chuv.ch

Alfio Marazzi
Institute of social and preventive medicine (IUMSP),
Lausanne University Hospital
Route de la Corniche 10
1010 Lausanne, Switzerland
Email: alfio.marazzi@chuv.ch

***Abstract*—A parametric robust estimate of the expected cost of a group of hospital stays is proposed. This estimate combines a robust estimate of the length of stay (LOS) distribution with a robust estimate of the conditional expected cost for a given LOS. It is a robust version of Lin's indirect estimates; in addition it takes covariate information into account.**

I. Introduction

We consider the problem of estimating the expected cost of a hospital stay as a function of patient characteristics, such as diagnosis, treatment, and other covariates. This is a basic task in hospital management. Estimation of a cost mean is complicated by different features of cost data. Cost distributions are skewed and stays may be censored because a patient may die or be transferred to a different hospital before the regular home discharge. In addition, the data may contain "outliers".

In order to treat censoring, some attempts to apply usual survival analysis techniques to censored costs have been made ([1], [2]). Unfortunately, this approach produces biased estimates, because of the intrinsic dependence between cost and censoring on the cost scale (Section 2). To overcome this problem, "indirect procedures"have been proposed ([3], [4], [6]) combining an estimate of the LOS distribution with a separate estimate of the expected cost for given LOS (Section 3). Indirect procedures give rise to consistent mean cost estimates, but they still have important limitations which are discussed in the following. We propose to encompass such limitations with a parametric version of the indirect approach (Section 4).

II. Framework and assumptions

Suppose that T is a random variable representing the complete (noncensored) LOS of a patient. The observed LOS is $T^* = \min(T, C)$, with C being an unobserved random censoring time. Let Y be the censored total cost accumulated until the LOS is completed. The observed cost is $Y^* = \min(Y, K)$, where K is the cost accumulated until the censoring time C. Finally, let Δ be the censoring indicator: $\Delta = 1$ if the observation is complete ($T < C$) and 0 otherwise.

We suppose that a sample (t_i^*, y_i^*, δ_i), $i = 1, ..., n$ of observations of (T^*, Y^*, Δ) is available. In addition, we suppose that a vector $\mathbf{x}_i$ of covariate values is observed for each i. We wish to estimate the expected conditional cost $\mu_{\mathbf{x}} = E(Y \mid \mathbf{x})$.

As usual, we assume that censoring is noninformative on the LOS scale, i.e. T is independent of C (A1). Moreover, we assume independence of C and Y (A2: "extended noninformative censoring"). One can show that, even under A1-A2, the censoring mechanism is informative on the cost scale, i.e. the total cost Y and the censoring cost K are not independent. For instance, defining a unit cost variable $U = Y/T$, the covariance of complete cost and censoring cost takes the form:

$$\mathrm{Cov}(Y, K) = E(C)\,\mathrm{Cov}(Y, U). \tag{1}$$

We note that $\mathrm{Cov}(Y, U)$ equals zero only when either the total or the unit cost is a constant. In general, Y and K are correlated and, therefore, direct application of usual survival analysis techniques to censored cost will lead to biased estimates.

III. Non parametric indirect estimates

Suppose that the time range $[0, \tau)$ is splitted into m little intervals $[\tau_j, \tau_{j+1})$, $j = 1, ..., m$, with $\tau_1 = 0$ and $\tau_{m+1} = \tau$. We set $\tau_{m+2} = \infty$. Then, under A1-A2, we have:

$$\mu = \sum_{j=1}^{m+1} [S(\tau_j) - S(\tau_{j+1})]\,\eta(\tau_j, \tau_{j,j+1}) \tag{2}$$

where $\eta(\tau_j, \tau_{j,j+1}) = E(Y \mid \tau_j \le T^* < \tau_{j+1},\ \Delta = 1)$ is the expected cost for subjects who experience the event (e.g. home-discharge) in the time interval $[\tau_j, \tau_{j+1})$ and $S(t) = P(T > t)$ is the survival function of T. On the ground of (2), Lin et al. (1997) [3] proposed the following "indirect estimate" of μ:

$$\bar{\mu} = \sum_{j=1}^{m+1} \left[\bar{S}(\tau_j) - \bar{S}(\tau_{j+1})\right] \bar{y}_j^*, \tag{3}$$

where $\bar{S}(t)$ is the Kaplan-Meier (KM) estimate of $S(t)$ and

$$\bar{y}_j^* = \frac{\sum_{i=1}^n y_i^* I(\tau_j \le t_i^* < \tau_{j+1}, \delta_i = 1)}{\sum_{i=1}^n I(\tau_j \le t_i^* < \tau_{j+1}, \delta_i = 1)}, \quad j = 1, ...m+1 \tag{4}$$

is a natural estimate of $\eta(\tau_j, \tau_{j,j+1})$.

A nice feature of $\bar{\mu}$ is that it is nonparametric. However, it does not include the covariate information provided by $\mathbf{x}_i$. Moreover, it is well known that both the arithmetic means $\bar{y}_j^*$ and the Kaplan-Meier estimate $\bar{S}(t)$ can be heavily affected by outliers (e.g., Locatelli et al., 2011) [5].

IV. Parametric indirect estimates

The procedure proposed here uses a parametric accelerated failure time (AFT) regression model to estimate $S(t)$ and a parametric log-linear regression model to estimate the conditional expectation in (2). More precisely, we assume that the relationship between log(LOS) and the covariate vector $\mathbf{x}_i$ can be described by the model

$$\log t_i = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \sigma e_i, \quad i = 1, ..., n, \tag{5}$$

where $\boldsymbol{\beta}$ is a vector of unknown coefficients, σ is an unknown scale parameter, and the errors e_i are independently distributed according to a given model distribution. We also assume that the observed total costs depend upon durations, censoring, and covariates according to the log-linear model

$$\log y_i^* = \mathbf{x}_i^{\mathrm{T}}\mathbf{b} + c\log t_i^* + d\delta_i + s\varepsilon_i, \quad i = 1, ..., n, \tag{6}$$

where $\mathbf{b}$, c, d and s are parameters to be estimated and ε_i is an error term, which follows another given model distribution.

Suppose that $\hat{\boldsymbol{\beta}}$, and $\hat{\sigma}$ are the estimates of $\boldsymbol{\beta}$ and σ based on (5), that $\hat{S}(\cdot; \mathbf{x}, \hat{\boldsymbol{\beta}}, \hat{\sigma})$ is the corresponding estimate of the LOS survival function, that $\hat{\mathbf{b}}$, $\hat{c}$, $\hat{d}$ and $\hat{s}$ are the estimates of the parameters in (6), and that $\hat{\eta}(\tau_{j,j+1}; \mathbf{x}, \hat{\mathbf{b}}, \hat{c}, \hat{d}, \hat{s})$ is the corresponding estimate of $E(Y^* |\ \tau_j < T^* < \tau_{j+1},\ \Delta = 1,\ \mathbf{x})$. Then, the parametric version of (3) is given by

$$\hat{\mu}_{\mathbf{x}} = \sum_{j=1}^{m+1} [\hat{S}(\tau_j; \mathbf{x}, \hat{\boldsymbol{\beta}}, \hat{\sigma}) - \hat{S}(\tau_{j+1}; \mathbf{x}, \hat{\boldsymbol{\beta}}, \hat{\sigma})]\hat{\eta}(\tau_{j,j+1}; \mathbf{x}, \hat{\mathbf{b}}, \hat{c}, \hat{d}, \hat{s}) \tag{7}$$

Explicit forms of $\hat{\mu}_{\mathbf{x}}$ can be obtained for particular distributions of the error terms e_i and ε_i. For example, if both e_i and ε_i follow standard normal distributions, we have

$$\hat{\mu}_{\mathbf{x}} = \exp[\mathbf{x}^{\mathrm{T}}(\hat{\mathbf{b}} + \hat{c}\hat{\boldsymbol{\beta}}) + \hat{d} + (\hat{c}^2\hat{\sigma}^2 + \hat{s}^2)/2]. \tag{8}$$

In order to bound the sensitivity of the estimate $\hat{\mu}_{\mathbf{x}}$ with respect to outliers we propose to use the robust estimates of $\boldsymbol{\beta}$, σ described in Locatelli et al. (2011) [5] to compute $\hat{S}(t; \mathbf{x}, \hat{\boldsymbol{\beta}}, \hat{\sigma})$, and the robust regression estimates of Marazzi and Yohai (2004) [7] to compute $\hat{\eta}(t; \mathbf{x}, \hat{\mathbf{b}}, \hat{c}, \hat{d}, \hat{s})$. Both these procedures can be applied with symmetric or asymmetric location-scale error distributions and are asymptotically consistent and fully efficient when the data are generated according to the models. Formulae to estimate their variances and covariances can be derived with the help of the influence functions.

TABLE I
1000 samples of size $n = 100$. Root of the 10% tmse of the estimates of $\mu_{0.3}$

	Nominal model		5% Outliers on LOS		5% Outliers on costs	
censoring	ML	Robust	ML	Robust	ML	Robust
34%	23.6	23.8	33.5	22.5	>1000	23.9
11%	18.4	19.0	19.2	18.7	>1000	19.2

V. Simulations

An extensive simulation study has been performed in order to compare the robust estimate of the expected cost proposed here with the one obtained with the classical ML approach for models (5) and (6). Here, we provide some of the results.

In a first experiment, the censored times were generated according to the model

$$\log t_i = \beta_1 + \beta_2 x_i + \sigma e_i, \quad e_i \sim N(0,1), \tag{9}$$

$$\log x_i \sim N(0,1), \quad t_i^* = \min(t_i, c_i), \quad \log c_i \sim N(\nu, 1), \tag{10}$$

for $i = 1, ..., n$ and total costs according to

$$\log y_i^* = b_1 + b_2 x_i + c\log t_i^* + d\delta_i + s\varepsilon_i; \quad \varepsilon_i \sim N(0,1). \tag{11}$$

We choosed $\beta_1 = 0$, $\beta_2 = 2$, $\sigma = 1$ in (9) and $b_1 = 0$, $b_2 = 1$, $c = 2$, $d = 0$, and $s = 1$ in (11). Thus, the true conditional expected cost was

$$\mu_x = \exp[(b_2 + c\beta_2)x + (c^2\sigma^2 + s^2)/2]. \tag{12}$$

Two values of parameter ν were used: $\nu = 1$, corresponding to 34% of censored data and $\nu = 3$ with 11% of censoring. We simulated 1000 samples of size $n = 100$. In the first two columns of Table 1 we provide the root of the 10% trimmed mean squared error (tmse) of the estimates of μ_x for $x = 0.3$ ($\mu_{0.3} = 54.60$, according to (12)) as a global measure of performace. As expected, ML provides the best performing estimates. However, the robust method shows very low efficiency losses of about 3%.

In a second experiment we considered 5% of outliers in the durations: 5 of the log durations $\log t_i$ were replaced by $\log t_i + 15$ (columns 3 and 4). Finally we introduced 5% of outliers in the costs: 5 of the log-costs $\log y_i$ were replaced by $\log y_i + 15$ (columns 5 and 6). One can see that the robust method provides smaller tmse than ML, especially when outliers are in the costs.

References

[1] Quesenberry CP, Fireman B, Hiatt RA, Selby JV. A survival analysis of hospitalization among patients with acquired immunodefincency syndrome. *American Journal of Public Health* 1989; **79**: 1643-1647.

[2] MaWhinney S, Brown ER, Malcolm J, Villanueva C, Groves BM, Quaife RA, Lindenfeld J, Warner BA, Hammermeister KE, Grover FL, Shroyer ALW. Identification of risk factors for increased cost, charges and length of stay for cardiac patients. *Annals of Thoracic Surgery* 2000; **70**: 702-710.

[3] Lin DY, Feuer EJ, Etzioni R, Wax Y. Estimating Medical Costs from Incomplete Follow-Up Data. *Biometrics* 1997; **53**: 419-434.

[4] Hallstrom AP, Sullivan SD. On Estimating Costs for Economic Evaluation in Failure Time Studies. *Medical Care* 1998; **36**: 433-436.
[5] Locatelli I., Marazzi A., Yohai V.J. (2011). Robust Accelerated Failure Time Regression. Computational Statistics & Data Analysis, 55, 874-887.
[6] Etzioni RD, Feuer EJ, Sullivan SD, Lin DY, Hu C, Ramsey SD. On the use of survival analysis techniques to estimate medical care costs. *Journal of Health Economics* 1999; **18**: 365-380.
[7] Marazzi A, Yohai VJ. Adaptively truncated maximum likelihood regression with asymmetric errors. *Journal of Statistical Planning and Inference* 2004; **122**: 271-291.

Robust estimation of the accelerated failure time model

Alfio Marazzi
Institute of social and preventive medicine (IUMSP),
Lausanne University Hospital
Route de la Corniche 10
1010 Lausanne, Switzerland
Email: alfio.marazzi@chuv.ch

Isabella Locatelli
Institute of social and preventive medicine (IUMSP),
Lausanne University Hospital
Route de la Corniche 10
1010 Lausanne, Switzerland
Email: Isabella.Locatelli@chuv.ch

***Abstract*—Parametric robust estimates of the accelerated failure time model are described. The estimates are based on two steps. In the first step a new S estimate for censored data is computed. In the second step a weighted likelihood estimate is computed, where observations that are unlikely under the model estimated in the first step are are rejected. The final estimate is highly robust in presence of contaminated data and fully efficient at the model.**

I. Introduction

Positive random variables with asymmetric distributions arise in many applications (e.g., analysis of income and expenditures, failure times, output of biological systems). Often the population mean is the parameter of interest and depends upon a number of covariates. The data may contain censored observations as well as outliers; these features make the mean a difficult parameter to estimate.

Parametric robust estimation provides powerful tools to detect outliers and compute stable and efficient inferences for regression problems with positive responses. This includes the estimation of the expected response. In particular, the use of high breakdown point (bdp) methods [6] for regression problems with positive noncensored responses has been considered in [7]. The proposed estimates, called truncated maximum likelihood (TML) estimates, are based on two steps. In the first step, an initial high breakdown point estimate is computed. In the second step, observations that are unlikely under the estimated model are identified and a weighted maximum likelihood estimate is computed, where outliers are rejected. The rejection rule is based on an adaptive cut-off that, asymptotically, does not reject any observation when the data are generated according to the model. Therefore, the final estimate attains full efficiency (at the model), while maintaining the breakdown point of the initial estimator.

The TML has been extended to the case of censored observations in [4]. In the following sections, we summarize the basic steps of the TML estimate for censored observations, review its properties, and discuss current developments. An application of the method to the estimation of the expected hospital length and cost of stay is presence of censoring is discussed in the companion paper [5].

II. The model

We consider an accelerated failure time model for n pairs of variates $(\mathbf{x}_i, y_i)$

$$y_i = \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x}_i + \sigma_0 u_i, \quad i = 1, \ldots, n, \tag{1}$$

where y_i represents the duration on the logarithmic scale. The errors u_i are i.i.d. with cdf F and independent of $\mathbf{x}_i$; $\boldsymbol{\beta}_0 \in \mathbb{R}^p$ is an unknown vector of coefficients, the first component being an intercept term, and σ_0 an unknown scale parameter. The distribution of the covariate vectors $\mathbf{x}_i$ is unknown. Let $v_1, \ldots, v_n$ be i.i.d. censoring times, which are independent of the y_i's and $y_i^* = \min(y_i, v_i)$ be the censored responses. We define $\delta_i = 1$ if $y_i^* = y_i$ and $\delta_i = 0$ if $y_i^* = v_i$.

We use a hypothetical model cdf F_0 as an approximation of the real unknown error distribution F. We assume that $F_0(z) = F_{0,1}(z)$, where $F_{0,1}$ is the standard member of a parametric location-scale family of asymmetric (or symmetric) distributions with cdf $F_{\mu,\sigma}(z) = F_{0,1}((z-\mu)/\sigma)$. We denote by $f_{\mu,\sigma}$ and f_0 the densities of $F_{\mu,\sigma}$ and F_0 and by $H_{\boldsymbol{\beta},\sigma}$ the corresponding cdf of $(y, \mathbf{x})$ when $(\boldsymbol{\beta},\sigma)$ are the true parameters. Examples are the Gaussian model with mean μ and variance σ^2 and the Log-Weibull model $f_{\mu,\sigma}(z) = \exp[(z-\mu)/\sigma - \exp((z-\mu)/\sigma)]/\sigma$. We define $\rho_{\mu,\sigma}(u) = -\ln f_{\mu,\sigma}(u)$, put $\rho_0(u) = \rho_{0,1}(u)$, and assume that this function is convex. Finally, let $\psi_0(u) = \rho_0'(u) = -f_0'(u)/f_0(u)$ and $\psi_1(u) = u\psi_0(u)$. The triplets $(y_i^*, \mathbf{x}_i, \delta_i)$ are observed and we want to estimate $(\boldsymbol{\beta}_0, \sigma_0)$.

Define an *empirical cdf for censored observations* $(y_i^*, \mathbf{x}_i)$ as

$$H_{n,\boldsymbol{\beta},\sigma}(z, \mathbf{z}) = \frac{1}{n}\sum_{i=1}^{n} E_{\boldsymbol{\beta},\sigma}\left[I\left(y \le z\right) \middle| y_i^*, \mathbf{x}_i, \delta_i\right] I\left(\mathbf{x}_i \le \mathbf{z}\right) \tag{2}$$

where, for any measurable function $h(y, \mathbf{x})$,

$$E_{\boldsymbol{\beta},\sigma}[h(y, \mathbf{x}) \,|\, y > y_i^*, \mathbf{x}_i] = \int_{(y_i^* - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})/\sigma}^{\infty} h(\sigma u + \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}, \mathbf{x}_i) f_0(u)du \;/\; (1 - F_0((y_i^* - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})/\sigma)). \tag{3}$$

When there is no censoring, $H_{n,\boldsymbol{\beta},\sigma}(z, \mathbf{z})$ coincides with the usual empirical cdf H_n. In addition, $H_{n,\boldsymbol{\beta}_0,\sigma_0}(z, \mathbf{z})$ is

a consistent estimate of $H_{\boldsymbol{\beta}_0,\sigma_0}(z,\mathbf{z})$. We note that the ML equations of the estimates of $\boldsymbol{\beta}_0$ and σ_0 can be written as follows,

$$E_{n,\boldsymbol{\beta},\sigma}\left[\psi_0\left(\left(y-\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}\right)/\sigma\right)\mathbf{x}\right]=0, \quad (4)$$

$$E_{n,\boldsymbol{\beta},\sigma}\left[\psi_1\left(\left(y-\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}\right)/\sigma\right)\right]=1. \quad (5)$$

III. The estimates

The initial estimate. The initial step is the computation of a high bdp S estimate. S estimates were introduced in [11] for noncensored data. We extend the S estimates to the case of censored observations.

Suppose that ρ is a given function $\mathbb{R}\to\mathbb{R}^+$ such that (i) $\rho(0)=0$; (ii) ρ is even; (iii) if $|z_1|<|z_2|$, then $\rho(z_1)\le\rho(z_2)$; (iv) ρ is bounded; (v) ρ is continuous at 0. For example, ρ is a member of the Tukey's biweight family

$$\rho^T(z,k)=\begin{cases}3v^2-3v^4+v^6 & \text{if} \quad |z|\le k,\\ 1 & \text{if} \quad |z|>k,\end{cases} \quad (6)$$

where $v=z/k$ and k is a user chosen tuning parameter. For any μ, let the function $S(\mu)$ be the M-scale [3] defined by $E_0\left[\rho\left((u-\mu)/S(\mu)\right)\right]=b$, where expectation is based on F_0 and $b=\max\rho(z)/2$. Thus, $b=0.5$ for Tukey's biweight. Under certain conditions [10], there exists a unique μ_0 such that $\mu_0=\arg\min_\mu S(\mu)$. Then, we define the *S-scale* of F_0 as $s_0=S(\mu_0)$ and, without loss of generality, assume that $s_0=1$ and $\mu_0=0$. For any $\boldsymbol{\beta}\in\mathbb{R}^p$, let the residual scale $s(\boldsymbol{\beta})$ be defined by

$$E_0\left[\rho\left(\left(y-\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}\right)/s(\boldsymbol{\beta})\right)\right]=b. \quad (7)$$

It is shown in [8] (Lemma 2), that $\sigma_0=\min_{\boldsymbol{\beta}} s(\boldsymbol{\beta})$ and $\boldsymbol{\beta}_0=\arg\min_{\boldsymbol{\beta}} s(\boldsymbol{\beta})$. Then, for the noncensored case, the S estimate $(\tilde{\boldsymbol{\beta}}_n,\tilde{\sigma}_n)$ is defined by $\tilde{\boldsymbol{\beta}}_n=\arg\min s_n(\boldsymbol{\beta})$, and $\tilde{\sigma}_n=s_n(\tilde{\boldsymbol{\beta}}_n)$, where $s_n(\boldsymbol{\beta})$ solves

$$E_{H_n}\left[\rho\left(\left(y-\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}\right)/s_n(\boldsymbol{\beta})\right)\right]=b. \quad (8)$$

Since H_n is a consistent estimate of $H_{\boldsymbol{\beta}_0,\sigma_0}$, $(\tilde{\boldsymbol{\beta}}_n,\tilde{\sigma}_n)$ is consistent for $(\boldsymbol{\beta}_0,\sigma_0)$.

In the censored case, H_n is not available. However, if the scale $s_n(\boldsymbol{\beta})$ is defined by

$$E_{n,\boldsymbol{\beta},s_n(\boldsymbol{\beta})}\left[\rho\left(\left(y-\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}\right)/s_n(\boldsymbol{\beta})\right)\right]=b, \quad (9)$$

one can show that $H_{n,\boldsymbol{\beta}_0,s_n(\boldsymbol{\beta}_0)}$ is a consistent estimate of $H_{\boldsymbol{\beta}_0,\sigma_0}$. Therefore, if we replace H_n by $H_{n,\boldsymbol{\beta}_0,s_n(\boldsymbol{\beta}_0)}$ in (8), the estimate $\tilde{\boldsymbol{\beta}}_n$ will converge to $\boldsymbol{\beta}_0$ for $n\to\infty$. This motivates the following definition of S estimate for censored observations.

1. For any $\boldsymbol{\beta}\in\mathbb{R}^p$, let $s_n(\boldsymbol{\beta})$ be a solution of (9).
2. For any $\boldsymbol{\beta}\in\mathbb{R}^p$ and $\boldsymbol{\gamma}\in\mathbb{R}^p$, let $S_n(\boldsymbol{\beta},\boldsymbol{\gamma})$ be given by $E_{n,\boldsymbol{\beta},s_n(\boldsymbol{\beta})}\left[\rho\left(\left(y-\mathbf{x}^{\mathrm{T}}\boldsymbol{\gamma}\right)/S_n(\boldsymbol{\beta},\boldsymbol{\gamma})\right)\right]=b$.
3. Let $\tilde{\boldsymbol{\gamma}}_n(\boldsymbol{\beta})=\arg\min_{\boldsymbol{\gamma}} S(\boldsymbol{\beta},\boldsymbol{\gamma})$ and note that $\boldsymbol{\gamma}_n(\boldsymbol{\beta}_0)\to\boldsymbol{\beta}_0$.
4. Define the S estimate $\tilde{\boldsymbol{\beta}}_n$ of $\boldsymbol{\beta}_0$ by $\tilde{\boldsymbol{\gamma}}_n(\tilde{\boldsymbol{\beta}}_n)=\tilde{\boldsymbol{\beta}}_n$.
5. Define the S estimate $\tilde{\sigma}_n$ of σ_0 by $\tilde{\sigma}_n=s_n(\tilde{\boldsymbol{\beta}}_n)$.

The final estimate. We now suppose that $(\tilde{\boldsymbol{\beta}}_n,\tilde{\sigma}_n)$ is an "initial"high bdp and consistent but maybe inefficient estimate, such as the parametric S estimate defined above. To obtain a "final "estimate that keeps the bdp point of the initial estimate but which is highly efficient, we reject the outliers.

Suppose that we want to reject observations whose likelihoods under the initial model are smaller than a given cut-off value. Then, for this purpose, we can consider the cdf G_0 of the negative log-likelihood $l=\rho_0(u)$ under the model and the estimate of G_0 given by

$$G_{n,\tilde{\boldsymbol{\beta}}_n,\tilde{\sigma}_n}(z)=\frac{1}{n}\sum_{i=1}^n \left[\delta_i I\left(l_i^*\le z\right)+(1-\delta_i)P_{\tilde{\boldsymbol{\beta}}_n,\tilde{\sigma}_n}(l\le z|y>y_i^*)\right]$$

where $l_i^*=\rho_0(\tilde{r}_i^*)$ and $\tilde{r}_i^*=(y_i^*-\mathbf{x}_i^{\mathrm{T}}\tilde{\boldsymbol{\beta}}_n)/\tilde{\sigma}_n$. We write G_n in place of $G_{n,\tilde{\boldsymbol{\beta}}_n,\tilde{\sigma}_n}$. One can show that G_n is a consistent estimate of G_0. A fixed cut-off ζ on the likelihood scale can be defined as a large quantile of G_0, e.g., $\zeta=G_0^{-1}(0.99)$. To define an adaptive cut-off ϑ_n, that depends on the observed degree of contamination, we compare the tails of G_0 and G_n. Let $G_{n,\vartheta}$ denote G_n truncated at ϑ, i.e.,

$$G_{n,\vartheta}(z)=\begin{cases}G_n(z)/G_n(\vartheta) & \text{if } z\le\vartheta,\\ 1 & \text{otherwise.}\end{cases} \quad (10)$$

We look for the largest ϑ such that $G_{n,\vartheta}(z)\ge G_0(z)$ for all $z\ge\zeta$, i.e. $\vartheta_n=\sup\{\vartheta \mid G_{n,\vartheta}(z)\geqq G_0(z) \text{ for all } z\ge\zeta\}$. One can prove that, if the sample does not contain outliers, $\vartheta_n\to\infty$ a.s.

Suppose now that $\omega(z)$ is a a function such that (a) $\omega(z)$ is nonincreasing; (b) $\lim_{z\to-\infty}\omega(z)=1$; (c) $\omega(z)=0$ for $z>0$. For example, let $c>0$ and consider the function $\omega(z)=\rho^T(z,c)\cdot I(z\le 0)$, where $\rho^T(z,c)$ is in the biweight family (6). Then, define the weight function $w_{\vartheta_n}(z)=\omega\left(\rho_0(z)-\vartheta_n\right)$, where ϑ_n is a fixed or adaptive cut-off for outlier rejection. The "final"weighted estimate $(\hat{\boldsymbol{\beta}}_n,\hat{\sigma}_n)$ is defined by

$$E_{n,\hat{\boldsymbol{\beta}}_n,\hat{\sigma}_n}\left[w_{\vartheta_n}\left(\tilde{u}\right)\psi_0\left(\hat{u}\right)\mathbf{x}\right]=\mathbf{0}, \quad (11)$$

$$E_{n,\hat{\boldsymbol{\beta}}_n,\hat{\sigma}_n}\left[w_{\vartheta_n}\left(\tilde{u}\right)\psi_1\left(\hat{u}\right)\right]=b_{\vartheta_n} \quad (12)$$

where $\tilde{u}=\left(y-\mathbf{x}^{\mathrm{T}}\tilde{\boldsymbol{\beta}}_n\right)$, $\hat{u}=\left(y-\mathbf{x}^{\mathrm{T}}\hat{\boldsymbol{\beta}}_n\right)/\hat{\sigma}_n$, $b_{\vartheta_n}=E_0\left[w_{\vartheta_n}(u)\psi_1(u)\right]$. The estimate $(\hat{\boldsymbol{\beta}}_n,\hat{\sigma}_n)$ is a natural extension of the TML estimate for noncensored observations proposed in [7], which uses $\omega(z)=I(z\le 0)$. Note that when the sample does not contain outliers and $\vartheta_n\to\infty$, $w_{\vartheta_n}(u)\to 1$ for all u and the estimating equations (11)-(12) become the ML equations (4)-(5).

IV. Results

The following results are proved in [4] under suitable conditios.

1. The bdp of both the initial S estimate and the TML estimate for censored observations is not lower that $0.5-0.5m/n$, where m is the number censored observations.
2. At the model, the asymptotic distribution of $n^{1/2}(\hat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}_0,\hat{\sigma}_n-\sigma_0)$ is the same as the one of the maximum likelihood

TABLE I
SIMULATED ROOT MEAN SQUARE ERRORS AT THE NOMINAL GAUSSIAN MODEL

		Sample size n			
Parameter	Estimate	100	200	500	1000
intercept	WML	0.118	0.083	0.055	0.037
	ML	0.116	0.082	0.054	0.036
slope	WML	0.123	0.092	0.055	0.038
	ML	0.124	0.090	0.054	0.037
scale	WML	0.097	0.070	0.044	0.031
	ML	0.090	0.063	0.040	0.029

TABLE II
SIMULATED MAXIMUM ROOT MEAN SQUARE ERRORS (OVER $m = 1.0, 1.5, ..., 6.0$) UNDER POINT CONTAMINATION AT (x_0, mx_0), $\epsilon = 10\%$, $n = 100$, AND GAUSSIAN ERRORS.

	$x_0 = 1$		$x_0 = 10$	
Parameter	WML	ML	WML	ML
intercept	0.417	1.007	0.310	2.399
slope	0.434	0.998	0.652	4.661
scale	0.304	1.018	0.122	3.357

estimate defined by equations (4)-(5). It follows that the TML estimate for censored observations is fully efficient.

These theoretical results were confirmed by several Monte Carlo experiments in [4]. For example, using the simple regression Gaussian model with 35% of censored observations,

$$y_i = \alpha_0 + \beta_0 x_i + \sigma_0 u_i, \quad i = 1, ..., n,$$
$$v_i \sim N(\mu, 1),\ x_i \sim N(0, 1),\ u_i \sim N(0.668, 1),$$

with $\alpha_0 = 0$, $\beta_0 = 1$ and $\sigma_0 = 1$ we obtain the mean squared errors reported in Table 1. As expected, the TML estimate attains a very high performance, approaching the ML values when n increases. Similar results are obtained with a log-Weibull error model.

In order to investigate the behavior of the estimates in the presence of outliers, the simulated samples were contaminated with a fixed fraction $\epsilon = 10\%$ of outlying observations at (x_0, mx_0) for $x_0 = 1$ (low leverage point) and $x_0 = 10$ (high leverage point) and m varying over the regular grid $1.0, 1.5, ..., 6.0$. Table 2 reports the maximum rMSEs over the grid of the estimated parameters for $n = 100$.

V. DISCUSSION

The robust procedure described above may be considered the parametric counterpart of the estimates presented in [12] which used a conditional expectation approach introduced in [2], where censored residuals are replaced by their estimated conditional expectation given that the response is larger than the recorded censored value. The conditional expectation was based on the Kaplan-Meier distribution of the residuals. From the point of view of robustness, the KM estimate has an important drawback, because it distributes the mass of each censored residual r_i^* among all the noncensored residuals r_j such that $r_j > r_i^*$. Therefore, outliers may receive a very large mass and the regression estimate is more affected by outliers than in the uncensored case. The parametric TML estimates avoid this shortcoming. They are more more robust and more efficient than their nonparametric counterparts.

At present, the parametric TML estimates for AFT regression with Gaussian and logWeibull error distributions have been implemented in the package RobustAFT [9]. Extensions of the two step procedure to more flexible error models, such as the generalized loggamma family of distributions, are under study [1].

REFERENCES

[1] Agostinelli C., Marazzi A., Yohai V.J. (2011). Robust estimates of the generalized loggamma model. Submitted.

[2] Buckley, J. and James, I. (1979). Linear regression with censored data. *Biometrika*, 66, 429-436.

[3] Huber, P.J. (1981). Robust statistics. Wiley, New York.

[4] Locatelli I., Marazzi A., Yohai V.J. (2010). Robust accelerated failure time regression. *Computational Statistics & Data Analysis*, 55, 874-887.

[5] Locatelli I., Marazzi A. (2012). Robust estimates of hospital cost means with covariates, censored data, and outliers. Submitted.

[6] Maronna, R.A., Martin, R.D. and Yohai, V.J. (2007). Robust Statistics: Theory and Methods, Wiley, New York.

[7] Marazzi, A. and Yohai, V. J. (2004). Adaptively truncated maximum likelihood regression with asymmetric errors. *Journal of Statistical Planning and Inference*, 122, 271-291.

[8] Marazzi, A., Villar, A.J., Yohai, V.J., 2009. Supplemental material for Robust response transformations based on optimal prediction. Available at the Jasa Site for supplemental material.

[9] Marazzi A., Muralti J.L. (2011). RobustAFT: Robust Accelerated Failure Time Model Fitting, The Comprehensive R Archive Network.

[10] Mizera, I. (1993), On consistent M-estimators: tuning constants, unimodality and breakdown, *Kybernetika*, 30, 289-300.

[11] Rousseeuw P.J. and Yohai V.J. (1987). Robust regression by means of S-estimates. In Robust and Nonlinear Time Series, J. Franke, W. Härdle and R. Martin (Eds). *Lecture Notes in Statistics* 26, Springer, New York, pp. 256-272.

[12] Salibian-Barrera M. and Yohai V.J. (2008). High breakdown point robust regression with censored data. *The Annals of Statistics*, 36, 118-146.

Multiple imputation for estimating predictive ability in case-cohort surveys

Helena Marti
Inserm, CESP Centre for Research in Epidemiology and Population Health, U1018, Biostatistics team, F-94807 Villejuif, France
Email: helena.marti-soler@inserm.fr

Laure Carcaillon
Inserm, CESP Centre for Research in Epidemiology and Population Health, U1018, Hormones and Cardiovascular Disease team, F-94807 Villejuif, France
Email: laure.carcaillon@inserm.fr

***Abstract*—The weighted estimators generally used for analyzing case-cohort studies are not fully efficient and naive estimates of the predictive ability of a model from case-cohort data depend on the subcohort size. However, case-cohort studies represent a special type of incomplete data, and methods for analyzing incomplete data should be appropriate, in particular multiple imputation, which provides unbiased estimates of hazard ratios if data are missing at random and if the imputation model is correctly specified. Any tool proposed in the framework of cohort studies can be applied to case-cohort data using multiple imputation. In particular, the estimation of the predictive ability of a model or an additional variable. We performed simulations to validate the multiple imputation approach for estimating predictive ability of a model or of an additional variable in case-cohort surveys. When the imputation model of the phase-2 variable was correctly specified, multiple imputation estimates of predictive abilities were similar to those obtained with full data. By contrast, the naive estimates of the predictive ability provided by the case-cohort complete data sensibly differed from the full cohort values. So, multiple imputation is a simple approach for analyzing case-cohort data and provides an easy evaluation of the predictive ability of a model or of an additional variable. This approach was applied to a case-cohort survey about the effects of D-dimer levels on the risks of coronary heart disease and vascular dementia.**

I. Introduction

Case-cohort surveys produce incomplete data by design. A subcohort is selected by simple or stratified random sampling, all subjects are followed up and the events of interest are recorded. The phase-1 variables are observed for the entire cohort, while the phase-2 variables are only known for the case-cohort sample, i.e., subjects belonging to the subcohort and all those presenting the event of interest [1]. Thus, in case-cohort studies, the non-cases who do not belong to the subcohort are incompletely observed by design, enabling cost reduction with a small loss of efficiency. However, the weighted estimators generally used for analyzing case-cohort studies are not fully efficient.

On the other hand, case-cohort studies represent a special type of incomplete data where observations are missing at random [2]. Marti and Chavance [3] showed that multiple imputation is a good alternative to classical weighted methods for the analysis of case-cohort data. When the imputation model is correct, the multiple imputation approach provides unbiased estimates of the log hazard ratios and of the variance of their estimators. The multiple imputation approach was more precise than the usual weighted estimators for the parameters associated with phase-1 variables. The former was also slightly more precise than the latter for the phase-2 variable.

Moreover, because multiple imputation reconstitutes whole cohorts, any tool developed for cohorts can be applied to case-cohort data. No standard method exists for quantifying the usefulness or predictive ability of a model or an additional variable in the framework of case-cohort surveys. As shown below, a naive measurement of predictive ability from case-cohort data often leads to a biased estimate of the predictive ability because it varies with the censoring rate and thus depends on the subcohort size. So, we propose using the multiple imputation approach to perform inferences on the predictive ability of a model or of an additional variable.

II. Multiple imputation

In case-cohort surveys, incomplete observations are missing at random [2] by design, as the probability of being completely observed only depends on the case status, with simple random sampling, and on some phase-1 variables with stratified sampling. Multiple imputation is a simple and efficient method for analyzing missing at random observations, providing an approximation of the maximum likelihood estimator. This method relies on the generation of several plausibly completed data sets ($M \geq 2$), accounting for all levels of uncertainty concerning the missing values. A prediction model must be built, taking into consideration the relationships between the incomplete variable, the outcome and the other variables, as observed in the complete part of the data. The missing data are not replaced by their expectation but by a value drawn from the distribution posited by the model. To take into account the uncertainty concerning the parameters of the imputation model, several imputations are performed with parameters drawn from the asymptotic distribution of their estimator. An estimate of the parameter vector of interest, $\hat{\theta}_m$, $m = \{1, \ldots, M\}$, and an estimate of the covariance matrix of its estimator, $\widehat{V}(\hat{\theta}_m)$, are obtained from each completed data set. If the imputation model is correct, these estimators are not biased. The multiple imputation estimate, also unbiased, is the

mean of the M estimates:

$$\hat{\theta}_{MI} = \frac{1}{M}\sum_{m=1}^{M}\hat{\theta}_m \quad (1)$$

The multiplicity of imputations enables correct estimation of the variance of this single estimator, which is the sum of 2 components: the within-imputations component, W_{MI}, and the between-imputations component, B_{MI}:

$$\begin{aligned}\widehat{V}(\hat{\theta}_{MI}) &= \widehat{W}_{MI} + \widehat{B}_{MI} \\ &= \frac{1}{M}\sum_{m=1}^{M}\widehat{V}(\hat{\theta}_m) + (1+\frac{1}{M})\frac{\sum_{m=1}^{M}(\hat{\theta}_m - \hat{\theta}_{MI})(\hat{\theta}_m - \hat{\theta}_{MI})'}{M-1} \quad (2)\end{aligned}$$

where the factor $(1+M^{-1})$ is an adjustment for using a finite number of imputations [4].

In case-cohort surveys, we need to impute phase-2 variable values for the non-cases who do not belong to the subcohort. Under the rare disease assumption, we have shown that a generalized linear model, using all the complete data (cases and non-cases) and including the case indicator among the explanatory variables, has to be considered [3]. Practically, in addition to the case indicator and the stratification variables, when the subcohort has been selected by stratified sampling, it is necessary to include in the imputation model all the variables appearing in the proportional hazard model. Because imputations are based on asymptotic distributions, caution is necessary, since if too few subjects present the event of interest, the distributions of some estimators can differ from their asymptotic one. As a consequence, the maximum likelihood estimator of the imputation model parameters could be biased or not normally distributed.

III. Predictive ability of a model and of an additional variable

Harrell *et al.* [5] proposed the C index which measures the predictive ability of a model in cohort studies as the agreement between the order of the predicted and observed survival times in all pairs of subjects from the target population (the event of interest is assumed to be death, leading to the use of survival terminology). However, with censored data, it is not possible to consider all the pairs of subjects because survival time is not observed for censored subjects. Let T_i be the survival time for subject i, $i = 1,\ldots,N$, where N is the cohort size, and C_i the censoring time for subject i. We observe $X_i = min(T_i, C_i)$. Usable pairs are those for which the order of the predicted survival times can be compared to the order of the true survival times, i.e., pairs formed by 2 uncensored subjects or an uncensored subject and a subject censored after the uncensored subject's death. A pair of censored subjects carries no information about its agreement with the expected survival according to the model since the order of the survival times is not known. Similarly a pair formed by a subject whose survival time is observed and a subject censored before this survival time provides no information on this agreement since the unknown survival time could be anterior or posterior to the observed one. Harrell *et al.* [6] showed that, in the common models used for survival analysis, such as the Cox model, the predicted survival times and the predicted survival probabilities at a fixed time t can be interchanged for the comparison. The Harrell's C index is defined as:

$$C = \frac{\pi_c}{\pi_c + \pi_d} \quad (3)$$

where π_c is the probability of concordance for a pair (i,j) and π_d is the probability of discordance. We assume continuous survival times and continuous predicted survival probabilities, so $P(X_i = X_j) = P(Y_i = Y_j) = 0$, thus $\pi_c + \pi_d = 1$. C is estimated by the proportion of concordant pairs among the usable pairs. The estimated variance was given by Kremers [7].

In practice, we are often interested in estimating the predictive ability of an additional phase-2 variable. Let M_1 be a Cox model including only phase-1 variables and C_1 the C index of M_1. Let M_2 be a Cox model adding the phase-2 variable to M_1 and C_2 the C index of M_2. Harrell's predictive ability of the added phase-2 variable is $\Delta = C_2 - C_1$. Complementary measures of predictive ability of a new variable, such as the net reclassification improvement (NRI) and the integrated discrimination index (IDI), were proposed by Pencina [8]. NRI needs some a priori meaningful risk categories. It quantifies the correct reclassification introduced by using a model with the added variable as compared to the classification obtained without this variable. The IDI can be viewed as a continuous version of the NRI with probabilities used instead of categories. To estimate the predictive ability of a model or of an additional variable, we reconstructed plausible whole cohorts using multiple imputation. For each reconstructed whole cohort, we could directly obtain C_1, C_2, Δ, NRI, IDI and their respective variances. Using equations (1) and (2), we obtained the multiple imputation estimates of these quantities. Concerning the variance of Δ, the between-imputation component is estimated by the empirical variance of the M estimates of Δ provided by the M completed data sets. However, for the within-imputation component, the asymptotic variance of the estimator provided by a complete data set, does not have an analytical form. With a fully observed cohort, bootstrapping is a way to estimate the variance of the corresponding Δ. Therefore, each whole cohort reconstructed by multiple imputation has to be resampled.

IV. Simulations

Two phase-1 variables were simulated: a binary variable, Z_1, and a Gaussian variable, Z_3, observed for the entire cohort. Also simulated was a phase-2 standard Gaussian variable, Z_2, which was independent of Z_1, but having a correlation coefficient of 0.2 with Z_3. The survival time had an exponential distribution, with $\lambda = exp(\beta_1 Z_1 + \beta_2 Z_2 + \beta_3 Z_3)$. Three scenarios were considered: $\beta_1 = \beta_2 = \beta_3 = 0$, $\beta_1 = \beta_3 = log(2)$ and $\beta_2 = log(1.5)$, and $\beta_1 = \beta_2 = \beta_3 = log(2)$. The censoring time followed a uniform distribution over the interval $[0,\tau]$, where τ was chosen so that the probability of

an event was approximately 0.03 (τ = 0.025). The cohort size was 10,000. We also simulated a phase-1 variable predictive of Z_2, $\widetilde{Z}_2 \equiv Z_2 + \varepsilon$ with $\varepsilon \sim N(0, \sigma^2)$ independent of Z_2. The variance σ^2 was fixed at 1 which corresponds to a correlation between Z_2 and $\widetilde{Z}_2$ of approximately 0.7. We wanted to estimate the predictive ability of Z_2. The cohort was divided into 9 strata based on the tertiles of $\widetilde{Z}_2$ and Z_3, and the non-cases were chosen by stratified sampling. We simulated case-cohort samples with the subcohort size set at 300 or 1,000 subjects. The phase-2 variable was not available for non-cases not included in the subcohort, so multiple imputation was used to complete the data set. The imputation model was: $Z_2 = \alpha_0 + \alpha_1 I_{case} + \alpha_2 Strata + \epsilon$, where α_0 and α_1 are scalar, α_2 is a vector coefficient, I_{case} is the case indicator, $Strata$ is the vector of stratum indicators and ϵ is the vector of errors independently and identically distributed $\sim N(0, \sigma)$. Five imputations were performed and 5 complete data sets were generated for each cohort. We estimated the predictive ability of models with and without the phase-2 variable, the predictive ability of the phase-2 variable, NRI and IDI. We also applied the naive estimators of Harrell's C index, NRI and IDI to the complete data of the simulated case-cohort samples.

A. Results

The results are reported in Table I. In the scenario $\beta_1 = \beta_2 = \beta_3 = 0$, the mean C index was around 0.52 for both models, without and with Z_2, whatever the analysis performed. In the scenarios $\beta_1 = \beta_3 = log(2)$ and $\beta_2 = log(1.5)$ or $\beta_2 = log(2)$, the naive computation of C with the case-cohort data led to lower predictive abilities than the full cohort, especially for the smaller subcohorts. By contrast, the Harrell's C indexes estimated by multiple imputation were similar to those computed from the full cohorts and did not depend on the subcohort size. The estimated dispersion of the C index was greater than the observed dispersion of the estimates even with the full cohort. The rejection percentages of the null hypothesis $\Delta = 0$ with the full cohort analysis and with multiple imputation were always similar. As a consequence of the standard error overestimation, the observed first type error rate was lower than 5%. Nevertheless, in the considered scenarios, the observed power was very high. As expected, the loss of power when comparing case-cohort with multiple imputation to full cohort analysis was small: with $\beta_2 = log(1.5)$, the observed power was 84.6% with a subsample size of 300, and 90.6% with a subsample size of 1,000 versus 91.6% with the full cohort. Multiple imputation estimates of NRI and IDI indexes were close to those obtained with the full cohort analysis and did not depend on the subcohort size. By contrast, whatever the effect of the phase-2 variable, the estimation of NRI and IDI in the case-cohort sample provided larger measures of these indexes than the full cohort analysis.

V. Application

Briefly, the 3C-Study was designed to examine the relationship between risks factors of vascular diseases and dementia in a community housing 9,294 persons aged 65 years and over between 1999 and 2001 in three French cities. The detailed methodology has been previously described [10]. A case-cohort substudy was conducted [11], to investigate the relationship between biomarkers, such as plasma levels of D-dimer (a marker of coagulation and fibrinolysis) and the 4-year incidence of coronary heart disease (CHD), stroke and all subtypes of dementia, including vascular dementia (VaD). The phase-1 variables provided information on socio-demographic characteristics, education, medical history, diet, alcohol and tobacco use. Blood pressure, height and weight were also available. A subcohort of size $n = 1,254$, (13.5% of the full cohort) was randomly selected, stratifying on age, sex and recruitment center. Observed cumulated incidences of CHD and VaD were approximately 2% and 0.6%, respectively. Plasma D-dimer levels were only available for phase-2 subjects. Carcaillon *et al.* [11] treated quintiles of D-dimer level both qualitatively and linearly. They reported a linear increase in the risk of VaD according to D-dimer quintiles.

We re-assessed the relationship between plasma D-dimer levels and the risk of CHD and VaD, using multiple imputation and weighted estimators, and evaluated the predictive ability of D-dimer levels on both risks. We included the same explanatory variables as Carcaillon *et al.* [11] although we used tertiles of D-dimer rather than quintiles, to estimate CHD and VaD risks, due to the small number of events and to the necessity of obtaining an estimator of the imputation model asymptotically distributed. Therefore, to estimate the risk of CHD, the proportional hazard model included the phase-1 variables: age, sex, center, body mass index, hypertension, hypercholesterolemia, diabetes, tobacco use, diabetes drugs, and as phase-2 variables, indicators of D-dimer tertiles. To estimate the risk of VaD, the proportional hazard model included the phase-1 variables: age, sex, centre, educational level, body mass index, the presence or absence of an apolipoprotein $\epsilon 4$ allele and indicators of D-dimer tertiles.

For each outcome (CHD or VaD), it was necessary to reproduce the relationships among the incomplete variable, the outcomes and the confounder variables. For each outcome, we built an imputation model of tertiles of D-dimer levels, including the variables used in the proportional hazard model and the case-indicator. We estimated the predictive ability of proportional hazard models, without (C_1) and with (C_2) D-dimer levels, $\Delta = C_2 - C_1$, and IDI for CHD and VaD risks. The NRI requires that some a priori meaningful risk categories be known. Based on the Third Adult Treatment Panel [ATP III] [12] risk classification for the 10-year risk of CHD, we adapted the cut-offs to 4-year risk. For VaD, we do not know a priori meaningful risk categories and did not compute NRI.

A. Results

Table II gives the estimated hazard ratios associated with D-dimer tertiles. The multiple imputation and the weighted approaches yielded similar estimates. The CI of the hazard ratio associated with the linear effect of a one-tertile difference were respectively (0.94 - 1.38) versus (0.92 - 1.38) for CHD and (1.13 - 2.53) versus (1.13 - 2.67) for VaD. For phase-1

variables, both estimators provided similar results, but multiple imputation was always the more precise (data not shown).

Harrell's C for the models including only phase-1 variables were above 0.69 for CHD risk and above 0.86 for VaD risk (Table III). Hence, CHD and VaD risks were largely explained by standard risk factors, and the inclusion of plasma D-dimer levels did not significantly improve the predictive ability of the model, despite the fact that elevated D-dimer levels significantly increased the VaD risk. For CHD also, the index did not significantly differ from 0.

VI. Conclusion

As far as we know, this is the first time that predictive values are computed from case-cohort data. We showed that the naive application of the C index to case-cohort data provided estimates sensibly different from the full cohort analysis for the predictive ability of a model as for that of an additional variable. Harrell's C index could theoretically be estimated with a weighted approach, but this can be computationally difficult because it requires weighting each pair by the pairwise sampling probabilities, i.e., using a square matrix of size $N'(N'-1)$, where N' is the size of the case-cohort sample. Computing the variance of this Horvitz-Thompson estimator requires weighting each quadruplet by the quadruple-wise sampling probabilities, i.e., working with a matrix of size $N'(N'-1)(N'-2)(N'-3)$. By contrast, multiple imputation easily allows estimation of the predictive ability of a model or of an additional phase-2 variable and of their variances in the context of case-cohort data, only requiring bootstrapping to estimate the variance of the predictive ability of the phase-2 variable. Multiple imputation provided estimates of Harrell C, NRI and IDI indexes similar to those obtained with the full cohort analysis. Note, however, that the predictive abilities were always overestimated because the same data were used to estimate the model and its predictive ability.

Acknowledgment

This study was supported by a grant from the Région Île-de-France. It used data from the Three-City study which is conducted under an agreement between the Institut National de la Santé et de la Recherche Médicale and the Université Victor Segalen-Bordeaux 2.

References

[1] Prentice R: **A case-cohort design for epidemiologic cohort studies and disease prevention trials**. *Biometrika* 1986, **73**:1–11.

[2] Little R, Rubin D: *Statistical analysis with missing data*. New York: Wiley 1987.

[3] Marti H, Chavance M: **Multiple imputation analysis of case-cohort studies**. *Stat Med* 2011, **30**(13):1595–1607.

[4] Rubin DB, Schenker N: **Multiple imputation in health-care databases: an overview and some applications.** *Stat Med* 1991, **10**(4):585–598.

[5] Harrell FE, Califf RM, Pryor DB, Lee KL, Rosati RA: **Evaluating the yield of medical tests**. *Journal of the American Medical Association* 1982, **247**(18):2543–2546.

[6] Harrell F, Lee K, Mark D: **Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors**. *Stat Med* 1996, **15**(4):361–387.

[7] Kremers WK: **Concordance for survival time data: fixed and time-dependent covariates and possible ties in predictor and time**. Technical Report Series No. 80, Departement of Health Science Research, Mayo Clinic, Rochester, Minnesota 2007.

[8] Pencina M, D'Agostino R Sr, D'Agostino R Jr, Vasan R: **Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond**. *Stat Med* 2008, **27**(2):157–172.

[9] Marti H, Carcaillon L, Chavance M: **Multiple imputation for estimating hazard ratios and predictive abilities in case-cohort surveys**. *BMC Medical Research Methodology* 2012, **12**(1):24. doi:10.1186/1471-2288-12-24

[10] Alperovitch A, 3C Study Grp: **Vascular factors and risk of dementia: Design of the three-city study and baseline characteristics of the study population**. *Neuroepidemiology* 2003, **22**(6):316–325.

[11] Carcaillon L, Gaussem P, Ducimetiere P, Giroud M, Ritchie K, Dartigues JF, Scarabin PY: **Elevated plasma fibrin D-dimer as a risk factor for vascular dementia: the Three-City cohort study**. *J Thromb Haemost* 2009, **7**(12):1972–1978.

[12] **Executive Summary of The Third Report of The National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, And Treatment of High Blood Cholesterol In Adults (Adult Treatment Panel III)**. *J Am Med Assoc* 2001, **285**(19):2486–2497.

TABLE I

MEAN OF THE PREDICTIVE ABILITY ESTIMATES (EST), MEAN OF THE STANDARD ERROR ESTIMATES ($\widehat{SE}$) AND STANDARD ERROR OF THE ESTIMATES (SE). RESULTS FROM 1,000 SIMULATIONS.

	$\beta_1 = \beta_2 = \beta_3 = 0$				$\beta_1 = \beta_3 = log(2)$, $\beta_2 = log(1.5)$				$\beta_1 = \beta_2 = \beta_3 = log(2)$			
	Est	$\widehat{SE}$	SE	% H_0 rejected	Est	$\widehat{SE}$	SE	% H_0 rejected	Est	$\widehat{SE}$	SE	% H_0 rejected
Cohort												
C_1	0.518	0.033	0.012		0.727	0.032	0.015		0.733	0.029	0.014	
C_2	0.524	0.033	0.013		0.747	0.031	0.015		0.733	0.029	0.014	
Δ	0.006	0.010	0.009	3.7	0.020	0.007	0.007	91.6	0.049	0.010	0.010	100
NRI	0.007	0.017	0.019	4.8	0.071	0.030	0.033	52.5	0.167	0.034	0.035	99.9
IDI	$2e^{-4}$	$2e^{-4}$	$3e^{-4}$	6.0	0.014	0.003	0.005	99.9	0.048	0.006	0.009	99.9
MI1000												
C_1	0.518	0.033	0.012		0.724	0.032	0.016		0.733	0.029	0.014	
C_2	0.526	0.033	0.013		0.745	0.031	0.016		0.783	0.027	0.014	
Δ	0.008	0.012	0.010	3.4	0.021	0.008	0.008	90.6	0.049	0.010	0.011	100
NRI	0.009	0.019	0.017	1.5	0.076	0.033	0.033	64.8	0.172	0.037	0.036	100
IDI	$3e^{-4}$	$3e^{-4}$	$4e^{-4}$	3.5	0.014	0.004	0.005	99.0	0.045	0.008	0.010	100
MI300												
C_1	0.518	0.033	0.012		0.724	0.032	0.016		0.733	0.029	0.014	
C_2	0.528	0.033	0.012		0.745	0.031	0.017		0.783	0.027	0.015	
Δ	0.010	0.014	0.011	3.0	0.021	0.008	0.009	84.6	0.050	0.011	0.012	100
NRI	0.013	0.023	0.018	1.3	0.076	0.035	0.035	57.0	0.172	0.039	0.039	99.7
IDI	$4e^{-4}$	$4e^{-4}$	$5e^{-4}$	1.8	0.014	0.005	0.006	87.5	0.046	0.010	0.012	100
CC1000												
C_1	0.528	0.032	0.013		0.667	0.033	0.015		0.670	0.031	0.014	
C_2	0.534	0.033	0.015		0.709	0.032	0.022		0.737	0.029	0.014	
Δ	0.006	0.010	0.010	4.7	0.043	0.011	0.017	100	0.067	0.012	0.012	100
NRI	0.017	0.031	0.033	6.7	0.147	0.039	0.043	96.7	0.261	0.041	0.043	100
IDI	0.002	0.001	0.003	15.2	0.058	0.009	0.014	100	0.114	0.011	0.017	100
CC300												
C_1	0.523	0.034	0.013		0.620	0.037	0.016		0.620	0.034	0.015	
C_2	0.529	0.034	0.015		0.647	0.036	0.016		0.668	0.032	0.015	
Δ	0.006	0.010	0.009	3.6	0.027	0.011	0.011	83.3	0.048	0.013	0.013	99.8
NRI	0.019	0.039	0.043	6.2	0.154	0.043	0.050	94.4	0.257	0.046	0.051	99.9
IDI	0.002	0.001	0.003	13.9	0.040	0.008	0.014	99.8	0.078	0.010	0.017	100

Cohort, full cohort estimates; MI300, MI1000: multiple imputation estimates with subcohort sizes set, respectively, at 300 and 1,000; CC300, CC1000, case-cohort estimates with subcohort sizes set, respectively, at 300 and 1,000.

C_1, Harrell's C index of the Cox model without the phase-2 variable.

C_2, Harrell's C index of the Cox model with the phase-2 variable.

Δ, Harrell's predictive value of the phase-2 variable, $H_0 : \Delta = 0$.

NRI, Net reclassification index by adding the phase-2 variable, H_0 : NRI = 0.

IDI, Integrated discrimination index by adding the phase-2 variable, H_0 : IDI = 0.

TABLE II
ESTIMATES OF HAZARD RATIOS (HR) AND 95% CONFIDENCE INTERVAL (CI) ASSOCIATED WITH D-DIMER TERTILES.

	Multiple imputation estimates	Weighted estimates
	HR (95% CI)	HR (95% CI)
Risk of CHD and D-Dimer [a]		
T1	1.00 (reference)	1.00 (reference)
T2	1.42 (0.99 - 2.04)	1.40 (0.97 - 2.04)
T3	1.32 (0.89 - 1.97)	1.30 (0.84 - 1.99)
Linear trend	1.14 (0.94 - 1.38)	1.13 (0.92 - 1.38)
Risk of VaD and D-Dimer [b]		
T1	1.00 (reference)	1.00 (reference)
T2	1.57 (0.63 - 3.93)	1.60 (0.63 - 4.09)
T3	2.77 (1.17 - 6.57)	2.93 (1.22 - 7.06)
Linear trend	1.69 (1.13 - 2.53)	1.74 (1.13 - 2.67)

CHD, cardiovascular heart disease; VaD, vascular dementia;
T1, tertile 1; T2, tertile 2; T3, tertile 3;
[a] Adjusted for age, center, sex, body mass index, hypertension, hypercholesterolemia, diabetes, diabetes drugs, tobacco use.
[b] Adjusted for age, center, sex, educational level, body mass index, apolipoprotein $\epsilon 4$.

TABLE III
PREDICTIVE ABILITY AND 95% CONFIDENCE INTERVAL (CI) OF D-DIMER TERTILES ON CARDIOVASCULAR HEART DISEASE (CHD) AND VASCULAR DEMENTIA (VAD) RISKS.

	CHD		VaD	
	Estimate	95% CI	Estimate	95% CI
C_1	0.693	(0.622 - 0.764)	0.865	(0.787 - 0.943)
C_2	0.694	(0.621 - 0.767)	0.874	(0.798 - 0.950)
Δ	0.002	(-0.004 - 0.008)	0.009	(-0.011 - 0.029)
NRI	0.009	(-0.049 - 0.066)	-	-
IDI	0.001	(-0.001 - 0.003)	0.0004	(-0.0002 - 0.0010)

C_1, C index of the Cox model without the phase-2 variable.
C_2, C index of the Cox model with the phase-2 variable.
Δ, Harrell's predictive ability of the phase-2 variable.
NRI, net reclassification improvement by adding the phase-2 variable.
IDI, integrated discrimination index by adding the phase-2 variable.

Cramér-von Mises test with estimated parameters

G. Martynov
Institute for Information Transmission Problems,
Russian Academy of Sciences &
Higher School of Economics, Moscow Russia
email: martynov@iitp.ru

We will consider the problem of the testing a hypothesis that a distribution function of observations of a random variable belong to the parametric class $\{G(x,\theta), \theta \in \Theta\}$. For testing such hypothesis, it can be used the weighted Cramér-von Mises statistic

$$\omega_n^2 = n \int_{-\infty}^{\infty} \psi^2(G(x,\theta_n))(F_n(x) - G(x,\theta_n))^2 \, dG(x,\theta_n).$$

This statistics asymptotically does not depend of estimated parameters vector θ_n for the families $G((x-\theta_1)/\theta_2)$ and $G((x/\theta_2)^{\theta_1})$, $\theta_2 > 0$ [3]. Both families include the most parametric distribution families used on the practice. It is presented method for the exact calculation of the limit distributions of such statistics (see [4]). The method is based on previously found formulas for the eigenvalues and eigenfunctions for the covariance operators for the limit processes without estimated parameters [1]. The described methods are competitive in practice, compared with the methods presented in [2].

References

[1] Deheuvels, P. and Martynov, G. V. (1996). Cramér–Von Mises–type tests with applications to tests of independence for multivariate extreme–value distributions. *Communications in Statistics - Theory and Methods.* V. 25 871–908.

[2] Khmaladze E.V. (1981) A martingale approach in the theory of parametric goodness-of-fit tests,*Theor. Prob. Appl.* V. 26, N 2

[3] Martynov G. (2010) Note on the Cramer-von Mises test with estimated parameters. *Publ. Math Debrezen*, V. 76/3 341-346.

[4] Martynov G. (2011) Weighted Cramer-von Mizes test with estimated parameters. *Communications in Statistics-Theory and Methods*, V. 40. Issue 19-20. P. 3569- 3586.

Multivariate frailty models for two types of recurrent events with a dependent terminal event: Application to breast cancer data

Yassin Mazroui
INSERM U897 and
Université Bordeaux Segalen
Bordeaux, France
Email: Yassin.Mazroui@isped.u-bordeaux2.fr

Virginie Rondeau
INSERM U897 and
Université Bordeaux Segalen
Bordeaux, France
Email: Virginie.Rondeau@isped.u-bordeaux2.fr

***Abstract*—Individuals may experience more than one type of recurrent event and a terminal event during the life course of a disease. Follow-up may be interrupted for several reasons, including the end of a study, or patients lost-to-follow-up, which are non-informative censoring events. Death could also mean that stop follow-up is stopped, hence, it is considered as a dependent terminal event. We propose a multivariate frailty model that jointly analyzes two types of recurrent events with a dependent terminal event. Two estimation methods are proposed: a semi-parametrical approach using penalized likelihood estimation where baseline hazard functions are approximated by M-splines, and another one with piecewise constant baseline hazard functions. Finally, we derived martingale residuals to check the goodness-of-fit. We illustrate our proposals with a real data set on breast cancer. The main objective was to model the dependency between the two types of recurrent events (locoregional and metastatic) and the terminal event (death) after a breast cancer.**

I. Introduction

Relapses and death are often the events of interest for long term diseases. In statistics, relapses are treated as recurrent events and death as a terminal event. These last decades, researchers have developed methods for analyzing such events separately or jointly. Proportional hazard (PH) models [1] were developed first for an event of interest that could happen once. This model implicitly assumes a homogeneous population to be studied. Frailty models [2], [3] that are extensions of PH models to one type of recurrent event, aim to account for potential heterogeneity caused by unmeasured prognostic factors and inter-recurrence dependency. The occurrence of one event may produce biological weakening, damage or strengthening, which is why the notion of frailty was introduced. A review of the existing methods for the analysis of recurrent events is detailed in [4]. These methods assume that the follow-up period is independent of the underlying recurrent process. Later, joint modeling approaches relying on frailty models [5], [6] or marginal models [7] appeared to consider informative censoring and terminal event as a dependent processes. [8]–[10] and [11] have proposed a more flexible approach allowing frailty to differ between the risk of recurrences and the risk of terminal event. Nevertheless, for some disease, there are several, different but related types of recurrent events of interest. [12] proposed a class of semiparametric marginal rate models for multiple type recurrent events without considering a terminal event. [13] proposed a generalization of [14]'s approach that proposed arbitrary structures for both the relationship between the recurrent events and the terminal event and the effect of covariates on terminal event. [15] proposed a semiparametric additive rates model where the dependency between recurrent events and the terminal event is nonparametric.

In these recent approaches, dependency between events is considered, but not of interest. In this article, we develop a multivariate frailty model for recurrent events in the presence of a dependent terminal event, with right censored survival data. Hence, the relationship between disease recurrences and survival can be assessed with random effects, frailties, or other dependency parameters that add more flexibility. Our model considers the natural history of the disease beyond the first diagnosis and also accounts for covariate effects. For instance, after a first breast cancer, the evolution toward locoregional recurrences or metastases and their association with death is studied. To analyze recurrent event data, the focus can be placed on time-between-events (i.e. gap times) or time-to-events (i.e. calendar times). The proposed approach can deal with both timescales. [8] and [10] used a Monte Carlo EM algorithm to estimate the hazard functions and the parameters, which could be time-consuming. Furthermore, these methods proposed to estimate the cumulative hazard functions, but we can not directly estimate a smooth hazard function, which often has a meaningful interpretation in epidemiological studies. Most of the time, the baseline intensity estimate is based on Breslow's estimate, leading to a piecewise-constant baseline hazard function or an unspecified baseline hazard function. [9] have proposed a semi-parametric estimation method using penalized likelihood. We propose two estimation methods: maximization of likelihood for model with a parametric (piecewise constant) baseline hazard function and maximization of the penalized likelihood for models with baseline hazard functions approximated by splines.

The aim of this article is to provide methods for jointly analyzing two types of recurrent events with a terminal event, to measure potential dependency between each type of event and to estimate the influence of prognostic factors. The work was motivated by the analysis of different types of recurrences, locoregional or metastatic, and death for patients after a breast cancer diagnosis. Researchers have shown that the risk of death is increased after a metastatic relapse [16]. Further after a locoregional relapse, the risk of metastatic recurrence is also increased [17]. The dependency between locoregional relapse and the risk of death is observed for young women [18] but it is not clearly established for women of any age [19].

II. Joint modeling framework for different types of events

In this section, we present the multivariate frailty model and the estimation methods.

A. The model

For each individual i, $\quad i = 1, ..., N$, we consider the two types of observed recurrent event times $X_{ij}^{(l)}, j = 1, ..., n_i^{(l)}$ since the initiation of the processes; $l \in \{1, 2\}$ indicates the type of recurrent events. Each individual is censored by the terminal event time $T_i^* = \min(C_i, D_i)$ which could be a non-informative censoring C_i or the death D_i. The considered event time vector is $T_{ij} = min\{X_{ij^{(1)}}^{(1)}, X_{ij^{(2)}}^{(2)}\}, j = 1, ..., n_i, j^{(l)} = 1, ..., n_i^{(l)}$ and $T_{in_i+1} = T_i^*$. The time of study entry for each individual i is T_{i0} assumed equal to 0, we are indeed interested in the history of the disease since the diagnosis. We denote the event indicators $\delta_{ij}^{(l)} = \mathcal{I}_{\{T_{ij}=X_{ij}^{(l)}\}}$, and the death indicator $\delta_i^* = \mathcal{I}_{\{T_i^*=D_i\}}$. The recurrent gap times S_{ij} represents the duration between two consecutive events: $S_{ij} = T_{ij} - T_{i(j-1)}$. We actually observe $\{T_{ij}$ (or S_{ij} if gap-times); T_i^*; $\delta_{ij}^{(l)}, l \in \{1, 2\}; \delta_i^*\}$. Let $N_i^{R(l)^*}(t)$ count the number of the recurrent events of type l for individual i over the interval $(0, t], i = 1, ..., N$. Because of censoring, it is impossible to observe these numbers. Actually, we observe the processes $N_i^{R(l)}(t) = N_i^{R(l)^*}(min(T_i^*, t))$ which count the observed numbers of recurrent events of type l. Similarly, denote by $N_i^{D^*}(t) = \mathcal{I}_{\{D_i \le t\}}$ and $N_i^D(t) = \mathcal{I}_{\{T_i^* \le t, \delta_i^* = 1\}}$ the actual and the observed death indicator by time t, respectively. Furthermore, let $Y_i(t) = \mathcal{I}_{\{t \le T_i^*\}}$ denote whether or not the individual i is at risk of an event at time t. The number of recurrent events that occur for subject i over the small interval $[t, t + dt)$ is $dN_i^{R(l)^*}(t) = N_i^{R(l)^*}((t + dt)^-) - N_i^{R(l)^*}(t^-)$ and we have $dN_i^{R(l)}(t) = Y_i(t) dN_i^{R(l)^*}(t), l \in \{1, 2\}$. We can notice that $n_i^{(l)} = N_i^{R(l)}(T_i^*)$. The history of the i^{th} process up to time t is denoted by: $\mathcal{H}_{it} = \sigma\{Y_i(h), N_i^{R(l)}(h), l \in \{1, 2\}, N_i^D(h), Z_i(h), 0 \le h \le t\}, \quad i = 1, ..., N$ where $Z_i(h)$ is a vector of possibly time-dependent covariates. We denote the filtration : $\mathcal{F}_{it} = \sigma\{H_{it}, u_i, v_i\}, \quad i = 1, ..., N$. The underlying intensity processes are jointly dependent through two correlated random effects u_i, v_i. The random effects account for the non-observed heterogeneity, the inter-recurrences dependency and the dependency between different event types. We assume that the two recurrent and the terminating processes are continuous, which means that a recurrent event and death cannot happen at the same time. In the case of simultaneous death and recurrent event types, we consider that the recurrent event happens first in the small interval [t, t+dt). We consider that the actual recurrent event processes $N_i^{R(l)^*}(t), l \in \{1, 2\}$, are constant after death time D_i but can increase after the censoring time C_i. That means death precludes the observation of new recurrent events but on the contrary censoring, such as end of study or lost to follow-up, does not interrupt the occurrence of new recurrent events, they are simply not observed. The recurrent event intensity processes at time t are, for ($l \in \{1, 2\}$): $Y_i(t) r_i^{(l)}(t) dt = P(dN_i^{R(l)^*}(t) = 1 | \mathcal{F}_{it^-})$, where $r_i^{(l)}(t) dt = P(dN_i^{R(l)^*}(t) = 1 | Z_i(t), u_i, v_i, D_i > t^-)$. The death intensity process at time t is: $Y_i(t)\lambda_i(t) dt = P(dN_i^D(t) = 1 | \mathcal{F}_{it^-})$, where $\lambda_i(t) dt = P(dN_i^{D^*}(t) = 1 | Z_i(t), u_i, v_i, D_i > t^-)$. Finally, we model the intensity functions of counting processes for the two types of recurrent events and the terminal event observed processes given that the individual is still alive. The multivariate joint frailty model for two types of recurrent events with a terminal event is (in the calendar or time-to-event timescale):

$$\begin{cases} r_i^{(1)}(t | u_i, v_i) & = r_0^{(1)}(t) \exp(\beta_{\mathbf{1}}' Z_i(t) + u_i) \\ r_i^{(2)}(t | u_i, v_i) & = r_0^{(2)}(t) \exp(\beta_{\mathbf{2}}' Z_i(t) + v_i) \\ \lambda_i(t | u_i, v_i) & = \lambda_0(t) \exp(\beta_{\mathbf{3}}' Z_i(t) + \alpha_1 u_i + \alpha_2 v_i) \end{cases} \tag{1}$$

where $r_0^{(l)}(t)$, $l \in 1, 2$ and $\lambda_0(t)$ are respectively the recurrent and terminal event baseline hazard functions, and $\beta_{\mathbf{1}}, \beta_{\mathbf{2}}, \beta_{\mathbf{3}}$ the regression coefficient vectors associated with $Z_i(t)$ the covariate vector. The covariates could be different for the recurrent event rates and death rate. The covariates may be time-dependent with for instance a function of the number of previous events. We consider that death stops new occurrences of recurrent events of any type, hence given $t > D$, $dN^{R(l)*}(t), l \in 1, 2$ takes the value 0. Note that death is not an informative censoring. An individual can not experience a recurrent event after death whereas it is possible after informative or non-informative censoring. Thus, the terminal and the two recurrent event processes are not independent or even conditional upon frailties and covariates. We consider the rates of recurrent events among individuals still alive. The three components in the above multivariate model are linked together by two random effects u_i, v_i. We assume that u_i and v_i are two Gaussian correlated random effects with :

$$(u_i, v_i)^T \sim \mathcal{MVN}\left(\mathbf{0}, \begin{pmatrix} \theta & \rho\sqrt{\theta\eta} \\ \rho\sqrt{\theta\eta} & \eta \end{pmatrix}\right), \rho = Corr(u_i, v_i)$$

The variance of u_i (θ) specifies the dependency between occurrences of the recurrent events of type 1 and the terminal event and also the inter-recurrence dependency. Similarly, the variance of v_i (η) specifies the dependency between occurrences of the recurrent events of type 2 and the terminal event and also the inter-recurrence dependency. The parameter α_l,

$l \in 1, 2$, assesses the sign and the strength of the dependency between the l^{th} type of recurrent events and the terminal event, and informs whether the terminal event and these recurrent events are really dependent. A high value of α_1 (resp. α_2) and the variances θ (resp. η) illustrates a strong dependency between recurrent events of type 1 (resp. 2) and terminal event. A high absolute value of the correlation coefficient ρ corresponds to a strong dependency between the two types of recurrent events. Note that $\rho \in [-1, 1]$. A correlation coefficient $\rho = 0$ means that the occurrences of these two recurrent events for the same subject are independent of each other. If the α_1 (resp. α_2) and/or the variance θ (resp. η) are not significantly different from 0 then this means that the terminal event and the recurrent events of type 1 (resp. 2) are independent.

