Statistical Methods in Credit Risk Modeling

by

Aijun Zhang

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Statistics)
in The University of Michigan
2009

Doctoral Committee:

Professor Vijayan N. Nair, Co-Chair
Agus Sudjianto, Co-Chair, Bank of America
Professor Tailen Hsing
Associate Professor Jionghua Jin
Associate Professor Ji Zhu

To my elementary school, high school and university teachers

ACKNOWLEDGEMENTS

First of all, I would express my gratitude to my advisor Prof. Vijay Nair for guiding me during the entire PhD research. I appreciate his inspiration, encouragement and protection through these valuable years at the University of Michigan. I am thankful to Julian Faraway for his encouragement during the first years of my PhD journey. I would also like to thank Ji Zhu, Judy Jin and Tailen Hsing for serving on my doctoral committee and helpful discussions on this thesis and other research works.

I am grateful to Dr. Agus Sudjianto, my co-advisor from Bank of America, for giving me the opportunity to work with him during the summers of 2006 and 2007 and for offering me a full-time position. I appreciate his guidance, active support and his many illuminating ideas. I would also like to thank Tony Nobili, Mike Bonn, Ruilong He, Shelly Ennis, Xuejun Zhou, Arun Pinto, and others I first met in 2006 at the Bank. They all persuaded me to jump into the area of credit risk research; I did it a year later and finally came up with this thesis within two more years.

I would extend my acknowledgement to Prof. Kai-Tai Fang for his consistent encouragement ever since I graduated from his class in Hong Kong 5 years ago.

This thesis is impossible without the love and remote support of my parents in China. To them, I am most indebted.

TABLE OF CONTENTS

CHAPTER

LIST OF FIGURES

Figure

LIST OF TABLES

<u>**Table**</u>

ABSTRACT

This research deals with some statistical modeling problems that are motivated by credit risk analysis. Credit risk modeling has been the subject of considerable research interest in finance and has recently drawn the attention of statistical researchers. In the first chapter, we provide an up-to-date review of credit risk models and demonstrate their close connection to survival analysis.

The first statistical problem considered is the development of adaptive smoothing spline (AdaSS) for heterogeneously smooth function estimation. Two challenging issues that arise in this context are evaluation of reproducing kernel and determination of local penalty, for which we derive an explicit solution based on piecewise type of local adaptation. Our nonparametric AdaSS technique is capable of fitting a diverse set of 'smooth' functions including possible jumps, and it plays a key role in subsequent work in the thesis.

The second topic is the development of dual-time analytics for observations involving both lifetime and calendar timescale. It includes "vintage data analysis" (VDA) for continuous type of responses in the third chapter, and "dual-time survival analysis" (DtSA) in the fourth chapter. We propose a maturation-exogenous-vintage (MEV) decomposition strategy in order to understand the risk determinants in terms of self-maturation in lifetime, exogenous influence by macroeconomic conditions, and heterogeneity induced from vintage originations. The intrinsic identification problem is discussed for both VDA and DtSA. Specifically, we consider VDA under Gaussian process models, provide an efficient MEV backfitting algorithm and assess its performance with both simulation and real examples.

DtSA on Lexis diagram is of particular importance in credit risk modeling where the default events could be triggered by both endogenous and exogenous hazards. We consider nonparametric estimators, first-passage-time parameterization and semi-parametric Cox regression. These developments extend the family of models for both credit risk modeling and survival analysis. We demonstrate the application of DtSA to credit card and mortgage risk analysis in retail banking, and shed some light on understanding the ongoing credit crisis.

CHAPTER I

An Introduction to Credit Risk Modeling

Credit risk is a critical area in banking and is of concern to a variety of stakeholders: institutions, consumers and regulators. It has been the subject of considerable research interest in banking and finance communities, and has recently drawn the attention of statistical researchers. By Wikipedia's definition,

> "*Credit risk* is the risk of loss due to a debtor's non-payment of a loan or other line of credit." (Wikipedia.org, as of March 2009)

Central to credit risk is the *default event*, which occurs if the debtor is unable to meet its legal obligation according to the debt contract. The examples of default event include the bond default, the corporate bankruptcy, the credit card charge-off, and the mortgage foreclosure. Other forms of credit risk include the repayment delinquency in retail loans, the loss severity upon the default event, as well as the unexpected change of credit rating.

An enormous literature in credit risk has been fostered by both academics in finance and practitioners in industry. There are two parallel worlds based upon a simple dichotomous rule of data availability: (a) the direct measurements of credit performance and (b) the prices observed from credit market. The data availability leads to two streams of credit risk modeling that have key distinctions.

1.1 Two Worlds of Credit Risk

The two worlds of credit risk can be simply characterized by the types of default probability, one being *actual* and the other being *implied*. The former corresponds to the direct observations of defaults, also known as the physical default probability in finance. The latter refers to the risk-neutral default probability implied from the credit market data, e.g. corporate bond yields.

The academic literature of corporate credit risk has been inclined to the study of the implied defaults, which is yet a puzzling world.

1.1.1 Credit Spread Puzzle

The *credit spread* of a defaultable corporate bond is its excess yield over the default-free Treasury bond of the same time to maturity. Consider a zero-coupon corporate bond with unit face value and maturity date T. The *yield-to-maturity* at $t < T$ is defined by

$$Y(t,T) = -\frac{\log B(t,T)}{T-t} \tag{1.1}$$

where $B(t,T)$ is the bond price of the form

$$B(t,T) = \mathbb{E}\left[\exp\left\{-\int_t^T (r(s)+\lambda(s))ds\right\}\right], \tag{1.2}$$

given the independent term structures of the *interest rate* $r(\cdot)$ and the *default rate* $\lambda(\cdot)$. Setting $\lambda(t) \equiv 0$ gives the benchmark price $B_0(t,T)$ and the yield $Y_0(t,T)$ of Treasury bond. Then, the credit spread can be calculated as

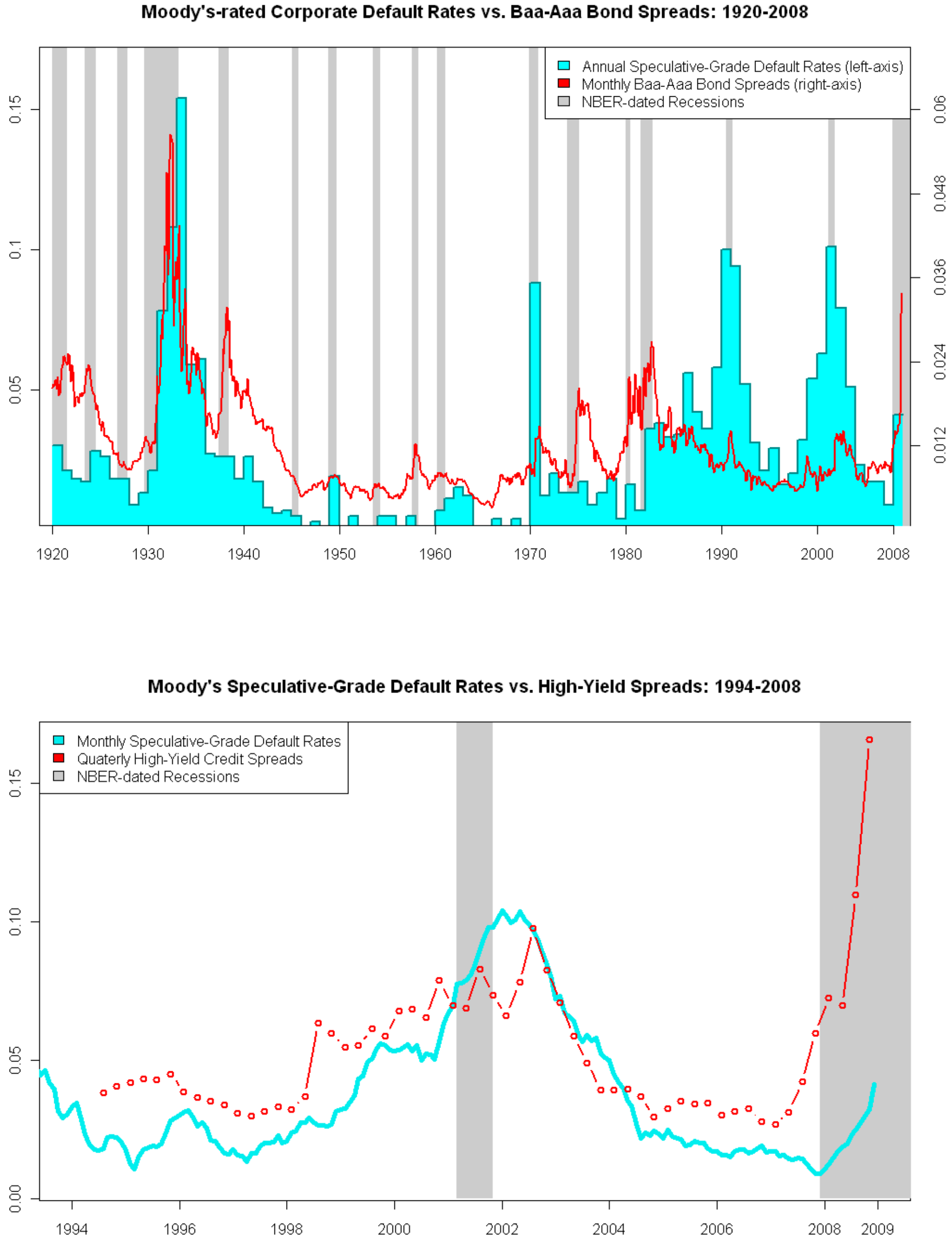
$$\mathsf{Spread}(t,T) = Y(t,T) - Y_0(t,T) = -\frac{\log(B(t,T)/B_0(t,T))}{T-t} = -\frac{\log(1-q(t,T))}{T-t} \tag{1.3}$$

where $q(t,T) = \mathbb{E}\big[e^{-\int_t^T \lambda(t)ds}\big]$ is the *conditional default probability* $\mathbb{P}[\tau \le T|\tau > t]$ and τ denotes the time-to-default, to be detailed in the next section.

The credit spread is supposed to co-move with the default rate. For illustration, Figure 1.1 (top panel) plots the Moody's-rated corporate default rates and Baa-Aaa bond spreads ranging from 1920-2008. The shaded backgrounds correspond to NBER's latest announcement of recession dates. Most spikes of the default rates and the credit spreads coincide with the recessions, but it is clear that the movements of two time series differ in both level and change. Such a lack of match is the so-called *credit spread puzzle* in the latest literature of corporate finance; the actual default rates could be successfully implied from the market data of credit spreads by none of the existing structural credit risk models.

The default rates implied from credit spreads mostly overestimate the expected default rates. One factor that helps explain the gap is the *liquidity risk* – a security cannot be traded quickly enough in the market to prevent the loss. Figure 1.1 (bottom panel) plots, in fine resolution, the Moody's speculative-grade default rates versus the high-yield credit spreads: 1994-2008. It illustrates the phenomenon where the spread changed in response to liquidity-risk events, while the default rates did not react significantly until quarters later; see the liquidity crises triggered in 1998 (Russian/LTCM shock) and 2007 (Subprime Mortgage meltdown). Besides the liquidity risk, there exist other known (e.g. tax treatments of corporate bonds vs. government bonds) and unknown factors that make incomplete the implied world of credit risk. As of today, we lack a thorough understanding of the credit spread puzzle; see e.g. Chen, Lesmond and Wei (2007) and references therein.

The shaky foundation of the default risk implied from market credit spreads without looking at the historical defaults leads to further questions about the credit derivatives, e.g. Credit Default Swap (CDS) and Collateralized Debt Obligation (CDO). The 2007-08 collapse of credit market in Wall Street is partly due to over-complication in "innovating" such complex financial instruments on the one hand, and over-simplification in quantifying their embedded risks on the other.



Moody's-rated Corporate Default Rates vs. Baa-Aaa Bond Spreads: 1920-2008

Moody's Speculative-Grade Default Rates vs. High-Yield Spreads: 1994-2008

Figure 1.1: Moody's-rated corporate default rates, bond spreads and NBER-dated recessions. Data sources: a) Moody's Baa & Aaa corporate bond yields (http://research.stlouisfed.org/fred2/categories/119); b) Moody's Special Comment on Corporate Default and Recovery Rates, 1920-2008 (http://www.moodys.com/); c) NBER-dated recessions (http://www.nber.org/cycles/).

1.1.2 Actual Defaults

The other world of credit risk is the study of default probability bottom up from the actual credit performance. It includes the popular industry practices of

a) credit rating in corporate finance, by e.g. the three major U.S. rating agencies: Moody's, Standard & Poor's, and Fitch.

b) credit scoring in consumer lending, by e.g. the three major U.S. credit bureaus: Equifax, Experian and TransUnion.

Both credit ratings and scores represent the creditworthiness of individual corporations and consumers. The final evaluations are based on statistical models of the expected default probability, as well as judgement by rating/scoring specialists. Let us describe very briefly some rating and scoring basics related to the thesis.

The letters Aaa and Baa in Figure 1.1 are examples of Moody's rating system, which use Aaa, Aa, A, Baa, Ba, B, Caa, Ca, C to represent the likelihood of default from the lowest to the highest. The speculative grade in Figure 1.1 refers to Ba and the worse ratings. The speculative-grade corporate bonds are sometimes said to be high-yield or junk.

FICO, developed by Fair Isaac Corporation, is the best-known consumer credit score and it is the most widely used by U.S. banks and other credit card or mortgage lenders. It ranges from 300 (very poor) to 850 (best), and intends to represent the creditworthiness of a borrower such that he or she will repay the debt. For the same borrower, the three major U.S. credit bureaus often report inconsistent FICO scores based on their own proprietary models.

Compared to either the industry practices mentioned above or the academics of the implied default probability, the academic literature based on the actual defaults is much smaller, which we believe is largely due to the limited access for an academic researcher to the proprietary internal data of historical defaults. A few academic works will be reviewed later in Section 1.3. In this thesis, we make an attempt to develop statistical methods based on the actual credit performance data. For demonstration, we shall use the synthetic or tweaked samples of retail credit portfolios, as well as the public release of corporate default rates by Moody's.

1.2 Credit Risk Models

This section reviews the finance literature of credit risk models, including both structural and intensity-based approaches. Our focus is placed on the probability of default and the hazard rate of time-to-default.

1.2.1 Structural Approach

In credit risk modeling, structural approach is also known as the firm-value approach since a firm's inability to meet the contractual debt is assumed to be determined by its asset value. It was inspired by the 1970s Black-Scholes-Merton methodology for financial option pricing. Two classic structural models are the *Merton model* (Merton, 1974) and the *first-passage-time model* (Black and Cox, 1976).

The Merton model assumes that the default event occurs at the maturity date of debt if the asset value is less than the debt level. Let D be the debt level with maturity date T, and let $V(t)$ be the latent asset value following a geometric Brownian motion

$$dV(t) = \mu V(t)dt + \sigma V(t)dW(t), \tag{1.4}$$

with drift μ, volatility σ and the standard Wiener process $W(t)$. Recall that $\mathbb{E}W(t) = 0$, $\mathbb{E}W(t)W(s) = \min(t, s)$. Given the initial asset value $V(0) > D$, by Itó's lemma,

$$\frac{V(t)}{V(0)} = \exp\left\{\Big(\mu - \frac{1}{2}\sigma^2\Big)t + \sigma W_t\right\} \sim \mathsf{Lognormal}\left(\Big(\mu - \frac{1}{2}\sigma^2\Big)t, \sigma^2 t\right), \tag{1.5}$$

from which one may evaluate the default probability $\mathbb{P}(V(T) \le D)$.

The notion of *distance-to-default* facilitates the computation of conditional default probability. Given the sample path of asset values up to t, one may first estimate the unknown parameters in (1.4) by maximum likelihood method. According to Duffie and Singleton (2003), let us define the distance-to-default $X(t)$ by the number of standard deviations such that $\log V_t$ exceeds $\log D$, i.e.

$$X(t) = (\log V(t) - \log D)/\sigma. \tag{1.6}$$

Clearly, $X(t)$ is a drifted Wiener process of the form

$$X(t) = c + bt + W(t), \quad t \geq 0 \tag{1.7}$$

with $b = \frac{\mu - \sigma^2/2}{\sigma}$ and $c = \frac{\log V(0) - \log D}{\sigma}$. Then, it is easy to verify that the conditional probability of default at maturity date T is

$$\mathbb{P}(V(T) \leq D | V(t) > D) = \mathbb{P}(X(T) \leq 0 | X(t) > 0) = \Phi\left(\frac{X(t) + b(T-t)}{\sqrt{T-t}}\right), \tag{1.8}$$

where $\Phi(\cdot)$ is the cumulative normal distribution function.

The first-passage-time model by Black and Cox (1976) extends the Merton model so that the default event could occur as soon as the asset value reaches a pre-specified debt barrier. Figure 1.2 (left panel) illustrates the first-passage-time of a drifted Wiener process by simulation, where we set the parameters $b = -0.02$ and $c = 8$ (s.t. a constant debt barrier). The key difference of Merton model and Black-Cox model lies in the green path (the middle realization viewed at T), which is treated as default in one model but not the other.

By (1.7) and (1.8), $V(t)$ hits the debt barrier once the distance-to-default process $X(t)$ hits zero. Given the initial distance-to-default $c \equiv X(0) > 0$, consider the

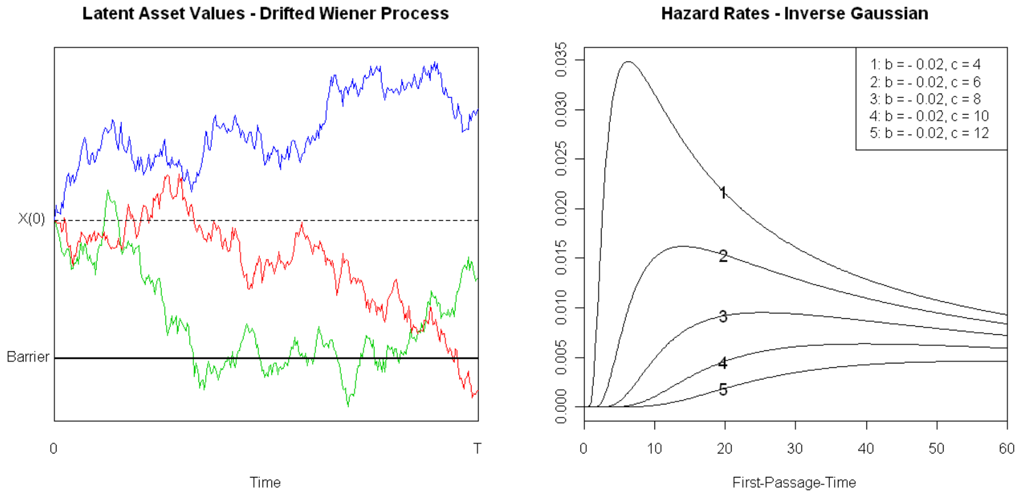


Figure 1.2: Simulated drifted Wiener process, first-passage-time and hazard rate.

first-passage-time

$$\tau = \inf\{t \geq 0 : \; X(t) \leq 0\}, \tag{1.9}$$

where $\inf \emptyset = \infty$ as usual. It is well known that τ follows the *inverse Gaussian distribution* (Schrödinger, 1915; Tweedie, 1957; Chhikara and Folks, 1989) with the density

$$f(t) = \frac{c}{\sqrt{2\pi}} t^{-3/2} \exp\left\{ -\frac{(c+bt)^2}{2t} \right\}, \quad t \geq 0. \tag{1.10}$$

See also other types of parametrization in Marshall and Olkin (2007; §13). The survival function $S(t)$ is defined by $\mathbb{P}(\tau > t)$ for any $t \geq 0$ and is given by

$$S(t) = \Phi\left(\frac{c+bt}{\sqrt{t}}\right) - e^{-2bc}\Phi\left(\frac{-c+bt}{\sqrt{t}}\right). \tag{1.11}$$

The *hazard rate*, or the conditional default rate, is defined by the instantaneous rate of default conditional on the survivorship,

$$\lambda(t) \;=\; \lim_{\Delta t \downarrow 0} \frac{1}{\Delta t} \mathbb{P}(t \leq \tau < t + \Delta t | \tau \geq t) = \frac{f(t)}{S(t)}. \tag{1.12}$$

Using the inverse Gaussian density and survival functions, we obtain the form of the first-passage-time hazard rate:

$$\lambda(t; c, b) = \frac{\dfrac{c}{\sqrt{2\pi t^3}} \exp\left\{-\dfrac{(c+bt)^2}{2t}\right\}}{\Phi\left(\dfrac{c+bt}{\sqrt{t}}\right) - e^{-2bc}\Phi\left(\dfrac{-c+bt}{\sqrt{t}}\right)}, \quad c > 0. \tag{1.13}$$

This is one of the most important forms of hazard function in structural approach to credit risk modeling. Figure 1.2 (right panel) plots $\lambda(t; c, b)$ for $b = -0.02$ and $c = 4, 6, 8, 10, 12$, which resemble the default rates from low to high credit qualities in terms of credit ratings or FICO scores. Both the trend parameter b and the initial distance-to-default parameter c provide insights to understanding the shape of the hazard rate; see the details in Chapter IV or Aalen, Borgan and Gjessing (2008; §10).

Modern developments of structural models based on Merton and Black-Cox models can be referred to Bielecki and Rutkowski (2004; §3). Later in Chapter IV, we will discuss the dual-time extension of first-passage-time parameterization with both endogenous and exogenous hazards, as well as non-constant default barrier and incomplete information about structural parameters.

1.2.2 Intensity-based Approach

The intensity-based approach is also called the reduced-form approach, proposed independently by Jarrow and Turnbull (1995) and Madan and Unal (1998). Many follow-up papers can be found in Lando (2004), Bielecki and Rutkowski (2004) and references therein. Unlike the structural approach that assumes the default to be completely determined by the asset value subject to a barrier, the default event in the reduced-form approach is governed by an externally specified intensity process that may or may not be related to the asset value. The default is treated as an unexpected event that comes 'by surprise'. This is a practically appealing feature, since in the real world the default event (e.g. Year 2001 bankruptcy of Enron Corporation) is often all of a sudden happening without announcement.

The *default intensity* corresponds to the hazard rate $\lambda(t) = f(t)/S(t)$ defined in (1.12) and it has roots in statistical reliability and survival analysis of time-to-failure. When $S(t)$ is absolutely continuous with $f(t) = d(1 - S(t))/dt$, we have that

$$\lambda(t) = \frac{-dS(t)}{S(t)dt} = -\frac{d[\log S(t)]}{dt}, \quad S(t) = \exp\left\{-\int_0^t \lambda(s)ds\right\}, \quad t \geq 0.$$

In survival analysis, $\lambda(t)$ is usually assumed to be a deterministic function in time. In credit risk modeling, $\lambda(t)$ is often treated as stochastic. Thus, the default time τ is *doubly stochastic*. Note that Lando (1998) adopted the term "doubly stochastic Poisson process" (or, Cox process) that refers to a counting process with possibly recurrent events. What matters in modeling defaults is only the first jump of the counting process, in which case the default intensity is equivalent to the hazard rate.

In finance, the intensity-based models are mostly the term-structure models borrowed from the literature of interest-rate modeling. Below is an incomplete list:

$$\begin{aligned} \text{Vasicek:} \quad & d\lambda(t) = \kappa(\theta - \lambda(t))dt + \sigma dW_t \\ \text{Cox-Ingersoll-Roll:} \quad & d\lambda(t) = \kappa(\theta - \lambda(t))dt + \sigma\sqrt{\lambda(t)}dW_t \\ \text{Affine jump:} \quad & d\lambda(t) = \mu(\lambda(t))dt + \sigma(\lambda(t))dW_t + dJ_t \end{aligned} \tag{1.14}$$

with reference to Vasicek (1977), Cox, Ingersoll and Roll (1985) and Duffie, Pan and Singleton (2000). The last model involves a pure-jump process J_t, and it covers both the mean-reverting Vasicek and CIR models by setting $\mu(\lambda(t)) = \kappa(\theta - \lambda(t))$ and $\sigma(\lambda(t)) = \sqrt{\sigma_0^2 + \sigma_1^2\lambda(t)}$.

The term-structure models provide straightforward ways to simulate the future default intensity for the purpose of predicting the conditional default probability. However, they are *ad hoc* models lacking fundamental interpretation of the default event. The choices (1.14) are popular because they could yield closed-form pricing formulas for (1.2), while the real-world default intensities deserve more flexible and meaningful forms. For instance, the intensity models in (1.14) cannot be used to model the endogenous shapes of first-passage-time hazards illustrated in Figure 1.2.

The dependence of default intensity on state variables $\mathbf{z}(t)$ (e.g. macroeconomic covariates) is usually treated through a multivariate term-structure model for the joint of $(\lambda(t), \mathbf{z}(t))$. This approach essentially presumes a linear dependence in the diffusion components, e.g. by correlated Wiener processes. In practice, the effect of a state variable on the default intensity can be non-linear in many other ways.

The intensity-based approach also includes the duration models for econometric analysis of actual historical defaults. They correspond to the classical survival analysis in statistics, which opens another door for approaching credit risk.

1.3 Survival Models

Credit risk modeling in finance is closely related to survival analysis in statistics, including both the first-passage-time structural models and the duration type of intensity-based models. In a rough sense, default risk models based on the actual credit performance data exclusively belong to survival analysis, since the latter by definition is the analysis of time-to-failure data. Here, the failure refers to the default event in either corporate or retail risk exposures; see e.g. Duffie, Saita and Wang (2007) and Deng, Quigley and Van Order (2000).

We find that the survival models have at least four-fold advantages in approach to credit risk modeling:

1. flexibility in parametrizing the default intensity,

2. flexibility in incorporating various types of covariates,

3. effectiveness in modeling the credit portfolios, and

4. being straightforward to enrich and extend the family of credit risk models.

The following materials are organized in a way to make these points one by one.

1.3.1 Parametrizing Default Intensity

In survival analysis, there are a rich set of lifetime distributions for parametrizing the default intensity (i.e. hazard rate) $\lambda(t)$, including

$$\begin{aligned}
\textsf{Exponential:} \quad & \lambda(t) = \alpha \\
\textsf{Weibull:} \quad & \lambda(t) = \alpha t^{\beta-1} \qquad (1.15)\\
\textsf{Lognormal:} \quad & \lambda(t) = \frac{1}{t\alpha}\frac{\phi(\log t/\alpha)}{\Phi(-\log t/\alpha)} \\
\textsf{Log-logistic:} \quad & \lambda(t) = \frac{(\beta/\alpha)(t/\alpha)^{\beta-1}}{1+(t/\alpha)^{\beta}}, \quad \alpha, \beta > 0
\end{aligned}$$

where $\phi(x) = \Phi'(x)$ is the density function of normal distribution. Another example is the inverse Gaussian type of hazard rate function in (1.13). More examples of parametric models can be found in Lawless (2003) and Marshall and Olkin (2007) among many other texts. They all correspond to the distributional assumptions of the default time τ, while the inverse Gaussian distribution has a beautiful first-passage-time interpretation.

The log-location-scale transform can be used to model the hazard rates. Given a latent default time τ_0 with baseline hazard rate $\lambda_0(t)$ and survival function $S_0(t) = \exp\big\{-\int_0^t \lambda_0(s)ds\big\}$, one may model the firm-specific default time τ_i by

$$\log \tau_i = \mu_i + \sigma_i \log \tau_0, \quad -\infty < \mu_i < \infty, \ \sigma_i > 0 \qquad (1.16)$$

with individual risk parameters (μ_i, σ_i). Then, the firm-specific survival function takes the form

$$S_i(t) = S_0\left(\left[\frac{t}{e^{\mu_i}}\right]^{1/\sigma_i}\right) \qquad (1.17)$$

and the firm-specific hazard rate is

$$\lambda_i(t) = -\frac{d\log S_i(t)}{dt} = \frac{1}{\sigma_i t}\left[\frac{t}{e^{\mu_i}}\right]^{1/\sigma_i}\lambda_0\left(\left[\frac{t}{e^{\mu_i}}\right]^{1/\sigma_i}\right). \tag{1.18}$$

For example, given the Weibull baseline hazard rate $\lambda_0(t) = \alpha t^{\beta-1}$, the log-location-scale transformed hazard rate is given by

$$\lambda_i(t) = \frac{\alpha}{\sigma_i}t^{\frac{\beta}{\sigma_i}-1}\exp\left\{-\frac{\beta\mu_i}{\sigma_i}\right\}. \tag{1.19}$$

1.3.2 Incorporating Covariates

Consider firm-specific (or consumer-specific) covariates $\mathbf{z}_i \in \mathbb{R}^p$ that are either static or time-varying. The dependence of default intensity on $\mathbf{z}_i$ can be studied by regression models. In the context of survival analysis, there are two popular classes of regression models, namely

1. the multiplicative hazards regression models, and

2. the accelerated failure time (AFT) regression models.

The hazard rate is taken as the modeling basis in the first class of regression models:

$$\lambda_i(t) = \lambda_0(t)r(\mathbf{z}_i(t)) \tag{1.20}$$

where $\lambda_0(t)$ is a *baseline hazard* function and $r(\mathbf{z}_i(t))$ is a firm-specific *relative-risk multiplier* (both must be positive). For example, one may use the inverse Gaussian hazard rate (1.13) or pick one from (1.15) as the baseline $\lambda_0(t)$. The relative risk term is often specified by $r(\mathbf{z}_i(t)) = \exp\{\boldsymbol{\theta}^T\mathbf{z}_i(t)\}$ with parameter $\boldsymbol{\theta} \in \mathbb{R}^p$. Then, the *hazard ratio* between any two firms, $\lambda_i(t)/\lambda_j(t) = \exp\{\boldsymbol{\theta}^T([\mathbf{z}_i(t) - \mathbf{z}_j(t)]\}$, is constant if the covariates difference $\mathbf{z}_i(t) - \mathbf{z}_j(t)$ is constant in time, thus ending up with proportional hazard rates. Note that the covariates $\mathbf{z}_i(t)$ are sometimes transformed before entering the model. For example, an econometric modeler usually takes difference of Gross Domestic Product (GDP) index and uses the GDP growth as the entering covariate.

For a survival analyst, the introduction of the multiplicative hazards model (1.20) is incomplete without introducing the Cox's proportional hazards (CoxPH) model. The latter is among the most popular models in modern survival analysis. Rather than using the parametric form of baseline hazard, Cox (1972, 1975) considered $\lambda_i(t) = \lambda_0(t)\exp\{\boldsymbol{\theta}^T\mathbf{z}_i(t)\}$ by leaving the baseline hazard $\lambda_0(t)$ unspecified. Such semiparametric modeling could reduce the estimation bias of the covariate effects due to baseline misspecification. One may use the partial likelihood to estimate $\boldsymbol{\theta}$, and the nonparametric likelihood to estimate $\lambda_0(t)$. The estimated baseline $\hat{\lambda}_0(t)$ can be further smoothed upon parametric, term-structure or data-driven techniques.

