QUASI-LIKELIHOOD FUNCTIONS

By Peter McCullagh, 1983

Henrique Laureano (.github.io)
LEG @ UFPR

July 23, 2021

The Annals of Statistics
1983, Vol. 11, No. 1, 59–67

QUASI-LIKELIHOOD FUNCTIONS

BY PETER MCCULLAGH

Imperial College, London and University of British Columbia

The connection between quasi-likelihood functions, exponential family models and nonlinear weighted least squares is examined. Consistency and asymptotic normality of the parameter estimates are discussed under second moment assumptions. The parameter estimates are shown to satisfy a property of asymptotic optimality similar in spirit to, but more general than, the corresponding optimal property of Gauss-Markov estimators.

1. Distinguished Professor
   in the Department of Statistics @ University of Chicago;
2. Completed his PhD at Imperial College London,
   supervised by David Cox and Anthony Atkinson;
3. Also at Imperial College London,
   was the PhD supervisor of Gauss Cordeiro.

1. Introduction
2. A class of likelihood functions
3. Quasi-likelihood functions
4. Properties of quasi-likelihood functions
5. Estimation of σ^2
6. Examples of quasi-likelihood functions
7. A higher order theory

Introduction

1. Likelihood fucntion with **exponential family form**
   ↳ MLE through **weighted least squares**
   - **variance (assumed) constant**: we minimize a sum of squared residuals;
   - **variance not constant**:
     estimating equations can be thought as a generalization of the scoring method.
2. Likelihood function without exponential family form
   ↳ In some cases: weighted least squares
     ↳ Jorgensen, B. (1983). Maximum likelihood estimation and large sample inference for generalized linear and non-linear regression models. *Biometrika* 70

Paper purposes

1. Maximize the likelihood function through weighted least squares
   ↳ In which class of problems;
2. Weighted least squares under 2nd moment assumptions (**quasi-likelihood**).

1. Introduction
2. A class of likelihood functions
3. Quasi-likelihood functions
4. Properties of quasi-likelihood functions
5. Estimation of σ^2
6. Examples of quasi-likelihood functions
7. A higher order theory

A class of likelihood functions

$$\text{Log likelihood written in the form :} \qquad \sigma^{-2}\{\mathbf{y}^\top\boldsymbol{\theta} - b(\boldsymbol{\theta}) - c(\mathbf{y}, \sigma)\} \tag{1}$$

By differentiating it and assuming that the support does not depend on θ

$\Lsh$

Modeling failure time data

First of all, we have to choose which scale we model the **survival experience**.

1. Usually, is in the

$$\text{hazard (failure rate) scale}: \quad \lambda(t \mid \text{features}) = \lambda_0(t) \times c(\text{features}). \qquad (2)$$

We have a Equation 2 for each competing cause.

The cluster dependence is something actually not measured...

Not measured dependence $\rightarrow$ **random/latent effects** $\rightarrow$ Frailty models.

Frailty-based models for (multiple) survival experiences turn out in challengeable likelihood functions with inference routines mostly done via

- Elaborated and slow expectation–maximization (EM) algorithms;
- Inefficient Markov chain Monte Carlo (MCMC) schemes.

2. *Not* usually, the probability scale.

Probability scale $\rightarrow$ Cause-specific CIF

i.e., $\quad$ CIF $= \mathbb{P}[$ failure time $\leqslant t,$ a given cause $\mid$ features & latent effects $]$.

Common applications: *family studies*.
↳ Keywords: *within-family/cluster dependence*; *age at disease onset*; *populations*.