B. Estimation

Maximization of the log-likelihood and maximization of the penalized log-likelihood are the estimation methods proposed to estimate the different parameters $(\beta_1, \beta_2, \beta_3, \theta, \eta, \rho, \alpha_1, \alpha_2)$ and the baseline hazard functions $r_0^{(l)}(t), l \in \{1, 2\}$ for recurrent events or $\lambda_0(t)$ for death times. The first one is used for parametric baseline hazard functions (piecewise or Weibull) and the second one is when baseline hazard functions are non-parametric and approximated by M-splines. Let Φ denote $(r_0^{(1)}(.), r_0^{(2)}(.), \lambda_0(.), \beta_1, \beta_2, \beta_3, \theta, \eta, \rho, \alpha_1, \alpha_2)$ the parameters to estimate. We directly use $\widehat{H}^{-1}$ as a variance estimator, where H is minus the converged Hessian of the penalized log-likelihood or the log-likelihood, depending on the estimation method. We also use the robust sandwich estimator $\widehat{H^{-1}IH^{-1}}$, I is the Fisher information matrix of the non-penalized likelihood. $\widehat{H^{-1}IH^{-1}}$ gives almost same results as $\widehat{H}^{-1}$. Furthermore, to deal with the constraint on the correlation parameter ρ, we estimate uu where $\rho = \frac{2\exp(uu)}{1+\exp(uu)} - 1 \in [-1, 1]$ for $uu \in \mathbb{R}$ and a positivity constraint is imposed to the baseline hazard function parameters $(r_0^{(l)}, \lambda_0)$ and to the variances of the random effects (θ, η). The standard errors for the parameters estimated with constraint were computed by the Δ-method [20].

C. Goodness-of-fit

1) Likelihood cross validation criterion: The approximated likelihood cross validation criterion LCV [21] is used as a choice model criterion:
LCV $= \dfrac{trace(\widehat{H^{-1}I}) - l(\widehat{\Phi})}{\sum_{i=1}^{N} n_i}$. It allows to compare between models estimated with maximum likelihood and penalized likelihood. In the case of maximum penalized likelihood estimation, the AIC is not applicable. Note that for maximum likelihood estimation $trace(\widehat{H^{-1}I}) = np$ (np: number of parameters in the model), and LCV $= \dfrac{np - l(\widehat{\Phi})}{\sum_{i=1}^{N} n_i}$

2) Martingale residuals: The use of martingale residuals have been proposed for model checking in survival data [22]. It enables us to check whether the model predicts correctly the number of observed events. Martingale residuals applied to our proposed model become :

$$\begin{cases} M^{R(1)}(t) &= N^{R(1)}(t) - \mathcal{I}_{\{t \le T_i*\}} \exp(\widehat{u_i}) \int_0^t \widehat{r_i^{(1)}}(u)du \\ M^{R(2)}(t) &= N^{R(2)}(t) - \mathcal{I}_{\{t \le T_i*\}} \exp(\widehat{v_i}) \int_0^t \widehat{r_i^{(2)}}(u)du \\ M^{D}(t) &= N^{D}(t) - \mathcal{I}_{\{t \le T_i*\}} \exp(\widehat{\alpha_1}\widehat{u_i} + \widehat{\alpha_2}\widehat{v_i}) \int_0^t \widehat{\lambda_i}(u)du \end{cases} \quad (2)$$

The parameters $\widehat{\Phi} = \{\widehat{r_0^{(1)}}, \widehat{r_0^{(2)}}, \widehat{\lambda_0}, \widehat{\theta}, \widehat{\eta}, \widehat{\rho}, \widehat{\alpha_2}, \widehat{\alpha_1}\}$ used here are the estimated parameters, obtained after the maximization of the log-likelihood or the penalized log-likelihood. For the calculus, we need individual estimates of the random effects $\widehat{u_i}, \widehat{v_i}$. We used empirical Bayesian estimators obtained by maximizing the posterior probability density function of the random effects (details can be found in the Appendix B of the supplementary material). Then we use the Lowess function of R software to have smoothed curves for these martingale residuals.

III. Application to Breast cancer patients

This study aimed to estimate the prognostic factors (hazard ratios) associated with the occurrences of breast cancer locoregional relapses, metastatic relapses and death. It also aimed to study the dependencies between these three events of interest. Several papers [17] showed that after a locoregional relapse, there is a high risk of experiencing a metastatic event. Other papers [16] showed a strong link between a metastatic relapse and death. The link between locoregional relapses and death have not been clearly established yet [19]. We consider here two different types of recurrent events which could be associated. Moreover, death is considered as a dependent terminating event for the relapses, it is necessary to analyze these recurrent events jointly with the terminal event to make valid inferences. The use of the proposed multivariate frailty is justified with such data. The coefficients α_1 and α_2 indicate the sign of the association whether a type of recurrent events, locoregional or/and metastatic relapses and death are significantly positively or negatively associated. The variances of the random effects (u_i, v_i) measure the dependency between the two types of recurrent events and death and also whether there are inter-relapse dependencies. With such an approach, we are able to assess the association between breast cancer locoregional relapses, metastatic relapses and death, and secondly the intra-recurrence dependency. We concluded that the risk of locoregional recurrences is associated with the risk of metastatic recurrences and the risk of metastatic recurrences is also associated with death. The risk of death is not associated with the risk of locoregional recurrence.

IV. Software

All of our proposals have been implemented in R with the freely available package frailtypack [23]. This package can be used to fit a variety of joint frailty models or other frailty models for recurrent or clustered time-to-event data with several different options for the baseline risk functions. The package can be downloaded from the Comprehensive R Archive Network accessible via http://cran.r-project.org/package=frailtypack.

V. Conclusion

We presented a multivariate frailty model with two correlated random effects to simultaneously model two types of recurrent events with a dependent terminal event. The proposed model for possibly right-censored data was able to express the dependency among the two types of recurrent events, but also to deal with the association between recurrent and terminal events. This approach accounts for non-observed heterogeneity and inter-recurrence dependency. Simulation studies indicated that the approach works well for practical situations and was slightly better than using three separate shared frailty models (one for each type of event). One advantage of the abovementioned approach is that different covariate effects may be assessed for the two types of recurrence rates or death rates, these covariates may be time-dependent or time-independent. In total, there are many reasons to use multivariate frailty models for three survival endpoints, including giving a general description of the data, correcting for bias in survival analysis due to dependent terminal events and improving efficiency of survival analyses thanks to the use of supplementary information. Furthermore, they also provide information on whether one or both recurrence times can be used as surrogate endpoints for overall survival.

References

[1] D. Cox, "Regression models and life-tables," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 34, no. 2, pp. 187–220, 1972.

[2] D. Clayton, "A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence," *Biometrika*, vol. 65, no. 1, pp. 141–151, 1978.

[3] J. Vaupel, K. Manton, and E. Stallard, "The impact of heterogeneity in individual frailty on the dynamics of mortality," *Demography*, vol. 16, no. 3, pp. 439–454, 1979.

[4] R. Cook and J. Lawless, *The statistical analysis of recurrent events*. Springer Verlag, New-York, 2007.

[5] T. Lancaster and O. Intrator, "Panel Data with Survival: Hospitalization of HIV-Positive Patients." *Journal of the American Statistical Association*, vol. 93, no. 441, p. 4653, 1998.

[6] X. Huang and R. Wolfe, "A frailty model for informative censoring," *Biometrics*, vol. 58, no. 3, pp. 510–520, 2002.

[7] D. Schaubel and J. Cai, "Analysis of clustered recurrent event data with application to hospitalization rates among renal failure patients," *Biostatistics*, vol. 6, no. 3, pp. 404–419, 2005.

[8] L. Liu, R. Wolfe, and X. Huang, "Shared frailty models for recurrent events and a terminal event," *Biometrics*, vol. 60, no. 3, pp. 747–756, 2004.

[9] V. Rondeau, S. Mathoulin-Pelissier, H. Jacqmin-Gadda, V. Brouste, and P. Soubeyran, "Joint frailty models for recurring events and death using maximum penalized likelihood estimation: application on cancer events," *Biostatistics*, vol. 8, no. 4, pp. 708–721, 2007.

[10] X. Huang and L. Liu, "A joint frailty model for survival and gap times between recurrent events," *Biometrics*, vol. 63, no. 2, pp. 389–397, 2007.

[11] Y. Mazroui, S. Mathoulin-Pélissier, P. Soubeyran, and V. Rondeau, "General joint frailty model for recurrent event data with a dependent terminal event: Application to follicular lymphoma data," *Statistics in Medicine*, In press 2012.

[12] J. Cai and D. Schaubel, "Marginal means/rates models for multiple type recurrent event data," *Lifetime data analysis*, vol. 10, no. 2, pp. 121–138, 2004.

[13] L. Zhu, J. Sun, X. Tong, and D. Srivastava, "Regression analysis of multivariate recurrent event data with a dependent terminal event," *Lifetime data analysis*, vol. 16, no. 4, pp. 478–490, 2010.

[14] M. Wang, J. Qin, and C. Chiang, "Analyzing recurrent event data with informative censoring," *Journal of the American Statistical Association*, vol. 96, no. 455, pp. 1057–1065, 2001.

[15] D. Zeng and J. Cai, "A semiparametric additive rate model for recurrent events with an informative terminal event," *Biometrika*, vol. 97, no. 3, pp. 699–712, 2010.

[16] J. O'Shaughnessy, "Extending survival with chemotherapy in metastatic breast cancer," *The Oncologist*, vol. 10, no. 3, pp. 20–29, 2005.

[17] E. Montagna, V. Bagnardi, N. Rotmensz, G. Viale, G. Renne, G. Cancello, A. Balduzzi, E. Scarano, P. Veronesi, A. Luini, S. Zurrida, S. Monti, M. G. Mastropasqua, L. Bottiglieri, A. Goldhirsch, and M. Colleoni, "Breast cancer subtypes and outcome after local and regional relapse," *Annals of Oncology*, 2011.

[18] P. Elkhuizen, M. van de Vijver, J. Hermans, H. Zonderland, C. van de Velde, and J. Leer, "Local recurrence after breast-conserving therapy for invasive breast cancer: high incidence in young patients and association with poor survival," *International Journal of Radiation Oncology* Biology* Physics*, vol. 40, no. 4, pp. 859–867, 1998.

[19] I. Monteiro Grillo, M. Jorge, P. Marques Vidal, M. Ortiz, and P. Ravasco, "The effect of locoregional recurrence on survival and distant metastasis after conservative treatment for invasive breast carcinoma," *Clinical oncology*, vol. 17, no. 2, pp. 111–117, 2005.

[20] K. Knight, "Mathematical Statistics, Texts in Statistical Science," 2000.

[21] D. Commenges, P. Joly, A. Gégout-Petit, and B. Liquet, "Choice between semi-parametric estimators of Markov and non-Markov multi-state models from generally coarsened observations," *Scandinavian Journal of Statistics*, vol. 34, no. 1, pp. 33–52, 2007.

[22] D. Commenges and V. Rondeau, "Standardized martingale residuals applied to grouped left truncated observations of dementia cases," *Lifetime Data Analysis*, vol. 6, no. 3, pp. 229–235, 2000.

[23] V. Rondeau, Y. Mazroui, and J. Gonzalez, "FRAILTYPACK: An R package for the analysis of correlated survival data with frailty models using the penalized likelihood estimation," *Journal of Statistical Software*, In press 2012.

Field-Failure Predictions Based on Failure-time Data with Dynamic Covariate Information

William Q. Meeker

Department of Statistics
Center for Nondestructive Evaluation
Iowa State University
Ames, Iowa 50011

Today, there are more and more products installed with automatic data-collecting devices such as smart chips and sensors as well as cellular and network communications capabilities that track how and under which environments the product is being used. While there is tremendous amount of such dynamic data being collected, there is little research on using such data to provide more accurate reliability information for products and systems. Motivated by a consulting problem, this paper focuses on using failure-time data with dynamic covariate information to make warranty and other field-failure predictions. The dynamic covariate information is incorporated into a parametric failure-time model through a cumulative exposure model. A prediction procedure that accounts for unit-to-unit and temporal variability in the use rate is developed to predict field-failure returns at a future time. We also define a metric to quantify the improvements obtained using dynamic information in the prediction accuracy. Simulation studies were conducted to study the effect of different sources of covariate process variability on predictions.

Key Words: Accelerated failure time model; Cumulative exposure model; Dynamic data; Lifetime data; Usage history, Warranty returns.

This is Joint work with Yili Hong.

A nonparametric test for comparing treatments with missing data and dependent censoring

A. Mezaouer
Faculté des Sciences
Université Saad Dahleb
Blida, Algeria
Email: a_mezaouer@univ-blida.dz

J.-F. Dupuy
Institut de Recherche Mathématique de Rennes
INSA de Rennes, France
Email: Jean-Francois.Dupuy@insa-rennes.fr

K. Boukhetala
Université des Sciences et Technologie
Houari Boumédienne
Bab-Ezzouar, Algeria
Email: kboukhetala@usthb.dz

***Abstract*—The log-rank test is often used to compare randomized treatment groups with respect to the distribution of a failure time outcome. The so-called stratified log-rank test can be used when it is necessary to adjust for the effect of some discrete covariate that may be predictive of the outcome. In many applied situations, this discrete covariate is missing for some of the patients and moreover, the distribution of the censoring time depends on the treatment group. In this talk, we introduce a modified version of the stratified log-rank test, which accommodates both problems simultaneously. The asymptotic distribution of this new test under the null hypothesis of equality of the randomized treatment groups is established. A numerical study is conducted to examine the finite-sample behavior of this test under both the null and alternative hypotheses.**

I. Introduction

The log-rank test is widely used to compare randomized treatment groups with respect to the distribution of some failure time outcome. If $\lambda_k(\cdot)$ is the hazard rate of failure in the kth treatment group ($k = 1, \ldots, K$), the null hypothesis of interest is: $\lambda_1(\cdot) = \ldots = \lambda_K(\cdot)$, and the alternative is: $\lambda_j(\cdot) \neq \lambda_{j'}(\cdot)$ for some $j \neq j'$. The log-rank test is a nonparametric test, whose construction is based on the simple idea of comparing the Nelson-Aalen estimators of the $\int \lambda_k(t)\,dt$, $k = 1, \ldots, K$, with an estimator of the hypothesized common value $\int \lambda(t)\,dt$. The resulting test statistic is asymptotically χ^2 distributed under the null hypothesis, which can be proved by invoking a martingale central limit theorem [1].

If one needs to control for a covariate that may be predictive of the outcome, and if this covariate is discrete, the log-rank can be generalized to the so-called stratified log-rank test [1]. Consider a clinical trial where n patients are randomly assigned to K different treatment groups, and suppose that we wish to compare survival between groups, while adjusting for some discrete factor S with L modalities (also called strata, such as income groups or disease stages for example). If $\lambda_{k,l}(\cdot)$ is the hazard rate of failure in the kth treatment group and lth stratum, then the test for treatment effect can be formulated as

$$H_0 : \lambda_{1,l}(\cdot) = \ldots = \lambda_{K,l}(\cdot) \quad \text{for every } l = 1, \ldots, L$$

versus

$$H_1 : \text{there exists } j \text{ and } j' \text{ such that } \lambda_{j,l}(\cdot) \neq \lambda_{j',l}(\cdot) \text{ for some } l.$$

Let $T_1^0, \ldots, T_n^0$ be the times from randomization to failure observed in the K pooled groups. Let $C_1, \ldots, C_n$ be right-censoring times (the C_i are assumed to be independent of the T_i^0 and non-informative). For each patient i, we observe $T_i = \min(T_i^0, C_i)$ and $\Delta_i = 1(T_i^0 \leq C_i)$, where $1(\cdot)$ is the indicator function. Assume that the data consist of n independent and identically distributed quadruplets $(T_i, \Delta_i, G_i, S_i)$, $i = 1, \ldots, n$, where $G_i \in \{1, \ldots, K\}$ and $S_i \in \{1, \ldots, L\}$ respectively indicate the group and stratum of the ith patient. Let $N_i(t) = \Delta_i 1(T_i \leq t)$, $Y_i(t) = 1(T_i \geq t)$, and define

$$E_{k,l}^{(n)}(t) = \frac{\sum_{i=1}^n Y_i(t) 1(G_i = k) 1(S_i = l)}{\sum_{i=1}^n Y_i(t) 1(S_i = l)}.$$

Then the stratified log-rank statistic for the test of no randomized treatment effect is of the form

$$U = (Z_1, \ldots, Z_{K-1}) \hat{\Theta}^{-1} (Z_1, \ldots, Z_{K-1})',$$

where for every $k = 1, \ldots, K-1$,

$$Z_k = \sum_{i=1}^n \int_0^\tau \left\{ 1(G_i = k) - \sum_{l=1}^L 1(S_i = l) E_{k,l}^{(n)}(t) \right\} dN_i(t), \quad (1)$$

τ denotes the end of the study period, and $\hat{\Theta}$ is the estimated asymptotic covariance matrix of $(Z_1, \ldots, Z_{K-1})'$. Under H_0, U is asymptotically distributed as a χ^2 distribution with $K-1$ degrees of freedom.

In some applications, the stratum S may be missing for some patients. For example, consider the case where S represents the histological stage of patients included in a cancer clinical trial. The determination of S may require a biopsy, which due to expensiveness may not be performed on all the study subjects. One simple solution to handle such incomplete stratum information is to perform a complete-case analysis that is, to discard patients with unobserved stratum. This, however, may induce a substantial loss of power, as will be illustrated in our simulation study. [5] and [6] considered the distinct but related problem of estimation in the stratified proportional hazards model with missing strata. The authors proposed a modified version of the maximum partial likelihood estimator, in which the unobserved stratum indicators are replaced by an estimate of their conditional expectation given available auxiliary covariates (this is the so-called regression calibration

idea, see for example [3]). Simulation results provided some evidence that such a replacement substantially improves on the complete-case analysis.

In many applications, it also happens that the distribution of censoring time depends on the treatment group. This arises, for example, when censoring follows from a study dropout caused by treatment toxicity. The treatment group with the heaviest toxicity will be more likely to have a higher dropout rate, and thus a higher censoring rate, than the other groups. Inverse probability of censoring weighted (IPCW) procedures have been proposed to remedy this problem (see for example [2], [8], [10]).

In this talk, we propose and investigate a test of no randomized treatment effect, when the patients stratum information is only partially available *and* the distribution of censoring time depends on the treatment group. The test we propose combines the regression calibration and IPCW principles.

The rest of this abstract is organized as follows. In Section II, we introduce the new test statistic and we derive its asymptotic distribution under H_0. In Section III, we describe a short simulation study investigating the finite-sample behaviour of the proposed test. A discussion concludes our contribution.

II. The test statistic and its asymptotic distribution under H_0

Assume that n independent patients are randomly assigned to K treatment groups. Assume that the stratum value is missing for some of these patients. Thus, a subsample is available where all variables (T, Δ, G, S) are observed, while only (T, Δ, G) are observed for the other patients. We assume that some auxiliary variables $W \in R^p$ are observed for all patients, and that W provides a partial information about S when S is missing. Let R be the indicator variable which is 1 if S is observed and 0 otherwise. Throughout the paper, we assume that T^0 and C are independent given G, S, W and R, and that C is independent of S and W given G. However as mentioned above, the distribution of the censoring time depends on the treatment group. We assume that T^0 is independent of W given S (that is, the auxiliary variables W provide no additional information about failure when the true stratum S is known), and that G is independent of S and W, as is the case in randomized clinical trials. We assume that R is independent of T^0, C and G, and that $0 < P(R = 1) < 1$. Finally, we assume that R and S are independent given W, which is the so-called missing-at-random assumption.

We consider the problem of implementing the stratified log-rank test of H_0 based on n independent vectors $(T_i, \Delta_i, G_i, W_i, R_i, R_i S_i)$, $i = 1, \ldots, n$ of possibly incomplete data when moreover, the distribution of censoring time depends on the treatment group. To tackle simultaneously the missing strata and dependent censoring problems, we introduce a modified version of U, which is obtained by:

1) replacing any missing stratum indicator $1(S_i = l)$ in (1) by its conditional expectation given the auxiliary W (this idea is related to regression calibration methods, see, for example, [6] and [9]), and
2) weighting every patient by the inverse of the conditional (given the patient's treatment group) survival function of the censoring time (this idea is related to the inverse probability of censoring weighted principle, e.g., [2], [8], [10]).

Precisely, we propose to base our test statistic on the following modified version of (1):

$$\widetilde{Z}_k = \sum_{i=1}^{n} \int_0^\tau \mu(G_i, t) \left\{ G_i^k - \sum_{l=1}^{L} D_i^l \widetilde{E}_{k,l}^{(n)}(t) \right\} dN_i(t),$$

where for every $i = 1, \ldots, n$, $k = 1, \ldots, K$, $l = 1, \ldots, L$, and $t \in [0, \tau]$, $G_i^k = 1(G_i = k)$, $D_i^l = R_i 1(S_i = l) + (1 - R_i) E[1(S_i = l)|W_i]$, $\mu(G_i, t) = 1/P(C \geq t|G_i)$ (where $P(C \geq t|G)$ is the survival function of the censoring time in group G), and $\widetilde{E}_{k,l}^{(n)}(t) = \widetilde{S}_{k,l}^{(n)}(t)/\widetilde{S}_l^{(n)}(t)$, with

$$\widetilde{S}_{k,l}^{(n)}(t) = \frac{1}{n} \sum_{i=1}^{n} Y_i(t) G_i^k D_i^l \mu(G_i, t)$$

and

$$\widetilde{S}_l^{(n)}(t) = \frac{1}{n} \sum_{i=1}^{n} Y_i(t) D_i^l \mu(G_i, t).$$

Before stating our result, we need to introduce some further notations. For every $i = 1, \ldots, n$ and $k = 1, \ldots, K$, let

$$V_{i,k} = \int_0^\tau \mu(G_i, t) \left\{ G_i^k - \sum_{l=1}^{L} D_i^l \widetilde{E}_{k,l}^{(n)}(t) \right\} dN_i(t) - \frac{1}{n} \sum_{j=1}^{n} \sum_{l=1}^{L} \int_0^\tau \frac{Y_i(t) D_i^l D_j^l \mu(G_i, t) \mu(G_j, t)}{\widetilde{S}_l^{(n)}(t)} \times \left\{ G_i^k - \widetilde{E}_{k,l}^{(n)}(t) \right\} dN_j(t),$$

define $V_i = (V_{i,1}, \ldots, V_{i,K-1})'$, and $\widehat{\Sigma} = \sum_{i=1}^{n} V_i V_i'$. We are now in position to state our main result for the new test statistic:

$$\widetilde{U} := (\widetilde{Z}_1, \ldots, \widetilde{Z}_{K-1}) \widehat{\Sigma}^{-1} (\widetilde{Z}_1, \ldots, \widetilde{Z}_{K-1})',$$

under some regularity conditions that are described in [7]:

Theorem. *Under H_0, as $n \to \infty$, $\widetilde{U}$ converges in distribution to a χ^2 distribution with $K - 1$ degrees of freedom.*

Based on this theorem, the proposed test rejects H_0 if $\widetilde{U} \geq \chi^2_{1-\alpha}(K-1)$, where $\chi^2_{1-\alpha}(K-1)$ is the quantile of order $1 - \alpha$ of $\chi^2(K-1)$.

III. A simulation study

We conducted a simulation study to evaluate the finite sample behavior of the proposed test under various conditions. We considered the case of $K = 2$ randomized treatment groups and $L = 2$ strata. In each group and stratum, the event times T_i^0 were generated from a Weibull distribution $W(\alpha, \lambda)$ with hazard rate $\lambda(t) = \alpha \lambda t^{\alpha - 1}$ (the Weibull distribution is flexible and has a wide range of applications in survival analysis). The failure times of stratum 1 in group 1 were

generated from $W(\alpha_1, \lambda_1)$, and those of stratum 1 in group 2 from $W(\alpha_1, \lambda_1 r_1)$, where '$r_1$' denotes the hazard rates ratio of two patients in stratum 1 of groups 1 and 2 respectively. The failure times of stratum 2 in group 1 were generated from $W(\alpha_2, \lambda_2)$ and those of stratum 2 in group 2 from $W(\alpha_2, \lambda_2 r_2)$, with '$r_2$' being the hazard rates ratio of two patients in stratum 2 of groups 1 and 2 respectively.

We used $\alpha_1 = .5, \alpha_2 = .75, \lambda_1 = .75$, and $\lambda_2 = 1.5$. Three cases were considered for the pairs (r_1, r_2) of hazard ratios: (a) $(r_1, r_2) = (1, 1)$, (b) $(r_1, r_2) = (1.5, 1.5)$, (c) $(r_1, r_2) = (1.25, 2)$. Case (a) corresponds to the null case of no difference between treatment groups, within each stratum. Cases (b) and (c) correspond to various magnitudes of difference between groups. In each case, the censoring times were generated from exponential distributions with parameters θ_1 in group 1 and θ_2 in group 2, with θ_1 and θ_2 chosen to yield censoring percentages equal to c_1 in group 1 and c_2 in group 2 (letting $\theta_1 \neq \theta_2$ ensures that the distribution of censoring depends on the treatment group). We considered $(c_1, c_2) = (5, 20)$ and $(c_1, c_2) = (20, 50)$ (more results are given in [7])

Let n_1 and n_2 denote respectively the sample size in group 1 and 2 (with $n = n_1 + n_2$). We considered various values for (n_1, n_2): $(n_1, n_2) = (50, 50)$, $(n_1, n_2) = (100, 100)$, and $(n_1, n_2) = (150, 150)$. The auxiliary variable W was taken to be 2-dimensional ($W = (W_1, W_2)'$) with W_1 (respectively W_2) generated from the uniform distribution on $[-1, 1]$ (respectively the normal distribution with mean 0 and standard deviation 0.5). A logistic regression model

$$P(S = 1|W) = \frac{\exp(b_0 + b_1 W_1 + b_2 W_2^2)}{1 + \exp(b_0 + b_1 W_1 + b_2 W_2^2)}$$

was taken for the relationship between S and W, with (b_0, b_1, b_2) chosen so that within each treatment group, each stratum contains approximately half of the patients.

The following stratum missingness percentages were considered: $20\%, 40\%$. For each patient, the missingness indicator R was obtained by randomly drawing a Bernoulli random variable, with parameter chosen to yield the prescribed overall missingness percentage.

As mentioned previously, in practical situations the weighting functions $\mu(G_i, \cdot)$ and/or the conditional probabilities $P(S_i = l|W_i) = E[1(S_i = l)|W_i]$ may either be known (from previous studies, for example), or completely unknown. In this latter case, one would estimate them and substitute the estimated values in the proposed test statistic $\widetilde{U}$ (the resulting statistic will be denoted by $\widehat{U}$ in the sequel). However, the null asymptotic distribution of $\widehat{U}$ may be somewhat distorted from the $\chi^2(K-1)$ distribution. This, in turn, may affect the size and power of the test based on $\widehat{U}$. In fact, our simulation results (see [7] for full details) show that as long as the $\mu(G_i, \cdot)$ and $P(S_i = l|W_i)$ are reasonably estimated, the prescribed level of the test is nearly maintained by $\widehat{U}$, and that $\widehat{U}$ outperforms the complete-case log-rank test in term of power.

Various methods may be used to estimate $P(S_i = l|W_i)$. Here, we used local logistic regression, and we used non-parametric Kaplan-Meier estimators within each treatment group to estimate the censoring survival functions $P(C \geq t|G_i)$. See [7] for full details.

For each configuration of the design parameters, 1000 replications were obtained using the software R. Based on these 1000 repetitions, we obtained the empirical size (case (a)) and power (cases (b) and (c)) of the "estimated version" $\widehat{U}$ of the proposed test $\widetilde{U}$, at the significance level 0.05. For comparison, we included the results of the stratified log-rank test based on complete cases only (*i.e.* on individuals with known stratum). In the sequel, we shall refer this latter test to as U_{cc} for short. Table 1 summarizes the results for an overall stratum missingness percentage equal to 40% (the results for 20% are similar and are therefore not presented).

Table 1. Empirical size and power of $\widehat{U}$ and U_{cc}, based on 1000 replicates.

		(5,20)		(20,50)	
n	(r_1, r_2)	$\widehat{U}$	U_{cc}	$\widehat{U}$	U_{cc}
100	(1, 1)	.067	.045	.070	.067
	(1.5, 1.5)	.354	.228	.272	.204
	(1.25, 2)	.349	.227	.303	.220
200	(1, 1)	.074	.055	.065	.051
	(1.5, 1.5)	.692	.440	.565	.369
	(1.25, 2)	.592	.386	.605	.418
300	(1, 1)	.060	.046	.076	.048
	(1.5, 1.5)	.823	.596	.762	.472
	(1.25, 2)	.772	.535	.813	.527

From these results, it appears that the proposed test $\widehat{U}$ performs well and clearly outperforms the stratified test based on the complete cases. The empirical level of $\widehat{U}$ tends to exceed (but only slightly) 0.05 in all cases. This may be due to the replacement of the unknown $P(S_i = l|W_i)$ and $\mu(G_i, \cdot)$ by their estimations, which causes the null asymptotic distribution of $\widehat{U}$ to be slightly distorted from the $\chi^2(K-1)$. But as expected, in cases (b) and (c), the powers of $\widehat{U}$ are greater than those of U_{cc} for every sample size and censoring percentages. In particular, $\widehat{U}$ maintains a high power even when the censoring percentage heavily depends on the treatment group (20% in group 1, 50% in group 2), while at the same time the powers of U_{cc} substantially decrease.

IV. Discussion

We have constructed and investigated a modified version of the stratified log-rank test of no randomized treatment effect. This new test statistic is useful when the stratum information is missing at random for some patients and the distribution of the censoring time depends on the treatment group. From our simulations, we have found that this test performs well

compared to the only alternative available so far, namely a complete-case based stratified log-rank test. Now, several questions still deserve attention.

First, we have assumed a missing-at-random mechanism for the stratum missingness. Investigating the robustness of the proposed test to a deviation to this assumption constitutes a topic for further numerical investigations. Extending the proposed method to non-ignorable missingness may be a non-trivial task however: the missing-at-random assumption is central in our proofs and for estimating the stratum belonging probabilities $P(S_i = l|W_i)$ when they are unknown.

Second, in order to accommodate group-dependent censoring, we have used the inverse probability of censoring weighted principle, with weight function $\mu(G_i, t) = h(\widehat{P}(C \geq t|G_i))$ and $h(x) = 1/x$. Alternative test statistics $\widetilde{Z}_k$ may be obtained by choosing other forms for h (we refer to [4] for alternative forms of weighting functions in the different context of bias correction for score tests arising from misspecified proportional hazards regression models). Searching for the function h which yields the most efficient testing procedure constitutes another non-trivial but very interesting task.

REFERENCES

[1] Andersen, P.K., Borgan, Ø., Gill, R.D., Keiding, N., 1993. Statistical models based on counting processes. Springer: New York.

[2] Cain, L.E., Cole, S.R., 2009. Inverse probability-of-censoring weights for the correction of time-varying noncompliance in the effect of randomized highly active antiretroviral therapy on incident AIDS or death. *Statistics in Medicine* 28, 1725-1738.

[3] Carroll, R.J., Ruppert, D., Stefanski, L.A., 1995. Measurement Error in Nonlinear Models. Monographs on Statistics and Applied Probability vol. 63, Chapman & Hall: London.

[4] DiRienzo, A.G., Lagakos, S.W., 2001. Bias Correction for Score Tests Arising from Misspecified Proportional Hazards Regression Models. *Biometrika* 88, 421–434.

[5] Dupuy, J.-F., Leconte, E., 2008. Cox regression with missing values of a covariate having a non-proportional effect on hazard of failure. In: Mathematical methods in survival analysis, reliability and quality of life (Applied Stochastic Methods Series), 133-150, ISTE: London.

[6] Dupuy, J.-F., Leconte, E., 2009. A study of regression calibration in a partially observed stratified Cox model. *Journal of Statistical Planning and Inference* 139, 317–328.

[7] Mezaouer, M., Dupuy, J.-F., Boukhetala, K., 2011. A nonparametric K-sample test with missing data and dependent censoring. Submitted.

[8] Robins, J.M., Finkelstein, D.M., 2000. Correcting for noncompliance and dependent censoring in an AIDS Clinical Trial with inverse probability of censoring weighted (IPCW) log-rank tests. *Biometrics* 56, 779–788.

[9] Weller, E.A., Milton, D.K., Eisen, E.A., Spiegelman, D., 2007. Regression calibration for logistic regression with multiple surrogates for one exposure. *Journal of Statistical Planning and Inference* 137, 449–461.

[10] Yoshida, M., Matsuyama, Y., Ohashi, Y., 2007. Estimation of treatment effect adjusting for dependent censoring using the IPCW method: an application to a large primary prevention study for coronary events (MEGA study). *Clinical Trials* 4, 318–328.

Multi-state model for prostate cancer development

Michalski A.I.
Institute of Control Sciences, Russian Academy of Sciences
Moscow, Russia
ipuran@yandex.ru

Zharinov G.M.
Russian Research Centre for Radiology and Surgical Technologies
St. Petersburg, Russia

***Abstract*—Prostate cancer development is considered as jumps between states with different PSA levels. The moments of jumps are unobserved. Likelihood for observed trajectories is derived for censored times of transitions in the presence of two competing risks of death. Maximum likelihood estimates for transition intensity matrix calculated using real data about 1981 males. Expected times spent at different states are estimated. The proportions of these times in total life expectancy are estimated as well.**

Keywords-multistate Markov-model; censored observations; prostate cancer; survival in state.

I. INTRODUCTION

Multi-state models are often used in investigations of epidemiological processes and processes of health changes [1]. The structure of such models ranges from a "simple" "illness-death" model with one absorbing terminal state [2] till the complex models with many states, corresponding to different states of the disease progression, and several absorbing states, reflecting competing risks of different causes of death. Mathematical background for multi-state modeling is continuous-time Markov chains, which are described in terms of transitions intensity matrix [3].

Estimation of a transition intensity is simple if one observes the time of transition. The maximum likelihood estimate for stationary transition intensity is given in this case by the number of transitions divided by the time, spent at the initial state. In reality observations of transition times are censored and one is to apply different approaches for transitions intensity matrix estimation [4].

II. MODEL FOR PROSTATE CANCER

The report presents a multi-state model which describes prostate cancer development based on the observations of 1981 males aged from 39 till 89 years. Mean age at the start of observation was 65.0 years, mean duration of observation was 4.0 years and did not exceed 26 years. In this group 613 cases of prostate cancer death and 170 cases of death of other causes were registered. Mean age at prostate cancer death was 67.6 years, mean age at death of other causes was 76.0 years. Three states of the multi-state model were determined by the level of PSA at the moment of investigation. The states were "normal" (PSA≤4ng/mL), "upper norm" (4ng/mL<PSA≤40ng/mL) and "high risk" (PSA>40ng/mL). Two absorbing states were included in the model, corresponding to prostate cancer death and death of cause different from prostate cancer. Transition to any absorbing state was allowed from any alive state. Transitions between states "normal" and "high risk" were allowed only via state "upper norm".

Likelihood function for presented data was constructed as a product of probabilities to observe personal trajectories of transitions between states. Probability of a given trajectory was calculated as product of conditional probabilities to observe a person at the two successive times in the two fixed states. Given the transitions intensity matrix these probabilities are calculated as solutions of Chapman-Kolmogorov equations with different initial conditions. If the last observed state was absorbing one the last term in the product was calculated as sum of products between conditional probabilities to get the 'alive' state from the previous one and corresponding transition intensity to the absorbing state. This reflects the hypotheses that the time of death was reported without censoring.

III. RESULTS

Maximum likelihood estimates for transitions intensity matrix are presented in table I. These intensities were used for estimation of expected time, spent in the three 'alive' states and life expectancy for a person, observed in the specific state. The estimates are shown in table II. For a person, observed in the specific state, the proportion of life expectancy, spent in the different states, was calculated and are shown in table II.

From the table II one can see that a man with "normal" PSA level not exceeding 4ng/mL may have such PSA level during 3.7 years in average. After this period a person either get higher PSA level or die of prostate cancer or die of other cause of death. Estimates in table I show that chances for these two last events are equal.

A man with PSA level ranges from 4ng/mL till 40ng/mL may stay in this state 1.3 years in average and after this he takes one of four options. A man can get lower PSA level, can get higher PSA level, can die of prostate cancer or can die of other cause of death. Estimates in table I show that the chances to get the "normal" PSA level are more that two times higher the chances to develop the disease state, which corresponds to "high risk" PSA level. Chance of death of prostate cancer increases almost twice in comparison with the "normal" PSA level state while death of other cause of death has the same likelihood.

The average time spent in the state with PSA level higher then 40ng/mL is estimated as 1.6 years and it is close to the average time spent in the previous state. Chance to return to the state with lower PSA level is approximately 6% higher than chance to die of prostate cancer, which is almost 5 times higher than that in the previous state. Chances of death of other cause is stay the same as in two previous states.

In addition to expected time period spent in the state, table II presents life expectancy for a person with given PSA level. The difference between these two times, presented in table II, is that the last includes time period spend in any not absorbing state. The life expectance decreases with increase of PSA level and equals 14.1 years for people with initial PSA level not exceeding 4ng/mL, 12.6 years for people with initial PSA level in range 4- 40ng/mL and 8.2 years for people with initial PSA level higher then 40ng/mL. The last means that high PSA level is attributed to loss of approximately 40% of life expectancy at advanced age.

Table III presents the composition of life expectancy by time, spent in different states. One can see that men, which had initial PSA level not exceeding 4ng/mL, spent in average 71 % of their life span having the same level of PSA, 21% with PSA level in range 4- 40ng/mL and 8% with PSA level higher then 40ng/mL. Men, which had initial PSA level in range 4-40ng/mL spent in average 55 % of their life span having lower level of PSA, 38% of the life span they have the same level of PSA, and 45% of the life span they have PSA level higher then 40ng/mL. Those men, which had initial PSA level higher then 40ng/mL spent in average half of their life span having level of PSA not exceeding 4ng/mL, 22% of the life span they have PSA level in range 4- 40ng/mL, and 28% of the life span they have PSA level higher then 40ng/mL.

IV. DISCUSSION

The data analyzed in this report are rather heterogeneous. The age at which a person enters the observation is between 39 and 89 years. The ages at which persons enter states with different levels of PSA are different as well. This means that estimated values of transition rates are not accurate and reflect only general trends in health state changes. Nevertheless these trends are meaningful. The mortality rate of death of other than prostate cancer is estimated as 0.02(1/year) and it does not depend on the level of PSA. The mortality rate of death of prostate cancer increases with increase of PSA level, which shows that level of PSA is the reliable indicator for progression of prostate cancer.

Future life expectancy among men with "normal" level of PSA at the beginning of the observation period is estimated to be 14.1 years while age adjusted future life expectancy for men of the same ages in Russia is approximated by 12.3 years. These calculations are based on the data provided by Human Mortality Database [5]. Such difference shows the necessity to incorporate age dependence in the rates of transitions to the absorbing terminal states. Future life expectancy among men with PSA level in range 4-40ng/mL at the beginning of the observation period is close to age adjusted future life expectancy for men of the same ages in Russia. This shows that increase in risk of death attributed with high level of PSA is close to increase in risk because of aging. For men with PSA level higher then 40ng/mL future life expectancy is approximately 30% lower than age adjusted future life expectancy for men of the same ages in Russia. This demonstrates that the high level of PSA is more "risky" for life than ageing.

REFERENCES

[1] P.K. Andersen and N. Keiding. "Multi-state models for event history analysis,". Stat Methods Med Res.; vol. 11, pp.: 91-115, 2002.

[2] H. Frydman and M. Szarek, "Nonparametric Estimationin a Markov "Illness–Death" Process from Interval Censored Observations with Missing Intermediate Transition Status,". Biometrics; vol. 65, pp.: 143-151, 2009.

[3] D.R. Cox and H.D. Miller, The theory of stohastic processes. London: Chapman and Hall, 1965.

[4] R. Sutradhar, L. Barbera, H. Seow, D. Howell, A. Husain, and D. Dudgeon, "Multistate Analysis of Interval-Censored Longitudinal Data: Application to a Cohort Study on Performance Status Among Patients Diagnosed With Cancer". Am. J. Epidemiol.; vol. 173, p. 468-475, 2011.

[5] Human Mortality Database. University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany). Available at www.mortality.org (data downloaded on 01.02.2012)

TABLE I. MLE ESTIMATES FOR TRANSITIONS INTENSITY MATRIX (1/YEAR)

	PSA level (ng/mL)		
	≤4	***4-40***	***>40***
PSA level ≤4 (ng/mL)	-0.27	0.49	---
PSA level 4-40 (ng/mL)	0.22	-0.75	0.34
PSA level >40 (ng/mL)	---	0.19	-0.64
Death of prostate cancer	0.02	0.05	0.27
Death of other cause	0.02	0.02	0.02

TABLE II. TIME TO STAY IN STATE AND LIFE EXPECTANCY (YEARS)

	PSA level (ng/mL)		
	≤4	***4-40***	***>40***
Expected time to stay in state	3.7	1.3	1.6
Life expectancy	14.1	12.6	8.2

TABLE III. EXPECTED LIFE SPAN, SPENT IN DIFFERENT STATES (YEARS)

	PSA level (ng/mL)		
	≤4	***4-40***	***>40***
PSA level ≤4 (ng/mL)	10.0 (71%)	6.9 (55%)	4.1 (50%)
PSA level 4-40 (ng/mL)	3.0 (21%)	4.8 (38%)	1.8 (22%)
PSA level >40 (ng/mL)	1.1 (8%)	0.9 (45%)	2.3 (28%)
TOTAL	14.1 (100%)	12.6 (100%)	8.2 (100%)

Regression modeling of the cumulative incidence function with missing causes of failure using pseudo-values

Margarita Moreno-Betancur
Inserm, CESP U1018, Biostatistics Team
Université Paris Sud 11, France
Email: margarita.moreno@inserm.fr

Aurelien Latouche
CNAM, Paris, France and
Inserm, CESP U1018, Biostatistics Team
Email: aurelien.latouche@cnam.fr

***Abstract*—Methods for estimating the effects of prognostic factors on the risk of death from a given cause in the competing risks setting often assume that the cause of death is known for all individuals. This is seldom the case. Exclusion of individuals with missing cause of death information might lead to biased and unprecise estimates. Some authors have proposed methods taking into account the mechanism leading to missing causes, particularly for modeling the cause-specific hazards. However, little attention has been given to direct modeling of the cumulative incidence function which is of prime interest with competing risks. This function expresses the probability of occurrence of a given event before a given time. We propose to extend the pseudo-value approach for modeling the cumulative incidence function to the missing cause setting, by considering inverse probability weighting and multiple imputation approaches. The proposed methods will serve to analyse the data from an ECOG breast cancer treatment clinical trial.**

I. Introduction

In a competing risks setting, patients may fail from several causes. For example, in a cancer study, the time to treatment failure is determined by the time to two competing events: relapse or death in remission (see Figure 1). More generally, suppose that the time T to treatment failure is determined by the time to any of the two competing events $\varepsilon \in \{1, 2\}$. In this situation, two identifiable quantities of main interest are the cause-specific hazard rates (CSH) defined by

$$\lambda_j(t) = \lim_{h \to 0} h^{-1} P(t \leq T < t + h, \varepsilon = j | T \geq t), \quad j = 1, 2$$

and the cumulative incidence functions (CIF) given by

$$F_j(t) = P(T \leq t, \varepsilon = j), \qquad j = 1, 2.$$

The CSH expresses the rate at which the event occurs and the CIF expresses the probability of occurrence of that event before a given time. These two quantities provide complementary information about the competing risks and there is often an interest to determine the effects of prognostic factors on them.

Regression modeling of the CSH and the CIF in the presence of right censoring was at first considered by (18) and (7), respectively, and has been extensively studied since. These approaches generally require that the cause of failure is known for all patients who have failed, a prerequisite which is seldom met in practice (1; 16). For example, some causes of failure might be missing in studies with a long follow-up because collection of information tends to deteriorate with time, in studies with good prognosis patients where death due to the disease is less likely or with elderly patients whose death is not easily attributed to only one cause.

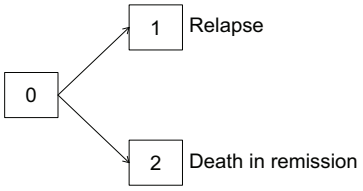


Fig. 1. Competing risks in a cancer study where patients may fail from relapse or death in remission.

In this missing data setting, we can use the typology of missing data introduced by Rubin (19) to characterize the mechanism leading to missing causes. The missingness mechanism is said to be MCAR (for *missing completely at random*) when the probability that a cause is missing is constant; MAR (for *missing at random*) when the probability that the cause is missing depends on covariates or failure time but not on the cause of failure when conditioning on these; and MNAR (for *missing not at random*) when missingness probability depends on the cause of failure even when conditioning on covariates and failure time. One can never test from the data whether the missingness mechanism is MAR or MNAR. Nevertheless, the CSH and CIF are not identifiable from the observed data when there are missing causes of failure without making assumptions about the missingness mechanism. Hence, this classification is useful to identify the assumptions underlying any modeling strategy in this setting, particularly those required to obtain unbiased estimates.

A common ad-hoc technique for modeling the CSH or the CIF when there are missing causes consists in performing

a *complete case* analysis, that is, an analysis excluding the individuals with missing cause of failure. Another approach is to *recode* all the missing causes as due to only one cause according to the observed distribution of the competing events. For example, one may assign all individuals with a missing cause the cause of interest if the disease under study is known to be very lethal, or a competing cause if not. To obtain unbiased estimates using these methods, very strong assumptions regarding the missingness mechanism must hold. For example, MCAR is required for the complete case estimator to be unbiased. Methods that require more relaxed and realistic missingness assumptions, acknowledging possible (observed or unobserved) relationships between the variables in the data and missingness probability, should be preferred.

Several authors have addressed the problem of modeling the cause-specific hazards when there are missing causes of failure (8; 9; 14; 15). A starting point for methodological development has been a stage II breast cancer treatment clinical trial from the Eastern Cooperative Oncology Group (ECOG) with 23% missing causes of failure among observed deaths (see 9; 14; 17). Since the CIF is a function of the CSH of each of the competing events, the latter methods provide an indirect way of estimating the CIF conditionally on covariates. For example, the authors of (12) combined the approach of (5) with the multiple imputation method of (14) to provide estimates of the CIF. Recently, estimation of the CIF based on estimates of other related quantities in this setting was considered in (17) through vertical modeling.

If the main interest is determining the effects of prognostic factors on the CIF, the methods based on estimates of the CSH are not suitable because (i) they require modeling the CSH of each of the competing events and (ii) covariate effects on the CIF will be hard to interpret since the CIF will typically be a complex non-linear function of these CSH's. Indeed the effect of a covariate on the CSH may be very different from the effect of the covariate on the corresponding CIF. Similar remarks apply to estimates obtained through vertical modeling. Thus, direct regression modeling of the CIF is desirable (7). However, in the missing cause setting this problem has only been addressed in (3) using multiple imputation in the case of the Fine and Gray model. No generic approach to this problem has yet been proposed.

We propose to extend the pseudo-value approach of (11) for modeling the CIF to the missing cause setting, by considering inverse probability weighting and multiple imputation approaches. In Section II we introduce the model and notation. In Section III we briefly describe the pseudo-value approach. The extension based on inverse probability weighting is presented in Section IV, and the multiple imputation method is described in Section V. In Section VI, we describe the ECOG data. Concluding remarks are presented in Section VII.

II. Notation and model

The observed data for each patient $i \in \{1, \ldots, n\}$ consist thus of the minimum $\tilde{T}_i$ of a failure time T_i and a censoring time C_i, i.e. $\tilde{T}_i = \min\{T_i, C_i\}$, and a status indicator ε_i such that $\varepsilon_i = 0$ if the observation is censored, $\varepsilon_i = 1$ for uncensored patients who failed from the cause of interest and $\varepsilon_i = 2$ for uncensored patients failing from a competing cause. Furthermore, the cause of failure is missing for some of the patients known to have failed. Thus for uncensored patients we observe an indicator M_i of whether the cause is known ($M_i = 1$) or not ($M_i = 0$). Finally, we let $Z_i = (Z_{i1}, \ldots, Z_{ip})$ be a fully observed p-vector of discrete finite-ranged covariates. We assume that data from different patients are independent and identically distributed (i.i.d.).

We model the CIF by means of a generalized linear model. That is, we assume that

$$g\{F_1(t|Z_i)\} = \beta_0(t) + \sum_{j=1}^{p} \beta_j Z_{ij}, t > 0, \qquad (1)$$

where g is a monotone link function, β_j is the effect of Z_{ij}, the j^{th} component of Z_i, and $\beta_0(t)$ is the baseline cumulative incidence function corresponding to individuals with $Z_{ij} = 0$ for all j. Model (1) encompasses models such as the Fine and Gray model if g is the complementary log-log (cloglog) function and the additive model if g is the identity function.

III. The pseudo-value approach

When the cause of failure is known for all patients who have failed, Klein and Andersen (11) proposed using the pseudo-value approach to fit model (1) (2). The approach consists in using pseudo-values from a jackknife statistic, constructed from the Aalen-Johansen estimator of the CIF, as the outcome variables. The properties of these pseudo-values enable the use of the generalized estimating equation (GEE) approach of (13) to estimate the regression coefficients of model (1). More precisely, the procedure is as follows:

1) Fix a grid of time-points $\tau_1, \ldots, \tau_S$ and let $\theta = (F_1(\tau_1), \ldots, F_1(\tau_S))$. We may take the grid to be the deciles or quintiles of the observed failure times.
2) Let $\hat{F}_1$ be the Aalen-Johansen estimator of F_1 and set $\hat{\theta} = (\hat{F}_1(\tau_1), \ldots, \hat{F}_1(\tau_S))$.
3) Calculate the pseudo-observation i which is given by:
$$\hat{\theta}_i = n\hat{\theta} - (n-1)\hat{\theta}^{-i}$$
where $\hat{\theta}^{-i}$ is obtained by excluding individual i from the sample.
4) Model (1) has the following representation in terms of pseudo-observations:
$$E(\hat{\theta}_i(\tau_s)|Z_i) = g^{-1}\left(\beta_0(\tau_s) + \sum_{j=1}^{p} \beta_j Z_{ij}\right)$$
for $i = 1, \ldots, n$ and $s = 1, \ldots, S$. Thus, setting the pseudo-observations as outcome variables for all individuals, GEE can be used to fit this model and obtain coefficients estimates.

The consistency and asymptotic normality of the estimates obtained through this procedure were established in (10, Theorem 2). The variance of the estimates can be consistently estimated using the usual sandwich estimator.

IV. Inverse probability weighted pseudo-values

We consider the following MAR-type assumption about the mechanism driving missingness:

$$P(M = 1|Z, T \leq t, \varepsilon) = P(M = 1|Z, T \leq t) =: \pi_t(Z).$$

That is, at each time t, the probability that the cause of failure is observed among individuals who have already failed is independent of the cause of failure when conditioning on covariates. Under this assumption, the following factorization holds:

$$\tilde{F}_1(t|Z) = \pi_t(Z) \times F_1(t|Z), \tag{2}$$

where $\tilde{F}_1(t|Z) = P(T \leq t, \varepsilon = 1, M = 1|Z)$. Thus, $F_1(t|Z)$ becomes identifiable as the quotient of two identifiable quantities, $\tilde{F}_1(t|Z)$ and $\pi_t(Z)$. Actually, $\tilde{F}_1$ is the CIF of cause 1 when treating a missing cause of failure as an additional competing event, i.e. $\tilde{F}_1(t|Z) = P(T \leq t, \tilde{\varepsilon} = 1|Z)$ where $\tilde{\varepsilon}$ is defined as follows: for censored individuals, $\tilde{\varepsilon} = 0$; for uncensored individuals, $\tilde{\varepsilon} = \varepsilon$ if $M = 1$ and $\tilde{\varepsilon} = 2$ otherwise. Hence, $\tilde{F}_1$ can be modeled using complete data methods. Since M is fully observed among individuals who have failed, $\pi_t(Z)$ can also be estimated from the data.

We propose an estimation procedure to fit model (1) motivated by factorization (2). It consists in using *inverse probability weighted pseudo-values* (IPWpv) as alternative outcomes for performing regression via GEE. More precisely, we consider the pseudo-values $\tilde{\theta}_i(t)$ corresponding to $\tilde{F}_1$, i.e. those obtained when setting the missing causes of failure as due to a competing event. These pseudo-values are then weighted by the inverse of the estimated probability of the cause being observed:

$$\hat{\theta}_i(t) = \tilde{\theta}_i(t)/\hat{\pi}_t(Z_i),$$

where $\hat{\pi}_t(Z_i)$ is the proportion of observed causes among individuals with $Z = Z_i$ who failed before t. Setting these IPWpv's as outcome variables for all patients, GEE can be used to fit model (1) as in the complete data pseudo-value procedure presented in Section III. IPWpv's have the same convenient properties as the ordinary pseudo-observations, and thus the estimates obtained following this procedure are also consistent and asymptotically normal (the proof is analogous to that of 10).

Nevertheless, the usual sandwich variance estimator is no longer suitable as it does not account for the variability of the estimated weights. Alternatively, a bootstrap procedure in which individuals are resampled from the original data can be used because it results in updated weights at each resample and thus leads to correct variance estimation.

V. Multiple imputation

Multiple imputation (20) was considered by (3) in the case of the Fine and Gray model. We propose a multiple imputation method tailored for the pseudo-value approach.

The first step is the imputation procedure of (3) and consists in producing $m > 1$ completed datasets (usually, $m = 5$ or 10 is enough), obtained by imputing each missing cause m times in a way that accounts for all levels of uncertainty. For this purpose, a prediction model, called the imputation model, is built for $\Pi(Z, T) := P(\varepsilon = 1|M = 0, \varepsilon > 0, Z, T)$. Assuming that the missingness mechanism is MAR, the probability that individuals failed from the cause of interest is the same for the complete and incomplete cases, conditionally on covariates and failure time, i.e. $\Pi(Z, T) = P(\varepsilon = 1|M = 1, \varepsilon > 0, Z, T)$. Therefore, a model for $\Pi(Z, T)$ can be constructed from the complete cases. To this end, one can use a logistic regression model of the form $\text{logit}\{\Pi(Z, T)\} = h(Z, T)'\gamma$, where $h(Z, T)$ is a vector including Z, T and their interaction, as suggested in (14) and (3). Once the imputation model is fitted to the complete cases, one may impute the missing causes by, (i) drawing a vector of parameters $\tilde{\gamma}$ of this model from their estimated asymptotic distribution, and (ii) drawing a cause of failure for each patient with a missing cause from the distribution posited by the imputation model when $\gamma = \tilde{\gamma}$. These two steps are repeated m times to obtain m completed datasets.

The second step is to fit the analysis model to each of the m completed datasets by applying an appropriate complete data method. In our case, we fit model (1) to each dataset by applying the pseudo-value approach described in Section III and obtain m estimates of the regression coefficients, $\hat{\beta}^{(1)}, \ldots, \hat{\beta}^{(m)}$, and m sandwich variance estimates, $\hat{\text{var}}(\hat{\beta})^{(1)}, \ldots, \hat{\text{var}}(\hat{\beta})^{(m)}$. These estimates can then be combined using the formulas of (20) to obtain the multiple imputation coefficient and variance estimates, $\hat{\beta} = \frac{1}{m}\sum_{l=1}^{m} \hat{\beta}^{(l)}$ and $\hat{\text{var}}(\hat{\beta}) = \hat{W} + \left(1 + m^{-1}\right)\hat{B}$, respectively, where $\hat{W}$ is the arithmetic mean of the variance estimates across imputations (the within-imputation component) and $\hat{B}$ is the sample variance of the m coefficient estimates (the between-imputation component).

To apply Rubin's formulas, some conditions must be met: the imputation procedure must be *proper*, the imputation model must be correctly specified and the imputation and analysis models should be compatible. If these conditions are met, the multiple imputation coefficient estimator is consistent, unbiased and asymptotically normal, and the variance estimator is unbiased. In practice these conditions may be easily violated but there are ways to improve coefficient estimates in that case and alternative variance estimators such as the bootstrap.

VI. The ECOG data example

The proposed methods will be used to analyse the data from the ECOG clinical trial on breast cancer treatment. The study involved 169 elderly women of which 79 had died at the time of analysis (53% censored). Because of their advanced age, these women were at high risk of death from other non-cancer competing events. Indeed, among the deceased patients, 44 had died from cancer whereas 17 had died from other causes. For the remaining 18 the cause of death was unknown (23% of deaths with missing cause). The authors of (6) reported the presence of 4 or more positive nodes and having an estrogen-receptor negative (ER-negative) primary tumor as two covariates that are significantly associated with

overall survival. We will conduct a regression analysis of the effects of these factors on the CIF of death from cancer to illustrate the practical value of the proposed methodology.

VII. Conclusion

We examine two extensions of the pseudo-value approach for regression modeling of the cumulative incidence function to the case where the cause of failure is missing for some patients. Asymptotic properties for these estimators are verified and variance estimators are suggested. Simulation studies are underway to evaluate the small-sample performance of the proposed estimators in terms of bias correction when compared to the complete case estimator under relaxed assumptions about the missingness mechanism. Preliminary results are promising. The ECOG data example will elucidate the practical value of these approaches.

The proposed methods are related to two paradigms of dealing with missing data, inverse probability weighting and multiple imputation, of which there exists several comparisons in the literature (see for example 4). When choosing between the two approaches, there is a trade-off between modeling missingness probability and the incomplete outcome. In our context, both require modeling a binary variable (M or ε, the latter being binary when restricted to uncensored patients) so the complexity of the task is comparable.

Acknowledgment

The authors thank the Eastern Cooperative Oncology Group and Professor Robert Gray for supplying the ECOG data.

References

[1] J. Andersen, E. Goetghebeur, and L. Ryan, "Missing cause of death information in the analysis of survival data," *Statistics in Medicine*, vol. 15, pp. 2191–2201, 1996.

[2] P. K. Andersen, J. P. Klein, and S. Rosthoj, "Generalised linear models for correlated pseudo-observations, with applications to multi-state models," *Biometrika*, vol. 90, pp. 15–27, 2003. [Online]. Available: http://biomet.oxfordjournals.org/content/90/1/15.abstract

[3] G. Bakoyannis, F. Siannis, and G. Touloumi, "Modelling competing risks data with missing cause of failure," *Statistics in Medicine*, vol. 29, pp. 3172–3185, 2010.

[4] J. R. Carpenter, M. G. Kenward, and S. Vansteelandt, "A comparison of multiple imputation and inverse probability weighting for analyses with missing data," *Journal of the Royal Statistical Society Series A (Statistics in Society)*, vol. 169, pp. 571–584, 2006.

[5] S. C. Cheng, J. P. Fine, and L. J. Wei, "Prediction of cumulative incidence function under the proportional hazards model," *Biometrics*, vol. 54, pp. 219–228, 1998.

[6] F. J. Cummings, R. Gray, T. E. Davis, D. C. Tormey, J. E. Harris, G. G. Falkson, and J. Arseneau, "Tamoxifen versus placebo: double-blind adjuvant trial in elderly women with stage II breast cancer." *NCI monographs : a publication of the National Cancer Institute*, pp. 119–123, 1986. [Online]. Available: http://view.ncbi.nlm.nih.gov/pubmed/3534584

[7] J. P. Fine and R. J. Gray, "A proportional hazards model for the subdistribution of a competing risk," *Journal of the American Statistical Association*, vol. 94, pp. 496–509, 1999.

[8] G. Z. Gao and A. A. Tsiatis, "Semiparametric estimators for the regression coefficients in the linear transformation competing risks model with missing cause of failure," *Biometrika*, vol. 92, pp. 875–891, 2005.

[9] E. Goetghebeur and L. Ryan, "Analysis of competing risks survival data when some failure types are missing," *Biometrika*, vol. 82, pp. 821–833, 1995.

[10] F. Graw, T. A. Gerds, and M. Schumacher, "On pseudo-values for regression analysis in competing risks models," *Lifetime Data Analysis*, vol. 15, pp. 241–255, 2009.

[11] J. P. Klein and P. K. Andersen, "Regression modeling of competing risks data based on pseudovalues of the cumulative incidence function," *Biometrics*, vol. 61, pp. 223–229, 2005.

[12] M. Lee, K. A. Cronin, M. H. Gail, J. J. Dignam, and E. J. Feuer, "Multiple imputation methods for inference on cumulative incidence with missing cause of failure," *Biometrical Journal*, vol. 53, pp. 974–993, 2011. [Online]. Available: http://dx.doi.org/10.1002/bimj.201000175

[13] K.-Y. Liang and S. L. Zeger, "Longitudinal data analysis using generalized linear models," *Biometrika*, vol. 73, pp. 13–22, 1986. [Online]. Available: http://biomet.oxfordjournals.org/content/73/1/13.abstract

[14] K. F. Lu and A. A. Tsiatis, "Multiple imputation methods for estimating regression coefficients in the competing risks model with missing cause of failure," *Biometrics*, vol. 57, pp. 1191–1197, 2001.

[15] W. Lu and Y. Liang, "Analysis of competing risks data with missing cause of failure under additive hazards model," *Statistica Sinica*, vol. 18, pp. 219–234, 2008.

[16] J. B. Manola and R. J. Gray, "When bad things happen to good studies," *Journal of Clinical Oncology*, vol. 29, pp. 3497–3499, 2011.

[17] M. A. Nicolaie, H. C. v. Houwelingen, and H. Putter, "Vertical modeling: Analysis of competing risks data with missing causes of failure," *Statistical Methods in Medical Research*, 2011, [Epub ahead of print].

[18] R. Prentice, J. Kalbfleisch, A. Peterson, N. Flournoy, V. Farewell, and N. Breslow, "Analysis of failure times in presence of competing risks," *Biometrics*, vol. 34, pp. 541–554, 1978.

[19] D. B. Rubin, "Inference and missing data," *Biometrika*, vol. 63, pp. 581–592, 1976.

[20] ——, *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons., 1987.

Non-parametric Tests for Comparing the Progression of an Epidemic

Uttara Naik-Nimbalkar
Department of Statistics
University of Pune
Pune, 411007, India
Email: uvnaik@stats.unipune.ac.in

***Abstract*—The severity of an epidemic is quantified by measures like the death rate, the cure rate and the case fatality rate (CFR). These rates may change with time as the treatment of the disease and management procedures improve. To capture the time-varying nature of the CFR, Yip et al. [12] introduced the concept of the real-time fatality rate. The fatality rates may differ in different populations due to some factors like the etiological characteristics, geographical conditions or the treatment and intervention procedures. In this talk, we consider some measures related to the real-time fatality rate for this comparison. We propose two-sample tests, one for the proportionality of the crude transition death rates and one for the equality of the crude reverse hazard rates and derive the asymptotic properties of the test statistics. We propose sequential tests that can be carried out as the epidemic progresses.**

I. Introduction

During an outbreak of an emerging infectious disease such as the severe acute respiratory syndrome(SARS) or influenza A(H1N1), one of one of the most important epidemiologic quantities to be determined is the case fatality rate (CFR). The CFR is the fraction of cases that die from the disease over a certain period of time and provides a measure of severity of the epidemic. The higher the CFR, the more virulent is the virus. Estimation of the CFR has been considered by many authors. For some recent references we refer to Ghani et al. [4], Jewell et al. [5], Chen et al. [3], Reich et al. [10] and references therein. Yip et al. [12] showed that the CFR fails to capture the time varying nature of the severity of the epidemic and proposed the 'real-time fatality rate' based on a counting process in a competing risks set-up. The real-time fatality rate is the probability of death at time t conditioned on a transition to death or recovery at time t. An estimator for the real-time fatality rate via a counting process and kernel function is given in Yip et al. [12] and an estimator using the chain multinomial model is considered by Yip et al. [13].

The fatality of an outbreak of an infectious disease should decrease with time if proper intervention policies to control the epidemic are implemented. In order to plan the allocation of resources that may be limited or costly or to assess the intervention policies, it is of interest to compare the progress of the epidemic in two different groups. Comparison of the CFRs can also help in assessing the treatment effects in the absence of randomized clinical trials, which are not possible during a severe outbreak. Implementing clinical treatments that are specific to the groups, formed according to say gender or region may be more efficient than giving the same treatment to all. Peiris et al. [9] and Karlberg et al. [6] discuss this situation for the SARS epidemic of 2003. Reich et al. [10] consider estimation of the relative CFR, which is defined as the CFR of one group divided by that of a reference group. The CFR's in different groups can be compared by including covariates in a generalized linear model framework. This framework has been discussed in Chen et al. [3], Reich et al. [10] but it is assumed that the CFR is constant. Ongoing comparisons of the fatality are one way of knowing how the different groups change as the epidemic progresses. If at any time, the fatality for specific disease exceeds the reference or the ideal rate, an immediate evaluation of the current control and treatment efforts should be undertaken.

Lam et al. [7] propose a test for testing the null hypothesis of constant real-time fatality rate against the alternative hypothesis of decreasing fatality rate. Their test has the ability to reject the null hypothesis as soon as there is enough evidence of a decreasing fatality as information accumulates over time. Xu [11] proposes a test for equality of the fatality rate using the Multinomial model assuming constant death and recovery rates in a given time interval.

In this article we propose tests based on two measures related to the real-time fatality rate for the comparison of the progression of an epidemic. The assumption of constant fatality is not required for our procedure. The tests proposed are sequential in nature and can be carried out during the progression of the epidemic. The tests are based on data which are readily available as the epidemic progresses.

A. Quantitative measures for comparing the progress of an epidemic

Suppose a healthy individual gets infected with a certain infectious disease and is admitted to a hospital at a rate $\gamma_0(t)$ and then either dies at a rate $\gamma_1(t)$ (crude transition death rate), or recovers at a rate $\gamma_2(t)$ (crude transition recovery rate) where t is the calendar time since the beginning of the epidemic.

The real-time fatality rate (hereafter referred to as the fatality rate) proposed by Yip et al. [12] is:

$$\pi(t) = \frac{\gamma_1(t)}{\gamma_1(t) + \gamma_2(t)}.$$

It as the probability of death conditioned on a transition to death or recovery at time t.

Let $\Gamma_{\cdot}(t) = \int_0^t (\gamma_1(s) + \gamma_2(s))ds$, denote the cumulative intensity. The following measure and a test based on it is discussed in Lam et al [8].

Definition:'Crude Reverse Hazard Rate' $v(t)$ of an epidemic is defined as:

$$v(t) = \frac{\gamma_1(t)}{\Gamma_{\cdot}(t)}.$$

This measure indicates the contribution of the fatalities at time t to the total accumulation of fatalities and discharges up to time t and is analogous to the crude reverse hazard rate function due to one of two risks in the competing risk scenario.

For group i, let the death rate, the recovery rate, the cumulative intensity and the crude reverse hazard rate, respectively be denoted by $\gamma_{i1}(t),\ \gamma_{i2}(t),\ \Gamma_{i\cdot}(t)$ and $v_i(t)$. We note the following relations between these rates. In what follows t denotes the calendar time since the beginning of the epidemic and τ is a pre-specified time.

Lemma 1. If $\gamma_{11}(t) = \alpha\gamma_{21}(t)$ for some constant α, then $\pi_1(t) = \pi_2(t)$ if and only if $\gamma_{12}(t) = \alpha\gamma_{22}(t)$. That is if the death rates in the two groups are proportional then the real-time fatality rates of the two groups are equal iff the recovery rates are proportional with the same constant of proportionality.

Lemma 2. Suppose $v_1(t) = v_2(t)$ and $\pi_1(t) = \pi_2(t)$ for all $t,\ 0 \le t \le \tau$. Further suppose that $\gamma_{11}(0)$ and $\gamma_{21}(0)$ are both not equal to zero. Then $\gamma_{11}(t) = \alpha\gamma_{21}(t)$ for $0 \le t \le \tau$ and with $\alpha = \gamma_{11}(0)/\gamma_{21}(0)$.

The assumption $\gamma_{11}(0)$ and $\gamma_{21}(0)$ are both not equal to zero is reasonable, if we start measuring time as soon as there is at least one death in each of the two groups.

Lemma 3. If $\gamma_{11}(t) = \alpha\gamma_{21}(t)$ for some constant α and $\pi_1(t) = \pi_2(t)$ for $0 \le t \le \tau$, then $v_1(t) = v_2(t)$, $0 \le t \le \tau$.

In the next section we propose two-sample tests for testing the proportionality of the crude transition death rates and the equality of the crude reversed hazard rates in the two groups. Both the tests are rank tests and only the order of occurrences of the events is required to compute the test statistics. The exact times of occurrences of the events, which in practice may be unobservable are not required. Both the tests are sequential and based on data that become readily available as the epidemic progresses.

In view of the above Lemmas, the proposed tests can help to assess whether the fatality rates within the two groups are equal.

B. Two-sample Tests for comparing the progress of an epidemic

Consider two distinct groups, classified according to gender, age, location or professional status. Suppose each group consists of n_i (i = 1,2) initially healthy individuals subject to infection during an epidemic. When an individual is infected, he/she will experience an incubation period, which is disease specific, after which he/she will exhibit the symptoms of the infectious disease and be admitted to a hospital to receive treatment. We assume that for population i each of the n_i healthy individuals follows a Markov chain with states: healthy, inpatient, dead or recovery (healthy) and with the respective transition rates $\gamma_{i0}(t),\ \gamma_{i1}(t)$ and $\gamma_{i2}(t)$ at time t. Here t denotes the calendar time as stated in the earlier section.

The data consist of the n_i independent Markov chains for $i = 1, 2$ and we can form the following counting processes,
$N_{i1}(t)$ = the cumulative number of deaths up to time t for population i,
$N_{i2}(t)$ = the cumulative number of recoveries up to time t for population i
and
$I_i(t)$ = the number of inpatients just before time t for population i.

We further assume that the filtration

$$\mathcal{F}_t = \{I_i(s),\ N_{ij}(s),\ j = 1, 2,\ i = 1, 2,\ 0 \le s \le t\}$$

satisfies the usual regularity conditions and no two events occur simultaneously. Define

$$M_{ij}(t) = N_{ij}(t) - \int_0^t \gamma_{ij}(s)I_i(s)ds, \quad j = 1, 2,\ i = 1, 2. \quad (1)$$

These are zero mean martingales w.r.t. the filtration $\{\mathcal{F}_t,\ t \ge 0\}$, (Andersen, Borgan, Gill, and Keiding [1], pp.94). Also, define $N_{i\cdot}(s) = N_{i1}(s) + N_{i2}(s)$, $J_i(s) = I\{I_i(s) > 0\}$, $\Gamma_{ij}(t) = \int_0^t \gamma_{ij}(t)dt$ and $\Gamma_{i\cdot}(t) = \int_0^t (\gamma_{i1}(s) + \gamma_{i2}(s))ds$, where $I\{A\}$ denotes the indicator function of the event in the braces.

A consistent estimator of $\Gamma_{ij}(t)$ is then given by the Nelson - Aalen estimator

$$\hat{\Gamma}_{ij}(t) = \int_0^t \frac{J_i(s)}{I_i(s)} dN_{ij}(s). \quad (2)$$

1) Test 1: A test for proportionality of crude transition death rates: The crude instantaneous transition rates in the two populations could be used to compare the progress of the epidemics. We consider the null hypothesis,

$$H_0^{(1)} : \frac{\gamma_{11}(t)}{\gamma_{21}(t)} = \alpha, \text{for all } \ 0 \le \mathrm{t} \le \tau,$$

for some unknown constant α against the alternative against

$$H_1^{(1)} : \frac{\gamma_{11}(t)}{\gamma_{21}(t)} \uparrow \text{ in } t, \qquad 0 \le t \le \tau.$$

Proportionality of the crude death rates means that the death rate in the first group does not change relative to that of the second group. Whereas if the above ratio shows a monotonic increasing trend, then the deaths are more frequent in the first group compared to the second, which indicates that the control measures should be increased for the first group.