The second class of AFT regression models are based on the aforementioned log-location-scale transform of default time. Suppose that $\mathbf{z}_i$ is not time-varying. Given τ_0 with baseline hazard rate $\lambda_0(t)$, consider (1.16) with the location $\mu_i = \boldsymbol{\theta}^T\mathbf{z}_i$ and the scale $\sigma_i = 1$ (for simplicity), i.e. $\log\tau_i = \boldsymbol{\theta}^T\mathbf{z}_i + \log\tau_0$. Then, it is straightforward to get the survival function and hazard rate by (1.17) and (1.18):

$$S_i(t) = S_0\Big(t\exp\{-\boldsymbol{\theta}^T\mathbf{z}_i\}\Big), \quad \lambda_i(t) = \lambda_0\Big(t\exp\{-\boldsymbol{\theta}^T\mathbf{z}_i\}\Big)\exp\{-\boldsymbol{\theta}^T\mathbf{z}_i\} \tag{1.21}$$

An interesting phenomenon is when using the Weibull baseline, by (1.19), we have that $\lambda_i(t) = \lambda_0(t)\exp\{-\beta\boldsymbol{\theta}^T\mathbf{z}_i\}$. This is equivalent to use the multiplicative hazards model (1.20), as a unique property of the Weibull lifetime distribution. For more details about AFT regression, see e.g. Lawless (2003; §6).

Thus, the covariates $\mathbf{z}_i$ could induce the firm-specific heterogeneity via either a relative-risk multiplier $\lambda_i(t)/\lambda_0(t) = e^{\boldsymbol{\theta}^T\mathbf{z}_i(t)}$ or the default time acceleration $\tau_i/\tau_0 = e^{\boldsymbol{\theta}^T\mathbf{z}_i}$. In situations where certain hidden covariates are not accessible, the unexplained variation also leads to heterogeneity. The *frailty models* introduced by Vaupel, et al.

(1979) are designed to model such unobserved heterogeneity as a random quantity, say, $Z \sim \mathsf{Gamma}(\delta^{-1}, \delta)$ with shape δ^{-1} and scale $\delta > 0$, where a $\mathsf{Gamma}(k, \delta)$ distribution has the density

$$p(z; k, \delta) = \frac{z^{k-1} \exp\{-z/\delta\}}{\Gamma(k)\delta^k}, \quad z > 0 \tag{1.22}$$

with mean $k\delta$ and variance $k\delta^2$. Then, the proportional frailty model extends (1.20) to be

$$\lambda_i(t) = Z \cdot \lambda_0(t) r(\mathbf{z}_i(t)), \tag{1.23}$$

for each single obligor i subject to the odd effect Z; see Aalen, et al. (2008; §6). On the other hand, the frailty models can be also used to characterize the default correlation among multiple obligors, which is the next topic.

1.3.3 Correlating Credit Defaults

The issue of default correlation embedded in credit portfolios has drawn intense discussions in the recent credit risk literature, in particular along with today's credit crisis involving problematic basket CDS or CDO financial instruments; see e.g. Bluhm and Overbeck (2007). In eyes of a fund manager, the positive correlation of defaults would increase the level of total volatility given the same level of total expectation (cf. Markowitz portfolio theory). In academics, among other works, Das, et al. (2007) performed an empirical analysis of default times for U.S. corporations and provided evidence for the importance of default correlation.

The default correlation could be effectively characterized by multivariate survival analysis. In a broad sense, there exists two different approaches:

1. correlating the default intensities through the common covariates,
2. correlating the default times through the copulas.

The first approach is better known as the *conditionally independent* intensity-based approach, in the sense that the default times are independent conditional on the common covariates. Examples of common covariates include the market-wide variables, e.g. the GDP growth, the short-term interest rate, and the house-price appreciation index. Here, the default correlation induced by the common covariates is in contrast to the aforementioned default heterogeneity induced by the firm-specific covariates.

As announced earlier, the frailty models (1.23) can also be used to capture the dependence of multi-obligor default intensities in case of *hidden* common covariates. They are usually rephrased as *shared frailty models*, for differentiation from single-obligor modeling purpose. The frailty effect Z is usually assumed to be time-invariant. One may refer to Hougaard (2000) for a comprehensive treatment of multivariate survival analysis from a frailty point of view. See also Aalen et al. (2008; §11) for some intriguing discussion of dynamic frailty modeled by diffusion and Lévy processes. In finance, Duffie, et al. (2008) applied the shared frailty effect across firms, as well as a dynamic frailty effect to represent the unobservable macroeconomic covariates.

The copula approach to default correlation considers the default times as the modeling basis. A copula $C : [0,1]^n \mapsto [0,1]$ is a function used to formulate the multivariate joint distribution based on the marginal distributions, e.g.,

$$\begin{aligned}
&\textsf{Gaussian:} && C_{\boldsymbol{\Sigma}}(u_1,\ldots,u_n) = \Phi_{\boldsymbol{\Sigma}}(\Phi^{-1}(u_1),\ldots,\Phi^{-1}(u_n)) \\
&\textsf{Student-t:} && C_{\boldsymbol{\Sigma},\nu}(u_1,\ldots,u_n) = \Theta_{\boldsymbol{\Sigma},\nu}(\Theta_\nu^{-1}(u_1),\ldots,\Theta_\nu^{-1}(u_n)) \qquad (1.24)\\
&\textsf{Archimedean:} && C_\Psi(u_1,\ldots,u_n) = \Psi^{-1}\left(\sum_{i=1}^n \Psi(u_i)\right)
\end{aligned}$$

where $\boldsymbol{\Sigma} \in \mathbb{R}^{n\times n}$, $\Phi_{\boldsymbol{\Sigma}}$ denotes the multivariate normal distribution, $\Theta_{\boldsymbol{\Sigma},\nu}$ denotes the multivariate Student-t distribution with degrees of freedom ν, and Ψ is the generator of Archimedean copulas; see Nelsen (2006) for details. When $\Psi(u) = -\log u$, the Archimedean copula reduces to $\prod_{i=1}^n u_i$ (independence copula). By Sklar's theorem, for a multivariate joint distribution, there always exists a copula that can link the joint distribution to its univariate marginals. Therefore, the joint survival distribution of $(\tau_1, \ldots, \tau_n)$ can be characterized by $S_{\mathsf{joint}}(t_1, \ldots, t_n) = C(S_1(t_1), \ldots, S_n(t_n))$, upon an appropriate selection of copula C.

Interestingly enough, the shared frailty models correspond to the Archimedean copula, as observed by Hougaard (2000; §13) and recast by Aalen, et al. (2008; §7). For each τ_i with underlying hazard rates $Z \cdot \lambda_i(t)$, assume $\lambda_i(t)$ to be deterministic and Z to be random. Then, the marginal survival distributions are found by integrating over the distribution of Z,

$$S_i(t) = \mathbb{E}_Z \left[\exp \left\{ - \int_0^t \lambda_i(s) ds \right\} \right] = \mathscr{L}(\Lambda_i(t)), \quad i = 1, \ldots, n \tag{1.25}$$

where $\mathscr{L}(x) = \mathbb{E}_Z[\exp\{-xZ\}]$ is the *Laplace transform* of Z and $\Lambda_i(t)$ is the cumulative hazard. Similarly, the joint survival distribution is given by

$$S_{\mathsf{joint}}(t_1, \ldots, t_n) = \mathscr{L}\left(\sum_{i=1}^n \Lambda_i(t_i) \right) = \mathscr{L}\left(\sum_{i=1}^n \mathscr{L}^{-1}(S_i(t_i)) \right), \tag{1.26}$$

where the second equality follows (1.25) and leads to an Archimedean copula. For example, given the Gamma frailty $Z \sim \mathsf{Gamma}(\delta^{-1}, \delta)$, we have that $\mathscr{L}(x) = (1 + \delta x)^{-1/\delta}$, so

$$S_{\mathsf{joint}}(t_1, \ldots, t_n) = \left(1 + \delta \sum_{i=1}^n \int_0^{t_i} \lambda_i(s) ds \right)^{-1/\delta}.$$

In practice of modeling credit portfolios, one must first of all check the appropriateness of the assumption behind either the frailty approach or the copula approach. An obvious counter example is the recent collapse of the basket CDS and CDO market for which the rating agencies have once abused the use of the over-simplified Gaussian copula. Besides, the correlation risk in these complex financial instruments is highly contingent upon the macroeconomic conditions and *black-swan* events, which are rather dynamic, volatile and unpredictable.

Generalization

The above survey of survival models is too brief to cover the vast literature of survival analysis that deserves potential applications in credit risk modeling. To be selective, we have only cited Hougaard (2000), Lawless (2003) and Aalen, et al. (2008), but there are numerous texts and monographs on survival analysis and reliability.

It is straightforward to enrich and extend the family of credit risk models, by either the existing devices in survival analysis, or the new developments that would benefit both fields. The latter is one of our objectives in this thesis, and we will focus on the dual-time generalization.

1.4 Scope of the Thesis

Both old and new challenges arising in the context of credit risk modeling call for developments of statistical methodologies. Let us fist preview the complex data structures in real-world credit risk problems before defining the scope of the thesis.

The vintage time series are among the most popular versions of economic and risk management data; see e.g. ALFRED digital archive[1] hosted by U.S. Federal Reserve Bank of St. Louis. For risk management in retail banking, consider for instance the revolving exposures of credit cards. Subject to the market share of the card-issuing bank, thousands to hundred thousands of new cards are issued monthly, ranging from product types (e.g. speciality, reward, co-brand and secured cards), acquisition channels (e.g. banking center, direct mail, telephone and internet), and other distinguishable characteristics (e.g. geographic region where the card holder resides). By convention, the accounts originated from the same month constitute a *monthly vintage* (quarterly and yearly vintages can be defined in the same fashion),

[1] http://alfred.stlouisfed.org/

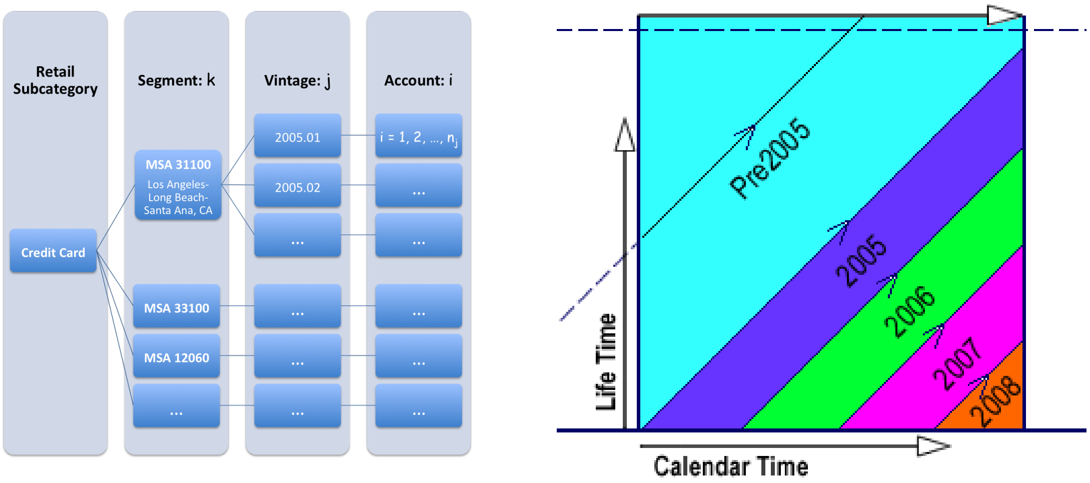


Figure 1.3: Illustration of retail credit portfolios and vintage diagram.

where "vintage" is a borrowed term from wine-making industry to account for the sharing origination. Then, a vintage time series refers to the longitudinal observations and performance measurements for the cohort of accounts that share the same origination date and therefore the same age.

Figure 1.3 provides an illustration of retail credit portfolios and the scheme of vintage data collection, where we use only the geographic MSA (U.S. Metropolitan Statistical Area) as a representative of segmentation variables. Each segment consists of multiple vintages that have different origination dates, and each vintage consists of varying numbers of accounts. The most important feature of the vintage data is its *dual-time coordinates*. Let the calendar time be denoted by t, each vintage be denoted by its origination time v, then the lifetime (i.e. age, or called the "month-on-book" for a monthly vintage) is calculated by $m = t - v$ for $t \geq v$. Such $(m, t; v)$ tuples lie in the dual-time domain, called the *vintage diagram*, with x-axes representing the calendar time t and y-axis representing the lifetime m. Unlike a usual 2D space, the vintage diagram consists of unit-slope trajectories each representing the evolvement of a different vintage. On a vintage diagram, all kinds of data can be collected including the continuous, binary or count observations. In particular, when the dual-time observations are the individual lives together with birth and death time points, and when these observations are subject to truncation and censoring, one may draw a line segment for each individual to represent the vintage data graphically. Such a graph is known as the *Lexis diagram* in demography and survival analysis; see Figure 4.1 based on a dual-time-to-default simulation.

For corporate bond and loan issuers, the observations of default events or default rates share the same vintage data structure. Table 1.1 shows the Moody's speculative-grade default rates for annual cohorts 1970-2008, which are calculated from the cumulative rates released recently by Moody's Global Credit Policy (February 2009). Each cohort is observed by year end of 2008, and truncated by 20 years maximum in lifetime. Thus, the performance window for the cohorts originated in the latest years becomes shorter and shorter. The default rates in Table 1.1 are aligned in lifetime, while they can be also aligned in calendar time. To compare the cohort performances in dual-time coordinates, the marginal plots in both lifetime and calendar time are given Figure 1.4, together with the side-by-side box-plots for checking the vintage (i.e. cohort) heterogeneity. It is clear that the average default rates gradually decrease in lifetime, but very volatile in calendar time. Note that the calendar dynamics follow closely the NBER-dated recessions shown in Figure 1.1. Besides, there seems to be heterogeneity among vintages. These projection views are only an initial exploration. Further analysis of this vintage data will be performed in Chapter III.

The main purpose of the thesis is to develop a dual-time modeling framework based on observations on the vintage diagram, or "dual-time analytics". First of all, it is worth mentioning that the vintage data structure is by no means an exclusive feature of economic and financial data. It also appears in clinical trials with staggered entry, demography, epidemiology, dendrochronology, etc. As an interesting example in dendrochronology, Esper, Cook and Schweingruber (*Science*, 2002) analyzed the long historical tree-ring records in a dual-time manner in order to reconstruct the

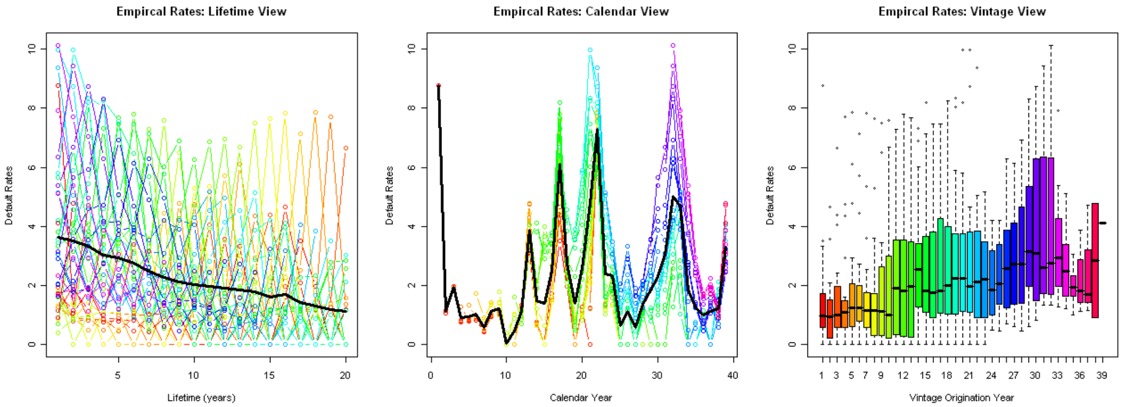


Figure 1.4: Moody's speculative-grade default rates for annual cohorts 1970-2008: projection views in lifetime, calendar and vintage origination time.

past temperature variability. We believe that the dual-time analytics developed in the thesis can be also applied to non-financial fields.

The road map in Figure 1.5 outlines the scope of the thesis. In Chapter II, we discuss the adaptive smoothing spline (AdaSS) for heterogeneously smooth function estimation. Two challenging issues are evaluation of reproducing kernel and determination of local penalty, for which we derive an explicit solution based upon piecewise type of local adaptation. Four different examples, each having a specific feature of heterogeneous smoothness, are used to demonstrate the new AdaSS technique. Note that the AdaSS may estimate 'smooth' functions including possible jumps, and it plays a key role in the subsequent work in the thesis.

In Chapter III, we develop the vintage data analysis (VDA) for continuos type of dual-time responses (e.g. loss rates). We propose an MEV decomposition framework based on Gaussian process models, where M stands for the maturation curve, E for the exogenous influence and V for the vintage heterogeneity. The intrinsic identification problem is discussed. Also discussed is the semiparametric extension of MEV models in the presence of dual-time covariates. Such models are motivated from the practical needs in financial risk management. An efficient MEV Backfitting algorithm is provided for model estimation, and its performance is assessed by a sim-

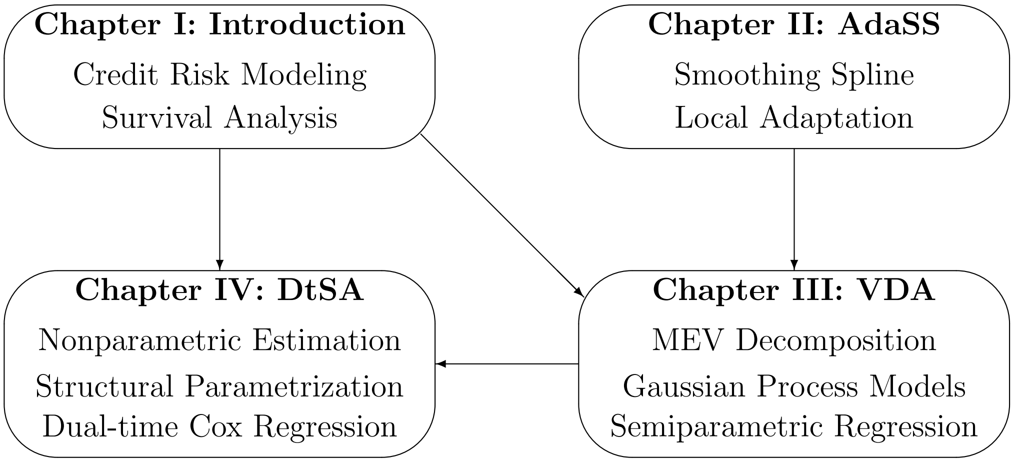


Figure 1.5: A road map of thesis developments of statistical methods in credit risk modeling.

ulation study. Then, we apply the MEV risk decomposition strategy to analyze both corporate default rates and retail loan loss rates.

In Chapter IV, we study the dual-time survival analysis (DtSA) for time-to-default data observed on Lexis diagram. It is of particular importance in credit risk modeling where the default events could be triggered by both endogenous and exogenous hazards. We consider (a) nonparametric estimators under one-way, two-way and three-way underlying hazards models, (b) structural parameterization via the first-passage-time triggering system with an endogenous distance-to-default process associated with exogenous time-transformation, and (c) dual-time semiparametric Cox regression with both endogenous and exogenous baseline hazards and covariate effects, for which the method of partial likelihood estimation is discussed. Also discussed is the random-effect vintage heterogeneity modeled by the shared frailty. Finally, we demonstrate the application of DtSA to credit card and mortgage risk analysis in retail banking, and shed some light on understanding the ongoing credit crisis from a new dual-time analytic perspective.

Table 1.1: Moody's speculative-grade default rates. Data source: Moody's special comment (release: February 2009) on corporate default and recovery rates, 1920-2008 (http://www.moodys.com/) and author's calculations.

Cohort	Year1	Year2	Year3	Year4	Year5	Year6	Year7	Year8	Year9	Year10	Year11	Year12	Year13	Year14	Year15	Year16	Year17	Year18	Year19	Year20
1970	8.772	1.082	1.866	0.778	0.804	0.840	0.443	0.941	1.000	0.000	0.000	1.250	3.329	0.711	0.000	1.612	3.500	1.975	0.000	1.205
1971	1.152	1.989	0.829	0.859	0.898	0.474	1.009	1.072	0.000	0.000	1.339	3.573	0.761	0.000	1.721	4.659	2.110	0.000	1.293	0.000
1972	1.957	0.812	0.840	0.878	0.463	0.977	1.034	0.000	0.000	1.266	4.049	0.719	0.000	2.439	4.382	1.967	0.000	1.185	1.263	6.647
1973	1.271	0.876	0.916	0.484	1.018	1.069	0.000	0.000	1.289	4.755	0.718	0.000	1.608	4.343	2.934	0.000	1.157	1.231	7.698	1.372
1974	1.330	0.931	0.493	1.037	1.090	0.000	0.000	1.303	4.794	0.726	0.000	1.620	6.098	2.930	0.000	1.177	2.507	7.855	1.406	1.578
1975	1.735	0.912	1.439	1.007	0.000	0.000	1.205	4.462	0.676	0.000	1.521	5.721	2.741	1.001	1.078	2.283	7.125	1.266	1.386	0.000
1976	0.864	1.361	1.424	0.000	0.531	1.134	3.607	0.639	0.000	1.430	5.348	2.561	0.943	1.009	2.124	7.814	1.204	1.308	0.000	0.000
1977	1.339	1.406	0.000	1.052	1.128	3.599	0.638	0.000	1.419	5.305	2.536	0.927	0.983	2.070	7.650	1.186	1.297	0.000	0.000	0.000
1978	1.798	0.000	1.012	1.086	3.463	1.223	0.651	2.709	6.496	2.411	0.884	0.943	2.994	7.486	1.173	2.575	0.000	0.000	0.000	0.000
1979	0.420	0.897	0.958	3.031	2.128	3.396	2.963	7.579	2.087	0.765	0.824	3.507	6.668	1.045	2.276	0.000	0.000	0.000	0.000	0.000
1980	1.613	0.859	4.045	1.884	3.993	3.138	7.272	2.451	0.680	2.222	3.925	6.966	1.967	2.125	0.000	0.000	0.000	1.223	0.000	0.000
1981	0.701	4.053	1.932	3.675	3.423	7.796	2.514	0.560	2.470	4.702	6.855	1.752	1.906	0.000	0.000	0.000	1.120	1.154	0.000	0.000
1982	3.571	4.104	2.902	3.403	7.673	2.213	0.495	2.208	4.266	6.256	1.603	1.739	0.000	0.000	0.000	1.103	1.169	0.000	0.000	2.575
1983	3.824	3.153	3.681	7.308	2.871	2.315	3.180	6.013	6.904	1.631	1.833	0.000	0.000	0.000	1.181	1.261	1.306	1.355	2.783	2.864
1984	3.333	3.797	8.199	2.793	3.535	4.052	7.148	5.788	1.349	2.278	0.000	0.000	0.000	1.050	1.157	1.194	1.234	1.280	2.623	0.000
1985	3.448	6.668	3.860	3.753	5.041	7.454	5.696	1.657	1.873	0.000	1.477	0.000	0.856	0.954	1.014	1.071	2.275	2.326	0.000	0.000
1986	5.644	4.563	3.282	4.603	7.460	6.458	2.966	2.435	1.084	1.194	0.000	1.419	0.813	1.773	1.861	3.960	1.030	0.000	0.000	1.150
1987	4.222	3.894	5.441	8.261	7.547	3.644	2.653	1.132	1.264	0.471	1.065	1.239	1.376	2.889	3.072	4.113	0.000	0.000	1.016	0.000
1988	3.580	5.890	8.335	8.032	3.618	3.078	1.191	1.678	0.754	1.268	2.435	1.640	2.364	3.871	4.210	0.000	0.000	0.890	0.000	2.119
1989	5.794	9.980	8.176	3.955	3.238	1.280	1.759	0.997	1.890	2.631	1.477	2.674	3.527	5.134	0.000	0.000	0.798	0.000	1.887	3.020
1990	9.976	8.732	4.654	3.237	1.349	1.811	0.876	1.982	2.310	1.730	2.815	4.139	4.534	0.000	0.000	0.735	0.000	1.663	3.501	
1991	9.370	5.059	3.371	1.401	2.155	0.914	2.078	2.440	1.859	3.069	4.534	4.951	0.000	0.759	0.829	0.000	0.923	3.849		
1992	5.154	3.495	1.596	2.362	1.185	2.017	2.338	2.225	2.949	4.290	5.170	0.000	0.712	0.791	0.000	0.926	4.049			
1993	3.072	1.848	3.340	1.399	1.608	1.871	3.146	3.101	4.210	4.056	0.501	0.559	0.607	1.303	0.715	3.194				
1994	2.073	2.869	1.901	1.265	2.327	3.883	3.553	4.197	3.896	1.474	1.269	0.471	1.568	0.586	3.255					
1995	2.920	1.785	1.874	2.763	3.862	3.963	6.297	4.960	2.696	1.247	0.695	1.165	0.877	2.448						
1996	1.637	2.145	3.544	4.187	4.119	6.152	4.894	2.780	1.362	0.617	1.057	1.213	2.715							
1997	2.028	3.570	4.597	4.727	6.921	4.697	2.617	1.444	0.468	1.351	1.261	2.829								
1998	3.152	5.642	5.933	8.324	5.068	3.792	2.415	0.620	1.466	1.522	2.468									
1999	5.384	6.667	8.743	6.299	3.723	2.449	0.832	1.398	1.494	2.482										
2000	6.339	9.440	7.048	4.105	2.562	1.365	1.744	1.348	2.607											
2001	10.124	7.713	4.947	2.730	1.542	1.851	1.341	2.780												
2002	7.921	5.375	2.927	1.691	2.247	1.281	3.159													
2003	5.123	2.955	1.654	2.050	1.224	3.392														
2004	2.346	1.652	1.943	1.013	3.070															
2005	1.732	1.895	1.141	3.880																
2006	1.688	1.169	4.717																	
2007	0.918	4.792																		
2008	4.129																			

CHAPTER II

Adaptive Smoothing Spline

2.1 Introduction

For observational data, statisticians are generally interested in smoothing rather than interpolation. Let the observations be generated by the signal plus noise model

$$y_i = f(t_i) + \varepsilon(t_i), \quad i = 1, \ldots, n \tag{2.1}$$

where f is an unknown smooth function with unit interval support $[0, 1]$, and $\varepsilon(t_i)$ is the random error. Function estimation from noisy data is a central problem in modern nonparametric statistics, and includes the methods of kernel smoothing [Wand and Jones (1995)], local polynomial smoothing [Fan and Gijbels (1996)] and wavelet shrinkage [Vidakovic (1999)]. In this thesis, we concentrate on the method of *smoothing spline* that has an elegant formulation through the mathematical device of *reproducing kernel Hilbert space* (RKHS). There is a rich body of literature on smoothing spline since the pioneering work of Kimeldorf and Wahba (1970); see the monographs by Wahba (1990), Green and Silverman (1994) and Gu (2002). Ordinarily, the smoothing spline is formulated by the following variation problem

$$\min_{f \in \mathcal{W}_m} \left\{ \frac{1}{n} \sum_{i=1}^{n} w(t_i)[y_i - f(t_i)]^2 + \lambda \int_0^1 [f^{(m)}(t)]^2 dt \right\}, \quad \lambda > 0 \tag{2.2}$$

where the target function is an element of the m-order Sobolev space

$$\mathcal{W}_m = \{f : f, f', \ldots, f^{(m-1)} \text{ absolutely continuous}, f^{(m)} \in \mathcal{L}_2[0,1]\}. \tag{2.3}$$

The objective in (2.2) is a sum of two functionals, the mean squared error $\texttt{MSE} = \frac{1}{n}\sum_{i=1}^n w(t_i)[y_i - f(t_i)]^2$ and the roughness penalty $\texttt{PEN} = \int_0^1 [f^{(m)}(t)]^2 dt$; hence the tuning parameter λ controls the trade-off between fidelity to the data and the smoothness of the estimate. Let us call the ordinary smoothing spline (2.2) OrdSS in short.

The weights $w(t_i)$ in $\texttt{MSE}$ come from the non-stationary error assumption of (2.1) where $\varepsilon(t) \sim N(0, \sigma^2/w(t))$. They can be also used for aggregating data with replicates. Suppose the data are generated by (2.1) with stationary error but time-varying sampling frequency r_k for $k = 1, \ldots, K$ distinct points, then

$$\texttt{MSE} = \frac{1}{n}\sum_{k=1}^{K}\sum_{j=1}^{r_k}[y_{kj} - f(t_k)]^2 = \frac{1}{K}\sum_{k=1}^{K} w(t_k)\big[\bar{y}_k - f(t_k)\big]^2 + \texttt{Const}. \tag{2.4}$$

where $\bar{y}_k \sim N(f(t_k), \sigma^2/r_k)$ for $k = 1, \ldots, K$ are the pointwise averages. To train OrdSS, the raw data can be replaced by $\bar{y}_k's$ together with the weights $w(t_k) \propto r_k$.

The OrdSS assumes that $f(t)$ is homogeneously smooth (or rough), then performs regularization based on the simple integral of squared derivatives $[f^{(m)}(t)]^2$. It excludes the target functions with spatially varying degrees of smoothness, as noted by Wahba (1985) and Nycha (1988). In practice, there are various types of functions of non-homogeneous smoothness, and four such scenarios are illustrated in Figure 2.1. The first two are the simulation examples following exactly the same setting as Donoho and Johnstone (1994), by adding the $N(0,1)$ noise to $n = 2048$ equally spaced signal responses that are re-scaled to attain signal-to-noise ratio 7. The last two are the sampled data from automotive engineering and financial engineering.