Formally,

for a cause-specific of failure k,
the cumulative incidence function (CIF) is defined as

$$
\begin{aligned}
F_k(t \mid \boldsymbol{x}) &= \mathbb{P}[T \leqslant t,\ K = k \mid \boldsymbol{x}] \\
&= \int_0^t f_k(z \mid \boldsymbol{x})\, \mathrm{d}z \quad (f_k(t \mid \boldsymbol{x}) \text{ is the (sub)density for the time to a type } k \text{ failure}) \\
&= \int_0^t \underbrace{\lambda_k(z \mid \boldsymbol{x})}_{\substack{\text{cause-specific} \\ \text{hazard function}}} \underbrace{S(z \mid \boldsymbol{x})}_{\substack{\text{overall} \\ \text{survival} \\ \text{function}}} \mathrm{d}z, \quad t > 0, \quad k = 1, \ \ldots, \ K.
\end{aligned}
$$

Again, a comprehensive reference is @kalb&prentice's book.

Here, we use the same CIF specification of @SCHEIKE.

SCHEIKE's CIF specification

For two competing causes of failure,
the cause-specific CIFs are specified in the following manner

$$F_k(t \mid \boldsymbol{x},\ u_1,\ u_2,\ \eta_k) = \underbrace{\pi_k(\boldsymbol{x},\ u_1,\ u_2)}_{\substack{\text{cluster-specific}\\ \text{risk level}}} \times \underbrace{\Phi[w_k g(t) - \boldsymbol{x}\gamma_k - \eta_k]}_{\substack{\text{cluster-specific}\\ \text{failure time trajectory}}}, \quad t > 0, \quad k = 1,\ 2, \quad (3)$$

with

1. $\pi_k(\boldsymbol{x}, \boldsymbol{u}) = \exp\{\boldsymbol{x}\beta_k + u_k\} \Big/ \left(1 + \sum_{m=1}^{K-1} \exp\{\boldsymbol{x}\beta_m + u_m\}\right), \quad k = 1,\ 2, \quad K = 3;$
2. $\Phi(\cdot)$ is the cumulative distribution function of a standard Gaussian distribution;
3. $g(t) = \text{arctanh}(2t/\delta - 1), \quad t \in (0,\ \delta), \quad g(t) \in (-\infty,\ \infty).$

In @SCHEIKE, this CIF specification is modeled under a *challengeable* pairwise composite likelihood approach [@lindsay88; @varin11].

Our contribution: a full likelihood analysis

For two competing causes of failure, a subject i, in the cluster j, in time t, we have

$$y_{ijt} \mid \underbrace{\{u_{1j}, u_{2j}, \eta_{1j}, \eta_{2j}\}}_{\text{latent effects}} \sim \text{Multinomial}(p_{1ijt}, p_{2ijt}, p_{3ijt})$$

$$\begin{bmatrix} u_1 \\ u_2 \\ \eta_1 \\ \eta_2 \end{bmatrix} \sim \begin{matrix}\text{Multivariate}\\ \text{Normal}\end{matrix} \left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2_{u_1} & \text{cov}(u_1, u_2) & \text{cov}(u_1, \eta_1) & \text{cov}(u_1, \eta_2) \\ & \sigma^2_{u_2} & \text{cov}(u_2, \eta_1) & \text{cov}(u_2, \eta_2) \\ & & \sigma^2_{\eta_1} & \text{cov}(\eta_1, \eta_2) \\ & & & \sigma^2_{\eta_2} \end{bmatrix} \right)$$

$$\begin{aligned} p_{kijt} &= \frac{\partial}{\partial t} F_k(t \mid \boldsymbol{x}, \boldsymbol{u}, \eta_k) \\ &= \frac{\exp\{\boldsymbol{x}_{kij}\beta_k + u_{kj}\}}{1 + \sum_{m=1}^{K-1} \exp\{\boldsymbol{x}_{mij}\beta_m + u_{mj}\}} \\ &\times w_k \frac{\delta}{2\delta t - 2t^2} \phi \left(w_k \text{arctanh}\left(\frac{t - \delta/2}{\delta/2} \right) - \boldsymbol{x}_{kij}\gamma_k - \eta_{kj} \right), \quad k = 1,\ 2. \end{aligned} \tag{4}$$