We base the test on the statistic:

$$W_n(t) = \frac{1}{\sqrt{n}}\left[\int_0^\tau \frac{J(s)N_{21}(s-)}{I(s)}dN_{11}(s) - \int_0^\tau \frac{J(s)N_{11}(s-)}{I(s)}dN_{21}(s)\right], \quad (3)$$

where $I(s) = I_1(s) + I_2(s)$, $J(s) = I\{(s) > 0\}$ and $n = n_1 + n_2$. Define

Define $\Lambda_{i1}(t) = \int_0^t \gamma_{i1}(s)I(s)ds$. Using integration by parts, the test statistic (3) may be rewritten as

$$\begin{aligned} W_n(t) &= \frac{1}{\sqrt{n}}[\int_0^t \{\frac{J(s)N_{21}(s-)}{I(s)} \\ &\quad -(\Gamma_{21}(t)-\Gamma_{21}(s))\}dM_{11}(s) \\ &\quad -\int_0^t \{\frac{J(s)N_{11}(s-)}{I(s)} \\ &\quad -(\Gamma_{11}(t)-\Gamma_{11}(s))\}dM_{21}(s)] \\ &\quad +\frac{1}{\sqrt{n}}[\int_0^t \Lambda_{21}(s)\gamma_{11}(s)ds \\ &\quad -\int_0^t \Lambda_{11}(s)\gamma_{21}(s)ds]. \end{aligned} \quad (4)$$

The first two integrals in (4) have zero mean and the last term in square bracket is 0 under $H_0^{(1)}$ and is strictly positive under $H_1^{(1)}$. Hence, the proposed test statistic is consistent against any increasing ratio. Thus the null hypothesis is rejected for large values of the test statistic.

2) Test 2: Test for equality of 'crude reverse hazard rates': Let $v_i(t) = \gamma_{i1}(t)/\Gamma_{i\cdot}(t), \quad i = 1, 2$. We next propose a test for the null hypothesis

$$H_0^{(2)} : v_1(t) = v_2(t), \quad \text{for all} \quad 0 \le \mathrm{t} \le \tau,$$

against

$$H_1^{(2)} : v_1(t) \ge v_2(t), \ 0 \le t \le \tau, \quad \text{with strict inequality on some interval.}$$

Consider the statistic

$$Z_n(t) = \sqrt{n}\left[\int_0^t \frac{\hat{\Gamma}_{2\cdot}(s)J_1(s)}{I_1(s)}dN_{11}(s) - \int_0^t \frac{\hat{\Gamma}_{1\cdot}(s)J_2(s)}{I_2(s)}dN_{21}(s)\right]. \quad (5)$$

Using integration by parts, the above statistic may be rewritten as

$$\begin{aligned} Z_n(t) &= \sqrt{n}\left[\int_0^t \left(\frac{\hat{\Gamma}_{2\cdot}(s)J_1(s)}{I_1(s)} - \frac{(\Gamma_{21}(t)-\Gamma_{21}(s))J_1(s)}{I_1(s)}dM_{11}(s)\right)\right. \\ &\quad -\int_0^t \frac{(\Gamma_{21}(t)-\Gamma_{21}(s))J_1(s)}{I_1(s)}dM_{12}(s) \\ &\quad -\int_0^t \left(\frac{\hat{\Gamma}_{1\cdot}(s)J_2(s)}{I_2(s)} - \frac{(\Gamma_{11}(t)-\Gamma_{11}(s))J_2(s)}{I_2(s)}dM_{21}(s)\right) \\ &\quad \left.+\int_0^t \frac{(\Gamma_{11}(t)-\Gamma_{11}(s))J_2(s)}{I_2(s)}dM_{22}(s)\right] \\ &\quad +\sqrt{n}\left[\int_0^t \Gamma_{2\cdot}(s)\gamma_{11}(s)ds - \int_0^t \Gamma_{1\cdot}(s)\gamma_{21}(s)ds\right]. \end{aligned}$$

The first four integrals in the above expression have mean zero always, whereas the last term above is 0 under $H_0^{(2)}$ and is strictly positive under $H_1^{(2)}$. This property leads to the consistency of the test for $H_1^{(2)}$. The null hypothesis is rejected for large values of the test statistic.

3) Asymptotic distributions: The asymptotic distribution of both the test statistics is obtained using the Rebolledo martingale central limit theorem (Andersen et al. [1], pp. 83). We assume the following:

Assumption 1:

(i) As $n \to \infty$, $n_i/n \to \lambda_i$, with $\lambda_i > 0$, $i = 1, 2$,
(ii) $\lim_{n\to\infty} \frac{N_{11}s-)}{n} = p_{11}(s)$ in probability,
(iii) $\lim_{n\to\infty} \frac{N_{21}(s-)}{n} = p_{21}(s)$ in probability,
(iv) $\lim_{n\to\infty} \frac{I(s)}{n} = p_I(s) > 0$ in probability,
(v) $\lim_{n\to\infty} \frac{I_1(s)J_1(s)}{n_1} = p_1(s) > 0$, in probability,
(vi) $\lim_{n\to\infty} \frac{I_2(s)J_2(s)}{n_2} = p_2(s) > 0$, in probability,

where (ii) to (vi) hold for all s, $0 < s < \tau$.

Below we give the asymptotic variances and their estimators for the two statistics. Let

$$\begin{aligned} &\sigma_1^2(t) \\ &= \int_0^t \left(\frac{p_{21}(s)}{p_I(s)} - (\Gamma_{21}(\tau)-\Gamma_{21}(s))\right)^2 p_I(s)\gamma_{11}(s)ds \\ &+ \int_0^t \left(\frac{p_{11}(s)}{p_I(s)} - (\Gamma_{11}(\tau)-\Gamma_{11}(s))\right)^2 p_I(s)\gamma_{21}(s)ds, \end{aligned}$$

$$\begin{aligned} &\widehat{\sigma}_{1n}^2(t) \\ &= \int_0^t \left(\frac{J(s)N_{21}(s-)}{I(s)} - (\hat{\Gamma}_{21}(t)-\hat{\Gamma}_{21}(s))\right)^2 dN_{11}(s) \\ &+ \int_0^t \left(\frac{J(s)N_{11}(s-)}{I(s)} - (\hat{\Gamma}_{11}(t)-\hat{\Gamma}_{11}(s))\right)^2 dN_{21}(s), \end{aligned}$$

$$\begin{aligned}\sigma_2^2(t) &= \int_0^t \frac{(\Gamma_{2\cdot}(s)-(\Gamma_{21}(t)-\Gamma_{21}(s))^2}{\lambda_1\ p_1(s)}\gamma_{11}(s)ds \\ &+ \int_0^t \frac{(\Gamma_{21}(t)-\Gamma_{21}(s))^2}{\lambda_1\ p_1(s)}\gamma_{12}(s)ds \\ &+ \int_0^t \frac{(\Gamma_{1\cdot}(s)-(\Gamma_{11}(t)-\Gamma_{11}(s))^2}{\lambda_2\ p_2(s)}\gamma_{21}(s)ds \\ &+ \int_0^t \frac{(\Gamma_{11}(t)-\Gamma_{11}(s))^2}{\lambda_2\ p_2(s)}\gamma_{22}(s)ds,\end{aligned}$$

and let

$$\begin{aligned}\hat{\sigma}_{2n}^2(t) &= n\left[\int_0^t \left(\frac{\hat{\Gamma}_{2\cdot}(s)-(\hat{\Gamma}_{21}(t)-\hat{\Gamma}_{21}(s))}{I_1(s)}\right)^2 dN_{11}(s)\right. \\ &+ \int_0^t \left(\frac{\hat{\Gamma}_{21}(t)-\hat{\Gamma}_{21}(s)}{I_1(s)}\right)^2 dN_{12}(s) \\ &+ \int_0^t \left(\frac{\hat{\Gamma}_{1\cdot}(s)-(\hat{\Gamma}_{11}(t)-\hat{\Gamma}_{11}(s)}{I_2(s)}\right)^2 dN_{21}(s) \\ &\left.+ \int_0^t \left(\frac{\hat{\Gamma}_{11}(t)-\hat{\Gamma}_{11}(s)}{I_2(s)}\right)^2 dN_{22}(s)\right],\end{aligned}$$

where $\hat{\Gamma}_{ij}(t)$ is the Nelson-Aalen estimator given in (2) and the p_{11}, p_{21}, p_I, p_1 and p_2 are as in Assumption 1.

Using the Rebolledo martingale central limit theorem Andersen et al. [1], pp. 83) we obtain the following result.

Theorem 1.: If the Assumption 1 holds, then as $n \to \infty$,
(i) Under $H_0^{(1)}$, the process $W_n = \{W_n(t), \quad 0 \le t \le \tau\}$ converges weakly to a process $V = \{V(t), \quad 0 \le t \le \tau\}$ in $D[0,\tau]$ with the Skorohod topology, where $\{V(t), \quad 0 \le t \le \tau\}$ is a continuous mean-zero Gaussian martingale and $cov(V(s),V(t)) = \sigma_1^2(min(t,s))$, and $\sup_{t\in[0,\tau]} |\hat{\sigma}_{1n}^2(t) - \sigma_1^2(t)| \to 0$ in probability,
(iii) Under $H_0^{(2)}$, the process $Z_n = \{Z_n(t), \quad 0 \le t \le \tau\}$ converges weakly to a process $U = \{U(t), \quad 0 \le t \le \tau\}$ in $D[0,\tau]$ with the Skorohod topology, where $\{U(t), \quad 0 \le t \le \tau\}$ is a continuous mean-zero Gaussian martingale and $cov(U(s),U(t)) = \sigma_2^2(min(t,s))$, and $sup_{t\in[0,\tau]}|\hat{\sigma}_{2n}^2(t) - \sigma_2^2(t)| \to 0$ in probability.

If the test is to be conducted at time τ, and for level α the critical region for Test 1 is given by $W_n(\tau)/\hat{\sigma}_{1n}(\tau) > z_\alpha$, where z_α is the upper $(1-\alpha)\times 100 - th$ percentile of the standard normal distribution. Similarly the critical region for the Test 2 is given by $Z_n(\tau)/\hat{\sigma}_{3n}(\tau) > z_\alpha$, where z_α is the upper $(1-\alpha)\times 100 - th$ percentile of the standard normal distribution.

In view of the Lemmas stated in the previous section we propose that the null hypothesis of equality of the fatality rates of the two groups, that is $\pi_1(t) = \pi_2(t)$ is not rejected if the null hypotheses $H_0^{(1)}$ and $H_0^{(2)}$ are not rejected. If either of the two hypotheses $H_0^{(1)}$ or $H_0^{(2)}$ are rejected, the preventive measures for Group 1 should be increased.

It is important to know whether the epidemic is progressing faster in a given group so that the appropriate intervention measures can be taken and thus it should be possible to reject the null hypothesis at an early epoch than having to wait up to a given time τ. In order for this to be possible, we propose sequential test procedures similar to the one discussed in Lam et al. [7]. The sequential version of the tests will enable us to reject the null hypotheses when the ratio of the death rates decreases in some sub-interval of $(0,\tau)$ or the functions $v_1(t)$ and $v_2(t)$ cross at time τ_1 $(< \ \tau)$. The result stated below is helpful in computing the critical regions and handling the significance level of tests carried out at different time points.

Let $g_n(s) = inf\{t \ge 0 | \hat{\sigma}_{1n}^2(t) \ge s\}$, $h_n(s) = inf\{t \ge 0 | \hat{\sigma}_{2n}^2(t) \ge s\}$, $g(s) = inf\{t \ge 0 | \sigma_1^2(t) \ge s\}$ and $h(s) = inf\{t \ge 0 | \sigma_2^2(t) \ge s\}$. We note that g and h are nondecreasing, continuous, non-random functions on $[0, \sigma_1^2(\tau)]$ and $[0, \sigma_2^2(\tau)]$, respectively.

From the Theorem 1 above and using results from Billingsley ([2], Theorem 3.9 and an argument on P. 151; see also Lam et al. [7]) we obtain the following result.

Theorem 2: If the Assumption 1 holds, then as $n \to \infty$
(i) under $H_0^{(1)}$, the process $W_n(g_n) = \{W_n(g_n(s)), \quad 0 \le s \le \sigma_1^2(\tau)\}$ converges weakly to a Brownian Motion,
(ii) under $H_0^{(2)}$, the process $Z_n(h_n) = \{Z_n(h_n(s)), \quad 0 \le s \le \sigma_2^2(\tau)\}$ converges weakly to a Brownian Motion.

If the process is observed on a daily basis, in order to carry out the sequential procedure for Test 1, define a partition $0 \le s_0 \le s_1 \le \cdots \le s_\tau$ where $s_t = \hat{\sigma}_{1n}^2(t)$ if $\hat{\sigma}_{1n}^2(t) > \hat{\sigma}_{1n}^2(t-1)$ or else $s_t = \hat{\sigma}_{1n}^2(t-1)$ for $t = 1, \cdots, \tau$. A sequential test then can be based on $\{W_n(g_n(s_1)), \cdots, W_n(g_n(s_\tau))\}$. First compute the statistic $W_n(t)$ at the end of the t-th day and reject the null hypothesis $H_0^{(1)}$ at the end of day t if $W_n(t) > c_t$. The $c_t's$, $t = 1, \cdots, \tau$ are obtained using the above Theorem 2 so that the overall significance level is α. We refer to Lam et al. [7] for the details. Similarly for testing the null hypothesis $H_0^{(2}$ at the end of day t a sequential test can be based on $\{Z_n(h_n(s_1)), \cdots, Z_n(h_n(s_\tau))\}$. For this test, compute the statistic $Z_n(t)$ at the end of the t-th day and reject the null hypothesis at the end of day t if $Z_n(t) > a_t$. The $a_t's$, $t = 1, \cdots, \tau$ are obtained using the Theorem 2 so that the overall significance level is α.

Remarks: Technically, for the asymptotic results to hold, we do require that no two events occur simultaneously and the processes $N_{ij}(t)$ and $I_i(t)$ $(i = 1,2; j = 1,2)$ are observed continuously. However, in practice, the data are usually collected on a daily basis therefore, the exact transition times are generally unobservable. In such situation, we propose to pool all the deaths and recoveries within the day for two

populations together and then randomly impute the death and recovery times, or simply impute the order of deaths and recoveries on that day. This method works because each test is a rank test and the test statistic only requires the order of occurrences of the events rather than the exact time of the occurrences of events.

In the talk we will report some simulation results and an application of the tests.

II. Conclusion

In this article we have proposed a testing procedure to investigate the difference in the progress of an epidemic in two groups, where one group (namely the Group 2) can be considered as the reference group. The procedure consists of two tests based on measures related with the real-time fatality rate. When either of the test rejects the null hypothesis, we conclude that the epidemic in Group 1 is more severe and control measures for this group should be increased. The tests proposed are sequential in nature and can be implemented as the data accumulate.

Further the test which is carried out at the end of the epidemic may not be able to discriminate between the rates which intersect. However a sequential test should be able to detect these alternatives as well.

References

[1] Andersen, P. K., Borgan, Ø., Gill, R. D., and Keiding, N. (1993). *Statistical models based on counting processes*. New York: Springer.

[2] Billingsley, P. (1999). *Convergence of Probability Measures*. 2nd edition. New York: Wiley.

[3] Chen Z., Akazawa, K. and Nakamura T. (2009). Estimating the case fatality rate using a constant cure-death hazard ratio. *Lifetime Data Anal* **15** 316329.

[4] Ghani, A. C., Donnelly, C. A., Cox, D. R., Griffin, J. T., Fraser C., Lam, T. H., Ho, L. M., Chan, W. S., Anderson, R. M., Hedley, A. J. and Leung, G. M. (2005). Methods for Estimating the Case Fatality Ratio for a Novel, Emerging Infectious Disease *American J. of Epidemiology* **162**, 479-486.

[5] Jewell, N. P., Lei, X. D., Ghani, A. C., Donnelly, A. C., Leung, G. M., Ho,, L. M., Cowling,B. J. and Hedley, A. J. (2007) Non-parametric estimation of the case fatality ratio with competing risks data: an application to severe acute respiratory syndrome (SARS). *Stat. Med.* **26** 19821998.

[6] Karlberg, J.,Chong, D. S.Y., Lai, W. Y. Y. (2004). Do men have a higher case fatality rate of severe acute respiratory syndrome than women do? *American J. of Epidemiology* **159**, 229-31.

[7] Lam, K. F., Deshpande, J. V., Lau, E. H. Y., Naik-Nimbalkar, U. V., Yip, P. S. F. and Xu Ying. (2009).A Test for Constant Fatality Rate of an Emerging Epidemic: With Applications to Severe Acute Respiratory Syndrome in Hong Kong and Beijing. *Biometrics* **64**, 869876.

[8] Lam, K. F., Deshpande, J. V., Naik-Nimbalkar, U. V. and Tina Xu. (2011). On Comparing Fatality of an epidemic in Different Compartments. Working Paper.

[9] Peiris, J. S. M., Lai, S. T., Poon, L. L. M. et al(2003). Coronavirus as a possible cause of severe acute respiratory syndrome. **Lancet 361**, 1319-25.

[10] Reich, N. G., Lessler, J., Cummings, D. A. T. and Brookmeyer R. (2012). Estimating Absolute and Relative Case Fatality Ratios from Infectious Disease Surveillance Data. *Biometrics,* DOI: 10.1111/j.1541-0420.2011.01709.x.

[11] Xu Ying, (2009). *Statistical Analysis of the Infectivity and Fatality of an Emerging Epidemic.* Doctoral Thesis, University of Hong Kong.

[12] Yip, P. S. F., Lam, K. F., Lau, E. H. Y., Chau, P. H., Chao, A. and Tsang, K. W. (2005a). A comparison study of realtime fatality rates: severe acute respiratory syndrome in Hong Kong, Singapore, Taiwan, Toronto and Beijing, China. *J. Roy. Statist. Soc. Ser A* **168**, 233-43.

[13] Yip, P. S. F., Lau, E. H. Y., Lam, K. F., and Huggins, R. M. (2005b). A chain multinomial model for estimating the realtime fatality rate of a disease, with an application to severe acute respiratory syndrome. *American J. of Epidemiology* **161**, 700-6.

Lifetime Density and Failure Rate Estimation

Sarah Ouadah

Abstract—We provide laws of the iterated logarithm for the lifetime density and failure rate estimators, in a right-censorship model. These results are uniform in both bandwidth and kernel and are established in the framework of convergence in probability, in the complete range for which the estimators are consistent.

1 Introduction and Results

Let $X, X_1, X_2, \ldots$ be independent and identically distributed [iid] nonnegative lifetimes, having common distribution function [df] $F(\cdot) := \mathbb{P}(X \leq \cdot)$ and density $f(\cdot) := \frac{\partial}{\partial \cdot} F(\cdot)$ continuous and positive on $J := [A, B] \subseteq \mathbb{R}$. Denote by $C, C_1, C_2, \ldots$ the iid nonnegative censoring times having df $G(\cdot) := \mathbb{P}(C \leq \cdot)$. Let S_F (resp. S_G) be the upper endpoint of F (resp. G) defined by $S_F := \sup\{x : F(x) < 1\}$ (resp. S_G), and fix $[A, B] \subseteq [0, \Theta]$, with $\Theta := \min(S_F, S_G) > 0$. We assume that F, G are such that $F(0) = G(0) = 0$ and are continuous on J. We suppose X right-censored by C, in an independant way. The data set is given by $\{(T_i, \delta_i) : 1 \leq i \leq n\}$, where

$$\begin{cases} T_i = X_i \wedge C_i, \\ \delta_i = \mathbb{1}_{X_i \leq C_i}, \end{cases}$$

with $\mathbb{1}_E$ denoting the indicator function of E and T having df $H(\cdot) := \mathbb{P}(T \leq \cdot) = 1 - (1 - F(\cdot))(1 - G(\cdot))$.

The nonparametric maximum likelihood estimator of $F(\cdot)$ (*product-limit* estimator) based upon the data set is due to Kaplan and Meier [9] and is defined, for $x \in \mathbb{R}$, by

$$F_n(x) := 1 - \prod_{\substack{T_{i,n} \leq x \\ 1 \leq i \leq n}} \left\{ 1 - \frac{\delta_{i,n}}{n - i + 1} \right\} \qquad (1)$$

where, for all $n \geq 1$, $T_{1,n} \leq \ldots \leq T_{n,n}$ are the ordered $T_1, \ldots, T_n$, and, for each $i = 1, \ldots, n$, $\delta_{i,n}$ is the δ_j corresponding to $T_{i,n} = T_j, 1 \leq j \leq n$. Throughout, ψ will denote a specified continuous and positive function on J. We assume that ψ_n is an estimator of ψ such as, we have, as $n \to \infty$,

$$\sup_{x \in I} |\psi_n(x)/\psi(x) - 1| \to 0 \text{ in probability},$$

with $I := [C, D] \subseteq J$.

1.1 Uniform in Bandwidth and Kernel Lifetime Density Estimation.

Consider the right censorship model previously introduced. Assume that the lifetime density $f(\cdot)$ of X is defined and continuous on J. Let $\mathcal{K}$ denote the set of all right-continuous distributions functions of totally bounded signed measures $K(dt)$ on $\mathbb{R}$, of the form $K(dt) = K_+(dt) - K_-(dt)$, where $K_+(dt)$ and $K_-(dt)$ denote two orthogonal positive bounded Radon measures on $\mathbb{R}$. Denote the total variation of $K(dt)$, by

$$\|dK\| = \int_{\mathbb{R}} (K_+(dt) + K_-(dt)).$$

For all $K \in \mathcal{K}$, there exists $\kappa > 0$ such that $\|dK\| \leq \kappa$. The kernel estimator of $f(x)$ (see, e.g., Watson and Leadbetter [14], [15], Tanner and Wong [12])) is defined, for $K \in \mathcal{K}$, $h > 0$ and $x \in \mathbb{R}$, by

$$f_{n,K,h}(x) := \frac{1}{h} \int_{-\infty}^{\infty} K\left(\frac{x - t}{h}\right) dF_n(t). \qquad (2)$$

Our next theorem, describes the *uniform in bandwidth and kernel* consistency of $f_{n,K,h}(\cdot)$.

• *S. Ouadah, L.S.T.A., Université Pierre et Marie Curie (Paris 6), T15-25, E2, 4 Place Jussieu, 75252 Paris Cedex 05, France. E-mail: sarah.ouadah@upmc.fr*

Theorem 1: Let $0 < a_n \leq b_n \leq 1$ be such that, as $n \to \infty$,

$$b_n \to 0 \quad \text{and} \quad \frac{na_n}{\log n} \to \infty. \tag{3}$$

Then, with $\mathcal{H}_n = [a_n, b_n]$, we have, as $n \to \infty$,

$$\sup_{h\in\mathcal{H}_n} \left| \sup_{K\in\mathcal{K}} \left| \left\{ \frac{nh}{2\log_+(1/h)} \right\}^{1/2} \sup_{x\in I} \pm A_{n,K,h}(x) \right.\right.$$
$$\left.\left. \times \left\{ \psi_n(x) \times \frac{1-G(x)}{f(x)} \right\}^{1/2} - \sigma(\psi, K) \right| \right| = o_{\mathbb{P}}(1), \tag{4}$$

with

$$A_{n,K,h}(x) := f_{n,K,h}(x) - \mathbb{E}\left(f_{n,K,h}(x)\right),$$
$$\text{and } \sigma(\psi, K) := \left\{ \sup_{x\in I} \psi(x) \int_{\mathbb{R}} K^2(t)dt \right\}^{1/2}.$$

Remark 1: 1°) In the uncensored case, where $G \equiv 0$, $\psi \equiv 1$ and where in addition X follows the uniform distribution on $[0, 1]$, Theorem 1 reduces to Theorem 2 of Deheuvels and Ouadah [5].

2°) Under (3), the limit law (4) holds with the formal replacement of $\pm\{f_{n,K,h}(x) - \mathbb{E}\left(f_{n,K,h}(x)\right)\}$ by $|f_{n,K,h}(x) - \mathbb{E}\left(f_{n,K,h}(x)\right)|$.

3°) Our theorem can be used to construct uniform asymptotic simultaneous confidence bands for $f(\cdot)$, in the spirit of that given in Deheuvels and Mason [4].

4°) Recall that ψ denote a specified continuous and positive function on J. It could be defined in many ways, among which the following ones (see (1.13) in [3]):

$$\psi_1(x) = 1, \quad \psi_2(x) = \frac{1}{1-G(x)},$$
$$\psi_3(x) = f(x), \quad \psi_4(x) = \frac{f(x)}{1-G(x)},$$
$$\psi_5(x) = \frac{f(x)\psi(x)}{(1-F(x))^2(1-G(x))}.$$

5°) The replacement of $F(\cdot)$ by $F_n(\cdot)$ in $\psi_5(\cdot)$ corresponds to an estimator of the hazard rate function $\lambda_{n,K,h}(\cdot)$, we will consider in §1.2.

A motivation for uniform in bandwidth results such as that given Theorem 1, is to describe the limiting behavior of estimators when their bandwidth is possibly random or data dependent. Many elaborate schemes have been proposed in the statistical literature for constructing such bandwidth sequences (see, e.g., sections 2.4.1 and 2.4.2 in Deheuvels and Mason [4], and Berlinet and Devroye [1]). The use of bandwidths h of the form $h_n := Z_n n^{-1/5}$ where Z_n is a random sequence, stochastically bounded away from 0 and ∞, is usually suggested. It turns out that Theorem 1 allows the description of the limiting behavior of the corresponding kernel estimator. In the uncensored case, we refer to Einmahl and Mason [7], and Deheuvels and Ouadah [5], for discussions and references on uniform in bandwidth functional estimators. To illustrate the sharpness of the conditions (3) implying (4), we set $\mathcal{H}_n = [h_n, h_n]$ in Theorem 1, and observe that, whenever $\{h_n : n \geq 1\}$ are constants fulfilling, as $n \to \infty$,

$$nh_n/\log n \to \infty, \text{ and } h_n \to 0, \tag{5}$$

and the kernel function being a specified $K \in \mathcal{K}$, then, as $n \to \infty$,

$$\left\{ \frac{nh_n}{2\log_+(1/h_n)} \right\}^{1/2} \sup_{x\in I} \pm A_{n,K,h}(x)$$
$$\times \left\{ \psi_n(x) \times \frac{1-G(x)}{f(x)} \right\}^{1/2}$$
$$\stackrel{\mathbb{P}}{\to} \left\{ \sup_{x\in I} \psi(x) \int_{\mathbb{R}} K^2(t)dt \right\}^{1/2}. \tag{6}$$

Almost sure versions of (6) have been established, under various sets of assumptions, by Diehl and Stute [6] (taken with $\psi \equiv 1$ and $c = \infty$), Giné and Guillou [8], and Deheuvels and Eimahl [2], [3]. We note that (6) (and hence, (4)), does not hold almost surely [a.s.] for arbitrary continuous $f(\cdot)$ on J, and bandwidth sequences $\{h_n : n \geq 1\}$ fulfilling (5). If we assume, in addition to (5), that

$$\frac{\log(1/h_n)}{\log\log n} \to c \in (0, \infty], \quad h_n \downarrow 0, \text{ and } nh_n \uparrow \infty, \tag{7}$$

then, setting $(c+1)/c := 1$ when $c = \infty$ (see, e.g., Theorem 1.1, pp. 1304-1305 in [3]), we have,

a.s.,

$$\limsup_{n\to\infty}\left\{\frac{nh_n}{2\{\log_+(1/h_n)+\log\log n\}}\right\}^{1/2}$$
$$\times\sup_{x\in I}\pm A_{n,K,h}(x)\left\{\psi_n(x)\times\frac{1-G(x)}{f(x)}\right\}^{1/2}$$
$$=\left(\frac{c+1}{c}\right)^{1/2}\left\{\sup_{x\in I}\psi(x)\int_{\mathbb{R}}K^2(t)dt\right\}^{1/2},$$
$$\text{and }\ \liminf_{n\to\infty}\left\{\frac{nh_n}{2\{\log_+(1/h_n)+\log\log n\}}\right\}^{1/2}$$
$$\sup_{x\in I}\pm A_{n,K,h}(x)\left\{\psi_n(x)\times\frac{1-G(x)}{f(x)}\right\}^{1/2}$$
$$=\left\{\sup_{x\in I}\psi(x)\int_{\mathbb{R}}K^2(t)dt\right\}^{1/2}.$$

This last result is known not to hold in general when the first condition in (7) is not fulfilled. Viallon [13] (see, e.g., Maillot and Viallon [10]) has used the theory of *empirical processes indexed by functions* to obtain uniform in bandwidth convergence theorems in the spirit of (4), without the uniformity in kernel. He showed that, for a specified $K\in\mathcal{K}$, a.s.,

$$\lim_{n\to\infty}\sup_{h\in\mathcal{H}_n}\left\{\frac{nh}{2\log_+(1/h)}\right\}^{1/2}\sup_{x\in I}\left(\pm A_{n,K,h}(x)\right.$$
$$\left.\times\left\{\psi_n(x)\times\frac{1-G(x)}{f(x)}\right\}^{1/2}\right)=\sigma(\psi,K),\tag{8}$$

where $\mathcal{H}_n=[h_n',h_n'']$, and h_n',h_n'' are sequences of constants fulfilling (5)–(7) together with the additionnal condition $h_n''\le[(B-D)\wedge(1-H_1(D))]$ for each $n\ge1$ (J=[A,B], I=[C,D]), with $H_1(\cdot)=\mathbb{P}(T\le\cdot\text{ and }\delta=1)$. His method of proof is appropriate to provide an a.s. version, under suitable conditions on $\mathcal{H}_n$, as $n\to\infty$, of

$$\sup_{h\in\mathcal{H}_n}\left\{\frac{nh}{2\log_+(1/h)}\right\}^{1/2}\sup_{x\in I}\left(\pm A_{n,K,h}(x)\right.$$
$$\left.\times\left\{\psi_n(x)\times\frac{1-G(x)}{f(x)}\right\}^{1/2}\right)\xrightarrow{\mathbb{P}}\sigma(\psi,K).$$

Independently of the conditions imposed on $\mathcal{H}_n$ in either [13] or (7) (which are more strenuous than (3)), we should point out that this last result is a much weaker statement than (4) in terms of uniformity in bandwidth and kernel. Besides it is not clear whether the method Viallon maked use of, may be used or not, to give alternate proofs of our result.

1.2 Uniform in Bandwidth and Kernel Failure Rate Estimation.

Denote the failure rate function pertaining to $F(\cdot)$ by

$$\lambda(x):=\frac{f(x)}{(1-F(x))}\text{ for }x\in J.$$

We consider $\lambda_{n,K,h}(\cdot)$ the estimator of $\lambda(\cdot)$ defined by

$$\lambda_{n,K,h}(x):=\frac{f_{n,K,h}(x)}{(1-F_n(x))}\text{ for }x\in J,$$

where $f_{n,K,h}(\cdot)$ is as in (2) and $F_n(\cdot)$ as in (1). The following theorem, describes the *uniform in bandwidth and kernel* consistency of $\lambda_{n,K,h}(\cdot)$.

Theorem 2: Let $0<a_n\le b_n\le1$ be such that, as $n\to\infty$,

$$b_n\to0\quad\text{and}\quad\frac{na_n}{\log n}\to\infty.$$

Then, with $\mathcal{H}_n=[a_n,b_n]$, we have, as $n\to\infty$,

$$\sup_{h\in\mathcal{H}_n}\left|\sup_{K\in\mathcal{K}}\left|\left\{\frac{nh}{2\log_+(1/h)}\right\}^{1/2}\sup_{x\in I}\pm B_{n,K,h}(x)\right.\right.$$
$$\left.\left.\times\left\{\psi_n(x)\times\frac{1-H(x)}{\lambda(x)}\right\}^{1/2}-\sigma(\psi,K)\right|\right|=o_{\mathbb{P}}(1),$$

with

$$B_{n,K,h}(x):=\lambda_{n,K,h}(x)-\frac{\mathbb{E}(\lambda_{n,K,h}(x))}{1-F(x)},$$
$$\text{and }\sigma(\psi,K):=\left\{\sup_{x\in I}\psi(x)\int_{\mathbb{R}}K^2(t)dt\right\}^{1/2}.$$

Remark 2: 1°) The uniform consistency of $\lambda_{n,K,h}(\cdot)$ over bounded intervals was treated by Zhang [16], and Deheuvels and Einmahl [3].

2 MAIN TOOL OF PROOF.

2.1 A Functional Limit Law in the Uncensored Case.

To prove theorems 1 and 2, we will make use of a functional limit law due to Deheuvels and Ouadah [5]. The following notation is needed for the statement of this result, stated in Fact 1 below.

Let $U_1, U_2, \ldots$ be iid random variables with a uniform distribution on $(0,1)$. Denote by $\mathbb{U}_n(\cdot) := n^{-1}\#\{U_i \leq u : 1 \leq i \leq n\}$, the empirical df based upon the first $n \geq 1$ of these observations, with $\#$ denoting cardinality. Let,

$$\alpha_n(u) := n^{1/2}\left(\mathbb{U}_n(u) - u\right) \quad \text{for} \quad u \in \mathbb{R}\ ,$$

denote the uniform empirical process. For each choice of $h > 0$ and $t \in [0,1]$, consider, the increment function

$$\xi_n(h;t;u) := \alpha_n(t+hu) - \alpha_n(t) \text{ for } u \in \mathbb{R}\ .$$

together with the set of functions, defined, for $h > 0$, by

$$\mathcal{F}_{n;\mathcal{I},\gamma}(h) := \left\{ \frac{\xi_n(\gamma h;t;\cdot)}{\sqrt{2h\log_+(1/h)}} : t \in [0,1-h] \cap \mathcal{I} \right\},$$

where $\gamma > 0$ and $\mathcal{I} := [r,s] \subseteq [0,1]$ is a specified interval, with $r < s$. Denote, by $(B[0,1],\mathcal{U})$ (resp. $(AC[0,1],\mathcal{U})$) the set of bounded (resp. absolutely continuous) functions on $[0,1]$, endowed with the uniform topology $\mathcal{U}$, induced by the sup-norm $\|f\| := \sup_{u\in[0,1]}|f(u)|$. Let $\Delta(A,B)$, denote the Hausdorff set-distance of $A, B \subseteq B[0,1]$ (refer to the definition in Deheuvels and Ouadah [5]). For each $f \in AC[0,1]$, denote by $\dot{f}(u) = \frac{d}{du}f(u)$ the Lebesgue derivative of f for $u \in [0,1]$. Consider the Hilbertian norm defined on $B[0,1]$ by

$$|f|_{\mathbb{H}} := \left\{\int_0^1 \dot{f}(u)^2 du\right\}^{1/2}, \text{ when } f(0) = 0 \text{ and } f \in AC[0,1],$$
$$|f|_{\mathbb{H}} := \infty \text{ otherwise.}$$

For each $\lambda > 0$, set

$$\mathbb{S}_\lambda := \{f \in B[0,1] : |f|_{\mathbb{H}} \leq \lambda\} = \{\lambda^{1/2} f : f \in \mathbb{S}_1\}.$$

Notice that $\mathbb{S}_1 = \mathbb{S}$ is the unit ball of the reproducing kernel Hilbert space of the usual Wiener process on $[0,1]$, shown by Strassen [11] to be the limit set in the functional law of the iterated logarithm for Wiener processes. Given these notations, the functional limit law stated in Fact 1 below, is a direct consequence of Theorem 1 (i) in Deheuvels and Ouadah [5].

Fact 1: Assume that $0 < a_n \leq b_n \leq 1$ are such that, as $n \to \infty$,

$$b_n \to 0 \quad \text{and} \quad \frac{na_n}{\log n} \to \infty.$$

Then, with $\mathcal{H}_n = [a_n, b_n]$, for any $\gamma > 0$ and $\mathcal{I} = [u,v] \subseteq [0,1]$ with $u < v$, we have, as $n \to \infty$,

$$\sup_{h\in\mathcal{H}_n} \Delta\left(\mathcal{F}_{n;\mathcal{I},\gamma}(h), \mathbb{S}_\gamma\right) = o_{\mathbb{P}}(1).$$

Remark 3: Detailed proofs of theorems 1 and 2 will be given elsewhere.

REFERENCES

[1] Berlinet, A. and Devroye, L. A. (1994). Comparison of kernel density estimates. *Publ. Inst. Statist. Univ. Paris* **38**, 3–59.

[2] Deheuvels, P. and Einmahl, J. H. J. (1996) On the strong limiting behavior of local functionals of empirical processes based upon censored data. *Ann. Probab.* **24**, 504–525.

[3] Deheuvels, P. and Einmahl, J. H. J. (2000). Functional limit laws for the increments of Kaplan-Meier product-limit processes and applications. *Ann. Probab.* **28**, 1301–1335.

[4] Deheuvels, P. and Mason, D. M. (2004). General asymptotic confidence bands based on kernel-type function estimators. *Statist. Infer. Stoch. Processes.* **7**, 225–277.

[5] Deheuvels, P. and Ouadah, S. (2011). Uniform in bandwidth functionnal limit laws. *J. Theor. Probab.*, accepted for publication.

[6] Diehl, S. and Stute, W. (1988). Kernel density and hazard function estimation in the presence of censoring. *J. Multivariate Anal.* **25**, 299–310.

[7] Einmahl, U. and Mason, D. M. (2005). Uniform in bandwidth consistency of kernel-type function estimators. *Ann. Statist.* **33**, 1380–1403.

[8] Giné, E. and Guillou, A. (2001). On consistency of kernel density estimators for randomly censored data: rates holding uniformly over adaptive intervals. *Ann. Inst. H. Poincar Probab. Statist.* **37**, 503–522

[9] Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *J. Amer. Statist. Assoc.* **53**, 457–481

[10] Maillot, B. and Viallon, V. (2009). Uniform limit laws of the logarithm for nonparametric estimators of the regression function in presence of censored data. *Math. Methods Statist.* **18**, 159–184.

[11] Strassen, V. (1964). An invariance principle for the law of the iterated logarithm. *Z. Wahrscheinlichkeit. Verw. Gebiete.* **3**, 211–226.

[12] Tanner, M. A. and Wong, W.H. (1983). The estimation of the hazard function from randomly censored data by the kernel method. *Ann. Statist.* **11**, 989–993.

[13] Viallon, V. (2006). Processus empiriques, estimation non paramtrique et donnes censures. *Doctoral Dissertation, Université Pierre et Marie Curie, Dec. 2, 2006*. Paris, France.
[14] Watson, G. S. and Leadbetter, M. R. (1964a). Hazard analysis. II. *Sankhya- Ser. A* **26**, 101–116
[15] Watson, G. S. and Leadbetter, M. R. (1964b). Hazard analysis. I. *Biometrika* **51**, 175–184.
[16] Zhang, B. (1996) A law of the iterated logarithm for kernel estimators of hazard functions under random censorship. *Scand. J. Statist.* **23**, 37–47.

Estimation and Inference for Censored Linear Regression with Heteroscedastic Errors

Lei Pang
Department of Statistics
North Carolina State University
Raleigh, North Carolina, USA

Wenbin Lu
Department of Statistics
North Carolina State University
Raleigh, North Carolina, USA
Email: lu@stat.ncsu.edu

Huixia Judy Wang
Department of Statistics
North Carolina State University
Raleigh, North Carolina, USA

***Abstract*—In survival analysis, the accelerated failure time model is a useful alternative to the popular Cox proportional hazards model due to its easy interpretation. Current estimation methods for the accelerated failure time model mostly assume independent and identically distributed random errors. However, as shown in our two motivating examples, the conditional variance of log survival times may often depend on covariates exhibiting some form of heteroscedasticity. In this paper, we develop a local Buckley- James estimator for the accelerated failure time model with heteroscedastic errors. We establish the consistency and asymptotic normality of the proposed estimator and propose a resampling approach for inference. Simulation study demonstrates that the proposed method is flexible and leads to more efficient estimation when heteroscedasticity is present. The value of the proposed method is further assessed by the analysis of the two motivating survival data sets.**

I. Introduction

In survival analysis, the accelerated failure time (AFT) model is an attractive alternative to the popular proportional hazards model for its simplicity and ease of interpretability. The conventional AFT model assumes a direct linear relationship between T_i, the survival time or some transformation thereof, and the covariates X_i:

$$T_i = \alpha + X_i^T \beta + \epsilon_i, \ i = 1, \cdots, n, \tag{1}$$

where α is the intercept, β is the p-dimensional vector of regression coefficients, and ϵ_i is the independent and identically distributed ($i.i.d.$) random error with mean zero. The AFT model is semiparametric as the distribution of ϵ_i is not specified.

A large number of estimation methods have been proposed for the semiparametric AFT model. By assuming unconditional independence between survival and censoring times, the "synthetic data" approaches via the inverse probability of censoring weighted (IPCW) technique have been studied by Koul, Susarla, and Van Ryzin (1981), Leurgans (1987), and Fan and Gijbels (1994), among others. Many methods were also developed based on the more relaxed assumption of conditional independence of survival and censoring times, including the Buckley-James estimator (Buckley and James, 1979; Lai and Ying, 1991; Zhou and Li, 2008), weighted rank estimators (Tsiatis, 1990; Ritov, 1990; Ying, 1993; Zhou, 2005), and nonparametric maximum profile likelihood estimator (Zeng and Lin, 2007), to name a few. The asymptotic properties of these estimators and their associated inference procedures have been formally studied. In particular, Jin, Lin, Wei, and Ying (2003) and Jin, Lin, and Ying (2006) developed proper resampling procedures for variance estimation of the rank and Buckley-James estimators, respectively.

The semiparametric AFT model assumes that the random errors are $i.i.d.$ and independent of the covariates. Most existing estimation methods rely on this assumption. However, in many applications, the random errors tend to depend on the covariates and exhibit some form of heteroscedasticity. One of our motivating examples is the breast cancer data from a clinical study with three treatment arms of adjunct therapies for breast cancer (Farewell, 1986). The population is highly heterogenous with some subjects having extremely long survival times while others having very short survivals. Our exploratory analysis reveals that the conditional variance of the log survival times tends to be decreasing as the fitted means of the log survival times increase (here the Buckley-James estimator was used to fit the data); see Figure 1(b) in Section 5. A similar pattern is observed in the acute myocardial infarction data set (Pohar and Stare, 2006); see Figure 1(a).

In situations where the homoscedasticity assumption is violated, it is necessary to account for the heteroscedasticity in order to obtain valid and efficient inference. For example, Stare, Heinzl, and Harrell (2000) showed through simulation that the Buckley-James estimator is biased for the AFT model with heteroscedastic errors. The issue of heteroscedasticity in survival analysis was discussed in various contexts. Zhou, Bathke, and Kim (2007) discussed an empirical likelihood inference approach for a heteroscedastic AFT model. Wu, Hsieh, and Chen (2002) studied a Cox-type regression model accommodating heteroscedasticity. Heuchenne (2010) proposed a nonparametric method to estimate the conditional variance function when the response variable is subject to censoring. Regression of quantiles offered an alternative approach to capture the survival heterogeneity, and some related work can be found in Koenker and Geling (2001), Portnoy (2003), Peng and Huang (2008), Wang and Wang (2009), among others.

Little work has been done for the AFT model with heteroscedastic errors. Recently Liu and Lu (2009) proposed a Weighted Least Squares (WLS) method based on the "New

Class" (Fan and Gijbels, 1994) type data transformation , a linear combination of the transformations proposed in Koul, Susarla, and Van Ryzin (1981) and Leurgans (1987). Conditional variances of the transformed survival times are then estimated nonparametrically and used as weights for a weighted least squares fit. However, the WLS method requires the censoring times to be independent of the covariates and thus could be restrictive in practice.

In this paper, we develop a new local Buckley-James estimator for a heteroscedastic AFT model. The proposed method can be viewed as an extension of the Buckley-James estimator to account for heteroscedasticity. The main idea is to recursively impute the censored survival times by estimating the conditional mean based on the local Kaplan-Meier estimate of the conditional survival functions. The regression coefficients are then estimated through ordinary least squares fit based on the imputed survival times. Compared with the WLS method, our proposed method makes use of the censoring survival times more effectively and it allows the censoring time to depend on the covariates.

II. Proposed Method

A. The Heteroscedastic AFT Model

In light of the motivating examples, we consider the following heteroscedastic AFT model:

$$T_i = \alpha + X_i^T\beta + \sigma(X_i^T\beta)\epsilon_i,\ \ i = 1, 2, \ldots, n, \tag{2}$$

where T_i is the survival time or some transformation thereof, and ϵ_i are $i.i.d.$ random errors with mean zero and standard deviation one. The function $\sigma(X_i^T\beta)$ describes the error heteroscedasticity with $\sigma(\cdot)$ being an unspecified nonparametric function.

The above model is inspired by the two motivating data sets, where the conditional variance of log-transformed survival times tends to depend on the covariates through the fitted mean log-survival. Another motivation is the generalized linear model, where the variance of the response variable is often a function of the mean. Further relaxation on the form of heteroscedasticity, for example, letting $\sigma = \sigma(X_i)$, an arbitrary function of the covariate vector X_i's, is theoretically enticing but will impose technical difficulties due to the curse of dimensionality. In this paper, we focus on the estimation of β in model (2).

B. The Local Buckley-James Estimation

Due to right censoring, we only observe the triplets (Y_i, δ_i, X_i), where $Y_i = \min(T_i, C_i)$, $\delta_i = I(T_i \le C_i)$, and C_i is the corresponding censoring time. Throughout, we assume that T_i and C_i are conditionally independent given X_i.

When there is no censoring, the classical ordinary least squares (OLS) estimator $\hat\beta$ can be obtained by solving the following estimating equation for β:

$$n^{-1}\sum_{i=1}^{n}(X_i - \bar X_n)(T_i - X_i^T\beta) = 0, \tag{3}$$

where $\bar X_n = n^{-1}\sum_{i=1}^n X_i$. The intercept α can be estimated by $\hat\alpha = n^{-1}\sum_{i=1}^n e_i(\hat\beta)$, where $e_i(b) = T_i - X_i^T b$ for a given vector b.

In the presence of censoring, the exact survival times T_i are unknown for censored cases with $\delta_i = 0$. Buckley and James (1979) proposed to replace the censored T_i in (3) with the estimate of $E(T_i|T_i \ge C_i, Y_i, X_i)$, the conditional expectation of T_i. Buckley-James (BJ) estimator $\hat\beta$ can then be obtained by fitting least squares regression with the imputed survival times. However, the conventional BJ estimator assumes homoscedastic errors and thus does not work for model (2), where the variance of T_i depends on X_i.

To account for the heteroscedasticity, we propose a local Buckley-James (LBJ) estimator. The method shares the same spirit as the BJ estimator by imputing the censored survival time T_i by its estimated conditional mean. Under the heteroscedastic AFT model (2),

$$\begin{aligned} E(T_i|T_i \ge C_i, Y_i, X_i) &= E(e_i|T_i \ge C_i, Y_i, X_i^T\beta) + X_i^T\beta \\ &= \frac{\int_{Y_i - X_i^T\beta}^{\infty} u dF_\beta(u|X_i^T\beta)}{1 - F_\beta(Y_i - X_i^T\beta|X_i^T\beta)} + X_i^T\beta, \end{aligned} \tag{4}$$

where $F_\beta(u|v)$ is the unknown conditional cumulative distribution function (CDF) of the residual $e_i \equiv e_i(\beta) = T_i - X_i^T\beta$ given $X_i^T\beta = v$, i.e., $F_\beta(u|v) = P(e_i \le u|X_i^T\beta = v)$. In the presence of heteroscedastic errors, $F_\beta(u|X_i^T\beta)$ depends on $X_i^T\beta$ and thus cannot be estimated by the Kaplan-Meier estimate, as done in the conventional Buckley-James method. To account for such heteroscedasticity, we adopt the local Kaplan-Meier estimator (Dabrowska, 1987) to estimate $F_\beta(u|X_i^T\beta)$, which is then used to estimate β through iteration. The proposed local Buckley-James algorithm is as follows.

Step 1. Obtain an initial coefficient estimator $\hat\beta_0$, for instance, the Buckley-James estimator.

Step 2. At the ath iteration, impute the censored survival times T_i by

$$\tilde Y_i(\hat\beta_a) = \delta_i Y_i + (1-\delta_i)\hat E(T_i|T_i \ge C_i, Y_i, X_i^T\hat\beta_a), i = 1, 2, \ldots, n,$$

where

$$\hat E(T_i|T_i \ge C_i, Y_i, X_i^T\hat\beta_a) = X_i^T\hat\beta_a + \frac{\int_{\tilde e_i(\hat\beta_a)}^{\infty} u d\hat F_{\hat\beta_a}(u|X_i^T\hat\beta_a)}{1 - \hat F_{\hat\beta_a}\{\tilde e_i(\hat\beta_a)|X_i^T\hat\beta_a\}}, \tag{5}$$

where $\tilde e_i(b) = Y_i - X_i^T b$. In (5), $\hat F_b(u|X_i^T b)$ is the local Kaplan-Meier estimate of $F_b(u|X_i^T b)$, the conditional CDF of the residual $e_i(b)$ given $X_i^T b$. That is,

$$\hat F_b(t|X_i^T b) = 1 - \prod_{j:\tilde e_j(b) < t}^{n}\left\{1 - \frac{B_{nj}(X_i^T b)\delta_j}{\sum_{k=1}^n I\{\tilde e_k(b) \ge \tilde e_j(b)\}B_{nk}(X_i^T b)}\right\},$$

where $B_{nk}, k = 1, 2, \ldots, n$, is a sequence of non-negative weights with $\sum_{k=1}^n B_{nk} = 1$. We choose the Nadaraya-Watson type of weight for $B_{nk}(X_i^T b)$ (Nadaraya, 1964), namely,

$$B_{nk}(X_i^T b) = \frac{K(\frac{X_i^T b - X_k^T b}{h_n})}{\sum_{l=1}^n K(\frac{X_i^T b - X_l^T b}{h_n})},$$

where h_n is the bandwidth such that $h_n \to 0$ as $n \to \infty$ and $K(\cdot)$ is a symmetric kernel function.

Step 3. Fit least squares regression using the imputed survival times $\tilde{Y}_i(\hat{\beta}_a)$ to obtain the updated estimator

$$\hat{\beta}_{a+1} = \left\{ \sum_{i=1}^{n} (X_i - \bar{X}_n)^{\otimes 2} \right\}^{-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)\{\tilde{Y}_i(\hat{\beta}_a) - \bar{Y}_n(\hat{\beta}_a)\},$$

where $\bar{Y}_n(\hat{\beta}_a) = n^{-1} \sum_{i=1}^{n} \tilde{Y}_i(\hat{\beta}_a)$.

Step 4. Repeat Steps 2 and 3 until convergence is achieved. We denote the converged estimator as $\hat{\beta}_{LBJ}$.

III. Asymptotic Properties and Inference

A. Asymptotic Properties

Note that the local Buckley-James (LBJ) estimator $\hat{\beta}_{LBJ}$ is the solution to

$$\begin{aligned} & U_n(b) \\ = & \sum_{i=1}^{n} \left\{ \int_{-\infty}^{\infty} t dY_i^x(t,b) + \int_{-\infty}^{\infty} \int_{t}^{\infty} \frac{1 - \hat{F}_{ib}(s)}{1 - \hat{F}_{ib}(t)} ds dJ_i^x(t,b) \right\} \\ = & 0, \end{aligned}$$

where $Y_i^x(t,b) = (X_i - \bar{X}_n) I\{\tilde{e}_i(b) \geq t\}$ and $J_i^x(t,b) = (X_i - \bar{X}_n) I\{\tilde{e}_i(b) \geq t, \delta_i = 0\}$ are the indicator processes related to the ith observation, and $\hat{F}_{ib}(t)$ is the shorthand notation for $\hat{F}_b(t|X_i^T b)$. Since $U_n(b)$ is not a continuous function of b, to facilitate theoretical investigation, we define the LBJ estimator $\hat{\beta}_{LBJ}$ as a zero-crossing of the estimating function $U_n(b)$. Define $V_n(b)$ as a smooth approximation of $U_n(b)$,

$$\begin{aligned} & V_n(b) \\ = & \sum_{i=1}^{n} \left\{ \int_{-\infty}^{\infty} t dE Y_i^x(t,b) + \int_{-\infty}^{\infty} \int_{t}^{\infty} \frac{1 - F_{ib}(s)}{1 - F_{ib}(t)} ds dE J_i^x(t,b) \right\}, \end{aligned}$$

where $F_{ib}(t)$ is defined in the Appendix as the limit of $\hat{F}_{ib}(t)$. We assume the following regularity conditions.

- A1. $\sup_i \|X_i\| \leq M$, where M is a positive constant, and $\beta \in B_p(0,\rho)$, a p-dimensional ball in R^p centered at zero and with radius ρ. In addition, $X_i^T \beta$ has a differentiable and bounded density function $f_\mu(\cdot)$ and $\sigma(\cdot)$ is differentiable.
- A2. For all v, $F_\beta(u|v)$ has a bounded twice-differentiable density $f_\beta(u|v)$. In addition, $\int_{-\infty}^{\infty} u^2 dF_\beta(u|v) < \infty$ and $\int_{-\infty}^{\infty} \{\dot{f}_\beta(u|v)\}^2 / f_\beta(u|v) du < \infty$, where $\dot{f}_\beta(u|v)$ is the first derivative of $f_\beta(u|v)$ with respect to u.
- A3. The bandwidth satisfies $h_n = O(n^{-1/2+\kappa})$, where $0 < \kappa \leq 1/6$.
- A4. The kernel function $K(\cdot)$ is Lipschitz continuous of order one and satisfies $\int K(u) du = 1$, $\int uK(u) du = 0$, $\int K^2(u) du < \infty$, and $\int u^2 K(u) du < \infty$.
- A5. There exist some constants ν_1 and $\nu_2 > 0$ such that $P(C_i - X_i'\beta > \nu_1) = 0$ and $\inf_v F_\beta(\nu_1|v) > \nu_2$ for all v.
- A6. For $0 < \lambda < \frac{1}{12}$, $\lim_{n\to\infty} n^{-3/4} \{ \inf_{\|b\| \leq \rho, \|b-\beta\| \geq n^{-\lambda}} \|V_n(b)\| \} = \infty$.
- A7. The first order derivative matrix $\mathbf{\Gamma}_n$ of $n^{-1} V_n(b)$ at β converges to a positive definite matrix $\mathbf{\Gamma}$, as n goes to infinity.

Conditions A1-A5 are standard conditions that are widely used in the literature for simplifying the derivation of asymptotic properties including those of the local Kaplan-Meier estimator. Condition A6 is assumed to ensure the consistency of the LBJ estimator as the zero-crossing of $U_n(b)$ for $b \in B_p(0,\rho)$. A similar condition was also assumed in Lai and Ying (1991) for the regular BJ estimator. Condition A7 is needed to establish the asymptotic normality of the LBJ estimator.

Theorem III.1 *Under assumptions A1-A7, we have, as* $n \to \infty$,

(i) $\hat{\beta}_{LBJ} - \beta = o(n^{-\lambda})$ *a.s.;*

(ii) $\sqrt{n}(\hat{\beta}_{LBJ} - \beta) \xrightarrow{d} N(0, \mathbf{\Gamma}^{-1} \mathbf{\Sigma} \mathbf{\Gamma}^{-1})$, *where* $\mathbf{\Sigma}$ *is the asymptotic covariance matrix of* $n^{-1/2} U_n(\beta)$ *as defined in Lemma 7.4 of the Appendix.*

B. Inference via Resampling

Since the matrices $\mathbf{\Gamma}$ and $\mathbf{\Sigma}$ take complicated analytical forms and involve the unknown conditional error density function, it is impractical to directly estimate them to obtain the variance estimate of $\hat{\beta}_{LBJ}$. Jin, Lin, and Ying (2006) proposed a resampling procedure for estimating the variance of the regular BJ estimator and showed its validity. Here, we adopt a similar resampling approach for variance estimation of the LBJ estimator. The resampling procedure for the LBJ method works as follows.

First, we generate positive random variables $W_i, i = 1, 2, \ldots, n$, with $E(W_i) = Var(W_i) = 1$, which are used to introduce random perturbation into the LBJ estimation. To be specific, define

$$L^*(b) = \left\{ \sum_{i=1}^{n} W_i (X_i - \bar{X}_n)^{\otimes 2} \right\}^{-1} \left[\sum_{i=1}^{n} W_i (X_i - \bar{X}_n)\{\tilde{Y}_i^*(b) - \bar{Y}_n^*(b)\} \right],$$

where

$$\begin{aligned} \tilde{Y}_i^*(b) &= \delta_i Y_i + (1 - \delta_i) \left[\frac{\int_{\tilde{e}_i(b)}^{\infty} u d\hat{F}_b^*(u|X_i^T b)}{1 - \hat{F}_b^*\{\tilde{e}_i(b)|X_i^T b\}} + X_i^T b \right], \\ \hat{F}_b^*(t|X_i^T b) &= 1 - \prod_{j:\tilde{e}_j(b) \leq t}^{n} \left[1 - \frac{W_j B_{nj}(X_i^T b) \delta_j}{\sum_{k=1}^{n} W_k I\{\tilde{e}_k(b) \geq \tilde{e}_j(b)\} B_{nk}(X_i^T b)} \right], \end{aligned}$$

and $\bar{Y}_n^*(b) = n^{-1} \sum_{i=1}^{n} \tilde{Y}_i^*(b)$. Note that $\hat{F}_b^*(t|X_i^T b)$ is a randomly perturbed version of the local Kaplan-Meier estimate, $\tilde{Y}_i^*(b)$ is the imputed survival time based on $\hat{F}_b^*(t|X_i^T b)$, and $L^*(b)$ is the result of a perturbed least squares estimation.

Then, starting from the initial estimate $\hat{\beta}_{LBJ}$, we can obtain one resampled estimate $\hat{\beta}^*$ by running the above perturbed LBJ estimation until convergence. That is, starting from $\hat{\beta}_{LBJ}$, $\hat{\beta}^*$ is obtained by iteratively updating $L^*(\cdot)$ from the previous estimate till the algorithm converges. As in Jin, Lin, and Ying (2006), it can be shown that given the observed data, $\sqrt{n}(\hat{\beta}^* - \hat{\beta}_{LBJ})$ converges to the same limiting distribution as $\sqrt{n}(\hat{\beta}_{LBJ} - \beta)$. The proof of this result mainly follows the proof of Theorem 1, and thus is omitted in the paper. By repeating the above resampling scheme N times, we obtain the resampled estimates $\hat{\beta}_k^*, k = 1, 2, \ldots, N$. The sample variance

of $\{\hat{\beta}_k^*, k = 1, 2, \ldots, N\}$ provides a consistent variance estimate of $\hat{\beta}_{LBJ}$.

IV. Simulation Study

In the simulation study, we compare the finite sample performance of the Buckley-James (BJ) estimator, the Weighted Least Squares (WLS) estimator of Liu and Lu (2009) and our proposed local Buckley-James (LBJ) estimator under various situations.

We consider four scenarios. In scenario 1, data are generated with homoscedastic errors and covariate-independent censoring. In scenario 2, data are generated with heteroscedastic errors with covariate-independent censoring. Scenario 3 is based on the simulation setting in Liu and Lu (2009), which also assumes heteroscedastic error and covariate-independent censoring but with more covariates. In scenario 4, data are generated with heteroscedastic errors with covariate-dependent censoring.

For scenarios 1, 2 and 4, we generate T_i, the log survival time, from the following model

$$T_i = X_i^T\beta + \sigma(X_i^T\beta)\epsilon_i,\ i = 1, 2, \ldots, n, \tag{6}$$

where $X_i = (X_{i1}, X_{i2})^T$, $X_{i1} \sim$ Unif(-1,1) and $X_{i2} \sim$ Bernoulli(0.5). Here we choose $\beta = (\beta_1, \beta_2)^T = (1, 1)^T$ and $\sigma(X_i^T\beta) = \exp(-0.3 - X_i^T\beta)$ for scenarios 2 and 4, while $\sigma(X_i^T\beta) = 0.7$ for scenario 1. We consider two different families of error distribution for ϵ_i: standard normal and centered standard extreme distributions.

For scenario 3, T_i is generated from the same model as (6) with: $X_i = (1, X_{i1}, X_{i2}, X_{i3}, X_{i4})^T$, where $X_{i1} \sim$ Unif(-1,1), $X_{i2} = X_{i1}/3 + 2X_{i5}/3$ with $X_{i5} \sim$ Triangle(-2,2) being independent of X_{i1}, and X_{i3} and X_{i4} are independent Bernoulli random variables with success probability 0.5 and both are independent of X_{i1} and X_{i2}. Here $\beta = (6, -1, 2, 1, -1)^T$, $\sigma(X_i^T\beta) = \exp(3.52 - X_i^T\beta)$ and $\epsilon_i \sim \mathrm{N}(0, 1)$.

For scenarios with covariate-independent censoring (scenarios 1, 2 and 3), the censoring time C_i is generated from a normal distribution $\mathrm{N}(c_1, c_2)$. For covariate-dependent censoring (scenario 4), the censoring time C_i is generated from a distribution that depends on the covariate X_{i2}. Specifically, C_i is generated from $\mathrm{N}(c_3, c_4)$ if $X_{i2} = 1$, and from $\mathrm{N}(c_5, c_6)$ if $X_{i2} = 0$. Constants $c_i, i = 1, \cdots, 6$ are chosen to yield two censoring proportions, 20% and 40%. For each setting, we consider two sample sizes: $n = 200$ and 400. For the proposed LBJ method, the bandwidth parameter is chosen by $h_n = 4\mathrm{sd}(X^T\hat{\beta}_0)n^{-1/3}$, where $\mathrm{sd}(X^T\hat{\beta}_0)$ is the standard deviation of the linear index $X^T\hat{\beta}_0$, and $\hat{\beta}_0$ refers to the initial estimator in the LBJ algorithm.

Tables 1–4 summarize the simulation results of three different methods in the four scenarios, respectively. In the tables, $bias$ is the mean bias averaged over 500 simulations, sd is the Monte Carlo standard deviation, se is the mean estimated standard error obtained from the resampling procedure, and $covp$ is the empirical coverage probability of the Wald-type 95% confidence interval (the results are omitted here).

For the simplest scenario 1 with homoscedastic error and covariate independent censoring, all three methods give essentially unbiased estimation. The resampling procedure for the LBJ method works reasonably well. The resampling standard errors are close to the Monte Carlo standard deviations, and the confidence intervals have coverage probabilities close to the 95% nominal level. Not surprisingly, BJ is slightly more efficient than LBJ as the $i.i.d.$ error assumption is satisfied in this scenario. Both BJ and LBJ estimators tend to be more efficient than WLS. One possible explanation is that the imputation procedure in LBJ utilizes the information from the censored data more efficiently than WLS, which is partly dependent on the inverse probability weighting principle.

For scenario 2 with heteroscedastic errors, the estimations from BJ are clearly biased, and the bias is more prominent with heavier censoring. In contrast, both LBJ and WLS still give unbiased estimations. As observed in scenario 1, LBJ tends to be more efficient than WLS.

The design of scenario 3 is the same as the example used in Liu and Lu (2009). For comparison, in Table 3 we also include the results of the unweighted version (LS) of the weighted least squares method. Compared with LS, WLS accounts for the heteroscedasticity by performing a weighted least squares based on the nonparametric estimates of the error conditional variance. Therefore, WLS leads to more efficient estimation than LS in this scenario with heteroscedastic errors. For the light censoring (20%), WLS and LBJ estimates have similar performance. However, for heavier censoring (40%), LBJ estimates have clearly smaller variances than WLS, and this agrees with our observations in scenarios 1-2.

Scenario 4 presents a more complicated design, where the errors are heteroscedastic and the censoring depends on the covariates. Therefore, the assumptions required by BJ and WLS are both violated. As observed in scenarios 2-3, BJ estimates show systematic biases. The WLS estimates are also biased, especially for the estimation of β_2, the coefficient of the covariate that affects the censoring distribution.

In summary, the proposed LBJ method works competitively well in all the simulation scenarios considered, for both homoscedastic or heteroscedastic error models, and for both covariate-dependent and covariate-independent censoring. We also observe that LBJ tends to be more efficient than WLS, and this is possibly due to the efficient imputation step of the LBJ method.

V. Data Analysis

To further illustrate the proposed method, we analyze the two motivating examples: the acute myocardial infarction data set and the breast cancer data set.

A. Acute Myocardial Infarction Data Analysis

The acute myocardial infarction data is based on a clinical study conducted at the University Clinical Center in Ljubljana. The purpose of the study was to explore the relationship between patients' survival times after acute myocardial infarction and patients' characteristics such as age and gender. The data

set contains records on 1040 patients with age ranging from 24 to 95, and female to male ratio approximately 1:3. The data set can be obtained in the R package "relsurv". More detailed information about the study could be found in Pohar and Stare (2006).

As a preliminary analysis, we use the BJ method to fit the data and obtain the estimator $\hat{\beta}_{BJ}$. Let $Y_i - X_i^T\hat{\beta}_{BJ}$ be the estimated residual and $X_i^T\hat{\beta}_{BJ}$ be the estimated linear index, where Y_i is the observed log survival time. We plot the centered residuals against the estimated linear indices, as shown in Figure 1(a). Although the residual plot does not truthfully describe the error heteroscedasticity since it plots censored residuals together with uncensored ones, it suggests the existence of error heteroscedasticity, and the conditional error variance tends to decrease as the mean log survival time increases.

Analysis results from methods BJ, LBJ and WLS are summarized in the top part of Table 5. The results from the proposed LBJ method suggest that age is significantly associated with log survival (p-value< 0.001) while gender is marginally significant (p-value$= 0.047$). Compared to LBJ, BJ gives similar coefficient estimates, but fails to identify gender as significant due to slightly larger standard error estimates. In terms of p-values, WLS gives the same results as LBJ, and WLS's coefficient estimates are of a smaller scale compared to LBJ and BJ. This data set has been analyzed by Wang and Wang (2009) using a censored quantile regression method. Their results suggested that increased age and being a female is associated with significantly shorter median log survival time, and this agrees with our analysis about the mean log survival time based on the LBJ method.

B. Breast Cancer Data Analysis

The breast cancer data is from a clinical trial with three treatment arms of adjunct therapies for breast cancer (Farewell, 1986). Besides the observed survival times and censoring indicator, the data set also contains the following information: indicators for two treatments, trt1 and trt2 (control arm is set as the baseline), clinical stage I indicator (an early stage indicator, equals 1 with tumor size smaller than 2 cm and no positive movable axillary nodes), and the number of lymph nodes having disease involvement. The data set contains 139 records with 44 events (censoring proportion is about 68%). To simplify the analysis, we transform the number of lymph nodes to a binary variable, i.e., with or without lymph nodes having disease involvement.

We perform a similar preliminary analysis for the breast cancer data. Figure 1(b) suggests that the conditional variance of the log survival also changes with fitted linear indices. The results from methods BJ, LBJ and WLS are summarized in the bottom part of Table 5. Methods BJ and LBJ give similar coefficient estimates. For predictors trt2 and clinical stage I, the WLS estimates are different from those of BJ and LBJ. In general, the LBJ estimates have smaller standard errors than the BJ estimates, resulting in smaller p-values across all variables. Both methods LBJ and BJ identify all covariates as significant, while WLS fails to identify the significant effectiveness of treatment 2.

Using the same data set, previous analyses by Farewell (1986), Peng and Dear (2000), and Lu (2010) suggested that treatment 1 is significantly beneficial in short-term survival while treatment 2 is significantly beneficial in long-term survival, and clinical stage I is significantly associated with both short-term and long-term survivals. Compared to WLS, the results from the proposed LBJ method are more in line with these previous analyses.

VI. Conclusion

In this paper, we developed a new estimation method for the semiparametric AFT model with heteroscedastic random errors. Compared with the existing methods, our proposed method is more flexible, and it allows both heteroscedasticity and covariate-dependent censoring. Motivated by real survival studies and by the generalized linear model, we assumed that the error variance is related to the linear mean survival function $X_i^T\beta$ through a nonparametric link function $\sigma(\cdot)$. The essence of the proposed idea can be adapted for models with more general forms of heteroscedasticity, for instance, by allowing the conditional variance to depend on the index $X_i^T\gamma$ with γ a p-dimensional vector possibly different from β. Further research is needed in this direction.

Acknowledgment

This research was partially supported by the NIH award R01 CA-140632.

References

Buckley, J. and James, I. (1979), "Linear regression with censored data," *Biometrika*, 66, 429–436.

Dabrowska, D. (1987), "Non-parametric regression with censored survival time data," *Scandinavian Journal of Statistics*, 181–197.

Fan, J. and Gijbels, I. (1994), "Censored Regression: Local Linear Approximations and Their Applications," *Journal of the American Statistical Association*, 89, 560–570.

Farewell, V. (1986), "Mixture models in survival analysis: Are they worth the risk?" *Canadian Journal of Statistics*, 14, 257–262.

Heuchenne, C. (2010), "Conditional variance estimation in censored regression models," *Technical Report*.

Jin, Z., Lin, D. Y., Wei, L. J., and Ying, Z. (2003), "Rank-based inference for the accelerated failure time model," *Biometrika*, 90, 341.

Jin, Z., Lin, D. Y., and Ying, Z. (2006), "On least-squares regression with censored data," *Biometrika*, 93, 147.

Koenker, R. and Geling, O. (2001), "Reappraising Medfly Longevity," *Journal of the American Statistical Association*, 96, 458–468.

Koul, H., Susarla, V., and Van Ryzin, J. (1981), "Regression Analysis with Randomly Right-Censored Data," *The Annals of Statistics*, 9, 1276–1288.

Lai, T. L. and Ying, Z. (1991), "Large Sample Theory of a Modified Buckley-James Estimator for Regression Analysis with Censored Data," *The Annals of Statistics*, 19, pp. 1370–1402.

Leurgans, S. (1987), "Linear models, random censoring, and synthetic data," *Biometrika*, 74, 301–309.

Liu, W. and Lu, X. (2009), "Weighted least squares method for censored linear models," *Journal of Nonparametric Statistics*, 21:7, 787–799.

Lu, W. (2010), "Efficient estimation for an accelerated failure time model with a cure fraction," *Statistica Sinica*, 20, 661.

Nadaraya, E. (1964), "On Estimating Regression," *Theory of Probability and its Applications*, 9, 141–142.

Peng, L. and Huang, Y. (2008), "Survival analysis with quantile regression models," *Journal of the American Statistical Association*, 103, 637–649.

Peng, Y. and Dear, K. (2000), "A nonparametric mixture model for cure rate estimation," *Biometrics*, 56, 237–243.

Pohar, M. and Stare, J. (2006), "Relative survival analysis in R," *Computer methods and programs in biomedicine*, 81, 272–278.

Portnoy, S. (2003), "Censored regression quantiles," *Journal of the American Statistical Association*, 98, 1001–1012.

Ritov, Y. (1990), "Estimation in a Linear Regression Model with Censored Data," *The Annals of Statistics*, 18, pp. 303–328.

Stare, J., Heinzl, H., and Harrell, F. (2000), "On the use of Buckley and James least squares regression for survival data," *New Approaches in Applied Statistics*, 16, 125–134.

Tsiatis, A. (1990), "Estimating regression parameters using linear rank tests for censored data," *Annals of Statistics*, 18, 354–372.

Wang, H. and Wang, L. (2009), "Locally weighted censored quantile regression," *Journal of the American Statistical Association*, 104, 1117–1128.

Wu, H., Hsieh, F., and Chen, C. (2002), "Validation of a heteroscedastic hazards regression model," *Lifetime Data Analysis*, 8, 21–34.

Ying, Z. (1993), "A large sample study of rank estimation for censored regression data," *The Annals of Statistics*, 76–99.