1. Doppler function: $f(t) = \sqrt{t(1-t)}\sin(2\pi(1+a)/(t+a)), a = 0.05, t \in [0,1]$. It

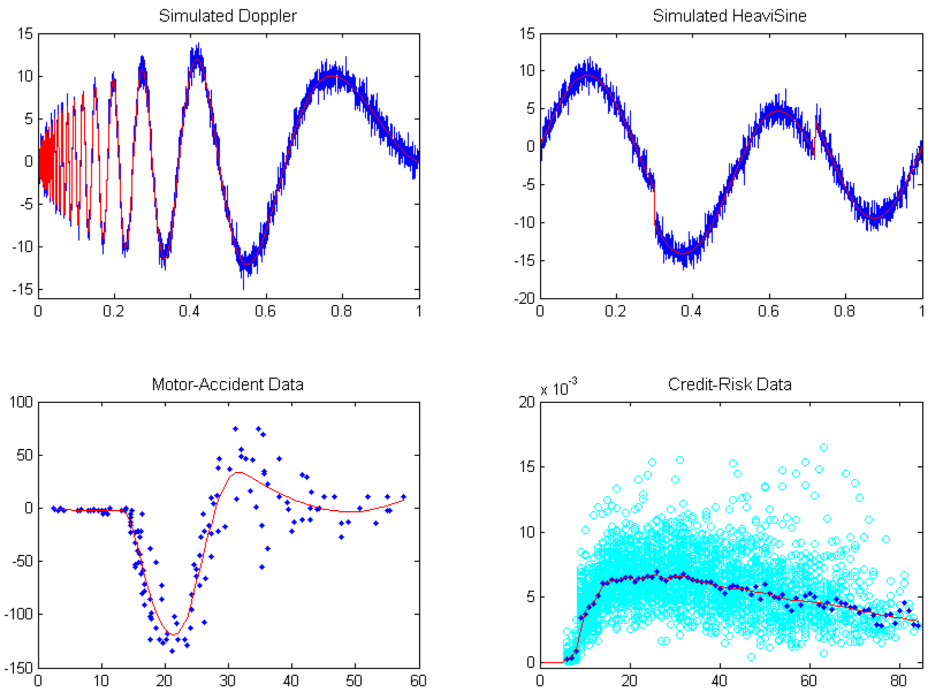


Figure 2.1: Heterogeneous smooth functions: (1) Doppler function simulated with noise, (2) HeaviSine function simulated with noise, (3) Motorcycle-Accident experimental data, and (4) Credit-Risk tweaked sample.

has both time-varying magnitudes and time-varying frequencies.

2. HeaviSine function: $f(t) = 3 \sin 4\pi t - \text{sgn}(t - 0.3) - \text{sgn}(0.72 - t), t \in [0, 1]$. It has a downward jump at $t = 0.3$ and an upward jump at 0.72.

3. Motorcycle-Accident experiment: a classic data from Silverman (1985). It consists of accelerometer readings taken through time from a simulated motorcycle crash experiment, in which the time points are irregularly spaced and the noises vary over time. In Figure 2.1, a hypothesized smooth shape is overlaid on the observations, and it illustrates the fact that the acceleration stays relatively constant until the crash impact.

4. Credit-Risk tweaked sample: retail loan loss rates for the revolving credit exposures in retail banking. For academic use, business background is removed and

> the sample is tweaked. The tweaked sample includes monthly vintages booked in 7 years (2000-2006), where a vintage refers to a group of credit accounts originated from the same month. They are all observed up to the end of 2006, so a vintage booked earlier has a longer observation length. In Figure 2.1, the x-axis represents the months-on-book, and the circles are the vintage loss rates. The point-wise averages are plotted as solid dots, which are wiggly on large months-on-book due to decreasing sample sizes. Our purpose is to estimate an underlying shape (illustrated by the solid line), whose degree of smoothness, by assumption, increases upon growth.

The first three examples have often been discussed in the nonparametric analysis literature, where function estimation with non-homogeneous smoothness has been of interest for decades. A variety of adaptive methods has been proposed, such as variable-bandwidth kernel smoothing (Müller and Stadtmüller, 1987), multivariate adaptive regression splines (Friedman, 1991), adaptive wavelet shrinkage (Donoho and Johnstone, 1994), local-penalty regression spline (Ruppert and Carroll, 2000), as well as the treed Gaussian process modeling (Gramacy and Lee, 2008). In the context of smoothing spline, Luo and Wahba (1997) proposed a hybrid adaptive spline with knot selection, rather than using every sampling point as a knot. More naturally, Wahba (1995), in a discussion, suggested to use the local penalty

$$\texttt{PEN} = \int_0^1 \rho(t)[f^{(m)}(t)]^2 dt, \quad \text{s.t.} \quad \rho(t) > 0 \text{ and } \int_0^1 \rho(t)dt = 1 \tag{2.5}$$

where the add-on function $\rho(t)$ is adaptive in response to the local behavior of $f(t)$ so that heavier penalty is imposed to the regions of lower curvature. This local penalty functional is the foundation of the whole chapter. Relevant works include Abramovich and Steinberg (1996) and Pintore, Speckman and Holmes (2006).

This chapter is organized as follows. In Section 2.2 we formulate the AdaSS (adaptive smoothing spline) through the local penalty (2.5), derive its solution via RKHS and generalized ridge regression, then discuss how to determine both the smoothing parameter and the local penalty adaptation by the method of cross validation. Section 2.3 covers some basic properties of the AdaSS. The experimental results for the aforementioned examples are presented in Section 2.4. We summarize the chapter with Section 2.5 and give technical proofs in the last section.

2.2 AdaSS: Adaptive Smoothing Spline

The AdaSS extends the OrdSS (2.2) based on the local penalty (2.5). It is formulated as

$$\min_{f\in\mathcal{W}_m}\left\{\frac{1}{n}\sum_{i=1}^{n} w(t_i)\big[y_i - f(t_i)\big]^2 + \lambda\int_0^1 \rho(t)\big[f^{(m)}(t)\big]^2 dt\right\}, \quad \lambda > 0. \tag{2.6}$$

The following proposition has its root in Kimeldorf and Wahba (1971); or see Wahba (1990; §1).

Proposition 2.1 (AdaSS)**.** *Given a bounded integrable positive function $\rho(t)$, define the r.k.*

$$\mathbb{K}(t,s) = \int_0^1 \rho^{-1}(u)G_m(t,u)G_m(s,u)du, \quad t,s\in[0,1] \tag{2.7}$$

where $G_m(t,u) = \frac{(t-u)_+^{m-1}}{(m-1)!}$. Then, the solution of (2.6) can be expressed as

$$f(t) = \sum_{j=0}^{m-1}\alpha_j\phi_j(t) + \sum_{i=1}^{n} c_i\mathbb{K}(t,t_i) \equiv \boldsymbol{\alpha}^T\boldsymbol{\phi}(t) + \mathbf{c}^T\boldsymbol{\xi}_n(t) \tag{2.8}$$

where $\phi_j(t) = t^j/j!$ for $j = 0,\ldots,m-1$, $\boldsymbol{\alpha}\in\mathbb{R}^m$, $\mathbf{c}\in\mathbb{R}^n$.

Since the AdaSS has a finite-dimensional linear expression (2.8), it can be solved by generalized linear regression. Write $\mathbf{y} = (y_1,\ldots,y_n)^T$, $\mathbf{B} = (\boldsymbol{\phi}(t_1),\cdots,\boldsymbol{\phi}(t_n))^T$, $\mathbf{W} = \mathsf{diag}(w(t_1),\ldots,w(t_n))$ and $\boldsymbol{\Sigma} = [\mathbb{K}(t_i,t_j)]_{n\times n}$. Substituting (2.8) into (2.6), we have the regularized least squares problem

$$\min_{\boldsymbol{\alpha},\mathbf{c}} \left\{ \frac{1}{n} \big\| \mathbf{y} - \mathbf{B}\boldsymbol{\alpha} - \boldsymbol{\Sigma}\mathbf{c} \big\|_{\mathbf{W}}^2 + \lambda \big\|\mathbf{c}\big\|_{\boldsymbol{\Sigma}}^2 \right\}, \tag{2.9}$$

where $\|\mathbf{x}\|_{\mathbf{A}}^2 = \mathbf{x}^T\mathbf{A}\mathbf{x}$ for $\mathbf{A} \geq 0$. By setting partial derivatives w.r.t. $\boldsymbol{\alpha}$ and $\mathbf{c}$ to be zero, we obtain the equations

$$\mathbf{B}^T\mathbf{W}\big(\mathbf{B}\boldsymbol{\alpha} + \boldsymbol{\Sigma}\mathbf{c}\big) = \mathbf{B}^T\mathbf{W}\mathbf{y}, \quad \boldsymbol{\Sigma}\mathbf{W}\big(\mathbf{B}\boldsymbol{\alpha} + (\boldsymbol{\Sigma} + n\lambda\mathbf{W}^{-1})\mathbf{c}\big) = \boldsymbol{\Sigma}\mathbf{W}\mathbf{y},$$

or

$$\mathbf{B}\boldsymbol{\alpha} + (\boldsymbol{\Sigma} + n\lambda\mathbf{W}^{-1})\mathbf{c} = \mathbf{y}, \quad \mathbf{B}^T\mathbf{c} = \mathbf{0}. \tag{2.10}$$

Consider the QR-decomposition $\mathbf{B} = (\mathbf{Q}_1 \vdots \mathbf{Q}_2)(\mathbf{R}^T \vdots \mathbf{0}^T)^T = \mathbf{Q}_1\mathbf{R}$ with orthogonal $(\mathbf{Q}_1 \vdots \mathbf{Q}_2)_{n\times n}$ and upper-triangular $\mathbf{R}_{m\times m}$, where $\mathbf{Q}_1$ is $n \times m$ and $\mathbf{Q}_2$ is $n \times (n-m)$. Multiplying both sides of $\mathbf{y} = \mathbf{B}\boldsymbol{\alpha} + (\boldsymbol{\Sigma} + n\lambda\mathbf{W}^{-1})\mathbf{c}$ by $\mathbf{Q}_2^T$ and using the fact that $\mathbf{c} = \mathbf{Q}_2\mathbf{Q}_2^T\mathbf{c}$ (since $\mathbf{B}^T\mathbf{c} = \mathbf{0}$), we have that $\mathbf{Q}_2^T\mathbf{y} = \mathbf{Q}_2^T(\boldsymbol{\Sigma} + n\lambda\mathbf{W}^{-1})\mathbf{c} = \mathbf{Q}_2^T(\boldsymbol{\Sigma} + n\lambda\mathbf{W}^{-1})\mathbf{Q}_2\mathbf{Q}_2^T\mathbf{c}$. Then, simple algebra yields

$$\widehat{\mathbf{c}} = \mathbf{Q}_2\big(\mathbf{Q}_2^T(\boldsymbol{\Sigma} + n\lambda\mathbf{W}^{-1})\mathbf{Q}_2\big)^{-1}\mathbf{Q}_2^T\mathbf{y}, \quad \widehat{\boldsymbol{\alpha}} = \mathbf{R}^{-1}\mathbf{Q}_1^T\big(\mathbf{y} - (\boldsymbol{\Sigma} + n\lambda\mathbf{W}^{-1})\widehat{\mathbf{c}}\big). \tag{2.11}$$

The fitted values at the design points are given by $\hat{\mathbf{y}} = \mathbf{B}\widehat{\boldsymbol{\alpha}} + \boldsymbol{\Sigma}\widehat{\mathbf{c}}$. By (2.10), $\mathbf{y} = \mathbf{B}\widehat{\boldsymbol{\alpha}} + (\boldsymbol{\Sigma} + n\lambda\mathbf{W}^{-1})\widehat{\mathbf{c}}$, so the residuals are given by

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = n\lambda\mathbf{W}^{-1}\widehat{\mathbf{c}} \equiv \big(\mathbf{I} - \mathbf{S}_{\lambda,\rho}\big)\mathbf{y} \tag{2.12}$$

where $\mathbf{S}_{\lambda,\rho} = \mathbf{I} - n\lambda\mathbf{W}^{-1}\mathbf{Q}_2\big(\mathbf{Q}_2^T(\boldsymbol{\Sigma} + n\lambda\mathbf{W}^{-1})\mathbf{Q}_2^T\big)^{-1}\mathbf{Q}_2^T$ is often called the *smoothing matrix* in the sense that

$$\hat{\mathbf{y}} = \mathbf{S}_{\lambda,\rho}\mathbf{y} = n\lambda\mathbf{W}^{-1}\mathbf{Q}_2\big(\mathbf{Q}_2^T(\boldsymbol{\Sigma} + n\lambda\mathbf{W}^{-1})\mathbf{Q}_2^T\big)^{-1}\mathbf{Q}_2^T\mathbf{y}. \tag{2.13}$$

We use the notation $\mathbf{S}_{\lambda,\rho}$ to denote its explicit dependence on both the smoothing parameter λ and the local adaptation $\rho(\cdot)$, where ρ is needed for the evaluation of the matrix $\boldsymbol{\Sigma} = [\mathbb{K}(t_i, t_j)]$ based on the r.k. (2.7).

The AdaSS is analogous to OrdSS in constructing the confidence interval under a Bayesian interpretation by Wahba (1983). For the sampling time points in particular, we may construct the 95% interval pointwise by

$$\hat{f}(t_i) \pm z_{0.025}\sqrt{\widehat{\sigma^2} w^{-1}(t_i)\mathbf{S}_{\lambda,\rho}[i,i]}, \quad i = 1, \ldots, n \tag{2.14}$$

where $\mathbf{S}_{\lambda,\rho}[i,i]$ represents the i-th diagonal element of the smoothing matrix. The estimate of noise variance is given by $\widehat{\sigma^2} = \left\|(\mathbf{I} - \mathbf{S}_{\lambda,\rho})\mathbf{y}\right\|^2_{\mathbf{W}}/(n-p)$, where $p = \mathsf{trace}(\mathbf{S}_{\lambda,\rho})$ can be interpreted as the equivalent degrees of freedom.

2.2.1 Reproducing Kernels

The AdaSS depends on the evaluation of the r.k. of the integral form (2.7). For general $\rho(t)$, we must resort to numerical integration techniques. In what follows we present a closed-form expression of $\mathbb{K}(t,s)$ when $\rho^{-1}(t)$ is piecewise linear.

Proposition 2.2 (Reproducing Kernel)**.** *Let $\rho^{-1}(t) = a_k + b_k t,\ t \in [\tau_{k-1}, \tau_k)$ for some fixed (a_k, b_k) and the knots $\{0 \equiv \tau_0 < \tau_1 < \cdots < \tau_K \equiv 1\}$. Then, the AdaSS kernel (2.7) can be explicitly evaluated by*

$$\begin{aligned}\mathbb{K}(t,s) \;=\; & \sum_{k=1}^{\kappa_{t\wedge s}}\sum_{j=0}^{m-1}(-1)^j\Bigg\{(a_k + b_k u)\frac{(u-t)^{m-1-j}(u-s)^{m+j}}{(m-1-j)!(m+j)!} \\ & -b_k\sum_{l=0}^{m-1-j}(-1)^l\frac{(u-t)^{m-1-j-l}(u-s)^{m+1+j+l}}{(m-1-j-l)!(m+1+j+l)!}\Bigg\}\Bigg|_{\tau_{k-1}}^{t\wedge s\wedge\tau_k^-}\end{aligned} \tag{2.15}$$

where $\kappa_{t\wedge s} = \min\{k : \tau_k > t \wedge s\}$, $\kappa_{1\wedge 1} = K$, and $\{F(u)\}|_{\tau}^{\tau'} = F(\tau') - F(\tau)$.

Proposition 2.2 is general enough to cover many types of r.k.'s. For example, setting $b_k \equiv 0$ gives the piecewise-constant results for $\rho^{-1}(t)$ obtained by Pintore, et al. (2006). For another example, set $m = 2$, $K = 1$ and $\rho^{-1}(t) = a + bt$ for $t \in [0, 1]$. (By (2.5), $a = b/(e^b - 1)$ and $b \in \mathbb{R}$.) Then,

$$\mathbb{K}(t,s) = \begin{cases} \frac{a}{6}(3t^2s - t^3) + \frac{b}{12}(2t^3s - t^4), & \text{if } t \leq s, \\ \frac{a}{6}(3ts^2 - s^3) + \frac{b}{12}(2ts^3 - s^4), & \text{otherwise.} \end{cases} \tag{2.16}$$

When $b = 0$, (2.16) reduces to the OrdSS kernel for fitting cubic smoothing splines.

The derivatives of the r.k. are also of interest. In solving boundary value differential equations, the Green's function $G_m(t, u)$ has the property that $f(t) = \int_0^1 G_m(t, u) f^{(m)}(u) du$ for $f \in \mathcal{W}_m$ with zero $f^{(j)}(0)$, $j = 0, \ldots, m-1$. The equation (2.7) is such an example of applying $G_m(t, \cdot)$ to $\mathbb{K}(t, s)$ (s fixed), implying that

$$\frac{\partial^m \mathbb{K}(t,s)}{\partial t^m} = \rho^{-1}(t) G_m(s,t) = \begin{cases} (a_{\kappa_t} + b_{\kappa_t} t)\frac{(s-t)^{m-1}}{(m-1)!}, & \text{if } t \leq s \\ 0, & \text{otherwise.} \end{cases} \tag{2.17}$$

The well-defined higher-order derivatives of the kernel can be derived from (2.17), and they should be treated from knot to knot if $\rho^{-1}(t)$ is piecewise.

Lower-order l-th derivatives ($l < m$) of (2.15) can be obtained by straightforward but tedious calculations. Here, we present only the OrdSS kernels (by setting $\rho(t) \equiv 1$ in Proposition 2.2) and their derivatives to be used in next section:

$$\mathbb{K}(t,s) = \sum_{j=0}^{m-1} \frac{(-1)^j t^{m-1-j} s^{m+j}}{(m-1-j)!(m+j)!} + \frac{(-1)^m (s-t)^{2m-1}}{(2m-1)!} I(t \leq s), \tag{2.18}$$

$$\frac{\partial^l \mathbb{K}(t,s)}{\partial t^l} = \sum_{j=0}^{m-1-l} \frac{(-1)^j t^{m-1-j-l} s^{m+j}}{(m-1-j-l)!(m+j)!} + \frac{(-1)^{m+l} (s-t)^{2m-1-l}}{(2m-1-l)!} I(t \leq s). \tag{2.19}$$

Note that the kernel derivative coincides with (2.17) when $l = m$.

Remarks: Numerical methods are inevitable for approximating (2.7) when $\rho^{-1}(t)$

takes general forms. Rather than directly approximating (2.7), one may approximate $\rho^{-1}(t)$ first by a piecewise function with knot selection, then use the closed-form results derived in Proposition 2.2.

2.2.2 Local Penalty Adaptation

One of the most challenging problems for the AdaSS is selection of smoothing parameter λ and local penalty adaptation function $\rho(t)$. We use the method of cross-validation based on the leave-one-out scheme.

Conditional on $(\lambda, \rho(\cdot))$, let $\hat{f}(t)$ denote the AdaSS trained on the complete sample of size n, and $\hat{f}^{[k]}(t)$ denote the AdaSS trained on the leave-one-out sample $\{(t_i, y_i)\}_{i \neq k}$ for $k = 1, \ldots, n$. Define the score of cross-validation by

$$\mathrm{CV}(\lambda, \rho) = \frac{1}{n} \sum_{i=1}^{n} w(t_i) \big(y_i - \hat{f}^{[i]}(t_i)\big)^2. \tag{2.20}$$

For the linear predictor (2.13) with smoothing matrix $\mathbf{S}_{\lambda,\rho}$, Craven and Wahba (1979) justified that $\hat{f}(t_i) - \hat{f}^{[i]}(t_i) = \mathbf{S}_{\lambda,\rho}(y_i - \hat{f}^{[i]}(t_i))$ for $i = 1, \ldots, n$. It follows that

$$\mathrm{CV}(\lambda, \rho) = \frac{1}{n} \sum_{i=1}^{n} \frac{w(t_i)(y_i - \hat{f}(t_i))^2}{(1 - \mathbf{S}_{\lambda,\rho}[i,i])^2} = \frac{1}{n} \mathbf{y}^T \mathbf{W} (\mathbf{I} - \mathsf{diag}(\mathbf{S}_{\lambda,\rho}))^{-2} (\mathbf{I} - \mathbf{S}_{\lambda,\rho})^2 \mathbf{y}, \tag{2.21}$$

where $\mathsf{diag}(\mathbf{S}_{\lambda,\rho})$ is the diagonal part of the smoothing matrix. Replacing the elements of $\mathsf{diag}(\mathbf{S}_\lambda)$ by their average $n^{-1}\mathsf{trace}(\mathbf{S}_{\lambda,\rho})$, we obtain the so-called generalized cross-validation score

$$\mathrm{GCV}(\lambda, \rho) = \frac{1}{n} \sum_{i=1}^{n} \frac{w(t_i)(y_i - \hat{f}_\lambda(t_i))^2}{(1 - n^{-1}\mathsf{trace}(\mathbf{S}_{\lambda,\rho}))^2}. \tag{2.22}$$

For fixed $\rho(\cdot)$, the best tuning parameter λ is usually found by minimizing (2.21) or (2.22) on its log scale, i.e. by varying $\log \lambda$ on the real line.

To select $\rho(\cdot)$, consider the rules: (1) choose a smaller value of $\rho(t)$ if $f(\cdot)$ is less smooth at t; and (2) the local roughness of $f(\cdot)$ is quantified by the squared m-th derivative $\left[f^{(m)}(t)\right]^2$. If the true $f^{(m)}(t)$ were known, it is reasonable to determine the shape of $\rho(t)$ by

$$\rho^{-1}(t) \propto \left[f^{(m)}(t)\right]^2, \tag{2.23}$$

Since the true function $f(t)$ is unknown a priori, so are its derivatives. Let us consider the estimation of the derivatives. In the context of smoothing splines, nonparametric derivative estimation is usually obtained by taking derivatives of the spline estimate. Using that as an initial estimate of local roughness, we may adopt a two-step procedure for selection of $\rho(t)$, as follows.

> Step 1: run the OrdSS with $\tilde{m} = m + a$ $(a = 0, 1, 2)$ and cross-validated λ to obtain the initial function estimate: $\tilde{f}(t) = \sum_{j=0}^{m+l-1} \tilde{\alpha}_j \phi_j(t) + \sum_{i=1}^n \tilde{c}_i \widetilde{\mathbb{K}}(t, t_i)$, where $\widetilde{\mathbb{K}}(t, s) = \int_0^1 G_{m+a}(t, u) G_{m+a}(s, u) du$.
>
> Step 2: calculate $\tilde{f}^{(m)}(t)$ and use it to determine the shape of $\rho^{-1}(t)$.

In Step 1, with no prior knowledge, we start with OrdSS upon the *improper* adaptation $\tilde{\rho}(t) \equiv 1$. The extra order $a = 0, 1, 2$ imposed on penalization of derivatives $\int_0^1 [f^{(m+a)}(t)]^2 dt$ is to control the smoothness of the m-th derivative estimate. So, the choice of a depends on how smooth the estimated $\tilde{f}^{(m)}(t)$ is desired to be.

In Step 2, the m-th derivative of $\tilde{f}(t)$ can be obtained by differentiating the bases,

$$\tilde{f}^{(m)}(t) = \sum_{j=0}^{a-1} \tilde{\alpha}_{m+j} \phi_j(t) + \sum_{i=1}^{n} \tilde{c}_i \frac{\partial^m}{\partial t^m} \widetilde{\mathbb{K}}(t, t_i) \tag{2.24}$$

where $\frac{\partial^m}{\partial t^m} \widetilde{\mathbb{K}}(t, t_i)$ is calculated according to (2.19) after replacing m by $m + a$. The goodness of the spline derivative depends largely on the spline estimate itself. Rice and Rosenblatt (1983) studied the large-sample properties of spline derivatives (2.24) that have slower convergence rates than the spline itself. Given the finite sample, we find by simulation study that (2.24) tends to under-smooth, but it may capture the rough shape of the derivatives upon standardization. Figure 2.2 demonstrates the

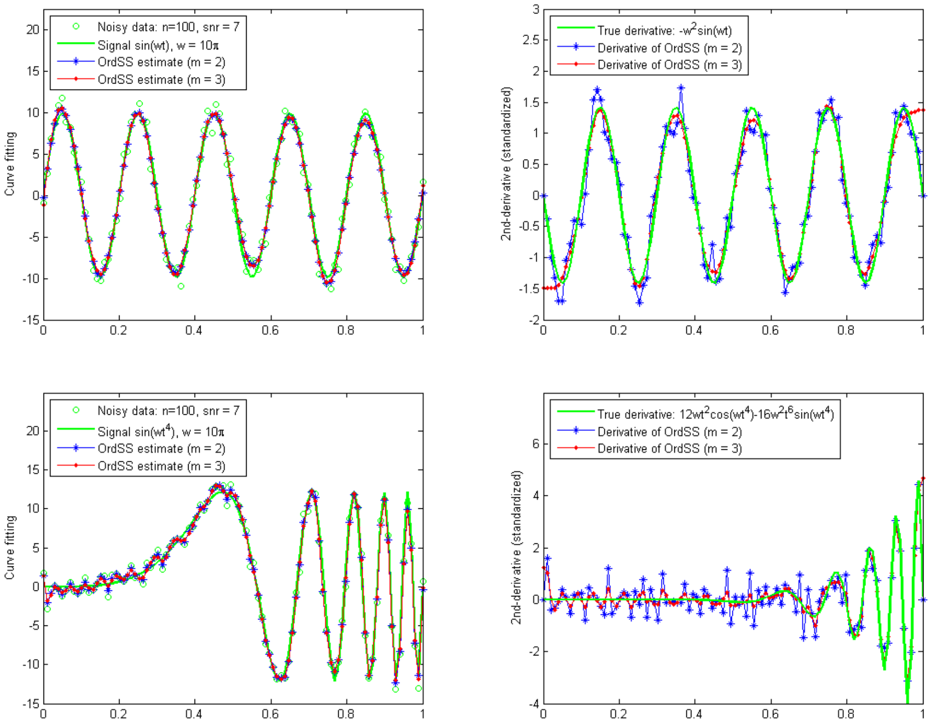


Figure 2.2: OrdSS function estimate and its 2nd-order derivative (upon standardization): scaled signal from $f(t) = \sin(\omega t)$ (upper panel) and $f(t) = \sin(\omega t^4)$ (lower panel), $\omega = 10\pi$, $n = 100$ and $\mathsf{snr} = 7$. The $\sin(\omega t^4)$ signal resembles the Doppler function in Figure 2.1; both have time-varying frequency.

estimation of the 2nd-order derivative for $f(t) = \sin(\omega t)$ and $f(t) = \sin(\omega t^4)$ using the OrdSS with $m = 2, 3$. In both sine-wave examples, the order-2 OrdSS could yield rough estimates of derivative, and the order-3 OrdSS could give smooth estimates but large bias on the boundary. Despite these downside effects, the shape of the derivative can be more or less captured, meeting our needs for estimating $\rho(t)$ below.

In determining the shape of $\rho^{-1}(t)$ from $\tilde{f}^{(m)}(t)$, we restrict ourselves to the piecewise class of functions, as in Proposition 2.2. By (2.23), we suggest to estimate $\rho^{-1}(t)$ piecewise by the method of constrained least squares:

$$\left|\tilde{f}^{(m)}(t)\right|^{2\gamma} + \varepsilon = a_0\rho^{-1}(t) = a_{\kappa_t} + b_{\kappa_t}t, \quad \kappa_t = \min\{k : \tau_k \geq t\} \tag{2.25}$$

subject to the constraints (1) positivity: $\rho^{-1}(t) > 0$ for $t \in [0, 1]$, (2) continuity: $a_k +$

$b_k\tau_k = a_{k+1} + b_{k+1}\tau_k$ for $k = 1, \ldots, K-1$ and (3) normalization: $a_0 = a_0 \int_0^1 \rho(t)dt = \sum_{k=1}^K \int_{\tau_{k-1}}^{\tau_k^-} \frac{dt}{a_k+b_k t} > 0$, for given $\gamma \geq 1$ and pre-specified knots $\{0 \equiv \tau_0 < \tau_1 < \ldots < \tau_{K-1} < 1\}$ and $\tau_K^- = 1$. Clearly, $\rho^{-1}(t)$ is an example of linear regression spline. If the continuity condition is relaxed and $b_k \equiv 0$, we obtain the piecewise-constant $\rho^{-1}(t)$.

The power transform $\left|\tilde{f}^{(m)}(t)\right|^{2\gamma}$, $\gamma \geq 1$ is used to stretch $[\tilde{f}^{(m)}(t)]^2$, in order to make up for (a) overestimation of $[f^{(m)}(t)]^2$ in slightly oscillating regions and (b) underestimation of $[f^{(m)}(t)]^2$ in highly oscillating regions, since the initial OrdSS is trained under the improper adaptation $\tilde{\rho}(t) \equiv 1$. The lower-right panel of Figure 2.2 is such an example of overestimating the roughness by the OrdSS.

The knots $\{\tau_k\}_{k=1}^K \in [0, 1]$ can be specified either regularly or irregularly. In most situations, we may choose equally spaced knots with K to be determined. For the third example in Figure 2.1 that shows flat vs. rugged heterogeneity, one may choose sparse knots for the flat region and dense knots for the rugged region, and may also choose regular knots such that $\rho^{-1}(t)$ is nearly constant in the flat region. However, for the second example in Figure 2.1 that shows jump-type heterogeneity, one should shrink the size of the interval containing the jump.

Given the knots, the parameters (λ, γ) can be jointly determined by the cross-validation criteria (2.21) or (2.22).

Remarks: The selection of the local adaptation $\rho^{-1}(t)$ is a non-trivial problem, for which we suggest to use $\rho^{-1}(t) \propto \left|\tilde{f}^{(m)}(t)\right|^{2\gamma}$ together with piecewise approximation. This can be viewed as a joint force of Abramovich and Steinberg (1996) and Pintore, et al. (2006), where the former directly executed (2.23) on each sampled t_i, and the latter directly impose the piecewise constant $\rho^{-1}(t)$ without utilizing the fact (2.23). Note that the approach of Abramovich and Steinberg (1996) may have the estimation bias of $\left|\tilde{f}^{(m)}(t)\right|^2$ as discussed above, and the r.k. evaluation by approximation based on a discrete set of $\rho^{-1}(t_i)$ remains another issue. As for Pintore, et al. (2006), the high-dimensional parametrization of $\rho^{-1}(t)$ requires nonlinear optimization whose stability is usually of concern. All these issues can be resolved by our approach, which therefore has both analytical and computational advantages.

2.3 AdaSS Properties

Recall that the OrdSS (2.2) has many interesting statistical properties, e.g.

(a) it is a natural polynomial spline of degree $(2m-1)$ with knots placed $t_1, \ldots, t_n$ (hence, the OrdSS with $m = 2$ is usually referred to the *cubic* smoothing spline);

(b) it has a Bayesian interpretation through the duality between RKHS and Gaussian stochastic processes (Wahba, 1990, §1.4-1.5);

(c) its large sample properties of consistency and rate of convergence depends on the smoothing parameter (Wahba, 1990, §4 and references therein);

(d) it has an equivalent kernel smoother with variable bandwidth (Silverman, 1984).

Similar properties are desired to be established for the AdaSS. We leave (c) and (d) to future investigation. For (a), the AdaSS properties depend on the r.k. (2.7) and the structure of local adaptation $\rho^{-1}(t)$. Pintore, et al. (2006) investigated the piecewise-constant case of $\rho^{-1}(t)$ and argued that the AdaSS is also a piecewise polynomial spline of degree $(2m-1)$ with knots on $t_1, \ldots, t_n$, while it allows for more flexible lower-order derivatives than the OrdSS. The piecewise-linear case of $\rho^{-1}(t)$ can be analyzed in a similar manner. For the AdaSS (2.8) with the r.k. (2.15),

$$f(t) = \sum_{j=0}^{m-1} \alpha_j \phi_j(t) + \sum_{i:t_i \le t} c_i \mathbb{K}(t, t_i) + \sum_{i:t_i > t} c_i \mathbb{K}(t, t_i)$$

with degree $(m-1)$ in the first two terms, and degree $2m$ in the third term (after one degree increase due to the trend of $\rho^{-1}(t)$).