Simulating from the model

Marginal likelihood function for two competing causes

$$L(\boldsymbol{\theta}; y) = \prod_{j=1}^{J} \int_{\mathfrak{R}^4} \pi(y_j \mid \boldsymbol{r}_j) \times \pi(\boldsymbol{r}_j)\, \mathrm{d}\boldsymbol{r}_j$$

$$= \prod_{j=1}^{J} \int_{\mathfrak{R}^4} \underbrace{\left\{ \prod_{i=1}^{n_j} \prod_{t=1}^{n_{ij}} \left(\frac{(\sum_{k=1}^{K} y_{kijt})!}{y_{1ijt}!\, y_{2ijt}!\, y_{3ijt}!} \prod_{k=1}^{K} p_{kijt}^{y_{kijt}} \right) \right\}}_{\text{fixed effect component}} \times$$

$$\underbrace{(2\pi)^{-2} |\Sigma|^{-1/2} \exp\left\{ -\frac{1}{2} \boldsymbol{r}_j^{\top} \Sigma^{-1} \boldsymbol{r}_j \right\}}_{\text{latent effect component}} \mathrm{d}\boldsymbol{r}_j$$

$$= \prod_{j=1}^{J} \int_{\mathfrak{R}^4} \underbrace{\left\{ \prod_{i=1}^{n_j} \prod_{t=1}^{n_{ij}} \prod_{k=1}^{K} p_{kijt}^{y_{kijt}} \right\}}_{\text{fixed effect}} \underbrace{(2\pi)^{-2} |\Sigma|^{-1/2} \exp\left\{ -\frac{1}{2} \boldsymbol{r}_j^{\top} \Sigma^{-1} \boldsymbol{r}_j \right\}}_{\text{latent effect component}} \mathrm{d}\boldsymbol{r}_j, \qquad (5)$$

with p_{kijt} from Equation 4 and where $\boldsymbol{\theta} = [\boldsymbol{\beta}\ \boldsymbol{\gamma}\ \boldsymbol{w}\ \boldsymbol{\sigma}^2\ \boldsymbol{\rho}]^{\top}$ is the parameters vector.

1. Introduction
2. A class of likelihood functions
3. Quasi-likelihood functions
4. Properties of quasi-likelihood functions
5. Estimation of σ^2
6. Examples of quasi-likelihood functions
7. A higher order theory

TMB: Automatic Differentiation and Laplace Approximation

@TMB.

An R [@R21] package for the quickly implementation of complex random effect models through simple C++ templates.

Workflow

1. Write your objective function in a `.cpp` through a `#include <TMB.hpp>`;
2. Compile and load it in R via `TMB::compile()` and `base::dyn.load(TMB::dynlib())`;
3. Compute your objective function derivatives with `obj <- TMB::MakeADFun()`;
4. Perform the model fitting, `opt <- base::nlminb(obj$par, obj$fn, obj$gr)`;
5. Compute the parameters standard deviations, `TMB::sdreport(obj)`.

For details about TMB, AD, and Laplace approximation: @laurence.

1 Introduction

2 A class of likelihood functions

3 Quasi-likelihood functions

4 Properties of quasi-likelihood functions

5 Estimation of σ^2

6 Examples of quasi-likelihood functions

7 A higher order theory

Simulation study model designs

Risk model

Latent effects only on the risk level i.e.,

$$\Sigma = \begin{bmatrix} \sigma^2_{u_1} & \text{cov}_{u_1,u_2} \\ & \sigma^2_{u_2} \end{bmatrix}.$$

Time model

Latent effects only on the failure time trajectory level i.e.,

$$\Sigma = \begin{bmatrix} \sigma^2_{\eta_1} & \text{cov}_{\eta_1,\eta_2} \\ & \sigma^2_{\eta_2} \end{bmatrix}.$$