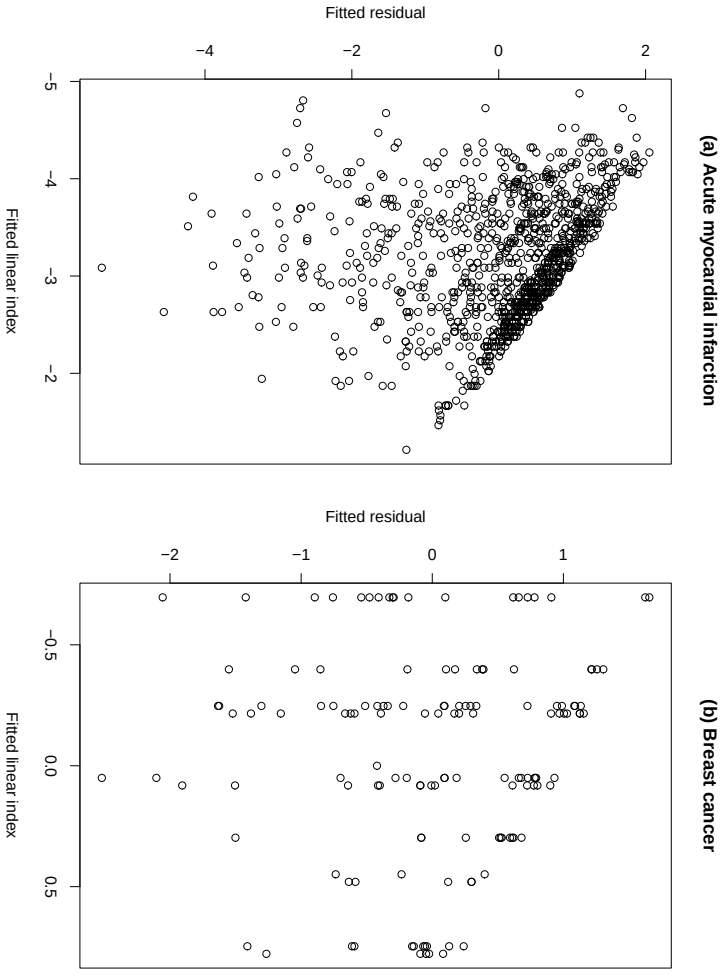
Zeng, D. and Lin, D. Y. (2007), "Efficient estimation for the accelerated failure time model," *Journal of the American Statistical Association*, 102, 1387–1396.

Zhou, M. (2005), "Empirical likelihood analysis of the rank estimator for the censored accelerated failure time model," *Biometrika*, 92, 492.

Zhou, M., Bathke, A., and Kim, M. (2007), "Empirical likelihood for Heteroscedastic AFT model," *Technical Report*.

Zhou, M. and Li, G. (2008), "Empirical likelihood analysis of the Buckley-James estimator," *Journal of multivariate analysis*, 99, 649–664.

Fig. 1. Estimated residuals against the estimated linear indices in the acute myocardial infarction data and the breast cancer data



Age replacement policy for a gamma process modulated by a Markov jump process

Christian Paroissin
Université de Pau et des Pays de l'Adour
LMAP – UMR CNRS 5142
Avenue de l'Université
64013 Pau cedex, France
Email: cparoiss@univ-pau.fr

Landy Rabehasaina
Université de de Franche-Comté
LMB – UMR CNRS 6623
16 Route de Gray
25030 Besançon cedex, France
Email: lrabehas@univ-fcomte.fr

***Abstract*—In this paper we consider a Markov modulated gamma process as a degradation model. Such model corresponds to the degradation of a device subject to a Markovian covariate. We focus essentially on the case of a single binary covariate, but it can be easily extended to the case of multi-state covariate or to the case of multiple covariates. For such degradation process the block-replacement policy has been considered previously and numerical studies show that a certain stochastic comparison property is preserved [8]. Here we consider a different maintenance policy, namely the age replacement policy, and we raise the following question: is the stochastic comparison property also preserved under this policy?**

I. Introduction and model

Gamma process is one of the most popular stochastic process to model degradation of device in reliability theory (see the review by van Noortwijk [13]). Briefly one says that a stochastic process is a stationary gamma process if and only if its increments are stationary, independent and gamma distributed.

In several papers, covariates or random effects have been incorporated to a gamma process in order to take into account the environment or the individual heterogeneity. For instance random effects have been considered by Lawless and Crowder [5] assuming that the scale parameter is random and may depend on covariates. Bagdonavičius and Nikulin [1] proposed an accelerated life test model by scaling the time index by time-dependent covariates. However (time-dependent or not) covariates are all assumed to be deterministic.

As in a previous paper [8], we consider a gamma process influenced by some covariates which evolves according to a Markov jump process (which is assumed to be independent of the underlying gamma processes). For lack of simplicity we restrict ourselves to a two-states Markov process (or binary Markov process), but it can be extended to a multi-state Markov process. This Markov process aims at describing the environment in which the device is used. For instance assume that the device could be used under nominal stress (state 0) or accelerated stress (state 1). When the transition rate from 1 to 0 is null, then it turns to be a special case of the model studied by Saassouh *et al.* [9] for which one will assume an exponential duration in the nominal state. Another related model is the one studied by Zhao *et al.* [15] (see also [14]). Such kind of models has been considered earlier [12].

A. Covariate process

The covariates (or solicitation conditions) are modelled by a binary Markov process $(J(t))$. We will denote λ (resp. μ) the rate from state 0 to state 1 (resp. from state 1 to state 0). We recall that a Markov process is characterized by its infinitesimal generator Q and its initial distribution ν. For any $t \geqslant 0$, the transition matrix between instants 0 et t is given by $P_t = \exp(tQ)$ and thus the distribution of $J(t)$ is $\nu \exp(tQ)$. For any $t \geqslant 0$:

$$P_t = \frac{1}{\lambda+\mu}\begin{pmatrix}\mu & \lambda\\ \mu & \lambda\end{pmatrix} + \frac{e^{-(\lambda+\mu)t}}{\lambda+\mu}\begin{pmatrix}\lambda & -\lambda\\ -\mu & \mu\end{pmatrix}.$$

It follows that its unique stationary distribution π equals to:

$$\pi = \begin{pmatrix}\frac{\mu}{\lambda+\mu}\\ \frac{\lambda}{\lambda+\mu}\end{pmatrix}.$$

In the general context, one has to assume that $(J(t))$ is an homogeneous, irreducible and recurrent Markov process. Most of the times the transition matrix cannot be computed explicitly.

B. Degradation process

The degradation process $(D(t))$ is described through the increments of two independent gamma processes whose parameters depend on the state of covariates. If the covariates are in state 0, then the degradation process will be governed by a gamma process with parameter (ξ, α_0) where ξ is the scale parameter and α_0 the shape parameter, and if the covariates are in state 1, then the degradation process will be governed by a gamma process with parameter (ξ, α_1) with $\alpha_0 \leqslant \alpha_1$ (the average degradation is larger under higher solicitation use than under nominal condition). More precisely let (T_n) be the instant of jumps of the Markov process $(J(t))$, with the convention that $T_0 = 0$. We have the following decomposition:

$$D(t) = \sum_{k=1}^{n}(D(T_k) - D(T_{k-1})) + (D(t) - D(T_n)),$$

if $T_n \leqslant t < T_{n+1}$. In other words, we have:

$$D(t) = \sum_{n\geq 0} \left[\sum_{k=1}^{n} (D(T_k) - D(T_{k-1})) + (D(t) - D(T_n)) \right] \mathbb{I}_{T_n \leqslant t < T_{n+1}}.$$

Moreover, conditionally to (T_n), $D(T_k) - D(T_{k-1})$ is gamma distributed with parameter $(\xi, \alpha_0(T_k - T_{k-1}))$ if k is odd and $(\xi, \alpha_1(T_k - T_{k-1}))$ otherwise.

II. Failure time distribution

We recall here some results about the hitting time distribution of a fixed level $c > 0$ by a Markov modulated gamma process:

$$T_c = \inf \{t \geqslant 0 \; ; \; D(t) \geqslant c\}.$$

For more details, the reader could refer to [8]. Since $(D(t))$ has increasing paths, it follows that:

$$\forall t \geqslant 0 \, , \quad \mathbb{P}[T_c > t] = \mathbb{P}[D(t) < c].$$

Thus it follows that it is sufficient to study the distribution of $D(t)$ for any $t \geqslant 0$.

A. Case of a non-modulated gamma process

First we consider the distribution of the hitting time for a non-modulated gamma process. According to Park and Padgett [7], the cumulative distribution function of T_c is, for any $t \geqslant 0$,

$$F_{T_c}(t) = \mathbb{P}[D(t) \geqslant c] = \frac{\Gamma(\alpha t, c/\xi)}{\Gamma(\alpha t)}, \tag{1}$$

where $\Gamma(\cdot, \cdot)$ is the upper incomplete gamma function. Moreover, the probability distribution function (pdf) of T_c is, for any $t \geqslant 0$,

$$f_{T_c}(t) = \alpha \left(\Psi(\alpha t) - \log\left(\frac{c}{\xi}\right) \right) \frac{\gamma(\alpha t, c/\xi)}{\Gamma(\alpha t)} + \frac{\alpha}{(\alpha t)^2 \Gamma(\alpha t)} \left(\frac{c}{\xi}\right)^{\alpha t} {}_2F_2(\alpha t, \alpha t; \alpha t + 1, \alpha t + 1; -c/\xi),$$

where Ψ is the di-gamma function (or logarithmic derivative of the gamma function), $\gamma(a, x)$ is the lower incomplete gamma function and ${}_2F_2$ the generalized hypergeometric function of order $(2, 2)$.

B. Case of a Markov modulated gamma process

Let us turn to the case of a Markov modulated gamma process. Consider $\Delta_0(t)$ is the occupation time of state 0 between the interval $[0, t]$. Such random variables have been already studied in the related field of computers reliability [3], [10]. Here the random variable $\Delta_0(t)$ depends on the time t and is given by:

$$\Delta_0(t) = \int_0^t \mathbb{I}_{J(u)=0} du.$$

The independent increment property of the gamma process, as well its independence from the Markov process, implies

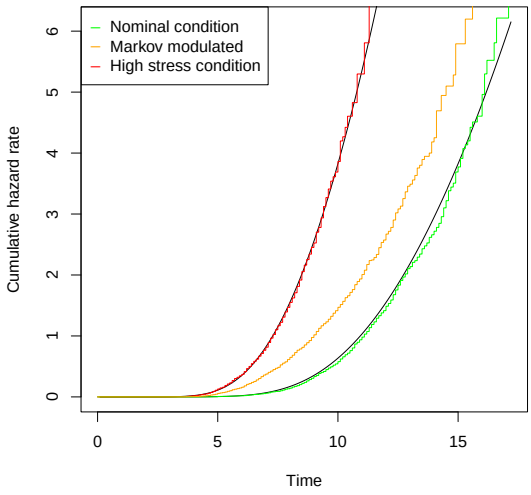


Fig. 1. Cumulative hazard rates

that we have equality in distribution $D(t) \stackrel{d}{=} D^0(\Delta_0(t)) + D^1(t - \Delta_0(t))$ where $(D^0(t))$ and $(D^1(t))$ are independent gamma processes with respective parameters (ξ, α_0) and (ξ, α_1). Thus, after conditioning on $\Delta_0(t)$, one can decompose the integral as previously and get:

$$\mathbb{P}[D(t) \leqslant c] = \int_0^t \left[1 - \frac{\Gamma(\alpha_0 u + \alpha_1(t-u), c/\xi)}{\Gamma(\alpha_0 u + \alpha_1(t-u))} \right] dF_{\Delta_0(t)}(u), \tag{2}$$

where $F_{\Delta_0(t)}$ is the cdf of $\Delta_0(t)$. The expression of $F_{\Delta_0(t)}$ is quite complex but has been derived by Sericola [10], [11]: it can be viewed roughly speaking as a mixture of a Dirac distribution (corresponding to the event that no jump occurs between 0 and t) and an absolutely continuous distribution.

Simulations were carried out in order to compare hitting time distributions (with 1000 repetitions). The parameters were set as follows: $\lambda = \mu = 1/10$, $\xi = 0.1$, $\alpha_0 = 2$, $\alpha_1 = 3$ and $c = 2$. On Figure 1 we represent cumulative hazard rate of hitting times. The green curve corresponds to the cumulative hazard rate of the hitting time when the covariate is always equal to 0 (i.e. when $\mu = 0$) while the red curve corresponds to the cumulative hazard rate function of the hitting time when the covariate is always equal to 1 (i.e. when $\lambda = 0$). For these cumulative hazard rates, we have indeed the explicit expression since the corresponding degradation process is simply a gamma process (see at the begin of this section). These exact cumulative hazard rates are plotted in black. At least the orange curve corresponds to the empirical cumulative hazard rate of the hitting time for the Markov modulated gamma process.

TABLE I
APPROXIMATIONS AND BOUNDS FOR THE MTTF

MTTF	$\mathbb{E}[T_c^{(1)}]$	$\mathbb{E}[T_c]$	$\mathbb{E}[T_c^{(0)}]$
Numerical comp.	6.84	8.39	10.42
Approximation	6.83	-	10.25

C. Stochastic comparison

Let us denote by $T_c^{(0)}$ the hitting time when $\mu = 0$ (deterioration only under the nominal mode) and by $T_c^{(1)}$ the hitting time when $\lambda = 0$ (deterioration only under the high-stress mode). As confirmed by simulations, the following stochastic orders hold (in the usual sense):

Proposition II.1. *We have the following relations:*

$$T_c^{(1)} \preceq_{st} T_c \preceq_{st} T_c^{(0)}.$$

See [8] for the proof. As it is well known, the above result implies the following order between mean time to failures (MTTF):

$$\mathbb{E}[T_c^{(1)}] \leqslant \mathbb{E}[T_c] \leqslant \mathbb{E}[T_c^{(0)}].$$

Using the approximation of the MTTF (in the non-modulated case) obtained by Bérenguer *et al.* [2], we can deduce lower and upper bounds for the MTTF in the modulated case (provided that c/ξ is large enough):

$$\frac{1}{\alpha_1}\left(\frac{c}{\xi}+\frac{1}{2}\right) \lessapprox \mathbb{E}[T_c] \lessapprox \frac{1}{\alpha_0}\left(\frac{c}{\xi}+\frac{1}{2}\right).$$

We illustrate these inequalities by considering the same simulations as previously. Results are reported on Table I.

III. AGE REPLACEMENT POLICY

Here we consider the age replacement policy (see Chapter 3 of [6] for additional details). A device is replaced either at failure or at age δ if no failure occurs between 0 and δ ($\delta = \infty$ means that replacement occurs only at failure). Hence there are two different costs: c_f the one incurred by the replacement of a failed device and $c_{nf} < c_f$ the one incurred by the replacement of a non-failed device. A natural problem is to determine the optimal value of δ for such policy. Using also the renewal theory, the asymptotic cost per unit of time is equal to:

$$C_{arp}(\delta) = \frac{c_f F_{T_c}(\delta) + c_{nf} R_{T_c}(\delta)}{\int_0^\delta R_{T_c}(u)du},$$

where R_{T_c} is the reliability function of T_c. When δ tends to infinity, since it corresponds to the case where replacement occurs only at failure, then it can easily seen that $C_{arp}(\delta)$ tends to $\frac{c_f}{\mathbb{E}[T_c]} := C_{arp}(\infty)$.

From the stochastic order properties between failure times shown previously and since we have assumed that $c_{nf} < c_f$, it follows that the same relationship holds also for the cost functions:

$$\forall \delta > 0, \quad C_{arp}^{(0)}(\delta) \leqslant C_{arp}(\delta) \leqslant C_{arp}^{(1)}(\delta),$$

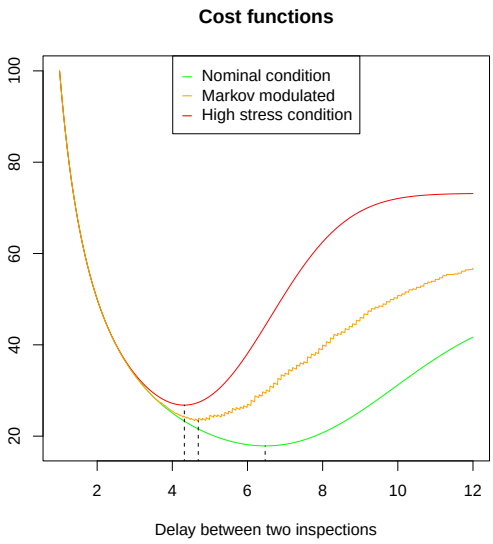


Fig. 2. Cost functions and optimal parameter values

where $C_{arp}^{(0)}$ and $C_{arp}^{(1)}$ denote the cost function of the age replacement policy when the device is respectively under nominal stress and high stress. Such inequalities between cost functions were also observed for the block replacement policy [8]. Moreover, from the comparison between MTTF, it follows that :

$$C_{arp}^{(0)}(\infty) \leqslant C_{arp}(\infty) \leqslant C_{arp}^{(1)}(\infty).$$

However it says nothing about the location of the minima. To end with this section, we provide a numerical illustration of this problem. We consider the same parameters value as above and in addition we set $c_{nf} = 100$ (replacement cost of a non-failed component) and $c_f = 500$ (replacement cost of a failed component). On Figure 2 we have plotted the three cost functions. Since the expression of the cdf in the modulated case is quite too complex to be used for numerical purposes, we have used empirical distribution computed from 1000 simulations. This explains why the cost function in orange is piecewise constant.

We obtain that $\delta_\star^{(1)} = 4.32$, $\delta_\star = 4.69$ and $\delta_\star^{(0)} = 6.47$. As one can expect, the order between optimal inter-inspection delays is also preserved. Similar numerical results were obtained for block replacement policy [8].

IV. CONCLUSION

As the numerical studies seems to point out, the stochastic order between failure times induces an order between optimal inter-inspection delays for two policies, the block replacement policy and the age replacement policy. The proof of these results is still an open problem. Moreover the same problem can be addressed for other maintenance policy and one can expect a similar issue.

REFERENCES

[1] Bagdonavičius, V. and M. Nikulin (2001). Estimation in degradation models with explanatory variables. *Lifetime Data Anal. 7*, 85–103.

[2] C. Bérenguer, A. Grall, L. Dieulle and M. Roussignol. Maintenance policy for a continuously monitored deteriorating system. *Probab. Engrg. Inform. Sci.*, 17(2): 235–250, 2003.

[3] de Souza e Silva, E. and H. Gail (1986). Calculating cumulative operational time distributions of repairable computer systems. *IEEE Trans. Computers C-35*(4), 322–332.

[4] Grall, A., L. Dieulle, C. Bérenguer, and M. Roussignol (2002). Continuous-time predictive-maintenance scheduling for a deteriorating system. *IEEE Trans. Reliab. 51*(2), 141–150.

[5] Lawless, J. and M. Crowder (2005). Accelerated degradation models for failure based on geometric Brownian motion and gamma process. *Lifetime Data Anal. 11*, 511–527.

[6] T. Nakagawa. *Maintenance theory of reliability*. Springer-Verlag, London, 2005.

[7] Park, C. and W. Padgett (2004). Covariates and random effects in a gamma process model with application to degradation and failure. *Lifetime Data Anal. 10*, 213–227.

[8] C. Paroissin and L. Rabehasaina. On the gamma process modulated by a Markov jump process. In C. Bérenguer, A. Grall, and C.G. Soares, editors, *Risk, Reliability and Societal Safety, Proceedings of the European Safety and Reliability Conference 2011 (ESREL 2011), Troyes, France, 18-22 September 2011*. Taylor and Francis, 2011.

[9] Saassouh, B., L. Dieulle, and A. Grall (2007). Online maintenance policy for a deteriorating system with random change of mode. *Reliab. Eng. Syst. Safety 92*, 1677–1685.

[10] Sericola, B. (2000). Occupation times in Markov processes. *Comm. Statist. Stoch. Models 16*(5), 479–510.

[11] Sericola, B. (43). Interval-availability distribution of 2-states systems. *IEEE Trans. Reliability 2*(335–343), 1994.

[12] Singpurwalla, N. (1995). Survival in dynamic environments. *Statistical Scienc 10*(1), 96–103.

[13] van Noortwijk, J. (2009). A survey of the application of gamma processes in maintenance. *Reliab. Eng. Syst. Safety 94*, 2–21.

[14] Zhao, X., M. Fouladirad, and C. Bérenguer (2010). Residual based inspection/replacement policy for a deteriorating system with Markovian covariates. In *Proceeding of international conference on industrial engineering and engineering management, 2010 December 7-12, Macau*, 636–64.

[15] Zhao, X., M. Fouladirad, C. Bérenguer, and L. Bordes (2010). Condition-based inspection/replacement policies for monotone deteriorating systems with environmental covariates. *Reliab. Eng. Syst. Safety 95*(8), 921–934.

A Note on Optimal Allocations for the Second Elementary Symmetric Function with Applications for Optimal Reliability Design

Chien-Yu Peng
Institute of Statistical Science
Academia Sinica
Taipei, 11529, Taiwan
E-mail: chienyu@stat.sinica.edu.tw

***Abstract*—This paper considers the problem of determining the optimal size allocation and optimal number of experimental conditions for the second elementary symmetric function with different coefficients. The current study derives analytical solutions for several practical applications and uses the general formulation to elucidate the foundation between different parametric models found in recent studies. This approach makes some complex problems more tractable than numerical search algorithms use.**

I. Introduction

Before launching a new product into the marketplace, the manufacturer must decide the optimal method to estimate product reliability. Experimental design is a popular methodology that enables engineers and researchers to conduct better experiments, analyze data efficiently and to establish the connection between the original objectives of the investigation and the conclusions of the analysis. Thus, the design of a reliability experiment, such as accelerated tests, may be required. Conducting accelerated tests requires determining the total number of experimental conditions and sample size allocations. Hence, optimal sample size allocation is a major issue in real experiments for efficient equipment use and cost reduction. How to determine the optimal sample size allocation for each test condition before implementing experiments is a practical issue for engineers and researchers.

Many scientific and industrial experiments formulate the optimal allocation problem as follows. Let s be the number of experimental conditions and π_i $(1 \le i \le s)$ be the proportion allocated to the ith experimental condition. For $1 \le r \le s$, consider the constrained optimization problem,

$$\max_{\boldsymbol{\pi}_s \ge \mathbf{0}_s,\ \mathbf{1}_s'\boldsymbol{\pi}_s = 1} f_s^r(\boldsymbol{\pi}_s) = \sum_{1 \le i_1 < \cdots < i_r \le s} \alpha_{i_1,\ldots,i_r}^2 \prod_{j=i_1}^{i_r} \pi_j, \quad (1)$$

where $\boldsymbol{\pi}_s = (\pi_1, \ldots, \pi_s)'$, the $s \times 1$ column vector $\mathbf{1}_s$ $(\mathbf{0}_s)$ has each component equal to 1 (0), the coefficients $\alpha_{i_1,\ldots,i_r} \in \mathbb{R}^+ \equiv (0, \infty)$ contain the information of the i_1th, $\ldots$, i_rth experimental conditions and $\alpha_{i_1,\ldots,i_r} = \alpha_{\Pi(i_1),\ldots,\Pi(i_r)}$ for any permutation Π of $\{i_1, \ldots, i_r\}$. The optimality of an allocation depends on a parametric model and is assessed according to a statistical criterion. For instance, the widely used D-optimality maximizes the determinant of the expected Fisher's information matrix, defined by taking the expectations of the negative second partial derivatives of the log-likelihood function with respect to model parameters. The determinant of the expected Fisher's information matrix can be expressed as a function of $\boldsymbol{\pi}_s$ in the form $f_s^r(\boldsymbol{\pi}_s)$ described in (1).

In accelerated life tests (ALT) with an extreme value regression model, Ng, Balakrishnan and Chan [3] investigated the case of general s and $r = 2$ to obtain the optimal sample size allocation under the D-criterion. For two competing failure modes, Suzuki, Nakamoto and Matsuo [4] studied the case of general s and $r = 2$ to obtain optimal specimen sizes and optimal allocation while minimizing the asymptotic variance of the estimated Weibull shape parameter. However, the authors assumed specific conditions for the coefficients, α_{i_1,i_2}. In accelerated degradation tests (ADT) based on a Wiener process, Tseng, Tsai and Balakrishnan [5] dealt with the case of $s = 2$ and $r = 2$, and derived the exact solution for the optimal allocation under the D-criterion. Yang, Mandal and Majumdar [6] provided some analytical results for the case of $s = 4$ and $r = 3$ to obtain D-optimal allocations with binary responses. For general parametric models with correlated or uncorrelated errors and general criteria, the optimal allocation problems studied previously can be expressed as special cases of the optimization problem in (1).

An explicit solution is highly valued and attractive when numerical methods are computationally intensive for finding an optimal solution. Closed form solutions not only avoid time-consuming algorithms, but also provide direct insights as to how the allocations depend on the primitives of the model. In addition, they can verify the solution obtained by exhaustive search combined with numerical optimization. Although the objective function in (1) with equal coefficients has an attractive algebraic structure, no explicit solution involving general coefficients $\alpha_{i_1,\ldots,i_r}$ is available. Exact solutions can be obtained with small r or under specific structures on the coefficients. For example, if the coefficients $\alpha_{i_1,\ldots,i_r}$ in (1) are equal, it is well-known that the objective function $f_s^r(\boldsymbol{\pi}_s)$ reduces to the rth elementary symmetric function of $\boldsymbol{\pi}_s$. This is strictly Schur-concave (Marshall and Olkin [2, p. 78]). The optimal allocation $\boldsymbol{\pi}_s^*$ is then $\pi_1^* = \cdots = \pi_s^* = 1/s$

by the majorization theory. This means that each condition contributes the same information, and the experiment needs to be done with equal resources. This result is often used as a benchmark to compare with other competing allocations. For $r = 1$, under the assumption $\alpha_1^2 < \cdots < \alpha_s^2$, the optimal allocation is easily seen to be $\pi_1^* = \cdots = \pi_{s-1}^* = 0$ and $\pi_s^* = 1$. If the maximum coefficient is not unique, there are many optimal solutions with the same maximum value. In many practical applications, however, the coefficients $\alpha_{i_1,\ldots,i_r}$ are not equal and the optimization problem is not as simple as linear programming (e.g., $r = 1$). The coefficients may also depend on the design points and model parameters, so that the objective function may not be convex. Hence, traditional methodologies such as the equivalence theorem developed by Kiefer [1] may not be able to solve this type of problem.

For the optimal allocation problem, if any one of the π_i^* is equal to zero, this means that the ith experimental condition is not included and plays no role in the experiments. Reducing the total number of experimental conditions can save manufacturing cost and testing time. Therefore, it is recommended that engineers and researchers reconsider the setting of experimental conditions to obtain additional experiment information (see Example 1 in Section 4). This paper first deals with the case of $r = 2$ in (1) to determine the optimal number of experimental conditions and the optimal sample size allocation. More precisely, the constrained optimization problem is

$$\max_{\boldsymbol{\pi}_s \geq \mathbf{0}_s,\ \mathbf{1}_s'\boldsymbol{\pi}_s = 1} f_s^2(\boldsymbol{\pi}_s) = \sum_{1 \leq i < j \leq s} \alpha_{i,j}^2 \pi_i \pi_j. \tag{2}$$

For $s = 2$, the function $f_2^2(\boldsymbol{\pi}_2)$ is maximized at the optimal allocation $\boldsymbol{\pi}_2^* = (\pi_1^*, \pi_2^*)' = (1/2, 1/2)'$ because the coefficient $\alpha_{1,2}^2$ is independent of π_1 and π_2. The optimal number of experimental conditions for $s = 2$ is not a problem because the difference can only be compared using at least two or more experimental conditions. Hence, we shall only consider (2) with $s \geq 3$.

The remainder of this paper is organized as follows. Section 2 reformulates the constrained optimization problem. Section 3 provides explicit solutions of optimization problems in some practical situations. Section 4 uses two examples to illustrate the proposed optimal allocations. Concluding remarks are given in Section 5.

II. Reformulation of the problem

It may be observed that the objective function in (2) can be written as

$$f_s^2(\boldsymbol{\pi}_s) = \boldsymbol{\pi}_s' \mathcal{M} \boldsymbol{\pi}_s,$$

where $\mathcal{M} = [m_{i,j}]$ is an $s \times s$ symmetric matrix, $m_{i,j} = \alpha_{i,j}^2/2$ for $1 \leq i \neq j \leq s$ and $m_{i,j} = 0$ for $1 \leq i = j \leq s$. But the matrix $\mathcal{M}$ is neither semi-positive definite nor semi-negative definite when $s \geq 3$. Hence, a mathematically tractable result may not be easy to obtain and we need to transform the problem (2) into a more tractable form.

To study the constrained optimization problem, additional notation is needed. Let $\mathcal{H}_{s,k} = [h_{i,j}]$ be an $(s-1) \times (s-1)$ matrix with

$$h_{i,j} = \begin{cases} 2\alpha_{i,k}^2, & \text{for } 1 \leq i = j \neq k \leq s, \\ \alpha_{i,k}^2 + \alpha_{j,k}^2 - \alpha_{i,j}^2, & \text{for } 1 \leq i \neq j \neq k \leq s, \end{cases} \tag{3}$$

and $\boldsymbol{c}_{s,k} = (h_{1,1}, \ldots, h_{s-1,s-1})'/2$. Clearly, for $1 \leq k \leq s$, $\mathcal{H}_{s,k}$ is a symmetric matrix and so $\mathcal{H}_{s,s} = \mathcal{H}_s$ and $\boldsymbol{c}_{s,s} = \boldsymbol{c}_s$ can be used for short. Thus, by substituting $\pi_k = 1 - (\pi_1 + \cdots + \pi_{k-1} + \pi_{k+1} + \cdots + \pi_s)$ into (2), the non-linear constrained optimization problem immediately becomes the following quadratic programming (QP).

$$\text{Minimize } \mathcal{Q}_s(\boldsymbol{\pi}_{s-1}) = \frac{1}{2}\boldsymbol{\pi}_{s-1}' \mathcal{H}_{s,k} \boldsymbol{\pi}_{s-1} - \boldsymbol{c}_{s,k}' \boldsymbol{\pi}_{s-1} \tag{4}$$

subject to the inequality constraint $\mathbf{1}_{s-1}'\boldsymbol{\pi}_{s-1} \leq 1$ and non-negativity of the variables $\boldsymbol{\pi}_{s-1} \geq \mathbf{0}_{s-1}$, where $\boldsymbol{\pi}_{s-1} = (\pi_1, \ldots, \pi_{k-1}, \pi_{k+1}, \ldots, \pi_s)'$.

In our problem, there is an easy and direct way of obtaining analytical solutions in some practical situations. In particular, the case of $s = 3$ is widely considered in numerous applications, and specific conditions on the coefficients $\alpha_{i,j}$ $(1 \leq i < j \leq s)$ are satisfied in practice. Hence, we shall provide optimal allocations in these cases in the following section.

III. Analytical Solutions

For very expensive or newly developed products which only have a few available test units on hand, the optimal allocation for $s = 3$ is useful in initial tests. The following first derives exact results for the optimal allocation and optimal number of experimental conditions for $s = 3$.

Theorem 1. *Let $\alpha_{i,j} \in \mathbb{R}^+$ and $\alpha_{i,j} = \alpha_{j,i}$ for $1 \leq i < j \leq 3$. Under the condition $\alpha_{i,k}, \alpha_{j,k} < \alpha_{i,j}$ for $1 \leq i \neq j \neq k \leq 3$, if $\alpha_{i,k}^2 + \alpha_{j,k}^2 > \alpha_{i,j}^2$, then the optimal sample size allocation is*

$$\pi_i^* = \alpha_{j,k}^2(\alpha_{i,j}^2 + \alpha_{i,k}^2 - \alpha_{j,k}^2)/|\mathcal{H}_{3,k}| \in (0, 1/2),$$

where the subscripts i, j, $k = 1, 2, 3$ are distinct and the corresponding optimal value is

$$f_3^2(\boldsymbol{\pi}_3^*) = \alpha_{1,2}^2 \alpha_{1,3}^2 \alpha_{2,3}^2/|\mathcal{H}_{3,k}|;$$

otherwise $\pi_i^* = \pi_j^* = 1/2$ and $\pi_k^* = 0$, where $|\mathcal{H}_{3,k}| = 4\alpha_{i,k}^2\alpha_{j,k}^2 - (\alpha_{i,j}^2 - \alpha_{i,k}^2 - \alpha_{j,k}^2)^2$.

It can be clearly seen that there are simple sufficient conditions $\alpha_{i,k}, \alpha_{j,k} < \alpha_{i,j}$ and $\alpha_{i,k}^2 + \alpha_{j,k}^2 \leq \alpha_{i,j}^2$ $(1 \leq i \neq j \neq k \leq 3)$ for the optimal number of experimental conditions needing to be two. According to the value of the index k in $\alpha_{i,k}^2 + \alpha_{j,k}^2 \leq \alpha_{i,j}^2$, one gets $\pi_k^* = 0$. For example, let $(i, j, k) = (1, 2, 3)$. Then under the condition $\alpha_{1,3}, \alpha_{2,3} < \alpha_{1,2}$, if $\alpha_{1,3}^2 + \alpha_{2,3}^2 > \alpha_{1,2}^2$, then the optimal allocations are

$$\pi_1^* = \alpha_{2,3}^2(\alpha_{1,2}^2 + \alpha_{1,3}^2 - \alpha_{2,3}^2)/|\mathcal{H}_3|,$$
$$\pi_2^* = \alpha_{1,3}^2(\alpha_{1,2}^2 + \alpha_{2,3}^2 - \alpha_{1,3}^2)/|\mathcal{H}_3|$$

and

$$\pi_3^* = \alpha_{1,2}^2(\alpha_{1,3}^2 + \alpha_{2,3}^2 - \alpha_{1,2}^2)/|\mathcal{H}_3|;$$

otherwise $\boldsymbol{\pi}_3^* = (1/2, 1/2, 0)'$.

Unfortunately, for $s \geq 4$, simple sufficient conditions in terms of a set of inequalities of lower degree does not exist for optimizing the original problem because the determinant $|\mathcal{H}_{s,k}|$ (or $|\mathcal{M}|$) is irreducible. But in the resource allocation and balancing problems of accelerated tests and engineering, there are some structures about the coefficients $\alpha_{i,j}$ ($1 \leq i < j \leq s$). Hence, the analytical solution $\boldsymbol{\pi}_s^*$ can be obtained in the following cases.

Let $\mathcal{A} = \{\alpha_{i,j} | 1 \leq i < j \leq s\}$, $\mathcal{A}_k = \{\alpha_{i,j} | 1 \leq i < j \leq s, i = k \text{ or } j = k\}$ for $1 \leq k \leq s$ and $\alpha_{l_1,l_2} = \max_{1 \leq i \neq j \leq s} \alpha_{i,j}$.

(I) There is a subset $\mathcal{A}_k$ and $\alpha_{i,k} - \alpha_{j,k} = \alpha_{i,j}$ for the distinct subscripts i, j, $k = 1, \ldots, s$.

(II) The coefficients $\alpha_{i,j}$ satisfy $\alpha_{i,k}^2 + \alpha_{j,k}^2 = \alpha_{i,j}^2$ for $1 \leq i < k < j \leq s$.

Let $f(\cdot|\text{I})$ and $f(\cdot|\text{II})$ denote the objective function $f(\cdot)$ in (2) under assumptions (I) and (II), respectively. We have the following results.

Theorem 2. *Let $\alpha_{i,j} \in \mathbb{R}^+$ and $\alpha_{i,j} = \alpha_{j,i}$ for $1 \leq i \neq j \leq s$. If the coefficient $\alpha_{i,j}$ satisfies assumption (I), then the optimal number of experimental conditions is two and the corresponding optimal sample size allocation is $\pi_{l_1}^* = \pi_{l_2}^* = 1/2$ and $\pi_l^* = 0$ for the distinct subscripts l, l_1, $l_2 = 1, \ldots, s$, where $k = l_1$ or l_2.*

The result is a generalization of Theorem 3 in Ng et al. [3], which assumes a specific structure on the coefficients. According to Theorem 2, we do not have to assign units at any experimental condition other than the l_1th and the l_2th experimental conditions. One should allocate 50% of the units to the l_1th experimental condition and the remaining 50% of the units to the l_2th experimental condition. This means that the remaining experimental conditions play no role in this optimization problem.

The following case demonstrates an interesting result under assumption (II). Without loss of generality, assume that $\alpha_{1,s}^2 \leq \cdots \leq \alpha_{s-1,s}^2$ under assumption (II).

Theorem 3. *Let $\alpha_{i,j} \in \mathbb{R}^+$ and $\alpha_{i,j} = \alpha_{j,i}$ for $1 \leq i \neq j \leq s$. Under $s \geq 4$ and assumption (II), there exists an integer j ($1 \leq j \leq s-3$) such that for $j \leq i \leq s-1$,*

$$\begin{cases} \pi_i^* = \dfrac{1}{2} - \dfrac{s-2-j}{2\alpha_{i,s-j+1}^2 \mathbf{1}_{s-j}' \boldsymbol{c}_{s-j+1}^{-1}}, \\ \qquad\qquad \text{if} \quad \mathbf{1}_{s-j}' \boldsymbol{c}_{s-j+1}^{-1} > \dfrac{s-2-j}{\alpha_{i,s-j+1}^2}, \\ \pi_1^* = \cdots = \pi_j^* = \pi_s^* = 0, \\ \qquad \text{if} \quad \dfrac{s-2-j}{\alpha_{i+1,s-j+1}^2} < \mathbf{1}_{s-j}' \boldsymbol{c}_{s-j+1}^{-1} \leq \dfrac{s-2-j}{\alpha_{i,s-j+1}^2}. \end{cases} \tag{5}$$

Moreover, the maximum is

$$f_s^2(\boldsymbol{\pi}_s^*|\text{II}) = \mathbf{1}_{s-j}' \boldsymbol{c}_{s-j+1}/4 - (s-j-2)^2/(4\mathbf{1}_{s-j}' \boldsymbol{c}_{s-j+1}^{-1}).$$

From Theorem 3, we have $s-3$ candidate optimal allocations and the optimal number of experimental conditions depends on the value of the index j in (5). Because the order of $\alpha_{i,s}^2$ is fixed, the sequence $\{\pi_i^*\}_{i=1}^{s-1}$ is increasing and no π_i^* is greater than $1/2$. In addition, $\pi_s^* = 0$ means that the sth experimental condition is redundant and the middle experimental conditions provide the most information for the corresponding proportions.

IV. Applications

To show the applicability and effectiveness of the theory described in this paper, two examples are presented.

Example 1. Tseng et al. [5] discuss the optimal sample size allocation problem under ADT with multiple levels of stress when a Wiener process is used to describe the LED's degradation path. For 3-stress levels (i.e., $s = 3$) and the D-criterion, comparing our settings with theirs, we have

$$\alpha_{i,j} = \beta_0 \exp\left\{\beta_1 (S_i^{-1} + S_j^{-1})\right\} (S_i^{-1} - S_j^{-1}). \tag{6}$$

The estimates of parameters and the setting of stress levels are given by $(\beta_0, \beta_1) = (0.5335, -1552.5)$ and $(S_1, S_2, S_3) = (338.15, 353.15, 378.15)$, respectively. The corresponding $\alpha_{i,j}^2$s are

$$(\alpha_{1,2}^2, \alpha_{1,3}^2, \alpha_{2,3}^2) = (7.015 \times 10^{-17}, 7.781 \times 10^{-16}, 4.116 \times 10^{-16}).$$

It is easily seen that $\max_{1 \leq i < j \leq 3} \alpha_{i,j} = \alpha_{1,3}$ and $\alpha_{1,2}^2 + \alpha_{2,3}^2 < \alpha_{1,3}^2$. By Theorem 1, we only need to assign the same number of test units at the lowest and highest stress levels. This result coincides with that obtained by the extensive discrete search procedure in Tseng et al. [5].

Note that when an optimal sample size allocation is used in practice, the optimal number $N\boldsymbol{\pi}_s^*$ of test units at each experimental condition must be rounded to an integer, where N denotes the total number of test units. Hence, the allocation may be one unit away from the optimum.

We calculate the efficiency of an allocation defined by

$$\text{eff}(\boldsymbol{\pi}_3) = f_3^2(\boldsymbol{\pi}_3)/f_3^2(\boldsymbol{\pi}_3^*).$$

Efficiencies of the different allocations are shown in Table I. We observe that the benchmark allocation $(\pi_1, \pi_2, \pi_3) = \frac{1}{3}(1, 1, 1)$ yields reasonable efficiency (i.e., 71.96%) with respect to all cases, and the efficiency of the statistical analysis could have been improved by allocating units according to optimal allocation (see the first row in Table I). Moreover, the $4:2:1$ plan, which allocates $4/7$, $2/7$ and $1/7$ of all units, respectively, is widely used as a compromise plan in accelerated tests. In this case in terms of parameter estimates and stress-level setting, the $4:2:1$ plan has poor efficiency (i.e., 47.18%) relative to the optimal allocation.

If we take proportional allocations as a function of the median stress level S_2, Fig. 1 illustrates the characteristics of the multiple optimal allocations. There is no change of the optimal allocation for different values of S_2. The optimal allocation is $\pi_1^* = \pi_3^* = 1/2$ and $\pi_2^* = 0$. But if we change S_1 from 338.15 to 273.15, Fig. 2 shows the changing trend of proportional allocations for different values of S_2.

TABLE I
EFFICIENCIES OF ALLOCATIONS

(π_1, π_2, π_3)	eff($\boldsymbol{\pi}_3$)
$(1, 0, 1)/2$	100.00%
$(0, 1, 1)/2$	52.90%
$(1, 1, 0)/2$	9.02%
$(1, 1, 1)/3$	71.96%
$(4, 2, 1)/7$	47.18%

As shown in this figure, we only need two stresses for the range $(273.15, 326.09) \cup (339.72, 378.15)$. More precisely, the median and largest stresses are needed in the interval $(273.15, 326.09)$ and the smallest and largest stresses are needed in the interval $(339.72, 378.15)$. In the interval $(326.09, 339.72)$, as the stress level S_2 increases, π_1^* increases and π_2^* decreases. This may be because the stress level S_2 is farther away from S_1, so that more units are allocated to the smallest stress to get more information. In other words, when the stress level S_2 is close to S_1, more units are allocated to the median stress instead of the smallest stress. In addition, the optimal proportion π_3^* is almost 0.5, no matter what the median stress level is. This means that the level of the highest stress has valuable information for the ADT with Wiener process. Comparing Fig. 1 with Fig. 2, one can see that the chosen range of the stress level in the original setting is not wide enough.

Example 2. Ng et al. [3] deal with the optimal sample size allocation for ALT with an extreme value regression model. Suppose we have s ordered stress levels (S), say $S_1 < \cdots < S_s$. According to their setting and the D-optimal criterion, the coefficient $\alpha_{i,j}$ is expressed as $\alpha_{i,j} = S_j - S_i$ for $1 \le i < j \le s$. It is easily seen that $\max_{1 \le i < j \le s} \alpha_{i,j} = \alpha_{1,s}$ is in the subset $\mathcal{A}_s$ and $\alpha_{i,s} - \alpha_{j,s} = (S_s - S_i) - (S_s - S_j) = \alpha_{i,j}$ for $1 \le i < j \le s-1$. By Theorem 2, the optimal sample size allocation is then $\pi_1^* = \pi_s^* = 1/2$ and $\pi_k^* = 0$ for $k = 2, \ldots, s-1$. This means that we do not have to assign units at any experimental condition other than the first and last stress levels. The middle stress levels play no role in the ALT.

Note that, although the optimal allocation (i.e., the first and last stress levels) in Example 2 is the same as in Example 1, the structure of coefficients in (6) has a nonlinear relationship between the parameters.

V. CONCLUSION

The results presented in this paper provide sufficient conditions for analytical solutions of the second elementary symmetric function with different coefficients. Many practical applications typically use optimal sample size allocation, particularly for data analysis from experiments in early scientific studies. Hence, the implicit results from this paper are important for a comparative or screening experiment, which may involve a large number of experimental conditions. Whatever parametric model, criterion, and number of parameters used, the analytical results presented herein go beyond solving problems on a case-by-case basis. The use of the theorems enables tackling more complex problems than numerical search algorithms currently allow. The connection between the optimization problem (1) and the geometrical implication raises many other interesting issues that deserve further investigation.

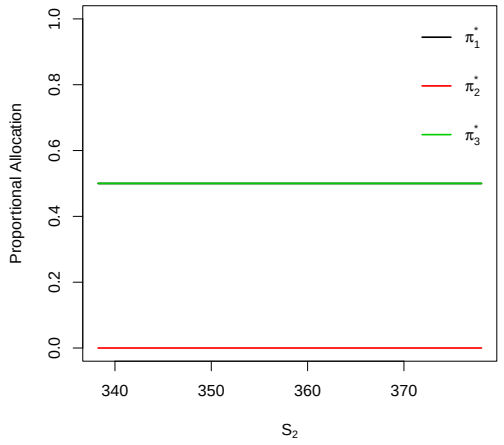


Fig. 1. Proportional allocations as a function of the median stress level showing different optimal allocations ($S_1 = 338.15$).

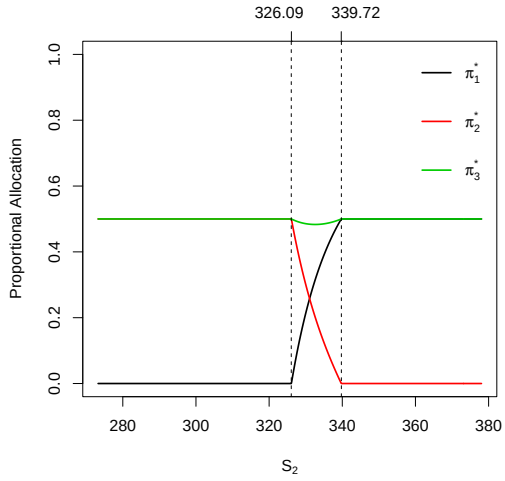


Fig. 2. Proportional allocations as a function of the median stress level showing different optimal allocations ($S_1 = 273.15$).

REFERENCES

[1] J. Kiefer, "Optimal designs in regression problems, II," *Annals of mathematical Statistics*, vol. 32, pp. 298–325, 1961.

[2] A. W. Marshall and I. Olkin, *Inequalities: Theory of Majorization and Its Application*, New York: Academic Press, 1979.

[3] H. K. T. Ng, N. Balakrishnan and P. S. Chan, "Optimal sample size allocation for tests with multiple levels of stress with extreme value regression," *Naval Research Logistics*, vol. 54, pp. 237–249, 2007.

[4] K. Suzuki, T. Nakamoto and Y. Matsuo, "Optimum specimen sizes and sample allocation for estimating Weibull shape parameters for two competing failure modes," *Technometrics*, vol. 52, pp. 209–220, 2010.

[5] S. T. Tseng, C. C. Tsai and N. Balakrishnan, *Optimal sample size allocation for accelerated degradation test based on Wiener process*, *Methods and Applications of Statistics in Engineering, Quality Control, and the Physical Sciences*, N. Balakrishnan (Editor), New York: Addison-Wesley, 2011, pp. 330–343.
[6] J. Yang, A. Mandal and D. Majumdar, "Optimal designs for two-level factorial experiments with binary response," *Statistica Sinica*, April 2012, to be published.

A Viterbi algorithm for hidden semi-Markov models

Pertsinidou Christina-Elisavet and Limnios Nikolaos

Université de Technologie de Compiégne,
Laboratoire de Mathématiques Appliquées de Compiègne,
Centre de Recherches de Royallieu, BP 20529,
60205 Compiègne, Cedex, France
pertsict@math.auth.gr nikolaos.limnios@utc.fr

***Abstract*—Hidden Markov models (HMM) are important tools in estimation and analysis of biological sequences and many other systems. The hidden semi-Markov models (HSMM) are an important generalization of HMM. See, e.g. Barbu and Limnios [1].**

We present a Viterbi algorithm for hidden semi-Markov models of SM1-M0 type. In our knowledge this kind of algorithm for HSMM is proposed for the first time. In HSMM the unobserved process is a semi-Markov chain and the observed process is independent conditionally on the values of the semi-Markov chain. Let $E = \{e_1, ..., e_d\}$ and $A = \{a_1, ..., a_s\}$ two finite sets. Consider also the double chain $(Z_k, Y_k)_{k\geq 0}$, where (Z_k) is an E-valued semi-Markov chain, with semi-Markov kernel $q(k) = (q_{ij}(k); i, j \in E)$, $k \in \mathbb{N}$, and initial probability α. Let $(J_n, S_n)_{n\geq 0}$ the corresponding to Z Markov renewal chain, where $S_0 = 0 < S_1 < ... < S_n < ...$ are the $\mathbb{N}$-valued jump times, and (J_n) the successive visited states, we have:

$$q_{ij}(k) = \mathbb{P}(J_n = j, S_n - S_{n-1} = k \mid J_{n-1} = i),$$

where the imbedded Markov chain (J_n) is determined by the probability transition matrix

$$P = (p_{ij}, i, j \in E)$$

with

$$P_{ij} = P(J_{n+1} = j | J_n = i)$$

The associated backward recurrence Markov chain (Z, U) transition probabilities are given by the equation:

$$\tilde{p}_{(i,t_1)(j,t_2)} = \begin{cases} q_{ij}(t_1+1)/\overline{H}_i(t_1) & \text{if } i \neq j,\ t_2 = 0 \\ \overline{H}_i(t_1+1)/\overline{H}_i(t_1) & \text{if } i = j,\ t_2 - t_1 = 1 \\ 0 & \text{elsewhere} \end{cases},$$

The process (Y_k) is the observed process, with distribution depending on the values of (Z_k), i.e,

$$\mathbb{P}(Y_k = y \mid Z_k = x) = R(x, y), \quad (x, y) \in E \times A.$$

In that case, the process $(Z_k, Y_k)_{k\geq 0}$ is called a HSMM. The aim is to observe a realisation of $(Y_k, 0 \leq k \leq n)$ i.e, $y_0^n = (y_0, ...y_n) \in A^{n+1}$ and try to find out the corresponding hidden regime, i.e, $q_0^n = (q_0, ...q_n) \in E^{n+1}$ a realisation of $(Z_k, 0 \leq k \leq n)$. The proposed algorithm an example on DNA analysis and a second example are given.

***Index Terms*—Hidden semi-Markov model, hidden markov model, Viterbi algorithm.**

I. Algorithm

Algorithm

The proposed new algorithm is the following one

1) Initial conditions. For $k = 0$,

$$d_0(i_0) = \log_2(\tilde{\alpha}(i_0)) + \log_2(R(i_0, y_0))$$
$$b_0(i_0) = 0.$$

2) For $k \geq 1$

$$d_k(i_k) = \max_{i_{k-1}\in E}[d_{k-1}(i_{k-1}) + \log_2(a_k(i_0^k))] + \\ + \log_2(R(i_k, y_k))$$
$$b_k(i_k) = \arg\max_{i_{k-1}\in E}[d_{k-1}(i_{k-1}) + \log_2(a_k(i_0^k))]$$

3) Termination
If $k = T$ (T is the number of observations),

$$P = \max_{i_T\in E}[d_T(i_T)],$$
$$P^* = 2^{max_{i_T\in E}[d_T(i_T)]},$$
$$q_T = \arg\max_{i_T\in E}[d_T(i_T)],$$

where

$$a_k(i_0^k) = \sum_{l_1=0}^{k-1} \sum_{l_2\in\{0,l_1+1\}} \tilde{p}_{\{(i_{k-1},l_1)(i_k,l_2)\}} \times \\ \times \mathbf{1}_{\{i_{k-1}=...=i_{k-l_1-1}\neq i_{k-l_1-2}\}}, \quad k \geq 1.$$

4) Sequence of states with backward-forward steps

$$q_k = b_{k+1}(q_{k+1}), \quad k = T-1, \ldots, 0.$$

A. Example

Real DNA sequences are inhomogeneous and can be described by a hidden Markov model with hidden states representing different types of nucleotide composition. Consider an HMM that includes two hidden states H and L for higher and lower $C+G$ content, respectively. Initial probabilities for both H and L are equal to 0.5, while transition probabilities are as follows: $p_{HH} = 0.5$, $p_{HL} = 0.5$, $p_{LL} = 0.6$, $p_{LH} = 0.4$. Nucleotides T, C, A, G are emitted from states H and L with probabilities 0.2, 0.3, 0.2, 0.3, and 0.3, 0.2,

0.3, 0.2, respectively. We use the Viterbi algorithm to define the most likely sequence of hidden states for the sequence $x = GGCACTGAA$.

The Viterbi algorithm for the hidden Markov models, see [2], [3] , in logarithmic mode (base 2) is of the form

Algorithm

1) Initial conditions. For $k = 1$,

$$d_1(i_1) = \log_2(\tilde{\alpha}(i_1)) + \log_2(R(i_1, y_1))$$

$$b_1(i_1) = 0.$$

2) For $k > 1$

$$d_k(i_k) = \max_{i_{k-1} \in E}[d_{k-1}(i_{k-1}) + \log_2(p_{i_{k-1},i_k})] + \\ + \log_2(R(i_k, y_k))$$

$$b_k(i_k) = \arg \max_{i_{k-1} \in E}[d_{k-1}(i_{k-1}) + \log_2(p_{i_{k-1},i_k})]$$

3) Termination
If $k = T$ (T is the number of observations),

$$P = \max_{i_T \in E}[d_T(i_T)],$$

$$P^* = 2^{max_{i_T \in E}[d_T(i_T)]},$$

$$q_T = \arg \max_{i_T \in E}[d_T(i_T)],$$

where p_{i_{k-1},i_k} is the transition probability matrix.

4) Sequence of states with backward-forward steps

$$q_k = b_{k+1}(q_{k+1}), \quad k = T-1, \ldots, 0.$$

The implementation of the hidden Markov Viterbi algorithm and the hidden semi-Markov Viterbi algorithm as described above, gives exactly the same results. In the sequel we will analyze the results from the hidden semi-Markov Viterbi algorithm implementation.

The semi-Markov kernel of a Markov chain is given by the following equation

$$q_{ij}(k) = \begin{cases} p_{ij}(p_{ii})^{k-1}, & \text{if } i \neq j \text{ and } k \in \mathbb{N}^*, \\ 0, & \text{elsewhere .} \end{cases}$$

So for our example it is

$$q_{12}(k) = 0.5(0.5)^{k-1}$$

$$q_{21}(k) = 0.4(0.6)^{k-1}$$

$$\bar{H}_1(k) = 1 - \sum_{m=1}^{k} q_{12}(m)$$

$$\bar{H}_2(k) = 1 - \sum_{m=1}^{k} q_{21}(m).$$

For $k = 0$

$$d_0(1) = \log_2(R(1, G)) + log_2\tilde{a}_{(1)} = \\ = -2.736967$$

$$b_0(1) = 0$$

$$d_0(2) = \log_2(R(2, G)) + log_2\tilde{a}_{(2)} = \\ = -3.32193$$

$$b_0(2) = 0$$

For $k = 1$

$$d_1(1) = \log_2(R(1, G)) + d_0(1) + log_2\tilde{p}_{(1,0)(1,1)} = \\ = -5.47393$$

$$b_1(1) = 1$$

$$d_1(2) = \log_2(R(2, G)) + d_0(1) + log_2\tilde{p}_{(1,0)(2,0)} = \\ = -6.05889$$

$$b_1(2) = 1$$

For $k = 2$

$$d_2(1) = \log_2(R(1, C)) + d_1(1) + log_2\tilde{p}_{(1,1)(1,2)} = \\ = -8.2109$$

$$b_2(1) = 1$$

$$d_2(2) = \log_2(R(2, C)) + d_1(1) + log_2\tilde{p}_{(1,1)(2,0)} = \\ = -8.79586$$

$$b_2(2) = 1$$

For $k = 3$

$$d_3(1) = \log_2(R(1, A)) + d_2(1) + log_2\tilde{p}_{(1,2)(1,3)} = \\ = -11.5328$$

$$b_3(1) = 1$$

$$d_3(2) = \log_2(R(2, A)) + d_2(1) + log_2\tilde{p}_{(1,2)(2,0)} = \\ = -10.9479$$

$$b_3(2) = 1$$

For $k = 4$

$$d_4(1) = \log_2(R(1, C)) + d_3(2) + log_2\tilde{p}_{(2,0)(1,0)} = \\ = -14.0068$$

$$b_4(1) = 2$$

$$d_4(2) = \log_2(R(2, C)) + d_3(2) + log_2\tilde{p}_{(2,0)(2,1)} = \\ = -14.0068$$

$$b_4(2) = 2$$

For $k = 5$

$$d_5(1) = \log_2(R(1,T)) + d_4(1) + log_2\tilde{p}_{(1,0)(1,1)} = \\ = \ -17.3287$$

$$b_5(1) = 1$$

$$d_5(2) = \log_2(R(2,T)) + d_4(2) + log_2\tilde{p}_{(2,1)(2,2)} = \\ = \ -16.4807$$

$$b_5(2) = 2$$

For $k = 6$

$$d_6(1) = \log_2(R(1,G)) + d_5(2) + log_2\tilde{p}_{(2,2)(1,0)} = \\ = \ -19.5396$$

$$b_6(1) = 2$$

$$d_6(2) = \log_2(R(2,G)) + d_5(2) + log_2\tilde{p}_{(2,2)(2,3)} = \\ = \ -19.5396$$

$$b_6(2) = 2$$

For $k = 7$

$$d_7(1) = \log_2(R(1,A)) + d_6(1) + log_2\tilde{p}_{(2,2)(1,0)} = \\ = \ -22.8615$$

$$b_7(1) = 1$$

$$d_7(2) = \log_2(R(2,A)) + d_6(2) + log_2\tilde{p}_{(2,3)(2,4)} = \\ = \ -22.0135$$

$$b_7(2) = 2$$

For $k = 8$

$$d_8(1) = \log_2(R(1,A)) + d_7(2) + log_2\tilde{p}_{(2,4)(1,0)} = \\ = -25.6574$$

$$b_8(1) = 2$$

$$d_8(2) = \log_2(R(2,A)) + d_7(2) + log_2\tilde{p}_{(2,4)(2,5)} = \\ = \ -24.4874$$

$$b_8(2) = 2$$

By the traceback procedure we have the most likely hidden state sequence

$$q_0^8 = \{1,1,1,2,2,2,2,2,2\}$$

with probability

$$P_8^* = 4.25153 \times 10^{-8}$$

B. Example

We assume that we have two hidden states 1 and 2. The transition matrix is of the standard form and the distributions of the waiting times are $q_{12}(k) = \frac{e^{-5}5^k}{k!}, k \geq 0$ and $q_{21}(k) = \frac{e^{-3}3^k}{k!}, k \geq 0$ respectively, where by q_{ij} we denote the semi-Markov kernel. The observations H and T are emitted from the hidden states 1,2 with emission probabilities $R(1,H) = 0.2$, $R(1,T) = 0.8$, $R(2,H) = 0.7$ and $R(2,T) = 0.3$. We use the Viterbi algorithm for hidden semi-Markov models, as described above, to define the most likely sequence of hidden states for the observed sequence $TTTTTTTHHHTHHTTT$.We got this sequence by simulation and it will be the data for our algorithm. We have the following results

$$d_0(1) = -1.32193, d_0(2) = -2.73697$$

$$b_0(1) = 0, b_0(2) = 0$$

$$d_1(1) = -1.6933, d_1(2) = -4.70731$$

$$b_1(1) = 1, b_1(2) = 2$$

$$d_2(1) = -2.14679, d_2(2) = -6.88529$$

$$b_2(1) = 1, b_2(2) = 2$$

$$d_3(1) = -2.71878, d_3(2) = -6.5354$$

$$b_3(1) = 1, b_3(2) = 1$$

$$d_4(1) = -3.43014, d_4(2) = -6.5354$$

$$b_4(1) = 1, b_4(2) = 1$$

$$d_5(1) = -4.28715, d_5(2) = -6.85733$$

$$b_5(1) = 1, b_5(2) = 1$$

$$d_6(1) = -5.28527, d_6(2) = -7.44229$$

$$b_6(1) = 1, b_6(2) = 1$$

$$d_7(1) = -8.4108, d_7(2) = -7.02726$$

$$b_7(1) = 1, b_7(2) = 1$$

$$d_8(1) = -11.6376, d_8(2) = -7.77521$$

$$b_8(1) = 1, b_8(2) = 2$$

$$d_9(1) = -12.0219, d_9(2) = -8.73079$$

$$b_9(1) = 2, b_9(2) = 2$$

$$d_{10}(1) = -10.5365, d_{10}(2) = -11.1061$$

$$b_{10}(1) = 2, b_{10}(2) = 2$$

$$d_{11}(1) = -12.9079, d_{11}(2) = -12.4001$$

$$b_{11}(1) = 1, b_{11}(2) = 2$$

$$d_{12}(1) = -15.3614, d_{12}(2) = -13.7254$$

$$b_{12}(1) = 1, b_{12}(2) = 2$$

$$d_{13}(1) = -15.4546, d_{13}(2) = -16.1451$$
$$b_{13}(1) = 2, b_{13}(2) = 2$$
$$d_{14}(1) = -15.826, d_{14}(2) = -18.3152$$
$$b_{14}(1) = 1, b_{14}(2) = 2$$
$$d_{15}(1) = -16.2795, d_{15}(2) = -20.2553$$
$$b_{15}(1) = 1, b_{15}(2) = 2$$

And by the traceback procedure we have the most likely hidden state sequence,

$$q_0^{15} = \{1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 1, 1, 1\}$$

with probability

$$P_{15}^* = 0.0000125716.$$

REFERENCES

[1] V.S. Barbu and N. Limnios, *Semi-Markov Chains and Hidden Semi-Markov Models toward Applications*. New York: Springer, 2008.

[2] L.R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc.IEEE*, vol. **77**: pp. 257–286, 1989.

[3] T.Koski, *Hidden Markov Models for Bioinformatics*. Dordrecht: Kluwer, 2001.

RELIABILITY MODELING AND OPTIMIZATION USING FUZZY LOGIC AND CHAOS THEORY

A. Rotshtein[1], L. Pustylnik[2]

[1] Dep. of Industrial Engineering and Management, Jerusalem College of Technology - Machon Lev, Israel. E-mail: rot@jct.ac.il

[2] Mechanical Engineering, Afeka -Tel Aviv Academic College of Engineering, Israel.

ABSTRACT

Fuzzy sets membership functions integrated with logistic map as the chaos generator were used to create reliability bifurcations diagrams of the componentwise redundancy system. This paper shows that increasing in the number of redundant components results in a postponement of the moment of the first bifurcation which is considered as most contributing to the loss of the reliability. The increasing of redundancy also provides the shrinkage of the oscillation orbit of the level of the system's membership to reliable state. The paper includes the problem statement of redundancy optimization under conditions of chaotic behavior of influencing parameters and genetic algorithm of this problem solving. The paper shows the possibility of chaos-tolerant systems design with the required level of reliability.

Keywords: system reliability, redundancy, fuzzy logic, chaos theory, reliability bifurcations, optimization, genetic algorithm.

1. INTRODUCTION

The classical reliability theory [1, 2] is based on the probabilistic approach. Essential limitations of this approach are connected with "the problem of the raw data" which depend on many factors and which may not correspond to the real conditions of the system's action. Besides, the statistical data used in the probabilistic models of the reliability fix only the facts of real failures, and don't contain the information about the causes of these failures. Whereas the causes of these failures are connected with the elements' variables (temperature, humidity, tension, etc.), which become more (or less) than a certain critical level. So, we can affirm that the probabilistic theory [1, 2] models the reliability in the space of events-consequences (i.e. failures) and suits badly for the reliability modelling in the space of events-causes (i.e. variables).

The alternative for the probabilistic modeling of the reliability is the approach based on the fuzzy logic [3] and related possibility theory [4]. In this case the classical "failure probability" is replaced by "failure possibility" which is modeled by the membership function of the system (or the element) variables to the reliable state.

The explicit dependence of the membership function on the variables (failure causes) makes convenient the integration of the fuzzy model of reliability with the technique of time series [5], which allows observing the change of the reliability level in the real time.

The chaos theory is a new approach to the analysis of nonlinear time series [6]. It uses the conceptual apparatus of the theory of nonlinear oscillations [7] and purposes to study the phase portrait of the dynamical system with its intrinsic states of stability (attractors) and bifurcations, i.e. "jumpings" between stable states. Unlike the classical oscillation theory [7] where the phase portrait is formed on the base of the system description by means of differential equations, the chaos theory [6] offers the methods of the phase-portrait extraction from the experimental data, i.e. directly from the time series.

The integration of the fuzzy model of the reliability with the phase portraits of variables creates preconditions for the construction of the phase portrait reflecting *the system*

reliability dynamics. The pattern of bifurcations which may be interpreted as failure instants is of particular interest.

The works on the fuzzy reliability theory began in 90s of the last century. The first specialized collection of articles in this field is the work [8]. The first monographs containing the approaches to the construction of the fuzzy reliability theory are [9-11]. In the works [12, 13] there is fuzzy algorithm approach to the reliability modeling based on the algebra of regular algorithms [14] and expert assessments of the performance correctness of operators and conditions by means of membership functions [3]. The tasks of the reliable optimization of the control resources on the base of fuzzy algorithm approach are solved in [15].

In the work [16] there is an approach to the on-line reliability evaluation on the base of the integration of the fuzzy logic and forecasting methods of time series: the exponential smoothing and Kalman filtering.

The idea of the application of the chaos theory in reliability modeling appears in [17]. Two real data bases about software failures are processed by the methods of chaos theory in the work [18]. It was shown that the deterministic model of failures is more adequate to the experimental data than the traditional stochastic models, for example, the modified Poisson's law etc. The results of the work [18] can be considered as a new approach (alternative to the statistical) to the data processing about the failures on the level of elements. We don't know the publications about the application of chaos theory for the reliability modeling on the level of the system taking into account its structure.

In this connection there are following questions:

1) How does the system structure influence the phase portrait of reliability?
2) Is it possible to solve problems of the redundancy optimization with the deterministic (chaotic) order of the occurrences of failures?

As far as the failures are connected with the oscillations of variables, then the answers for these questions should be searched on the base of the integration of the fuzzy logic and chaos theory.

In this article it is used a fuzzy algorithm approach to the modeling of the parametric reliability proposed in [12, 13], and the simple generator of chaotic oscillations of variables in the form of logistic function [19]. The further description is organized in a following way.

The section 2 describes the principles of the modelling of the reliability dynamics by means of the composition of the membership function and chaos generator.

The section 3 examines the fuzzy model of the element reliability with the multiple redundancy. There are the results of computer experiments on the analysis of bifurcations of the reliability level depending on the number of backup components.

The section 4 considers the task of the redundancy optimization under the conditions of chaotic oscillations of the element variables.

2. BASIC PRINCIPLES

The fuzzy chaotic approach to the modeling of the reliability dynamics has the following principles.

2.1. Fuzzy correctness

According to this principle introduced in [12, 13] there is not a distinguished boundary between "correct" (1) and "incorrect" (0) results of the functioning of the system and its elements. For the formal evaluation of the correctness level it is used the multidimensional (by the number of variables) membership function $\mu^1(x_1, x_2, ...)$ which depends on the

measured variables (input variables). The correctness of each variable is determined by the membership function $\mu^1(x_i)$ of the variable x_i to the correct value.

The function $\mu^1(x_i)$ can be interpreted as the *correctness distribution of the variable* x_i: extreme cases $\mu^1(x_i)=1$ (0) correspond to the maximal (minimal) level of the correctness of the *variable* x_i. Pay attention that the correctness distribution $\mu^1(x_i)$ satisfies the axiomatics of the fuzzy-set theory [3], in contrast to the probabilistic distributions used in the classical reliability theory [1, 2].

For example, the typical correctness distributions (membership functions) correspond to three possible cases of fuzzy boundaries between "correct" (1) and "incorrect" (0):

a) correct (1) - incorrect (0),

b) incorrect (0) - correct (1) - incorrect (0),

c) incorrect (0) correct (1).

2.2. Integration of membership functions and time series

It is assumed that for the variable x it is known the time series of its values $(x_1, x_2, x_3,...)$ in discrete moments of time $(t_1, t_2, t_3,...)$. Putting these values in the membership function $\mu^1(x)$, we receive the dynamics of the correctness level of the variable x in the form of the function $\mu^1[x(t)]$.

2.3. Chaos generator

Chaos means the oscillations which seem random but in truth they are generated by the deterministic nonlinear model. In [6] about 40 models of the chaos are described. Each model contains variables whose values must be fitted on the base of the experimental data. Algorithm of the chaos generation is explained by means of iterative Lamerey diagram, widely used in the classical theory of nonlinear oscillations [7].

There is known function $x_{m+1} = f(x_m)$, connecting two neighboring elements of time series: x_m and x_{m+1}. Iterative diagram consists of this function and the bisector $x_m = x_{m+1}$. Choosing the initial point x_0 by means of vertical and horizontal lines we obtain the points on the axis x_i:

$x_1 = f(x_0)$, $x_2 = f(x_1)$, $x_3 = f(x_2)$, ...

We use in this article the logistic function [19] as a chaos generator:

$$x_{m+1} = a x_m (1 - x_m), m = 0,1,2,..., \quad (1)$$

where a is a parameter determining the nature of chaotic orbits.

With the corresponding values of the parameter a it is possible to get different types of attractors by means of iteration algorithm:

a) stable focus ($a = 2.9$), b) stable orbit ($a = 3.3$), c) double orbit ($a = 3.6$), d) chaotic orbit ($a = 4$).

Increasing gradually the parameter a, it is possible to observe the moments of bifurcations, i.e. transitions from one type of the attractor to another. Fig. 6 shows that in the moment $a = 3$ there is a jump from one stable state to two other stable states. In the moment $a = 3.449$ the number of stable states is doubled, etc. The moment $a = 4$ corresponds to the complete chaos [19]. The described chaos generator will be used further for the integration with the fuzzy model of the reliability.

3. FUZZY-CHAOTIC MODEL OF RELIABILITY

It is examined the simple system with the redundancy which is modeled on the base of the fuzzy-algorithm approach proposed in [12, 13] and logistic function (1).

3.1. Element with the reiterative redundancy

For the system model in the form of the parallel circuit where the primary element (A) has y of redundant elements $(y = 0,1,2,..)$, all the elements are supposed to be homogeneous.

The quality of the functioning of the element A depends on the variable x, which varies during the time $x = x(t)$. To evaluate the reliability of the element A it is used:

$\mu_A^1(x)$ - the membership function which determines the correctness distribution of the variable x during the functioning of A.

The parallel circuit assumes that the failure of the system requires the failure of all $(y+1)$ elements similar to A. That's why the correctness of the element functioning with the multiple redundant is evaluated by the formula:

$$\mu_B^1(y) = 1 - \left[1 - \mu_A^1(x)\right]^{y+1} . \quad (2)$$

3.2. Reliability bifurcation

The model (2) allows observing the dynamics of the system reliability level, i.e. of the factor $\mu_B^1(y)$ during the chaotic oscillations of the variable x in conformity with the logistic function (1).

The purpose of the computer experiment consisted of the research of bifurcations of the factor $\mu_B^1(y)$ with different correctness distributions of the element A and different redundancy rates ($y = 1,3,5,7$).

The experiment is carried out with two correctness distributions shown in the fig. 8: triangle (a) and threshold (b). During the chaos generation the parameter a of the logistic function (1) was changed in the range from 2.5 to 4. For each distribution 4 bifurcation diagrams were obtained, each of them corresponds to different redundancy rates (y). The results are represented in the fig. 9 and 10, where the horizontal axis is the chaos parameter (a), and vertical axis is the reliability index $\mu_B^1(y)$.

It can be showing that in spite of the chaos growth (parameter a), at the expense of the increment of the redundancy rate (y) it is possible:

а) to postpone the moment of the first bifurcation which is associated with the reliability loss;

б) to decrease the orbit diameter around which there are oscillations of the level of the system membership to the stable state.

Thereby, there is a task of the optimal reliability assurance in chaos conditions.

4. REDUNDENCY OPTOMIZATION IN CHAOS CONDITIONS

It is examined the sequential system where each component has multiply redundancy. This system is described by the series-parallel chart, where A_i is a component which depends

on the variable x_i, y_i is the redundancy rate of the component A_i, $\mathbf{Y} = (y_1, y_2, ..., y_n)$ is the vector of the redundancy rates y_i, $i = 1,2,...,n$.

It is supposed to be known:

$\mu_{A_i}^1(x_i)$ is the correctness distribution of the variable x_i during the functioning of the component A_i,

$[\underline{x_i}, \overline{x_i}]$ is the range of possible values of the variable x_i,

c_i is the cost of one redundant component like A_i.

We get for the system taking into account (2):

$$\mu_B^1(\mathbf{Y}) = \prod_{i=1}^{n} \left\{1 - \left[1 - \mu_{A_i}^1(x_i)\right]^{y_i+1}\right\}, \quad (3)$$

$$C_B(\mathbf{Y}) = \sum_{i=1}^{n} c_i y_i, \quad (4)$$

where $\mu_B^1(\mathbf{Y})$ is the correctness of the system functioning with the redundancy rates vector $\mathbf{Y}$,

$C_B(\mathbf{Y})$ is the total expenses for the redundancy.