For (b), the AdaSS based on the r.k. (2.7) corresponds to the following zero-mean Gaussian process

$$Z(t) = \int_0^1 \rho^{-1/2}(u) G_m(t,u) dW(u), \quad t \in [0,1] \tag{2.26}$$

where $W(t)$ denotes the standard Wiener process. The covariance kernel is given by $\mathbb{E}[Z(t)Z(s)] = \mathbb{K}(t,s)$. When $\rho(t) \equiv 1$, the Gaussian process (2.26) is the $(m-1)$-fold integrated Wiener process studied by Shepp (1966). Following Wahba (1990; §1.5), consider a random effect model $F(t) = \boldsymbol{\alpha}^T \boldsymbol{\phi}(t) + b^{1/2} Z(t)$, then the best linear unbiased estimate of $F(t)$ from noisy data $Y(t) = F(t) + \varepsilon$ corresponds to the AdaSS solution (2.8), provided that the smoothing parameter $\lambda = \sigma^2 / nb$. Moreover, one may set an improper prior on the coefficients $\boldsymbol{\alpha}$, derive the posterior of $F(t)$, then obtain a Bayesian type of confidence interval estimate; see Wahba (1990; §5). Such confidence intervals on the sampling points are given in (2.14).

The inverse function $\rho^{-1}(t)$ in (2.26) can be viewed as the measurement of nonstationary volatility (or, variance) of the m-th derivative process such that

$$\frac{d^m Z(t)}{dt^m} = \rho^{-1/2}(t) \frac{dW(t)}{dt}. \tag{2.27}$$

By the duality between Gaussian processes and RKHS, $d^m Z(t)/dt^m$ corresponds to $f^{(m)}(t)$ in (2.6). This provides an alternative reasoning for (2.23) in the sense that $\rho^{-1}(t)$ can be determined locally by the second-order moments of $d^m Z(t)/dt^m$.

To illustrate the connection between the AdaSS and the Gaussian process with nonstationary volatility, consider the case $m = 1$ and piecewise-constant $\rho^{-1}(t) = a_k, t \in [\tau_{k-1}, \tau_k)$ for some $a_k > 0$ and pre-specified knots $0 \equiv \tau_0 < \tau_1 < \cdots < \tau_K \equiv 1$. By (2.15), the r.k. is given by

$$\mathbb{K}(t,s) = \sum_{k=1}^{\kappa_{t\wedge s}} a_k \Big(t \wedge s - \tau_{k-1} \Big), \quad t,s \in [0,1]. \tag{2.28}$$

By (2.26) and (2.27), the corresponding Gaussian process satisfies that

$$dZ(t) = \sqrt{a_k} dW(t), \quad \text{for } t \in [\tau_{k-1}, \tau_k),\ k = 1, \ldots, K \tag{2.29}$$

where a_k is the local volatility from knot to knot. Clearly, $Z(t)$ reduces to the standard Wiener process $W(t)$ when $a_k \equiv 1$. As one of the most appealing features, the construction (2.29) takes into account the jump diffusion, which occurs in $[\tau_{k-1}, \tau_k)$ when the interval shrinks and the volatility a_k blows up.

2.4 Experimental Results

The implementation of OrdSS and AdaSS differs only in the formation of $\mathbf{\Sigma}$ through the r.k. $\mathbb{K}(t, s)$. It is most crucial for the AdaSS to select the local adaptation $\rho^{-1}(t)$. Consider the four examples in Figure 2.1 with various scenarios of heterogeneous smoothness. The curve fitting results by the OrdSS with $m = 2$ (i.e. cubic smoothing spline) are shown in Figure 2.3, as well as the 95% confidence intervals for the last two cases. In what follows, we discuss the need of local adaptation in each case, and compare the performance of OrdSS and AdaSS.

Simulation Study.

The Doppler and HeaviSine functions are often taken as the benchmark examples for testing adaptive smoothing methods. The global smoothing by the OrdSS (2.2) may fail to capture the heterogeneous smoothness, and as a consequence, over-smooth the region where the signal is relatively smooth and under-smooth where the signal is highly oscillating. For the Doppler case in Figure 2.3, the underlying signal increases its smoothness with time, but the OrdSS estimate shows lots of wiggles for $t > 0.5$. In the HeaviSine case, both jumps at $t = 0.3$ and $t = 0.72$ could be captured by the OrdSS, but the smoothness elsewhere is sacrificed as a compromise.

The AdaSS fitting results with piecewise-constant $\rho^{-1}(t)$ are shown in Figure 2.4.

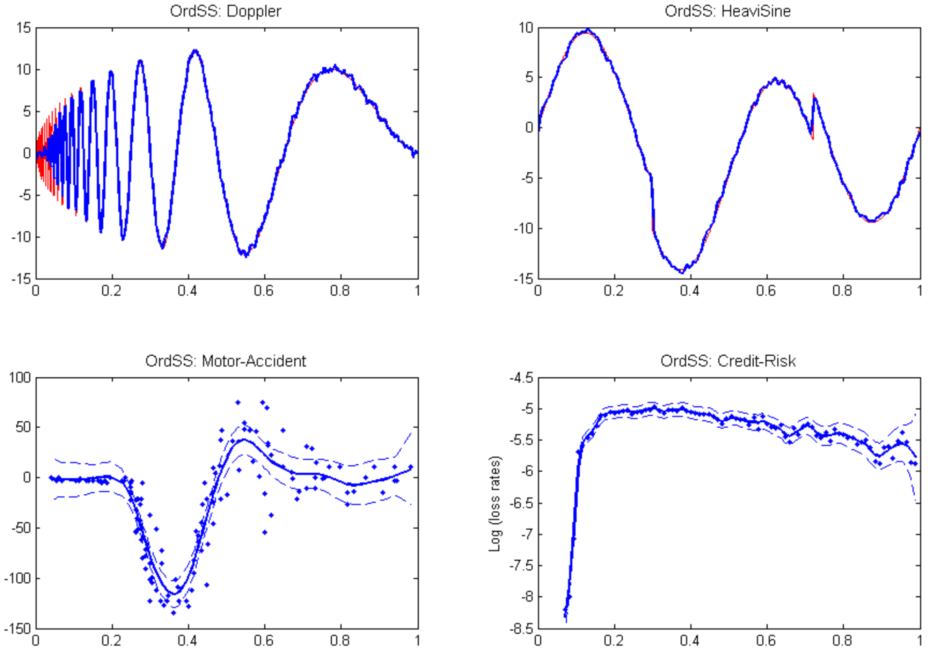


Figure 2.3: OrdSS curve fitting with $m = 2$ (shown by solid lines). The dashed lines represent 95% confidence intervals. In the credit-risk case, the log loss rates are considered as the responses, and the time-dependent weights are specified proportional to the number of replicates.

Initially, we fitted the OrdSS with cross-validated tuning parameter λ, then estimated the 2nd-order derivatives by (2.24). To estimate the piecewise-constant local adaptation $\rho^{-1}(t)$, we pre-specified 6 equally spaced knots (including 0 and 1) for the Doppler function, and pre-specified the knots $(0, 0.295, 0.305, 0.5, 0.715, 0.725, 1)$ for the HeaviSine function. (Note that such knots can be determined graphically based on the plots of the squared derivative estimates.) Given the knots, the function $\rho^{-1}(t)$ is estimated by constrained least squares (2.25) conditional on the power transform parameter γ, where the best γ^* is found jointly with λ by minimizing the score of cross-validation (2.21). In our implementation, we assumed that both tuning parameters are bounded such that $\log \lambda \in [-40, 10]$ and $\gamma \in [1, 5]$, then performed the bounded nonlinear optimization. Table 2.1 gives the numerical results in both simulation cases, including the cross-validated (λ, γ), the estimated variance of noise (cf. true $\sigma^2 = 1$), and the

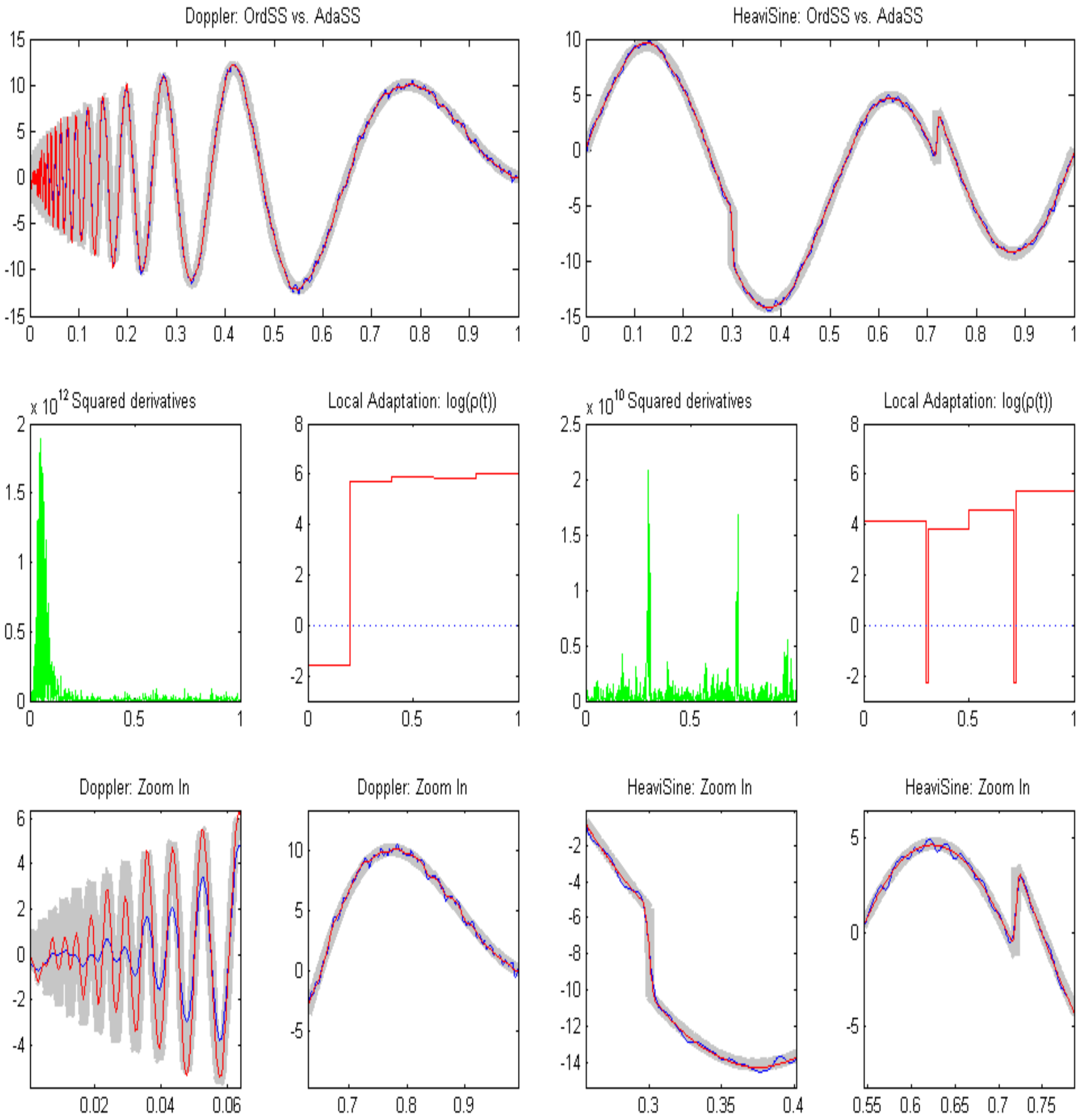


Figure 2.4: Simulation study of Doppler and HeaviSine functions: OrdSS (blue), AdaSS (red) and the heterogeneous truth (light background).

Table 2.1: AdaSS parameters in Doppler and HeaviSine simulation study.

Simulation	n	snr	λ	γ	$\hat{\sigma}^2$	dof	CV
Doppler	2048	7	1.21e-012	2.13	0.9662	164.19	2180.6
HeaviSine	2048	7	1.86e-010	4.95	1.0175	35.07	2123.9

equivalent degrees of freedom.

Figure 2.4 (bottom panel) shows the improvement of AdaSS over OrdSS after zooming in. The middle panel shows the data-driven estimates of the adaptation function $\rho(t)$ plotted at log scale. For the Doppler function, the imposed penalty $\rho(t)$ in the first interval $[0, 0.2)$ is much less than that in the other four intervals, resulting in heterogenous curve fitting. For the HeaviSine function, low penalty is assigned to where the jump occurs, while the relatively same high level of penalty is assigned to other regions. Both simulation examples demonstrate the advantage of using the AdaSS. A similar simulation study of the Doppler function can be found in Pintore, et al.(2006) who used a small sample size ($n = 128$) and high-dimensional optimization techniques.

Motorcycle-Accident Data.

Silverman (1985) originally used the motorcycle-accident experimental data to test the non-stationary OrdSS curve fitting; see the confidence intervals in Figure 2.3. Besides error non-stationarity, smoothness non-homogeneity is another important feature of this data set. To fit such data, the treatment from a heterogenous smoothness point of view is more appropriate than using only the non-stationary error treatment.

Recall that the data consists of acceleration measurements (subject to noise) of the head of a motorcycle rider during the course of a simulated crash experiment. Upon the crash occurrence, it decelerates to about negative 120 gal (a unit gal is 1 centimeter per second squared). The data indicates clearly that the acceleration rebounds well above its original level before setting back. Let us focus on the period prior to the crash happening at $t \approx 0.24$ (after time re-scaling), for which it is reasonable to

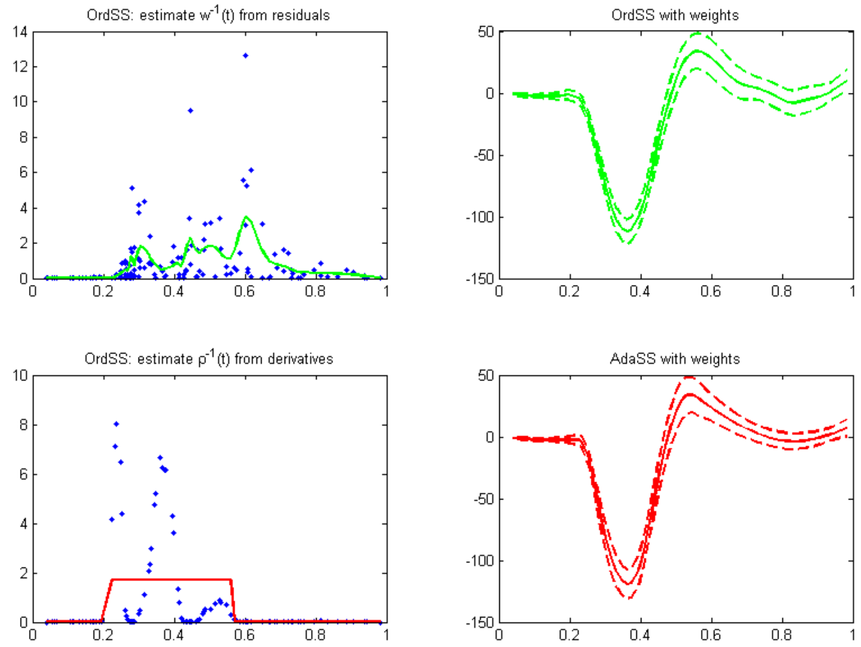


Figure 2.5: Non-stationary OrdSS and AdaSS for Motorcycle-Accident Data.

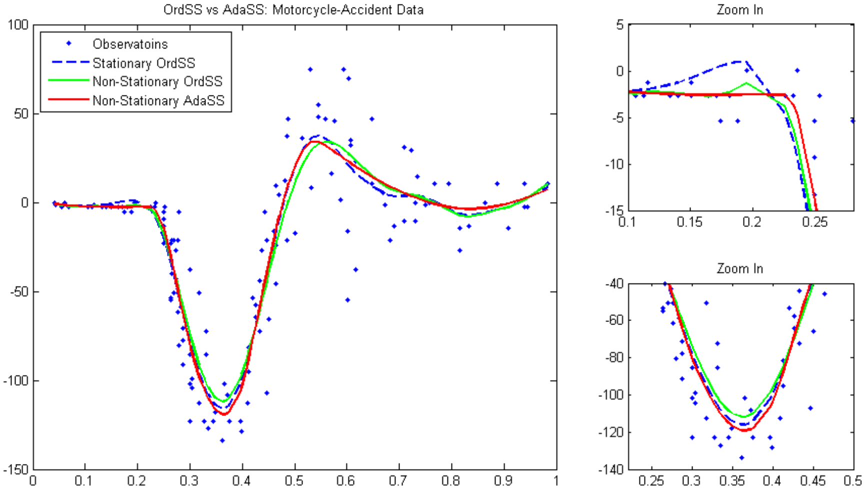


Figure 2.6: OrdSS, non-stationary OrdSS and AdaSS: performance in comparison.

assume the acceleration stays constant. However, the OrdSS estimate in Figure 2.3 shows a slight bump around $t = 0.2$, a false discovery.

In presence of non-stationary errors $N(0, \sigma^2/w(t_i))$ with unknown weights $w(t_i)$, a common approach is to begin with an unweighted OrdSS and estimate $w^{-1}(t_i)$ from the local residual sums of squares; see Silverman (1985). The local moving averaging (e.g. loess) is usually performed; see the upper-left panel in Figure 2.5. Then, the smoothed weight estimates are plugged into (2.2) to run the non-stationary OrdSS. To perform AdaSS (2.6) with the same weights, we estimated the local adaptation $\rho^{-1}(t)$ from the derivatives of non-stationary OrdSS. Both results are shown in Figure 2.5. Compared to Figure 2.3, both non-stationary spline models result in tighter estimates of 95% confidence intervals than the unweighted OrdSS.

Figure 2.6 compares the performances of stationary OrdSS, non-stationary OrdSS and non-stationary AdaSS. After zooming in around $t \in [0.1, 0.28]$, it is clear that AdaSS performs better than both versions of OrdSS in modeling the constant acceleration before the crash happening. For the setting-back process after attaining the maximum acceleration at $t \approx 0.55$, AdaSS gives a more smooth estimate than OrdSS. Besides, they differ in estimating the minimum of the acceleration curve at $t \approx 0.36$, where the non-stationary OrdSS seems to have underestimated the magnitude of deceleration, while the estimate by the non-stationary AdaSS looks reasonable.

Credit-Risk Data.

The last plot in Figure 2.1 represents a tweaked sample in retail credit risk management. It is of our interest to model the growth or the *maturation curve* of the log of loss rates in lifetime (month-on-book). Consider the sequence of vintages booked in order from 2000 to 2006. Their monthly loss rates are observed up to December 2006 (right truncation). Aligning them to the same origin, we have the observations accumulated more in smaller month-on-book and less in larger month-on-book, i.e., the sample size decreases in lifetime. As a consequence, the simple pointwise or moving

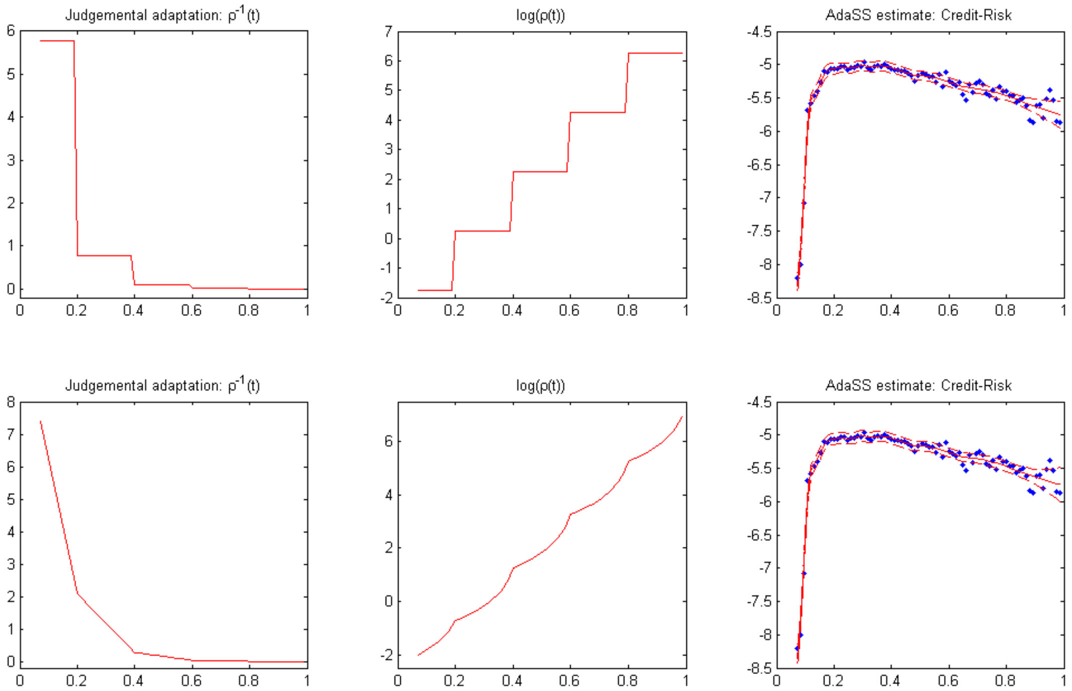


Figure 2.7: AdaSS estimate of maturation curve for Credit-Risk sample: piecewise-constant $\rho^{-1}(t)$ (upper panel) and piecewise-linear $\rho^{-1}(t)$ (lower panel).

averages have increasing variability. Nevertheless, our experiences tell that the maturation curve is expected to have increasing degrees of smoothness once the lifetime exceeds certain month-on-book. This is also a desired property of the maturation curve for the purpose of extrapolation.

By (2.4), spline smoothing may work on the pointwise averages subject to non-stationary errors $N(0, \sigma^2/w(t_i))$ with weights determined by the number of replicates. The non-stationary OrdSS estimate is shown in Figure 2.3, which however does not smooth out the maturation curve for large month-on-book. The reasons of the unstability in the rightmost region could be (a) OrdSS has the intrinsic boundary bias for especially small $w(t_i)$, and (b) the raw inputs for training purpose are contaminated and have abnormal behavior. The exogenous cause of contamination in (b) will be explained in Chapter III.

Alternatively, we applied the AdaSS with judgemental prior on heterogeneous smoothness of the target function. For demonstration, we purposefully specified

$\rho^{-1}(t)$ to take the forms of Figure 2.7, including (a) piecewise-constant and (b) piecewise-linear. Both specifications are non-decreasing in lifetime and they span about the same range. The curve fitting results for both $\rho^{-1}(t)$ are nearly identical upon visual inspection. Compared with the OrdSS performance in the large month-on-book region, the AdaSS yields smoother estimate of maturation curve, as well as tighter confidence intervals based on judgemental prior. The AdaSS estimate of maturation curve, transformed to the original scale, is plotted by the red solid line in Figure 2.1.

This credit-risk sample will appear again in Chapter III of vintage data analysis. The really interesting story is yet to unfold.

2.5 Summary

In this chapter we have studied the adaptive smoothing spline for modeling heterogeneously 'smooth' functions including possible jumps. In particular, we studied the AdaSS with piecewise adaptation of local penalty and derived the closed-form reproducing kernels. To determine the local adaptation function, we proposed a shape estimation procedure based the function derivatives, together with the method of cross-validation for parameter tuning. Four different examples, each having a specific feature of heterogeneous smoothness, were used to demonstrate the improvement of AdaSS over OrdSS.

Future works include the AdaSS properties w.r.t. (c) and (d) listed in Section 2.3. Its connection with a time-warping smoothing spline is also under our investigation. However, it is not clear how the AdaSS can be extended to multivariate smoothing in a tensor-product space, as an analogy of the smoothing spline ANOVA models in Gu (2002). In next chapter, we will consider the application of AdaSS in a dual-time domain, where the notion of smoothing spline will be generalized via other Gaussian processes.

2.6 Technical Proofs

Proof to Proposition 2.1: by minor modification on Wahba (1990; §1.2-1.3). By Taylor's theorem with remainder, any function in $\mathcal{W}_m$ (2.3) can be expressed by

$$f(t) = \sum_{j=0}^{m-1} \frac{t^j}{j!} f^{(j)}(0) + \int_0^1 G_m(t,u) f^{(m)}(u) du \tag{2.30}$$

where $G_m(t,u) = \frac{(t-u)_+^{(m-1)}}{(m-1)!}$ is the Green's function for solving the boundary value problem: if $f_1^{(m)} = g$ with $f_1 \in \mathcal{B}_m = \{f : f^{(k)}(0) = 0, \text{ for } k = 0, 1, \ldots, m-1\}$, then $f_1(t) = \int_0^1 G_m(t,u) g(u) du$. The remainder functions $f_1(t) = \int_0^1 G_m(t,u) f^{(m)}(u) du$ lie in the subspace $\mathcal{H} = \mathcal{W}_m \cap \mathcal{B}_m$. Given the bounded integrable function $\rho(t) > 0$ on $t \in [0,1]$, associate $\mathcal{H}$ with the inner product

$$\langle f_1, g_1 \rangle_{\mathcal{H}} = \int_0^1 f_1^{(m)}(u) g_1^{(m)}(u) \lambda(u) du, \tag{2.31}$$

and the induced square norm $\|f_1\|_{\mathcal{H}}^2 = \int_0^1 \lambda(u)[f_1^{(m)}(u)]^2 du$. By similar arguments to Wahba (1990; §1.2), one may verify that $\mathcal{H}$ is an RKHS with r.k. (2.7), where the reproducing property

$$\langle f_1(\cdot), \mathbb{K}(t, \cdot) \rangle_{\mathcal{H}} = f_1(t), \quad \forall f_1 \in \mathcal{H} \tag{2.32}$$

follows that $\frac{\partial^m}{\partial t^m} \mathbb{K}(t,s) = \rho(t)^{-1} G_m(s,t)$.

Then, using the arguments of Wahba (1990; §1.3) would prove the finite-dimensional expression (2.8) that minimizes the AdaSS objective functional (2.6).

Proof to Proposition 2.2: The r.k. with piecewise $\rho^{-1}(t)$ equals

$$\mathbb{K}(t,s) = \sum_{k=1}^{\kappa_{t \wedge s}} \int_{\tau_{k-1}}^{t \wedge s \wedge \tau_k^-} (a_k + b_k u) \frac{(u-t)^{m-1}}{(m-1)!} \frac{(u-s)^{m-1}}{(m-1)!} du.$$

For any $\tau_{k-1} \le t_1 \le t_2 < \tau_k$, using the successive integration by parts,

$$\begin{aligned}
\int_{t_1}^{t_2} (a_k + b_k u)\frac{(u-t)^{m-1}}{(m-1)!}\frac{(u-s)^{m-1}}{(m-1)!}du &= \int_{t_1}^{t_2} (a_k + b_k u)\frac{(u-t)^{m-1}}{(m-1)!}d\frac{(u-s)^{m}}{m!} \\
= \quad (a_k + b_k u)\frac{(u-t)^{m-1}}{(m-1)!}\frac{(u-s)^{m}}{m!}\Big|_{t_1}^{t_2} &- \int_{t_1}^{t_2} (a_k + b_k u)\frac{(u-t)^{m-2}}{(m-2)!}\frac{(u-s)^{m}}{m!}du \\
-b_k \int_{t_1}^{t_2} \frac{(u-t)^{m-1}}{(m-1)!}\frac{(u-s)^{m}}{m!}du &= \cdots ,
\end{aligned}$$

it is straightforward (but tedious) to derive that

$$\begin{aligned}
&\int_{t_1}^{t_2} (a_k + b_k u)\frac{(u-t)^{m-1}}{(m-1)!}\frac{(u-s)^{m-1}}{(m-1)!}du \\
&= \sum_{j=0}^{m-1}(-1)^j (a_k + b_k u)\frac{(u-t)^{m-1-j}}{(m-1-j)!}\frac{(u-s)^{m+j}}{(m+j)!}\Big|_{t_1}^{t_2} - b_k \sum_{j=0}^{m-1}(-1)^j T_j
\end{aligned}$$

in which T_j denotes the integral evaluations for $j = 0, 1, \ldots, m-1$

$$T_j \equiv \int_{t_1}^{t_2} \frac{(u-t)^{m-1-j}}{(m-1-j)!}\frac{(u-s)^{m+j}}{(m+j)!}du = \sum_{l=0}^{m-1-j}(-1)^l \frac{(u-t)^{m-1-j-l}(u-s)^{m+1+j+l}}{(m-1-j-l)!(m+1+j+l)!}\Big|_{t_1}^{t_2}.$$

Choosing $\tau_1 = \tau_{k-1}, \tau_2 = t \wedge s \wedge \tau_k^-$ for $k = 1, \ldots, \kappa_{t\wedge s}$ and adding them all gives the final r.k. evaluation by

$$\begin{aligned}
\mathbb{K}(t,s) \quad = \quad & \sum_{k=1}^{\kappa_{t\wedge s}}\sum_{j=0}^{m-1}(-1)^j \Bigg\{ (a_k + b_k u)\frac{(u-t)^{m-1-j}(u-s)^{m+j}}{(m-1-j)!(m+j)!} \\
& -b_k \sum_{l=0}^{m-1-j}(-1)^l \frac{(u-t)^{m-1-j-l}(u-s)^{m+1+j+l}}{(m-1-j-l)!(m+1+j+l)!} \Bigg\}\Bigg|_{\tau_{k-1}}^{t\wedge s\wedge\tau_k^-}.
\end{aligned}$$

CHAPTER III

Vintage Data Analysis

3.1 Introduction

The vintage data represents a special class of functional, longitudinal or panel data with dual-time characteristics. It shares the cross-sectional time series feature such that there are J subjects each observed at calendar time points $\{t_{jl},\ l = 1, \ldots, L_j\}$, where L_j is the total number of observations on the j-th subject. Unlike purely longitudinal setup, each subject j we consider corresponds to a vintage originated at time v_j such that $v_1 < v_2 < \ldots < v_J$ and the elapsed time $m_{jl} = t_{jl} - v_j$ measures the lifetime or age of the subject. To this end, we denote the vintage data by

$$\Big\{y(m_{jl}, t_{jl}; v_j), \quad l = 1, \ldots, L_j,\ j = 1, \ldots, J\Big\}, \tag{3.1}$$

where the dual-time points are usually sampled from the regular grids such that for each vintage j, $m_{jl} = 0, 1, 2, \ldots$ and $t_{jl} = v_j, v_j + 1, v_j + 2, \ldots$. We call the grids of such $(m_{jl}, t_{jl}; v_j)$ a *vintage diagram*. Also observed on the vintage diagram could be the covariate vectors $\mathbf{x}(m_{jl}, t_{jl}; v_j) \in \mathbb{R}^p$. In situations where more than one segment of dual-time observations are available, we may denote the (static) segmentation variables by $\mathbf{z}_k \in \mathbb{R}^q$, and denote the vintage data by $\{y(m_{kjl}, t_{kjl}; v_{kj}), \mathbf{x}(m_{kjl}, t_{kjl}; v_{kj})\}$ for segment $k = 1, \ldots, K$. See Figure 1.3 for an illustration of vintage diagram and dual-time collection of observations, as well as the hierarchical structure of multi-segment credit portfolios.