Block-diag model

Latent effects on the risk and time levels without cross-correlations i.e.,

$$\Sigma = \begin{bmatrix} \sigma^2_{u_1} & \text{cov}_{u_1,u_2} & 0 & 0 \\ & \sigma^2_{u_2} & 0 & 0 \\ & & \sigma^2_{\eta_1} & \text{cov}_{\eta_1,\eta_2} \\ & & & \sigma^2_{\eta_2} \end{bmatrix}.$$

Complete model

A *complete* latent effects structure i.e.,

$$\Sigma = \begin{bmatrix} \sigma^2_{u_1} & \text{cov}_{u_1,u_2} & \text{cov}_{u_1,\eta_1} & \text{cov}_{u_1,\eta_2} \\ & \sigma^2_{u_2} & \text{cov}_{u_2,\eta_1} & \text{cov}_{u_2,\eta_2} \\ & & \sigma^2_{\eta_1} & \text{cov}_{\eta_1,\eta_2} \\ & & & \sigma^2_{\eta_2} \end{bmatrix}.$$

Simulation study setup

Four latent effects structures:

1. Risk model;
2. Time model;
3. Block-diag model;
4. Complete model.

Two CIF configurations:

Low max incidence ≈ 0.15;

High max incidence ≈ 0.60.

For each of those $\mathbf{4} \times \mathbf{2} = \mathbf{8}$ scenarios, we vary the sample and cluster sizes:

5000 data points

- 2500 clusters of **size 2**;
- 1000 clusters of **size 5**;
- 500 clusters of **size 10**.

30000 data points

- 15000 clusters of **size 2**;
- 6000 clusters of **size 5**;
- 3000 clusters of **size 10**.

60000 data points

- 30000 clusters of **size 2**;
- 12000 clusters of **size 5**;
- 6000 clusters of **size 10**.

Totalizing, $\mathbf{8} \times \mathbf{3} \times \mathbf{3} = \mathbf{72}$ scenarios.
For each scenario, we simulate **500** samples, totalizing $\mathbf{72} \times \mathbf{500} = \mathbf{36000}$ model fittings.

Simulation study results

First of all, the **time**.

- The *non-complete* models (2D Laplace aprox.) are kind of fast, taking always **less than 5 min**.
- In the most expensive scenarios (30K 4D Laplaces), **the complete model takes 30 min**. In a **full R** implementation with 10K 4D Laplaces, it **took 30hrs**. **TMB is fast**.
- We also did a Bayesian analysis via Stan/NUTS-HMC [@RStan].
  - **1 week of parallelized processing** for a 2500 size 2 clusters scenario with tuned NUTS. This just reinforces the MCMC impracticability for some complex models.

Parameters estimation.

- The *non-complete* models fail to learn the data. They appear to be *not structured enough* to capture the data characteristics.

1. Introduction

2. A class of likelihood functions

3. Quasi-likelihood functions

4. Properties of quasi-likelihood functions

5. Estimation of σ^2

6. Examples of quasi-likelihood functions

7. A higher order theory

1 Introduction

2 A class of likelihood functions

3 Quasi-likelihood functions

4 Properties of quasi-likelihood functions

5 Estimation of σ^2

6 Examples of quasi-likelihood functions

7 A higher order theory

1 Introduction

2 A class of likelihood functions

3 Quasi-likelihood functions

4 Properties of quasi-likelihood functions

5 Estimation of σ^2

6 Examples of quasi-likelihood functions

7 A higher order theory

Take-home message

The complete model works. It's not magnificent, but it works.

1. It works better in the high CIF scenarios;
2. As expected, as the sample size increases the results get better;
3. We do not see any considerable performance difference between cluster/family sizes;
4. Satisfactory full likelihood analysis under the maximum likelihood estimation framework (the estimates bias-variance could be smaller).

What else can we do?

1. Instead of a conditional approach (latent effects model), we can try a marginal approach e.g., an McGLM [@mcglm];
2. We can also try a copula [@copulas], on maybe two fronts: 1) for a full specification; 2) to accommodate the within-cluster dependence.

For more read @laurence master thesis.

Thanks for watching and have a great day

Special thanks to