It is supposed that the variable x_i during the functioning of the component A_i makes the chaotic oscillations according to the law (1):

$$x_i(m+1) = a x_i(m)[1 - x_i(m)], \quad (5)$$

where $x_i(m)$ is the value of the variable x_i on the m-step of the logistic function.

One of the practically important tasks of the reliable optimization in the chaos conditions can be formulated as follows:

To find the vector $\mathbf{Y} = (y_1, y_2, ..., y_n)$ which provides

$$C_B(\mathbf{Y}) \longrightarrow \min \quad (6)$$

with conditions

$$\mu_B^1(\mathbf{Y}) \in [\mu_B^*, 1], \; a = a^*, \quad (7)$$

where μ_B^* is the minimum acceptable level of the correctness of the system functioning with the elementwise redundancy.

a^* is the parameter of the logistic function (5) determining the chaos level.

The task was solved for $n = 5$, $c_i = 1$, $i = 1,2,...,5$ and the correctness distributions from the Table 1. It was supposed that the variables x_i for all the components A_i have the same level of chaos (parameter a^*).

In order to solve the task of optimization the genetic algorithm from the MATLAB was used – to see the Appendix.

The results of the task solution for two minimum acceptable levels of the correctness of the system functioning (0.8 and 0.9), four levels of chaos (a^*) shown that increasing the redundancy rate in the conditions of the chaotic oscillations of variables it is possible to achieve the stabilization of the correctness level of the system functioning in the required interval $[\mu_B^*, 1]$. During the system design"for the worst case" the chaos variable a^* can be chosen as the biggest.

It is appropriate to mention here that the classical task of the optimal redundancy [20] was solved without the statistic data about the reliability of system components. Instead there were used the membership functions determining the expert correctness distributions of variables which influence the reliability.

5. CONCLUSION

The joint application of the fuzzy logic and the chaos theory is a convenient modeler of the dynamics of the system reliability within the space of variables connected with the failure causes.

On the base of the integration of membership functions and chaos generator in a form of the logistic map, there were obtained the bifurcation diagrams of the system reliability with the elementwise redundancy.

It is shown that the increment of the number of redundant components postpones the moment of the first bifurcation connected with a loss of the reliability, and decreases the orbit size around which there is the oscillation of the level of system membership to the reliable state.

It was stated the task of the optimization of the number of redundant components in the conditions of chaotic oscillations of variables, and it was provided the example of its solution. It was shown the possibility of the design of chaotic-reliable systems with the required reliability level.

The dynamics modeling of the reliability of structurally complicated systems on the base of the chaos theory and bifurcations may become a prospective direction for the further research.

References

1. Gnedenko, B.V., Belyaev, Yu. K., Solovyev, A.D. (1969). Mathematical. Methods of Reliability, Theory. New York: Academic Press, 1969. 506 p.

2. Barlow R.E., Proshan F. (1965) Mathematical theory of reliability. Wiley, New York, London, Sydney.

3. Zadeh, L. The concept of a linguistic variable and its application to approximate reasoning, I-III, Information Sciences 8 (1975) 199–251, 301–357; 9 (1976) 43–80.

4. Zadeh L.A. Fuzzy Sets as a Basic for a Theory of Possibility // Fuzzy Sets and Systems. – 1978. Vol.1, pp.3-28.

5. Box G.E.P., Jenkins G.M. and Reinsel G.C. Time Series Analysis: Forecasting and Control. Third Edition, Prentice Hall, Englewood Cliffs, N.J., 1994.

6. Sprott J.C. Chaos and Time-Series Analysis. Oxford University Press, 2003.

7. N.V. Butenin, Y.I. Nejmark, N.A. Fufaev: An Introduction to the Theory of Nonlinear Oscillations. Nauka, Moscow, 1987. (In Russian).

8. Reliability and Safety Analysis under Fuzziness / Eds T. Onisawa, J. Kasprzyk. Studies in Fuzzyness. V. 4. Phisika-Verlag, A Springer Verlag Company, 1995.

9. Cai K.Y. Introduction on Fuzzy Reliability. Boston: Kluwer Acad. Publ., 1996.

10. Rotshtein A., Shtovba S. Fuzzy Reliability Analysis of Algorithmic Processes, Vinnitsa: Continent-PRIM, 1997, 150pp (In Russian)..

11. Utkin L., Shubinsky I. Nontraditional Methods of the Informational Systems. Reliability Estimation. – St.P., 2000, - 173 p. (in Russian)

12. Rotshtein A. Fuzzy Reliability Analyses of Human's Algorithms Activity,Reliability, No 2 (21),2007, pp3-18. (In Russian).

13. Rotshtein A. Fuzzy-algorithmic analysis of complex systems reliability // Cybernetics and Systems Analysis. 2011. № 6. 14p. (In Russian).

14. Glushkov V.M. Automata theory and formal microprogram transformations // Cybernetics. 1965. № 5. p. 1–10. (In Russian).

15. D.I. Katelnikov, A.P. Rotshtein. Fuzzy Algorithmic Simulation of Reliability: Control and Correction Resource Optimization. Journal of Computer and Systems Sciences International, 2010, Vol. 49, No. 6, pp. 967–971.

16. Kolarik W. and Woldstad J.C. Lu S. Lu H. Human Performance Reliability: On-Line Assessment Using Fuzzy Logic. // IIE Transactions, 36: 457-467, 2004, pp. 457-467.

17. Zou F.-Z., Li C.-X. A Chaotic Model for Software Reliability. // Chin. J. Comput., Vol. 24, № 3, Mar 2001, pp. 281-291.

18. Dick S. Bethel C. L. Kandel A. Software-Reliability Modeling: The Case for Deterministic Behavior. // IEEE Transactions on Systems, Man, and Cybernetics - TSMC , vol. 37, no. 1, 2007, pp. 106-119.

19. May R. M. Simple Mathematical Models With Very Complicated Dynamics. //Nature, Vol. 261, June 10, 1976, pp.459.

20. Kozlov B.A. , Ushakov I.A. (1975). Handbook on Reliability of Radio and Automation Systems. Moscow: Soviet Radio. (In Russian).

A Modified Chi-squared Goodness-of-Fit Test for the Inverse Gaussian Distribution

N. Saaidia
Department of Mathematics
University Badji Mokhtar
Annaba, Algérie

Ramzan Tahir
Univ. Bordeaux
IMB, UMR 5251,
F-33400 Talence, France

N. Seddik-Ameur
Department of Mathematics
University Badji Mokhtar
Annaba, Algérie

***Abstract*—In this paper, a modified Chi-squared goodness-of-fit test for the inverse Gaussian distribution based on the maximum liklihood estimators (MLE) is described and analyzed. We study the power of the test against some alternatives for equiprobable grouping random intervals.**

Key-words: Pearson's Chi-squared test, Modified Chi-squared test, Inverse Gaussian distribution, Power of the test, Estimation, Reliability.

I. Introduction

Over a century the inverse Gaussian distribution (IGD) had attracted the attention of many researchers in several fields. The origin of this distribution goes back to the famous botanist Robert Brown (1773-1858). He interested in the study of particles motion (which now is well-known Brownian motion). In 1905, Albert Einstein derived the normal distribution as the model for Brownian motion, also in 1915 Schrödinger has obtained the distribution of first passage time as inverse Gaussian. For more details see (Chhikara and Folks (1989), Seshadri (1993), Seshadri (1999). Use of the IGD as a life time model is particularly appealing (Chhikara and Folks (1989). The hazard rate of the IGD has $\cap -$ shape like lognormal, generalized Weibull, Birnbaum-Saunders and loglogistic distributions. The IGD offers certain advantages over these distributions, because for these three distributions the hazard rate increases from 0 to its maximum value and then decreases to 0. For the IGD, the hazard rate increases from 0 to its maximum value and then decreases asymptotically to a constant which implies that the occurrence of a failure eventually becomes purely random and independent of past life. In contrast, the other three distributions vanishing hazard rate implies that eventually almost no possibility of failure remains, which is not reasonable for most real system (Gunes and al. (1996).

The purpose of statistical modeling is to give the probability distribution an appropriate place to fit the data to be used in the analysis, among the tools used is the chi-square test.

Let us consider the sample $\mathbb{X} = (X_1, X_2, \cdots, X_n)^T$. We say that X_i follow the IGD if the density function is defined by:

$$f(x,\theta) = \left(\frac{\lambda}{2\pi x^3}\right)^{\frac{1}{2}} \exp\left\{-\frac{\lambda(x-\mu)^2}{2\mu^2 x}\right\},$$
$$x \geq 0, \quad \theta = (\mu,\lambda)^T \in \mathbb{R}_*^+ \times \mathbb{R}_*^+ \subset \mathbb{R}^2,$$

where μ is the mean and λ is the shape parameter.

The distribution function is

$$F(x,\theta) \quad = \Phi\left(\sqrt{\tfrac{\lambda}{x}}\left(\tfrac{x}{\mu}-1\right)\right) + \exp\left(\tfrac{2\lambda}{\mu}\right)\Phi\left(-\sqrt{\tfrac{\lambda}{x}}\left(\tfrac{x}{\mu}+1\right)\right).$$

The hazard rate function of IGD is

$$h(x,\theta) = \frac{\left(\frac{\lambda}{2\pi x^3}\right)^{\frac{1}{2}} \exp\{-\frac{\lambda(x-\mu)^2}{2\mu^2 x}\}}{\Phi\left(-\sqrt{\frac{\lambda}{x}}(\frac{x}{\mu}-1)\right) - \exp(\frac{2\lambda}{\mu})\Phi\left(-\sqrt{\frac{\lambda}{x}}(\frac{x}{\mu}+1)\right)},$$

$x \geq 0$.

The MLE's of μ and λ are

$$\hat{\mu}_n = \overline{X} \quad \text{and} \quad \hat{\lambda}_n = \frac{n}{\sum_{i=1}^n (X_i^{-1} - \overline{X}^{-1})}$$

respectively.

Now we consider the problem of testing the hypothesis H_0 that the distribution of the sample X_i belongs to the family of IGD

$$H_0 : \mathbf{P}(X_i \leq x) = F(x,\theta) \quad , \theta = (\mu,\lambda)^T, x \geq 0.$$

We divide the real line into r intervals $I_1, I_2, \cdots, I_r$ by the points

$$-\infty = a_0 < a_1 < \cdots < a_{r-1} < a_r = +\infty, \quad I_i =]a_{i-1}, a_i],$$
$$I_i \cap I_j = \emptyset, i \neq j, \quad \cup_{i=1}^r I_i = \mathbf{R}^1,$$

and we group the sample over these intervals, we obtain the vector of frequencies $\nu = (\nu_1, \nu_2, \cdots, \nu_r)^T$ and the probability vector $p(\theta) = (p_1(\theta), p_2(\theta), \cdots, p_r(\theta))^T$, where $p_j(\theta) = \mathbf{P}(X_i \in I_j | H_0), \quad j = 1, 2, \cdots, r$.

To test the hypothesis H_0, Pearson proposed a test based on the so-called quadratic form of Pearson

$$X_n^2(\theta) = X_n^T(\theta) X_n(\theta) = \sum_{i=1}^r \frac{(\nu_i - np_i(\theta))^2}{np_i(\theta)}, \qquad (1)$$

where

$$X_n(\theta)=\left(\frac{\nu_1-np_1(\theta)}{\sqrt{np_1(\theta)}},\frac{\nu_2-np_2(\theta)}{\sqrt{np_2(\theta)}},\cdots,\frac{\nu_r-np_r(\theta)}{\sqrt{np_r(\theta)}}\right)^T.$$

Under H_0, if θ is known, it was shown by K. Pearson in 1900 (see e.g. Drost (1998)) that

$$\lim_{n\to\infty}\mathbf{P}(X_n^2(\theta))\leq x|H_0)=\mathbf{P}(\chi^2_{r-1}\leq x).$$

The hypothesis H_0 must be rejected at a significance level α, whenever $X_n^2(\theta)>C_\alpha$, where C_α is the critical value of the Pearson's test, $C_\alpha=\chi^2_{r-1,\alpha}$ is the upper $\alpha-$ quantile of the χ^2 distribution with $r-1$ degrees of freedom.

But generally θ is unknown and must be estimated using the sample $\mathbb{X}$, if we replace θ in(1) by any $\sqrt{n}-$ consistent estimate θ_n^*, the limit distribution of (1) will not be χ^2_{r-1} and depends on the method of estimation of θ and the properties of the estimator θ_n^*.

1) In 1973, Nikulin (1973a, 1973b) proposed a modification in the standard chi-squared Pearson's test for continuous distribution with shift and scale parameters, also Rao and Robson (1974) had obtain the same result for exponential family , and since 1998 , the test is well known as the Rao-Robson-Nikulin (RRN) test (van der vaart (1998), Drost (1988)) and it can be written as

$$\begin{aligned}Y_n^2(\hat\theta_n) &= X_n^2(\hat\theta_n)+X_n^T(\hat\theta_n)B(\hat\theta_n)(I(\hat\theta_n)-\\ &\quad J(\hat\theta_n))^{-1}B^T(\hat\theta_n)X_n(\hat\theta_n),\end{aligned}$$

with

$$B(\theta)=\begin{pmatrix}b_{11}&b_{12}\\ b_{21}&b_{22}\\ \vdots&\vdots\\ b_{r1}&b_{r2}\end{pmatrix};$$

$$b_{i1}(\theta)=\frac{1}{\sqrt{p_i(\theta)}}\frac{\partial p_i(\theta)}{\partial\mu},\qquad b_{i2}(\theta)=\frac{1}{\sqrt{p_i(\theta)}}\frac{\partial p_i(\theta)}{\partial\lambda},$$

$i=1,2,\cdots,r$, and $J=B^T(\theta)B(\theta)$ is the Fisher's information matrix of the vector of frequencies ν and I is the Fisher's information matrix of X_i

$$I(\theta)=\begin{pmatrix}\frac{\lambda}{\mu^3}&0\\ 0&\frac{1}{2\lambda^2}\end{pmatrix}.$$

The asymptotic behavior of the statistics $Y_n^2(\hat\theta_n)$ is given by the following Nikulin (1973a, 1973b)

$$\lim_{n\to\infty}\mathbf{P}(Y_n^2(\hat\theta_n)\leq x|H_0)=\mathbf{P}(\chi^2_{r-1}\leq x).$$

For the inverse Gaussian distribution, the RRN test is very well studied (see Saaidia and Seddik-Ameur (2010), Lemeshko et al. (2010), Saaidia (2009), (ikulin and Saaidia (2009)).

2) In 1974 Dzhaparidze and Nikulin (1994) proposed a modification of the standard Pearson's test. The test is well known Dzhaparidze-Nikulin (DN) test $U_n^2(\hat\theta_n)$ which is

$$U_n^2(\hat\theta_n)=X_n^2(\hat\theta_n)-X_n^T(\hat\theta_n)B(\hat\theta_n)J^{-1}(\hat\theta_n)B^T(\hat\theta_n)X_n(\hat\theta_n),$$

and under some regularity conditions of Cramer, we have (Dzhaparidze and Nikulin (1994))

$$\lim_{n\to\infty}\mathbf{P}(U_n^2(\hat\theta_n)\leq x|H_0)=\mathbf{P}(\chi^2_{r-3}\leq x).$$

3) The statistic $Y_n^2(\hat\theta_n)-U_n^2(\hat\theta_n)$ is Asymptotically independent of the DN test (voinov (2009)) and we have

$$\lim_{n\to\infty}\mathbf{P}\left(Y_n^2(\hat\theta_n)-U_n^2(\hat\theta_n)\leq x|H_0\right)=\mathbf{P}(\chi^2_2\leq x).$$

Noting that Hsuan and Robson (1976), (also see Voinov et al. (2009), Voinov (2008a), (Mirvaliev 2001)) have proposed a modification of the standard Pearson's test (1) by using the methods of moments.

From the above theory it seems reasonable to investigate the power of three modified chi-squared type tests $Y_n^2(\hat\theta_n)$, $U_n^2(\hat\theta_n)$, and $Y_n^2(\hat\theta_n)-U_n^2(\hat\theta_n)$.

II. Power estimation

To investigate the power of $Y_n^2(\hat\theta_n)$, $U_n^2(\hat\theta_n)$ and $Y_n^2(\hat\theta_n)-U_n^2(\hat\theta_n)$ tests for the inverse Gaussian distribution as null hypothesis against alternatives lognormal, loglogistic, exponentiated Weibull (Mudholkar et al. (1995)) and power generalized Weibull distributions (Bagdonavicius and Nikulin (2002)). These distributions are generally used in reliability when the hazard rate function is unimodal (ie $\cap$-shape). Figures 1-4 depict the powers of the three tests with different alternatives.

III. Summary and conclusions

In this paper we studied the modified chi-squared of the standard Pearson's test (1) based on the MLE. It is clear that the DN $U_n^2(\hat\theta_n)$ test for equiprobable intervals possesses low power for all alternative distributions. In contrast the $Y_n^2(\hat\theta_n)$ and $Y_n^2(\hat\theta_n)-U_n^2(\hat\theta_n)$ tests are nearly the most powerful for all alternatives considered and for varying number of intervals r. Note that the case $r>40$ needs further investigation because the expected intervals probabilities become small and the limit distribution of the above tests will not follows the chi-squared distribution Voinov et al. (2008b).
It will be more interesting if we add to this study, the class of Neyman-Pearson and study the power of these proposed tests.

References

[1] Bagdonavičius, V., Nikukin, M. (2002) Accelerated Life Models: Modeling and Statistical Analysis. Chapman and Hall.

[2] Chhikara, R.S., Folks, J.L. (1977). The Inverse Gaussian Distribution as a Life Time Model. Technometrics, 19 , 461-468.

[3] Chhikara, R.S., Folks, J.L. (1989). The Inverse Gaussian Distribution. Marcel Dekker, New York.

[4] Cramer, U. (1946). Mathematical Methods of statistics. Princeton University Press, New York..

[5] Drost F. (1988), Asymptotics for Generalized Chi-squared Goodness-of-fit Tests. amsterdam: Centre for Mathematics and Computer Sciences,CWI Tracs, 48.

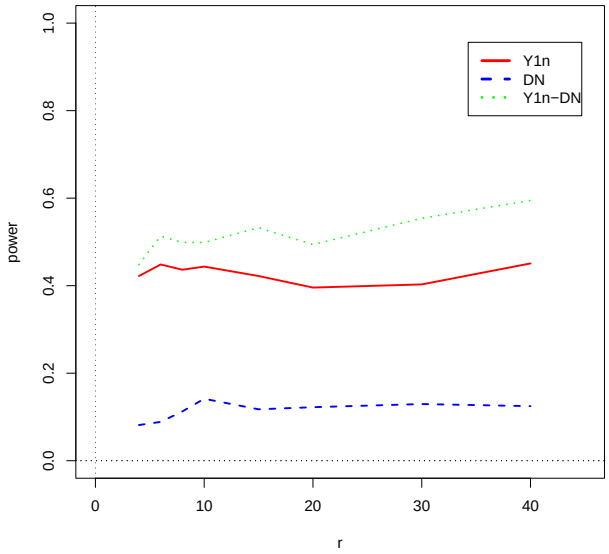


Fig. 1. Estimated powers of 3 tests $Y_n^2(\hat{\theta}_n)$, $U_n^2(\hat{\theta}_n)$ and $Y_n^2(\hat{\theta}_n) - U_n^2(\hat{\theta}_n)$ as function of the number of equiprobable intervals r against lognormal distribution as alternative. Sample size $n = 200$, significance level $\alpha = 0.05$.

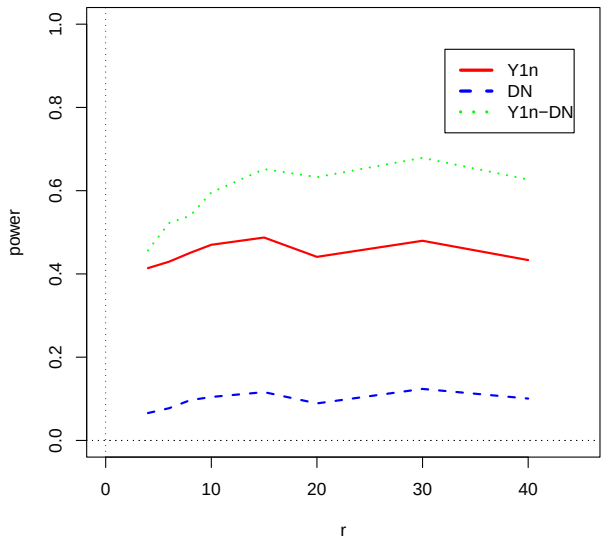


Fig. 2. Estimated powers of 3 tests $Y_n^2(\hat{\theta}_n)$, $U_n^2(\hat{\theta}_n)$ and $Y_n^2(\hat{\theta}_n) - U_n^2(\hat{\theta}_n)$ as function of the number of equiprobable intervals r against loglogistic distribution as alternative. Sample size $n = 200$, significance level $\alpha = 0.05$.

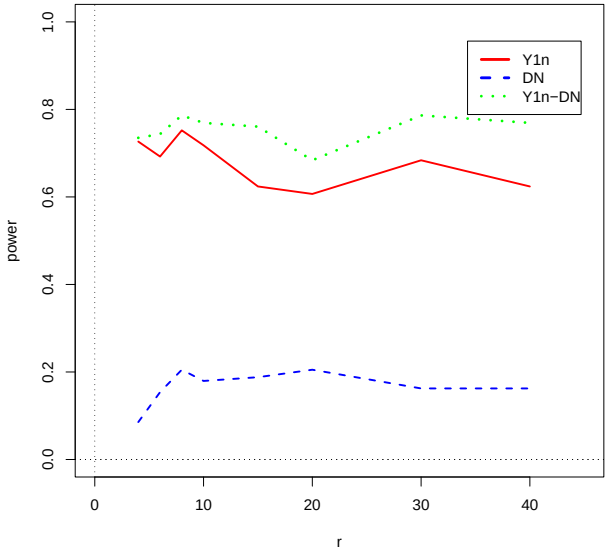


Fig. 3. Estimated powers of 3 tests $Y_n^2(\hat{\theta}_n)$, $U_n^2(\hat{\theta}_n)$ and $Y_n^2(\hat{\theta}_n) - U_n^2(\hat{\theta}_n)$ as function of the number of equiprobable intervals r against power generalized Weibull as distribution alternative. Sample size $n = 200$, significance level $\alpha = 0.05$.

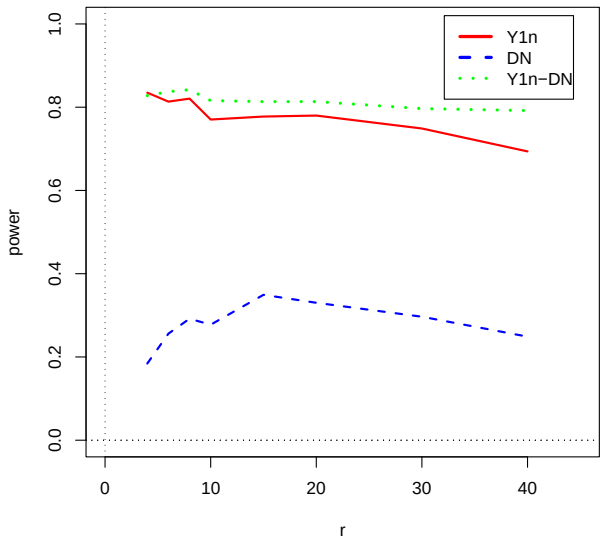


Fig. 4. Estimated powers of 3 tests $Y_n^2(\hat{\theta}_n)$, $U_n^2(\hat{\theta}_n)$ and $Y_n^2(\hat{\theta}_n) - U_n^2(\hat{\theta}_n)$ as function of the number of equiprobable intervals r against Exponentiated Weibull distribution as alternative. Sample size $n = 200$, significance level $\alpha = 0.05$.

[6] Greenwood, P. S. and Nikulin, M. (1996). A guide to Chi-squared Testing. John Wiley and Sons, New York.
[7] Gunes, H., Dietz, D.C., Auclai, P.F., Moor, A.H.,(1997), Modified goodness-of-fit tests for the inverse Gaussian distribution. Computational statistics and Data analysis 24, p.63-77.
[8] Hsuan T.A, Robson, D.S. (1976). The χ^2- Goodness-of-fit Tests with Moment Type Estimator. Communications in Statistics - Theory and Methods, 16 , 1509-1519.
[9] LeCam, L. (1956). Locally asymptotically normal families of distributions. In University of California, Publication in Statistics , vol.3, p. 37-98.
[10] Lemeshko, B.Y., Lemeshko, S.B., Akushkina, K.A., Nikulin, M., Saaidia, N. (2010). Inverse Gaussian Model and Its Applications in Reliability and Survival Analysis, In: Mathematical and Statistical Models and Methods in Reliability (Eds. Rykov, V.V., Balakrishnan, N., Nikulin, M.S.), Birkhäuser Boston, p 293-315.
[11] Mirvaliev M. (2001). An investigation of generalized chi-squared type statistics, Doctoral thesis, Academy of Science of the Republic of Uzbekistan, Tashkent.
[12] Meeker, W.Q., Escobar, L.A. (1998). Statistical Methods for reliability Data. John Wiley and Sons, INC.
[13] Mudholkar, G., Srivastava D., Freimer M. (1995). The exponentiated Weibull family: a reanalysis of the bus-motor-failure data, Technometrics, vol. 37, p. 436-445.
[14] Nikulin M.S. (1973a). Chi-square Test For Continuous Distributions with Shift and Scale Parameters. Teor. Veroyatn.Primen, 18, No. 3, 559-568.
[15] Nikulin M.S. (1973b). On a Chi-square Test For Continuous Distributions,Theory of Probability and Applications, 18 , p.638-639.
[16] Nikulin, M.S., Haghighi, F. (2004). A Chi-Squared Test for the Generalized Power Weibull Family for the Head-and-Neck Cancer Censored Data, journal of Mathematical Sciences Volume 133, Number 3, 1333-1341.
[17] Nikulin, M.S., Saaidia, N. (2009). Inverse Gaussian family and its applications in reliability, Study by simlulation. St. Petersburg, June 28–July 4. VVM comm. Ltd., St. Petersburg V.2, 657-661.
[18] Rao, K.C. , Robson, D.S. (1974). A chi-square statistic for goodness-of-fit tests within the exponential family. Communications in . Statistics, 3 , 1139-1153
[19] Saaidia, N. Seddik-Ameur, N. (2010). Chi-Squared Type Test for the Inverse Gaussian Distribution. les annales de l'ISUP, Volume LIV - Fascicule 3, pp. 67-84.
[20] Saaidia, N. (2009). Sur les applications de la famille des lois Gaussiennes inverses en fiabilit, 41mes Journes de Statistique, Bordeaux, 19-22 mai 2009.
[21] Seshadri, V. (1993). The Inverse Gaussian Distribution: A Case Study in Exponential Families. Clarendon Press:Oxford,
[22] Seshadri, V. (1999). The inverse Gaussian Distribution: statistical theory and applications. Springer, New York.
[23] Van Der Vaart, A.W. (1998). Asymptotic Statistics. Cambridge Series in Statistics and probabilistic Mathematics. Cambridge: Cambridge University Press.
[24] Voinov, V., Alloyarova, R., Pya, N. (2008a). Recent Achievements in Modified Chi-squared Goodness-of-fit Testing. In:Statistical Models and Methods for Biomedical and Technical Systems,(Eds. F. Vonta, M. Nikulin, N. Limnios, C. Huber.)Birkhäuser, Boston, p 241-258.
[25] Voinov, V., Alloyarova, R., Pya, N. (2008b). A Modified Chi-squared Goodness-of-fit Test for the Three-parameter Weibull Distribution and its Applications in Reliability. In:Mathematical Methods in Survival Analysis, Reliability and Quality of Life,(Eds. C. Huber, N. Limnios, M. Mounir, M. Nikulin.)John Wiley & Sons, Inc, pp 189-202.
[26] Voinov, Pya, N., V., Alloyarova, R. (2009). A Comparative Study of Some Modified Chi-squared Tests. Communi. in Stat.-Simul. and Comput.,38:2, 355-367.

Estimation of linear functionals of a multivariate distribution under multivariate censoring Conditional

Philippe Saint Pierre, Olivier Lopez
Email: philippe.saint_pierre@upmc.fr

Abstract: In this paper we study the non-parametric estimation of linear functionnals of a multivariate distribution under multivariate right censoring. We provide a new estimator of the multivariate distribution function of censored lifetimes. This estimator is based on a copula approach to modelize the dependance between multivariate censoring times. The estimators of the linear functionnals are shown to be asymptotically normal. Applications to linear regression estimation, empirical correlation estimation and Kendall's coefficient estimation are presented. A bootstrap procedure is also discussed.

Sliced Inverse Regression for Survival Data

Maya Shevlyakova
Ecole Polytechnique Fédérale de Lausanne
Switzerland
Email: maya.shevlyakova@epfl.ch

Stephan Morgenthaler
Ecole Polytechnique Fédérale de Lausanne
Switzerland
Email: stephan.morgenthaler@epfl.ch

***Abstract*—Dimension reduction is a technique for handling multivariate data. We use one of its versions, sliced inverse regression (SIR) and we study its performance on survival data. Our approach is a bit different from the original paper [1] on this subject. The right-censored observations are taken into account during the slicing of the survival times by assigning each of them with equal weight to all of the slices with longer survival. The classical version of the algorithm [2] is then applied to estimate the effective dimension-reduction directions.**

We test this method with different distributions for the two main survival data models: Accelerated Lifetime Model and Cox's proportional hazards model. In both cases and under different conditions of sparsity, sample size and dimension of parameters, it performs well as a variable selector.

I. Introduction

Dimension reduction aims to select a few new variables, which often are linear combinations of the original ones, and which describe the important aspects of the observed data well enough. We concentrate on this idea in the context of the regression of a vector y on a p-dimensional predictor x by estimating the central subspace (of dimension $k < p$) which contains most of the information about our response y. The regression model of the form

$$y = f(\beta_1^T x, \beta_2^T x, \ldots, \beta_k^T x, \epsilon), \tag{1}$$

where the β are unknown p-vectors, ϵ is a random variable independent of x, and the f is an arbitrary unknown function on $\mathbf{R}^{k+1}$, has a k-dimensional central subspace. Any linear combination of β's is called an effective dimension reduction (e.d.r) direction. We note here that the function f does not have to be linear in its components, the method is often able to estimate the e.d.r. directions even if the link between y and $\beta^T x$ is of more complex form. We concentrate on estimating the e.d.r. directions only, not the form of the function f. We refer to the linear space generated by the β's as the e.d.r. space.

Such an approach was suggested in 1991 by Li [2]. In 1999, a version for survival data was proposed in [1], in which the censoring was dealt with by a double slicing technique. Later on, an idea of inverse regression was generalized by Cook [3], introducing a minimum discrepancy approach in order to find the e.d.r. space. Finally, a recent study [4] adapted this method to survival data and investigated its performance on several datasets.

We rely on the classical method of SIR and deal with censoring in a somewhat naive and straightforward way. Nevertheless, the method shows quite good results for selecting the variables and in most cases can compare with the method based on the minimum discrepancy approach, all the while being easier to understand and not as computationally challenging.

A. Cox's proportional hazards model

Cox's proportional hazards model [5] assumes the following relation for the hazard function for an individual with characteristics Z:

$$\lambda_Z(t) = \lambda_0(t) e^{\beta^T Z}, \tag{2}$$

where $\lambda_0(t)$ is the baseline hazard function which is unknown and β contains our coefficients of interest. The matrix of covariates Z may contain the clinical and patient history as well as genomic information. The baseline hazard function corresponds to the probability of dying (or reaching an event) when all the explanatory variables are zero and is analogous to the intercept in ordinary regression (since $\exp(0) = 1$).

Cox's model is a semi-parametric model. While no assumptions about the form of $\lambda_0(t)$ are made, we assume a parametric form for the effect of the predictors on the hazard. The model can be estimated even though the baseline hazard is unspecified.

Parameter estimates are obtained by maximizing the partial likelihood function:

$$L(\beta) = \prod_{i=1}^{n} \frac{e^{\beta^T Z_i}}{\sum_{j \in R_i} e^{\beta^T Z_j}}.$$

The product is computed over all the events (failures) and for each event we define the set R_i as the list of individuals at risk (those who haven't experienced a failure yet) at the i-th failure time. Censored data are taken into account by being in R_i until censoring.

B. Accelerated lifetime model

Another model for survival data which seems reasonable in many studies is the accelerated lifetime (failure time) model. Cox's model suggests proportionality between hazards, while the accelerated lifetime model considers a different behavior for the hazards of different individuals. The survival time is expressed in formula (3) below:

$$T = \exp(\beta^T Z) T_0, \tag{3}$$

where T_0 denotes the base survival time and $\exp(\beta^T Z)$ estimates the effect of the covariates of the individual on his

survival time. If $\beta^T Z > 0$, lifetime is prolonged compared to the base T_0, that is, the failure has a tendency to occur later. Alternatively, when $\beta^T Z < 0$, one speaks of accelerated lifetime since the failure happens earlier.

II. Sliced Inverse Regression

We avoid dealing directly with a possible high-dimensional covariate vector by switching to the inverse problem. Instead of estimating y as a function of x, we regress x against y, which is a set of one-dimensional regressions.

The inverse regression curve under the condition (1) will fall into a k-dimensional affine subspace determined by the e.d.r. directions. The important step here is to standardize the x to have 0 mean and the identity covariance. We can rewrite the formula (1) as

$$y = f(\eta_1 z, \eta_2 z, \ldots, \eta_k z, \epsilon), \tag{4}$$

where $\eta_k = \Sigma_{xx}^{1/2}\beta_k$ and $z = \Sigma_{xx}^{-1/2}(x - \bar{x})$.

After standardization of x, we estimate the regression curve $\mathrm{E}(z|y)$. To do that, we slice the sorted y into several intervals and we compute the corresponding slice means of z. The principal component analysis on the slice means of z defines the most important k-dimensional subspace for tracking the inverse regression curve $\mathrm{E}(z|y)$. The original e.d.r. directions are found after re-transformation back to the original scale.

A. Algorithm

The sliced inverse regression algorithm [2], [6] can be described as follows:

1) Standardize x:

$$z_i = \hat{\Sigma}^{-1/2}(x_i - \bar{x}).$$

2) Divide the range of y_i into S nonoverlapping slices $H_s, s = 1, \ldots, S$. n_s denotes the number of observations within slice H_s, and I_{H_s} is the indicator function for this slice:

$$n_s = \sum_{i=1}^{n} I_{H_s}(y_i).$$

3) Compute the mean of z_i over all slices.

$$\bar{z}_s = \frac{1}{n_s}\sum_{i=1}^{n} z_i I_{H_s}(y_i).$$

4) Calculate the estimate for the covariance matrix

$$\hat{V} = S^{-1}\sum_{s=1}^{S} n_s \bar{z}_s \bar{z}_s^T.$$

5) Identify the eigenvalues $\hat{\lambda}_i$ and eigenvectors $\hat{\eta}_i$ of $\hat{V}$, where we choose the indices such that the eigenvalues are sorted in decreasing order..
6) Transform the standardized directions $\hat{\eta}_i$ back to the original scale.

$$\hat{\beta}_i = \hat{\Sigma}^{-1/2}\hat{\eta}_i.$$

The first eigenvector $\hat{\beta}_1$ contains the coefficients of interest, if we look for a single e.d.r.

B. Censored data

The classical method of SIR requires a relationship of some form between the response variable and the covariates. For the accelerated lifetime model, the logarithm of the survival time is a linear function of the variables. For the Cox's model, such a relationship exists for the hazards. In both cases, we perform the slicing on the survival time directly.

How do we take care of the censored observations? Ignoring them is not an option, because of the resulting bias. We use the following approach: given that the individual was right-censored at time t which falls in the slice s, the event for this individual could have taken place anytime after t. As a first approximation, we attribute this event to all consequent slices with weight proportional to the slice width. The total sum of the weights naturally equals one. This allows us to use the covariate information of the censored observations. This is inspired by [7]. Let us illustrate this procedure with a quick example:

Suppose we have 7 observations,

$$10, 11^*, 13, 15^*, 16^*, 18, 20,$$

and we choose 4 slices: 10-12, 13-15, 16-18 and 19-20. We then create a matrix of weights which shows in which slice each observation falls.

obs	slice 1	slice 2	slice 3	slice 4
1	1	0	0	0
2	0	0.33	0.33	0.33
3	0	1	0	0
4	0	0	0.5	0.5
5	0	0	0	1
6	0	0	1	0
7	0	0	0	1

The first censored observation 11^* is assigned to the next 3 slices, giving a weight of $1/3$ to each of them, the second censored observation 15^* will be taken into account in the slices 3 and 4 with the weight of $1/2$, while the 16^* will be only considered in the last slice. This matrix is used at the third step of the SIR algorithm, when computing the slice mean.

Such an approach is surely simplistic since it assumes the risk of the event to be equally distributed among all the remaining slices. In reality, an individual is most likely to experience an event with higher probability in some slices rather than others. We are currently exploring other options for distributing the weights to the slices.

III. Simulation results

A number of simulations were performed under different conditions to evaluate this approach.

A. Accelerated Lifetime model

For the accelerated lifetime model we simulated the covariate x from a multivariate normal distribution with identity covariance. The survival times were log-normal or Weibull. The Weibull distribution, which in fact has proportional hazard, was chosen for comparison. For most of the simulations a sparse solution was considered, with three nonzero components: $\beta_2 = 3.5, \beta_4 = -2.7$ and $\beta_7 = 7.2$ and it was tested under different conditions. Some of the results are listed in the tables below. They all show the components of the first eigenvector. All the situations were simulated 100 times.

Log-Normal, n= 50	0.08	**0.20**	-0.08	**0.16**	-0.08
	0.08	**0.41**	0.09	0.08	-0.07
Weibull, n = 50	0.06	**-0.14**	-0.06	**0.09**	-0.06
	-0.06	**0.20**	0.05	0.06	0.05
Log-Normal, n = 500	-0.02	**-0.21**	-0.02	**0.16**	0.02
	-0.02	**-0.43**	0.02	-0.02	0.02
Weibull, n = 500	0.01	**-0.13**	0.02	**-0.06**	-0.01
	- 0.01	**0.18**	-0.01	-0.01	0.01

TABLE I
RESULTS ON SAMPLE SIZE 50/500, 10 VARIABLES, 20% OF DATA CENSORED

The entries show the average over 100 replications of the components of the first eigenvector. There were 10 variables and 20% of the observations were right-censored. The censoring time was computed as a random uniform variable in order to keep the censoring uninformative. From Table I we can see that in both cases the correct coefficients are identified. Moreover, even on relatively small samples ($n = 50$), the method performs well as a variable selector. Having larger samples brings more accuracy, shrinking superfluous coefficients more towards zero. In general, the underlying distribution does not seem to play an important role in successful recovery of the coefficients. Results shown here were established for 10 variables but they remain identical for a larger number of variables as well.

Log-normal, n= 50	0.12	**0.15**	0.11	**-0.13**	-0.12
	0.12	**0.27**	0.11	0.13	-0.12
Weibull, n = 50	0.08	**-0.10**	0.08	**0.09**	0.07
	-0.07	**0.13**	0.08	-0.06	-0.06
Log-normal, n = 500	0.04	**-0.21**	0.04	**0.15**	-0.04
	0.04	**-0.41**	-0.04	0.04	-0.04
Weibull, n = 500	0.02	**-0.13**	0.02	**0.09**	-0.02
	-0.02	**-0.15**	0.02	0.02	0.02

TABLE II
RESULTS ON SAMPLE SIZE 50/500, 10 VARIABLES, 50% OF DATA CENSORED

Table II contains the same results as Table I, with the only difference that a larger proportion of the data was censored (50% instead of 20%). There is much more noise for the smaller sample size which makes the selection of the three non-zero variables (position two, four and seven) quite challenging. The larger samples do not seem to be influenced much by the severe censoring.

We see that the SIR procedure with slicing the survival time works well as a variable selector in the accelerated lifetime model disregarding the underlying distribution of the variables.

B. Cox's proportional hazards model

As a next step, we check whether a simple slicing of the survival time would work as well in Cox's model. We used the same distribution for the base hazard, log-normal and Weibull. The model's structure was also kept the same: $\beta_2 = 3.5, \beta_4 = -2.7$ and $\beta_7 = 7.2$, other coefficients are zero.

Log-normal, n= 50	0.45	**1.21**	-0.49	**-1.00**	0.51
	-0.58	**2.91**	-0.51	0.54	-0.47
Weibull, n= 50	-0.63	**-1.24**	0.58	**1.05**	-0.46
	0.53	**-2.85**	0.57	-0.57	-0.50
Log-normal, n= 500	-0.14	**-1.36**	0.15	**1.03**	-0.15
	0.14	**-2.99**	-0.13	-0.14	-0.14
Weibull, n = 500	0.15	**1.36**	-0.14	**-1.03**	0.16
	0.17	**3.01**	-0.18	0.18	-0.15

TABLE III
RESULTS ON SAMPLE SIZE 50/500, 10 VARIABLES, 20% OF DATA CENSORED

Log-normal, n= 50	0.81	**-1.14**	0.75	**0.98**	0.89
	0.82	**-1.93**	0.80	0.74	-0.86
Weibull, n= 50	0.87	**1.11**	0.87	**-1.02**	-0.71
	0.79	**1.84**	0.85	-0.92	0.86
Log-normal, n= 500	0.45	**1.21**	-0.49	**-1.00**	0.51
	-0.58	**2.91**	-0.51	0.54	-0.47
Weibull, n = 500	-0.22	**-1.28**	-0.28	**0.96**	0.28
	0.29	**-2.94**	0.23	-0.26	-0.24

TABLE IV
RESULTS ON SAMPLE SIZE 50/500, 10 VARIABLES, 50% OF DATA CENSORED

Tables III and IV show almost the same behaviour as in the accelerated lifetime model. The method selects the true variables, even in case of severe censoring. Both the degree of censoring and the sample size crucially influence the results. The larger the sample size, the better (and more accurate) estimates we get. The same pattern goes for the degree of censoring. But the sufficiently large sample size can compensate even for severe censoring. If we have a lot of data, we can get good results disregarding the fact that a major part of it has been censored.

C. Discussion

As stated earlier, the SIR method assumes a relationship of the form (1) between the response variable and the covariates. In Cox's model, something ressembling (1) holds for the hazard, but in general not for the survival times. To estimate the hazard, we could first perform a slicing of the survival times as above and then use the eigenvector to estimate the corresponding hazards of individuals. Afterwards, the algorithm is repeated, with estimated hazards used for the second round of slicing. Thus, the algorithm is run twice on the same dataset, with different response variables. The results

of the double slicing are shown below in the Table V. Two distributions were used for the base hazard, Weibull and log-normal, the true non-zero coefficients are the same as before (position two, four and seven), and all the results are averaged over 100 runs.

Log-normal, n= 50	-0.11	**0.16**	-0.11	**0.15**	-0.10
	-0.11	**0.29**	-0.11	0.12	-0.13
Weibull, n = 50	-0.08	**-0.12**	-0.08	**-0.09**	0.07
	-0.08	**-0.13**	0.07	0.08	0.08
Log-normal, n = 500	-0.04	**-0.21**	-0.04	**0.15**	- 0.04
	0.04	**-0.41**	-0.03	-0.03	0.04
Weibull, n = 500	-0.03	**0.13**	0.03	**-0.08**	0.03
	0.03	**-0.14**	0.03	0.03	0.03

TABLE V
RESULTS FOR THE COX'S MODEL ON SAMPLE SIZE 50/500, 10 VARIABLES, 50% OF DATA CENSORED

From Tables III and V, we can state that for the Weibull distribution results are comparable, if not better without the second run (slicing the hazards). As before, there is not much qualitative difference between using two different distributions for generating the survival time. We conclude that estimating hazards and slicing according to them is not really worth the effort, slicing the survival times does the job, at least in our examples.

Another open question is how the right-censored observations should be treated. Is it correct to simply put them into the higher hazard slices? We did not consider the conditional distributions, trying to keep the method simple but is the chance of experiencing an event equal in all the slices? We are currently investigating other ways to disperse such observations over the remaining slices.

Next, we compare our results to those in [4], where the authors also used the accelerated lifetime and Cox's model on the basis of inverse regression with quadratic discrepancy function. We used the same models for generating the survival and censoring times and compared the results on several models from their article. Since the authors of [4] were comparing their method to the double slicing method listed in [1] by computing the mean angle between the basis vector and the eigenvector estimate, we did the same for our method. Figures 1 and 2 below show mean angles from 100 simulations runs for different dimensions of x and different sample sizes for the Model 1 and Model 3 in [4] respectively.

Each figure has two plots, the part (a) shows the mean angles for a fixed number of predictors ($p = 6$) as the sample size grows, while the part (b) is a function of the p. We can see that our method definitely performs worse when the number of parameters grows. When it comes to a fixed p, it really depends on the model. While the computational cost of the inverse regression via minimum discrepancy approach is unknown to us, it is surely higher than the method we execute. And for purposes of variable selection such an approach seems to be valid.

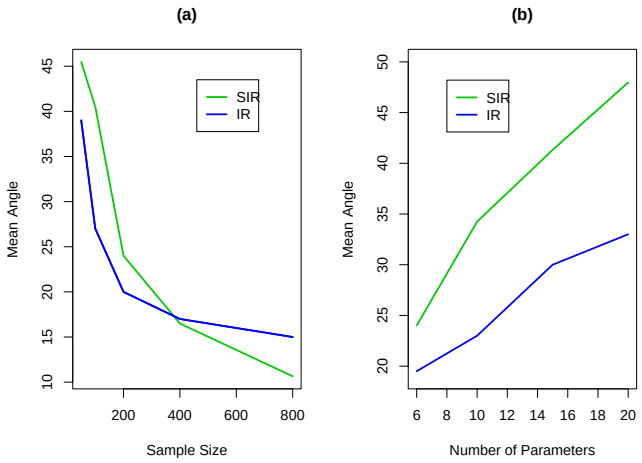


Fig. 1. Mean angles between the basis and SIR (our) and IR (alternative) estimates in Model 1.

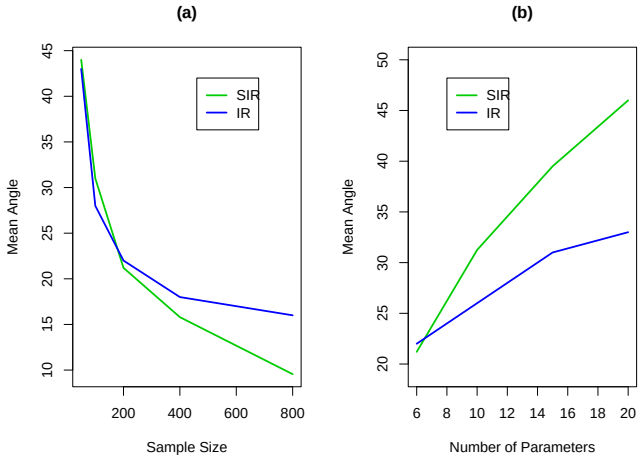


Fig. 2. Mean angles between the basis and SIR (our) and IR (alternative) estimates in Model 3.

D. Selecting the non-zero variables

It is often of interest to know which variables most influence survival. Many methods look for a sparse solution. Sliced inverse regression works for both sparse and non-sparse cases. A simple way to define zero coefficients in our SIR estimate is to use, for example, a trimmed variance. Computing the confidence interval around zero allowed us to select the true non-zero elements in the examples above. Once the variables are identified, they can be fitted by likelihood.
One could also base the selection on the variance estimate given by bootstrapping, a technique most commonly used in dimension reduction estimates. We are currently looking into a possibility of estimating an asymptotic variance for our method.

IV. Conclusion

Our procedure adapts the idea of sliced inverse regression and applies it to survival data. Censored observations are redistributed in posterior slices. For simplification and comparison purposes, only results on 10 variables with the same coefficients have been presented but they remain valid for different model structures. The method allows to pre-select the variables of interest which later can be fitted if necessary by any appropriate technique. Even under severe censoring the most important variables are uncovered. Although the main idea relies on dimension reduction, an underlying sparse solution in not necessarily required.

Acknowledgment

The authors gratefully acknowledge the support by the Swiss National Science Foundation.

References

[1] Ker-Chau Li, Jane-Ling Wang and Chun-Houh Chen, *Dimension Reduction for Censored Regression Data*, The Annals of Statistics, Vol. 27, No. 1, pp. 1-23, 1999.

[2] Ker-Chau Li, *Sliced Inverse Regression for Dimension Reduction*, Journal of the American Statistical Association, Vol. 86, No. 414, pp. 316-327, 1991.

[3] R. Dennis Cook and Liqiang Ni, *Sufficient Dimension Reduction via Inverse Regression: A Minimum Discrepancy Approach*, Journal of the American Statistical Association, Vol. 100, No. 470, pp. 410-428, 2005.

[4] Nivedita V. Nadkarni, Yingqi Zhao and Michael R. Kosorok, *Inverse Regression Estimation for Censored Data*, Journal of the American Statistical Association, Vol. 106, No. 493, 2011.

[5] D.R. Cox, *Regression Models and life tables (with discussion)*, Journal of Royal Statistical Society, Series B, vol. 34, pp. 187220, 1972.

[6] W. Haerdle and L. Simar, *Applied Multivariate Statistical Analysis*, Springer, 2007.

[7] B. Efron, *The Two Sample Problem with Censored Data*, 1967.

Survival analysis used for loan default of technology based firms

So Young Sohn
Department of Information & Industrial Engineering
Yonsei University
Seoul, Korea
sohns@yonsei.ac.kr

Yong Han Ju
Department of Information & Industrial Engineering
Yonsei University
Seoul, Korea
juyonghan@yonsei.ac.kr

Abstract—**Technology credit guarantee fund has been established to provide financial support to technology-based SMEs that have insufficient asset. Risk management of such fund is crucial. We develop a technology credit-scoring model based on survival analysis so that it can be used for stress test. We consider interaction effects among economic indicators, firm's characteristics, and technology evaluation attributes used at the time of application for loan of technology based firms. Our work is expected to contribute to reducing the risks associated with technology financing.**

Keywords-Technology credit guarantee policy; Risk management; Survival analysis; Stress test

I. INTRODUCTION

A technology credit guarantee policy has been established to provide financial support for SMEs based on their technology. The main purpose of this policy is to enhance the competitiveness of technology-based SMEs by increasing their accessibility to private financing sources [1]. However, one of the critical issues involved in this kind of funding is the high rate of its loan default of fund recipient. To reduce such loan defaults, Sohn et al. [2] first attempted to develop a technology credit scoring model based on historical fund recipient data that include their application information and funding result recorded in terms of the loan default. Additionally, many advanced models have been proposed to reduce recipients' loan default [3], [4], [5].

However, the economic environment and firm characteristics were not considered in their survival analysis. Also, the previous studies have not considered interaction effects among economic indicators, firm's characteristics, and evaluation attributes on the time to loan default of technology based firms. This information can be crucial for stress test.

In order to resolve the limitations of previous studies, in this paper we propose an interaction effects based survival model that can be used not only for the technology credit scoring but also for the stress test for technology financing under various scenarios. We expect that the proposed approach can contribute to the effective risk management of technology financing.

This paper is organized as follows. In section 2, we review previous studies regarding both technology credit scoring and risk management. In section 3, we define the data and variables for our survival analysis. In section 4, the new model for risk management is proposed for default predictions. We also conduct a stress test based on the proposed model. Finally, in section 5, we conclude the paper and discuss areas for further study.

II. LITERATURE REVIEW

Stress test is an important tool for risk management. In technology credit scoring model, Moon and Sohn [4] proposed an advanced scoring model considering both firm characteristics and economic conditions in addition to a logistic regression model. Their model also enables one to apply to a stress test by considering certain potential changes in those variables. In terms of bank's risk management, Jacobson et al. [6] used a credit risk stress test to assess the effects of certain specific conditions on internal ratings-based (IRB) regulatory capital requirements. Fabi et al. [7] provided an empirical evaluation of the possible impact of a new Basel accord proposal on the lending policies of Italian banks. The authors conducted a stress test to assess the effects of capital requirements on lending conditions in a negative economic scenario. Especially, in view of the fact that various macroeconomic variables such as interest rate and unemployment index are included in the survival model as time-varying covariates, survival analysis is known to be competitive for the prediction of default in comparison with logistic regression.

In credit scoring area, many studies considered interaction effects in their suggested system. Acharya et al. [8] suggested a regression model to confirm the effect of industry distress on creditor recoveries. In this paper, the authors found that the interaction effects between industry asset-specificity and poor industry conditions. Bahrammirzaee et al. [9] developed a hybrid credit ranking intelligent system using both expert system and artificial neural networks based on the criteria of personal customers. Khudnitskaya [10] suggested an improved credit scoring with multilevel statistical modeling. In this study, the authors used many interaction terms to explain the combined impact of the living area effects and individual-level characteristics on the probability of default.

However, these kinds of interaction terms have not been considered in the previous studies in relation to the risk management of technology credit fund in the context of survival analysis. Even under the same economic situation, loan default probability can vary depending on firm characteristics. In order to perform stress test, we suggest a survival analysis that accommodates interaction effects among

technology attributes, economic situation, and firm's characteristics.

III. SURVIVAL ANALYSIS & STRESS TEST

A. Data & Variabels

In order to develop a technology credit scoring model using survival analysis based on the interaction among technology evaluation attributes, economic attributes and SME's specific characteristics, we use an empirical data set. The data set consists of 4566 firms which obtained a credit guarantee based on their technology score in Korea during the years 1999-2004. Among these, 1336 firms defaulted within 1 to 5 years after they received the loan. Among 3230 non-default cases, 235 firms successfully repaid their loan. We consider these non-defaulting firms as censored cases.

Technology-oriented attributes used in the evaluation for these firms were divided into four groups: management, technology, marketability, and profitability, which cover 16 attributes as displayed in Table 1 [2], [4]. To resolve the problems of potential multi-collinearity in these 16 attributes, we conducted an exploratory factor analysis (EFA) first, which finds the factors for individual attributes [2], [11], [12], [13]. Before conducting the EFA, we rescaled all attributes on a five-point scale where 5 indicates the highest value while 1 represents the lowest value in a relevant attribute. As shown in Table 2, we found 12 factors from the EFA that explain more than 88% of the original sixteen evaluation attributes.

Table 1. Technology-oriented evaluation attributes [2]

Factors	Variable name	Attributes	Scale
Management	P1	Knowledge management	5
	P2	Technology experience	5
	P3	Management ability	5
	P4	Fund supply	5
	P5	Human resource	5
Technology	P6	Environment of technology development	5
	P7	Output of technology development (e.g. patents, certifications)	5
	P8	New technology development	5
	P9	Technology superiority	10
	P10	Technology commercialization potential	10
Marketability	P11	Market Potential	5
	P12	Market characteristic	5
	P13	Product competitiveness	10
Profitability	P14	Sales schedule	10
	P15	Business progress	5
	P16	Return on investment	5

Table 2. Naming of technology factors

Factor1 (F1)	**Factor2 (F2)**	**Factor3 (F3)**	**Factor4 (F4)**
Manager's knowledge and experience factor (P1 and P2)	Human resource and Environment of technology development (P5 and P6)	Product competitiveness and Technology superiority (P9 and P13)	Sales schedule and Return on investment (P14 and P16)
Factor5 (F5)	**Factor6 (F6)**	**Factor7 (F7)**	**Factor8 (F8)**
Business progress (P15)	Management ability (P3)	Output of technology development (P7)	Market Potential (P11)
Factor9 (F9)	**Factor10 (F10)**	**Factor11 (F11)**	**Factor12 (F12)**
Market characteristic (P12)	New technology development (P8)	Technology commercializati on potential (P10)	Fund supply (P4)

Next, we considered the economic environment situation. The economic environment can directly influence the firm's performance. Particularly, SMEs are more sensitive to changes in the economy than large enterprises because their financing ability is affected by the external environment [14], [15], [16]. Therefore, we included 5 economic environment variables at the time of loan application as shown in Table 3, following Moon and Sohn [4]. Like the technology-oriented evaluation attributes, the economic environment variables may have multi-collinearity. We conducted an EFA on these variables and extracted three factors that explain more than 87% of the five economic environment variables. These three factors comprise 'the contrast of the operation index of SMEs to the earning rate of the national bond factor', 'the contrast of KOSPI to the exchange rate factor', and 'Consumer price index' as displayed in Table 4. The first economic factor can have a high value when economic environment is active, while national bonds rate is low. In IMF crisis, the three-year earning rate of national bonds was14.96% while that of Nov-2011 is 3.39% in Korea. The second economic factor has a high value with high KOSPI and low exchange rate of the Korean Won to dollar. In domestic market, when KOSPI is high and the exchange rate is low, economic situation is considered to be healthy. In addition, we applied the SMEs' characteristic variables to our analysis. These are shown in Table 5.

Table 3. Economic variables

Variable name	Economic Indicators
E1	KOSPI (Korean Composite Stock Price Index)
E2	The operation index of SMEs
E3	Consumer price index
E4	The three-year earning rate of national bonds
E5	The won to dollar exchange rate

Table 4. Naming of economic factors

ECO1 (E1)	ECO2 (E2)	ECO3 (E3)
The contrast in the operation index of SMEs to the earning rate of the national bonds factor (E2 and E4)	The contrast of KOSPI to the exchange rate factor (E1 and E5)	Consumer price index (E3)

Table 5. SME specific characteristics [4]

Attributes	Explanation
1. Stock market listed	Listed in KOSPI, KOSDAQ, or other exchange =1, otherwise =0
2. External audit	External audit =1, otherwise =0
3. Investment by foreigners	Investment by foreigners =1, otherwise =0
4. Professional manager	Separation between capital and administration =1, otherwise =0
5. Venture company	Certified by SMBA* =1, otherwise =0
6. INNO-Biz	Certified by SMBA* =1, otherwise =0
7. Production stage	After pilot production stage =1, otherwise =0
8. Joint company	Consortium =1, otherwise =0

B. Survival Analysis

We used a lognormal regression model to estimate SMEs' loan survival time against not only 12 technology oriented factors, 3 economic indicators, and 8 firm's specific characteristics but also their interaction terms and the second order terms of technology factors. In the first step, 6 technology oriented factors, 2 squared terms of two technology oriented factors, 4 SME-specific characteristics variables, and 8 interaction effects were found to be significant at the level of

0.05. Based on only these significant variables, we re-fit a lognormal regression model. Table 6 shows the re-fitted lognormal regression model. A residual plot was drawn against the estimated cumulative hazard function to determine if the lognormal assumption is acceptable. The linear pattern of Figure 1 validates our assumption.

Table 6. The results of the lognormal regression

Analysis of Parameter Estimates					
Parameter	**Level**	**Estimate**	**Standard Error**	**Chi-Square**	**P value**
Intercept		3.8286	0.0269	20262.1	<.0001
E1*F3		0.0557	0.0225	6.13	0.0133
E1* Stock market listed	1	-0.431	0.1308	10.86	0.001
E1* Venture company	1	-0.3261	0.0781	17.45	<.0001
E1*INNO-Biz	1	-0.2338	0.0615	14.47	0.0001
E2*F7		-0.0414	0.0163	6.45	0.0111
E2*F8		0.0462	0.0157	8.65	0.0033
E3* F10		0.0443	0.0188	5.54	0.0186
F1		0.2171	0.0178	149.12	<.0001
F3		0.0446	0.0191	5.46	0.0195
F4		0.0507	0.0165	9.42	0.0021
F5		0.0657	0.018	13.35	0.0003
F8		-0.0619	0.0164	14.16	0.0002
F12		0.1041	0.0164	40.04	<.0001
Stock market listed	1	-0.5722	0.1184	23.34	<.0001
External audit	1	0.1622	0.0587	7.62	0.0058
Venture company	1	0.6701	0.072	86.62	<.0001
INNO-Biz factor	1	0.1323	0.0493	7.21	0.0072
F6*F6		0.0451	0.012	14.22	0.0002
F10*F10		0.0304	0.0139	4.81	0.0283
Scale		0.8082	0.0165		

**F: technology oriented factor, E: Economic factor*

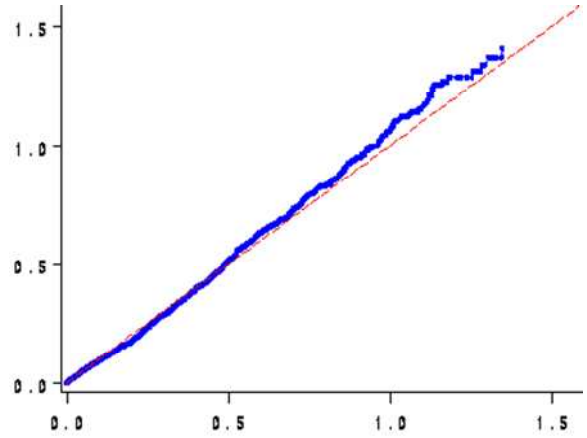


Figure 1. Cox-Snell's residual vs. cumulative hazard rate

In Table 6, one can find some significant interaction effects among economic factors, technology oriented factors, and firm's characteristics as follows. When the operation index of SMEs is high while the rate of national bond is low, if the SME evaluated as having high product competitiveness and technology superiority receives technology credit loan, then their loan survival probability is high.

When the firm which is either stock market listed, venture certified, or INNO-Biz certified receives fund when it is exposed under adverse economic condition in terms of the first economic factor, its loan survival probability will increase.

When KOSPI is high while exchange rate is low, if SME evaluated as having high market potential receives technology credit loan, then their loan survival probability is high. Meanwhile, when the firm which evaluated as having high output of technology development receives fund when it is exposed under adverse economic condition in terms of the second economic factor, its loan survival probability will increase.

When consumer price index is high, if SME evaluated as having high technology development receives technology credit loan, then their loan survival probability is high.

IV. STRESS TEST

A stress test determines potential risk both within the financial system and from the financial sector on the real economy [17]. In this section, we conduct a stress test for technology credit guarantee funds based on the survival model developed in section 3.

Based on various scenarios, it is possible to examine the potential patterns of loan default. In order to set up the scenarios, we considered the changes in the economy situation, technology oriented factor, and the characteristics of the loan applicant firms. We created 7 scenarios, and detailed information is shown in Table 6.

Scenario 1 represents general situation that represents average economic conditions and technology evaluation factors observed in our data and most frequent levels of firm characteristic. By comparing scenario 2 to 4, we wanted to examine the effects of t technology oriented factor on the loan survival probability for SMEs that received the fund under negative economic situation (E1=-5.83, E2=-3.37, and E3=-0.80) while the comparison between scenario 3 and 5 represent opposite economic situation (E1=8.34, E2=6.95, and E3=2.18).

Based on scenarios 6 and 7, we wanted to compare effects of technology oriented factors on loan survival time.

Table 6. The seven scenarios for stress test

Scenarios #	**Description**
1 (general situation)	All economic indicators and technology-oriented factors are set up at the average value, and firm is set as not listed in stock market, not conducting external audit, and neither certified with INNO-Biz nor venture company.
2 (negative situation)	All economic indicators are set at the level of the IMF crisis in Dec-1997, scores of product competitiveness and technology superiority, output of technology development, market potential, and new technology development in technology-oriented factor are set at the lowest value while other factors have the average value. Firm is set as listed in stock market, not conducting external audit, and certified with both INNO-Biz and venture company.
3 (positive situation)	All economic indicators are set at the level in Nov-2011, scores of product competitiveness and technology superiority, output of technology

	development, market potential, and new technology development in technology-oriented factor are set up at the highest value while other factors have the average value. Firm is set as not listed in stock market, conducting external audit, and neither certified with INNO-Biz nor venture company.
4 (negative situation in economic indicators, but positive situation in technology-oriented factor)	All economic indicators are set at the level of the IMF crisis in Dec-1997, scores of product competitiveness and technology superiority, output of technology development, market potential, and new technology development in technology-oriented factor are set at the highest value while other factors have the average value. Firm is set as listed in stock market, not conducting external audit, certified with both INNO-Biz, and venture company.
5 (positive situation in economic indicators, but negative situation in technology-oriented factor)	All economic indicators are set at the level in Nov-2011, scores of product competitiveness and technology superiority, output of technology development, market potential, and new technology development in technology-oriented factor are set at the lowest value while other factors have the average value. Firm is set as not listed in stock market, conducting external audit, and neither certified with neither INNO-Biz nor venture company.
6 (general situation in economic indicators, the highest score in output technology development and new technology development while the lowest score in market potential)	All economic indicators are set up at the average value. Scores of output of technology development, and new technology development in technology-oriented factor are set at the highest value while market potential is set at the lowest value. Other technology factors are set at the average value. Firm is set as not listed in stock market, not conducting external audit, and neither certified with INNO-Biz nor venture company.
7 (general situation in economic indicators, the lowest score in output technology development and new technology development while the highest score in market potential)	All economic indicators are set at the average value. Scores of output of technology development, and new technology development in technology-oriented factor are set at the lowest value while market potential is set at the highest value. Other technology factors have the average value. Firm is set as not listed in stock market, not conducting external audit, and neither certified with INNO-Biz nor venture company.

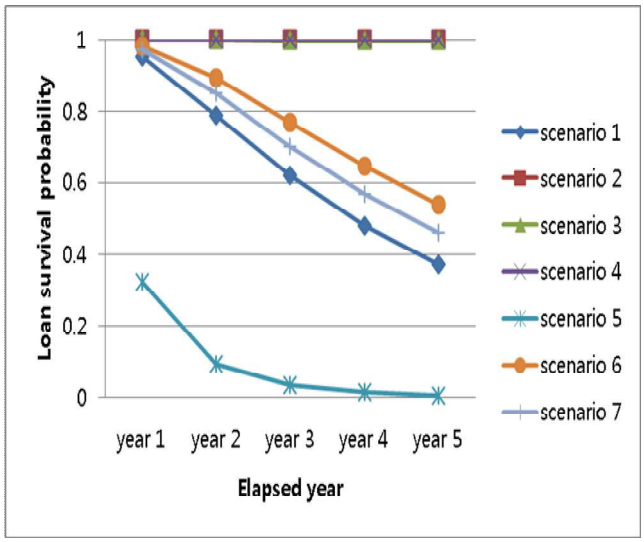


Figure 4. Stress tests result

Figure 4 shows the result of stress tests. Apparently scenarios 2 and 4 do not show significant difference in the probability of loan survival over the first five years since lending. Therefore we can conclude that when SMEs received credit guarantee under adverse economic situation (e.g., IMF crisis), their loan survival probability is very high, regardless of scores of product competitiveness and technology superiority, output of technology development, market potential, and new technology development. In contrast, although economic situation is positive (same as Nov 2011), if SME's scores of product competitiveness and technology superiority, output of technology development, market potential, and new technology development in technology-oriented factor are low, then their loan survival probability is very low. Last, among technology-oriented attributes, we can see that output of technology development and new technology development play more important roles than marketability factor in terms of increasing loan survival probability.

V. CONCOLUSION

Technology credit scoring model has been improved to increase either the accuracy of default prediction or that of loan survival time. However, previous researches have not considered interaction effects among technology-oriented factors, economic factors, and firm characteristics on recipient's loan survival probability. In this paper, we conducted a survival analysis that covered interaction effects among those variables.

In our analysis, we found several significant interaction effects with economic factors. This means that SME's loan survival time depends on the economic environments which interact with firm's technology-oriented score and firm characteristics. This implies that effect of economic conditions varies over different characteristics of firm and its technology attributes. Based on the result of 7 scenario analyses, we recommend two cases for lending: 1) when firm which is listed in stock market, but not conducting external audit with both INNO-Biz, and venture company certifications is exposed under adverse economic condition; and 2) when the firm that is not listed in stock market, but conducts external audit, and is neither certified with INNO-Biz nor venture company has high scores of product competitiveness and technology development, market potential, and new technology development in technology-oriented factor is exposed under positive economic condition. This kind of information can aid technology credit manager for appropriate funding that reflects economic environment that interacts with firm characteristics and technology.

The result and process of our model can be applied to many other areas. However, limitation also exists. In this paper, we did not conduct portfolio analysis based on an amount of guarantee fund. Based on not only the probability of default, but also loss given default, and exposure at default, further study should estimate the recipient's expected loss based on the result of survival analysis. This kind of extension is left for further areas of study.

REFERENCES

[1] Oh, I. Lee, J. D., Heshmati, A., Choi, G. G. Evaluation of credit guarantee policy using propensity score matching. Small Business Economics 33(3): 335-351.

[2] Sohn, S. Y., Moon, T. H., Kim, S. (2005). "Improved technology scoring model for credit guarantee fund." Expert Systems with Applications 28(2): 327-331.

[3] Jeon, H. J. and S. Y. Sohn (2008). "The risk management for technology credit guarantee fund." Journal of the Operational Research Society 59(12): 1624-1632.

[4] Moon, T. H., Sohn, S. Y. (2010a). Technology credit scoring model considering both SME characteristics and economic conditions: The Korean case, Journal of the Operational Research Society, 61(1): 666-675.

[5] Moon, T. H., Sohn, S. Y. (2010b). Survival analysis for technology credit scoring adjusting total perception. Journal of the Operational Research Society published online 16 June.

[6] Jacobson, T., Linde, J., Roszbach, K. (2006). "Internal ratings systems, implied credit risk and the consistency of banks' risk classification policies" Journal of Banking and Finance 30(7): 1899-1926.

[7] Fabi, F., Laviola, S., Reedtz, P. M. (2005). "The new capital accord and banks' lending decisions" Journal of Financial Stability 1(4): 501-521.

[8] Acharya, V. V., Bharath, S. T., Srinivasan, A. (2007), Does industry-wide distress affect defaulted firms? Evidence from creditor recoveries, Journal of Financial Economics 85, 787-821.

[9] Baharammirzaee, A., Ghatari, A. R., Ahmadi, P., Madani, K., (2011), Hybrid credit ranking intelligent system using expert system and artificail neurial networks. Applied Intelligence 34(1), 28-46.

[10] Khudnitskaya, A. S. (2010), Improved credit scorign with multilevel statistical modelling, PHD Thesis, Technischen Universität Dortmund.

[11] Henriksen, A. D. P., Traynor, A. J. (1999). "Practical R&D project-selection scoring tool." IEEE Transactions on Engineering Management 46(2): 158-170.

[12] Farrukh, C., Phaal, R., Probert, D., Gregory, M., Wright, J. (2000). "Developing a process for the relative valuation of R&D programmes." R&D Management 30(1): 43-54.

[13] Coldrick, S., Longhurst, P., Hannis, J. (2005). "An R&D options selection model for investment decisions." Technovation 25(3): 185-193.

[14] Stel, André, Carree, Martin Anthony and Thurik, Roy, (2005), The Effect of Entrepreneurial Activity on National Economic Growth, Small Business Economics, 24(3), p. 311-321.

[15] Wong, P. K., Y. P. Ho, et al. (2005). "Entrepreneurship, Innovation and Economic Growth: Evidence from GEM data." Small Business Economics 24(3): 335-350.

[16] Keiding, N., Andersen, P. K., Klein, J. P. (1997). The Role of Frailty Models and Accelerated Failure Time Models in Describing Heterogeneity Due to Omitted Covariates. Statistics in Medicine 16 (1-3): 215–224.

[17] Sorge, M. (2004). "Stress-testing Financial Systems: An Overview of Current Methodologies." BIS Working papers 165.

Surrogate computational unsteady fluid model for dynamic stall simulation

Vadim Surpin
Institute for Information Transmission Problems RAS
Moscow, Russia
vadim@iitp.ru

Alexander Bernstein
International Research Institute for Advanced Systems
Moscow, Russia
a.bernstein@irias.ru

Yuri Sviridenko
Central Aerohydrodynamic Institute
Moscow, Russia
ysviridenko@yandex.ru

***Abstract* — Numerical techniques based on Navier-Stokes equation to investigate a dynamic stall event are widely used. Applying numerical methods to solve these equations allows defining the pressure distribution as well as forces and moments affecting the aerodynamic airfoil in dynamic stall mode. Despite the significant progress in computing and numerical techniques, calculation viscous unsteady flow of aerodynamic units requires extensive calculating resources. It does not allow using CFD effectively in designing tasks and dynamic real-time simulation of objects.**

Keywords – numerical methods, surrogate model, aerodynamics, dynamic stall.