The vintage diagram could be truncated in either age or time, and several interesting prototypes are illustrated in Figure 3.1. The triangular prototype on the top-left panel is the most typical as the time series data are often subject to the systematic truncation from right (say, observations up to today). Each vintage series evolves in the 45° direction such that for the same vintage v_j, the calendar time t_j and the lifetime m_j move at the same speed along the diagonal — therefore, the vintage diagram corresponds to the hyperplane $\{(t, m, v) : v = t - m\}$ in the 3D space. In practice, there exist other prototypes of vintage data truncation. The rectangular diagram is obtained after truncation in both lifetime and calendar time. The corporate default rates released by Moody's, shown in Table 1.1, are truncated in lifetime to be 20 years maximum. The other trapezoidal prototype corresponds to the time-snapshot selection common to credit risk modeling exercises. Besides, the U.S. Treasury rates released for multiple maturity dates serve as an example of the parallelogram prototype.

The age, time and vintage dimensions are all of potential interest in financial risk management. The empirical evidences tell that the age effects of self-maturation nature are often subject to smoothness in variation. The time effects are often dynamic and volatile due to macroeconomic environment. Meanwhile, the vintage effects are conceived as the forward-looking measure of performance since origination, and they could be treated as random effects upon appropriate structural assumption. Bearing in mind these practical considerations, we aim to develop a formal statistical framework for vintage data analysis (VDA), by integrating ideas from functional data analysis (FDA), longitudinal data analysis (LDA)[1] and time series analysis (TSA). To achieve this, we will take a Gaussian process modeling approach, which is flexible

[1] Despite the overlap between FDA and LDA, we try to differentiate them methodologically by tagging FDA with nonparametric smoothing and tagging LDA with random-effect modeling.

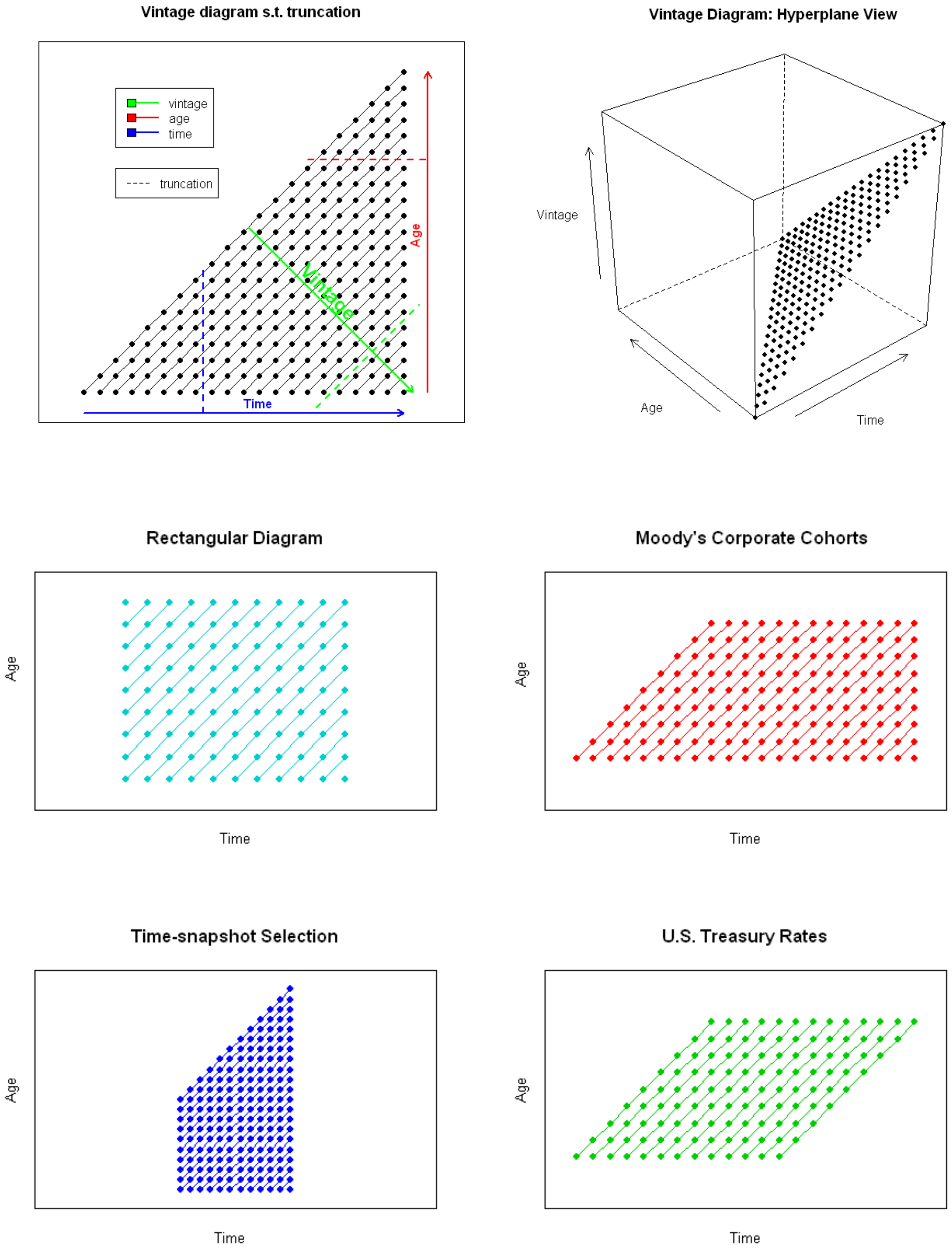


Figure 3.1: Vintage diagram upon truncation and exemplified prototypes.

enough to cover a diversity of scenarios.

This chapter develops the VDA framework for continuous type of responses (usually, rates). It is organized as follows. In Section 3.2, we discuss the maturation-exogenous-vintage (MEV) decomposition in general. The MEV models based on Gaussian processes are presented in Section 3.3, where we discuss Kriging, smoothing spline and kernel methods. An efficient backfitting algorithm is provided for model estimation. Section 3.4 covers the semiparametric regression given the dual-time covariates and segmentation variables. Computational results are presented in Section 3.5, including a simulation study for assessing the MEV decomposition technique, and real applications in both corporate and retail credit risk cases. We conclude the chapter in Section 3.6 and give technical proofs in the last section.

3.2 MEV Decomposition Framework

Let $\mathcal{V} = \{v_j : 0 \equiv v_1 < v_2 < \ldots < v_J < \infty\}$ be a discrete set of origination times of J vintages, and define the dual-time domain

$$\Omega = \Big\{(m, t; v) : \; m \geq 0, \; t = v + m, \; v \in \mathcal{V}\Big\},$$

where m denotes the lifetime and t denotes the calendar time. Unlike a usual 2D (m, t) space, the dual-time domain Ω consists of isolated 45° lines $\{(m, t) : t - m = v_j, m \geq 0\}$ for each fixed vintage $v_j \in \mathcal{V}$. The triangular vintage diagram in Figure 3.1 is the discrete counterpart of Ω upon right truncation in calendar time.

Let Ω be the support of the vintage data (3.1), for which we propose the *MEV (maturation-exogenous-vintage) decomposition* modeling framework

$$\eta(y(m, t; v)) = f(m) + g(t) + h(v) + \varepsilon, \tag{3.2}$$

where η is a pre-specified transform function, ε is the random error, and the three MEV components have the following practical interpretations:

a) $f(m)$ represents the *maturation curve* characterizing the endogenous growth,

b) $g(t)$ represents the *exogenous influence* by macroeconomic conditions,

c) $h(v)$ represents the *vintage heterogeneity* measuring the origination quality.

For the time being, let f, g, h be open to flexible choices of parametric, nonparametric or stochastic process models. We begin with some general remarks about model identification.

Suppose the maturation curve $f(m)$ in (3.2) absorbs the intercept effect, and the constraints $\mathbb{E}g = \mathbb{E}h = 0$ are set as usual:

$$\mathbb{E}\int_{t_{\min}}^{t_{\max}} g(t)dt = \mathbb{E}\int_{\mathcal{V}} h(v)dv = 0. \tag{3.3}$$

The question is, do these constraints suffice to ensure the model identifiability? It turns out for the dual-time domain Ω, all the MEV decomposition models are subject to the linear dependency, as stated formally in the following lemma.

Lemma 3.1 (Identification). *For the MEV models (3.2) defined on the dual-time domain Ω, write $f = f_0 + f_1$ with the linear part f_0 and the nonlinear part f_1 and similarly write $g = g_0 + g_1$, $h = h_0 + h_1$. Then, one of f_0, g_0, h_0 is not identifiable.*

The identification (or collinearity) problem in Lemma 3.1 is incurred by the hyperplane nature of the dual-time domain Ω in the 3D space: $\{(t, m, v) : v = t - m\}$ (see Figure 3.1 top right panel). We need to break up such linear dependency in order to make the MEV models estimable. A practical way is to perform binning in age, time or vintage direction. For example, given the vintage data of triangular prototype, one may group the large m region, the small t region and the large v region, as these regions have few observations. A commonly used binning procedure for $h(v)$ is to assume the piecewise-constant vintage effects by grouping monthly vintages every quarter or every year.

An alternative strategy for breaking up the linear dependency is to directly remove the trend for one of f, g, h functions. Indeed, this strategy is simple to implement, provided that f, g, h all could be expanded by certain basis functions. Then, one only needs to fix the linear bases for any two of f, g, h. By default, we remove the trend of h and let it measure only the nonlinear heterogeneity. A word of caution is in order in situations where there does exist an underlying trend $h_0(v)$, then it would be absorbed by $f(m)$ and $g(t)$. One may retrieve $h_0(v) = \gamma v$ by simultaneously adjusting $f(m) \rightarrow f(m) + \gamma m$ and $g(t) \rightarrow g(t) - \gamma t$. To determine the slope γ requires *extra* knowledge about the trend of at least one of f, g, h. The trend removal for $h(v)$ is needed only when there is no such extra knowledge. The MEV decomposition with zero-trend vintage effects may be referred to as the *ad hoc* approach.

In what follows we review several existing approaches that are related to MEV decomposition. The first example is a conventional age-period-cohort (APC) model in social sciences of demography and epidemiology. The second example is the generalized additive model (GAM) in nonparametric statistics, where the smoothing spline is used as the scatterplot smoother. The third example is an industry GAMEED approach that may reduce to a sequential type of MEV decomposition. The last example is another industrial DtD technique involving the interactions of MEV components. They all can be viewed as one or another type of vintage data analysis.

The APC Model.

Suppose there are I distinct m-values and L distinct t-values in the vintage data. By (3.2), setting $f(m) = \mu + \sum_{i=1}^{I} \alpha_i I(m = m_i)$, $g(t) = \sum_{l=1}^{L} \beta_l I(t = t_l)$ and $h(v) =$

$\sum_{j=1}^{J} \gamma_j I(v = v_j)$ gives us the the APC (age-period-cohort) model

$$\eta(y(m_i, t_l; v_j)) = \mu + \alpha_i + \beta_l + \gamma_j + \varepsilon, \tag{3.4}$$

where the vintage effects $\{\gamma_j' s\}$ used to be called the cohort effects; see Mason, et al. (1973) and Mason and Wolfinger (2002). Assume $\sum_i \alpha_i = \sum_l \beta_l = \sum_j v_j = 0$, as usual; then one may reformulate the APC model as $\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\varepsilon}$, where $\mathbf{y}$ is the the vector consisting of $\eta(y(m_i, t_l; v_j))$ and

$$\mathbf{X} = \left[\mathbf{1}, \mathbf{x}_1^{(\alpha)}, \ldots, \mathbf{x}_{I-1}^{(\alpha)}, \mathbf{x}_1^{(\beta)}, \ldots, \mathbf{x}_{L-1}^{(\beta)}, \mathbf{x}_1^{(\gamma)}, \ldots, \mathbf{x}_{J-1}^{(\gamma)}\right]$$

with dummy variables. For example, $\mathbf{x}_i^{(\alpha)}$ is defined by $\mathbf{x}_i^{(\alpha)}(m) = I(m = m_i) - I(m = m_I)$ for $i = 1, \ldots, I-1$. Without loss of generality, let us select the factorial levels (m_I, t_L, v_J) to satisfy $m_I - t_L + v_J = 0$ (after reshuffling the indices of v).

The identification problem of the APC model (3.4) follows Lemma 3.1, since the regression matrix is linearly dependent through

$$\sum_{i=1}^{I-1}(m_i - \bar{m})\mathbf{x}_i^{(\alpha)} - \sum_{l=1}^{L-1}(t_l - \bar{t}\ )\mathbf{x}_l^{(\beta)} + \sum_{j=1}^{J-1}(v_j - \bar{v})\mathbf{x}_j^{(\gamma)} = \bar{m} - \bar{t} + \bar{v}, \tag{3.5}$$

where $\bar{m} = \sum_{i=1}^{I} m_i/I, \bar{t} = \sum_{l=1}^{L} t_l/L$ and $\bar{v} = \sum_{j=1}^{J} v_j/J$. So, the ordinary least squares (OLS) cannot be used for fitting $\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\varepsilon}$. This is a key observation by Kupper (1985) that generated long-term discussions and debates on the estimability of the APC model. Among others, Fu (2000) suggested to estimate the APC model by ridge regression, whose limiting case leads to a so-called intrinsic estimator $\widehat{\boldsymbol{\theta}} = (\mathbf{X}^T\mathbf{X})^{+}\mathbf{X}^T\mathbf{y}$, where $\mathbf{A}^{+}$ is the Moore-Penrose generalized inverse of $\mathbf{A}$ satisfying that $\mathbf{A}\mathbf{A}^{+}\mathbf{A} = \mathbf{A}$.

The discrete APC model tends to massively identify the effects at the grid level and often ends up with unstable estimates. Since the vintage data are always unbalanced in m, t or v projection, there would be very limited observations for estimating certain factorial effects in (3.4). The APC model is formulated over-flexibly in the sense that it does not regularize the grid effect by neighbors. Fu (2008) suggested to use the spline model for $h(v)$; however, such spline regularization would not bypass the identification problem unless it does not involve a linear basis.

The Additive Splines.

If f, g, h are all assumed to be unspecified smooth functions, the MEV models would become the generalized additive models (GAM); see Hastie and Tibshirani (1990). For example, using the cubic smoothing spline as the scatterplot smoother, we have the *additive splines* model formulated by

$$\begin{aligned}\arg\min_{f,g,h}\Bigg\{\sum_{j=1}^{J}\sum_{l=1}^{L_j}\Big[\eta(y_{jl}) - f(m_{jl}) - g(t_{jl}) - h(v_j)\Big]^2 + \lambda_f \int \Big[f^{(2)}(m)\Big]^2 dm \\ + \lambda_g \int \Big[g^{(2)}(t)\Big]^2 dt + \lambda_h \int \Big[h^{(2)}(v)\Big]^2 dv\Bigg\}, \qquad (3.6)\end{aligned}$$

where $\lambda_f, \lambda_g, \lambda_h > 0$ are the tuning parameters separately controlling the smoothness of each target function; see (2.2).

Usually, the optimization (3.6) is solved by the backfitting algorithm that iterates

$$\begin{aligned}\hat{f} &\leftarrow \mathbf{S}_f\Big[\big\{\eta(y_{jl}) - \hat{g}(t_{jl}) - \hat{h}(v_j)\big\}_{j,l}\Big] \\ \hat{g} &\leftarrow \mathbf{S}_g\Big[\big\{\eta(y_{jl}) - \hat{f}(m_{jl}) - \hat{h}(v_j)\big\}_{j,l}\Big] \text{ (centered)} \qquad (3.7) \\ \hat{h} &\leftarrow \mathbf{S}_h\Big[\big\{\eta(y_{jl}) - \hat{f}(m_{jl}) - \hat{g}(t_{jl})\big\}_{j,l}\Big] \text{ (centered)}\end{aligned}$$

after initializing $\hat{g} = \hat{h} = 0$. See Chapter II about the formation of smoothing matrices $\mathbf{S}_f, \mathbf{S}_g, \mathbf{S}_h$, for which the corresponding smoothing parameters can be selected by the method of generalized cross-validation. However, the linear parts of f, g, h are not identifiable, by Lemma 3.1. To get around the identification problem, we may take the ad hoc approach to remove the trend of $h(v)$. The trend-removal can be achieved by appending to every iteration of (3.7):

$$\hat{h}(v) \leftarrow \hat{h}(v) - v\Big(\sum_{j=1}^{J} v_j^2\Big)^{-1} \sum_{j=1}^{J} v_j \hat{h}(v_j).$$

We provide more details on (3.6) to shed light on its further development in the next section. By Proposition 2.1 in Chapter II, each of f, g, h can be expressed by the basis expansion of the form (2.8). Taking into account the identifiability of both the intercept and the trend, we may express f, g, h by

$$\begin{aligned} f(m) &= \mu_0 + \mu_1 m + \sum\nolimits_{i=1}^{I} \alpha_i \mathbb{K}(m, m_i) \\ g(t) &= \mu_2 t + \sum\nolimits_{l=1}^{L} \beta_l \mathbb{K}(t, t_l) \\ h(v) &= \sum\nolimits_{j=1}^{J} \gamma_j \mathbb{K}(v, v_j). \end{aligned} \tag{3.8}$$

Denote $\boldsymbol{\mu} = (\mu_0, \mu_1, \mu_2)^T$, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_I)^T$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_L)^T$, and $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_J)^T$. We may equivalently write (3.6) as the regularized least squares problem

$$\min\left\{\frac{1}{n}\big\|\mathbf{y} - \mathbf{B}\boldsymbol{\mu} - \widetilde{\boldsymbol{\Sigma}}_f\boldsymbol{\alpha} - \widetilde{\boldsymbol{\Sigma}}_g\boldsymbol{\beta} - \widetilde{\boldsymbol{\Sigma}}_h\boldsymbol{\gamma}\big\|^2 + \lambda_f\big\|\boldsymbol{\alpha}\big\|^2_{\boldsymbol{\Sigma}_f} + \lambda_g\big\|\boldsymbol{\beta}\big\|^2_{\boldsymbol{\Sigma}_g} + \lambda_h\big\|\boldsymbol{\gamma}\big\|^2_{\boldsymbol{\Sigma}_h}\right\}, \tag{3.9}$$

where $\mathbf{y}$ is the vector of $n = \sum_{j=1}^{J} L_j$ observations $\{\eta(y_{jl})\}$, $\mathbf{B}, \widetilde{\boldsymbol{\Sigma}}_f, \boldsymbol{\Sigma}_f$ are formed by

$$\mathbf{B} = \big[\{(1, m_{jl}, t_{jl})\}_{j,l}\big]_{n\times 3}, \quad \widetilde{\boldsymbol{\Sigma}}_f = \big[\mathbb{K}(m_{jl}, m_i)\big]_{n\times I}, \quad \boldsymbol{\Sigma}_f = \big[\mathbb{K}(m_i, m_i)\big]_{I\times I}$$

and $\widetilde{\boldsymbol{\Sigma}}_g, \boldsymbol{\Sigma}_g, \widetilde{\boldsymbol{\Sigma}}_h, \boldsymbol{\Sigma}_h$ are formed similarly.

The GAMEED Approach.

The GAMEED refers to an industry approach, namely the generalized additive maturation and exogenous effects decomposition, adopted by Enterprise Quantitative Risk Management, Bank of America, N.A. (Sudjianto, et al., 2006). It has also been documented as part of the patented work “Risk and Reward Assessment Mechanism” (USPTO: 20090063361). Essentially, it has two practical steps:

1. Estimate the maturation curve $\hat{f}(m)$ and the exogenous effect $\hat{g}(t)$ from

$$\eta(y(m,t;v)) = f(m;\boldsymbol{\alpha}) + g(t;\boldsymbol{\beta}) + \varepsilon, \quad \mathbb{E}g = 0$$

   where $f(m;\boldsymbol{\alpha})$ can either follow a spline model or take a parametric form supported by prior knowledge, and $g(t;\boldsymbol{\beta})$ can be specified like in (3.4) upon necessary time binding while retaining the flexibility in modeling the macroeconomic influence. If necessary, one may further model $\hat{g}(t;\boldsymbol{\beta})$ by a time-series model in order to capture the autocorrelation and possible jumps.

2. Estimate the vintage-specific sensitivities to $\hat{f}(m)$ and $\hat{g}(t)$ by performing the regression fitting

$$\eta(y(m_{jl},t_{jl};v_j)) = \gamma_j + \gamma_j^{(f)}\hat{f}(m_{jl}) + \gamma_j^{(g)}\hat{g}(t_{jl}) + \varepsilon$$

   for each vintage j (upon necessary bucketing).

The first step of GAMEED is a reduced type of MEV decomposition without suffering theidentification problem. The estimates of $f(m)$ and $g(t)$ can be viewed as the common factors to all vintages. In the second step, the vintage effects are measured through not only the main (intercept) effects, but also the interactions with $\hat{f}(m)$ and $\hat{g}(t)$. Clearly, when the interaction sensitivities $\gamma_j^{(f)} = \gamma_j^{(g)} = 1$, the industry GAMEED approach becomes a sequential type of MEV decomposition.

The DtD Technique.

The DtD refers to the dual-time dynamics, adopted by Strategic Analytics Inc., for consumer behavior modeling and delinquency forecasting. It was brought to the public domain by Breeden (2007). Using our notations, the DtD model takes the form

$$\eta(y(m_{jl}, t_{jl}; v_j)) = \gamma_j^{(f)} f(m_{jl}) + \gamma_j^{(g)} g(t_{jl}) + \varepsilon \tag{3.10}$$

and $\eta^{-1}(\cdot)$ corresponds to the so-called superposition function in Breeden (2007). Clearly, the DtD model involves the interactions between the vintage effects and $(f(m), g(t))$. Unlike the sequential GAMEED above, Breeden (2007) suggested to iteratively fit $f(m), g(t)$ and $\gamma_j^{(f)}, \gamma_j^{(g)}$ in a simultaneous manner.

However, the identification problem could as well happen to the DtD technique, unless it imposes structural assumptions that could break up the trend dependency. To see this, one may use Taylor expansion for each nonparametric function in (3.10) then apply Lemma 3.1 to the linear terms. Unfortunately, such identification problem was not addressed by Breeden (2007), and the stability of the iterative estimation algorithm might be an issue.

3.3 Gaussian Process Models

In MEV decomposition framework (3.2), the maturation curve $f(m)$, the exogenous influence $g(t)$ and the vintage heterogeneity $h(v)$ are postulated to capture the endogenous, exogenous and origination effects on the dual-time responses. In this section, we propose a general approach to MEV decomposition modeling based on Gaussian processes and provide an efficient algorithm for model estimation.

3.3.1 Covariance Kernels

Gaussian processes are rather flexible in modeling a function $Z(x)$ through a mean function $\mu(\cdot)$ and a covariance kernel $\mathbb{K}(\cdot, \cdot)$,

$$\mathbb{E}Z(x) = \mu(x), \quad \mathbb{E}[Z(x) - \mu(x)][Z(x') - \mu(x')] = \sigma^2 \mathbb{K}(x, x') \tag{3.11}$$

where σ^2 is the variance scale. For ease of discussion, we will write $\mu(x)$ separately and only consider the zero-mean Gaussian process $Z(x)$. Then, any finite-dimensional collection of random variables $\mathbf{Z} = (Z(x_1), \ldots, Z(x_n))$ follows a multivariate normal distribution $\mathbf{Z} \sim N_n(\mathbf{0}, \sigma^2\mathbf{\Sigma})$ with $\mathbf{\Sigma} = [\mathbb{K}(x_i, x_j)]_{n\times n}$.

The covariance kernel $\mathbb{K}(\cdot,\cdot)$ plays a key role in constructing Gaussian processes. It is symmetric $\mathbb{K}(x,x') = \mathbb{K}(x',x)$ and positive definite: $\int_{x,x'} \mathbb{K}(x,x')f(x)f(x')dxdx' > 0$ for any non-zero $f \in \mathcal{L}_2$. The kernel is said to be stationary if $\mathbb{K}(x,x')$ depends only on $|x_i - x_i'|$, $i = 1,\ldots,d$ for $x \in \mathbb{R}^d$. It is further said to be isotropic or radial if $\mathbb{K}(x,x')$ depends only on the distance $\|x - x'\|$. For $x \in \mathbb{R}$, a stationary kernel is also isotropic. There is a parsimonious approach to using stationary kernels $\mathbb{K}(\cdot,\cdot)$ to model certain nonstationary Gaussian processes such that $\mathrm{Cov}[Z(x), Z(x')] = \sigma(x)\sigma(x')\mathbb{K}(x,x')$, where the variance scale $\sigma^2(x)$ requires a separate model of either parametric or non-parametric type. For example, Fan, Huang and Li (2007) used the local polynomial approach to smoothing $\sigma^2(x)$.

Several popular classes of covariance kernels are listed below, including the adaptive smoothing spline (AdaSS) kernels discussed in Chapter II.

$$\begin{aligned}
\textsf{Exponential family:}\quad & \mathbb{K}(x,x') = \exp\left\{-\left(\frac{|x-x'|}{\phi}\right)^{\kappa}\right\}, \quad 0 < \kappa \le 2,\ \phi > 0 \\
\textsf{Matérn family:}\quad & \mathbb{K}(x,x') = \frac{2^{1-\kappa}}{\Gamma(\kappa)}\left(\frac{|x-x'|}{\phi}\right)^{\kappa} K_\kappa\left(\frac{|x-x'|}{\phi}\right), \quad \kappa, \phi > 0 \qquad (3.12) \\
\textsf{AdaSS r.k. family:}\quad & \mathbb{K}(x,x') = \int_{\mathcal{X}} \phi^{-1}(u) G_\kappa(x,u) G_\kappa(x',u) du, \quad \phi(x) > 0.
\end{aligned}$$

More examples of covariance kernels can be referred to Stein (1999) and Rasmussen and Williams (2006).

Both the exponential and Matérn families consist of stationary kernels, for which ϕ is the scale parameter and κ can be viewed as the shape parameter. In the Matérn family, $K_\kappa(\cdot)$ denotes the modified Bessel function of order κ; see Abramowitz and Stegun (1972). Often used are the half-integer order $\kappa = 1/2, 3/2, 5/2$ and the corresponding Matérn covariance kernels take the following forms

$$
\begin{aligned}
\mathbb{K}(x,x') &= e^{-\frac{|x-x'|}{\phi}}, \quad \kappa = 1/2 \\
\mathbb{K}(x,x') &= e^{-\frac{|x-x'|}{\phi}}\Big(1+\frac{|x-x'|}{\phi}\Big), \quad \kappa = 3/2 \\
\mathbb{K}(x,x') &= e^{-\frac{|x-x'|}{\phi}}\Big(1+\frac{|x-x'|}{\phi}+\frac{|x-x'|^2}{3\phi^2}\Big), \quad \kappa = 5/2.
\end{aligned}
\tag{3.13}
$$

It is interesting that these half-integer Matérn kernels correspond to autocorrelated Gaussian processes of order $\kappa + 1/2$; see Stein (1999; p. 31). In particular, the $\kappa = 1/2$ Matérn kernel, which is equivalent to the exponential kernel with $\kappa = 1$, is the covariance function of the Ornstein-Uhlenbeck (OU) process.

The AdaSS family of reproducing kernels (r.k.'s) are based on the Green's function $G_\kappa(x,u) = (x-u)_+^{\kappa-1}/(\kappa-1)!$ with order κ and local adaptation function $\phi(x) > 0$. In Chapter 2.2.1, we have derived the explicit r.k. expressions with piecewise $\phi^{-1}(x)$. They are nonstationary covariance kernels with the corresponding Gaussian processes discussed in Chapter 2.3. In this chapter, we concentrate on a particular class of AdaSS kernels

$$
\mathbb{K}(x,x') = \sum_{k=1}^{\min\{k:\tau_k > x\wedge x'\}} \phi_k\Big(x\wedge x' - \tau_{k-1}\Big) \tag{3.14}
$$

with piecewise-constant adaptation $\phi^{-1}(x) = \phi_k$ for $x \in [\tau_k - 1, \tau_k)$ and knots $x_{\min}^- \equiv \tau_0 < \tau_1 < \cdots < \tau_K \equiv x_{\max}^+$. By (2.29), the Gaussian process with kernel (3.14) generalizes the standard Wiener process with nonstationary volatility, which is useful for modeling the heterogeneous behavior of an MEV component, e.g. $g(t)$.

3.3.2 Kriging, Spline and Kernel Methods

Using Gaussian processes to model the component functions in (3.2) would give us a class of Gaussian process MEV models. Among others, we consider

$$
\eta(y(m,t;v)) = \mu(m,t;v) + Z_f(m) + Z_g(t) + Z_h(v) + \varepsilon \tag{3.15}
$$

where $\varepsilon \sim N(0, \sigma^2)$ is the random error, $Z_f(m), Z_g(t), Z_h(v)$ are three zero-mean Gaussian processes with covariance kernels $\sigma_f^2\mathbb{K}_f, \sigma_g^2\mathbb{K}_g, \sigma_h^2\mathbb{K}_h$, respectively. To bypass the identification problem in Lemma 3.1, the overall mean function is postulated to be $\mu(m, t; v) \in \mathcal{H}_0$ on the hyperplane of the vintage diagram, where

$$\mathcal{H}_0 \subseteq \text{span}\{1, m, t\} \cap \text{null}\{Z_f(m), Z_g(t), Z_h(v)\}, \tag{3.16}$$

among other choices. For example, choosing cubic spline kernels for all f, g, h allows for $\mathcal{H}_0 = \text{span}\{0, m, t\}$, then $\mu(m, t; v) = \mu_0 + \mu_1 m + \mu_2 t$. In other situations, the functional space $\mathcal{H}_0$ might shrink to be span$\{1\}$ in order to satisfy (3.16), then $\mu(m, t; v) = \mu_0$. In general, let us denote $\mu(m, t; v) = \boldsymbol{\mu}^T\mathbf{b}(m, t; v)$, for $\boldsymbol{\mu} \in \mathbb{R}^d$.

For simplicity, let us assume that $Z_f(m), Z_g(t), Z_h(v)$ are mutually independent for $\mathbf{x} \equiv (m, t, v) \in \mathbb{R}^3$. Consider the sum of univariate Gaussian processes

$$Z(\mathbf{x}) = Z_f(m) + Z_g(t) + Z_h(v), \tag{3.17}$$

with mean zero and covariance $\text{Cov}[Z(\mathbf{x}), Z(\mathbf{x}')] = \sigma^2\mathbb{K}(\mathbf{x}, \mathbf{x}')$. By the independence assumption,

$$\mathbb{K}(\mathbf{x}, \mathbf{x}') = \lambda_f^{-1}\mathbb{K}_f(m, m') + \lambda_g^{-1}\mathbb{K}_g(t, t') + \lambda_h^{-1}\mathbb{K}_h(v, v'). \tag{3.18}$$

with $\lambda_f = \sigma^2/\sigma_f^2$, $\lambda_g = \sigma^2/\sigma_g^2$ and $\lambda_h = \sigma^2/\sigma_h^2$. Then, given the vintage data (3.1) with $n = \sum_{j=1}^J L_j$ observations, we may write (3.15) in the vector-matrix form

$$\mathbf{y} = \mathbf{B}\boldsymbol{\mu} + \widetilde{\mathbf{z}}, \quad \text{Cov}[\widetilde{\mathbf{z}}] = \sigma^2\big[\boldsymbol{\Sigma} + \mathbf{I}\big] \tag{3.19}$$

where $\mathbf{y} = \big[\eta(y_{jl})\big]_{n\times 1}$, $\mathbf{B} = \big[\mathbf{b}_{jl}^T\big]_{n\times d}$, $\widetilde{\mathbf{z}} = \big[Z(\mathbf{x}_{jl}) + \varepsilon\big]_{n\times 1}$ and $\boldsymbol{\Sigma} = \big[\mathbb{K}(\mathbf{x}_{jl}, \mathbf{x}_{j'l'})\big]_{n\times n}$. By either GLS (generalized least squares) or MLE (maximum likelihood estimation),

$\boldsymbol{\mu}$ is estimated to be

$$\widehat{\boldsymbol{\mu}} = \left(\mathbf{B}^T \left[\boldsymbol{\Sigma} + \mathbf{I}\right]^{-1}\mathbf{B}\right)^{-1}\mathbf{B}^T \left[\boldsymbol{\Sigma} + \mathbf{I}\right]^{-1}\mathbf{y}. \tag{3.20}$$

where the invertibility of $\boldsymbol{\Sigma} + \mathbf{I}$ is guaranteed by large values of $\lambda_f, \lambda_g, \lambda_h$ as in the context of ridge regression.

In the Gaussian process models there are additional unknown parameters for defining the covariance kernel (3.18), namely $(\boldsymbol{\lambda}, \boldsymbol{\theta}) = (\lambda_f, \lambda_g, \lambda_h, \theta_f, \theta_g, \theta_h)$, as well as the unknown variance parameter σ^2 (a.k.a. the nugget effect). The $\lambda_f, \lambda_g, \lambda_h$, as the ratios of σ^2 to the variance scales of Z_f, Z_g, Z_h, are usually referred to as the smoothing parameters. The $\theta_f, \theta_g, \theta_h$ denote the structural parameters, e.g scale parameter ϕ and shape parameter κ in exponential or Matérn family of kernels, or local adaptation $\phi(\cdot)$ in AdaSS kernel. By (3.19) and after dropping the constant term, the log-likelihood function is given by

$$\ell(\boldsymbol{\mu}, \boldsymbol{\lambda}, \boldsymbol{\theta}, \sigma^2) = -\frac{n}{2}\log(\sigma^2) - \frac{1}{2}\log\left|\boldsymbol{\Sigma} + \mathbf{I}\right| - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{B}\boldsymbol{\mu})^T(\boldsymbol{\Sigma} + \mathbf{I})^{-1}(\mathbf{y} - \mathbf{B}\boldsymbol{\mu})$$

in which $\boldsymbol{\Sigma}$ depends on $(\boldsymbol{\lambda}, \boldsymbol{\theta})$. Whenever $(\boldsymbol{\lambda}, \boldsymbol{\theta})$ are fixed, the $\boldsymbol{\mu}$ estimate is given by (3.20), then one may estimate the variance parameter to be

$$\hat{\sigma}^2 = \frac{1}{n-p}(\mathbf{y} - \mathbf{B}\widehat{\boldsymbol{\mu}})\boldsymbol{\Sigma}^{-1}(\mathbf{y} - \mathbf{B}\widehat{\boldsymbol{\mu}}), \tag{3.21}$$

where $p = \text{rank}(\mathbf{B})$ is plugged in order to make $\hat{\sigma}^2$ an unbiased estimator.

Usually, numerical procedures like Newton-Raphson algorithm are needed to iteratively estimate $(\boldsymbol{\lambda}, \boldsymbol{\theta})$ that admit no closed-form solutions; see Fang, Li and Sudjianto (2006, §5). Such a standard maximum likelihood estimation method is computationally intensive. In the next subsection, we will discuss how to iteratively estimate the parameters by an efficient backfitting algorithm.

MEV Kriging means the prediction based on the Gaussian process MEV model (3.15), where the naming of 'Kriging' follows the geostatistical convention; see Cressie (1993) in spatial statistics. See also Fang, Li and Sudjianto (2006) for the use of Kriging in the context of computer experiments. Kriging can be equally treated by either multivariate normal distribution theory or Bayesian approach, since the conditional expectation of multivariate normal distribution equals the Bayesian posterior mean. The former approach is taken here for making prediction from the *noisy* observations, for which the Kriging behaves as a smoother rather than an interpolator.

Let us denote the prediction of $\eta(\hat{y}(\mathbf{x}))$ at $\mathbf{x} = (m, t, v)$ by $\zeta(\mathbf{x}) = \widehat{\boldsymbol{\mu}}^T \mathbf{b}(\mathbf{x}) + \mathbb{E}\Big[Z_f(m) + Z_g(t) + Z_h(v) + \varepsilon \Big| \widetilde{\mathbf{z}}\Big]$, where $\mathbb{E}[Z|\widetilde{\mathbf{z}}]$ denotes the conditional expectation of Z given $\widetilde{\mathbf{z}} = \mathbf{y} - \mathbf{B}\widehat{\boldsymbol{\mu}}$. By the property of multivariate normal distribution,

$$\hat{\zeta}(\mathbf{x}) = \widehat{\boldsymbol{\mu}}^T \mathbf{b}(\mathbf{x}) + \boldsymbol{\xi}_n^T(\mathbf{x}) \big[\boldsymbol{\Sigma} + \mathbf{I}\big]^{-1} \Big(\mathbf{y} - \mathbf{B}\widehat{\boldsymbol{\mu}}\Big) \tag{3.22}$$

where $\boldsymbol{\xi}_n(\mathbf{x})$ denotes the vector of $\big[\mathbb{K}(\mathbf{x}, \mathbf{x}_{jl})\big]_{n\times 1}$ evaluated between $\mathbf{x}$ and each sampling point.

The Kriging (3.22) is known to be a best linear unbiased predictor. It has also a smoothing spline reformulation that covers the additive cubic smoothing splines model (3.6) as a special case. Formally, we have the following proposition.

Proposition 3.2 (MEV Kriging as Spline Estimator)**.** *The MEV Kriging (3.22) can be formulated as a smoothing spline estimator*

$$\arg\min_{\zeta \in \mathcal{H}} \left\{ \sum_{j=1}^{J} \sum_{l=1}^{L_j} \Big[\eta(y_{jl}) - \zeta(\mathbf{x}_{jl})\Big]^2 + \|\zeta\|^2_{\mathcal{H}_1} \right\} \tag{3.23}$$

where $\mathcal{H} = \mathcal{H}_0 \oplus \mathcal{H}_1$ *with* $\mathcal{H}_0 \subseteq \mathrm{span}\{1, m, t\}$ *and* $\mathcal{H}_1$ *is the reproducing kernel Hilbert space induced by (3.18) such that* $\mathcal{H}_0 \perp \mathcal{H}_1$.

The spline solution to (3.23) can be written as a finite-dimensional expression

$$\zeta(\mathbf{x}) = \boldsymbol{\mu}^T \mathbf{b}(\mathbf{x}) + \sum_{j=1}^{J} \sum_{l=1}^{L_j} c_{jl} \mathbb{K}(\mathbf{x}, \mathbf{x}_{jl}) \tag{3.24}$$

with coefficients determined by the regularized least squares,

$$\min \left\{ \left\| \mathbf{y} - \mathbf{B}\boldsymbol{\mu} - \boldsymbol{\Sigma}\mathbf{c} \right\|^2 + \left\| \mathbf{c} \right\|_{\boldsymbol{\Sigma}} \right\}. \tag{3.25}$$

Furthermore, we have the following simplification.

Proposition 3.3 (Additive Separability)**.** *The MEV Kriging (3.22) can be expressed additively as*

$$\zeta(\mathbf{x}) = \boldsymbol{\mu}^T \mathbf{b}(\mathbf{x}) + \sum_{i=1}^{I} \alpha_i \mathbb{K}_f(m, m_i) + \sum_{l=1}^{L} \beta_l \mathbb{K}_g(t, t_l) + \sum_{j=1}^{J} \gamma_j \mathbb{K}_g(v, v_j) \tag{3.26}$$

based on I distinct m-values, L distinct t-values and J distinct v-values in (3.1). The coefficients can be estimated by

$$\min \left\{ \left\| \mathbf{y} - \mathbf{B}\boldsymbol{\mu} - \widetilde{\boldsymbol{\Sigma}}_f \boldsymbol{\alpha} - \widetilde{\boldsymbol{\Sigma}}_g \boldsymbol{\beta} - \widetilde{\boldsymbol{\Sigma}}_h \boldsymbol{\gamma} \right\|^2 + \lambda_f \left\| \boldsymbol{\alpha} \right\|^2_{\boldsymbol{\Sigma}_f} + \lambda_g \left\| \boldsymbol{\beta} \right\|^2_{\boldsymbol{\Sigma}_g} + \lambda_h \left\| \boldsymbol{\gamma} \right\|^2_{\boldsymbol{\Sigma}_h} \right\}, \tag{3.27}$$

where the matrix notations are the same as in (3.9) up to mild modifications.

Proposition 3.3 converts the MEV Kriging (3.22) to a kernelized additive model of the form (3.26). It connects the *kernel methods* in machine learning that advocates the use of covariance kernels for mapping the original inputs to some high-dimensional feature space. Based on our previous discussions, such kernelized feature space corresponds to the family of Gaussian processes, or the RKHS; see e.g. Rasmussen and Williams (2006) for more details.

Proposition 3.3 breaks down the high-dimensional parameter estimation and it makes possible the development of a backfitting algorithm to deal with three univariate Gaussian processes separately. Such backfitting procedure provides a more efficient procedure than the standard method of maximum likelihood estimation. Besides, it can easily deal with the collinearity problem among $Z_f(m), Z_g(t), Z_h(v)$ caused by the hyperplane dependency of vintage diagram.

3.3.3 MEV Backfitting Algorithm

For given $(\boldsymbol{\lambda}, \boldsymbol{\theta})$, the covariance kernels in (3.18) are defined explicitly, so one may find the optimal solution to (3.27) by setting partial derivatives to zero and solving the linear system of algebraic equations. There are however complications for determining $(\boldsymbol{\lambda}, \boldsymbol{\theta})$. In this section we propose a backfitting algorithm to minimize the regularized least squares criterion (3.27), together with a generalized cross-validation (GCV) procedure for selecting the smoothing parameters $\lambda_f, \lambda_g, \lambda_h$, as well as the structural parameters $\theta_f, \theta_g, \theta_h$ used for defining the covariance kernels. Here, the GCV criterion follows our discussion in Chapter 2.2.2; see also Golub, Heath and Wahba (1979) in the context of ridge regression.

Denote the complete set of parameters by $\Xi = \{\boldsymbol{\mu};\ \boldsymbol{\alpha}, \lambda_f, \theta_f;\ \boldsymbol{\beta}, \lambda_g, \theta_g;\ \boldsymbol{\gamma}, \lambda_h, \theta_h;\ \sigma^2\}$ where λ's are smoothing parameter and θ's are structural parameters for defining the covariance kernels. Given the vintage data (3.1) with notations $\mathbf{y}, \mathbf{B}$ in (3.6):

Step 1: Set the initial value of $\boldsymbol{\mu}$ to be the ordinary least squares (OLS) estimate $(\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T\mathbf{y}$; Set the initial values of $\boldsymbol{\beta}, \boldsymbol{\gamma}$ to be zero.

Step 2a: Given Ξ with $(\boldsymbol{\alpha}, \lambda_f, \theta_f)$ unknown, get the pseudo data $\widetilde{\mathbf{y}} = \mathbf{y} - \mathbf{B}\boldsymbol{\mu} - \widetilde{\boldsymbol{\Sigma}}_g\boldsymbol{\beta} - \widetilde{\boldsymbol{\Sigma}}_h\boldsymbol{\gamma}$. For each input (λ_f, θ_f), estimate $\boldsymbol{\alpha}$ by ridge regression

$$\widehat{\boldsymbol{\alpha}} = \left(\widetilde{\boldsymbol{\Sigma}}_f^T\widetilde{\boldsymbol{\Sigma}}_f + \lambda_f\boldsymbol{\Sigma}_f\right)^{-1}\widetilde{\boldsymbol{\Sigma}}_f^T\widetilde{\mathbf{y}} \tag{3.28}$$

where $\widetilde{\boldsymbol{\Sigma}}_f = \big[\mathbb{K}_f(m_{jl}, m_i; \theta_f)\big]_{n\times I}$ and $\boldsymbol{\Sigma}_f = \big[\mathbb{K}_f(m_i, m_j; \theta_f)\big]_{I\times I}$. Determine the best choice of (λ_f, θ_f) by minimizing

$$\mathrm{GCV}(\lambda_f, \theta_f) = \frac{1}{n}\left\|(\mathbf{I} - \widetilde{\mathbf{S}}(\lambda_f, \theta_f))\widetilde{\mathbf{y}}\right\|^2 \Big/ \left[1 - \frac{1}{n}\mathsf{Trace}\big(\widetilde{\mathbf{S}}(\lambda_f, \theta_f)\big)\right]^2, \qquad (3.29)$$

where $\widetilde{\mathbf{S}}(\lambda_f, \theta_f) = \widetilde{\mathbf{\Sigma}}_f\left(\widetilde{\mathbf{\Sigma}}_f^T\widetilde{\mathbf{\Sigma}}_f + \lambda_f\mathbf{\Sigma}_f\right)^{-1}\widetilde{\mathbf{\Sigma}}_f^T$.

Step 2b: Run Step 2a with the unknown parameters replaced by $(\boldsymbol{\beta}, \lambda_g, \theta_g)$ and $\widetilde{\mathbf{y}} = \mathbf{y} - \mathbf{B}\boldsymbol{\mu} - \widetilde{\mathbf{\Sigma}}_f\boldsymbol{\alpha} - \widetilde{\mathbf{\Sigma}}_h\boldsymbol{\gamma}$. Form $\widetilde{\mathbf{\Sigma}}_g, \mathbf{\Sigma}_g$ and $\widetilde{\mathbf{S}}(\lambda_g, \theta_g)$. Select (λ_g, θ_g) by GCV, then estimate $\boldsymbol{\beta}$ by ridge regression.

Step 2c: Run Step 2a with the unknown parameters replaced by $(\boldsymbol{\gamma}, \lambda_h, \theta_h)$ and $\widetilde{\mathbf{y}} = \mathbf{y} - \mathbf{B}\boldsymbol{\mu} - \widetilde{\mathbf{\Sigma}}_f\boldsymbol{\alpha} - \widetilde{\mathbf{\Sigma}}_g\boldsymbol{\beta}$. (For the *ad hoc* approach, remove the trend of $\widetilde{\mathbf{y}}$ in the vintage direction.) Form $\widetilde{\mathbf{\Sigma}}_h, \mathbf{\Sigma}_h$ and $\widetilde{\mathbf{S}}(\lambda_h, \theta_h)$. Select (λ_h, θ_h) by GCV. Estimate $\boldsymbol{\gamma}$ by ridge regression.

Step 2d: re-estimate the $\boldsymbol{\mu}$ by GLS (3.20) for given $(\lambda_f, \lambda_g, \lambda_h, \theta_f, \theta_g, \theta_h)$.

Step 3: Repeat steps 2a–2d until convergence, say, when the estimates of $\boldsymbol{\mu}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}$ change less than a pre-specified tolerance. Obtain $\hat{\sigma}^2$ by (3.21).

For future reference, the above iterative procedures are named as the MEV Backfitting algorithm. It is efficient and can automatically take into account selection of both smoothing and structural parameters. After obtaining the parameter estimates, we may reconstruct the vintage data by $\widehat{\mathbf{y}} = \mathbf{B}\widehat{\boldsymbol{\mu}} + \widetilde{\mathbf{\Sigma}}_f\widehat{\boldsymbol{\alpha}} + \widetilde{\mathbf{\Sigma}}_g\widehat{\boldsymbol{\beta}} + \widetilde{\mathbf{\Sigma}}_h\widehat{\boldsymbol{\gamma}}$. The following are several remarks for practical implementation:

1. In order to capture a diversity of important features, a combination of different covariance kernels can be included in the same MEV model. The exponential, Matérn and AdaSS r.k. families listed in (3.12), (3.13), (3.14) are of our primary interest here, whereas other types of kernels can be readily employed. When the underlying function is heterogeneously smooth, e.g. involving jumps, one may choose the AdaSS kernel (3.14).

2. For the ridge regression (3.28), the matrix $\widetilde{\mathbf{\Sigma}}_f$ contains row-wise replicates and can be reduced by weighted least squares, in which case the response vector $\widetilde{\mathbf{y}}$ is reduced to the pointwise averages; see (2.4). The evaluation of the GCV criterion (3.29) can be modified accordingly; see (2.22).

3. When the mean function $\mu(m, t; v)$ is specified to be the intercept effect, i.e. $\mu(m, t; v) = \mu_0$, one may replace μ_0 by the mean of $\mathbf{y}$ in step 1 and step 2d. In this case, we may write the MEV Kriging as $\zeta(\mathbf{x}) = \hat{f}(m) + \hat{g}(t) + \hat{h}(v)$ with

$$\hat{f}(m) = \hat{\mu}_0 + \sum_{i=1}^{I} \hat{\alpha}_i \mathbb{K}_f(m, m_i), \quad \hat{g}(t) = \sum_{l=1}^{L} \hat{\beta}_l \mathbb{K}_g(t, t_l), \quad \hat{h}(v) = \sum_{j=1}^{J} \hat{\gamma}_j \mathbb{K}_h(v, v_j).$$

4. In principal, both $\hat{g}(t)$ and $\hat{h}(v)$ in step 2b and step 2c have mean zero, but in practice, they may need the centering adjustment due to machine rounding.

3.4 Semiparametric Regression

Having discussed the MEV decomposition models based on Gaussian processes, we consider in this section the inclusion of covariates in the following two scenarios:

1. $\mathbf{x}(m_{jl}, t_{jl}; v_j) \in \mathbb{R}^p$ (or $\mathbf{x}_{jl}$ in short) for a fixed segment, where $\mathbf{x}_{jl}$ denotes the covariates of j-th vintage at time t_{jl} and age $m_{jl} = t_{jl} - v_j$;

2. $\mathbf{x}(m_{kjl}, t_{kjl}; v_{kj}) \in \mathbb{R}^p$ and $\mathbf{z}_k \in \mathbb{R}^q$ for multiple segments that share the same vintage diagram, where $\mathbf{z}_k$ represents the static segmentation variables.

Of our interest are the the semiparametric MEV regression models with both nonparametric $f(m), g(t), h(v)$ and parametric covariate effects.

After a brief discussion of the first situation with a single segment, we will spend more time in the second situation with multiple segments. For simplicity, the covariates $\mathbf{x}_{jl}$ or $\mathbf{x}_{kjl}$ are assumed to be deterministic, i.e. no error in variables.

3.4.1 Single Segment

Given a single segment with dual-time observations $\{y(m_{jl}, t_{jl}; v_j), \mathbf{x}(m_{jl}, t_{jl}; v_j)\}$ for $j = 1, \ldots, J$ and $l = 1, \ldots, L_j$, we consider the partial linear model that adds the linear covariate effects to the MEV decomposition (3.2),

$$\begin{aligned} \eta(y_{jl}) &= f(m_{jl}) + g(t_{jl}) + h(v_j) + \boldsymbol{\pi}^T \mathbf{x}_{jl} + \varepsilon \qquad (3.30) \\ &= \boldsymbol{\mu}^T \mathbf{b}_{jl} + \boldsymbol{\pi}^T \mathbf{x}_{jl} + Z_f(m_{jl}) + Z_g(t_{jl}) + Z_h(v_j) + \varepsilon \end{aligned}$$

where $\mathbf{x}_{jl} \equiv \mathbf{x}(m_{jl}, t_{jl}; v_j)$, $(\boldsymbol{\mu}, \boldsymbol{\pi})$ are the parameters for the mean function, and f, g, h are nonparametric and modeled by Gaussian processes.

The semiparametric regression model (3.30) corresponds to the linear mixed-effect modeling, and the covariate effects $\boldsymbol{\pi}$ can be simply estimated in the same way we estimated $\boldsymbol{\mu}$ in (3.20). That is, conditional on $\mathbf{\Sigma}$,

$$(\widehat{\boldsymbol{\mu}}; \widehat{\boldsymbol{\pi}}) = \left(\widetilde{\mathbf{B}}^T \left[\mathbf{\Sigma} + \mathbf{I}\right]^{-1} \widetilde{\mathbf{B}}\right)^{-1} \widetilde{\mathbf{B}}^T \left[\mathbf{\Sigma} + \mathbf{I}\right]^{-1} \mathbf{y}. \qquad (3.31)$$

where $\widetilde{\mathbf{B}} = [\mathbf{B} \vdots \mathbf{X}]$ with $\mathbf{X} = [\mathbf{x}_{jl}]_{n \times p}$. This GLS estimator can be incorporated with the established MEV Backfitting algorithm discussed in Section 3.3.3. Only step 1 and step 2d need modified to be:

Step 1: Set $(\boldsymbol{\mu}, \boldsymbol{\pi})$ initially to be the OLS estimate, i.e. (3.31) with zero $\mathbf{\Sigma}$.

Step 2d: re-estimate $(\boldsymbol{\mu}; \boldsymbol{\pi})$ by (3.31) for given $(\lambda_f, \lambda_g, \lambda_h, \theta_f, \theta_g, \theta_h)$.

Then, the whole set of parameters in (3.30) can be iteratively estimated by the modified backfitting algorithm. There is a word of caution about multicollinearity, since the orthogonality condition in Proposition 3.2 may be not satisfied between the covariate space span$\{\mathbf{x}\}$ and the Gaussian processes. In this case, instead of starting from the fully nonparametric approach, we may assume the f, g, h components in (3.30) can be approximated by the kernel basis expansion through (3.26), then run backfitting. Fortunately, the smoothing parameters $\lambda_f, \lambda_g, \lambda_h$ may guard against possible multicollinearity and give the shrinkage estimate of the covariate effects $\boldsymbol{\pi}$. As a trade-off, part of Gaussian process components might be over-smoothed.

3.4.2 Multiple Segments

Suppose we are given K segments of data that share the same vintage diagram,

$$\big\{y(m_{kjl}, t_{kjl}; v_{kj}), \mathbf{x}(m_{kjl}, t_{kjl}; v_{kj}), \mathbf{z}_k\big\}, \tag{3.32}$$

for $k = 1, \ldots, K$, $j = 1, \ldots, J$ and $l = 1, \ldots, L_j$, where $\mathbf{x}_{kjl} \in \mathbb{R}^p$ and $\mathbf{z} \in \mathbb{R}^q$. Denote by N the total number of observations across K segments. There are various approaches to modeling cross-segment responses, e.g.

$$\begin{array}{ll} \text{a.} & \eta(y_{kjl}) = f(m_{kjl}) + g(t_{kjl}) + h(v_{kj}) + \boldsymbol{\pi}^T\mathbf{x}_{kjl} + \boldsymbol{\omega}^T\mathbf{z}_k + \varepsilon \\ \text{b.} & \eta(y_{kjl}) = f_k(m_{kjl}) + g_k(t_{kjl}) + h_k(v_{kj}) + \boldsymbol{\pi}_k^T\mathbf{x}_{kjl} + \varepsilon \\ \text{c.} & \eta(y_{kjl}) = u_f^{(k)}f(m_{kjl}) + u_g^{(k)}g(t_{kjl}) + u_h^{(k)}h(v_{kj}) + \boldsymbol{\pi}_k^T\mathbf{x}_{kjl} + W(\mathbf{z}_k) + \varepsilon. \end{array} \tag{3.33}$$

Here, the first approach assumes that the segments all perform the same MEV components and covariate effect $\boldsymbol{\pi}$, and differ only through $\boldsymbol{\omega}^T\mathbf{z}_k$. The second approach takes the other extreme assumption such that each segment performs marginally differently in terms of f, g, h and covariate effects, therefore equivalent to K separate single-segment models. Clearly, both of these approaches can be treated by the single-segment modeling technique discussed in the last section.

Of our interest of study is the third approach that allows for multiple segments to have different sensitivities $u_f^{(k)}, u_g^{(k)}, u_h^{(k)} \in \mathbb{R}$ to the same underlying $f(m), g(t), h(v)$. In another word, the functions $f(m), g(t), h(v)$ represent the common factors across segments, while the coefficients $u_f^{(k)}, u_g^{(k)}, u_h^{(k)}$ represent the idiosyncratic multipliers. Besides, it allows for the segment-specific add-on effects $W(\mathbf{z}_k)$ and segment-specific covariate effect $\boldsymbol{\pi}_k$. Thus, the model (c) is postulated as a balance between the over-stringent model (a) and the over-flexible model (b).

Different strategies can be applied to model the segment-specific effects $W(\mathbf{z}_k)$. Among others, we consider (c1) the linear modeling $W(\mathbf{z}_k) = \boldsymbol{\omega}^T\mathbf{z}_k$ with parameter $\boldsymbol{\omega} \in \mathbb{R}^q$ and (c2) the Gaussian process modeling

$$\mathbb{E}W(\mathbf{z}_k) = 0, \quad \mathbb{E}W(\mathbf{z}_k)W(\mathbf{z}_{k'}) = \lambda_W \mathbb{K}_W(\mathbf{z}_k, \mathbf{z}_{k'}) \tag{3.34}$$

for the smoothing parameter $\lambda_W > 0$ and some pre-specificed family of covariance kernels $\mathbb{K}_W$. For example, we may use the squared exponential covariance kernel in (3.12). The estimation of these two types of segment effects are detailed as follows.

Linear Segment Effects.

Given the multi-segment vintage data (3.32), consider

$$\eta(y_{kjl}) = u_f^{(k)} f(m_{kjl}) + u_g^{(k)} g(t_{kjl}) + u_h^{(k)} h(v_{kj}) + \boldsymbol{\pi}_k^T \mathbf{x}_{kjl} + \boldsymbol{\omega}^T\mathbf{z}_k + \varepsilon \tag{3.35}$$

with basis approximation

$$f(m) = \mu_0 + \sum_{i=1}^{I} \alpha_i \mathbb{K}_f(m, m_i),\ \ g(t) = \sum_{l=1}^{L} \beta_l \mathbb{K}_g(t, t_l),\ \ h(v) = \sum_{j=1}^{J} \gamma_j \mathbb{K}_g(v, v_j)$$

through the exponential or Matérn kernels. (The smoothing spline kernels can be also used after appropriate modification of the mean function.) Using the matrix notations, the model takes the form

$$\Big\langle \underline{\mathbf{u}}_f,\ \underline{\mathbf{B}}\boldsymbol{\mu} + \underline{\widetilde{\boldsymbol{\Sigma}}}_f \boldsymbol{\alpha} \Big\rangle + \Big\langle \underline{\mathbf{u}}_g,\ \underline{\widetilde{\boldsymbol{\Sigma}}}_g \boldsymbol{\beta} \Big\rangle + \Big\langle \underline{\mathbf{u}}_h,\ \underline{\widetilde{\boldsymbol{\Sigma}}}_h \boldsymbol{\gamma} \Big\rangle + \underline{\mathbf{X}}\boldsymbol{\pi} + \underline{\mathbf{Z}}\boldsymbol{\omega}$$

where the underscored vectors and matrices are formed based on multi-segment observations, $(\underline{\mathbf{u}}_f, \underline{\mathbf{u}}_g, \underline{\mathbf{u}}_h)$ are the extended vector of segment-wise multipliers, $\underline{\mathbf{X}}$ has

$K \times p$ columns corresponding to the vectorized coefficients $\boldsymbol{\pi} = (\boldsymbol{\pi}_1; \ldots; \boldsymbol{\pi}_K)$, and $\underline{\mathbf{Z}}$ is as usual. When both the kernel parameters $(\boldsymbol{\lambda}, \boldsymbol{\theta})$ for $\mathbb{K}_f, \mathbb{K}_g, \mathbb{K}_h$ and the (condensed K-vector) multipliers $\mathbf{u}_f, \mathbf{u}_g, \mathbf{u}_h$ are fixed, the regularized least squares can be used for parameter estimation, using the same penalty terms as in (3.27).

Iterative procedures can be used to estimate all the parameters simultaneously, including $\{\boldsymbol{\mu}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\lambda}, \boldsymbol{\theta}\}$ for the underlying f, g, h functions and $\mathbf{u}_f, \mathbf{u}_g, \mathbf{u}_h, \boldsymbol{\pi}, \boldsymbol{\omega}$ across segments. The following algorithm is constructed by utilizing the MEV Backfitting algorithm as an intermediate routine.

Step 1: Set the initial values of $(\boldsymbol{\pi}, \boldsymbol{\omega})$ to be the OLS estimates for fitting $\underline{\mathbf{y}} = \underline{\mathbf{X}}\boldsymbol{\pi} + \underline{\mathbf{Z}}\boldsymbol{\omega} + \underline{\boldsymbol{\varepsilon}}$ based on all N observations across K segments.

Step 2: Run the MEV Backfitting algorithm for the pseudo data $\underline{\mathbf{y}} - \underline{\mathbf{X}}\widehat{\boldsymbol{\pi}} - \underline{\mathbf{Z}}\widehat{\boldsymbol{\omega}}$ for obtaining the estimates of $\hat{f}(m), \hat{g}(t), \hat{h}(v)$

Step 3: Update $(\mathbf{u}_f, \mathbf{u}_g, \mathbf{u}_h)$ and $(\boldsymbol{\pi}, \boldsymbol{\omega})$ by OLS estimates for fitting

$$\underline{\mathbf{y}} = \underline{\mathbf{F}}\mathbf{u}_f + \underline{\mathbf{G}}\mathbf{u}_g + \underline{\mathbf{H}}\mathbf{u}_h + \underline{\mathbf{X}}\boldsymbol{\pi} + \underline{\mathbf{Z}}\boldsymbol{\omega} + \underline{\boldsymbol{\varepsilon}}$$

where $\underline{\mathbf{F}}, \underline{\mathbf{G}}, \underline{\mathbf{H}}$ are all $N \times K$ regression sub-matrices constructed based on $\hat{f}(m_{kjl})$, $\hat{g}(t_{kjl})$, $\hat{h}(v_{kj})$, respectively.