I. INTRODUCTION

Presently, numerical techniques based on Navier-Stokes equation to investigate a dynamic stall event are widely used [1, 2]. Applying numerical methods to solve these equations allows defining the pressure distribution as well as forces and moments affecting the aerodynamic airfoil in dynamic stall mode. Despite the significant progress in computing and numerical techniques in Computational Fluid Dynamics (CFD) software, application of CFD codes for calculation viscous unsteady flow of aerodynamic units requires extensive calculating resources. It does not allow using CFD effectively in designing tasks and dynamic real-time simulation of objects (e.g. flight test-rigs).

The Predictive Simulation Technology [3] based on Data Handling developed in IRIAS&IITP RAS is used in the research for constructing the fast surrogate models (or metamodels) for predicting the integral and distributed aerodynamic time characteristics of the airfoils under the unsteady flow condition with dynamic stall for a given airfoil geometry, unperturbed flow parameters and parameters describing time fluctuation of airfoil's Angle of Attack in time.

The constructed surrogate model is based on the data – the results of the computational experiments performed by the in-house CFD-solver FlowVision intended for modeling 3D laminar and turbulent steady and unsteady gas and liquid flows in complex geometries.

II. PROBLEM STATEMENT

The problem of predicting the integral and distributed aerodynamic time characteristics of airfoils (aircraft wing cross-sections or helicopter rotor blade) under the unsteady flow condition with dynamic stall is likely to arise in the industrial test cases. Thus as this poses one of the greatest challenges to surrogate modeling it became the subject of an initial study for the construction of a Fast Data-based Reduced Order Model on the basis of the results of the computational experiments with chosen CFD solver considered as Full Order Model.

The stall under unsteady flow (the so-called dynamic stall) is significantly different from the stall under steady flow. If the angle of attack is increasing at a noteworthy rate, the stall is delayed, as the dynamic stall angle of attack is greater than the angle of stall under the steady flow. After the onset of a dynamic stall, the lifting force and negative pitching moment originated in the resulting unsteady process are found greater than under the steady stall.

The dynamic stall event should be essentially considered under various industrial tasks and following conditions:

- Passenger aircraft penetration in high turbulence zones;
- Light aircraft maneuvers;
- Helicopter rotor blade flow under various regimes, etc.

For instance, one of the rotor blade aerodynamic characteristic features is that rotor blade cross-sections work under the unsteady flow at almost all helicopter flight regimes. It is known that the unsteady flow significantly influences airfoil aerodynamic characteristics [4].

Dynamic stall simulation is a complicated problem and the main investigation method employed to date is carrying out extremely expensive experiments in wind tunnels [5]. Moreover, due to the fact that the experimental determination of airfoil unsteady aerodynamic characteristics in wind tunnels is a very complicated task, airfoil characteristics, which are traditionally used were obtained from wind tunnels during

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7 / 2007-2013) under a grant agreement number 233665. FFAST (Future Fast Aeroelastic Simulation Technologies) is a collaborative research project aimed at developing, implementing and assessing a range of numerical simulation technologies to accelerate future aircraft design. Advances in critical load identification and reduced order modelling methods will potentially provide a step change in the efficiency and accuracy of the dynamic aeroelastic loads process. The partners in FFAST are: University of Bristol, INRIA, CSIR, TU Delft, DLR, IRIAS, University of Liverpool, Politecnico di Milano, NUMECA, Optimad Engineering, Airbus, EADS-MS, IITP and UCT.

steady blowing to calculate forces and moments of helicopter rotors [5]. For new helicopter airfoils such experimental data are not available. The corresponding problem fo passenger aircrafts is insignificant since almost never appears in today practice.

Due to the above difficulties, mathematical simulation under dynamic conditions is getting to be the main tool for the study of airfoil unsteady aerodynamic characteristics [6].

Therefore, simulation techniques enhancement is an important and challenging task. However, despite the significant progress in computing and numerical techniques, detailed CFD codes being applied to calculate detached viscous unsteady flow requires extensive computational resources, which limits its usefulness in real-time design and simulating object dynamics (e.g. flight test-rigs).

For the above reasons, generation of fast Reduced Order Model for the time-consuming CFD-code, which describes loading of aircraft wing cross-sections or helicopter rotor blade under dynamic stall, was selected as a test case for the research.

Meanwhile the surrogate model should correctly indicate influence of the main factors:

- Mach and Reynolds numbers of incoming flow,
- the law of variation of angle of attack,
- cross-section geometry,

on unsteady forces and moments affecting the airfoil as well as on pressure distribution along the wing airfoil or blade section.

III. Experiment Setup

Considered problem is behavior of an airfoil under the unsteady flow condition with dynamic stall. The stall under the unsteady flow (the so-called dynamic stall) is significantly different from the stall under the steady flow:

- If the angle of attack is being noteworthy increased, the stall is prolonged, as the dynamic stall angle of attack is greater than the angle of stall under the steady flow;
- After the beginning of a dynamic stall, the lifting force and negative pitching moment originated in the following unsteady process are found greater than under the steady.

To model of dynamic stall under unsteady flow the following experiment proposed. The airfoil performs a pitching motion (Fig. 1) expressed as

$$\alpha = \alpha(t) = \alpha_0 - \alpha_m \times \left(\cos\left(\frac{Sh \times V}{b} \times t\right) - 1\right), \qquad (1)$$

where:

- $\alpha(t)$ – represents the instantaneous angle of attack,
- α_0 – initial angle of attack,
- α_m – amplitude of airfoil's oscillation,
- Sh – Strouhal number,
- V – flow velocity,
- b – length of airfoil's chord.

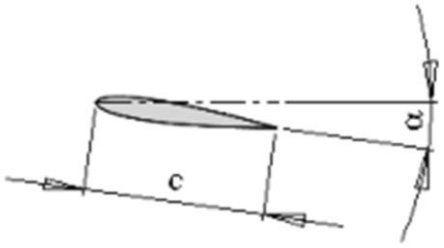


Figure 1. The schematic of airfoil's pitching motion.

A. Input data

The input data required for a detailed unsteady CFD analysis consists of:

- Description of airfoil geometry;
- Unperturbed flow parameters: Mach number (M) and Reynolds number (Re) of the incoming flow;
- Parameters (α_0, α_m, Sh) describing a fluctuation of airfoil's Angle of Attack α in time t - the current angle of attack $\alpha = \alpha(t)$.

B. Output Data

The output data consist of:

- The forces and moments affecting the airfoil;
- Pressure distribution along the airfoil.

Both as a function of time t.

C. CFD-solver

For solving the discussed problem, the CFD code FlowVision has been chosen as Full Order Model – a source of the experimental data for constructing the surrogate models. FlowVision is a general purpose CFD software for modeling 3D laminar and turbulent steady and unsteady gas and liquid flows in complex geometries. Finite Volume approach forms the basis for the software numerical contents. High-accuracy numerical schemes, efficient numerical methods, and robust physical models guarantee reliable results. FlowVision is an easy to use CFD code with intuitively straightforward interface. The post-processor provides a user with up-to-date visualization methods and data processing tools.

D. Settings

To get results that can be used for surrogate model training the number of time steps n of the full order CFD experiment was set to a constant value so that every experiment produced exactly two periods $2T$ of airfoil oscillations and every period contained fixed number of data points $n/2$. Every data point contains lift coefficient value for the full airfoil and pressure distribution over both sides of the airfoil.

The data acquired for the first period is transient, i.e. the first period is the time required for the CFD code to set up and it isn't used to surrogate model training. Data of the second and further periods considered to be valid.

As the process of surrogate model construction requires a large amount of experimental data for training, we decided to take two types of full order experiments. The first type is high-accuracy experiments that take much time to do. The CFD solver selected requires approximately 24 hours to produce required results if running a four-core desktop computer. The second type of experiment is fast CFD calculations on a rough grid with a large time step.

For the high accuracy models we set up computational grid of approximately 150 thousand cells and only 50 thousand cells for rough experiments.

IV. Experiment Results

An overview of the physical processes of dynamic stall is shown on Fig. 2. It is a velocity vector field which shows a complex structure of air flows near the airfoil in presence of dynamic stall effect. The dynamic stall effect

For the purposes of building surrogate model there were produced about 300 full order computational experiments. A large part of them is fast experiments on a rough grid. A smaller part of experiments is higher accuracy experiments. Their results are used as a reference to ensure the lower accuracy results reflect the physics of the dynamic stall effect (Fig. 3).

According to the figure, lower accuracy model reflects the general behavior of the flow but direct comparison to the high accuracy model using standard techniques such as root mean square error gives large values for residuals due to phase effects.

Despite this fact further research will be based on rough models as the reflect the basic physical effects and fast enough to produced required set of training data.

V. Surrogate Modelling Technology

The Predictive Simulation Technology based on Data Handling developed in IRIAS&IITP RAS is used for constructing the surrogate models (or metamodels) in this research. This technology was specifically dedicated to EADS Business Units (AIRBUS, EUROCOPTER, ASTRIUM, etc.), with a first successful phase called MACROS (Multidisciplinary Aeronautic Capability Research On Simulation) and completed in 2009. The technology [3] is based on advanced mathematical core and allows integrating the domain-specific knowledge, models and data into Surrogate models.

MACROS technology is implemented in MACROS Technological Tool for predictive modeling and simulation

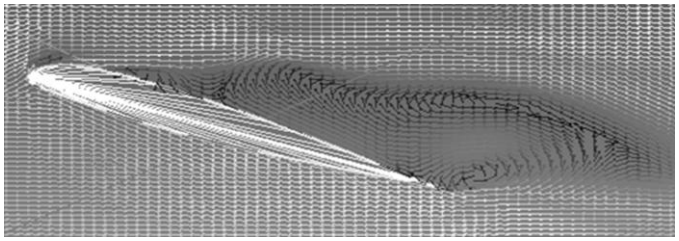


Figure 3. Flow speed vector field showing the dynamic stall effect.

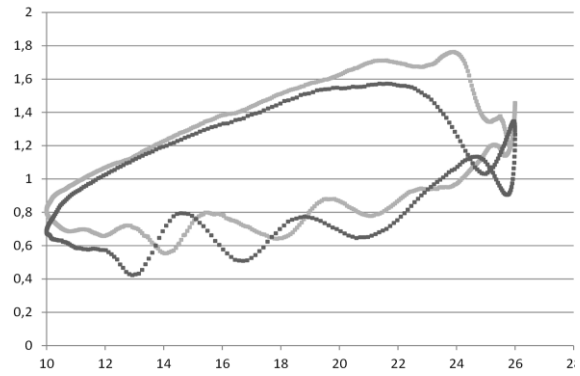


Figure 2. High accuracy (light line) and fast (dark line) lift coefficient CFD computation results for input parameters$Sh = 0.2$, $\alpha_0 = 10°$ and $\alpha_m = 8°$

based on data handling. The MACROS Technological Tool allows reducing the dimension of the input vectors, constructing the data-based dependencies and evaluating its accuracy and doing the other data-processing procedures in automatic mode for designing the surrogate models.

The main points of the Technology:

(1) The considered problem is to predict the values of the characteristic Y of some chosen object for the specified input vector X that includes the object's digital description and describes also its environment and control parameters.

(2) Let M be some chosen initial model (or method) considered as Full Order Model (FOM) that allows obtaining the value Y (response, output) for the specified input vector X. The model M determines the response function

$$Y = F_M(X), X \in \boldsymbol{X} \subset R^{\boldsymbol{p}}, Y \in R^q. \quad (2)$$

where R^p and R^q are p -dimensional and q -dimensional Euclidean spaces respectively.

(3) Usually FOM M is either a full-scale experiment or a computational experiment based on a solution of the differential equations and hence may be expensive and time-consuming. The problem is to construct the new Reduced Order Model (ROM) that is "close" to the FOM (it has the same accuracy) but essentially increases the calculation speed.

(4) The results of the experiments with the initial model M are available. These results form the Data Set

$$D_N = \{(X_i, Y_i = F_M(X_i)), i = 1, 2, \dots, N\}. \quad (3)$$

(5) The Learning Data Set D_N is used for constructing the Data-based dependency

$$Y = F_{SM}(X) = F_{SM}(X|DN) \quad (4)$$

providing an approximate equality

$$F_{SM}(X) \approx F_M(X) \quad (5)$$

that has to hold for all $X \in \boldsymbol{X}$ and not only $X \in X_N = \{X_i, i = 1, 2, \dots, N\}$.

The High Dimension Approximator Generic Tool (a part of the MACROS Technological tool) is used for constructing the Data-based dependency.

(6) Thus the new Data-based model S_M described by the constructed dependency $Y = F_{SM}(X)$ can replace the initial model M and may be regarded as a Reduced Order Model, or a surrogate model, or a metamodel (model over model).

VI. Application of the Technology to the Dynamic Stall Approximation

To build a dynamic stall surrogate model based on the MACROS high dimension approximation tool we used several approaches:

- Direct MACROS approximator exploitation;
- MACROS Approximator coupled with dimension reduction;
- Approach based on time-series analysis.

First experiments using direct approximator exploitation approach were carried out using a training set of 50 data vectors. Each data vector $\bar{x}_i = \{x_i^1, \dots, x_i^N\}$, $N = 360$, consists of N data points where N is a number of CFD solver steps per one period. As it was mentioned above N is fixed and equal to 360 steps per period. Index i in the notation of vector X stands for experiment number, so first training set had $i = 50$ data vectors.

Approximator build using $i = 50$ showed ability to catch the tendency but the error value was large. Increasing the number of experiments i to 100 gave better results and further growth of i leaded to error value decrease. We used root mean square method to calculate error value.

Along with extensive way of increasing number of data vectors used as a training base for the approximator, several alternative approaches where taken to increase accuracy.

The first alternative is a Dimension Reduction approach. It is based on the idea that our data vector is quite large – it has at least three hundred sixty points used as a training data. There should be interdependency between points which Dimension Reduction should eliminate. The first attempt was to do a Fourier transformation of the input data vector $\boldsymbol{x}$: $\boldsymbol{y}_i = \{Re\, F(\bar{x}_i), Im\, F(\bar{x}_i)\}$ and applying filtering to reduce number of its components $\{x_i^1, \dots, x_i^N\}$. The approach give approximately the same error value as direct approximator exploitation.

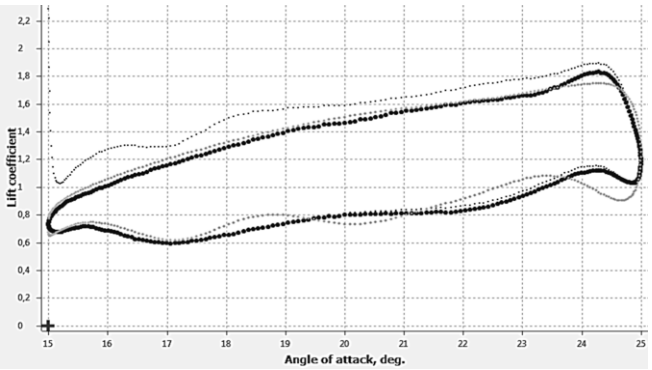


Figure 4. Full order (light line) and surrogate (dark line) lift coefficient computation results for input parameters $Sh = 0.4$, $\alpha_0 = 15°$ and $\alpha_m = 5°$

The second alternative approach is based on the fact that we know that interdependency exists. The data is a time-series data and we can use the fact explicit. To employ the fact we used a time-series analysis methods.

The first attempt was to transform a data vector to a difference vector where each successive point $y_i = x_i - x_{i-1}$ is a difference of original point value and previous point value. Index i stands for number of time step or index of i-th component of initial data vector X. The approach gave the same results as the basic one.

Another attempt was to employ autoregressive methods to set up an explicit dependency between points of our time series. In general, the approach states that each successive point is a function of N previous points. The traditional approach is to use linear function or apply some kernel function. Our approach is to use MACROS as a function approximator:

$$x_i = f(x_{i-1}, \dots, x_{i-n}) + \varepsilon, \qquad (6)$$

where f – non-linear function and ε – random error. In our case f was produced by the MACROS approximator and represented as executable computer code.

The approach demonstrated root mean square error in the range of 5-10 percent and calculation time less than one second. Fig. 4 shows a typical approximation result. The vertical axis is a lift coefficient and the horizontal axis is angle of attack. The dark line is the approximation result while the light one is full order CFD result. The upper tail of the light line is a plot of first period of oscillations which is transient (not used for training) since CFD code requires some time to set up. The figure shows result for the following values of input parameters: $Sh = 0.4$, $\alpha_0 = 15°$ and $\alpha_m = 5°$. Its relative standard deviation is nearly 6 percent which is typical for the approach.

VII. Conclusion

Proposed approach using time-series analysis methods for data preprocessing and MACROS approximator as surrogate model production tool showed ability to approximate lift coefficient data in presence of dynamic stall effect with standard deviation in range 5-10% and calculation time less than 1 second.

References

[1] A. Gonzales, X. Munduate (2007). Unsteady modeling of the oscillating S809 aerofoil and NREI phase VI parked blade using the Beddoes-Leishman dynamic stall model // In: Journal of Physics: Conference Series, 75, p. 1 – 8, IOP Publishing.

[2] A. Spentzos, G. Barakos, K. Badcock, B. Richardo (2006). Modeling three-dimensional dynamic stall of helicopter blades using computational fluid dynamics and neural networks // Proceedings of the Institute of Mechanical Engineers, vol. 220, Part G: J. Aerospace Engineering, p. 605 – 618

[3] A.P. Kuleshov, A.V. Bernstein (2009). Cognitive technologies in adaptive models of complex plants. // In: Keynote papers of 13th IFAC Symposium on Information Control Problems in Manufacturing (INCOM'09), June 3-5, Moscow, Russia. pp. 70 - 81.

[4] W.J. McCroskey, L.W. Carr, and K.W. McAlister (1976). Dynamic Stall Experiments on Oscillating Airfoils // AIAA Journal 1976 vol.14 no.1, p. 57-63.

[5] G. Leishman (1990). Dynamic stall experiments on the NACA 23012 aerofoil, Experiments in Fluids // Springer-Verlag.

[6] K. Nguyen, W. Johnson (1998). Evaluation of Dynamic Stall Models with UH-60A Airloads Flight Test Data // AHS International 54th Annual Forum Proceedings, Washington, D.C., May 20-22, pp. 576–587.

Optimal Design for Degradation Tests Based on Gamma Process with Random Effects

Chih-Chun Tsai
Department of Mathematics
Tamkang University
Tamsui, Taiwan
Email: chihchuntsai@mail.tku.edu.tw

Sheng-Tsaing Tseng
Institute of Statistics
National Tsing-Hua University
Hsinchu, Taiwan
Email: sttseng@stat.nthu.edu.tw

N. Balakrishnan
Department of Mathematics and Statistics
McMaster University
Hamilton, Ontario
Canada L8S 4K1
Email: bala@mcmaster.ca

***Abstract*—Degradation models are usually used to provide information on the reliability of highly reliable products that are not likely to fail within a reasonable period of time under the traditional life tests or accelerated life tests. Gamma process is a natural model for describing degradation paths which exhibit a monotone increasing pattern, while the commonly used Wiener process is not appropriate in such a case. In this paper, we discuss the problem of optimal design for degradation tests based on a gamma degradation process with random effects. In order to conduct a degradation experiment efficiently, several decision variables (such as the sample size, inspection frequency, and measurement numbers) need to be determined carefully. These decision variables affect not only the experimental cost, but also the precision of the estimates of lifetime parameters of interest. Under the constraint that the total experimental cost does not exceed a pre-specified budget, the optimal decision variables are found by minimizing the asymptotic variance of the estimate of the 100p-th percentile of the lifetime distribution of the product. A laser data is used to illustrate the proposed method.**

***Index Terms*—Optimal design, gamma process, random effects, asymptotic normality.**

I. Introduction

Due to the strong pressure from market, manufacturers are usually required to provide information on the reliability of their products (such as the 100p-th percentile of the lifetime distribution) to their customers. For highly reliable products, however, it is quite difficult to obtain the product's lifetime through traditional life tests within a reasonable period of time. Even though one may use accelerated life tests by testing the products at higher levels of stress than the normal use condition (such as elevated temperatures or voltages), these methods provide little assistance since either no or at most few failures are likely to occur within a reasonable testing duration. In such a situation, if there exists a QC whose degradation over time can be related to reliability, then the product's lifetime can be estimated well by collecting such a degradation data. Detailed discussions on degradation models and their applications can be found in Nelson [12], Singpurwalla [15], Chao [5], Meeker & Escobar [11], Bagdonavicius & Nikulin [3], Tseng & Peng [20], Wang [23], Wang $et\ al.$ [24], and Ye $et\ al.$ [27].

The performance of a degradation test strongly depends on the suitability of the assumed model of a product's degradation path. For degradation paths involving independent nonnegative increments, gamma processes are more suitable for describing the deterioration of the product. Bagdonavicius & Nikulin [2] modeled traumatic events by a gamma process, and discussed possibly time-dependent covariates. Lawless & Crowder [8] constructed a tractable gamma process by incorporating random effects. Park & Padgett [14] provided several new degradation models that incorporate an accelerated test variable based on stochastic processes such as a gamma process. Crowder & Lawless [6] used a gamma process to illustrate their single-inspection policy for the maintenance of automobile brake pads. For some recent applications of gamma degradation models, see Wang [22], Noortwijk [13], and Tsai $et\ al.$ [16], [17].

To conduct a degradation test efficiently, Tseng & Yu [21] proposed an intuitive rule for determining an appropriate termination time for a degradation experiment. Yu & Tseng [29] proposed a quasilinear model to address the associated optimal design problem. Marseguerra $et\ al.$ [10] designed optimal degradation tests by the use of multi-objective genetic algorithms. For some other pertinent work on the optimal design problems, one may refer to Tseng & Liao [19], Wu & Chang [26], Yu & Chiao [28], and Tseng $et\ al.$ [18].

In degradation modeling and analysis, random-effect formulation facilitates the handling of unit-to-unit variability in a convenient way. In this study, we first deal with the optimal design for a gamma degradation process with random effects. Under the constraint that the total experimental cost does not exceed a pre-specified budget, the optimal decision variables mentioned above are all determined by minimizing the asymptotic variance of the estimate of the 100p-th percentile of the lifetime of the product. Furthermore, in practical applications, one would also be interested in addressing the effects of wrongly treating a "random-effects" model as a "fixed-effects" model. More specifically, if the true gamma degradation path follows a random-effects model but gets wrongly assumed as a fixed-effects model, we will examine the penalty (effect) on the accuracy and precision of the estimate of the product's 100p-th percentile.

The rest of this paper is organized as follows. Section II presents the description and formulation of the problem. Section III deals with the estimation of the model parameters by the EM algorithm and describes the corresponding optimal

design. Section IV uses a laser data from the literature to illustrate the proposed method. Finally, some concluding remarks are made in Section V.

II. DESCRIPTION AND FORMULATION OF THE PROBLEM

Let $L(t)$ denote the degradation path of the product, where $t \geq 0$, and $L(0) = 0$. Then, the product's lifetime T can be suitably defined as the first passage time when $L(t)$ crosses the failure threshold level ω; that is,

$$T = \inf\{t | L(t) \geq \omega\}. \tag{1}$$

In the following, we assume that the degradation path of the product follows a gamma process with random effects (Lawless & Crowder, [8]). That is, for given t and Δt,

$$M_0 : \Delta L(t) \sim Ga(\eta \Delta t, \nu^{-1}), \tag{2}$$

where $\Delta L\,(t) = L(t + \Delta t) - L(t)$ and $Ga(\eta\Delta t, \nu^{-1})$ denotes a gamma distribution with shape parameter $\eta\Delta t$ and scale parameter ν^{-1}. Moreover, we assume that ν follows a gamma distribution $Ga(\delta, r^{-1})$ with pdf

$$g(\nu) = \frac{\nu^{\delta-1} r^{\delta} e^{-r\nu}}{\Gamma(\delta)}, \quad \nu > 0. \tag{3}$$

Since $L(t)$ is strictly increasing in t, and $\delta L(t)/(r\eta t)$ has an F-distribution, whose cdf is denoted by $F_{2\eta t, 2\delta}(x)$, the cdf of T for the model M_0 can be expressed as

$$F_0(t) = P(T \leq t) = \frac{B(\frac{\omega}{\omega + r}; \eta t, \delta)}{B(\eta t, \delta)}, \tag{4}$$

where $B\,(x; a, b) = \int_x^1 z^{a-1}(1-z)^{b-1} dz$ is the upper incomplete beta function, and $B(a, b)$ is the complete beta function, i.e., $B(a, b) = B(0, a, b)$. Let $\boldsymbol{\theta}_0 = (\eta, \delta, r)$ be the model parameter vector. Then, the $100p$-th percentile of the product's lifetime for the model M_0 can be obtained as $t_p(\boldsymbol{\theta}_0) = F_0^{-1}(p)$. Now, we are interested in designing an efficient degradation experiment in such a way that the $100p$-th percentile of the product can be estimated as precisely as possible.

Suppose n units are randomly selected for conducting a degradation experiment, and the measurements of each unit are made every f units of time until time $t_m = fmt_u$, where t_u is one unit of time. Let $L_i(t_j)$ denote the sample path of the i-th tested unit at time t_j, where $1 \leq i \leq n$ and $1 \leq j \leq m$, so that $ft_u = t_j - t_{j-1}$ and $t_0 = 0$. Based on $\{L_i(t_j)\}_{j=1}^m$, where $i = 1, \cdots, n$, let $\hat{\boldsymbol{\theta}}_0$ denote the MLE of the parameter vector $\boldsymbol{\theta}_0$. Clearly, the decision variables (n, f, m) affect the experimental cost as well as the precision of the estimate of the $100p$-th percentile, $t_p(\hat{\boldsymbol{\theta}}_0)$. Let $\mathrm{TC}(n, f, m)$ denote the total cost of conducting such a degradation experiment. Then, a typical decision problem of interest can be formulated as follows:

Minimize

$$\mathrm{AVar}(t_p(\hat{\boldsymbol{\theta}}_0) | n, f, m),$$

subject to

$$\mathrm{TC}(n, f, m) \leq C_b,$$

where $\mathrm{AVar}(t_p(\hat{\boldsymbol{\theta}}_0)|n, f, m)$ denotes the asymptotic variance of the MLE $t_p(\hat{\boldsymbol{\theta}}_0)$, and C_b is the total pre-specified budget for conducting the degradation experiment.

In the following section, we will address this optimization problem.

III. THE OPTIMAL TEST PLAN

To conduct a degradation test efficiently, the framework for solving the above described optimization problem consists of three parts:

(1) the MLE of the $100p$-th percentile,
(2) the asymptotic variance of the MLE of the $100p$-th percentile, and
(3) the total cost of the degradation experiment.

A. The Estimation of $t_p(\boldsymbol{\theta}_0)$

Let $Y_{ij} = L_i(t_j) - L_i(t_{j-1})$, where $1 \leq i \leq n$, and $1 \leq j \leq m$. From (2), it is seen that conditional on the random effect ν, $L(t)$ has a gamma process, and ν follows a gamma distribution. By the independent increment property of the gamma process, the likelihood function for the model M_0 can be expressed as

$$\mathcal{L}(\boldsymbol{\theta}_0) = \frac{r^{n\delta}\{\Gamma(\delta + \eta t_m)\}^n \prod_{i=1}^n \prod_{j=1}^m y_{ij}^{\eta f - 1}}{\{\Gamma(\eta f)\}^{nm}\{\Gamma(\delta)\}^n \left(\prod_{i=1}^n (y_{im} + r)\right)^{\eta t_m + \delta}}, \tag{5}$$

where $\Gamma(\cdot)$ is the complete gamma function, and $y_{im} = L_i(t_m)$. Maximization of the likelihood function in (5) is analytically intractable. In the following, we use the EM algorithm (which consists of the E- and M-steps) to obtain the MLE of $\boldsymbol{\theta}_0$, which works as detailed below.

E-Step. Calculate $Q(\boldsymbol{\theta}_0; \boldsymbol{\theta}_0^{(k)})$, where

$$\begin{aligned} Q(\boldsymbol{\theta}_0; \boldsymbol{\theta}_0^{(k)}) = {} & (\eta f - 1) \sum_{i=1}^n \sum_{j=1}^m \ln(y_{ij}) - r \sum_{i=1}^n \left(\frac{m\eta^{(k)} f + \delta^{(k)}}{\sum_{j=1}^m y_{ij} + r^{(k)}} \right) \\ & + (m\eta f + \delta - 1) \\ & \times \left(n\psi\left(m\eta^{(k)} f + \delta^{(k)}\right) - \sum_{i=1}^n \ln\left(\sum_{j=1}^m y_{ij} + r^{(k)}\right) \right) \\ & + n\delta \ln(r) - nm \ln\Gamma(\eta f) - n \ln\Gamma(\delta). \end{aligned} \tag{6}$$

M-Step. Choose $\boldsymbol{\theta}_0^{(k+1)} = (\eta^{(k+1)}, \delta^{(k+1)}, r^{(k+1)})$ to be the value of $\boldsymbol{\theta}_0$ that maximizes $Q(\boldsymbol{\theta}_0; \boldsymbol{\theta}_0^{(k)})$. Hence, we have

$$\begin{aligned} \eta^{(k+1)} = \frac{1}{f}\psi^{-1}\Bigg(& \frac{1}{nm}\sum_{i=1}^n \sum_{j=1}^m \ln(y_{ij}) + \psi\left(m\eta^{(k)} f + \delta^{(k)}\right) \\ & - \frac{1}{n}\sum_{i=1}^n \ln\left(\sum_{j=1}^m y_{ij} + r^{(k)}\right)\Bigg), \end{aligned}$$

where $\psi^{-1}(\cdot)$ is the inverse of the digamma function, and $\delta^{(k+1)}$ satisfies the equation

$$\psi(\delta^{(k+1)}) - \ln(\delta^{(k+1)}) = \psi\left(m\eta^{(k)}f + \delta^{(k)}\right) - \frac{1}{n}\sum_{i=1}^{n}\ln\left(\sum_{j=1}^{m} y_{ij} + r^{(k)}\right) - \ln\left(\frac{1}{n}\sum_{i=1}^{n}\left(\frac{m\eta^{(k)}f + \delta^{(k)}}{\sum_{j=1}^{m} y_{ij} + r^{(k)}}\right)\right),$$

and

$$r^{(k+1)} = n\delta^{(k+1)}\left(\sum_{i=1}^{n}\left(\frac{m\eta^{(k)}f + \delta^{(k)}}{\sum_{j=1}^{m} y_{ij} + r^{(k)}}\right)\right)^{-1}.$$

The MLE $\hat{\boldsymbol{\theta}}_0$ of $\boldsymbol{\theta}_0$ can be found by iteration until the difference, $\mathcal{L}(\boldsymbol{\theta}_0^{(k+1)}) - \mathcal{L}(\boldsymbol{\theta}_0^{(k)})$, is within a specified tolerance level, say, 10^{-6}. Then, the estimate of the $100p$-th percentile of the lifetime of the product, $t_p(\hat{\boldsymbol{\theta}}_0)$, can be found simply by substituting $\hat{\boldsymbol{\theta}}_0$ into (4).

B. Computation of AVar$(t_p(\hat{\boldsymbol{\theta}}_0))$

By using the BAN property of the MLE, the asymptotic distribution of $t_p(\hat{\boldsymbol{\theta}}_0)$ can be obtained as

$$\sqrt{n}\left(t_p(\hat{\boldsymbol{\theta}}_0) - t_p(\boldsymbol{\theta}_0)\right) \xrightarrow{d} N\left(0, \frac{\mathbf{h}_0^{'}\mathbf{I}(\boldsymbol{\theta}_0)^{-1}\mathbf{h}_0}{(f_0(t_p(\boldsymbol{\theta}_0)))^2}\right), \quad (7)$$

where

$$\mathbf{h}_0^{'} = \left(\frac{\partial F_0(t)}{\partial\eta}, \frac{\partial F_0(t)}{\partial\delta}, \frac{\partial F_0(t)}{\partial r}\right)\Big|_{t=t_p(\boldsymbol{\theta}_0)},$$

$\mathbf{I}(\boldsymbol{\theta}_0)$ is the Fisher information matrix, and $f_0(t)$ is the pdf of T for the model M_0. Explicit expressions for the elements of $\mathbf{h}_0^{'}$ and $\mathbf{I}(\boldsymbol{\theta}_0)$ can be derived easily. Thus, the asymptotic variance of $t_p(\hat{\boldsymbol{\theta}}_0)$ is seen to be

$$\text{AVar}(t_p(\hat{\boldsymbol{\theta}}_0)) = \frac{\mathbf{h}_0^{'}\mathbf{I}(\boldsymbol{\theta}_0)^{-1}\mathbf{h}_0}{n(f_0(t_p(\boldsymbol{\theta}_0)))^2}. \quad (8)$$

Now, to compute AVar$(t_p(\hat{\boldsymbol{\theta}}_0))$, an exact expression of $f_0(t)$ is required and the following lemma is useful for this purpose.

Lemma 1.

$$\frac{\partial}{\partial a}\left(\frac{B(x;a,b)}{B(a,b)}\right) = \frac{x^a}{aB(a,b)}\Bigg[\left(\psi(a) - \psi(a+b) - \ln(x) + \frac{1}{a}\right) \times {}_2F_1(\{a,1-b\},\{1+a\};x) - \frac{x(1-b)}{(1+a)^2} \times {}_3F_2(\{2-b,1+a,1+a\},\{2+a,2+a\};x)\Bigg],$$

where $\psi(z) = \frac{d\ln\Gamma(z)}{dz}$ is the digamma function, and ${}_gF_h$ is the confluent hypergeometric function defined by

$${}_gF_h(\{c_1,\cdots,c_g\},\{d_1,\cdots,d_h\};x) = \sum_{k=0}^{\infty}\frac{(c_1)_k\cdots(c_g)_k}{(d_1)_k\cdots(d_h)_k}\frac{x^k}{k!}$$

with Pochhammer symbol $(c)_k = \frac{\Gamma(c+k)}{\Gamma(c)}$, and $(c)_0 = 0$.

Now, let

$$W(x,a,b) = \frac{x^a}{aB(a,b)}\Bigg[\left(\psi(a) - \psi(a+b) - \ln(x) + \frac{1}{a}\right) \times {}_2F_1(\{a,1-b\},\{1+a\};x) - \frac{x(1-b)}{(1+a)^2}{}_3F_2(\{2-b,1+a,1+a\},\{2+a,2+a\};x)$$

Then, the pdf of T for the model M_0 is simply

$$f_0(t) = \eta W\left(\frac{\omega}{\omega+r}, \eta t, \delta\right).$$

C. Cost Function

The total cost of conducting the degradation test, TC(n,f,m), comprises the cost of conducting an experiment, the cost of measurement, and the cost of tested devices, and is given by

$$\text{TC}(n,f,m) = C_{op}fm + C_{mea}nm + C_{it}n, \quad (9)$$

where C_{op} denotes the unit cost of operation, C_{mea} denotes the unit cost of measurement, and C_{it} denotes the unit cost of device.

D. Optimization Problem

From all the statements above, the required optimization problem can now be stated as follows:

Minimize

$$\text{AVar}(t_p(\hat{\boldsymbol{\theta}}_0)|\boldsymbol{\xi}), \quad (10)$$

subject to

$$\text{TC}(\boldsymbol{\xi}) \le C_b,$$

where $\boldsymbol{\xi} = (n,f,m) \in \mathbb{N}^3$. Due to the complex form of the objective function, an analytic expression for the solution of this optimization problem does not exist. However, with the simplicity in the structure of the constraint and the integer restriction on the three decision variables, the optimal solution $\boldsymbol{\xi}^* = (n^*,f^*,m^*)$ can be easily evaluated by a complete enumeration method, as described in Yu & Tseng [29]. In the following section, we make use of a laser data from the literature to illustrate the proposed optimal design.

IV. ILLUSTRATIVE EXAMPLE

In this section, we illustrate the proposed optimal design with the laser data presented by Meeker & Escobar [11]. The QC of a laser device is its operating current, and when the operating current reaches a pre-specified failure level, the laser device is considered to have failed. Here, the failure level (ω) is 10, and Fig. 1 shows the plots of the operating current over time for 15 tested units. The measured frequency of its current is 250 hours, and the experiment was terminated at 4000 hours.

From (8), it is known that AVar$(t_p(\hat{\boldsymbol{\theta}}_0)|\boldsymbol{\xi})$ is a function of $\boldsymbol{\theta}_0$ when the test plan $\boldsymbol{\xi}$ is given. Hence, for determining an optimal degradation design, we need information on the

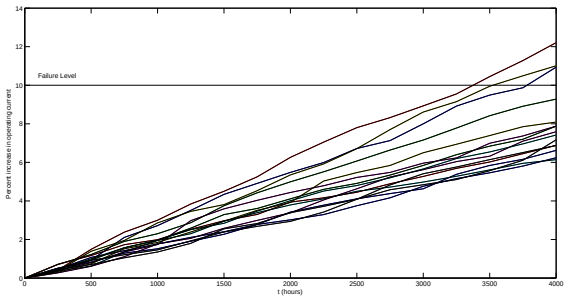


Fig. 1. Degradation paths for laser data.

parameter vector $\boldsymbol{\theta}_0$. By using the procedure described earlier in Section III-A, we determine the MLE of $\boldsymbol{\theta}_0$ to be

$$\hat{\boldsymbol{\theta}}_0 = (\hat{\eta}, \hat{\delta}, \hat{r}) = (0.0390, 28.8787, 1.4533). \qquad (11)$$

In what follows, we take the MLE of $\boldsymbol{\theta}_0$ in (11) to be the true model parameter. Now, we describe how the optimal test plan for the considered laser data can be constructed. Assume that the cost configurations are as follows: $C_{op} = \$13$/unit time, $C_{mea} = \$0.05$/measurement, $C_{it} = \$51$/unit, and the unit time is 24 hours. Further, suppose $p = 0.1$ is the lifetime percentile that is of interest. Then, under various budgets C_b, the optimal degradation plans can be obtained by using the algorithm described in Yu & Tseng [29], and the results so obtained are presented in Table I.

(1) **Optimal Test Plan**
For example, when $C_b = 1250$, the optimal test plan turns out to be $(n^*, f^*, m^*) = (15, 2, 18)$. In other words, the optimal sample size is 15, the optimal measurement frequency is $2 \times 24 = 48$ hours, and the optimal measurement number is 18. This means that the total test time for the degradation experiment is $2 \times 18 \times 24 = 864$ hours. Under such a test plan, the total cost is 1246.50, and the corresponding asymptotic variance of the 10-th percentile of the product's lifetime is 160366.07.

TABLE I
OPTIMAL TEST PLANS UNDER VARIOUS BUDGETS C_b.

C_b	n^*	f^*	m^*	AVar$(t_{0.1}(\hat{\boldsymbol{\theta}}_0)\|\boldsymbol{\xi}^*)$	total test cost
1000	11	3	11	232024.58	996.05
1250	15	2	18	160366.07	1246.50
1500	18	4	11	118674.66	1499.90

(2) **Sensitivity Analysis**
In practice, the estimated parameter $\hat{\boldsymbol{\theta}}_0 = (\hat{\eta}, \hat{\delta}, \hat{r})$ would depart from the true parameter $\boldsymbol{\theta}_0 = (\eta, \delta, r)$. Suppose ε_1, ε_2 and ε_3 denote the predicted errors for the parameters η, δ and r, respectively. Now, we study the effects of the precision of the estimates on the optimal degradation plan. Under the same cost configuration $(C_{op}, C_{mea}, C_{it}, C_b) = (13, 0.05, 51, 1250)$, Table II shows the optimal plans for various choices of $((1+\varepsilon_1)\,\eta, (1+\varepsilon_2)\,\delta, (1+\varepsilon_3)\,r)$ according to a $L_9(3^{3-1})$ orthogonal array with $\boldsymbol{\theta}_0 = (\eta, \delta, r)$. From these results, we observe that the optimal test plan (n^*, f^*, m^*) is quite robust for moderate departures from the assumed value of (η, δ, r), even when $\varepsilon_i = \pm 10\%$, for $\forall 1 \le i \le 3$.

TABLE II
OPTIMAL TEST PLANS UNDER VARIOUS CHOICES OF THE PARAMETERS $((1+\varepsilon_1)\,\eta, (1+\varepsilon_2)\,\delta, (1+\varepsilon_3)\,r)$.

ε_1	ε_2	ε_3	n^*	f^*	m^*	AVar$(t_{0.1}(\hat{\boldsymbol{\theta}}_0)\|\boldsymbol{\xi}^*)$	total cost
-10%	-10%	-10%	15	2	18	215844.19	1246.50
-10%	0	0	14	2	20	212060.96	1248.00
-10%	$+10\%$	$+10\%$	14	2	20	208400.70	1248.00
0	-10%	0	15	2	18	131506.76	1246.50
0	0	$+10\%$	15	2	18	131297.14	1246.50
0	$+10\%$	-10%	14	2	20	239746.49	1248.00
$+10\%$	-10%	$+10\%$	15	2	18	84266.11	1246.50
$+10\%$	0	-10%	15	2	18	155706.29	1246.50
$+10\%$	$+10\%$	0	15	2	18	149394.42	1246.50
0	0	0	15	2	18	160366.07	1246.50

(3) **The Case of Non-Cost Constraint**
From (8), it is easily seen that the asymptotic variance of $t_p(\hat{\boldsymbol{\theta}}_0)$ decreases when sample size n increases. Now, without imposing the budget constraint, we will investigate the effect of measurement frequency f and number of measurements m on the asymptotic variance of $t_p(\hat{\boldsymbol{\theta}}_0)$. With $n = 15$, Fig. 2 presents the mesh plot of the asymptotic variance of $t_p(\hat{\boldsymbol{\theta}}_0)$ for various combinations of $1 \le f \le 10$ and $10 \le m \le 50$. The plot demonstrates that the asymptotic variance of the estimated product's lifetime is a decreasing function of m and f. Comparing with the optimal test plan in Table I with $C_b = 1250$, we observe that the cost constraint has a critical impact on the test plans and its corresponding asymptotic variance of the estimated product's lifetime.

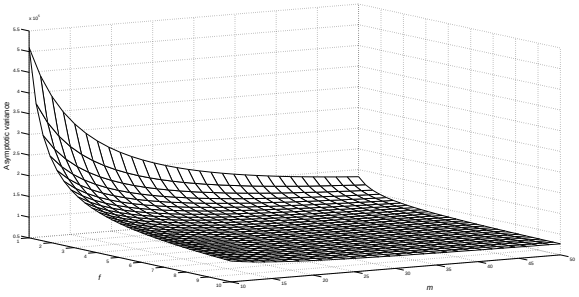


Fig. 2. Asymptotic variance of $t_p(\hat{\boldsymbol{\theta}}_0)$ for (f, m) with $n = 15$.

V. CONCLUDING REMARKS

This paper deals with the optimal design for degradation tests based on gamma degradation process with random effects. By minimizing the asymptotic variance of the estimated $100p$-th percentile subject to the total experimental cost not exceeding a pre-specified budget value, we derive the optimal settings of the sample size, measurement frequency, and number of measurements. The sensitivity analysis with respect to

moderate departures from the assumed values of the model parameters reveals that the optimal test plan is quite robust.

In this paper, by taking the rate parameter as a random effect (following a gamma distribution), we obtain closed-form expressions for the cdf and the pdf of the product's lifetime through the conjugate property. Even though the gamma degradation model can also be extended to allow random effects in the shape parameter, this problem is computationally quite complicated, and we hope to consider this issue for our future research.

Moreover, in the case of very highly reliable products, the degradation process may be quite slow and in such a case it would be impossible to obtain precise estimation within a reasonable test duration. It would be reasonable in such a scenario to collect the degradation data at higher levels of stress, and then extrapolate the product's lifetime at normal use condition. This problem is of great practical interest, as it would enable us to evaluate the lifetimes of very highly reliable products in an efficient way. We are currently working on this research problem and hope to report these findings in a future paper.

ACKNOWLEDGMENT

This work was partially supported by the National Center of Theoretical Science (NCTS), National Science Council of ROC, Taiwan (Contract No: NSC-96-2628-M-007-014-MY3 and NSC-100-2118-M-032-013), and 5Y5B MOE Project (Contract No: 101N2072E1) of ROC, Taiwan.

REFERENCES

[1] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables.* New York: Dover Publications, 1972.

[2] V. Bagdonavicius and M. Nikulin, "Estimation in degradation models with explanatory variables," *Lifetime Data Analysis,* vol. 7, pp. 85–103, 2001.

[3] V. Bagdonavicius and M. Nikulin, *Accelerated Life Models : Modeling and Statistical Analysis.* New York : Chapman & Hall, 2002.

[4] R. J. Boik and J. F. Robinson-Cox, "Derivatives of the incomplete beta function," *Journal of Statistical Software,* vol. 3, pp. 1–19, 1998.

[5] M. T. Chao, "Degradation analysis and related topics: some thoughts and a review," *The Proceedings of the National Science Council, Series A,* vol. 23, pp. 555–566, 1999.

[6] M. Crowder and J. F. Lawless, "On a scheme for predictive maintenance," *European Journal of Operational Research,* vol. 16, pp. 1713–1722, 2007.

[7] K. O. Geddes, M. L. Glasser, R. A. Moore, and T. C. Scott, "Evaluation of classes of definite integrals involving elementary functions via differentiation of special functions," *Applicable Algebra in Engineering, Communication and Computing,* vol. 1, pp. 149–165, 1990.

[8] J. Lawless and M. Crowder, "Covariates and random effects in a gamma process model with application to degradation and failure," *Lifetime Data Analysis,* vol. 10, pp. 213–227, 2004.

[9] Y. L. Luke, *The Special Functions and Their Approximations, Vol. 1.* New York : Academic Press, 1969.

[10] M. Marseguerra, E. Zio, and M. Cipollone, "Designing optimal degradation tests via multi-objective genetic algorithms," *Reliability Engineering & System Safety,* vol. 79, pp. 87–94, 2003.

[11] W. Q. Meeker and L. A. Escobar, *Statistical Methods for Reliability Data.* New York: John Wiley & Sons, 1998.

[12] W. Nelson, *Accelerated Testing: Statistical Models, Test Plans, and Data Analysis.* New York: John Wiley & Sons, 1990.

[13] J. M. V. Noortwijk, "A survey of the application of gamma processes in maintenance," *Reliability Engineering & System Safety,* vol. 94, pp. 2–21, 2009.

[14] C. Park and W. J. Padgett, "Accelerated degradation models for failure based on geometric Brownian motion and gamma process," *Lifetime Data Analysis,* vol. 11, pp. 511–527, 2005.

[15] N. D. Singpurwalla, "Survival in dynamic environments," *Statistical Science,* vol. 10, pp. 86–103, 1995.

[16] C. C. Tsai, S. T. Tseng, and N. Balakrishnan, "Optimal burn-in policy for highly reliable products using gamma degradation process," *IEEE Transactions on Reliability,* vol. 60, pp. 234–245, 2011.

[17] C. C. Tsai, S. T. Tseng, and N. Balakrishnan, "Mis-specification analyses of gamma and Wiener degradation processes," *Journal of Statistical Planning and Inference,* vol. 141, pp. 3725–3735, 2011.

[18] S. T. Tseng, N. Balakrishnan, and C. C. Tsai, "Optimal step-stress accelerated degradation test plan for gamma degradation processes," *IEEE Transactions on Reliability,* vol. 58, pp. 611–618, 2009.

[19] S. T. Tseng and C. M. Liao, "Optimal design for a degradation test," *International Journal of Operations and Quantitative Management,* vol. 4, pp. 293–301, 1998.

[20] S. T. Tseng and C. Y. Peng, "Stochastic diffusion modelling of degradation data," *Journal of Data Science,* vol. 5, pp. 315–333, 2007.

[21] S. T. Tseng and H. F. Yu, "A termination rule for degradation experiments," *IEEE Transactions on Reliability,* vol. 46, pp. 130–133, 1997.

[22] X. Wang, "A pseudo-likelihood estimation method for nonhomogeneous gamma process model with random effects," *Statistica Sinica,* vol. 18, pp. 1153–1163, 2008.

[23] X. Wang, "Wiener processes with random effects for degradation data," *Journal of Multivariate Analysis,* vol. 101, pp. 340–351, 2010.

[24] Z. Wang, H. Z. Huang, Y. Li, and N. C. Xiao, "An approach to reliability assessment under degradation and shock process," *IEEE Transactions on Reliability,* vol. 60, pp. 852–863, 2011.

[25] H. White, "Maximum likelihood estimation of misspecified models," *Econometrica,* vol. 50, pp. 1–25, 1982.

[26] S. J. Wu and C. T. Chang, "Optimal design of degradation tests in presence of cost constraint," *Reliability Engineering & System Safety,* vol. 76, pp. 109–115, 2002.

[27] Z. S. Ye, L. C. Tang, and H. Y. Xu, "A distribution-based systems reliability model under extreme shocks and natural degradation," *IEEE Transactions on Reliability,* vol. 60, pp. 246–256, 2011.

[28] H. F. Yu and C. H. Chiao, "An optimal designed degradation experiment for reliability improvement," *IEEE Transactions on Reliability,* vol. 51, pp. 427–433, 2002.

[29] H. F. Yu and S. T. Tseng, "Designing a degradation experiment," *Naval Research Logistics,* vol. 46, pp. 689–706, 1999.

Investigation of cancer mortality on the basis of historical comorbidity data

Varvara V. Tsurko, Anatoly I. Michalski
Institute of Control Sciences
Russian Academy of Sciences
Moscow, Russian Federation
Email: v.tsurko@gmail.com, ipuran@yandex.ru

***Abstract*—In this paper we investigate dependencies between associated diseases that a person has at the end of his live and the cause of death. We analyze public data about cause-specific mortality in conjunction with the problem of average risk estimation on empirical data. The use of the theory of Vapnik-Chervonenkis provides informative results about differences between distributions of associated diseases in group of people who died of cancer and group of people who died of another disease. This difference uncovers a relationship between some groups of associated diseases and risk of death of cancer.**

I. Introduction

The World Health Organization predicted cancer as an increasingly important cause of morbidity and mortality in the next few decades in all regions of the world. Even if current global cancer rates (in accordance with forecasted changes in population demographics) remain unchanged, the estimated incidence of 12.7 million new cancer cases in 2008 will rise to 21.4 million by 2030. Moreover, these cancer rates are expected to increase too [5].

Experts associate the increase in cancer prevalence with population ageing and improving the quality of life. According to the World Bank income groups, the cancer rates for all cancers combined (excluding non-melanoma skin cancers) rose with increasing levels of country income. High-income countries had more than double the rate of all cancers combined of low-income countries [5]. The questions are: why the cancer is more distributed in developed countries, how can it be connected with longevity increase observed from the second part of XX century?

The main goal of this research is to analyze relationships between cancer mortality and associated diseases that a person had at the end of his live on the basis of USA mortality-comorbidity data. We track these relationships during the second part of XX century and the beginning of XXI century, investigate differences and underline diseases that are associated with cancer mortality.

In the research the Multiple Cause-of-Death Public-Use Data by the National Center for Health Statistics USA [6] is analyzed. Distribution of associated diseases presented by the ICD10 codes among people who died of cancer (C00-C97) is compared with the same distribution among people who died of another disease. In order to select more "important" diseases associated with cancer mortality we solve a problem of contrasting the distributions. By the problem of contrasting we mean the selection of associated diseases for which we have the most distinguishable distributions.

We used symmetrized Kullback-Leibler divergence as a difference measure between the two distributions. For a set of associated diseases the symmetrized Kullback-Leibler divergence was estimated from the data as a half sum of mixed entropies corrected by a penalty term. This term takes into account both the amount of empirical data and the number of considered associated diseases and it construction based on the Vapnik-Chervonenkis dimension.

The results show associated diseases connected with cancer death, differences between associated diseases depending on age, dynamics of the development and distribution of different diseases depending on year of death. This research contains the medical interpretation of the results.

II. Definitions

In this section we introduce some definitions and notations that will be used throughout the paper.

We consider the problem of estimation a distance between two distributions $p_1(x)$ and $p_2(x)$ on empirical data, where $p_1(x)$ is a distribution of associated diseases among people who died of cancer (let's name this group as a cancer group), $p_2(x)$ – a distribution of associated diseases among people who died of another disease (non cancer group). Associated diseases are grouped into blocks according to their ICD10 classification, x is a block of associated diseases.

For each ith block $(i = 1, \ldots, k)$ of associated diseases a number of people of cancer group which had a disease from the ith block (n_i) and a number of people of non cancer group which had a disease from the ith block (m_i) are calculated. Cancer and non cancer groups have histograms of associated diseases: $g_1 = (n_{(1)}, n_{(2)}, \ldots, n_{(k)})$ and $g_2 = (m_{(1)}, m_{(2)}, \ldots, m_{(k)})$, blocks of associated diseases are sorted in descending order of the absolute difference between the values $n_i / \sum_{i=1}^{k} n_i$ and $m_i / \sum_{i=1}^{k} m_i$.

Our goal is to find a set of blocks of associated diseases which are the most important for difference between cancer and non cancer death. Let α be a variable which labels what set of blocks we use now, Σ is a set of all possible sets α. In an experimental part we create follow sequence of

variables α: $\alpha_{(1)}$ – the first block of associated disease, $\alpha_{(2)}$ – the first and the second blocks,..., $\alpha_{(k)}$ – all blocks, where the order of blocks is the same as in histograms above.

Let $\widehat{p}_1(x,\alpha)$, $\widehat{p}_2(x,\alpha)$ are estimates of distributions $p_1(x,\alpha)$ and $p_2(x,\alpha)$ on empirical data. We use symmetrized Kullback-Leibler divergence as a distance between empirical estimates $\widehat{p}_1(x,\alpha)$, $\widehat{p}_2(x,\alpha)$ and $p_2(x,\alpha)$, $p_1(x,\alpha)$:

$$D(\alpha) = -\frac{1}{2}\left(\sum_x p_2(x,\alpha)\ln\frac{\widehat{p}_1(x,\alpha)}{p_2(x,\alpha)} + \right.$$
$$\left. + \sum_x p_1(x,\alpha)\ln\frac{\widehat{p}_2(x,\alpha)}{p_1(x,\alpha)}\right)$$

III. Formalization of The Problem

Our goal is to find such set of blocks of associated diseases for which distribution of associated diseases in cancer group maximally differ from the distribution in non cancer group. In terms of Kullback-Leibler divergence: $D(\alpha) \stackrel{\alpha}{\to} \max$

In the rest of the article we consider a functional of average risk as a characterizing criterion of the distance $D(\alpha)$:

$$M(\alpha) = -\frac{1}{2}\left(\sum_x p_2(x,\alpha)\ln\widehat{p}_1(x,\alpha) + \right.$$
$$\left. + \sum_x p_1(x,\alpha)\ln\widehat{p}_2(x,\alpha)\right) \quad (1)$$

The distributions $p_1(x,\alpha)$ and $p_2(x,\alpha)$ are unknown and they are approximated by frequencies. We can use a trivial approximation by frequencies $\nu_1(x,\alpha)$ and $\nu_2(x,\alpha)$ which are equal to a portion of people who had an associated disease from block x and died of cancer or of another disease respectively. If x is an ith block of associated disease, α consists of k blocks, then frequencies are defined as:

$$\nu_1(x,\alpha) = \frac{n_i}{\sum\limits_{i=1}^{k} n_i}, \nu_2(x,\alpha) = \frac{m_i}{\sum\limits_{i=1}^{k} m_i},$$

To avoid zero value under logarithm in (1) we use empirical estimates $\widehat{p}_1(x)$ and $\widehat{p}_2(x)$ of distributions $p_1(x)$ and $p_2(x)$ in form:

$$\widehat{p}_1(x,\alpha) = \frac{n_i+1}{\sum\limits_{i=1}^{k} n_i + k}, \widehat{p}_2(x,\alpha) = \frac{m_i+1}{\sum\limits_{i=1}^{k} m_i + k} \quad (2)$$

These expressions are Bayes estimates of probabilities if a priori distribution of probabilities on the k-fold simplex given by $\Delta^k = \{p_1,\ldots,p_k\colon \sum\limits_{i=1}^{k} p_i = 1, p_i \geq 0, i = 1,\ldots,k\}$ is uniform.

By substitution of $\nu_1(x,\alpha)$ and $\nu_2(x,\alpha)$ instead of $p_1(x,\alpha)$ and $p_2(x,\alpha)$ in (1) we obtain so called empirical risk

$$M_e(\alpha) = -\frac{1}{2}\left(\sum_x \nu_2(x,\alpha)\ln\widehat{p}_1(x,\alpha) + \right.$$
$$\left. + \sum_x \nu_1(x,\alpha)\ln\widehat{p}_2(x,\alpha)\right) =$$
$$= -\frac{1}{2}\left(\frac{1}{\sum\limits_{j=1}^{k} m_j}\sum_{i=1}^{k} m_i \ln\frac{n_i+1}{\sum\limits_{j=1}^{k} n_j + k} + \right.$$
$$\left. + \frac{1}{\sum\limits_{j=1}^{k} n_j}\sum_{i=1}^{k} n_i \ln\frac{m_i+1}{\sum\limits_{j=1}^{k} m_j + k}\right) \quad (3)$$

The deviation between the average risk and the empirical risk can be estimated in form of an inequality

$$M(\alpha) > M_e(\alpha) - d(\alpha,\eta),$$

which is valid with probability $1-\eta$.

By maximizing on α the right part of the inequality we determine the set of blocks of associated diseases for which distribution of the associated diseases in cancer group maximal differs of the distribution of the associated diseases in non cancer group. The penalty term $d(\alpha,\eta)$ is estimated by Vapnik-Chervonenkis evaluation.

Let $x^{\alpha}_{1i}, i = 1,\ldots,L^{\alpha}_1$ denote a block of associated diseases which ith person form the cancer group had, where L^{α}_1 is a number of people who belonged to the cancer group and had an associated disease from a set α. In the same way, let $x^{\alpha}_{2i}, i = 1,\ldots,L^{\alpha}_2$ denote a block of associated diseases which ith person form the non cancer group had. Then we can obtain the following expression for the empirical risk (3):

$$M_e(\alpha) = -\frac{1}{2}\left(\frac{1}{L^{\alpha}_2}\sum_{i=1}^{L^{\alpha}_2}\ln\widehat{p}_1(x^{\alpha}_{2i},\alpha) + \frac{1}{L^{\alpha}_1}\sum_{i=1}^{L^{\alpha}_1}\ln\widehat{p}_2(x^{\alpha}_{1i},\alpha)\right) \quad (4)$$

To present the functional of average risk as an expectation and the functional of empirical risk as a mean assume additional notation $F_i(x,\alpha) = -\ln\widehat{p}_i(x,\alpha), i = 1,2$. Functions $F_i(x,\alpha), i = 1,2$ are bounded. Functionals of average and empirical risks take form:

$$M_e(\alpha) = \frac{1}{2}\left(\frac{1}{L^{\alpha}_2}\sum_{i=1}^{L^{\alpha}_2} F_1(x^{\alpha}_{2i},\alpha) + \frac{1}{L^{\alpha}_1}\sum_{i=1}^{L^{\alpha}_1} F_2(x^{\alpha}_{1i},\alpha)\right)$$
$$M(\alpha) = EF_1(x,\alpha) + EF_2(x,\alpha) =$$
$$= \int F_1(x,\alpha)dP(x) + \int F_2(x,\alpha)dP(x)$$

For such forms of functionals we can use the Vapnik-Chervonenkis result from [1] about the uniform convergence of means to expectations in a class of bounded functions and

obtain an equation for penalty term $d(\alpha, \eta)$:

$$d(\alpha,\eta) = 2\sqrt{\frac{r\left(\ln\frac{2l}{r}+1\right) - \ln\frac{\eta}{5}}{l-1}},$$

where l – the number of sample objects, r – the Vapnik-Chervonenkis dimension [1].

Substitution paraments of the task instead l and r leads to the inequality which holds with probability not less than $1-\eta$ for all sets of blocks of associated diseases composed not more than k blocks

$$M(\alpha) > M_e(\alpha) - 2\sqrt{\frac{2^{k-1}\left(\ln\frac{2(L_1^\alpha+L_2^\alpha)}{2^{k-1}}+1\right) - \ln\frac{\eta}{5}}{(L_1^\alpha+L_2^\alpha)-1}} \quad (5)$$

IV. Experimental Results

In this section we present the analysis of the data of human comorbidity and mortality. We are interested in differences between two groups of people: people who died of cancer and people who died of another disease. Usually a person in addition to underlying disease (the cause of death) has a list of associated diseases. Hence there are certain distributions of associated diseases in these two groups of people.

For the analysis the Multiply Cause-of-Death Public-Use data are used which was provided by the International Center of Health Statistics (USA) [6]. We considered the data for 1985-2008 years. Data are splitted into files yearly and for each particular year file contains the information about persons who died this year. For each person we have following information: age, date of death, a cause of death (one for each person), a list of associated diseases (up to 30 diseases for each person). Analysis is performed for three age groups: 15-34, 35-64, 65+ years old. All associated diseases are grouped into standard blocks by the first letter and two digits of their ICD10 codes and we work with more than 150 blocks.

On the figure 1 normalized histograms of associated diseases in cancer and non cancer groups are plotted (according to our definitions $n_i/\sum_{i=1}^{k} n_i$ and $m_i/\sum_{i=1}^{k} m_i, i = 1,\ldots,k$). There are three figures for different age groups. Histograms are plotted on the data about mortality and morbidity in 2008 year. Associated diseases are coded according their ICD10 classification. Blocks of associated diseases are sorted in descending order of the absolute difference between normalized histograms in cancer and non cancer groups.

Figure 1 illustrates that such associated diseases as bacterial diseases (A30-A49) and anaemias (D60-D64) are spread in the youngest age group (15-34 years old). Diseases of circulatory (I00-I99) and respiratory systems (J00-J99) are more common in elder age group. Over the age of 65 years old anaemias again become more significant in cancer death, and roles of obesity and other hyperalimentation (E65-E68) become less significant.

To evaluate the empirical and average risks on the experimental data and to find a set of associated diseases which are the most important for difference between cancer and non cancer death we consider different sets α of blocks

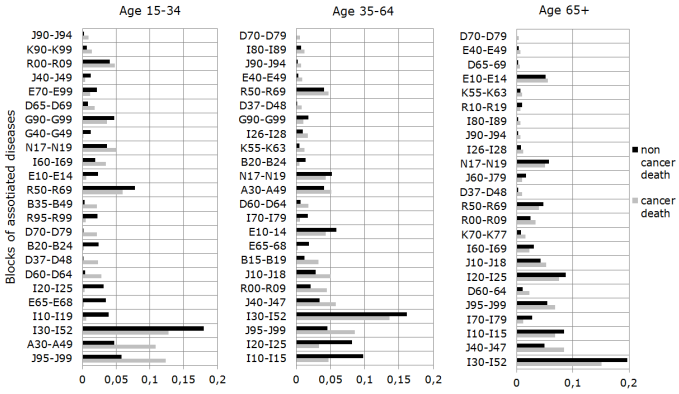


Fig. 1: Distributions of associated diseases in three age groups

of associated diseases. Then we create the sequences of sets $\alpha = (\alpha_{(1)}, \alpha_{(2)}, \ldots, \alpha_{(n)})$: for the age group 15-34 years old and 2008 year of death we have the following sets: $\alpha_{(1)} = \{\text{J95} - \text{J99}\}$, $\alpha_{(2)} = \{\text{J95} - \text{J99}, \text{A30} - \text{A49}\}$ and $\alpha_{(k)}$ — which includes all considered classes with the order of classes same as defined above.

Using the mortality data we calculate values of the functional of empirical risk (3) for each set α. According to inequality (5) we evaluate the lower bound of the average risk.

Figure 2 describes the lower bounds of the average risk $M(\alpha)$ for three age groups (data of 2008 year).

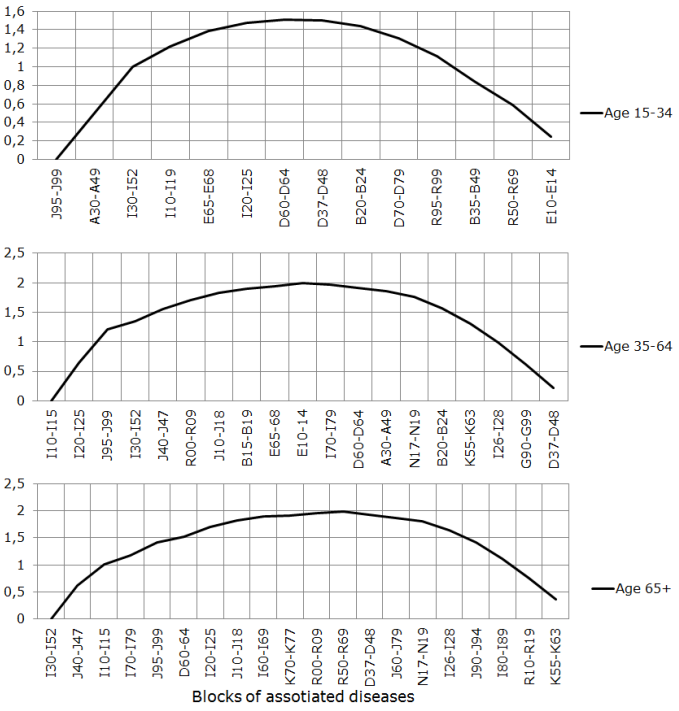


Fig. 2: Lower bound of average risk

Each curve of the lower bound of the average risk reaches its maximum on some set of blocks of associated diseases and

determines diseases for which the distribution of associated diseases in the cancer group maximally differs from the distribution of associated diseases in the non cancer group. Using these curves we calculate the portion of people died of cancer among all who had particular associated disease.

Obtained sets of blocks of associated diseases (for which cancer group maximally differs from non cancer group) and the portions of people died of cancer in these sets are compared for the different age groups and years of death.

First we observe the *youngest age group* (15-34 years old) in which among the others important blocks we have the bacterial diseases block. This block was selected in each performed analysis over the years from 1985 till 2008. Each 5th person who had bacterial disease and died at the age 15-34 years old died of cancer. However, it looks like this pattern has no medical interpretation and it can be explained by the widespread of bacterial diseases in the youngest age group.

The second block of associated diseases selected in the youngest age group is aplastic and other anaemias. The portion of people who had these diseases and died of cancer increased from 20% in 1985 year to 45% in 2008 year. It is a common fact that anaemia often is the first symptom of leukemia, so a cause-effect connection between these two diseases definitely exists.

Now proceeding to the *middle age group* (35-64 years old). In this group pneumonia and influenza are very "important" diseases in terms of difference between cancer and non cancer group. This block of associated diseases is selected in all considered years. The approximate portion of people who had these diseases and died of cancer is 30%, which means that each 3rd-4th person who had pneumonia or influenza at the end of his life died of cancer. The relationship between pneumonia and cancer is commonly known: on the one hand lung cancer can develop on the base of chronic inflammation caused by repeated pneumonia [8], [7]; on the other pneumonia can appear because of week immune system provoked by cancer or a treatment of cancer.

Aplastic and other anaemias were widely spread among people who died of cancer at 35-64 years old between 1985 and 1995 years. The portion of people who died of cancer and had anaemias was about 40% in 1985 and 1990 years and about 30% in 1995 year, but this block of associated diseases wasn't selected by the method of maximizing of average risk in this age group after 1995. We also can observe the proliferation of aplastic and other anaemias in the younger age groups after 1999.

Next selected associated disease connected with cancer death for middle age group is Viral hepatitis (B15-B19), which is the well-known factor of cancer risk. The portion of people who died of cancer at the age 35-64 years old is 35% in 2008 year.

Finally, in the *eldest age group* (65+) we can observe the aplastic and other anaemias are becoming significant in terms of differences between cancer and non cancer groups. About 30% of people who had a disease from this block died of cancer.

Diseases of liver are selected in the eldest group in all considered years. Approximately 20% of people in this group who had a disease of liver died of cancer and after 1999 year the portion increased to 30%. Cancer of liver can develop on the basis of inflammation provoked by these diseases.

These diseases are not the only diseases on which the functional of average risk reaches its maximum. Nevertheless, only them are selected in each considered year from 1985 to 2008 and the portions of people who had these diseases and died of cancer are more than 20%.

V. Conclusion

This paper is devoted to the problem of investigation of links between risk of cancer death and associated morbidity. It is mathematically formalized as the problem of contrasting the distributions of associated diseases among people died of cancer and among people died of another disease. To solve this problem we evaluate the average risk on the empirical data using the Vapnik-Chervonenkis inequalities. We select associated diseases for which distribution in cancer group maximally differs from the distribution in non cancer group. Three age groups and five years of death are analyzed separately.

Results for different years and ages are compared, common blocks of associated diseases are emphasized for different years of death. Thus anaemias are connected with cancer death for all considered years and age groups. Pneumonia is selected as linked with cancer only in middle age group in all analyzed years. Relationships between diseases of liver and cancer are obtained for eldest age group and all years of death. The performed connections are interpreted from the medical point of view.

The aim of the future investigations is consideration of the more "tiny" links between cancer mortality and associated diseases, division cancer incidence by different types of neoplasms. For such analysis one should use more precise estimation for the average risk than estimation based on the Vapnik-Chervonenkis approach.

References

[1] Vapnik V. (1998). *Statitical Learning Theory*. Wiley Interscience.

[2] Blagosklonny M.V. (2010). Why human lifespan is rapidly increasing: solving "longevity riddle" with "revealed-slow-aging" hypothesis. *Aging*. Vol. **4**, pp. 177-182.

[3] Michalski A.I., Ukraintseva S.V.,Arbeev K.G., Yashin A.I. (2005). Investigation of old age mortality structure in the presence of comorbidity. *In European Conference On Chronic Disease Prevention. Helsinki, Finland*, p. 60.

[4] Yashin A.I. et al. (2001). Have the oldest old adults ever been frail in the past? A hypothesis that explains modern trends in survival. *Gerontol. Biol. Sci.*. Vol. **56**, pp. 432-442.

[5] Global status report on noncommunicable diseases 2010, World Health Organization

[6] Mortality Data, Multiple Cause-of-Death Public-Use Data Files. http://www.cdc.gov/nchs/products/elec_prods/subject/mortmcd.htm

[7] Azad N., Rojanasakul Y., Vallyathan V. (2008). Inflammation and Lung Cancer: Roles of Reactive Oxygen/Nitrogen Species. *Journal of Toxicology and Environmental Health, Part B*. Vol. **11**, pp. 115.

[8] Mayne S., Buenconsejo J., Janerich D (1999). Previous Lung Disease and Risk of Lung Cancer among Men and Women Nonsmokers. *Epidemiol.*. Vol. **149**, pp. 13-20.

On Goodness of Fit Tests for Grouped Survival and Reliability Data

Ilia Vonta
Department of Mathematics
National Technical University of Athens
Athens, Greece
Email: vonta@math.ntua.gr

Alex Karagrigoriou
Department of Mathematics and Statistics
University of Cyprus
Nicosia, Cyprus
Email: alex@ucy.ac.cy

***Abstract*—Measures of divergence or discrepancy are used extensively in statistics in various fields. In this paper we are focusing on divergence measures that are based on a class of measures known as Csiszar's divergence measures. In particular, we propose a class of goodness of fit tests based on Csiszar's class of measures designed for censored survival or reliability data. Further, we derive the asymptotic distribution of the test statistic under the null hypothesis and under contiguous alternative hypotheses. We examine the case where the distribution under the null hypothesis is completely known, as well as the case where unknown parameters are involved in the assumed distribution. Simulations were conducted to show the performance of the proposed test.**

I. Introduction

Measures of divergence or discrepancy are used extensively in statistics. In this paper we are focusing on divergence measures that are based on a class of measures known as Csiszar's divergence measures (Csiszar (1963)). This class of measures is generated by a collection of functions φ with properties:

1. $\varphi(x)$ is continuous, differentiable and convex for $x \geq 0$
2. $\varphi(1) = 0$
3. $\varphi'(1) = 0$.

Measures of divergence between two probability distributions have a very long history initiated by the pioneer work of Pearson, Mahalanobis, Lévy and Kolmogorov. Among the most popular measures of divergence are the Kullback-Leibler measure of divergence (Kullback and Leibler, 1951) and the Csiszar's φ-divergence family of measures (Csiszár, 1963; Ali and Silvey, 1966). A unified analysis has been provided by Cressie and Read (1984) who introduced the power divergence family of statistics that depends on a parameter a and is used for goodness-of-fit tests for multinomial distributions. The Cressie and Read family includes among others the well known Pearson's X^2 divergence measure and for multinomial models the loglikelihood ratio statistic. Recently, the BHHJ divergence measure was proposed by Basu et *al.* (1998) and generalized to the BHHJ family of measures by Mattheou et *al.* (2009). The BHHJ family depends on an index a which controls the trade-off between robustness and efficiency when the measure is used as an estimating criterion for robust parameter estimation.

In Statistics, the problem of determining the appropriate distribution or the appropriate model for a given data set is extremely important for reducing the possibility of erroneous inference. Additional issues are raised in biomedicine and biostatistics. Indeed, the existence of censoring schemes in survival modelling makes the determination of the proper distribution or model an extremely challenging problem. An important aspect is how to check validity of a specific model assumption. Some research has been done in this regard. For example, Gail and Ware (1979) studied grouped censored survival data by comparing with a known survival distribution, while Akritas (1988) constructed a Pearson-type goodness-of-fit measure for one-sample data that allows for random censorship and Chen, et. al (2004) proposed a test of fit based on the Cressie and Read power divergence measure.

In the present work we focus on hypothesis testing and in particular on a class of goodness of fit tests based on Csiszar's family of measures and propose a general family of test statistics for treating the case of censored data with applications that can be extended from survival analysis to reliability theory. First we present the Csiszar's class of divergence measures, we formulate the problem and propose a class of test statistics based on the Csiszar's class of functions, for testing a null hypothesis about the true distribution function of lifetimes subject to right censoring. Then, we present theoretic results about the asymptotic distribution of the test statistic under the null hypothesis and under contiguous alternative hypotheses. Note that both the simple and the composite null hypothesis where the proposed distribution depends on a finite dimensional unknown parameter, are treated. Finally, we provide simulations and a real data example that illustrate the performance of the proposed test statistic.

II. The Proposed Test Statistics

Let X be an absolutely continuous, non-negative random variable that describes the lifetime of an individual or an item. Let $f(x)$, $F(x)$ and $h(x)$ be the density function, the cumulative distribution function and the hazard intensity function of X respectively. If we would like to test the null hypothesis that the distribution function of X is equal to F_0, we could utilize divergence measures for this purpose.

For example, the Kullback-Leibler distance between F and F_0 (Kullback and Leibler, 1951), is defined by

$$I^{KL}_{F,F_0} = \int_0^\infty f(x) \log\left(\frac{f(x)}{f_0(x)}\right) dx \tag{1}$$

where $\log$ denotes the natural logarithm and f_0 is the density under the null hypothesis. Without loss of generality we assume that the support of f and f_0 is $(0, +\infty)$.

A generalization of this distance is defined as (Csiszar, 1963)

$$I^{\varphi}_{F,F_0} = \int_0^\infty f_0(x)\varphi\left(\frac{f(x)}{f_0(x)}\right) dx \tag{2}$$

and is known as Csiszar's family of measures of divergence, for some function φ. Different $\varphi(\cdot)$ function give rise to different measures including the Kullback-Leibler measure and the Cressie and Read family of power divergences.

The proposed class of goodness of fit tests is based on Csiszar's class of measures which is indexed by the function φ and is designed for censored survival or reliability data which we treat as grouped data. The formulation is the following: suppose that we observe subjects during the time interval $[0, \tau]$ which is partitioned into k subintervals of the form $(\tau_{i-1}, \tau_i]$, $i = 1, \ldots, k$ with

$$0 = \tau_0 < \tau_1 < ... < \tau_k = \tau.$$

Let us assume also that censoring occurs only at τ_i. Let n_i be the number of subjects at risk at the beginning of the i^{th} interval, d_i the number of failures during the i^{th} interval and h_i represents the hazard rate in the i^{th} interval.

The test statistic for a simple null hypothesis under which the null distribution is completely known, is defined as

$$D^{\varphi}(\mathbf{d},\mathbf{n},\mathbf{h}) = \frac{2}{\varphi''(1)} \sum_{i=1}^{k} \left\{ n_i h_i \varphi\left(\frac{d_i}{n_i h_i}\right) + n_i h_i^c \varphi\left(\frac{d_i^c}{n_i h_i^c}\right) \right\} \tag{3}$$

where $\mathbf{d} = (d_1, \ldots, d_k)$, $\mathbf{n} = (n_1, \ldots, n_k)$, $\mathbf{h} = (h_1, \ldots, h_k)$, $d_i^c = n_i - d_i$ and $h_i^c = 1 - h_i$. This test statistic is an extension of the one proposed in Chen, Lai and Ying (2004) for the Cressie and Read function φ.

Although a simple null hypothesis suggested above appears frequently in practice, it is quite common to test the composite null hypothesis that the unknown distribution belongs to a parametric family $\{F_\theta\}_{\theta\in\Theta}$, where Θ is an open subset in R^m. In this case we can again consider a partition of the original sample space with k disjoint intervals. In such a case, the above general class of test statistics takes the form

$$D^{\varphi}(\mathbf{d},\mathbf{n},\mathbf{h}(\theta)) = \frac{2}{\varphi''(1)} \sum_{i=1}^{k} n_i \Big\{ h_i(\theta)\varphi\left(\frac{d_i/n_i}{h_i(\theta)}\right) + h_i^c(\theta)\varphi\left(\frac{d_i^c/n_i}{h_i^c(\theta)}\right) \Big\}. \tag{4}$$

Let

$$\mathcal{H} = \{\mathbf{h}(\theta) = (h_1(\theta), \ldots, h_k(\theta)) \ : \ \theta \in \Theta\}$$

and assume that $h_i(\theta)$ are twice continuously differentiable in θ. Let finally $\theta_{\mathbf{0}}$ denote the true parameter and $\mathbf{h}_0 = \mathbf{h}(\theta_{\mathbf{0}})$.

Observe that in the composite case, the probabilities of failure in the various intervals depend on the unknown $m-$dimensional parameter θ so that a consistent estimator $\hat{\theta}$ of θ is required. In regard to the estimating method applied for obtaining such an estimator, the traditional maximum likelihood estimator (MLE), under the null distribution, can be evaluated and implemented. Note though that one may alternatively consider the wider class of $\varphi-$divergence estimators. More specifically, for the partition $\{E_i\}_{i=1,\ldots,k}$ of the original sample space, the minimum $\varphi-$divergence estimator of θ under the null hypothesis

$$H_0 : F = F_0(\theta)$$

or equivalently

$$H_0 : \mathbf{h} = \mathbf{h}_0(\theta) = (h_{10}(\theta), \ldots, h_{k0}(\theta))'$$

is any $\hat{\theta}_\varphi \in \Theta$ satisfying

$$\hat{\theta}_\varphi = \arg\min_{\theta\in\Theta} \sum_{i=1}^{k} n_i \Big\{ h_{i0}(\theta)\varphi\left(\frac{d_i/n_i}{h_{i0}(\theta)}\right) + h_{i0}^c(\theta)\varphi\left(\frac{d_i^c/n_i}{h_{i0}^c(\theta)}\right) \Big\}. \tag{5}$$

Obviously, the resulting estimator depends on the φ-function chosen. Observe that for φ having the special form given by

$$\varphi(u) = 1 - (1 + \frac{1}{a})u + \frac{u^{1+a}}{a}, \quad a \neq 0.$$

and for $a \to 0$ the resulting estimator is the usual maximum likelihood estimator, for the grouped data.

For the asymptotic distribution under the alternative hypothesis we consider the contiguous alternative

$$H_0 : h_i = h_{i0} \ vs. \ H_1 : h_i = h_{ib}, \ i = 1, \ldots, k.$$

Suppose that the null hypothesis indicates that $h_i = h_{i0}$, $i = 1, 2, \ldots, k$ when in fact it is $h_i = h_{ib}$, $\forall i$. As it is well known if h_{i0} and h_{ib} are fixed then as the sample size tends to infinity the power of the test tends to 1. In order to examine the situation when the power is not close to 1, we must make it continually harder for the test as the sample size increases. This can be done by allowing the alternative hypothesis steadily closer to the null hypothesis. As a result we define a sequence of alternative hypotheses as follows

$$H_{1,N} : h_i = h_{iN} = h_{i0} + \beta_i \sqrt{n_i}, \ \forall i \tag{6}$$

where β_i are constants, which is known as Pitman transition alternative or Pitman (local) alternative or local contiguous alternative to the null hypothesis $H_0 : h_i = h_{i0}$. In vector notation the null hypothesis and the local contiguous alternative hypotheses take the form

$$H_0 : \mathbf{h} = \mathbf{h}_0 \ vs. \ H_{1,N} : \mathbf{h} = \mathbf{h}_N = \mathbf{h}_0 + \beta,$$

where $\mathbf{h}_N = (h_{1N}, h_{2N}, \ldots, h_{kN})'$ and $\beta = (\beta_1/\sqrt{n_1}, \ldots, \beta_k/\sqrt{n_k})'$ is a fixed vector such that

$\sum_{i=1}^{k} h_{i0} + \beta_i/\sqrt{n_i} \leq k$. Observe that as n_i tends to infinity $\forall i$, the local contiguous alternative converges to the null hypothesis at the rate $O(N^{-1/2})$.

In order to derive the asymptotic distribution of the test statistic (4) under the local contiguous alternatives $H_{1,N}$, we first obtain the asymptotic distribution of $\hat{p}_i \equiv d_i/n_i$.

For all the above cases we derive the asymptotic distribution of the proposed test statistics under the null hypothesis and under contiguous alternative hypotheses. We examine the case where the distribution under the null hypothesis is completely known, as well as the case where unknown parameters are involved in the assumed distribution.

III. Simulations and Real Data

Some of the most popular continuous distributions in biomedicine, engineering and reliability are the exponential, lognormal, Gamma, Inverse Gaussian, Weibull, Pareto, and Positive Stable distributions. The family of the two-parameter inverse Gaussian distribution is one of the basic models for describing positively skewed data which arise in a variety of fields of applied research as cardiology, hydrology, demography, linguistics, employment service, etc (see for instance Chhikara and Folks, 1977). Furthermore, distributions like the Weibull, the Positive Stable and the Pareto are frequently encountered in survival modelling. Finally, distributions like the exponential, the Gamma, the lognormal and others are very common in lifetime problems.

In order to assess the applicability and performance of the proposed tests of fit we will present a number of simulation studies. More specifically, first we test

$$H_0 : F(t) = 1 - \exp(-t)$$

vs.

$$H_1 : F(t) = 1 - \exp(-t/\gamma)$$

or equivalently

$$H_0 : h(t) = 1 \text{ vs. } H_1 : h(t) = 1/\gamma,$$

that is, we test under the null hypothesis the standard exponential model against exponential models with rate $1/\gamma$ with

$$\gamma = 1/(1 + b/\sqrt{n})$$

(see Akritas, 1988) with $b = -4, -3, -2, -1, 0, 1, 2, 3,$ and 4. Observe that for $b = 0$ the data come from the exponential null model so that the values appearing in the tables refer to the size of the test. We also use the test

$$H_0 : F(t) = 1 - \exp(-t)$$

vs.

$$H_1 : F(t) = 1 - \exp(t^{\gamma}).$$

Observe that in this case, the alternative is the Weibull distribution with shape parameter γ and scale parameter 1, denoted by $Weibull(1, \gamma)$. Note that equivalently, we can test the corresponding hazard or survival functions. In this case we choose

$$\gamma = 1/(1 + b/\sqrt{n}),$$

with $b = -4(1)4$. Observe again that for $b = 0$ the data come from the exponential null model so that the value $b = 0$ is associated with the size of the test.

For each test we simulate 10000 samples of various sample sizes ranging from $n = 10$ to $n = 500$. Censored observations are generated assuming that the hazard of the censoring distribution is proportional to the hazard of the failure distribution. The censoring proportions considered in the simulations range from 10% to 60% although selective results are presented here due to space constraints. For the same reasons we only present the results for sample sizes $n = 20, 50$ and $n = 100$.

The size of the tests was taken 5% although the size 1% was also examined but results are not given here. Among the members of the proposed family of tests we present the ones associated with the functions Kullback-Leibler, Pearson, Matusita, Cressie and Read and φ_2 given by

$$\varphi(u) \doteq \varphi_2(u) = u^{1+a} - (1 + \frac{1}{a})u^a + \frac{1}{a}, \quad a \neq 0.$$

For the C-R test we used the optimum value of

$$a = 2/3$$

(Cressie and Read (1984)) and for the φ_2 function we used the value

$$a = 10/9$$

which is optimal (Vonta & Tsanousa (2010)) for this function. See also Vonta & Chouchoumis (2010) for related issues.

Read applications with data sets investigated by Susarla and Van Ryzin (1978) and Nikulin and Haghighi (2004) have been used to test the performance of the proposed techniques in real situations.

References

[1] Akritas, M. G. (1988). Pearson-type goodness-of-fit tests: the univariate case. *J. Amer. Statist. Assoc.*, 83, 222-230.

[2] Ali, S. M. and Silvey, S. D. (1966). A general class of coefficients of divergence of one distribution from another, *J. Roy. Statist. Soc. B*, 28, 131-142.

[3] Basu, A., Harris, I. R., Hjort, N. L. and Jones, M. C. (1998). Robust and efficient estimation by minimising a density power divergence, *Biometrika*, *85*, 549-559.

[4] Cavanaugh, J. E. (2004). Criteria for linear model selection based on Kullback's symmetric divergence. *Australian and New Zealand Journal of Statistics* 46, 257-274.

[5] Chen H.S., Lai, K. and Ying, Z. (2004) Goodness of fit tests and minimum power divergence estimators for survival data, *Statistica Sinica*, 14: 231–248.

[6] Chhikara, R. S., Folks, J. L. (1977). The inverse Gaussian distribution as a lifetime model. *Technometrics*, 19, 461–468.

[7] Cox, D. R. (1972). Regression models and life tables, *J. Roy. Statist. Soc.*, *B 34*, 187-202.

[8] Cressie, N. and Read, T. R. C. (1984). Multinomial goodness-of-fit tests, *J. R. Statist. Soc*, *5*, 440-454.

[9] Csiszar, I. (1963). Eine Informationstheoretische Ungleichung und ihre Anwendung auf den Bewis der Ergodizitat on Markhoffschen Ketten, *Publ. of the Math. Inst. of the Hungarian Academy of Sc.*, *8*, 84–108.

[10] Gail, M. H. and Ware, J. H. (1979). Comparing observed life table data with a known survival curve in the presence of random censorship. *Biometrics*, 35, 385-391.

[11] Kagan, A. M. (1963). On the theory of Fisher's amount of information, *Sov. Math. Dokl.*, 4: 991–993.

[12] Kullback, S. & Leibler, R. (1951). On information and sufficiency, *Annals of Math. Statist.*, 22: 79–86.

[13] Mattheou, K., Lee, S., and Karagrigoriou, A. (2009). A model selection criterion based on the BHHJ measure of divergence, *J. of Statist. Plan. and Infer.*, 139, 128-135.
[14] Matusita, K. (1967). On the notion of affinity of several distributions and some of its applications, *Ann. Inst. Statist. Math.*, 19, 181–192.
[15] Nikulin, M.S. and Haghighi, F, (2004). A chi-squared test for the generalized power Weibull family for the head-and-neck cancer censored data, *J. Math. Sc.*, 133 (3), 1333-1341.
[16] Susarla, V. and van Ryzin, J. (1978). Empirical bayes estimation survival distribution function from right censored data , Anal. Statist., 6, 740-755.
[17] Vonta, F. and Chouchoumis I. (2010) *φ-measures of divergence and their applications to survivalk analysis and reliability*, Master thesis, National Technical University of Athens.
[18] Vonta F. and Tsanousa A. (2010). *Hypothesis testing via φ-measures of divergence*, Master thesis, National Technical University of Athens.

Nonparametric estimation of the rate of occurrence of failures for semi-Markov chains

Irene Votsi*,†, Nikolaos Limnios* and George Tsaklidis†
*Laboratoire de Mathématiques Appliquées de Compiègne
Université de Technologie de Compiègne, Compiègne, France 60205
Email: irene.votsi@utc.fr, nikolaos.limnios@utc.fr
† Department of Mathematics
Aristotle University of Thessaloniki, Thessaloniki, Greece 54124
Email: tsaklidi@math.auth.gr

***Abstract*—The problem of estimating the rate of occurrence of failures in semi-Markov chains is addressed for the first time. Firstly a simple formula for evaluating the rate of occurrence of failures for a semi-Markov chain is derived. As a consequence of this result, a statistical estimator of this function is proposed. The asymptotic properties of this estimator are studied, including the uniform strong consistency and the asymptotic normality.**

***Index Terms*—Semi-Markov chain; failure occurrence rate; asymptotic properties.**

I. Introduction

Consider a system S whose possible states during its evolution in time belong to $E = \{1, 2, ..., s\}$. Denote by $U = \{1, ..., r\}$ the subset of operational states of the system (the up states) and by $D = \{r + 1, ..., s\}$ the subset of failure states (the down states), with $0 < r < s$ (obviously, $E = U \bigcup D$ and $U \bigcap D = \emptyset$, $U \neq \emptyset$, $D \neq \emptyset$). One can think of the states of U as different operating modes or performance levels of the system, whereas the states of D can be seen as failures of the systems with different modes. The quantity that is studied here is the function $r(k)$ which is called the rate of occurrence of failures (ROCOF). At this point we have to refer that we examine the discrete-time case, whereas the continuous-time case was studied by Ouhbi and Limnios (2002).

II. Semi-Markov setting

In this section we define the discrete-time semi-Markov model and introduce the basic notation and definitions (Barbu and Limnios, 2008). Let $\mathbb{N}$ be the set of nonnegative integers. Consider a random system with finite state space E. We suppose that the evolution in time of the system is described by the following chains:

1) The chain $J = (J_n)_{n\in\mathbb{N}}$ with state space E, where J_n is the system state at the $n-$th jump time;
2) The chain $S = (S_n)_{n\in\mathbb{N}}$ with state space $\mathbb{N}$, where S_n is the $n-$th jump time. We suppose that $S_0 = 0$ and $0 < S_1 < S_2 < ... < S_n < S_{n+1} < ...$;
3) The chain $X = (X_n)_{n\in\mathbb{N}}$ with state space $\mathbb{N}$, $X_n := S_n - S_{n-1}$ for all $n \in \mathbb{N}^*$ and $X_0 := 0$ a.s. Thus for all $n \in \mathbb{N}^*$, X_n is the sojourn time in state J_{n-1}, before the $n-$th jump.

Definition 1. *The chain $(J_n, S_n)_{n\in\mathbb{N}}$ is said to be a Markov renewal chain (MRC) if for all $n \in \mathbb{N}$, for all $i, j \in E$, and for all $k \in \mathbb{N}$ it satisfies almost surely (a.s.)*

$$P(J_{n+1} = j, S_{n+1} - S_n = k|J_0, ..., J_n; S_0, ..., S_n) = P(J_{n+1} = j, S_{n+1} - S_n = k|J_n). \quad (1)$$

Moreover, if equation (1) is independent of n, then (J, S) is said to be homogeneous and the discrete-time semi-Markov kernel $\mathbf{q}$ is defined by

$$q_{ij}(k) := P(J_{n+1} = j, X_{n+1} = k|J_n = i),$$

where $i, j \in E$ and $k, n \in \mathbb{N}$. The process $(J_n)_{n\in\mathbb{N}}$ is the embedded Markov chain (EMC) of

the Markov renewal chain (MRC) (J_n, S_n) with transition kernel $P = (p_{ij})$. The semi-Markov kernel $\mathbf{q}$ is written as

$$q_{ij}(k) = p_{ij} f_{ij}(k),$$

where $p_{ij} := P(J_{n+1} = j | J_n = i)$ $i, j \in E$, $n \in \mathbb{N}$, are the transition probabilities of the EMC and $f_{ij}(k) := P(X_{n+1} = k | J_n = i, J_{n+1} = j)$ $i, j \in E$, $k \in \mathbb{N}$, is the conditional distribution of the sojourn time in the state i given that the next visited state is j. In this case the sojourn times are attached to transitions and when a sojourn time in a state i expires, we can determine the next visited state j by using the probability of the EMC as well as the duration of this time.

The initial distribution of the EMC a is written as $a(i) = P(J_0 = i)$, $i \in E$. Moreover, for all states $i, j \in E$, let us define:

1) $N_i(M) := \sum_{n=1}^{N(M)} \mathbf{1}_{\{J_{n-1}=i\}}$ is the number of visits to state i of the EMC, up to time M;
2) $N_{ij}(M) := \sum_{n=1}^{N(M)} \mathbf{1}_{\{J_{n-1}=i, J_n=j\}}$ is the number of transitions of the EMC from i to j, up to time M.

Definition 2. *Let (J, S) be a Markov renewal chain. The chain $Z = (Z_k)_{k\in\mathbb{N}}$ is said to be a semi-Markov chain associated to the MRC (J, S) if*

$$Z_k := J_{N(k)},\ k \in \mathbb{N},$$

where $N(k) := \max\{n \in \mathbb{N} | S_n \leq k\}$ is the discrete-time counting process of the number of jumps in $[1, k] \subset \mathbb{N}$. Thus Z_k gives the system state at time k. We have also $J_n = Z_{S_n}$ and $S_n = \min\{k > S_{n-1} | Z_k \neq Z_{k-1}\}$, $n \in \mathbb{N}$.

Let $U = (U_k)_{k\in\mathbb{N}}$ be the sequence of backward recurrence times of the SMC $(Z_k)_{k\in\mathbb{N}}$ defined as follows: $U_k = k - S_{N(k)}$. Let also $\overline{H}_i(\cdot)$ be the survival function of the sojourn time in state i defined by

$$\begin{aligned}\overline{H}_i(k) &:= P(S_{l+1} - S_l > k | J_l = i) \\ &= 1 - \sum_{j\in E} \sum_{n=0}^{k} q_{ij}(n),\end{aligned}$$

where $k \in \mathbb{N}$, $l \in \mathbb{N}^*$. It can been shown that the stochastic process $(Z, U) := (Z_k, U_k)_{k\in\mathbb{N}}$ is a Markov chain (Limnios and Oprişan, 2001). The transition probabilities of the Markov chain (Z_k, U_k) can be written as (Chryssaphinou et al., 2008):

$$\begin{aligned}&P(Z_{k+1} = j, U_{k+1} = t_2 | Z_k = i, U_k = t_1) \\ &= \begin{cases} q_{ij}(t_1+1)/\overline{H}_i(t_1) & \text{if } i \neq j,\ t_2 = 0 \\ \overline{H}_i(t_1+1)/\overline{H}_i(t_1) & \text{if } i = j,\ t_2 - t_1 = 1 \\ 0 & \text{elsewhere} \end{cases},\end{aligned}$$

for every $(i, t_1), (j, t_2) \in E \times \mathbb{N}$ and for every $k \in \mathbb{N}$ such that $P(Z_k = i, U_k = t_1) > 0$. It is worth noticing that the Markov chain $(Z_k, U_k)_{k\in\mathbb{N}}$ is time-homogeneous, so that for all $(i, t_1), (j, t_2) \in E \times \mathbb{N}$ we denote its transition probabilities by

$$\begin{aligned}&\widetilde{P}\big((i, t_1), (j, t_2)\big) \\ &:= P(Z_{k+1} = j, U_{k+1} = t_2 | Z_k = i, U_k = t_1),\end{aligned}$$

$\forall k \in \mathbb{N}$. The probability that the Markov chain $(Z_k, U_k)_{k\in\mathbb{N}}$ visits the state (i, m) is denoted by $\overline{P}_{im} = P(Z_k = i, U_k = m)$ for all $(i, m) \in E \times \mathbb{N}$, $k \in \mathbb{N}$. For a stochastic system with state space E described by a SMC $(Z_k)_{k\in\mathbb{N}}$, let us consider a partition U, D of E, i.e., $E = U \cup D$, with $U \cap D = \emptyset$, $U \neq \emptyset$, and $D \neq \emptyset$. The set U contains the up states and D contains the down states of the system. Let us further denote by $(\widetilde{a}\widetilde{P}^{k-1})(i, m)$ the element (i, m) of the vector $\widetilde{a}\widetilde{P}^{k-1}$ and by $(\widehat{\widetilde{a}}\widehat{\widetilde{P}}_M^{k-1})(i, m)$ the element (i, m) of the vector $\widehat{\widetilde{a}}\widehat{\widetilde{P}}_M^{k-1}$.

III. ROCOF of the Semi-Markov Systems

The following theorem gives a simple formula of the ROCOF of semi-Markov systems.

Theorem 1. *The ROCOF of the semi-Markov system at time k is given by*

$$\begin{aligned}\widetilde{r}(k) = \sum_{i\in U} \sum_{j\in D} \sum_{m=0}^{k-1} \sum_{l\in\{A_{k;m}\}} &[(\widetilde{a}\widetilde{P}^{k-1})(i, m)] \\ &\times \widetilde{P}\big((i, m), (j, l)\big),\end{aligned}$$

where $A_{k;m} = \{0, (m+1) \wedge k\}$.

IV. Statistical Estimation of the ROCOF of a Semi-Markov System

A natural estimator for ROCOF is

$$\widehat{\widetilde{r}}(k) = E[N_f(k) - N_f(k-1)],$$

which is based on the estimator of the transition probability matrix of the Markov chain $(Z_k, U_k)_{k\in\mathbb{N}}$.

Let us consider a sample path of an ergodic Markov renewal chain $(J_n, S_n)_{n\in\mathbb{N}}$, censored at fixed arbitrary time $M \in \mathbb{N}$,

$$H(M) := (J_0, X_1, ..., J_{N(M)-1}, X_{N(M)}, J_{N(M)}, u_M),$$

where $N(M)$ is the discrete-time counting process of the number of jumps in $[1, M]$, and $u_M := M - S_{N(M)}$ is the censored sojourn time in the last visited state $J_{N(M)}$.

On the basis of Theorem 1, we propose the following estimator for the ROCOF of the semi-Markov system,

$$\widehat{\widetilde{r}}(k,M) = \sum_{i\in U}\sum_{j\in D}\sum_{m=0}^{k-1}\sum_{l\in A_{k;m}} [(\widehat{\widetilde{a}}\widehat{\widetilde{P}}_M^{k-1})(i,m)] \times \widehat{\widetilde{P}}_M\big((i,m),(j,l)\big),$$

where $A_{k;m} = \{0, (m+1)\wedge k\}$.

The empirical estimator of the probability that the Markov chain (Z_k, U_k) visits the state (i, m) is defined by $\widehat{\overline{P}}_{im}(M) := \frac{N_{(i,m)}(M)}{N(M)}$ and the corresponding estimator of the transition probability matrix is defined by:

$$\widehat{\widetilde{P}}_M\big((i,t_1),(j,t_2)\big) = \begin{cases} \widehat{q}_{ij}(t_1+1,M)/\widehat{\overline{H}}_i(t_1,M) & \text{if } i\neq j,\ t_2=0 \\ \widehat{\overline{H}}_i(t_1+1,M)/\widehat{\overline{H}}_i(t_1,M) & \text{if } i=j,\ t'=1, \\ 0 & \text{elsewhere} \end{cases}$$

where $t' = t_2 - t_1$.

Moreover, the empirical estimator of the semi-Markov kernel is given by

$$\widehat{q}_{ij}(k,M) := \frac{1}{N_i(M)}\sum_{k=1}^{N(M)} \mathbf{1}_{\{J_{k-1}=i, J_k=j, X_k\le x\}}$$

along with the empirical estimator of the survival function in state $i \in E$ given by

$$\widehat{\overline{H}}_i(k,M) = 1 - \sum_{j\in E}\sum_{n=0}^{k}\widehat{q}_{ij}(k,M),\ k\in\mathbb{N}.$$

The following theorem gives the uniform strong consistency of the estimator of the ROCOF.

Theorem 2. *For any fixed arbitrary positive integer $k \in \mathbb{N}$, the estimator of the rate of a discrete-time system at instant k is strongly consistent in the sense that*

$$\widehat{\widetilde{r}}(k,M) \xrightarrow[M\to\infty]{a.s.} \widetilde{r}(k).$$

Theorem 3. *(see, e.g, Billingsley) In the stationary, ergodic case, the distribution of the s^2- dimensional random vector $\xi = (\xi_{((i,t_1),(j,t_2))})_{(i,t_1),(j,t_2)\in E\times\mathbb{N}}$, where*

$$\xi_{((i,t_1),(j,t_2))} = \frac{1}{\sqrt{N_{(i,t_1)}(M)}}\Big(N_{((i,t_1),(j,t_2))}(M) - N_{(i,t_1)}(M)\widetilde{P}\big((i,t_1),(j,t_2)\big)\Big),$$

converges, as M tends to infinity, to the normal distribution centered at the origin with covariance matrix

$$\lambda\Big(\big((i,t_1),(j,t_2)\big),\big((l,t_3),(r,t_4)\big)\Big),$$

where

$$\lambda\Big(\big((i,t_1),(j,t_2)\big),\big((l,t_3),(r,t_4)\big)\Big) = \delta_{(i,t_1)(l,t_3)}\Big(\delta_{(j,t_2)(r,t_4)}\widetilde{P}\big((i,t_1),(j,t_2)\big) - \widetilde{P}\big((i,t_1),(j,t_2)\big)\widetilde{P}\big((i,t_1),(r,t_4)\big)\Big)$$

$(i,t_1),(j,t_2),(l,t_3),(r,t_4) \in E\times\mathbb{N}$ and $\delta_{(i)(j)}$ is the Kronecker symbol, i.e, $\delta_{(i)(j)} = 1$ if $(i) = (j)$ and $\delta_{(i)(j)} = 0$ if $(i)\neq(j)$.

Proposition 1. *Let $\big(Z_k, U_k\big)_{k\in\mathbb{N}}$ be an homogeneous ergodic Markov chain. The random vector $F = (f_{(i,t_1)(j,t_2)})_{(i,t_1),(j,t_2)\in E\times\mathbb{N}}$, where*

$$f_{(i,t_1)(j,t_2)} = \sqrt{M}\Big(\widehat{\widetilde{P}}_M\big((i,t_1),(j,t_2)\big) - \widetilde{P}\big((i,t_1),(j,t_2)\big)\Big)$$

converges, as M tends to infinity, to the normal distribution $Z\sim\mathcal{N}(0,\Gamma)$, where Γ is a convariance matrix, of dimension $s^2\times s^2$, which defined with block diagonal form as follows

$$\Gamma = \begin{pmatrix} \frac{1}{\pi_1}\Lambda_{(i,t_1)_1} & \underline{0} & \cdots & \underline{0} \\ \underline{0} & \frac{1}{\pi_2}\Lambda_{(i,t_1)_2} & \cdots & \underline{0} \\ \vdots & \vdots & \ddots & \vdots \\ \underline{0} & \underline{0} & \cdots & \frac{1}{\pi_s}\Lambda_{(i,t_1)_s} \end{pmatrix}.$$

Theorem 4. *For any* $k \geq 1$, $\sqrt{M}\big(\widehat{\widetilde{r}}(k,M) - \widetilde{r}(k)\big)$ *converges in distribution, as* M *tends to infinity, to a zero mean normal random variable with variance* $\Phi' \Gamma_1 \Phi'^T$, *where* Γ_1 *is the limit of* $s^2 \times s^2$ *covariance matrix for the random vector* $\sqrt{M}\Big[\widehat{\widetilde{P}}_M\big((i,m),(j,l)\big) - \widetilde{P}\big((i,m),(j,l)\big)\Big]$, $(i,m) \in U \times \mathbb{N}$, $(j,l) \in D \times \mathbb{N}$ *and*

$$\Phi' = \Big(\frac{\partial \Phi}{\partial \widetilde{P}\big((i,m),(j,l)\big)}\Big)_{(i,m)\in U\times\mathbb{N},(j,l)\in D\times\mathbb{N}}.$$

A. Example

Let us consider a three state Markov chain $(J_n)_{n\in\mathbb{N}}$ with state space $E = \{1,2,3\}$ and transition probability matrix

$$P = \begin{pmatrix} 0 & 1 & 0 \\ 0.95 & 0 & 0.05 \\ 1 & 0 & 0 \end{pmatrix}$$

and let us suppose that the initial distribution α is known. We generate a trajectory of this Markov chain in $[0, M] \subseteq \mathbb{N}$ and we estimate its measurements of this system. We denote by $\mathcal{W}(q,b)$ the discrete-time Weibull distribution, i.e, $X \sim \mathcal{W}(q,b)$ if $P(X=0) = 0$ and $P(X=k) = q^{(k-1)^b} - q^{k^b}$, $k \in \mathbb{N}^*$.

We use a trajectory X_0^M, with $M = 500$, that was generated by assuming that $f_{12} := G(0.8)$, $f_{21} := \mathcal{W}(0.1, 0.9)$, $f_{23} := \mathcal{W}(0.1, 2.0)$ and $f_{31} := \mathcal{W}(0.6, 0.9)$. The system is defined by:

- The initial distribution $\alpha = (1\ 0\ 0)$
- The semi-Markov kernel $\mathbf{q}$ with

$$\mathbf{q}(k) = \begin{pmatrix} 0 & f_{12}(k) & 0 \\ a f_{21}(k) & 0 & b f_{23}(k) \\ f_{31}(k) & 0 & 0 \end{pmatrix},$$

$k \in \mathbb{N}$ with $a = 0.4$ and $b = 0.6$.

Figure 1 presents the true value of the ROCOF of the semi-Markov system, for $M = 500$. At this point we should refer that although we work at the discrete time case, we connected the values of the ROCOF continuously, in order to observe the evolution of the ROCOF in time more clearly.

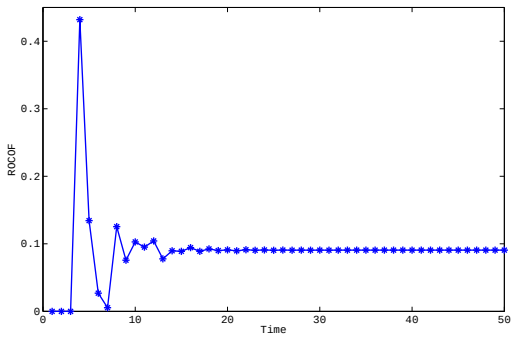


Fig. 1. True value of ROCOF versus time.

ACKNOWLEDGMENT

This research has been co-financed by the European Union (European Social Fund-ESF) and Greek national funds through the Operational Program Education and Lifelong Learning of the National Strategic Reference Framework (NSRF)-Research Funding Program: Heracleitus II. Investing in knowledge society through the European Social Fund.

REFERENCES

[1] Ascher, H. and Feingold, H. (1984): Repairable Systems Reliability. *Marcel Dekker, New York*.

[2] Barbu, V. and Limnios, N. (2008): Semi-Markov Chains and Hidden Semi-Markov Models towards Applications in Reliability and DNA Analysis. *Springer*.

[3] Billingsley, P. (1961): Statistical Inference for Markov Processes. The University of Chicago. *Chicago Press*.

[4] Bracquemond, C. and Gaudoin, O. (2003): A survey on discrete lifetime distributions. *Int. J. Reliabil., Qual., Safety Eng.*, *10(1)*, 69–98.

[5] Chryssaphinou, O., Karaliopoulou, M. and Limnios, N. (2008): On discrete time semi-Markov chains and applications in words occurrences. *Statist. Probab. Lett.*, *37*, 1306–1322.

[6] Limnios, N. and Oprişan, G. (2001): Semi-Markov Processes and Reliability, Birkhauser, Boston.

[7] Ouhbi, B. and Limnios, N. (2002): The rate of occurrence of failures for semi-Markov processes and estimation. *Statist. Probab. Lett.*, *59*, 245–255.

[8] Van Der Vaart, A.W. (2000): Asymptotic Statistics (Cambridge Series in Statistical and Probabilistic Mathematics, 3). *Cambridge University Press*.

Stable variable selection for right censored data: comparison of methods

Marie Walschaerts
Equipe d'accueil EA3694
Recherche en Fertilité Humaine
Hôpital Paule de Viguier
330 avenue de Grande-Bretagne
31059 Toulouse, France
Email: walschaerts.m@chu-toulouse.fr.

Eve leconte
TSE (GREMAQ)
Université Toulouse 1 Capitole
21 allée de Brienne
31000 Toulouse, France
Email: leconte.cict.fr

Philippe Besse
IMT, UMR CNRS 5219
Université de Toulouse, INSA
118 route de Narbonne
31062 Toulouse Cedex 9, France
Email: philippe.besse@insa-toulouse.fr

***Abstract*—The instability in the selection of models is a major concern with datasets containing a large number of covariates. This paper deals with methods of variable selection in the case of high-dimensional problems where the interest variable is censored. We focuse on new stable variable selection methods based on two different methodologies commonly used in survival analysis: the Cox model and the survival trees. For that, we investigate bootstrapping adapted to stepwise algorithm, $\mathcal{L}_1$-penalization of Lasso and survival trees. We review these different variable selection approaches and apply them to an original infertility dataset. We compare their prediction performance with this obtained from the random survival forest methodology known to give the smallest prediction error but difficult to interpret by non-statisticians. We also compare the relevance of their interpretation. The aim is to find a compromise between a good prediction performance and ease to use for clinicians.**

I. Introduction

Problems of variable selection arouse a growing interest in the processing of data sets containing more and more variables. In the last twenty years, many methods of variable selection have been proposed to handle these high-dimensional problems, especially when the number of covariates p exceeds the number of observations n. To avoid a wrong estimation due to collinearity problems and to improve interpretation, the scientific community has developed tools to select the most relevant variables. A large literature concentrates on the case of the linear regression [1], [2]. A classical well-known method is the stepwise algorithm based on the Akaike Information Criterion (AIC). Recently, another field of research has focused on optimization problems, such as $\mathcal{L}_1$-penalty approaches. On the other way, tree-based algorithms provide an interesting alternative to handle non-parametrically a large number of covariates.

We consider here the special case where the response variable is right censored and review and adapt stable selection methodology based on the Cox proportional hazards model and survival trees procedures.

In section II, a presentation of the variable selection methods based on the Cox model and on survival trees is given, followed, in section III, by the comparison of the different approaches on an original data set on infertility, giving the prediction error rate and the selected variables in the final model for each approach. Concluding remarks and perspectives are presented in the last section.

II. Stable variable selection methods

A. Methods based on the Cox model

In survival analysis, the Cox model [3] has become the most popular method to modelize the relationship between a survival time and one or more predictors (*i.e.* covariates). This model has the advantage to be semi-parametric in the sense that it does not require assumptions on the survival time distribution. Morevoer, it is of easy interpretation for clinicians in providing estimates of the effect of the covariates on survival time after adjustment on the other covariates. the hazard function for the failure time of an individual takes the form

$$\lambda(t|Z) = \lambda_0(t)\exp(\beta' Z), \tag{1}$$

where β is a p-vector of unknown regression parameters and $\lambda_0(t)$ is an unknown baseline hazard function. However, when p is high, it may be very unstable, even when stepwise selection or $\mathcal{L}_1$-penalization (Lasso) are added to the classical procedure. To remedy this problem, some authors have proposed to use bootstrapping to investigate the reliability of the choice of the variables in the final model.

Sauerbrei and Schumacher [4] developped a bootstrap selection procedure which combined the bootstrap method with stepwise selection in Cox regression. They examined the inclusion frequencies of the variables selected by the stepwise algorithm into the models derived from the bootstrapped samples and keep in the final model the variables for which the inclusion frequency exceeds a given cut-off value κ in $(0, 1)$. The choice of κ is arbitrary.

Bach [5] introduced stability in the selection by using bootstrapping in a Lasso algorithm, a method called Bolasso. Adapted by Tibshirani [6] to the Cox model, the Lasso method estimates the β parameter via maximising the log partial likelihood function with the constraint

$$\sum_{j=1}^{p} |\beta_j| \leq \lambda \tag{2}$$

where λ is a regularization parameter. The Lasso constraint selects variables by shrinking estimated coefficients towards 0. This leads to coefficients exactly equal to zero and allows a parcimonious and interpretable model. The bootstrapped Lasso method (BLS) have only been considered in the framework of linear regression: we propose to extend it to the stabilization of the selection of covariates in a Cox model.

Meinshausen and Buhlmann [2] proposed a generalisation of the bootstrap Lasso procedure called bootstrap randomized Lasso (BRLS) where the covariates are penalized by different values randomly chosen in the range $[\lambda, \lambda/\alpha]$ with α in $(0, 1)$. This turns out to estimate the β parameter with the constraint

$$\sum_{j=1}^{p} \left| \frac{\beta_j}{W_j} \right| \leq \lambda. \tag{3}$$

In practice, the set of covariates $\{Z_j : \ j = 1, \ldots, p\}$ are weighted by the set $\{W_j : \ j = 1, \ldots, p\}$ randomly generated where $P(W_j = \alpha) = p_w$ and $P(W_j = 1) = 1 - p_w$ with p_w in $(0, 1)$.

B. Methods based on survival trees

Although they are not so popular than the Cox model, tree-based methods in survival analysis (the so-called survival trees) have known a great development in the last decades. They provide a good alternative to the Cox regression model in identifying covariates which play a role on the survival outcome and in predicting the individual risk of failure. In addition to be easy to interpret in a large frame of applications, survival trees methods can incorporate non linear effects, and also take into account interactions between covariates. First developed for basic classification trees, the Classification and Regression Trees (CART) algorithm of Breiman [7] is based on binary recursive partitioning. This is an iterative process which splits the data into two subgroups (daughter nodes) according to the value of one of the predictors. The splitting rule maximises the difference between nodes. Instability in the selection of covariates by regression trees has been observed and demonstrated by many authors [8]–[11]. This instability may be due to an overfitting of data. The variance observed may also come from arbitrary cutpoints defined by the dichotomization of continuous covariates.

Dannegger [9] proposed a bootstrap node-level stabilization procedure (BNLS) for survival trees. The algorithm consists, at node h, in drawing bootstrapped samples from the original set, and for each of them, in finding the best split. The split which appears the most of the time at the node h is selected. For a continuous variable, the cut-off value b chosen in the set of realizations of the split variable is the median of all the b-values proposed at each bootstrap. As Dannegger [9] did not propose a choice of the cut-off for a categorical variable, we decide to affect to the daughter node the level of the categorical variable which was mainly chosen by the boostrapped samples.

Breiman [8] proposed a random selection approach which combines the bagging method (bootstrap aggregating) with a random selection of the covariates at each node of the tree. The method was adapted to the survival framework in an approach called "random survival forests" (RSF) by Ishwaran *et al.* [12]. However the model obtained with this latter method is not easy to use and interpret by non-statisticians, unlike BNLS.

III. Comparison of the methods

A. Comparison criteria

In order to compare the different variable selection approaches, we apply them to a fertility data set. The five following procedures have been compared: the bootstrap Cox stepwise procedure (BSS), the bootstrap Cox Lasso procedure (BLS), the bootstrap Cox randomized Lasso procedure (BRLS) with three different values of α (0.2, 0.4 and 0.6), the bootstrap node-level stabilization procedure (BNLS) and the random survival forest method (RSF). We graphically chose sensible values of the cut-off κ for the methods based on the Cox model and also values of λ for BLS and BLRS. For that, considering the whole data set, we plot the values of κ for each selected covariable for BSS and we plot the values of κ with respect to values of λ for BLS and BLRS. The aim is to identify the most relevant variables whose frequencies of inclusion κ are larger than those of the other covariates whatever the value of λ. Then we chose the values of λ and κ which maximize the gap between the two subgroups of covariates. For BNLS, the choice of the optimal complexity parameter cp is achieved by cross-validation techniques.

Concerning the number of bootstrapped samples, we take $N = 100$ for the Cox model based procedures, the default value $N = 1000$ for RSF, and $N = 1000$ for the BNLS procedure.

We compare the stability of the five procedures using a statistical criterion, the prediction error rate, which allows to quantify the prediction performance of the final model selected. To this aim, the original data set is divided into two subsets, the training set and the test set. The procedures, computed on the training set, give a final model which is then applied on the test set to calculate the prediction error rate, using the Harrell's concordance index C [13]. A total of 30 training sets and test sets were drawn to obtain a sample of error rates. Boxplots of the error rates are presented in order to compare the variability of the prediction performances.

B. Fertility data set

The fertility data were obtained from 2138 couples consulting for male infertility during the period from 2000 to 2004 at Toulouse Male Sterility Center (TMSC) located in Universitary Public Hospital in Toulouse (France) [14]. Patients were followed from entry and during treatment by an andrologist specialist until either discontinuation of treatment or delivery of an alive infant. The maximum follow-up duration is 9 years. The outcome assessment was based on the delivery of an alive infant obtained at TMSC (pregnancies after medical treatment - medecine and/or surgical treatment, or assisted reproductive technologies (ART) - as well as spontaneous pregnancies). The event considered here is the birth of an alive infant and right-censored events correspond to miscarriage

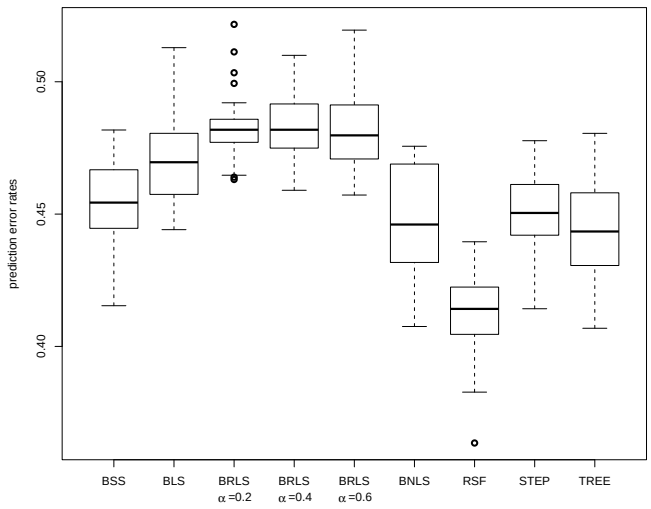


Fig. 1. Boxplots of the prediction error rates for the different methods : Bootstrap Stepwise selection, Booststrap Lasso selection, Bootstrap Randomized Lasso selection (three values of α), Bootstrap node-level selection, Random Survival Forest, Cox stepwise selection and a single survival tree.

or loss to follow-up. The "survival" time is the delay in months from the first visit of the couple to the birth of its alive infant. We will work on the subset of the 1773 couples with covariates without missing values. 40% of the couples succeed in their parental project, leading to a censoring rate of 60 %. In agreement with clinicians, we decided to keep 32 covariates, among which male age and his clinical investigation including medical histories and clinical examination, female age and her clinical investigation, if couple received a non-ART treatment (mediacl/surgical or hormonal treatment), fecundity type, infertility duration, type of ART including IUI (intra-uterine insemination), IVF (in vitro fertilization), ICSI (intracytoplasmic sperm injection) using male sperm cells, and ART with donor sperm.

The boxplots of the prediction error rates for the five procedures are presented in figure 1 and table I shows the summary statistics. Notice that the mean and median values obtained are high (between 0.41 and 0.45) which reflects the difficulty to predict the delay to the birth. As expected, the RSF method gives the best predictive model. However, it appears that RSF is similar in dispersion with other procedures. Moreover, we can notice that the BNLS procedure is not better than a single survival tree and the same remark can be done between the BSS procedure compared to a single stepwise Cox regression. These results can be explained by the size of the sample, which is sufficient to produce low error rates without bootstrapping. Regarding BLS and BRLS, these procedures show less variation in error rates but their means and medians are close to 0.5 which suggests that these models do no better prediction than random guessing.

If we compare the covariates selected by the different approaches, we can see in figure 2 that the first four

TABLE I

MEAN, STANDARD DEVIATION AND MEDIAN OF ERROR RATES FOR THE DIFFERENT PROCEDURES: BOOTSTRAP STEPWISE SELECTION, BOOSTSTRAP LASSO SELECTION, BOOTSTRAP RANDOMIZED LASSO SELECTION (THREE VALUES OF α), BOOTSTRAP NODE-LEVEL SELECTION, RANDOM SURVIVAL FOREST, COX STEPWISE SELECTION AND A SINGLE SURVIVAL TREE.

	Mean	Standard deviation	Median
BSS	0.453	0.017	0.454
BLS	0.471	0.017	0.470
BRLS $\alpha = 0.2$	0.483	0.013	0.482
BRLS $\alpha = 0.4$	0.484	0.013	0.482
BRLS $\alpha = 0.6$	0.482	0.015	0.480
BNLS	0.447	0.021	0.446
RSF	0.413	0.017	0.414
STEP	0.451	0.015	0.450
TREE	0.445	0.018	0.443

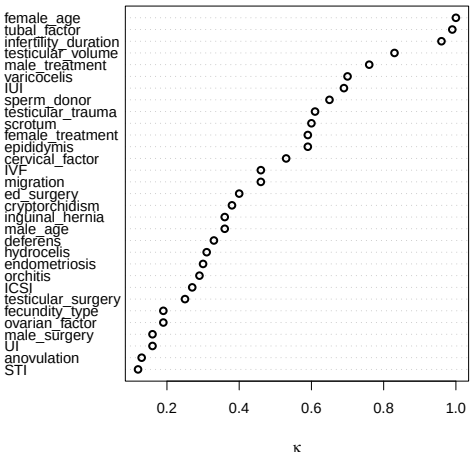
selected covariates do not differ for the BLS and BRLS procedures for a value of $\kappa = 0.3$: we find `tubal factor`, `IUI`, `sperm donor` and `infertility duration`. The additional covariates included by BLS are `epididymis`, `varicocelis`, `inguinal hernia`, `fecundity type`, `testicular trauma` and `testicular volume`. For the BSS procedure, for a value of $\kappa = 0.5$, we find in the selected covariates the first four covariates selected by BLS but also `female age`, `testicular volume`, `male treatment`, `varicocelis`, `testicular trauma`, `scrotum`, `female treatment` and `epididymis`. We observe also that the BSS procedure includes more variables than the BLS and BRLS procedures, which leads to a lower prediction error rate.

If we adjust a single Cox stepwise regression, we find the same covariates selected by the BSS procedure except for the variable `scrotum`. However, `epididymis`, `testicular trauma` and `male treatment` are not statistically significant at the 5% level in the Cox stepwise model.

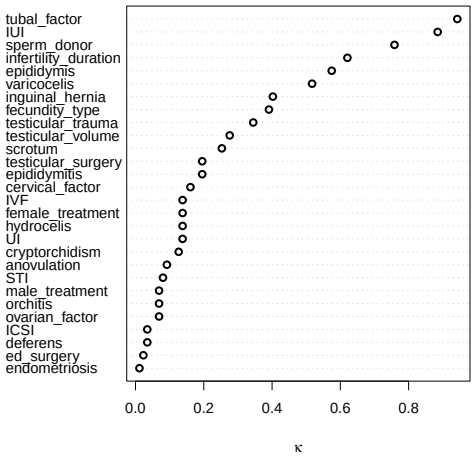
Regarding the tree-based RSF and BNLS procedures, we observe in figures 2 and 3 that the selected covariates are substantially different from the most relevant covariates selected by the Cox based procedures. The first split, which corresponds to `female age`, is the same for the two types of procedures but the following splits are different. This may be explained by the fact that the tree based procedures take into account interactions contrary to the Cox model based procedures. Moreover, the covariates selected for the BNLS procedure are found in the most important variables in RSF. The variables selected by a single survival tree are almost the same than those selected by BNLS, which can be explained by the large sample size of the fertility data set.

IV. CONCLUSION

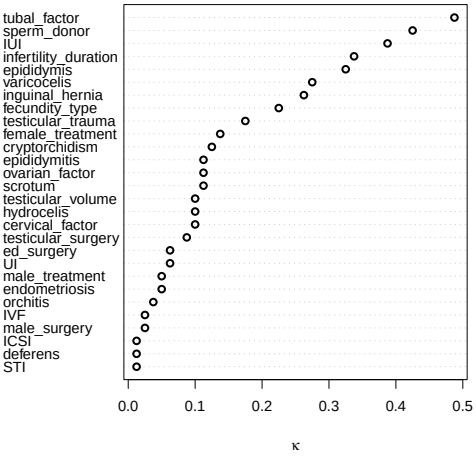
On a 2.4GHz processor, for the fertility data set with 1773 individuals and 32 covariates, running the RSF procedure is 200 times longer than running a single survival tree from the



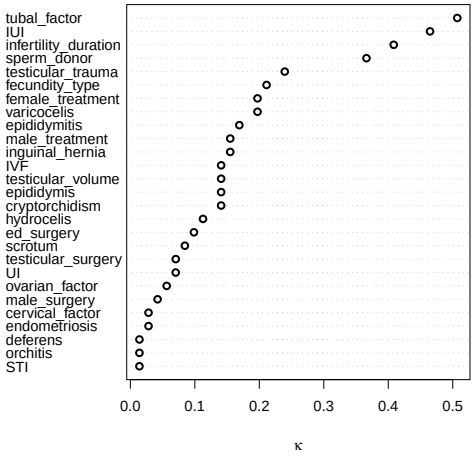
[A]



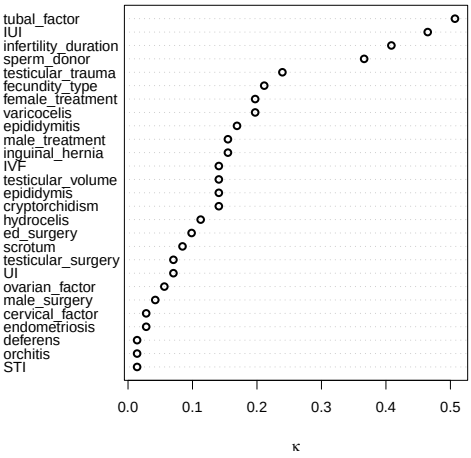
[B]



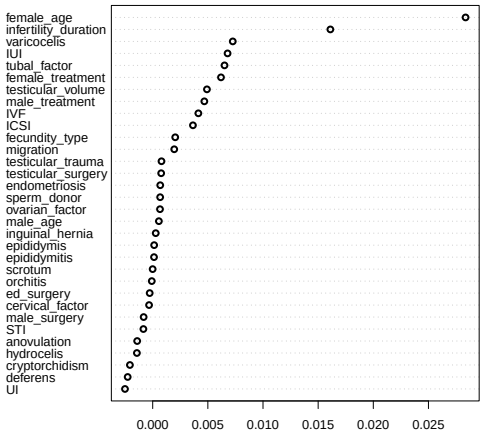
[C]



[D]



[E]



[F]

Fig. 2. Selected variables for [A] Bootstrap Stepwise Selection, [B] Bootstrap Lasso Selection, [C,D,E] Bootstrap Randomized Lasso Selection and [F] importance of variables for Random Survival Forest.

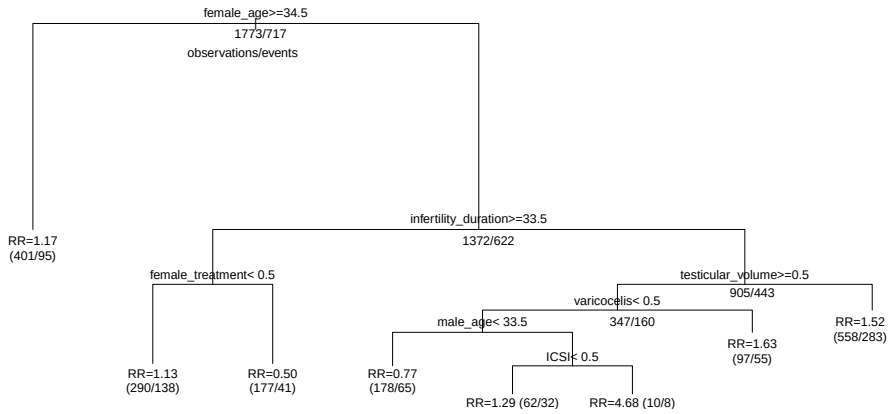


Fig. 3. The final tree obtained from the Bootstrap Node-Level selection.

CART procedure (which takes about 1 second), whereas the stepwise algorithm takes about 10 seconds. Compared to the RSF method, the BLS and BRLS procedures have similar computing times. The BNLS procedure is much longer than RSF (4 times) as a bootstrap is realized at each node. Finally, the BSS procedure, for which a bootstrap sample is used at each stepwise, is the most expensive procedure in time (8 times longer than RSF). Thus, the running times seem reasonable for each procedure taken separately. However, a systematic comparison of errors becomes heavy, as well as the running time to exhibit the optimal complexity parameter cp in the cross-validation procedure for BNLS, and also the penalty λ for the BLS and BRLS methods.

We find that RSF has the lowest prediction error rate. The RSF procedure is easy to use and does not require the choice of tuning values as do the BSS, BLS, BRLS and BNLS procedures. However, even if the selected covariates are identified and sorted by their importance, as no final tree is provided, the RSF results stay a black-box not easy to interpret and use for clinicians. We find that the BRLS procedure, whatever the value of α, does not seem to improve the BLS procedure, although the gap between the selected variables appears more clearly for BRLS. Moreover, the good results obtained with this method by Meinshausen and Buhlmann [2] may come from the fact that they presented an ad-hoc example. As noticed by these authors, we find on our data sets that choosing a value of κ equal to 1, as suggested by Bach [5], is too restrictive. Even if some authors [4], [15] showed that the bootstrap method adapted to the stepwise algorithm improves the stability of the variable selection, our results suggest no improvement with the BSS procedure: for the fertility data set whose size is sufficient to assure the convergence of the algorithm, it does not give better results than a single stepwise Cox model. As far as the BNLS procedure is concerned, it seems not perform better than a single survival tree, contrary to the results found by Dannegger [9]. It can be explained by the fact that it is difficult to tune a sensitive value for the complexity parameter by cross-validation. It may also result from the sufficient size of the fertility data set.

Finally, these results suggest that the Cox and tree based procedures should be performed in a complementary way to identify the most relevant covariates and provide to clinicians a stable and reliable model. Each procedure shows indeed a particular interest, either in terms of its prediction performance, either in the selection of the relevant covariates.

ACKNOWLEDGMENT

The authors would like to thank the Agence de la Biomedecine for grants.

REFERENCES

[1] F. E. JR. Harrell, K. L. Lee, R. M. Califf, D. B. Pryor, R. A. Rosati, "Regression modelling strategies for improved prognostic prediction", *Stat. in Med.*, vol. 3, pp. 143–152, 1984

[2] N. Meinshausen, P. Buhlmann P, "Stability selection", *J. Roy. Stat. Soc. B*, vol. 72, pp. 417–473, 2010.

[3] D. R. Cox, "Regression models and life tables (with discussion)", *J. Roy. Stat. Soc. B*, vol. 34, pp. 187–220, 1972.

[4] W. Sauerbrei, M. Schumacher, "A bootstrap resampling procedure for model building: application to the cox regression model", *Stat. in Med.*, vol. 11, pp. 2093–2109, 1992.

[5] F. Bach, "Model-consistent sparse estimation through the bootstrap", *Technical Report*, 2009.

[6] R. Tibshirani, "The lasso method for variable selection in the cox model", *Stat. in Med.* vol. 16, pp. 385–395, 1997.

[7] L. Breiman, *Classification and Regression Trees*, Chapman and Hall/CRC, 1984.

[8] L. Breiman, Bagging predictors, *Mach. Learn.*, vol. 24, pp. 123–140, 1996.

[9] F. Dannegger, "Tree stability diagnostics and some remedies for instability", *Stat. in Med.*, vol. 19, pp. 475–491, 2000.

[10] L. Ruey-Hsia, "Instability of decision tree classification algorithms", *PhD thesis*, 2001.

[11] S. Gey, J. M. Poggi, "Boosting and instability for regression trees", *Comput. Stat. Data An.*, vol. 50, pp. 533–550, 2006.

[12] H. Ishwaran, U. B. Kogalur, E. H. Blackstone, M. S. Lauer, "Random survival forests", *Ann. Appl. Stat.*, vol. 2, pp. 841–860, 2008.

[13] F. E. Harrell, C. E. Davis, "A new distribution-free quantile estimator", *Biometrika*, vol. 69, pp. 635–640, 1982.

[14] M. Walschaerts, L. Bujan, F. Isus, J. Parinaud, R. Mieusset, P. Thonneau, "Cumulative parenthood rates in 1735 couples: impact of male factor infertility", *Hum. Reprod.*, 10.1093/humrep/der466, 2012.

[15] C. H. Chen, S. L. George, "The bootstrap and identification of prognostic factors via cox's proportional hazards regression model", *Stat. in Med.*, vol. 4, pp. 39–46, 1985.

Employing the Diagnostic Matrix for Supporting the Reliability of the Aircraft Gas Turbine Engine in the Operating Process

Sergey Yunusov
Transport and Telecommunication Institute
Riga, Latvia
yunusov@inbox.lv

***Abstract* — the presented paper offers the approach of probability towards the determination of the level of reliability of the aircraft gas turbine blocks and assemblies. The described approach makes significant the choice of reference parameters imposing direct impact on the engine safety and allowing localising the block or the assembly, showing the decrease of the reliability level. It is offered to employ the diagnostic matrix for localising the engine component where the process of safety decline has begun. Implementing the diagnostic matrix permits to determine the moment of the very beginning of the engine parameters deviation, and allows controlling them up to the moment of the limit deviation achievement. The paper also considers the approach based on the Tikhonoff's method of regularisation and used for eliminating certain restrictions on implementing the diagnostic matrix.**

Keywords – reliability; gas turbine engine; probability of failure; method of minor deflection; diagnostic matrix; regularisation method; regularisation parameter

I. INTRODUCTION

The efficiency of the aviation equipment employment is substantially determined by the perfectness of the exploitation methods. The perfection of any method of servicing and maintenance is determined by the degree of providing the interaction between the actually existing process of changing the object technical state and the process of its technical employment. The conditioned methods of servicing and maintenance provide the closest connection between the above described processes. The principal maxim of these methods is the maxim of preventing the failures of the functional systems of the aircraft and engine under the condition of their maximum operating time and providing the sufficient level of the flight safety.

The great variety of the servicing and maintenance methods on condition can be divided for convenience into two groups: controlling the reliability level and controlling the parameters. The method of technical servicing and maintenance on condition, controlling the reliability level is the method of setting the margin of reliability for the homogeneous items. The objects are employed while the safety level is within the limits of set regulations.

The practical implementation of the method of technical servicing and maintenance on condition controlling the safety level allows decreasing the exploitation costs.

The high cost of the gas turbine engine (GTE) in comparison with the other components of the aircraft requires providing more resources, and it is achieved by implementing the optimal strategy of the engine employment till the pre-failure state of its details and blocks.

The concept of employing the aircraft engines till the pre-failure state stipulates the wide implementation of the methods and facilities providing the prognostication of the technical state of equipment. This fact distinguishes the methods of engine diagnostics compared to other technical systems.

The peculiarity of the aircraft gas turbine engine is the occurrence of close interconnection between various physical processes taking place within them. This interconnection results in complicated models, describing the diagnostic characteristics and the processes of the defects appearance and development. It is necessary to obtain the sufficient information on the physical parameters of the processes, having different nature for qualified diagnostics of the engine technical state. Moreover, there should take place the complex processing and analysis of measured diagnostic information with implementing the mathematical models, characterising the various physical factors interconnections, considering the factors, influencing and directly predetermining the development of defects and failures.

The diagnostic models of the aircraft engines are used for forming the informative diagnostic characteristics, and for numeric imitation of the functional processes for the regularly operating engine on the purpose of comparing the process computed characteristic parameters with the actual measured values and taking decision on the engine technical state.

The control of the safety level allows decreasing the operation costs of the engine service and improves the flights safety.

II. METHOD OF EVALUATING AND SUPPORTING THE ENGINE SAFETY

There are changes in the state of the aircraft engine in the process of its exploitation and technical servicing; these changes have impact on the reliability and safety of flights. At the same time, it is possible to distinguish two principal processes, influencing the engine state changes.

The first process of the engine state changing depends on its constructive perfectness, on the environment impact and on the conditions of exploitation and technical servicing. This process is determined by the actual activities, taking place in the engine components and systems and are described by transit from the operable state into non-operable. This transit is performed by means of accumulating the quantitative changes in the engine details and blocks, and then these changes are revealed via wearing out of the operating surfaces of mobile details, appearing cracks, alteration of the engine operation characteristics, etc. Under the certain conditions the quantitative alterations in the engine details and blocks turned to be the qualitative ones, and then they result in occurring defects and failures. The appearing failures and defects have the nature; consequently, they can be forecasted.

The second process, as distinct from the first one, is a subjective process, and the change of the engine state mostly depends on the human factor. Under the condition of the second process the change of the engine state mostly depends on the staff, maintaining the engine, its qualification, experience, operations arrangements, and technologies of preparation and servicing. This process also has a nature and can be analyzed by applying the methods.

Both processes are interconnected and interdependent, and they together impose impact on the engine state change, determining its safety. Nevertheless, as the existing experience of implementation demonstrates, the principal process, determining the engine safety is the first one. It is the reason why the control of the state of engine details and blocks is a determining condition in the procedure of evaluating its reliability. Information, registered by the engine control systems allows judging about the degree of deflection of certain parameters but not about the reliability of engine details and blocks. Consequently, there is a need in summarised information indicating the decrease in the level of reliability but not about failure appearance; in this case there is time for taking some actions directed on avoidance of failure. Summarised information on the engine state can be obtained if there are current values of engine control parameters in different modes of its operating. As a rule, the analysis of changing the engine parameters allows evaluating objectively the reliability level at the moment of taking control. In the process of exploitation on condition it is important to have information on the first appearance of defect, which is long before the failure itself. Moreover, it is important to have a parameter or parameters, capable of reacting on any changes or anomalies in the engine operating. If it is possible to obtain marginal accessible values or marginal accessible deflections from the computed values, describing the safety (the probability of non-failures in operations) on the basis of computations or experiments, it is possible to obtain the method of evaluating the engine reliability in the process of operating on condition by means of employing the dependences of the theory of probability. There is considering the method, allowing estimating the engine reliability level at any moment of operating it by employing facilities. The above described approach allows determining immediately the reserve of serviceability and establishing the necessity of implementation of the preventive measures oriented on supporting the specified level of flight safety and reliability. If the admissible limit value or deflection of the certain control parameter is designated as *y* and actual value as *x*, the reserve of serviceability is $z = y - x$. The measurable and computable engine thermo gas dynamic parameters can serve as the control parameters. All the admissible limit values of the control parameters *y* are determined beforehand and are implemented for estimating the engine state under the employment conditions. It is obvious, the values *x*, *y* and *z* are the variables, and as a rule they are distributed according to Gaussian law. For Gaussian law the distribution density for these values can be performed by the corresponding functions *f(x), f(y)* and *f(z).* For Gaussian law the probability of the fail-safe performance and probability of failure, employing Laplace's function, are relatively determined by the functions

$$P(t) = 0.5 - \Phi(U) \tag{1}$$

$$Q(t) = 0.5 + \Phi(U) \tag{2}$$

In the described case:

$$\Phi(U) = \Phi\left(\frac{\bar{z}}{\sigma_z}\right),$$

where $\bar{x}$, $\bar{y}, \bar{z}$ are average expectations of values

x, y, z, as well as σ_z are standard deviation *z.*

The transferred Laplace's function has the following view:

$$\frac{\bar{z}}{\sigma_z} = \frac{\bar{y} - \bar{x}}{\sigma_z} = \frac{\bar{x}}{\sigma_z}\left(\frac{\bar{y}}{\bar{x}} - 1\right) = \beta(k-1)$$

$\beta = \frac{\bar{x}}{\sigma_z}$ is the coefficient of the quality of processes and determines the dispersion of the values; $k = \frac{\bar{y}}{\bar{x}}$ characterises the serviceability reserve of details and blocks, then the formulae (1) and (2) can be presented as:

$$P(t) = 0.5 - \Phi(\beta(k-1)) \tag{3}$$

$$Q(t) = 0.5 + \Phi(\beta(k-1)) \tag{4}$$

Implementing Laplace's tables, it is possible to estimate the level of reliability *P(t)* and probability of failure *Q(t)* without any difficulty. If the value of the reliability level is lower than

it has been specified, it serves as a signal for examining the engine blocks state, aiming eliminating the probable failure. Under the condition of increasing the number of controlled parameters, the efficiency and performance capabilities of this method will grow.

However, alteration of the control parameters does not provide the accurate information about location of block, piece equipment, system where the processes of falling the functionality properties start.

In this case it is important to realise the fall in the safety level not in general but in the specific components. That is why there is a need in method permitting determining the blocks and units, where the process of falling reliability has started, according to the specified set of values of the control parameters. The method of diagnostic matrices is offered for this operation. The concept of this method is considered below.

III. FORMATION OF THE DIAGNOSTIC MATRIX FOR CONTROLLING THE STATE OF THE GAS TURBINE ENGINE AIR GAS CHANNEL

A. Developing the Adequate Mathematical Model of the Engine for Diagnostic Matrix Formation

All the peculiarities and properties of the diagnostic matrices useful for practice are the results of the analysis of the mathematical model of the processes taking place in the air gas channel of the gas turbine engines. The equations describing the processes in the components of the gas turbine engine are complicated; some of them are given graphically (for example, the compressor features). The operational process of the gas turbine engine represents the aggregate of the whole range of closely interconnected complex processes, and changing even single of the gas turbine engine parameters – sectional area, loss coefficient, coefficient of efficiency etc. – results in changing almost all other parameters of the flow (pressure, temperature, velocities) and further in the changes of the engine properties – thrust, power, fuel consumption. For simplifying the analysis of dependence between the increments in the interconnected equation parameters the approximate mathematical method is implemented – the method of minor deviations, allowing obtaining the linearization of the process original equations. As an example there is the way of obtaining the equations in minor deviations from the process principal equations. The work, spent for compressing 1 kilo of air in the compressor is expressed via the equation:

$$L_c = \frac{k}{k-1} R T_{t1} \left(\pi_c^{0.286} - 1 \right) \frac{1}{\eta_c},$$

where

T_{t1} – is the temperature of the stagnated flow before the compressor,

π_c – is the degree of increasing the full pressure in the compressor,

η_c – is the compressor adiabatic efficiency on the stagnated flow parameters.