Step 4: Repeat step 2 and step 3 until convergence.

Note that in Step 2, the common factors $\hat{f}(m)$, $\hat{g}(t)$ and $\hat{h}(v)$ are estimated first by assuming $u_k = \mathbb{E}u_k = 1$ for each of f, g, h. In Step 3, the segment-wise multipliers u_k's are updated by regressing over $\hat{f}$, $\hat{g}$ and $\hat{h}$. In so doing, the concern of stability is one reason. Disregarding the segment variability in Step 2 is also reasonable, since $f(m), g(t), h(v)$ are by definition the common factors in the overall sense. Then, including the segment variability in Step 3 improves both the estimation of $(\boldsymbol{\pi}, \boldsymbol{\omega})$

and the estimation of f, g, h in the next iteration of Step 2. Such fitting procedure is practically interesting and computationally tractable.

Spatial Segment Heterogeneity.

We move to consider the segment effect modeling by a Gaussian process

$$\eta(y_{kjl}) = u_f^{(k)} f(m_{kjl}) + u_g^{(k)} g(t_{kjl}) + u_h^{(k)} h(v_{kj}) + W(\mathbf{z}_k) + \boldsymbol{\pi}_k^T \mathbf{x}_{kjl} + \varepsilon \qquad (3.36)$$

where f, g, h are approximated in the same way as in (3.35) and $W(\mathbf{z}_k)$ is assumed to follow (3.34). The orthogonality constraint between $W(\mathbf{z})$ and $\{f(m), g(t), h(v)\}$ can be justified from a tensor-product space point of view, but the orthogonality is not guaranteed between $W(\mathbf{z}_k)$ and the span of $\mathbf{x}_{kjl}$. Despite of the orthogonality constraint, let us directly approximate $W(\mathbf{z})$ by the basis expansion $W(\mathbf{z}) = \sum_{k=1}^{K} \omega_k \mathbb{K}_W(\mathbf{z}, \mathbf{z}_k)$ associated with the unknown scale parameter (denoted by θ_W). Then, we have the following two iterative stages of model estimation.

Stage 1: for given $\boldsymbol{\pi}_k$ and $u_k \equiv 1$, $k = 1, \ldots, K$, estimate the common-factor functions $f(m), g(t), h(v)$ and the spatial segment effect $W(\mathbf{z}_k)$ in

$$\begin{aligned} \eta(y_{kjl}) - \boldsymbol{\pi}_k^T \mathbf{x}_{kjl} &= \mu_0 + \sum_{i=1}^{I} \alpha_i \mathbb{K}_f(m_{kjl}, m_i) + \sum_{i=1}^{L} \beta_i \mathbb{K}_g(t_{kjl}, t_i) \\ &\quad + \sum_{i=1}^{J} \gamma_i \mathbb{K}_g(v_{kj}, v_i) + \sum_{i=1}^{K} \omega_i \mathbb{K}_W(\mathbf{z}_k, \mathbf{z}_i) + \varepsilon \end{aligned}$$

by the regularized least squares

$$\begin{aligned} \min \Big\{ & \big\| \widetilde{\mathbf{y}} - \mathbf{B}\boldsymbol{\mu} - \widetilde{\boldsymbol{\Sigma}}_f \boldsymbol{\alpha} - \widetilde{\boldsymbol{\Sigma}}_g \boldsymbol{\beta} - \widetilde{\boldsymbol{\Sigma}}_h \boldsymbol{\gamma} - \widetilde{\boldsymbol{\Sigma}}_W \boldsymbol{\omega} \big\|^2 \\ & + \lambda_f \big\| \boldsymbol{\alpha} \big\|^2_{\boldsymbol{\Sigma}_f} + \lambda_g \big\| \boldsymbol{\beta} \big\|^2_{\boldsymbol{\Sigma}_g} + \lambda_h \big\| \boldsymbol{\gamma} \big\|^2_{\boldsymbol{\Sigma}_h} + \lambda_W \big\| \boldsymbol{\omega} \big\|^2_{\boldsymbol{\Sigma}_W} \Big\}, \end{aligned} \qquad (3.37)$$

where $\widetilde{\mathbf{y}}$ is the pseudo response vector of $\eta(y_{kjl}) - \boldsymbol{\pi}_k^T \mathbf{x}_{kjl}$ for all (k, j, l). The regression sub-matrices $\mathbf{B}, \widetilde{\boldsymbol{\Sigma}}_f, \widetilde{\boldsymbol{\Sigma}}_g, \widetilde{\boldsymbol{\Sigma}}_h, \widetilde{\boldsymbol{\Sigma}}_W$ are constructed accordingly, all having N rows. The matrices used for regularization are of size $I \times I$, $L \times L$, $J \times J$ and $K \times K$, respectively. Both the smoothing parameters $\lambda_f, \lambda_g, \lambda_h, \lambda_W > 0$ and the structural parameters $\theta_f, \theta_g, \theta_h, \theta_W$ need to be determined.

Stage 2: for given f, g, h, estimate the parameters $(\mathbf{u}_f, \mathbf{u}_g, \mathbf{u}_h)$ and $\boldsymbol{\pi}$ based on

$$\underline{\mathbf{y}} = \underline{\mathbf{F}}\mathbf{u}_f + \underline{\mathbf{G}}\mathbf{u}_g + \underline{\mathbf{H}}\mathbf{u}_h + \underline{\mathbf{X}}\boldsymbol{\pi} + \underline{\widetilde{\mathbf{w}}}, \quad \mathrm{Cov}[\underline{\widetilde{\mathbf{w}}}] = \sigma^2\big[\lambda_W \underline{\boldsymbol{\Sigma}}_W + \mathbf{I}\big]$$

where $\underline{\boldsymbol{\Sigma}}_W = \big[\mathbb{K}_W(\mathbf{z}_k, \mathbf{z}_{k'})\big]_{N\times N}$ is obtained by evaluating every pair of observations either within or between segments. The parameter estimation can be obtained by the generalized least squares.

Obviously, the regularized least squares in Stage 1 extends the MEV Backfitting algorithm to entertain a fourth kernel component $\mathbb{K}_W$. Let us call the extended algorithm *MEVS Backfitting*, where the added letter "S" means the segment effect. Iterate Stage 1 and Stage 2 until convergence, then we obtain the following list of effect estimates (as a summary):

1. Common-factor maturation curve: $\hat{f}(m) = \hat{\mu}_0 + \sum_{i=1}^{I} \hat{\alpha}_i \mathbb{K}_f(m, m_i)$
2. Common-factor exogenous influence: $\hat{g}(t) = \sum_{l=1}^{L} \hat{\beta}_l \mathbb{K}_g(t, t_l)$
3. Common-factor vintage heterogeneity: $\hat{h}(v) = \sum_{j=1}^{J} \hat{\gamma}_j \mathbb{K}_h(v, v_j)$
4. Idiosyncratic multipliers to $(\hat{f}, \hat{g}, \hat{h})$: $u_f^{(k)}, u_g^{(k)}, u_h^{(k)}$ for $k = 1, \ldots, K$
5. Segment-specific covariate effects: $\boldsymbol{\pi}_k$, for $k = 1, \ldots, K$
6. Spatial segment heterogeneity: $\widehat{W}(\mathbf{z}) = \sum_{k=1}^{K} \hat{\omega}_k \mathbb{K}_W(\mathbf{z}, \mathbf{z}_k)$.

3.5 Applications in Credit Risk Modeling

This section presents the applications of MEV decomposition methodology to the real data of (a) Moody's speculative-grade corporate default rates shown in Table 1.1, and (b) the tweaked sample of retail loan loss rates discussed in Chapter II. Both examples demonstrate the rising challenges for analysis of the credit risk data with dual-time coordinates, to which we make only an initial attempt based on the MEV Backfitting algorithm. Our focus is on understanding (or reconstruction) of the marginal effects first, then perform data smoothing on the dual-time domain. To test the MEV decomposition modeling, we begin with a simulation study based on the synthetic vintage data.

3.5.1 Simulation Study

Let us synthetize the year 2000 to 2008 monthly vintages according to the diagram shown in Figure 1.3. Assume each vintage is originated at the very beginning of the month, and its follow-up performance is observed at each month end. Let the horizontal observation window be left truncated from the beginning of 2005 (time 0) and right truncated by the end of 2008 (time 48). Vertically, all the vintages are observed up to 60 months-on-book. It thus ends with the rectangular diagram shown in Figure 3.1, consisting of vintages $j = -60, \ldots, -1$ originated before 2005 and $j = 0, 1, \ldots, 47$ originated after 2005. For each vintage j with origination time v_j, the dual-time responses are simulated by

$$\log(y_{jl}) = f_0(m_{jl}) + g_0(t_{jl}) + h_0(v_j) + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2) \tag{3.38}$$

where t_{jl} runs through $\max\{v_j, 0\}$ to 48 and $m_{jl} = t_{jl} - v_j$. Regardless of the noise, the responses y_{jl} is the product of three marginal effects, $\exp(f_0(m)) \cdot \exp(g_0(t)) \cdot \exp(h_0(v))$. In our simulation setting, the underlying truth $\exp(f_0(m))$ takes the smooth form of inverse Gaussian hazard rate function (1.13) with parameter $c = 6$ and $b = -0.02$, the underlying $\exp(g_0(t))$ is volatile with exponential growth and a jump at $t = 23$, and the underlying $h_0(v)$ is assumed to be a dominating sine wave within $[0, 40]$ and relatively flat elsewhere. See their plots in Figure 3.2 (top panel). Adding the noise with $\sigma = 0.1$, we obtain the synthetic vintage data, whose projection views in lifetime, calendar and vintage origination time are shown in the second panel of Figure 3.2. One may check the pointwise averages (or medians in the vintage box-plots) in each projection view, compare them with the underlying truths in order to see the contamination effects by other marginals.

This synthetic data can be used to demonstrate the flexibility of MEV decomposition technique such that a combination of different kernels can be employed. We manually choose the order-3/2 Matérn kernel (3.13) for $f(m)$, the AdaSS r.k. (3.14) for $g(t)$, and the squared exponential kernel (3.12) for $h(v)$, then run the MEV Backfitting algorithm based on the log-transformed data. Since our data has very few observations in both ends of the vintage direction, we performed binning for $v \geq 40$, as well as binning for $v \leq 0$ based on the prior knowledge that pre-2005 vintages can be treated as a single bucket. Then, the MEV Backtting algorithm outputs the estimated marginal effects plotted in Figure 3.2 (bottom panel). They match the underlying truth except for slight boundary bias.

The GCV-selected structural and smoothing parameters are tabulated in Table 3.1. In fitting $g(t)$ with AdaSS, we have specified the knots τ_k to handle the jump by diagnostic from the raw marginal plots; for simplicity we have enforced ϕ_k to be the same for regions without jump. It is interesting to check the ordering of cross-validated smoothing parameters. By (3.18), the smoothing parameters correspond to the reciprocal ratios of their variance scales to the nugget effect σ^2, therefore, the smaller the smoothing parameter is, the larger variation the corresponding Gaussian process holds. Thus, the ordering of cross-validated $\lambda_f > \lambda_h > \lambda_g$ confirms with the smoothness assumption for f_0, g_0, h_0 in our simulation setting.

Combining the three marginal estimates, and taking the inverse transform gives the fitted rates; see Figure 3.3 for the projection views. The noise-removed recon-

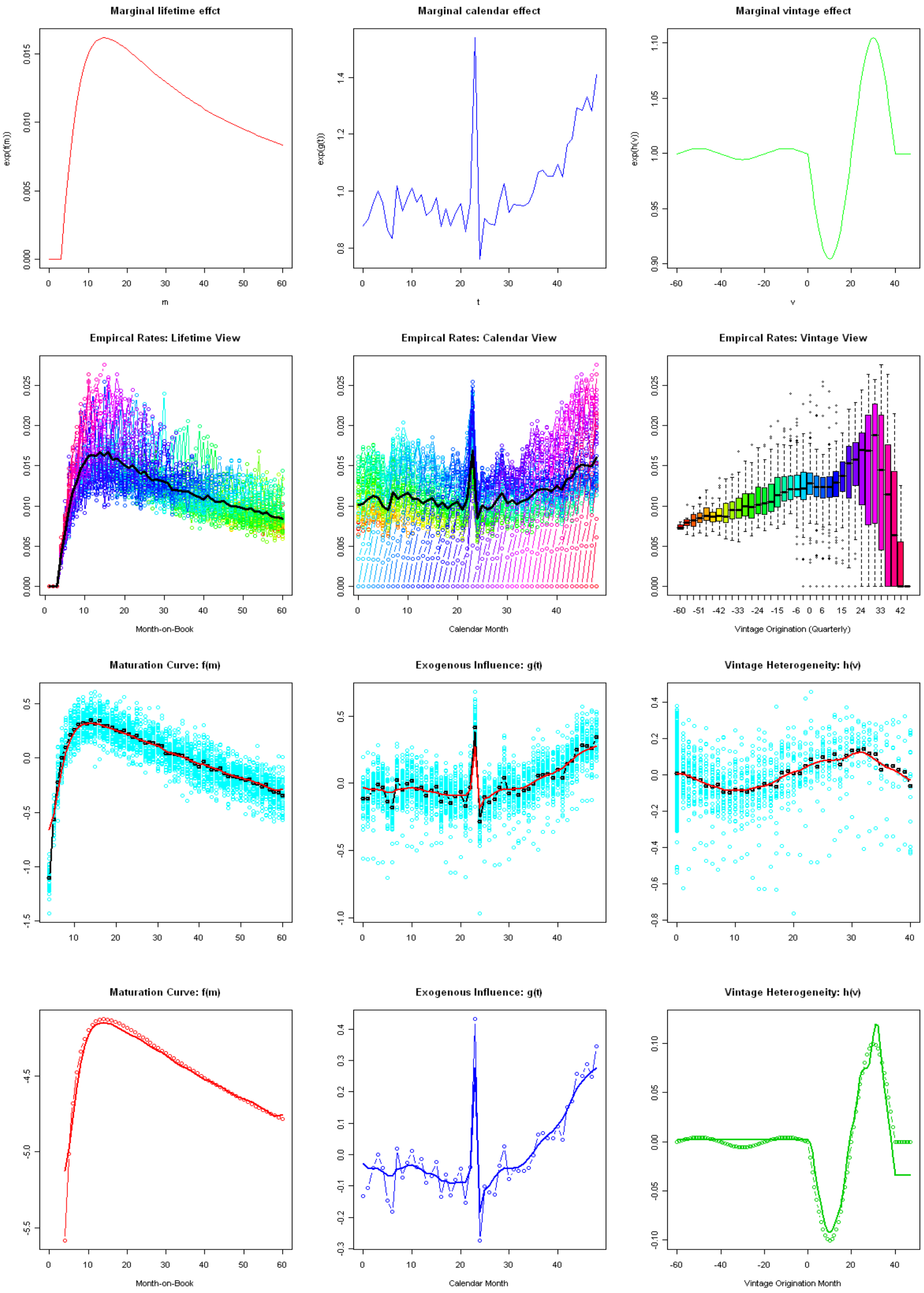


Figure 3.2: Synthetic vintage data analysis: (top) underlying true marginal effects; (2nd) simulation with noise; (3nd) MEV Backfitting algorithm upon convergence; (bottom) Estimation compared to the underlying truth.

Table 3.1: Synthetic vintage data analysis: MEV modeling exercise with GCV-selected structural and smoothing parameters.

MEV	Kernel choice	Structural parameter	Smoothing parameter
$f(m)$	Matérn ($\kappa = 3/2$)	$\phi = 2.872$	20.086
$g(t)$	AdaSS r.k. Eq.(3.14)	$\tau_k = 21, 25, 47$	1.182
		$\phi_k = 0.288, 8.650, 0.288$	
$h(v)$	Squared exponential	$\phi = 3.000$	7.389

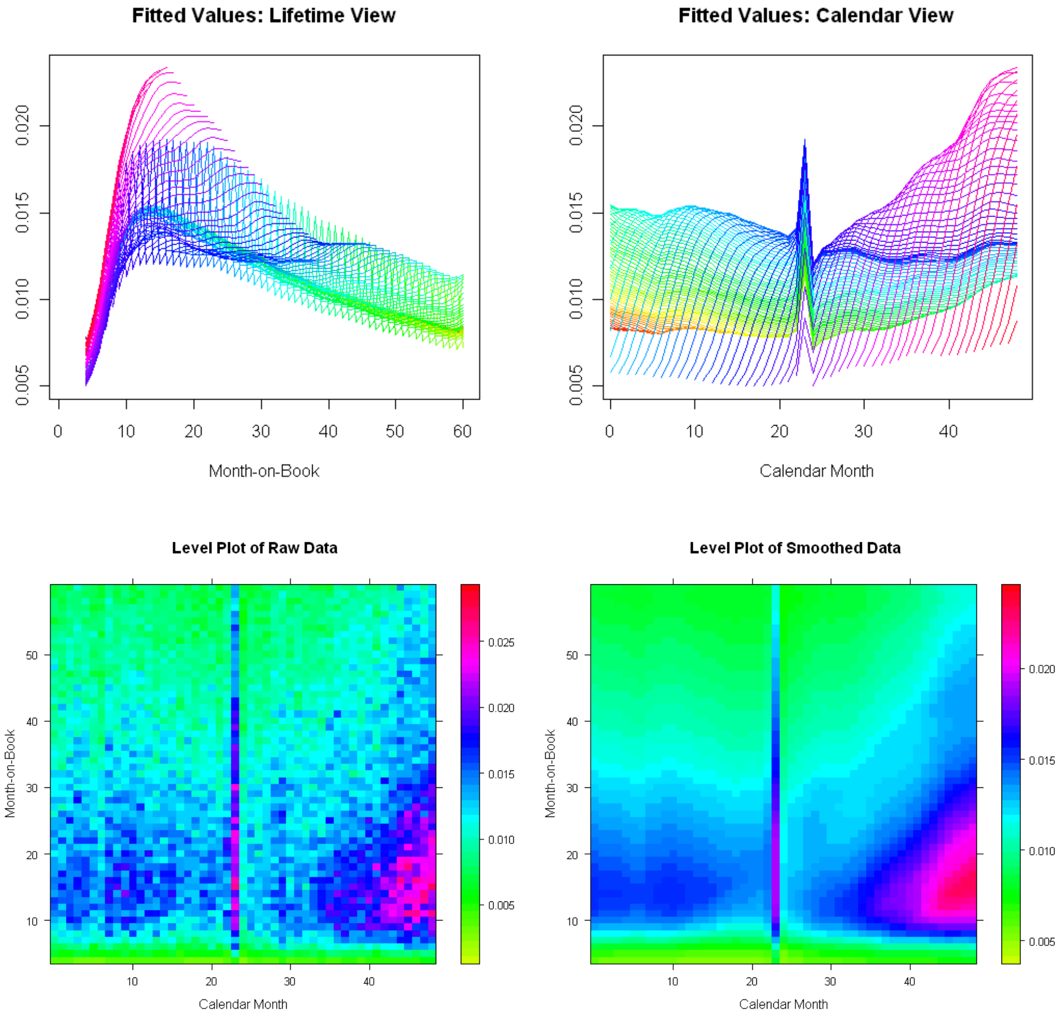


Figure 3.3: Synthetic vintage data analysis: (top) projection views of fitted values; (bottom) data smoothing on vintage diagram.

struction can reveal the prominent marginal features. It is clear that the MEV decomposition modeling performs like a smoother on the dual-time domain. One may also compare the level plots of the raw data versus the reconstructed data in Figure 3.3 to see the dual-time smoothing effect on the vintage diagram.

3.5.2 Corporate Default Rates

In the Figure 1.4 of Chapter 1 we have previewed the annual corporate default rates (in percentage) of Moody's speculative-grade cohorts: 1970-2008. Each yearly cohort is regarded as a vintage. For vintages originated earlier than 1988, the default rates are observed up to 20 years maximum in lifetime; for vintages originated after 1988, the default rates are observed up to the end of year 2008. Such truncation in both lifetime and calendar time corresponds to the trapezoidal prototype of vintage diagram illustrated in Figure 3.1.

Figure 1.4 gives the projection views of the empirical default rates in lifetime, calendar and vintage origination time. By checking the marginal averages or medians plotted on top of each view, it is evident that exogenous influence by macroeconomic environment is more fluctuating than both the maturation and vintage effects. However, such marginal plots are contaminated by each other; see e.g. the many spikes shown in the lifetime projection actually correspond to the exogenous effects. It is our purpose to separate these marginal effects.

We made slight data preprocessing. First, the transform function is pre-specified to be $\eta(y) = \log(y/100)$ where y is the raw percentage measurements and zero responses are treated as missing values (one may also match them to a small positive value). Second, since there are very limited number of observations for small calendar time, the exogenous effects $g(t)$ for $t = 2, 3$ are merged to $t = 4$, and the leftmost observation at $t = 1$ is removed because of its jumpy behavior. Besides, the vintage effects are merged for large $v = 30, \ldots, 39$.

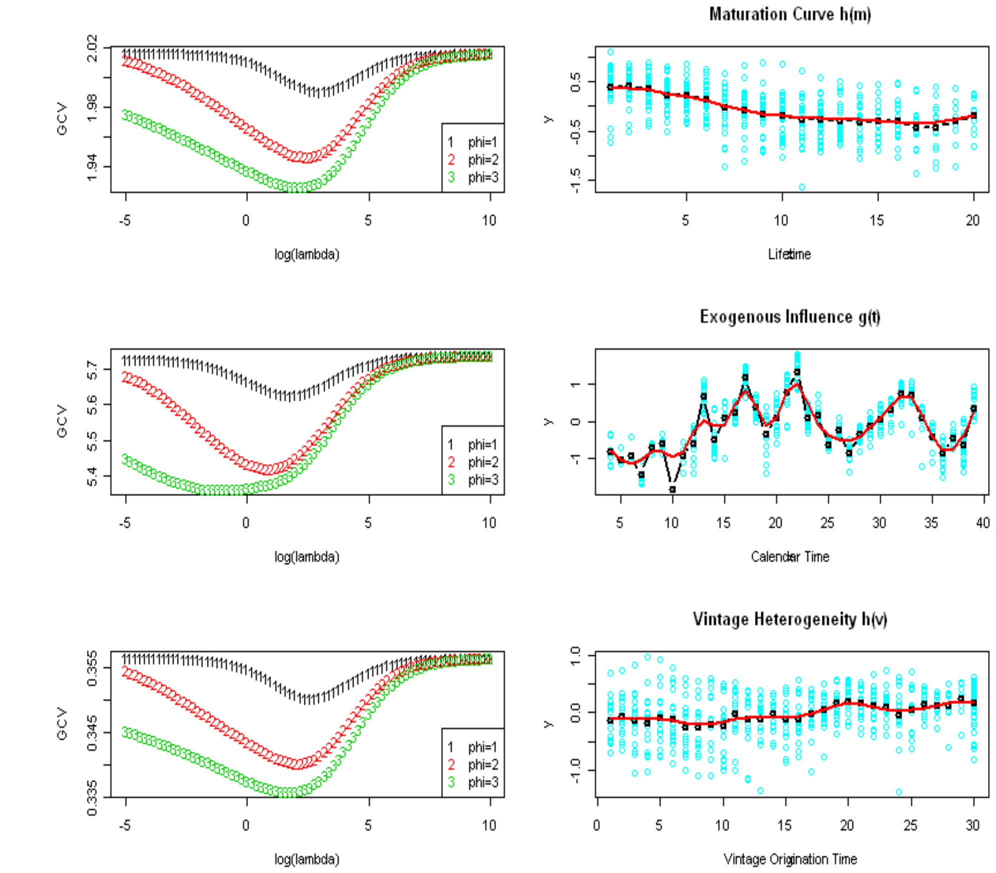


Figure 3.4: Results of MEV Backfitting algorithm upon convergence (right panel) based on the squared exponential kernels. Shown in the left panel is the GCV selection of smoothing and structural parameters.

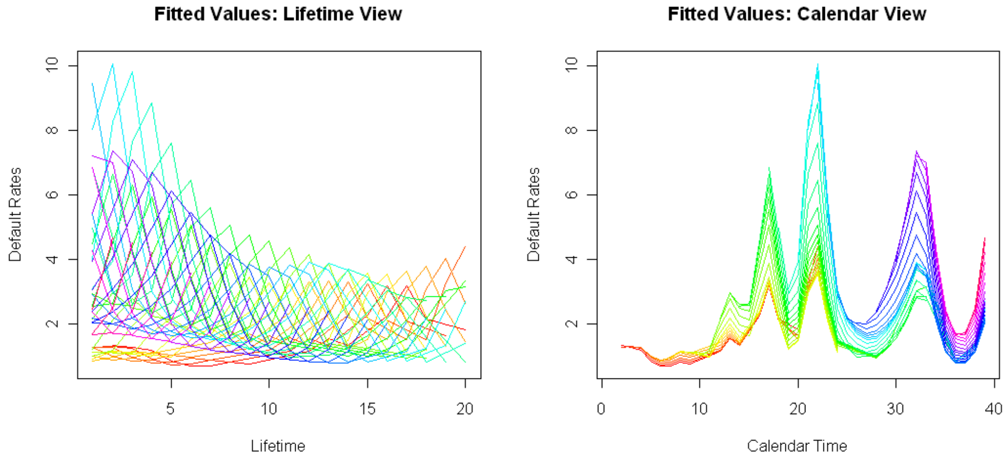


Figure 3.5: MEV fitted values: Moody's-rated Corporate Default Rates

To run the MEV Backfitting algorithm, $\mathbb{K}_f, \mathbb{K}_g, \mathbb{K}_h$ are all chosen to be the squared exponential covariance kernels, for which the mean function is fixed to be the intercept effect. For each univariate Gaussian process fitting, say $Z_f(m)$, the GCV criterion selects the smoothing parameter λ_f, as well as the structural parameter θ_f (i.e. the scale parameter ϕ in (3.12)). For demonstration, we used the grid search on the log scale of λ with only 3 different choices of $\phi = 1, 2, 3$ ($\phi = 3$ is set to be the largest scale for stability, otherwise the reciprocal conditional number of the ridged regression matrix would vanish). The backfitting results upon convergence are shown in Figure 3.4, where GCV selects the largest scale parameter in each iteration. An interesting result is the GCV-selected smoothing parameter λ, which turns out to be $9.025, 0.368, 6.050$ for Z_f, Z_g, Z_h, respectively. By (3.18), it is implied that the exogenous influence shares the largest variation, followed by the vintage heterogeneity, while the maturation curve has the least variation. Furthermore, the changes of estimated exogenous effects in Figure 3.4 follow the NBER recession dates shown in Figure 1.1.

The inverse transform $\eta^{-1}(\cdot)$ maps the fitted values to the original percentage scale. Figure 3.5 plots the fitted values in both lifetime and calendar views. Compared to the raw plots in Figure 1.4, the MEV modeling could retain some of most interesting features, while smoothing out others.

3.5.3 Retail Loan Loss Rates

Recall the tweaked credit-risk sample discussed in Chapter II. It was used as a motivating example for the development of adaptive smoothing spline in fitting the pointwise averages that suppose to smooth out in lifetime (month-on-book, in this case). However, the raw responses behave rather abnormally in large month-on-book, partly because of the small size of replicates, and partly because of the contamination by the exogenous macroeconomic influence. In this section, the same tweaked sample is analyzed on the dual-time domain, for which we model not only the maturation curve, but also the exogenous influence and the vintage heterogeneity effects.

The tweaked sample of dual-time observations are supported on the *triangular vintage diagram* illustrated in Figure 3.1, which is mostly typical in retail risk management. They measure the loss rates in retail revolving exposures from January 2000 to December 2006. Figure 3.6 (top panel) shows the projection views in lifetime, calendar and vintage origination time, where the new vintages originated in 2006 are excluded since they have too short window of performance measurements. Some of the interesting features for such vintage data are listed below:

1. The rates are fixed to be zero for small month-on-book $m \leq 4$.

2. There are less and less observations when the month-on-book m increases.

3. There are less and less observations when the calendar month t decreases.

4. The rates behave relatively smooth in m, from the lifetime view.

5. The rates behave rather dynamic and volatile in t, from the calendar view.

6. The rates show heterogeneity across vintages, from the vintage view.

To model the loss rates, timescale binding was performed first for large month-on-book $m \geq 72$ and small calendar time $t \leq 12$. We pre-specified the log transform for the positive responses (upon removal of zero responses). The MEV decomposition model takes the form of $\log(y_{jl}) = f(m_{jl}) + g(t_{jl}) + h(v_j) + \varepsilon$, subject to $\mathbb{E}g = \mathbb{E}h = 0$. In this way, the original responses are decomposed multiplicatively to be $e^{f(m)}$, $e^{g(t)}$ and $e^{h(v)}$, where the exponentiated maturation curve could be regarded as the baseline, together with the exponentiated exogenous and vintage heterogeneity multipliers.