After taking the logarithms of the right and the left parts of the equation

$$ln\, L_c = ln\left(\frac{k}{k-1} R \right) + ln\, T_{t1} + ln\left(\pi_c^{0.286} - 1 \right) - ln\, \eta_c$$

and then differentiating the received equation and taking into consideration that

$$d\left(ln\, x \right) = \frac{dx}{x},$$

it is possible to discover

$$\frac{dL_c}{L_c} = \frac{dT_{t1}}{T_{t1}} + \frac{0.286 \pi_c^{0.286}}{\pi_c^{0.286}} \frac{d\pi_c}{\pi_c} - \frac{d\eta_c}{\eta_c}$$

Though, the connection between the relative changes is found, and the following notions are introduced:

$$\frac{dL_c}{L_c} \approx \frac{\Delta L_c}{L_c} = \delta L_c$$
$$\frac{dT_{t1}}{T_{t1}} \approx \frac{\Delta T_{t1}}{T_{t1}} = \delta T_B$$
$$\frac{d\eta_c}{\eta_c} \approx \frac{\Delta \eta_c}{\eta_c} = \delta \eta_c$$
$$\frac{d\pi_c}{\pi_c} \approx \frac{\Delta \pi_c}{\pi_c} = \delta \pi_c$$

then

$$\delta L_c = \delta T_{t1} + K_1 \delta \pi_c - \delta \eta_c$$

This is the final equation of the process of compressing in minor deviations. The coefficients of $\delta T_{t1}, \delta \pi_c, \delta \eta_c$ in this equation are the coefficients of influence on L_c. It is obvious, that in δT_{t1} the coefficient is 1, and in $\delta \eta_c$ it is -1, and in $\delta \pi_c$ it is K_1

$$K_1 = \frac{0.286 \pi_c^{0.286}}{\pi_c^{0.286} - 1}$$

The equation in minor deviations should be understood in the following way: if the air temperature T_{t1} increases by 1%, the operation of compressing, ceteris paribus, $\left(\pi_c = const, \eta_c = const, \delta \pi_c = \delta \eta_c = 0 \right)$ increases by 1%. Similarly to this, in case of increasing η_c by 1% the operation of compressing decreases by 1%, and if π_c increases by 1% the operation of compressing increases by K_1 %. So the equation allows the determination of changing the operation of L_c with simultaneous alteration of T_{t1}, π_c, η_c, as well as finding the connection between the changes of T_{t1}, π_c and η_c under the determined change in the compression operation, in

other words finding solution for the reversed task. As an instance there is a mathematical model below. This model was obtained with the implementation of the method of minor deviations for the turbo propeller engine TB7-117C, presented in the following form

- the equation of parity of the compressor and the turbine operations (absence of bypassing and air bleeding):

$$K_1\delta\pi_c - \delta\eta_c = \delta T_{t4} + K_3\delta\pi_T + \delta\eta_T$$

- the correlation of pressures in channel:

$$\delta\pi_{out} = \delta\pi_c - \delta\pi_T + \delta\sigma_{in} + \delta\sigma_{ch} + \delta\sigma_{out}$$

- the equation of compression process in the compressor:

$$\delta T_{t2} = K_1 \cdot K_2 \cdot \delta\pi_c - K_2 \cdot \delta\eta_c$$

- the equation of the extension process in turbine:

$$\delta T_{t4} = \delta T_{t3} - K_3 \cdot K_4 \cdot \delta\pi_T - K_4 \cdot \delta\eta_T$$

- the equation of the continuity between the compressor inlet and the throat section of the turbine nozzle cascade:

$$\delta G_{air} = \delta\sigma_{in} - \delta\pi_c + \delta\sigma_{ch} + \delta F_T - 0.5\sigma T_{t3}$$

- the equation of the continuity between the turbine nozzle cascade and the jet nozzle:

$$\delta\pi_T - \delta\sigma_{out} + \delta F_T = \delta F_{out} + K_6\delta\pi_{out}$$

- the equations describing the compressor properties:

$$\delta G_B = K_{10}\delta\pi_C + \delta\sigma_{in}$$
$$\delta\eta_C = K_{11}\delta\pi_C + \delta\overline{\eta}_C$$

- the equation of the heat input:

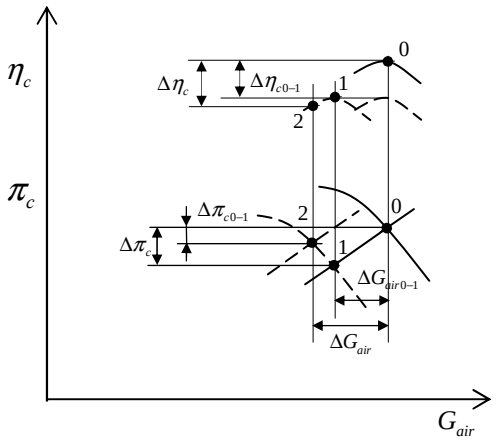
$$\delta G_F = \delta G_B + K_5\delta T_{t3} - (K_5 - 1)\delta T_C - \delta\xi_{ch}$$

- the equation of the jet thrust:

$$\delta R = K_9\delta F_{out} + K_7 \cdot K_8 \cdot K_9\delta\pi_{pc} - (K_9 - 1)\delta G_{air}$$

Nevertheless, in the obtained linear mathematical model the alteration of the engine air gas channel state is not taken into consideration, and this fact does not allow further forming the adequate diagnostic matrix. For taking into account the changes of the engine air gas channel it is suggested using the idea of introduction of the specific components into the compressor equations.

It is assumed here that under the condition of changing the state of the compressor air gas channel its head-capacity characteristics shift in the equidistant way by value ΔG_{air0-1} horizontally and $\Delta\pi_{C0-1}$ vertically (Fig.1.)



Figures 1: Diagram of Property Shifting

Besides the property shift there is also the shift of the operational modes line and the operational point on it (1-2) in the way that the cumulative change of the air consumption and the degree of pressure are correspondingly ΔG_{air} and $\Delta\pi_C$. Simultaneously there is the change in the efficiency by value $\Delta\eta_{C0-1}$, and taking into account the operational point shift - by $\Delta\eta_c$. Employing the approach [2], the full relative increments of the parameters in this case are determined as the sum of relative partial increments in the first (0-1) and the second (1-2) transitions:

$$\delta G_{air} = \delta G_{air0-1} + \delta G_{air1-2}$$
$$\delta\eta_c = \delta\eta_{c0-1} + \delta\eta_{c1-2}$$
$$\delta\pi_c = \delta\pi_{0-1} + \delta\pi_{c1-2}$$

Taking into consideration that

$$\delta G_{air1-2} = K_{10}\delta\pi_{c1-2};$$
$$\delta\eta_{c1-2} = K_{11}\delta\pi_{c1-2}$$
$$\delta\pi_{c1-2} = \delta\pi_c - \delta\pi_{c0-1}$$

the following correlations are received:

$$\delta G_{air} = \delta G_{air0-1} + K_{10}(\delta\pi_c - \delta\pi_{c0-1})$$
$$\delta\eta_c = \delta\eta_{c0-1} + K_{11}(\delta\pi_c - \delta\pi_{c0-1})$$

The parameters relative increments along the line (0-1) are marked as $\delta\overline{G}_{air}; \delta\overline{\eta}_C; \delta\overline{\pi}_C$, then

$$\delta G_{air} = \delta\overline{G}_{air} + K_{10}(\delta\pi_C - \delta\overline{\pi}_C)$$
$$\delta\eta_C = \delta\overline{\eta}_C + K_{11}(\delta\pi_C - \delta\overline{\pi}_C)$$

Is expressed $\delta\overline{G}_{air} = \overline{K}_{10}\delta\overline{\pi}_C$, where

$$\overline{K}_{10} = \frac{G_{air1} - G_{air0}}{G_{air0}} \frac{\pi_{C0}}{\pi_{C1} - \pi_{C0}} = \frac{\Delta\overline{G}_{air}}{G_{air}} \frac{\pi_{C0}}{\Delta\overline{\pi}_C}$$

This is the coefficient of the connection of alteration of the given air consumption and the degree of the compressor pressure increase along the line of shifting the head-capacity

curve (it is assumed, that this shift takes place along the operational modes line). Accordingly there is the additional parameter $\Delta\bar{G}_{air}$ in the equations, and this parameter characterizes the shift of the head-capacity curves of the compressor properties. The set of equations describing the compressor properties taking into consideration the malfunctions of the compressor and correspondingly the shift of its properties is the following:

$$\delta G_{air} = K_{10}\delta\pi_c + \left(1 - \frac{K_{10}}{\bar{K}_{10}}\right)\delta\bar{G}_{air} + \delta\sigma_{in}$$

$$\delta\eta_c = K_{11}\delta\pi_c + \delta\bar{\eta}_c - \frac{K_{11}}{\bar{K}_{10}}\delta\bar{G}_{air}$$

Accordingly the original equations describing the compressor properties, in the original mathematical model are exchanged by the newly obtained ones. The obtained mathematical model will be implemented for the diagnostic matrix formation.

B. Application of the Worked-out Algorithm for Finding the Steady Information-Diagnostic Matrix

The diagnostic matrix formation is demonstrated by setting the diagnostics of the air gas channel of the engine TB7-117C, for which the mathematical model was developed. It is worth mentioning that the depth of the engine air gas channel diagnostics is connected with the number of the registered fluid dynamics parameters. In this case the following parameters are supposed to be the registered (measured) ones:

- full temperature and air pressure at the engine inlet;
- speed of the compressor rotor;
- full gas temperature behind the turbine;
- full air pressure behind the compressor;
- fuel consumption per hour;
- speed of the free power turbine rotor.

For obtaining the diagnostic matrix the measured parameters are taken into the right part of the equation and the left part comprises the others. The necessary number of the measured parameters is determined as a difference between the number of variables and the number of equations. The number of the measured parameters needed for the diagnosis is determined by the number of independent criteria of calculation, the deviations of which characterize the state of the engine junctions. The diagnostic matrix is formed for the mode of maximum operation duration (under the condition of power N=2720 kW). The left parts of the equations are used for matrix A creation, and the right parts – for matrix B. Substituting them with the numerical values of the coefficients, calculated for this operation mode, and solving the set of equations: $C = A^{-1}B$.

The engine diagnostic matrix is obtained. Formation is described in details in the work [3]. The first four lines of the diagnostic matrix present the information on the gas generator state. The other lines have the referential features. For examining the diagnostic matrix it is possible to simulate the gas generator malfunctions, giving any deviation of the junctions state parameters $\left(\delta\bar{G}_{air}, \delta\bar{\eta}_C, \delta F_T, \delta\eta_T\right)$

C. Application of the Diagnostic Matrix for Fault Localization

The diagnostic matrix of TB7-117C engine was obtained with this worked-out algorithm, and the number of measured parameters is four. It is important, the matrix under consideration is the matrix of the definite operational and efficient engine.

TABLE 1: ENGINE DIAGNOSTIC MATRIX

	δn_{Tc}	δT_{Tc}	δG_f	$\delta\pi_c$
$\delta\bar{G}_{air}$	-2.48	-0.702	0	0.053
$\delta\bar{\eta}_C$	0.189	6.736	-8.803	0.482
δF_T	0	-1.397	1.287	-1
$\delta\eta_T$	0	-5.667	6.229	-0.551
$\delta\bar{\eta}_c$	0	6.809	-8.803	0.553
δG_{air}	0	-0.757	0	0
$\delta\pi_{Tc}$	0	0.257	0	1
$\delta\pi_{Tf}$	0	-0.257	0	0
δT_{t4}	0	-1.284	2.574	0
δT_{t2}	0	-3.871	5.005	0

The achieved diagnostic matrix (DM) allows analyzing the diagnostic properties, which are response for the fluid dynamic parameters' minor deviations. The diagnostic matrices allow locating mostly the charging set failure. At the engine operation process the gas generator failures develop as a rather specific, inherent for this trouble, set of deviations of measured parameters $\left(\delta G_f, \delta T_{t4}^*, \delta T_{t2}^*, \delta\pi_c^*\right)$. But even the experienced engineer, an expert, cannot estimate the whole variety of these parameters deviations. The diagnostic matrix allows locating the problem in the engine air-gas channel by means of defining trouble partial criteria (for compressor they are $\delta\bar{\eta}^*$ and $\delta\bar{G}_{air}$ and for turbine they are $\delta\eta_T^*$ and δF_T. Suppose at the process of engine operation the following set of parameter deviation is received: $\delta n_{Tc} = -0.074$,

$\delta T_{t4} = 1.659$, $\delta G_f = 1.472$, $\delta\pi_c^* = 0.426$, where δn_{Tc} is the engine speed deviation, δT_{t4}^* is the deviation of gas temperature behind the turbine, δG_f is the fuel consumption deviation, $\delta\pi_c^*$ is the engine pressure ratio deviation. Multiplying the coefficients of the efficient diagnostic matrix by the received deviations of the measured parameters, and summing these products up, we obtain the deviations $\delta\bar{\eta}_c^*$ and

$\delta\bar{G}_{air}$ in the engine compressor, and δF_T, $\delta\eta^*_{Tc}$ in the engine turbine.

$$\delta\bar{G}_{air} = (-2.48)\cdot(-0.074)+(-0.702)\cdot(1.659)+ +0\cdot 1.472+0.0503\cdot(-0.426)=-1.00368$$

$$\delta\bar{\eta}^*_c = 0.189\cdot(-0.074)+6.736\cdot(-0.702)+ +(-8.803)\cdot(1.472)+0.482\cdot(-0.426)=-2.00231$$

$$\delta F_T = 0\cdot(-0.074)+(-1.397)\cdot(1.659)+1.287\cdot 1.472+ +(-1)\cdot(-0.426)=0.00284$$

$$\delta\eta_{Tc} = 0\cdot(-0.074)+6.736\cdot 1.659+(6.229)\cdot(1.472)+ +(-0.553)\cdot(-0.426)=-0.003113$$

With this set of values the diagnostic matrix points the trouble in the compressor, as we see the compressor efficiency loss by 2% and air consumption rate by 1%, and the turbine is practically operational at this moment, as there is practically no any change in the passage area of the set of nozzles and no change in the turbine efficiency.

The values for every pair of criteria for every block allow introducing them as the defect field of vector, and then to estimate the development trends of these defects from flight to flight, and taking into consideration the access scope to anticipate the time of these defects dangerous development and engine malfunction. Certainly, the measured parameters deviations sets for every flight are average samples, obtained after flight information processing and referencing it to the definite engine work modes (typical operation, cruise rating, etc.).

IV. The Regularizing Algorithm of the Steady Inversion of the Calculation Identification Matrix

Nevertheless, the practice has shown that the diagnostic matrix development is not always possible, that is why certain papers [5] point the restrictions of their employment. It is connected with the fact that obtaining the reversed matrix A^{-1} is impossible, and consequently it is impossible to obtain the diagnostic matrix as well. There is an algorithm allowing sustained receiving the reversed matrix A^{-1} on the basis of Tikhonoff's regularisation method.

Stage1. First, the arbitrary vector u of order $m\times 1$ is found. Then the set of simultaneous linear algebraic equations is done.

$$Az = u, \tag{5}$$

where z a task is required solution vector (5).

Stage 2. In system (5) the matrix A and the right part of u are disturbed:

$$\left\|A^h - A\right\| \le h,$$

$$\left\|u_\delta - u\right\| \le \delta,$$

where h and δ are supposed to be invariably defined.

As a result, instead of (5) there is the following disturbed set of simultaneous linear algebraic equations:

$$A^h\tilde{z} = u_\delta, \tag{6}$$

where $\tilde{z}$ is the required approximate set (6) solution, dependent on the inaccuracies $\{h;\delta\}$.

Stage 3. The set (6) on the left is multiplied by the transposed matrix $\left(A^h\right)^T$:

$$\left(A^h\right)^T A^h\tilde{z} = \left(A^h\right)^T u_\delta.$$

Here indicating $\tilde{A}^h = \left(A^h\right)^T A^h;\ \tilde{u}_\delta = \left(A^h\right)^T u_\delta$, the following is found:

$$\tilde{A}^h\tilde{z} = \tilde{u}_\delta. \tag{7}$$

Stage 4. The matrix $B^\alpha = \left\{\beta^\alpha_{i,j}\right\}_{i,j=\overline{1,m}}$ is built up. The components of this matrix are detected according to the formulae

$$\beta_{i,j} = \sum_{k=1}^{m} a_{i,k}\cdot a_{j,k}, \qquad \forall j,\ j=\overline{1,m}.$$

The value α is added to the matrix components on the principal diagonal, it is the parameter of regularization.

$$B^\alpha = \begin{pmatrix} \alpha+\beta_{11} & \beta_{21} & \dots & \beta_{n1} \\ \beta_{12} & \alpha+\beta_{22} & \dots & \beta_{n2} \\ \dots & \dots & \dots & \dots \\ \beta_{1m} & \beta_{2m} & \dots & \beta_{nm} \end{pmatrix}$$

Stage 5. The regularizing parameter (matrix) $R^\alpha = \left(B^\alpha\right)^{-1}\cdot\tilde{A}^k$ is detected.

Stage 6. The approximate solution for the set of simultaneous linear algebraic equations is found.

$$\tilde{z} = R^\alpha u_\delta$$

Having the invariable value of the rate δ and changing the value of α, the optimal value is detected.

V. Conclusions

The presented paper considers the approach towards the estimation of the gas-turbine engine reliability level in the process of employing. The diagnostic matrix permits determining the control parameters describing the state of the engine. Deflection of these control parameters indicates the block or the unit where the reliability level decrease takes place. Knowing the limit values of the control parameters and their current values, it is possible to control the safety level of the engine blocks and units. The research presents the elimination of the diagnostic matrix drawbacks, restricting its

implementation; it provides the algorithm of sustained receiving the engine reversed calculated matrix. This algorithm has been obtained by employing Tikhonoff's regularisation method.

ACKNOWLEDGMENT

The article is written with the financial assistance of European Social Fund.

Project No. 009/0159/1DP/1.1.2.1.2/09/IPIA/VIAA/006 (The Support in Realization of the Doctoral Program "Telematics and Logistics" of the Transport and Telecommunication Institute).

REFERENCES

[1] Engl H.W. and Neubauer A. 1985. "Optimal discrepancy principles for the Tikhonov regularization of integral equations of the first kind." *Constructive Methods for the Practical Treatment of Integral Equations*. (G.Hämmerlin and K.-H.Hoffmann, eds.), Birkhäuser, Basel, Boston, Stuttgart, 120-141.

[2] Chercez. A. 1975. "Engineering Calculations of gas turbine engines by small deviations." Moscow, Mashinostrojenije. (in Russian)

[3] Labendik V. and Kuznetsov N. 1992. "On the development of adequate mathematical model of linear GTE for control of its flow part with diagnostic matrices." Riga, RAU. 14. (in Russian)

[4] Morozov V.A 1984. "Methods of Solving Incorrectly Posed Problems." *Springer-Verlag*, New York, Berlin Heidelberg.

[5] Novikov A.S., Paikin A.G., Sirotin N.N. 2007 "Control and Diagnostics of the Gas-Turbine Engine Performance". Moscow: Nauka, 469. (in Russian)

New approach of obtaining a stable diagnostic matrix to control the reliability level of gas turbine engine

Sergey M. Yunusov
Transport and Telecommunication Institute
Riga, Latvia
Email: yunusov@inbox.lv

Sharif E. Guseynov
*Institute of Mathematical Sciences and Information Technologies
University of Liepaja;
**Transport and Telecommunication Institute
*Liepaja, **Riga, Latvia
Email: sh.e.guseinov@inbox.lv

Shirmail G. Bagirov
Department of Mechanics and Mathematics
Baku State University
Baku, Azerbaijan
Email: sh.bagirov@yahoo.com

***Abstract*—The presented paper offers the approach of probability towards the determination of the level of reliability of the aircraft gas turbine blocks and assemblies. The described approach makes significant the choice of reference parameters imposing direct impact on the engine safety and allowing localizing the block or the assembly, showing the decrease of the reliability level. It is offered to employ the diagnostic matrix for localizing the engine component where the process of safety decline has begun. Implementing the diagnostic matrix permits to determine the moment of the very beginning of the engine parameters deviation, and allows controlling them up to the moment of the limit deviation achievement. The paper also considers the approach based on the Tikhonov's method of regularization and used for eliminating certain restrictions on implementing the diagnostic matrix.**

I. Introduction

The efficiency of the aviation equipment employment is substantially determined by the perfectness of the exploitation methods. The perfection of any method of servicing and maintenance is determined by the degree of providing the interaction between the actually existing process of changing the object technical state and the process of its technical employment. The conditioned methods of servicing and maintenance provide the closest connection between the above described processes. The principal maxim of these methods is the maxim of preventing the failures of the functional systems of the aircraft and engine under the condition of their maximum operating time and providing the sufficient level of the flight safety.

The great variety of the servicing and maintenance methods on condition can be divided for convenience into two groups: controlling the reliability level and controlling the parameters. The method of technical servicing and maintenance on condition, controlling the reliability level is the method of setting the margin of reliability for the homogeneous items. The objects are employed while the safety level is within the limits of set regulations.

The practical implementation of the method of technical servicing and maintenance on condition controlling the safety level allows decreasing the exploitation costs.

The high cost of the gas turbine engine (GTE) in comparison with the other components of the aircraft requires providing more resources, and it is achieved by implementing the optimal strategy of the engine employment till the pre-failure state of its details and blocks.

The concept of employing the aircraft engines till the pre-failure state stipulates the wide implementation of the methods and facilities providing the prognostication of the technical state of equipment. This fact distinguishes the methods of engine diagnostics compared to other technical systems.

The peculiarity of the aircraft GTE is the occurrence of close interconnection between various physical processes taking place within them. This interconnection results in complicated models, describing the diagnostic characteristics and the processes of the defects appearance and development. It is necessary to obtain the sufficient information on the physical parameters of the processes, having different nature for qualified diagnostics of the engine technical state. Moreover, there should take place the complex processing and analysis of measured diagnostic information with implementing the mathematical models, characterizing the various physical factors interconnections, considering the factors, influencing and directly predetermining the development of defects and failures.

The diagnostic models of the aircraft engines are used for forming the informative diagnostic characteristics, and for numeric imitation of the functional processes for the regularly

operating engine on the purpose of comparing the process computed characteristic parameters with the actual measured values and taking decision on the engine technical state.

The control of the safety level allows decreasing the operation costs of the engine service and improves the flights safety.

II. Formation of the diagnostic matrix for controlling the state of the gas turbine engine air gas channel

The contemporary GTE comprise a lot of components and have a complicated structure; these components interact continuously with the external environment in the process of operation, and cooperate with other sub-systems of the aircraft as well. The accurate forecast cannot be done for the further technical state of such complicated system as aircraft engine since it is impossible to present the full comprehensive description of the impact different factors can have on the engine in different situations. Accordingly, there is a necessity to take the decision on the engine state under the circumstance of indeterminacy.

Since L.A.Urban (see [1], [1]) introduced his gas path analysis in the 1970's, a substantial number of papers have been published in this area. These papers have proposed a wide variety of algorithms, employing linear (for instance, see [3], [4]) and non-linear methods (for instance, see [5]), genetic algorithms (for instance, see [6], [7]), neural networks (for instance, see [8]) and fuzzy expert systems (for instance, see [9], [10]). Comparisons of these methods were extensively reviewed in [11], [12]. Currently, gas path analysis is used in gas turbine analysis both widely and commercially. All the peculiarities and properties of the diagnostic matrices useful for practice are the results of the analysis of the mathematical model of the processes taking place in the air gas channel of the gas turbine engines. The equations describing the processes in the components of the gas turbine engine are complicated; some of them are given graphically (for example, the compressor features). The operational process of the gas turbine engine represents the aggregate of the whole range of closely interconnected complex processes, and changing even single of the gas turbine engine parameters - sectional area, loss coefficient, coefficient of efficiency etc. - results in changing almost all other parameters of the flow (pressure, temperature, velocities) and further in the changes of the engine properties - thrust, power, fuel consumption. For simplifying the analysis of dependence between the increments in the interconnected the equation parameters the approximate mathematical method is implemented - the method of minor deviations, allowing obtaining the linearization of the process original equations.

The problem solved by the diagnostics system in the operation process concerns receiving the data necessary for exact decision taking the further engine running as an object of diagnostics, and for possible maximal decreasing the existing indeterminacy. Nowadays there are different methods and diagnostic models for control and diagnostics of the gas turbine engines, but it is worth mentioning that there is no universal method or model. That is the reason why the diagnostics system comprises as a rule the whole complex of models and methods employed for determining and forecasting the engine technical state. Mathematical models as a means of forecasting the object state possess an important advantage - it is the possibility of examining the engine state under the influence of different factors under different conditions. The issue of developing the general mathematical model is identification and discovering connections between the fact of malfunction appearance and the engine parameters alteration as an object of diagnostics. The parameters under control are chosen on this purpose, then the algorithms of parameters transformation are worked out, the accuracy of measurement and the diagnostics regularity are specified, the complete association between the diagnostics results and the engine technical state are established. One of the methods of the engine air gas channel diagnostics is the method of diagnostic matrices. The essence of the method is the following: the equations in minor deviations describing the gas turbine engine become the basis for constructing and forming the diagnostic matrix $C = A^{-1}B$, where the matrix A components comprise the coefficients of the calculated parameters, and the matrix B components comprise the coefficients of measured parameters. In other words, for obtaining the diagnostic matrix the measured parameters are taken into the right hand sides of the equations and the left part comprises the others (for instance, see [4], [13]–[15] and respective references given in these). The necessary number of the measured parameters is determined as a difference between the number of variables and the number of the equations. The number of the measured parameters needed for the diagnosis is determined by the number of independent criteria of calculation, the deviations of which characterize the state of the engine junctions. The diagnostic matrix is formed for the mode of maximum operation duration (under the condition of power N=2720 kW). The left hand sides of the equations are used for matrix A creation, and the right parts - for matrix B (see [14], [15]). Let us notice that formation of the engine diagnostic matrix is described in details in the work [16]. The diagnostic matrix formation is demonstrated by setting the diagnostics of the air gas channel of the engine 7-117, for which the mathematical model was developed. It is worth mentioning that the depth of the engine air gas channel diagnostics is connected with the number of the registered fluid dynamics parameters. In this case the following parameters are supposed to be the registered (measured) ones: full temperature and air pressure at the engine inlet; speed of the compressor rotor; full gas temperature behind the turbine; full air pressure behind the compressor; fuel consumption per hour; speed of the free power turbine rotor. Finally, nevertheless let us notice that the practice has shown that the diagnostic matrix development is not always possible, that is why certain papers [17] point the restrictions of their employment.

The problem of the diagnostic matrix C construction for engines, especially complex schemes, is connected with the problem of finding the inversed calculation identification ma-

trix A. As a rule, matrix A is strongly sparsed, it is ill-conditioned matrix, and in general case the determinant of this matrix is zero. Consequently, the inverse matrix A^{-1} cannot be obtained by classical means, and this problem transits to the class of ill-conditioned problems (for instance, see [18]–[27]). In this connection, there is a problem of really sustained inversion of calculation identification matrix, but not receiving its pseudo-inversion. The point is that the method of pseudo-inversion is unable to function in the procedure of inversion of the constructed calculation identification matrix of engine; that is why it is important to work out the specific methods employing the apparatus of ill-conditioned problems theory. This problem becomes more urgent and more complicated taking into consideration the fact that the components of calculation-identification matrix A are based on the engine measurable parameters. In the process of measuring the parameters there is naturally existing equipment error; at the same time, the maximum admissible diapasons of changing the errors values are known a priori. That is why every component of calculation identification matrix is set by not a sole value, but by the interval, i.e. instead of $a_{ij}\ \ \left(i=\overline{1,\,n};\ j=\overline{1,\,m}\right)$ there is $\tilde{a}_{ij}\ \ \left(i=\overline{1,\,n};\ j=\overline{1,\,m}\right)$, which can take any value from the interval $[a_{ij}-\delta_{ij},\ \ a_{ij}+\delta_{ij}]\ \ \left(i=\overline{1,\,n};\ j=\overline{1,\,m}\right)$. Therefore, taking into consideration the errors of the engine parameters measurements, there is received not one ill-conditioned matrix A, but infinitely many ill-conditioned matrices $A(\delta)$. As consequence, there is another problem connected with the choice of the matrix, which is to be sustainable inversed, from the set of possible calculation identification matrices.

Next part of the paper under consideration proposes a new regularizing algorithm for determining the diagnostic matrix C, taking into account the sustained inversion of the engine calculation identification matrix. The idea of Tikhonov's regularization method is in the basis of the offered algorithm (for instance, see [18]–[27]). Nevertheless, as it is seen from the contest of the next part of this paper, the mentioned approach has the principal difference from the classical Tikhonov's approach towards the choice of the regularization parameter α in the Tikhonov's functional $M^{\alpha}[z,u_{\delta}]$.

It is worth mentioning that the idea and approaches presented in papers [14], [15], demonstrating the development and successful employment of the special regularizing algorithms (based on the results of [30]–[34] for adequate finding the diagnostic matrix C, also show the substantial differences with the new ideas and approaches, manifested in the next part of the paper under consideration.

III. Regulating algorithm of converting the calculation-identification matrix with empirical choice of the regularization parameter

The following operating equation is considered for the sustained conversion of the calculation-identification matrix A:

$$Az=u, \tag{1}$$

where $z\in Z$ is an unknown component; $u\in U$ is the given component; Z and U are Hilbert space; operator $A:Z\to U$ is the given linear restricted operator. As it has been said in part II, the components of the calculation-identification matrix A are based on the measured parameters of gas-turbine engine and consequently they are given with the certain inaccuracies. It is the reason to implement the approximate equation instead of the equation (1):

$$A^{\{h\}}z=u^{\{\delta\}}, \tag{2}$$

where

$$\left\|A-A^{\{h\}}\right\|\le h,\left\|u-u^{\{\delta\}}\right\|_{U}\le\delta,\delta>0,h\ge 0. \tag{3}$$

By noting $\Delta\overset{def}{=}(\delta;\,h)$, it is possible to formulate the aim as follows: it is required to find the solution $z^{\{\Delta\}}\in Z$ of the equation (2) according to $\left\{A^{\{h\}},\,u^{\{\delta\}};\ \Delta\right\}$, satisfying the conditions (3), to be sustained and satisfying the condition $\left\|z^{\{n.p.\}}-z^{\{\Delta\}}\right\|_{Z}\to 0$ under $\Delta\to 0$, where $z^{\{n.p.\}}\in Z$ means the normal pseudosolution (solution with the minimal norm in space Z) of the equation (1).

To demonstrate the substantial difference between the proposed by paper new approach towards the optimal regularization parameter α in the Tikhonov's functional $M^{\alpha}[z,u_{\delta}]$ and conventional regularizing methods there is a brief consideration of the Tikhonov's regularizing method (for instance, see [18]–[27]). In the procedure of considering this method a special attention is paid to the problem of choosing the optimal parameter of regularization employing the conventional methods (for instance, see [18]–[27], [30]–[33] and respective references given in these).

Tikhonov's regularization method implements and solves the below equation instead of the equation (2)

$$\left(A^{\{h\}}\right)^{*}A^{\{h\}}z^{\alpha}+\alpha\cdot z^{\alpha}=\left(A^{\{h\}}\right)^{*}u^{\{\delta\}}, \tag{4}$$

where $\alpha=\alpha(\Delta)>0$ is the regularization parameter; A^{*} is the operator conjugated to A. In [18]–[27], [30]–[34] and other numerous researchers present different in classification and in degree of accuracy methods of choosing optimal and/or quasi-optimal regularization parameter $\alpha=\alpha(\Delta)$, as well as estimating the error $\left\|z^{\{n.p.\}}-z^{\alpha}\right\|_{\bullet}$ of regularized solution z^{α}. All these approaches require proximity $\left\|z^{\{n.p.\}}-z^{\alpha}\right\|_{\bullet}$ to $\left\|z^{\{n.p.\}}-z^{\alpha_{\text{exact\_}optimal}}\right\|_{\bullet}=\min\limits_{\alpha(\Delta)>0}\left\|z^{\{n.p.\}}-z^{\alpha(\Delta)}\right\|_{\bullet}$ in asymptotic with $\Delta=(h;\,\delta)\to 0$, but not with final $h\ge 0$ and $\delta>0$. In other words, the conventional approaches with final h and δ present quite accessible results for the model problems, which have been specially selected for demonstrating opportunities of the certain choice of optimal and/or quasi-optimal regularization parameter. As it is shown in the book [27], employment of the conventional algorithms of choosing optimal and/or quasi-optimal regularization parameter with the final values (even at arbitrary small values) even in certain specially selected model problems, h and δ present the conservative value $\alpha(\Delta)$ compared to the accurately optimal $\alpha_{\text{exact\_}optimal}(\Delta)$ and, consequently, there at least two distortions: (a) the conservative value $\left\|z^{\{n.p.\}}-z^{\alpha}\right\|_{\bullet}$ compared to the desired value $\left\|z^{\{n.p.\}}-z^{\alpha_{\text{exact\_}optimal}}\right\|_{\bullet}$;

(b) decrease in resolvability (for instance, see [31], [32]) of Tikhonov's regularization method; it means that the conventionally obtained solution $z^{\alpha} \in Z$ is actually more smoothed comparing to the required solution $z^{\alpha_{\text{exact\_optimal}}} \in Z$. That is why there is only one conclusion: with the final values h and δ without substantial additional antecedent quantitative and qualitative assumptions about the required solution of the equation (1), it is impossible to receive the accurate optimal value of the regularization parameter employing the conventional methods $\alpha_{\text{exact\_optimal}}(\Delta)$, unless the specially selected modelled problems are analysed. The experience of solving the numerous model problems proves (for instance, see [27], [31]), the above mentioned distortions happen as soon as the relative errors of the principal operator and the right part of the equation (2) more than one percent, i.e. when the $\text{relative\_error\_}h > 1\%$ and $\text{relative\_error\_}\delta > 1\%$.

So, there is the problem (1), (3), and the following signs are introduced: $\bar{A}^{\{h\}} \stackrel{def}{=} \left(A^{\{h\}}\right)^{*} A^{\{h\}}$; $\bar{u}^{\{h;\,\delta\}} \stackrel{def}{=} \left(A^{\{h\}}\right)^{*} u^{\{\delta\}}$. Then Tikhonov's equation (4) takes the view:

$$\bar{A}^{\{h\}} z^{\alpha} + \alpha \cdot z^{\alpha} = \bar{u}^{\{h;\,\delta\}}. \tag{5}$$

Then, comparing the original approximate equation (2) and the equation (5), it is possible to conclude, that the classical Tikhonov's regularization method implement not the original equation (2), but the following one:

$$\bar{A}^{\{h\}} z^{\alpha} = \bar{u}^{\{h;\,\delta\}}, \tag{6}$$

i.e. the right part $u^{\{\delta\}} \in U$ of the initial equation (2) is not comprised in its explicit form in classical Tikhonov's regularization method, though the different variants of Residual method including the Generalized Residual principle (for instance, see [26]), employ actually error δ of the right part $u^{\{\delta\}} \in U$, but not the error $\bar{u}^{\{h;\,\delta\}} \in U$, of the right part, which depends not only on δ, but also on h as it is seen in (6). Consequently, accidental errors in $u^{\{\delta\}} \in U$ can be significantly smoothed that is why the relative error $\bar{u}^{\{h;\,\delta\}} \in U$, which is not taken into account in the classical Tikhonov's method of regularization, can be substantially (by several digits) different compared to the relative error $u^{\{\delta\}} \in U$ which is the only one considered in the classical Tikhonov's regularization method. This fact becomes obvious if the equation (2) employs, for instance, Fredholm integral operator functioning from $L_2[a,\,b]$ into $L_2[a,\,b]$, i.e. $A^{\{h\}}[\bullet] \stackrel{def}{=} \int\limits_a^b K^{\{h\}}(x,\,y) \cdot [\bullet]\, dy$: as a principal operator $A^{\{h\}}$; in this case $\bar{u}^{\{h;\,\delta\}}(x) = \int\limits_a^b K^{\{h\}}(x,\,y) \cdot u^{\{\delta\}}(y)\, dy$ and accordingly, since the integration operation is the smoothing filter, according to function $u^{\{\delta\}}(y) \in L_2[a,\,b]$ there is a rather smoothed function $\bar{u}^{\{h;\,\delta\}}(x) \in L_2[a,\,b]$. The above described way by smoothing the accidental errors is attributed to the equation (6) right part in the same degree, more specifically, the principal operator $\bar{A}^{\{h\}}$, in the left part of this equation, and its error can be significantly different from the error of the principal operator $A^{\{h\}}$ in the original approximate equation (2), though the classical Tikhonov's regularization method takes into account exactly operator $A^{\{h\}}$, error, but not an error of the actual (i.e. really existing in the equation (6)) operator $\bar{A}^{\{h\}} = \left(A^{\{h\}}\right)^{*} A^{\{h\}}$.

Summarizing all the above said, there is a conclusion that it is necessary to employ not errors $\Delta = (\delta;\, h)$ of the original data $\left\{A^{\{h\}};\, u^{\{\delta\}}\right\}$ in problems (2), (3) in Tikhonov's regularization method, but errors $\bar{\Delta}$ of the original data $\left\{\bar{A}^{\{h\}};\, \bar{u}^{\{h;\,\delta\}}\right\}$. Consequently, there is proposal to consider the equation (6) instead of the original problem (2)), (3)), and at the same time, take into account the errors of the actual original data of this equation in Tikhonov's regularization method, in other words the errors of the principal operator $\bar{A}^{\{h\}} \stackrel{def}{=} \left(A^{\{h\}}\right)^{*} A^{\{h\}}$ and the errors of the component $\bar{u}^{\{h;\,\delta\}} \stackrel{def}{=} \left(A^{\{h\}}\right)^{*} u^{\{\delta\}}$. This idea, as it follows from the below explanation, generates the principal difference in the methodology of choosing the optimal and/or quasi-optimal regularization parameter. There is this methodology. According to the summarised residual maxim (for instance, see [26] as well as [20], [21]), the regularization parameter $\alpha = \alpha(\Delta) > 0$ is the equation root:

$$\begin{aligned} &\left\|A^{\{h\}} z^{\alpha} - u^{\{\delta\}}\right\|_U = \left(\delta + h \cdot \|z^{\alpha}\|_Z\right)^2 \\ &+ \left(\inf_{z \in Z} \left\|A^{\{h\}} z - u^{\{\delta\}}\right\|_U\right)^2, \end{aligned} \tag{7}$$

where $\inf\limits_{z \in Z} \left\|A^{\{h\}} z - u^{\{\delta\}}\right\|_U$ is the measure of incompatibility of the original problem (2), (3). As it is presented in paper [21] (also see [22], [23]), the regularization parameter $\alpha = \alpha(\Delta) > 0$ is the equation root:

$$\begin{aligned} &\alpha^{k} \cdot \left\|\left(A^{\{h\}}\right)^{*} A^{\{h\}} z^{\alpha} - \left(A^{\{h\}}\right)^{*} u^{\{\delta\}}\right\|_U \\ &= \lambda \cdot \left\|A^{\{h\}}\right\| \cdot \left(\delta + h \cdot \|z^{\alpha}\|_Z\right), \end{aligned}$$

where $k \geq 0$ and $\lambda > 0$ are the certain constants.

Consequently, taking into account the equation (4), there is:

$$\alpha^{k+1} \cdot \|z^{\alpha}\|_Z = \lambda \cdot \left\|A^{\{h\}}\right\| \cdot \left(\delta + h \cdot \|z^{\alpha}\|_Z\right). \tag{8}$$

For discovering the optimal regularization parameter this paper offers using not the equation (8) but the following equation, the root of which is the required regularization parameter:

$$\begin{aligned} &\|z^{\alpha}\|_Z \cdot \left(\alpha^{k+1} - \lambda \cdot \sup_{h \geq 0} \bar{A}^{\{h\}}\right) \\ &= \lambda \cdot \sup_{h \geq 0,\, \delta > 0} \bar{u}^{\{h;\,\delta\}}, \;\; k \geq 0, \;\; \lambda > 0, \end{aligned} \tag{9}$$

where $\bar{A}^{def} \equiv A^{*} A$; $\bar{u}^{def} \equiv A^{*} u$.

The principal difference between the equations (8) and (9) is the fact that the value of "new" variable $\Sigma_{New} \stackrel{def}{=} \sup\limits_{h \geq 0,\, \delta > 0} \bar{u}^{\{h;\,\delta\}} + \|z^{\alpha}\|_Z \cdot \sup\limits_{h \geq 0} \bar{A}^{\{h\}}$ in the equation (9) cannot exceed the value $\Sigma_{Old} \stackrel{def}{=} \left\|A^{\{h\}}\right\| \cdot \left(\delta + h \cdot \|z^{\alpha}\|_Z\right)$ in the equation (8), i.e. $\Sigma_{New} \leq \Sigma_{Old}$, moreover, this inequality can be absolutely strict. This guarantees that the

root $\alpha_{optimal} \stackrel{def}{=} \alpha_{root} > 0$ of the equation (9) and the residual $\left\|z^{\{n.p.\}} - z^{\alpha_{\alpha_{optimal}}}\right\|_Z$ are not overrated.

Then, comparison of the equations (8) and (9) demonstrates that Tikhonov's regularization method employs the system of considering the errors of actual original data of the equation (6). The heart of this idea is the fact that considering the error of principal operator $\bar{A}^{\{h\}}$ of the equation (6) and error of the right part of the component $\bar{u}^{\{h;\,\delta\}}$ of the equation (6) provides the equality to zero of the incompatibility measure of the equation (6), i.e. $\inf\limits_{z\in Z}\left\|\bar{A}^{\{h\}}z - \bar{u}^{\{h;\,\delta\}}\right\|_U = 0$. This fact is substantial, it presents the principal difference between the offered the equation (9) and the equation (7) of the summarized residual principle.

There is the most important issue if the obtained root $\alpha_{root} > 0$ of the offered equation (9) generates the required regularizing operator? If the answer is positive, it arises two other issues: (a) under which conditions the solution of offered equation (9) exists and is the only? (b) what order of convergence of the obtained regularizing solution $z^{\alpha_{optimal}} = z^{\alpha_{root}}$ to the normal pseudo-solution $z^{\{n.p.\}}$? To answer these important questions it is necessary to set preliminary certain estimations for both root $\alpha_{root} > 0$ of the equation (9), and for residual $\left\|z^{\{n.p.\}} - z^{\alpha_{root}}\right\|_Z$. It is easy to show, that the fulfilment of the conditions

$$\begin{cases} \dfrac{\sup\limits_{h\geq 0,\,\delta>0}\left\|\bar{u}-\bar{u}^{\{h;\,\delta\}}\right\|_U}{\left\|\bar{u}^{\{h;\,\delta\}}\right\|_U} < \frac{1}{\lambda} \quad if \quad k=0; \\ \left\|\bar{u}^{\{h;\,\delta\}}\right\|_U \neq 0 \qquad if \quad k>0 \end{cases} \tag{10}$$

the function $\|z^\alpha\|_Z\alpha^{k+1}$ as a function of argument α is a monotonically increasing continuous function on semiaxis $(0,\,+\infty)$, and function $\lambda\left(\|z^\alpha\|_Z \sup\limits_{h\geq 0}\left\|\bar{A}-\bar{A}^{\{h\}}\right\| + \sup\limits_{h\geq 0,\,\delta>0}\left\|\bar{u}-\bar{u}^{\{h;\,\delta\}}\right\|_U\right)$ as a function of argument is a monotonically decreasing continuous function on the same semi-axis. Moreover, under the condition (10) there are following asymptotics:
$\|z^\alpha\|_Z\alpha^{k+1} \to +\infty$ at $\alpha\to+\infty$, if $k>0$, $\left\|\bar{u}^{\{h;\,\delta\}}\right\|_U \neq 0$;
$\|z^\alpha\|_Z\alpha^{k+1} \to \left\|\bar{u}^{\{h;\,\delta\}}\right\|_U$ at $\alpha\to+\infty$, if $k=0$;
$\|z^\alpha\|_Z\alpha^{k+1} \to 0$, if $k>0$, $\left\|\bar{u}^{\{h;\,\delta\}}\right\|_U = 0$;
$\|z^\alpha\|_Z\alpha^{k+1} \to 0$ at $\alpha\to 0+0$;
$\|z^\alpha\|_Z \sup\limits_{h\geq 0}\left\|A^*A-\bar{A}^{\{h\}}\right\| + \sup\limits_{h\geq 0,\,\delta>0}\left\|A^*u-\bar{u}^{\{h;\,\delta\}}\right\|_U \to \sup\limits_{h\geq 0,\,\delta>0}\left\|A^*u-\bar{u}^{\{h;\,\delta\}}\right\|_U$ at $\alpha\to+\infty$.

Then, since the offered equation (9) is equivalent to the equation

$$\|z^\alpha\|_Z\cdot\alpha^{k+1} = \lambda\cdot\left(\|z^\alpha\|_Z\cdot\sup_{h\geq 0}\left\|\bar{A}-\bar{A}^{\{h\}}\right\| + \sup_{h\geq 0,\,\delta>0}\left\|\bar{u}-\bar{u}^{\{h;\,\delta\}}\right\|_U\right). \tag{11}$$

where $k\geq 0$; $\lambda>0$, then the foregoing asymptotic estimates, together with the above-mentioned properties of strictly monotonicity and continuity of the functions $\|z^\alpha\|_Z\cdot\alpha^{k+1}$ and $\lambda\cdot\left(\|z^\alpha\|_Z\cdot\sup\limits_{h\geq 0}\left\|\bar{A}-\bar{A}^{\{h\}}\right\| + \sup\limits_{h\geq 0,\,\delta>0}\left\|\bar{u}-\bar{u}^{\{h;\,\delta\}}\right\|_U\right)$, which, as seen from (11), are correspondingly left and right parts of this equation. It makes possible to state the following important fact by virtue of the Brouwer-Schauder Fixed Point Theorem (for instance, see [35]): under fulfilling the conditions (10) the equation (11) (accordingly, its equivalent equation (9) also) has the only fixed point, i.e. the equation (9) has the only root, and this only root is taken as an optimal regularization parameter in Tikhonov's regularization method, or in Tikhonov's equation (5). There is a comprehensive answer for the above stated question, but there is no answer for the main question - Does the obtained unique root of the equation (9) generate the required regularizing operator? To answer this main question, first it is necessary to demonstrate several upper estimates, truth of which is easily stated under condition of employing the facts, that the limited linear operators A and B, reflecting Banach space Z to Banach space U, have true facts that $\|A\| = \|A^*\|$ and $\|AB\| \leq \|A\|\cdot\|B\|$:

$$\begin{cases} \sup\limits_{h\geq 0}\left\|\bar{A}-\bar{A}^{\{h\}}\right\| \leq 2\cdot\left\|A^{\{h\}}\right\|\cdot h; \\ \sup\limits_{h\geq 0,\,\delta>0}\left\|\bar{u}-\bar{u}^{\{h;\,\delta\}}\right\|_U \leq \left\|A^{\{h\}}\right\|\cdot\delta + \left\|u^{\{\delta\}}\right\|_U\cdot h. \end{cases} \tag{12}$$

Then there is assumption that

$$\frac{\sup\limits_{h\geq 0,\,\delta>0}\left\|\bar{u}-\bar{u}^{\{h;\,\delta\}}\right\|_U}{\left\|\bar{u}^{\{h;\,\delta\}}\right\|_U} \leq \frac{\left\|\bar{A}^{\{h\}}\right\|^k}{2\cdot\lambda} - \frac{\sup\limits_{h\geq 0}\left\|\bar{A}-\bar{A}^{\{h\}}\right\|}{\left\|\bar{A}^{\{h\}}\right\|}, \tag{13}$$

which actually is a summary of the first inequality of (10). Implementation of condition (13) guarantees the truthfulness of the following highly useful (especially in final errors $\Delta = (h;\,\delta)$ of original data of original problem (2), (3)) upper estimate for the root (as it has been proven, the only root) $\alpha_{root} > 0$ of the equation (9):

$$\alpha_{root}^{k+1} \leq \left(\sup_{h\geq 0,\,\delta>0}\left\|\bar{u}-\bar{u}^{\{h;\,\delta\}}\right\|_U + \sup_{h\geq 0}\left\|\bar{A}-\bar{A}^{\{h\}}\right\|\right) \times\left(\lambda\cdot\max\left\{1,\ \frac{2\cdot\left\|\bar{A}^{\{h\}}\right\|}{\left\|\bar{u}^{\{h;\,\delta\}}\right\|_U}\right\}\right), \tag{14}$$

from which under condition of $\sup\limits_{h\geq 0}\left\|\bar{A}-\bar{A}^{\{h\}}\right\| \to 0$, $\sup\limits_{h\geq 0,\,\delta>0}\left\|\bar{u}-\bar{u}^{\{h;\,\delta\}}\right\|_U \to 0$ there is direct asymptotic estimate $\alpha_{root} = O\left(\sup\limits_{h\geq 0,\,\delta>0}\left\|\bar{u}-\bar{u}^{\{h;\,\delta\}}\right\|_U + \sup\limits_{h\geq 0}\left\|\bar{A}-\bar{A}^{\{h\}}\right\|\right)^{\frac{1}{k+1}}$, which is less useful (rough, allowing overstating the values of the optimal regularization parameter) for solving the actual problem of determining the diagnostic matrix for sustained obtaining the calculation identification parameters of the gas turbine engine.

At this point, taking into account the above received results (exactly the equation (9); upper estimates (12) for the errors of the principal operator and the first part of the equation (6); the

conditions (10) and (13); upper estimate (14) for the equation (9) root) it is possible to estimate the error of solution $z^{\alpha_{root}}$ of Tikhonov's equation (5), where the choice of optimal and/or quasi-optimal regularization parameter is implemented not only via the conventional approaches, but also via the solution of offered and grounded equation (9). There is estimation of residual $\left\|z^{\{n.p.\}} - z^{\alpha_{root}}\right\|_Z \equiv \left\|z^{\{n.p.\}} - z^{\alpha_{optimal}}\right\|_Z$. On this purpose, together with the above obtained results, there are implemented the following results, obtained in researches [21]–[23]:

$$\begin{aligned}\left\|z^{\{n.p.\}} - z^{\alpha}\right\|_Z \le \alpha^{-1}\cdot\Bigg(\sup_h \left\|A^*A - \bar{A}^{\{h\}}\right\| \\ + \sup_{h,\,\delta}\left\|A^*u - \bar{u}^{\{h;\,\delta\}}\right\|_U\Bigg)\cdot\max\left\{1, \left\|z^{\{n.p.\}}\right\|_Z\right\} \\ +\alpha\cdot\left\|\left(\alpha\cdot E + \left(A^{\{h\}}\right)^* A^{\{h\}}\right)^{-1} z^{\{n.p.\}}\right\|_U,\end{aligned} \tag{15}$$

where E indicates the unit operator. The papers [21]–[23] also contain the following three useful estimates, determining the operator $\left(\alpha\cdot E + \left(A^{\{h\}}\right)^* A^{\{h\}}\right)^{-1}$ restriction:
$\left\|\left(\alpha\cdot E + \left(A^{\{h\}}\right)^* A^{\{h\}}\right)^{-1}\left(A^{\{h\}}\right)^* A^{\{h\}}\right\|_U \le 1;$
$\left\|\left(\alpha\cdot E + \left(A^{\{h\}}\right)^* A^{\{h\}}\right)^{-1}\left(A^{\{h\}}\right)^*\right\|_U \le \frac{1}{2\cdot\sqrt{\alpha}};$
$\left\|\left(\alpha\cdot E + \left(A^{\{h\}}\right)^* A^{\{h\}}\right)^{-1}\right\|_U \le \frac{1}{\alpha}.$

It is important the first two estimates can be successfully applied to residual (15). So, together with the above obtained results, there is implementation of estimate (15); there is the estimate of norm:

$$\begin{aligned}&\left\|\left(A^{\{h\}}\right)^* A^{\{h\}}\left(z^{\{n.p.\}} - z^{\alpha_{root}}\right)\right\|_U \\ &= \Bigg\|\left(A^{\{h\}}\right)^* A^{\{h\}}\left(\alpha_{root}\cdot E + \left(A^{\{h\}}\right)^* A^{\{h\}}\right)^{-1} \\ &\times\Big\{\left(A^{\{h\}}\right)^* u^{\{\delta\}} - A^* u\Big\} + \left(A^{\{h\}}\right)^* A^{\{h\}} \\ &\times\left(\alpha_{root}\cdot E + \left(A^{\{h\}}\right)^* A^{\{h\}}\right)^{-1}\{A^*A \\ &-\left(A^{\{h\}}\right)^* A^{\{h\}}\Big\} z^{\alpha_{root}} - \alpha_{root}\cdot\left(A^{\{h\}}\right)^* A^{\{h\}} \\ &\times\left(\alpha_{root}\cdot E + \left(A^{\{h\}}\right)^* A^{\{h\}}\right)^{-1} z^{\alpha_{root}}\Bigg\|_U \\ &\le \left\|\left(\alpha_{root}\cdot E + \left(A^{\{h\}}\right)^* A^{\{h\}}\right)^{-1}\left(A^{\{h\}}\right)^* A^{\{h\}}\right\| \\ &\times\left\|\left(A^{\{h\}}\right)^* u^{\{\delta\}} - A^* u\right\|_U + \left\|\left(A^{\{h\}}\right)^* A^{\{h\}} - A^*A\right\| \\ &\times\left\|\left(\alpha_{root}\cdot E + \left(A^{\{h\}}\right)^* A^{\{h\}}\right)^{-1}\left(A^{\{h\}}\right)^* A^{\{h\}}\right\| \\ &\times\|z^{\alpha_{root}}\|_Z \le \Bigg(\sup_{h\ge 0}\left\|A^*A - \bar{A}^{\{h\}}\right\| \\ &+ \sup_{h\ge 0,\,\delta>0}\left\|A^*u - \bar{u}^{\{h;\,\delta\}}\right\|_U\Bigg)\max\left\{1,\ \left\|z^{\{n.p.\}}\right\|_Z\right\} \\ &+\alpha_{root}\cdot\|z^{\alpha_{root}}\|_Z.\end{aligned}$$

Thus, there is the following upper estimate:

$$\begin{aligned}&\left\|\left(A^{\{h\}}\right)^* A^{\{h\}}\left(z^{\{n.p.\}} - z^{\alpha_{root}}\right)\right\|_U \\ &\le\left(\sup_{h\ge 0}\left\|A^*A - \bar{A}^{\{h\}}\right\| + \sup_{h\ge 0,\,\delta>0}\left\|A^*u - \bar{u}^{\{h;\,\delta\}}\right\|_U\right) \\ &\times\max\left\{1,\ \left\|z^{\{n.p.\}}\right\|_Z\right\} + \alpha_{root}\cdot\|z^{\alpha_{root}}\|_Z.\end{aligned} \tag{16}$$

The obtained inequality (16) allows answering positively the above-stated question: the offered and grounded empirical choice of the optimal regularization parameter $\alpha_{optimal}$ as solution α_{root} of the equation (9) generates Tikhonov's regularizing operator. Actually, taking into account the inequalities (12) in the obtained upper estimate (16), and then in the obtained inequality transiting to the limit and if $\Delta = (h;\,\delta) \to 0$, there is $\left\|z^{\{n.p.\}} - z^{\alpha_{root}}\right\|_Z = O(\delta + h)^{\frac{1}{k+1}}$, This fact proves the convergence on norm (strong convergence!) of the regularized solution $z^{\alpha_{optimal}} \equiv z^{\alpha_{root}}$, received from Tikhonov's equation (5), where the only root of the equation (9) is taken as an optimal regularization parameter $\alpha_{optimal}$, to the normal pseudo-solution $z^{\{n.p.\}}$. Moreover, there is an upper estimate (16) for the error of the received regularized solution, and the upper estimate (14) is true for the optimal regularization parameter.

IV. Conclusion

The presented paper suggests and grounds the new approach towards constructing the regularizing algorithm of determining the diagnostic matrix for sustained detection of the calculation identification parameters of bypass GTE. The idea of Tikhonov's regularization method lies in the basis of the suggested algorithm; nevertheless, distinguishing from the classical Tikhonov's approach the choice of optimal parameter and/or regularization parameter is accomplished by fundamentally different method, the heart of which lies in employing the errors of actual original data in Tikhonov's equation. Implementation of this method (a) provides the equity to zero the measures of incompatibility of the considered the first kind operator equation (this fact is substantial, and it distinguishes fundamentally the suggested approach from the generalized residual method); (b) allows constructing the equation for the regularization parameter. The research has discovered the conditions implementation of which provides the unique existence of the root of suggested equation for determination of the regularization parameter. It is proven that the implementation of the regularizing operator generates the unique root of this equation as an optimal parameter of regularization.

Moreover, the investigation also provides the estimations for both the obtained optimal regularization parameter and residual between the regularized solution and normal pseudo-solution of the first kind initial operator equation. The order of convergence of the obtained regularizing solution to the normal pseudo-solution is determined by application of the obtained values.

ACKNOWLEDGMENT

The authors would like to express their gratitude to Dr. Kojdecki Marek Andrzej from the Institute of Mathematics and Cryptology, Military University of Technology, Warsaw, Poland and Professor Sizikov Valery Sergeevich from the Department of Measurement Technologies and Computer Tomography, St.Petersburg State University of Information Technologies, Mechanics and Optics, St. Petersburg, Russian Federation for the opportunity to get acquainted with their significant results obtained in [21]–[24].

Present article was executed within the framework of the following two International Projects (Grants): (1) European Social Fund Project No. 009/0159/1DP/1.1.2.1.2/09/IPIA/VIAA/006 (for the first author resounding support in Realization of the Doctoral Program "Telematics and Logistics" of the Transport and Telecommunication Institute, Riga, Latvia); (2) The European Social Fund (ESF) Project No. 1DP/1.1.1.2.0/09/APIA/VIAA/142 *"Elaboration of effective analytical and numerical methods for solving of direct and inverse mathematical physics problems in material, economics and environment sciences"* (for the second author).

REFERENCES

[1] L. A. Urban, *Parameters election for multiple fault diagnostics of gas turbine engines*, Proceedings of the AGARD Conference, Liege, Belgium, 1974.

[2] L. A. Urban, *Gas path analysis of commercial aircraft engines*, Proceedings of the 11th Symposium on Aircraft Integrated Data Systems Cologne, West Germany, 1982.

[3] D. L. Doel, *TEMPER - a gas-path analysis tool for commercial jet engines*, Journal of Engineering for Gas Turbines and Power, Vol. 116, Issue 1, 1994, pp. 82-89.

[4] D. Xia, Y. Wang and Sh. Weng, *A new method to evaluate the influence coefficient matrix for gas path analysis*, Journal of Mechanical Sciences and Technology, Vol. 23, Issue 3, 2009, pp. 667-676.

[5] A. Stamatis, K. Mathioudakis and M. Smith, *Gas turbine component fault identification by means of adaptive performance modeling*, Proceedings of the ASME Turbo Expo, Brussels, Belgium, 1990.

[6] M. Zedda and R. Singh, *Gas-turbine engine and sensor diagnostics*, Proceedings of the XIV International Symposium on air-breathing engines, Florence, Italy, 1999, pp. 255-261.

[7] M. Zedda and R. Singh, *Gas turbin eengine and sensor fault diagnosis using optimization techniques*, Journal of Propulsion and Power, Vol. 18, Issue 5, 2002, pp. 1019-1025.

[8] G. Denny, *F-16 Jet Engine Trending and Diagnostics with Neural Networks*, Applications of Neural Networks, Vol. 4, 1965, pp. 419-422.

[9] P. Fuster, A. Ligeza and M. J. Aguilar, *Adductive diagnostic procedure based on an AND/OR/NOT graph for expected behaviour: Application to a gas turbine*, Proceedings of the 10th International Congress and Exhibition on Condition Monitoring and Diagnostic Engineering Management, Helsinki, Finland, 1997.

[10] C. Siu, Q. Shen and R. Milne, *A Fuzzy Rule- and Case-based Expert System for Turbomachinery Diagnosis*, Proceedings of the IFAC Symposium on Fault Detection, Supervision and Safety for Technical Processes, Vol. 1, Kingston Upon Hull, UK, 1997, pp. 556-563.

[11] Y. G. Li, *Performance-analysis-based gas turbine diagnostics: A review*, Journal of Power and Energy, Vol. 216, Issue 5, 2002, pp. 363-377.

[12] P. Kamboukos and K. Mathioudakis, *Comparison of linear and nonlinear gas turbine performance diagnostics*, Journal of Engineering for Gas Turbines and Power, Vol. 127, Issue 1, 2005, pp. 49-56.

[13] A. Chercez, *Engineering Calculations of gas turbine engines by small deviations*, Moscow: "Engineering Industry" Publishing House, 1975, 354 p.

[14] Sh. E. Guseynov and S. M. Yunusov, *New regularizing approach to determining the influence coefficient matrix for gas-turbine engines*, In Book: Dynamical Systems, Differential Equations and Applications, Volume I, Published by the American Institute of Mathematical Sciences (AIMS), ISBN: 978-1-60133-007-9, 2011, pp. 614-623, www.aimSciences.org

[15] S. M. Yunusov and Sh. E. Guseynov, *New approach to the formation of the adequate diagnostic matrix of the gas turbine engine*, Proceedings of the 25th European Conference on Modelling and Simulation, June 7-10, 2011, Krakow, Poland, pp. 362-369.

[16] V. Labendik and N. Kuznetsov, *On the development of adequate mathematical model of linear GTE for control of its flow part with diagnostic matrices*, Riga, RAU, 1992, 14 p.

[17] A. S. Novikov, A. G. Paikin and N. N. Sirotin, *Control and Diagnostics of the Gas-Turbine Engine Performance*, Moscow: "Science" Publishing House, 2007, 469 p.

[18] A. N. Tikhonov and V. Ya. *Arsenin, Solutions of Ill-Posed Problems*, New York: Wiley Publishing House, 1977, xiii+258 p.

[19] H. W. Engl and A. Neubauer, *Optimal discrepancy principles for the Tikhonov regularization of the first kind integral equations*, In Book: Constructive Methods for the Practical Treatment of Integral Equations (eds. G. Hammerlin and K. -H. Homann), Basel, Boston, Stuttgart: Birkhouser Publishing House, 1985, pp. 120-141.

[20] F. Bauer and M. A. Lukas, *Comparing parameter choice methods for regularization of ill-posed problems*, Journal of Mathematics and Computers in Simulation, Vol. 81, Issue 9, 2011, pp. 1795-1841.

[21] M. A. Kojdecki, *New criterion of regularization parameter choice in Tikhonov method*, Doctoral Dissertation, Institute of Mathematics of Polish Academy of Science, Warsaw, 1996, 127 p.

[22] M. A. Kojdecki, *New criterion of regularization parameter choice in Tikhonov method*, Bulletin of the Military University of Technology, Warsaw, Poland, Vol. 49, Issue 1, 2000, pp. 47-126.

[23] M. A. Kojdecki, *Examples of Saturated Convergence Rates for Tikhonov Regularization*, BIT Numerical Mathematics, Vol. 41, Issue 5, 2001, pp. 1059-1068, http://www.springerlink.com/content/t7762tj471l28730/

[24] V. S. Sizikov, *On methods of residual in the solution of ill-posed problems*, Journal of Computational Mathematics and Mathematical Physics, Vol. 23, Issue 9, 2003, pp. 1294-1312.

[25] A. G. Ramm, *Inverse Problems: Mathematical and Analytical Techniques with Applications to Engineering*, Boston: Springer Science Publishing House, 2005, xx+442 p.

[26] V. A. Morozov, *Methods of Solving Incorrectly Posed Problems*, New York, Berlin Heidelberg: Springer-Verlag Publishing House, 1984, xviii+257 p.

[27] A. F. Verlan and V. S. Sizikov, *Integral Equations: Methods, Algorithms, Computer Program*, Kyiv: "Naukova Dumka" , 1986, 544 p.

[28] E. N. Moore, *On the reciprocal of the general algebraic matrix*, Bulletin of the American Mathematical Society, Vol. 26, Issue 9, 1920, pp. 394-395.

[29] R. Penrose, *A generalized inverse for matrices*, Proceedings of the Cambridge Philosophical Society, Vol. 51, 1955, pp. 406-413.

[30] Sh. E. Guseynov and I. Volodko, *Convergence order of one regularization method*, Journal of Mathematical Modelling and Analysis, Vol. 8, Issue 1, 200), pp. 25-32.

[31] Sh. E. Guseynov and M. Okruzhnova, *Choice of a quasi-optimal regularization parameter for the first kind operator equations*, Journal of Transport and Telecommunication, Vol. 6, Issue 3, 2005, pp. 471-486.

[32] V. I. Dmitriev and Sh. E. Guseynov, *On conformance of the resolving power and the solution detailedness in an inverse problems*, Herald of the Moscow State University, Series 15: "Computational Mathematics and Cybernetics", Vol. 1, 1995, pp. 17-25.

[33] Sh. E. Guseynov, *Determination of the quasi-optimal regularization parameter for the solution of operator equation of the first kind*, Proceedings of the First International Conference on Computational Methods in Applied Mathematics, Minsk, Belarus, 2003, 25 p.

[34] Sh. E. Guseynov, *Conservative averaging method for solutions of inverse problems of mathematical physics*, In Book: Progress in Industrial Mathematics", Springer-Verlag Publishing House, 2004, pp. 241-246.

[35] V. C. L. Hutson and J. S. Pym, *Applications of Functional Analysis and Operator Theory*, London-New York-Toronto-Sydney-San Francisco: Academic Press, 1980, 404 p.

Robust Estimation for Longitudinal Data with Informative Observation Times

Xingqiu Zhao

Department of Applied Mathematics, The Hong Kong Polytechnic University, Hong Kong

xingqiu.zhao@polyu.edu.hk

Abstract. This article discusses regression analysis of longitudinal data that often occur in medical follow-up studies and observational investigations. The analysis of these data involves two processes. One is the underlying recurrent event process of interest and the other is the observation process that controls observation times. Most of the existing methods, however, rely on some restrictive models or assumptions such as the Poisson assumption. For this, we propose a more general and robust estimation approach for regression analysis of logitudinal data with related observation times. The asymptotic properties of the proposed estimates are established and numerical studies indicate that it works well for practical situations.

KEY WORDS: Estimating equation; Informative observation process; Longitudinal data; Recurrent event process; Robust estimation

1. Introduction

The analysis of longitudinal data has recently attracted considerable attention. These data frequently occur in medical follow-up studies and observational investigations. For the analysis of longitudinal data, a number of methods have been developed, mostly under the assumption that the longitudinal response process and the observation process are independent completely, or given covariates. For example, Diggle et al. (1994) presented an excellent summary about such commonly used methods as estimating equation and random-effect model approaches, and Lin and Ying (2001) and Welsh et al. (2002) discussed general semiparametric regression analysis of longitudinal data when both observation times and the censoring times may depend on covariates.

A common situation where informative observation times occur is that these times are

subject or response variable-dependent. For example, they may be hospitalization times of subjects in the study (Wang et al., 2001). In a bladder cancer study, Sun and Wei (2000) and Zhang (2002) discussed a set of longitudinal data arising from a bladder cancer follow-up study conducted by the Veterans Administration Cooperative Urological Research Group; in this study, some patients had significantly more clinical visits than others and thus the occurrence of bladder tumors of a patient and the visit times may be related. Lipsitz et al. (2002) presented a set of longitudinal data from a study of children with acute lymphoblastic leukemia that involved correlated response and observation processes. The same could be true for other medical follow-up studies, but there is limited research on the analysis of longitudinal data when the longitudinal response process of interest may be correlated with the observation process given covariates, that is, the observation times may be informative. Sun et al. (2005) studied semiparametric models that allow observation times to be correlated with the longitudinal process; Sun et al. (2007) proposed a joint model for the longitudinal process and the observation process, where both processes may be correlated through a shared latent variable or frailty, and used the estimating equation approach to estimate the regression parameters; Liang et al. (2009) discussed a joint model through two random effects, where the relationship between the random effects is specified and a parametric distribution assumption for a random effect is required. A common and key assumption of these methods is that they assumed that the observation process is a Poisson process. The aim of this paper is to consider more general joint models for longitudinal data with dependent observation times, to develop an estimating equation approach for estimation of regression parameters, and to establish the asymptotic properties of the resulting estimates.

2. Models

Consider a longitudianl study that consists of n independent subjects and let $Y_i(t)$ denote the longitudinal response variable of interest before or at time t for subject i. Suppose that for each subject, there exists a p-dimensional vector of covariates denoted by X_i. Given X_i and an unobserved positive random variable Z_i that is independent of X_i, the mean function of $Y_i(t)$ has the form

$$E\{Y_i(t)|X_i, Z_i\} = \mu_0(t) + X_i'\beta + g(Z_i) \quad (1)$$

Here, $\mu_0(t)$ is a completely unknown continuous baseline mean function, β is a vector of unknown regression parameters, and $g(\cdot)$ is a completely unspecified link function.

For subject i, suppose that $Y_i(\cdot)$ is observed only at finite time points $T_{i1} < \cdots < T_{iK_i}$, where K_i denotes the potential number of observation times, $i = 1, \ldots, n$. That is, only the values of $Y_i(t)$ at these observation times are known and we have panel count data on the $Y_i(t)$'s. Let C_i denote the follow-up time associated with subject i and thus $Y_i(t)$ is observed only at these T_{ij}'s with $T_{ij} \leq C_i$, $i = 1, \ldots, n$. Define $\tilde{O}_i(t) = O_i(\min(t, C_i))$, where $O_i(t) = \sum_{j=1}^{K_i} I(T_{ij} \leq t)$, $i = 1, \ldots, n$. Then $\tilde{O}_i(t)$ is a point process characterizing the ith subject's observation process and jumps only at the observation times.

For the observation process, we will assume that $O_i(t)$ satisfies the following rate function model

$$E\{dO_i(t)|X_i, Z_i\} = Z_i\, h(X_i)\, d\Lambda_0(t)\,, \tag{2}$$

where $h(\cdot)$ is a completely unspecified positive function as g and $\Lambda_0(\cdot)$ is a completely unknown continuous baseline function. Under model (2), one does not need Poisson assumption anymore. In the following, it will be assumed that given (X_i, Z_i), $Y_i(t)$ and $O_i(t)$ are independent. Also C_i is independent of $\{Y_i, O_i, X_i, Z_i\}$ and $\{\, Y_i(t), O_i(t), C_i, X_i, 0 \leq t \leq \tau \,\}_{i=1}^n$ are independent and identically distributed, where τ denotes the length of the study. Suppose that the main goal is to estimate regression parameter β.

3. Inference procedure

To estimate β, note that if the latent variables Z_i's are known, model (1) would become the usual linear mean model. Unfortunately, the Z_i's are unknown in practice. One natural way for this is to estimate the Z_i's first and then treat them known. In the following, we take a different approach motivated by that proposed in Sun and Wei (2000) among others.

Specifically, define

$$\bar{Y}_i = \sum_{j=1}^{m_i} Y_i(T_{ij}) I(T_{ij} \leq C_i) = \int_0^\tau Y_i(t) d\tilde{O}_i(t),$$

where $m_i = \tilde{O}_i(C_i)$, the total number of observations on subject i, $i = 1, ..., n$. Then, we

have

$$E\left(\bar{Y}_i|X_i\right) = E(Z_i)E(\Lambda_0(C_i))h(X_i)(\beta' X_i)+h(X_i)\int_0^\tau [E(Z_i)\mu_0(t)+E\{g(Z_i)Z_i\}]P(C_i \geq t)d\Lambda_0(t)$$

and

$$E(m_i|X_i) = E(Z_i)E\{\Lambda_0(C_i)\}h(X_i)\,.$$

These yield

$$E(\bar{Y}_i|X_i) = E(m_i|X_i)(\beta' X_i + \theta),$$

where

$$\theta = \frac{\int_0^\tau \{E(Z_i)\mu_0(t) + E(g(Z_i)Z_i)\}P(C_i \geq t)d\Lambda_0(t)}{E(Z_i)E\{\Lambda_0(C_i)\}}\,,$$

an unknown parameter. For estimation of β, motivated by the equation above, we propose to use the following class of estimating functions

$$U(\beta_1) \,=\, \sum_{i=1}^{n} W_i X_{1i}\{\bar{Y}_i - m_i\beta_1' X_{1i}\} \,=\, 0\,, \tag{3}$$

where the W_i's are some weights that could depend on X_i, $X_{1i}',= (X_i',\ 1)$ and $\beta_1' \,=\, (\beta', \theta)$.

Let $\hat{\beta}_1 \,=\, (\hat{\beta}', \hat{\theta})'$ denote the solution to equation (3). Then we have

$$\hat{\beta}_1 = \left[\sum_{i=1}^{n} W_i m_i X_{1i} X_{1i}'\right]^{-1} \sum_{i=1}^{n} W_i m_i X_{1i}\bar{Y}_i.$$

Acknowledgements

This research was supported in part by the Research Grant Council of Hong Kong (PolyU 5040/11P).

References

Diggle, P. J., Liang, K. Y., and Zeger, S. L. (1994). *The Analysis of Longitudinal Data*. Oxford University Press, Oxford.

Liang Y., Lu, W., and Ying, Z. (2009). Joint modeling and analysis of longitudinal data with informative observation times. *Biometrics* **65**, 377–384.

Lin, D. Y. and Ying, Z. (2001). Semiparametric and nonparametric regression analysis of longitudinal data. *J. Amer. Statist. Assoc.* **96**, 103–126.

Lipsitz, S. R., Fitzmaurice, G. M., Ibrahim, J. G., Gelber, R., and Lipshultz, S. (2002). Parameter estimation in longitudinal studies with outcome-dependent follow-up. *Biometrics* **58**, 621–630.

Sun, J., Park, D-H., Sun, L., and Zhao, X. (2005). Semiparametric regression analysis of longitudinal data with informative observation times. *J. Amer. Statist. Assoc.* **100**, 882–889.

Sun, J., Sun, L. and Liu, D. (2007). Regression analysis of longitudinal data in the presence of informative observation and censoring times. *J. Amer. Statist. Assoc.* **102**, 1397–1406.

Sun, J. and Wei, L. J. (2000). Regression analysis of panel count data with covariate-dependent observation and censoring times. *J. R. Statist. Soc. B* **62**, 293–302.

Wang, M. C., Qin, J., and Chiang, C. T. (2001). Analyzing recurrent event data with informative censoring. *J. Amer. Statist. Assoc.* **96**, 1057–1065.

Welsh, A. H., Lin, X., and Carroll, R. J. (2002). Marginal longitudinal nonparametric regression: locality and efficiency of spline and kernel Methods. *J. Amer. Statist. Assoc.* **97**, 482–493.

Zhang, Y. (2002). A semiparametric pseudolikelihood estimation method for panel count data. *Biometrika* **89**, 39–48.

Author Index