In this MEV modeling exercise, the Matérn kernels were chosen to run the MEV Backfitting algorithm, and the GCV criterion was used to select the smoothing and structural parameters. The fitting results are presented in Figure 3.6 (bottom panel),

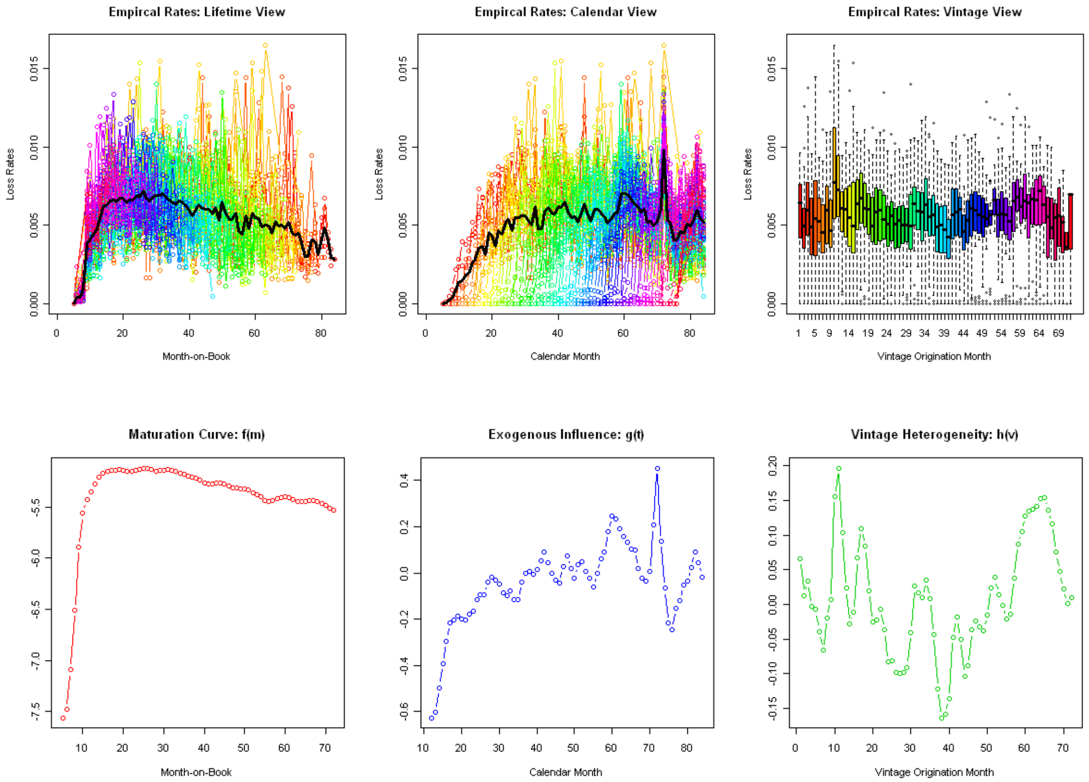


Figure 3.6: Vintage data analysis of retail loan loss rates: (top) projection views of emprical loss rates in lifetime m, calendar t and vintage origination time v; (bottom) MEV decomposition effects $\hat{f}(m)$, $\hat{g}(t)$ and $\hat{h}(v)$ (at log scale).

which shows the estimated $\hat{f}(m), \hat{g}(t), \hat{h}(v)$ in the log response scale. Each of these estimated functions follows the general dynamic pattern of the corresponding marginal averages in the empirical plots above. The differences in local regions reflect the improvements in each marginal direction after removing the contamination of other marginals. Compared to both the exogenous and vintage heterogeneity effects, the maturation curve is rather smooth. It is also worth pointing out that the exogenous spike could be captured relatively well by the Matérn kernel.

3.6 Discussion

For the vintage data that emerge in credit risk management, we propose an MEV decomposition framework to estimate the maturation curve, exogenous influence and vintage heterogeneity on the dual-time domain. One difficulty associated with the vintage data is the intrinsic identification problem due to their hyperplane nature, which also appears in the conventional cohort analysis under the age-period-cohort model. To regularize the three-way marginal effects, we have studied nonparametric smoothing based on Gaussian processes, and demonstrated its flexibility through choosing covariance kernels, structural and smoothing parameters. An efficient MEV Backfitting algorithm is provided for model estimation. The new technique is then tested through a simulation study and applied to analyze both examples of corporate default rates and retail loss rates, for which some preliminary results are presented.

Beyond our initial attempt made in this chapter, there are other open problems associated with the vintage data analysis (VDA) worthy of future investigation.

1. The dual-time model assessment and validation remains an issue. In the MEV Backfitting algorithm we have used the leave-one-out cross-validation for fitting each marginal component function. It is however not directly performing cross-validation on the dual-time domain, which would be an interesting subject of future study.

2. The model forecasting is often a practical need in financial risk management, as it is concerning about the future uncertainty. To make prediction within the MEV modeling framework, one needs to extrapolate $f(m), g(t), h(v)$ in each marginal direction in order to know about their behaviors out of the training bounds of m, t, v. By Kriging theory, for any $\mathbf{x} = (m, t; v)$ on the vintage diagram, the formula (3.22) provides the conditional expectation given the historical performance. Furthermore, it is straightforward to simulate the random paths for a sequence of inputs $(\underline{\mathbf{x}}_1, \ldots, \underline{\mathbf{x}}_p)$ based on the multivariate conditional normal distribution $N_p(\boldsymbol{\mu}^{(c)}, \boldsymbol{\Sigma}^{(c)})$, where the conditional mean and covariance

are given by

$$\begin{cases} \boldsymbol{\mu}^{(c)} = \underline{\mathbf{B}}\widehat{\boldsymbol{\mu}} + \underline{\Upsilon}\big[\boldsymbol{\Sigma} + \mathbf{I}\big]^{-1}\big(\mathbf{y} - \mathbf{B}\widehat{\boldsymbol{\mu}}\big) \\ \boldsymbol{\Sigma}^{(c)} = \underline{\boldsymbol{\Sigma}} - \underline{\Upsilon}\big[\boldsymbol{\Sigma} + \mathbf{I}\big]^{-1}\underline{\Upsilon}^T \end{cases} \text{with} \begin{cases} \underline{\Upsilon} = \big[\mathbb{K}(\underline{\mathbf{x}}_i, \mathbf{x}_{jl})\big]_{p\times n} \\ \underline{\boldsymbol{\Sigma}} = \big[\mathbb{K}(\underline{\mathbf{x}}_i, \underline{\mathbf{x}}_j)\big]_{p\times p} \end{cases} \quad (3.39)$$

using the aggregated kernel given by (3.18) and the notations in (3.19). However, there is a word of caution about making prediction of the future from the historical training data. In the MEV modeling framework, extrapolating the maturation curve $f(m)$ in lifetime is relatively safer than extrapolating the exogenous and vintage effects $g(t)$ and $h(v)$, since $f(m)$ is of self-maturation nature, while $g(t)$ can be interfered by future macroeconomic changes and $h(v)$ is also subject to the future changes of vintage origination policy. Before making forecasts in terms of calendar time, the out-of-sample test is recommended.

3. For extrapolation of MEV component functions, parametric modeling approach is sometimes more tractable than using Gaussian processes. For example, if given the prior knowledge that the maturation curve $f(m)$ follows the shape of the inverse Gaussian hazard rate (1.13), like our synthetic case in Section 3.5, one may estimate the parametric $f(m)$ with nonlinear regression technique. As for the exogenous effect $g(t)$ that is usually dynamic and volatile, one may use parametric time series models; see e.g. Fan and Yao (2003). Such parametric approach would benefit the loss forecasting as they become more and more validated through experiences.

4. The MEV decomposition we have considered so far is simplified to exclude the interaction effects involving two or all of age m, time t and vintage v arguments. In formulating the models by Gaussian processes, the mutual independence among $Z_f(m), Z_g(t), Z_h(v)$ is also postulated. One of the ideas for taking into account the interaction effects is to cross-correlate $Z_f(m), Z_g(t), Z_h(v)$. Consider

the combined Gaussian process (3.17) with covariance

$$\begin{aligned}&\mathrm{Cov}[Z_f(m)+Z_g(t)+Z_h(v), Z_f(m')+Z_g(t')+Z_h(v')]\\ &= \mathbb{E}Z_f(m)Z_f(m') + \mathbb{E}Z_f(m)Z_g(t') + \mathbb{E}Z_f(m)Z_h(v')\\ &\quad + \mathbb{E}Z_g(t)Z_f(m') + \mathbb{E}Z_g(t)Z_g(t') + \mathbb{E}Z_g(t)Z_h(v')\\ &\quad + \mathbb{E}Z_h(v)Z_f(m') + \mathbb{E}Z_h(v)Z_g(t') + \mathbb{E}Z_h(v)Z_h(v'),\end{aligned}$$

which reduces to (3.18) when the cross-correlations between $Z_f(m), Z_g(t)$ and $Z_h(v)$ are all zero. In general, it is straightforward to incorporate the cross-correlation structure and follow immediately the MEV Kriging procedure. For example, one of our works in process is to assume that

$$\mathbb{E}Z_g(t)Z_h(v) = \frac{\sigma^2}{\sqrt{\lambda_g\lambda_h}}\mathbb{K}_{gh}(t,v), \quad \text{where } \mathbb{K}_{gh}(t,v) = \rho I(t \geq v),\ \rho \geq 0$$

in order to study the interaction effect between the vintage heterogeneity and the exogenous influence. Results will be presented by a future report.

Besides the open problems listed above, the scope of VDA can be also generalized to study non-continuous types of observations. For examples we considered in Section 3.5, the Moody's default rates were actually calculated from some raw binary indicators (of default or not), while the loss rates in the synthetic retail data can be viewed as the ratio of loss units over the total units. Thus, if we are given the raw binary observations or unit counts, one may consider the generalized MEV models upon the specification of a link function, as an analogy to the generalized linear models (Mccullagh and Nelder, 1989). The transform η in (3.2) plays a similar role as such a link function, except for that it is functioning on the rates that are assumed to be directly observed.

One more extension of VDA is to analyze the time-to-default data on the vintage diagram, which is deferred to the next chapter where we will develop the dual-time survival analysis.

3.7 Technical Proofs

Proof of Lemma 3.1: By that $m-t+v=0$ on the vintage diagram Ω, at least one of the linear parts f_0, g_0, h_0 is not identifiable, since otherwise $f_0(m)-m, g_0(t)+t, h_0(v)-v$ would always satisfy (3.2).

Proof of Proposition 3.2: By applying the general spline theorem to (3.23), the target function has a finite-dimensional representation $\zeta(\mathbf{x}) = \mu_0+\mu_1 m+\mu_2 t+\mathbf{c}^T\boldsymbol{\xi}_n$, where $\mathbf{c}^T\boldsymbol{\xi}_n = \sum_{j=1}^{J}\sum_{l=1}^{L_j} c_{jl}\mathbb{K}(\mathbf{x},\mathbf{x}_{jl})$. Substitute it back into (3.23) and use the reproducing property $\langle \mathbb{K}(\cdot,\mathbf{x}_{jl}), \mathbb{K}(\cdot,\mathbf{x}_{j'l'})\rangle_{\mathcal{H}_1} = \mathbb{K}(\mathbf{x}_{jl},\mathbf{x}_{j'l'})$, then we get

$$\min\left\{\big\|\mathbf{y}-\mathbf{B}\boldsymbol{\mu}-\boldsymbol{\Sigma}\mathbf{c}\big\|^2 + \big\|\mathbf{c}\big\|_{\boldsymbol{\Sigma}}\right\}$$

using the notations in (3.19). Taking the partial derivatives w.r.t. $\boldsymbol{\mu}$ and $\mathbf{c}$ and setting them to zero gives the solution

$$\begin{aligned}\widehat{\boldsymbol{\mu}} &= \left(\mathbf{B}^T(\boldsymbol{\Sigma}+\mathbf{I})^{-1}\mathbf{B}\right)^{-1}\mathbf{B}^T(\boldsymbol{\Sigma}+\mathbf{I})^{-1}\mathbf{y}\\ \widehat{\mathbf{c}} &= \left(\boldsymbol{\Sigma}+\mathbf{I}\right)^{-1}(\mathbf{y}-\mathbf{B}\widehat{\boldsymbol{\mu}}),\end{aligned}$$

which leads $\zeta(\mathbf{x})$ to have the same form as the Kriging predictor (3.22).

Proof of Proposition 3.3: By (3.18) and comparing (3.24) and (3.26), it is not hard to derive the relationships between two sets of kernel coefficients,

$$\alpha_i = \lambda_f^{-1}\sum_{j,l} c_{jl}I(m_{jl}=m_i),\quad \beta_l = \lambda_g^{-1}\sum_{j,l} c_{jl}I(t_{jl}=t_l),\quad \gamma_j = \lambda_h^{-1}\sum_{l=1}^{L_j} c_{jl} \tag{3.40}$$

that aggregate c_{jl} for the replicated kernel bases for every m_i, t_l, v_j, respectively.

Then, using the new vectors of coefficients $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_I)^T, \boldsymbol{\beta} = (\beta_1, \ldots, \beta_L)^T$ and $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_J)^T$, we have that

$$\begin{aligned} \boldsymbol{\Sigma}\mathbf{c} &= \widetilde{\boldsymbol{\Sigma}}_f \boldsymbol{\alpha} + \widetilde{\boldsymbol{\Sigma}}_g \boldsymbol{\beta} + \widetilde{\boldsymbol{\Sigma}}_h \boldsymbol{\gamma} \\ \mathbf{c}^T \boldsymbol{\Sigma}\mathbf{c} &= \lambda_f \boldsymbol{\alpha}^T \boldsymbol{\Sigma}_f \boldsymbol{\alpha} + \lambda_g \boldsymbol{\beta}^T \boldsymbol{\Sigma}_g \boldsymbol{\beta} + \lambda_h \boldsymbol{\gamma}^T \boldsymbol{\Sigma}_h \boldsymbol{\gamma}. \end{aligned}$$

Plugging them into (3.25) leads to (3.27), as asserted.

CHAPTER IV

Dual-time Survival Analysis

4.1 Introduction

Consider in this chapter the default events naturally observed on the dual-time domain, lifetime (or, age) in one dimension and calendar time in the other. Of our interest is the time-to-default from multiple origins, such that the events happening at the same calendar time may differ in age. Age could be the lifetime of a person in a common sense, and it could also refer to the elapsed time since initiation of a treatment. In credit risk modeling, age refers to the duration of a credit account since its origination (either a corporate bond or a retail loan). For many accounts having the same origin, they share the same age at any follow-up calendar time, and group as a single vintage.

Refer to Chapter I about basics of survival analysis in lifetime scale, where the notion of hazard rate is also referred to be the default intensity in credit risk context. As introduced in Chapter 1.4, dual-time observations of survival data subject to truncation and censoring can be graphically represented by the Lexis diagram (Keiding, 1990). To begin with, let us simulate a Lexis diagram of *dual-time-to-default* observations:

$$(V_j, U_{ji}, M_{ji}, \Delta_{ji}), \quad i = 1, \ldots, n_j, \quad j = 1, \ldots, J \tag{4.1}$$

where V_j denotes the origination time of the jth vintage that consists of n_j accounts. For each account, U_{ji} denotes the time of left truncation, M_{ji} denotes the lifetime of event termination, and Δ_{ji} indicates the status of termination such that $\Delta_{ji} = 1$ if the default event occurs prior to getting censored and 0 otherwise. Let us fix the rectangular vintage diagram with age $m \in [0, 60]$, calendar time $t \in [0, 48]$ and vintage origination $V_j = -60, \ldots, -1, 0, 1, \ldots, 47$, same as the synthetic vintage data in Chapter 3.5. So, the observations are left truncated at time 0. For the censoring mechanism, besides the systematic censoring with $m_{\max} = 60$ and $t_{\max} = 48$, an independent random censoring with the constant hazard rate λ_{att} is used to represent the account attrition (i.e., a memoryless exponential lifetime distribution).

Denote by $\lambda_v(m, t)$ or $\lambda_j(m, t)$ the underlying dual-time hazard rate of vintage $v = V_j$ at lifetime m and calendar time t. Then, one may simulate the dual-time-to-default data (4.1) according to:

$$
\begin{aligned}
\tau_{ji} &= \inf\left\{m : \int_0^m \lambda_j(s, V_j + s)ds \geq -\log X,\ X \sim \mathsf{Unif}\big[0, 1\big]\right\} \text{ (default)} \\
C_{ji} &= \inf\left\{m : \lambda_{\text{att}} m \geq -\log X,\ X \sim \mathsf{Unif}\big[0, 1\big]\right\} \text{ (attrition)} \\
U_{ji} &= \max\{\bar{U}_{ji}, V_j\} - V_j, \quad \text{with } \bar{U}_{ji} \equiv 0 \text{ (left truncation)} \\
&\text{Book } (V_j, U_{ji}, M_{ji}, \Delta_{ji}) \text{ if } U_{ji} < \min\{\tau_{ji}, C_{ji}\}: \qquad (4.2) \\
M_{ji} &= \min\{\tau_{ji},\ C_{ji},\ m_{\max},\ t_{\max} - V_j\} \text{ (termination time)} \\
\Delta_{ji} &= I\Big(\tau_{ji} \leq \min\{C_{ji}, m_{\max}, t_{\max} - V_j\}\Big) \text{ (default indicator)}
\end{aligned}
$$

for each given $V_j = -60, \ldots, -1, 0, 1, \ldots, 47$. For each vintage j, one may book n_j number of origination accounts by repeatedly running (4.2) n_j times independently. Let us begin with $n_j \equiv 1000$ and specify the attrition rate $\lambda_{\text{att}} = 50$bp (a basis point is 1%). Among other types of dual-time default intensities, assume the multiplicative form

$$
\lambda_v(m, t) = \lambda_f(m)\lambda_g(t)\lambda_h(v) = \exp\Big\{f(m) + g(t) + h(v)\Big\} \qquad (4.3)
$$

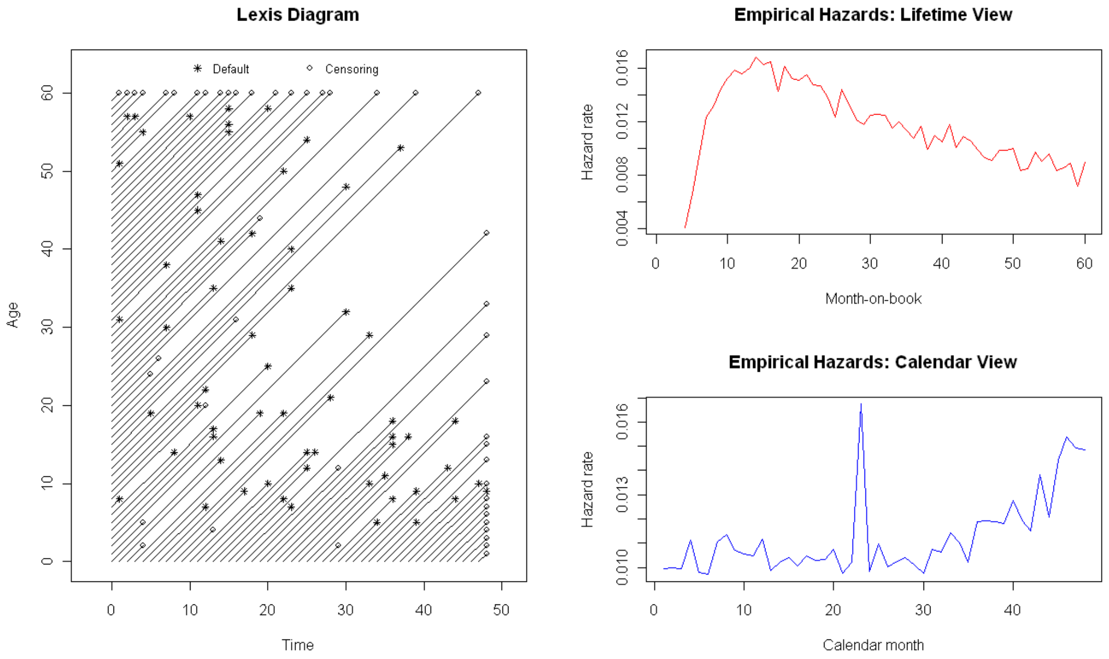


Figure 4.1: Dual-time-to-default: (left) Lexis diagram of sub-sampled simulations; (right) empirical hazard rates in either lifetime or calendar time.

where the maturation curve $f(m)$, the exogenous influence $g(t)$ and the unobserved heterogeneity $h(v)$ are interpreted in the sense of Chapter III and they are assumed to follow Figure 3.2 (top panel). For simulating the left-truncated survival accounts, we assume the zero exogenous influence on their historical hazards.

Figure 4.1 (left) is an illustration of the Lexis diagram for one random simulation per vintage, where each asterisk at the end of a line segment denotes a default event and the circle denotes the censorship. To see default behavior in dual-time coordinates, we calculated the empirical hazard rates by (4.14) in lifetime and (4.15) in calendar time, respectively, and the results are plotted in Figure 4.1 (right). Each marginal view of the hazard rates shows clearly the pattern matching our underlying assumption, namely

1. The endogenous (or maturation) performance in lifetime is relatively smooth;

2. The exogenous performance in calendar time is relatively dynamic and volatile.

This is typically the case in credit risk modeling, where the default risk is greatly affected by macroeconomic conditions. It is our purpose to develop the dual-time survival analysis in not only lifetime, but also simultaneously in calendar time.

Despite that the Lexis diagram has appeared since Lexis (1875), statistical literature of survival analysis has dominantly focused on the lifetime only rather than the dual-time scales; see e.g. Andersen, et al. (1993) with nine to one chapter coverage. In another word, lifetime has been taken for granted as the basic timescale of survival, with very few exceptions that take either calendar time as the primary scale (e.g. Arjas (1986)) or both time scales (see the survey by Keiding (1990)). There is also discussion on selection of the appropriate timescale in the context of multiple timescale reduction; see e.g. Farewell and Cox (1979) about a rotation technique based on a naive Cox proportional hazards (CoxPH) model involving only the linearized time-covariate effect. Such timescale dimension reduction approach can be also referred to Oakes (1995) and Duchesne and Lawless (2000), given the multiple dimensions that may include usage scales like mileage of a vehicle.

The literature of regression models based on dual-time-to-default data (a.k.a. survival data with staggered entry) has also concentrated on the use of lifetime as the basic timescale together with time-dependent covariates. Mostly used is the CoxPH regression model with an arbitrary baseline hazard function in lifetime. Some weak convergence results can be referred to Sellke and Siegmund (1983) by martingale approximation and Billias, Gu and Ying (1997) by empirical process theory; see also the related works of Slud (1984) and Gu and Lai (1991) on two-sample tests. However, such one-way baseline model specification is unable to capture the baseline hazard in calendar time. Then, it comes to Efron (2002) with symmetric treatment of dual timescales under a two-way proportional hazards model

$$\lambda_{ji}(m,t) = \lambda_f(m)\lambda_g(t)\exp\{\boldsymbol{\theta}^T\mathbf{z}_{ji}(m)\}, \quad i = 1,\dots,n_j,\ j = 1,\dots,J. \tag{4.4}$$

Note that Efron (2002) considered only the special case of (4.4) with cubic polynomial expansion for $\log \lambda_f$ and $\log \lambda_g$, as well as the time-independent covariates. In general, we may allow for arbitrary (non-negative) $\lambda_f(m)$ and $\lambda_g(t)$, hence name (4.4) to be a *two-way Cox regression* model.

In credit risk, there seems to be no formal literature on dual-time-to-default risk modeling. It is our objective of this chapter to develop the dual-time survival analysis (DtSA) for credit risk modeling, including the methods of nonparametric estimation, structural parameterization, intensity-based semiparametric regression, and frailty specification accounting for vintage heterogeneity effects. The aforementioned dual-time Cox model (4.4) will be one of such developments, whereas the Efron's approach with polynomial baselines will fail to capture the rather different behaviors of endogenous and exogenous hazards illustrated in Figure 4.1.

This chapter is organized as follows. In Section 4.2 we start with the generic form of likelihood function based on the dual-time-to-default data (4.1), then consider the one-way, two-way and three-way nonparametric estimators on Lexis diagram. The simulation data described above will be used to assess the performance of nonparametric estimators. In Section 4.3 we discuss the dual-time feasibility of structural models based on the first-passage-time triggering system, which is defined by an endogenous distance-to-default process associated with the exogenous time-transformation. Section 4.4 is devoted to the development of dual-time semiparametric Cox regression with both endogenous and exogenous baseline hazards and covariate effects, for which the method of partial likelihood estimation plays a key role. Also discussed is the frailty type of vintage heterogeneity by a random-effect formulation. In Section 4.5, we demonstrate the applications of both nonparametric and semiparametric dual-time survival analysis to credit card and mortgage risk modeling, based on our real data-analytic experiences in retail banking. We conclude the chapter in Section 4.6 and provide some supplementary materials at the end.

4.2 Nonparametric Methods

Denote by $\lambda_{ji}(m) \equiv \lambda_{ji}(m,t)$ with $t \equiv V_j + m$ the hazard rates of lifetime-to-default τ_{ji} on the Lexis diagram. Given the dual-time-to-default data (4.1) with independent left-truncation and right-censoring, consider the joint likelihood under either continuous or discrete setting:

$$\prod_{j=1}^{J}\prod_{i=1}^{n_j}\left[\frac{p_{ji}(M_{ji})}{S_{ji}(U_{ji}^+)}\right]^{\Delta_{ji}}\left[\frac{S_{ji}(M_{ji}^+)}{S_{ji}(U_{ji}^+)}\right]^{1-\Delta_{ji}} \tag{4.5}$$

where $p_{ji}(m)$ denotes the density and $S_{ji}(m)$ denotes the survival function of τ_{ji}. On the continuous domain, $m^+ = m$, $S_{ji}(m) = e^{-\Lambda_{ji}(m)}$ with the cumulative hazard $\Lambda_{ji}(m) = \int_0^m \lambda_{ji}(s)ds$, then the log-likelihood function has the generic form

$$\begin{aligned}\ell &= \sum_{j=1}^{J}\sum_{i=1}^{n_j}\Delta_{ji}\log\lambda_{ji}(M_{ji}) + \log S_{ji}(M_{ji}) - \log S_{ji}(U_{ji})\\ &= \sum_{j=1}^{J}\sum_{i=1}^{n_j}\left\{\Delta_{ji}\log\lambda_{ji}(M_{ji}) - \int_{U_{ji}}^{M_{ji}}\lambda_{ji}(m)dm\right\}. \end{aligned}\tag{4.6}$$

On the discrete domain of m, $S_{ji}(m) = \prod_{s<m}[1-\lambda_{ji}(s)]$ and the likelihood function (4.5) can be rewritten as

$$\prod_{j=1}^{J}\prod_{i=1}^{n_j}[\lambda_{ji}(M_{ji})]^{\Delta_{ji}}[1-\lambda_{ji}(M_{ji})]^{1-\Delta_{ji}}\prod_{m\in(U_{ji},M_{ji})}[1-\lambda_{ji}(m)] \tag{4.7}$$

and the log-likelihood function is given by

$$\ell = \sum_{j=1}^{J}\sum_{i=1}^{n_j}\left\{\Delta_{ji}\log\left(\frac{\lambda_{ji}(M_{ji})}{1-\lambda_{ji}(M_{ji})}\right) + \sum_{m\in(U_{ji},M_{ji}]}\log[1-\lambda_{ji}(m)]\right\}. \tag{4.8}$$

In what follows we discuss one-way, two-way and three-way nonparametric approach to the maximum likelihood estimation (MLE) of the underlying hazard rates.

4.2.1 Empirical Hazards

Prior to discussion of one-way nonparametric estimation, let us review the classical approach to the empirical hazards vintage by vintage. For each fixed V_j, let $\lambda_j(m)$ be the underlying hazard rate in lifetime. Given the data (4.1),

$$\begin{cases} \text{number of defaults:} & \mathsf{nevent}_j(m) = \sum_{i=1}^{n_j} I(M_{ji} = m, \Delta_{ji} = 1) \\ \text{number of at-risk's:} & \mathsf{nrisk}_j(m) = \sum_{i=1}^{n_j} I(U_{ji} < m \leq M_{ji}) \end{cases} \tag{4.9}$$

for a finite number of lifetime points m with $\mathsf{nevent}_j(m) \geq 1$. Then, the jth-vintage contribution to the log-likelihood (4.8) can be rewritten as

$$\ell_j = \sum_m \mathsf{nevent}_j(m) \log \lambda_j(m) + \big(\mathsf{nrisk}_j(m) - \mathsf{nevent}_j(m)\big) \log[1 - \lambda_j(m)] \tag{4.10}$$

on the discrete domain of m. It is easy to show that the MLE of $\lambda_j(m)$ is given by the following empirical hazard rates

$$\widehat{\lambda}_j(m) = \frac{\mathsf{nevent}_j(m)}{\mathsf{nrisk}_j(m)} \quad \text{if } \mathsf{nevent}_j(m) \geq 1 \text{ and } \mathsf{NaN} \text{ otherwise.} \tag{4.11}$$

They can be used to express the well-known nonparametric estimators for cumulative hazard and survival function

$$\text{(Nelson-Aalen estimator)} \qquad \widehat{\Lambda}_j(m) = \sum_{m_i \leq m} \widehat{\lambda}_j(m_i); \tag{4.12}$$

$$\text{(Kaplan-Meier estimator)} \qquad \widehat{S}_j(m) = \prod_{m_i \leq m^-} \big[1 - \widehat{\lambda}_j(m_i)\big] \tag{4.13}$$

for any $m \geq 0$ under either continuous or discrete setting. Clearly, the Kaplan-Meier estimator and the Nelson-Aalen estimator are related empirically by $\widehat{S}_j(m) = \prod_{m_i \leq m^-} \big[1 - \Delta\widehat{\Lambda}(m_i)\big]$, where the empirical hazard rates $\widehat{\lambda}(m_i) \equiv \Delta\widehat{\Lambda}(m_i)$ are also called the Nelson-Aalen increments.

Now, we are ready to present the one-way nonparametric estimation based on the complete dual-time-to-default data (4.1). In the lifetime dimension (vertical of Lexis diagram), assume $\lambda_{ji}(m,t) = \lambda(m)$ for all (j,i), then we may immediately extend (4.11) by simple aggregation to obtain the one-way estimator in lifetime:

$$\widehat{\lambda}(m) = \frac{\sum_{j=1}^{J} \mathsf{nevent}_j(m)}{\sum_{j=1}^{J} \mathsf{nrisk}_j(m)}, \tag{4.14}$$

whenever the numerator is positive. Alternatively, in the calendar time (horizontal of Lexis diagram) assume $\lambda_{ji}(m,t) = \lambda(t)$ for all (j,i), then we may obtain the one-way nonparametric estimator in calendar time:

$$\widehat{\lambda}(t) = \frac{\sum_{j=1}^{J} \mathsf{nevent}_j(t - V_j)}{\sum_{j=1}^{J} \mathsf{nrisk}_j(t - V_j)}, \tag{4.15}$$

whenever the numerator is positive. Accordingly, one may obtain the empirical cumulative hazard and survival function by the Nelson-Aalen estimator (4.12) and the Kaplan-Meier estimator (4.13) in each marginal dimension.

For example, consider the simulation data introduced in Section 4.1. Both one-way nonparametric estimates are plotted in the right panel of Figure 4.1 (upon linear interpolation), which may capture roughly the endogenous and exogenous hazards. However, they may be biased in other cases we shall discuss next.

4.2.2 DtBreslow Estimator

The two-way nonparametric approach is to estimate the lifetime hazards $\lambda_f(m)$ and the calendar hazards $\lambda_g(t)$ simultaneously under the multiplicative model:

$$\lambda_{ji}(m,t) = \lambda_f(m)\lambda_g(t) = \exp\{f(m) + g(t)\}, \quad \text{for all } (j,i). \tag{4.16}$$

For identifiability, assume that $\mathbb{E}\log\lambda_g = \mathbb{E}g = 0$.

Following the one-way empirical hazards (4.11) above, we may restrict to the pointwise parametrization

$$\hat{f}(m) = \sum_{l=1}^{L_m} \alpha_l I(m = m_{[l]}), \quad \hat{g}(t) = \sum_{l=1}^{L_t} \beta_l I(t = t_{[l]}) \tag{4.17}$$

based on L_m distinct lifetimes $m_{[1]} < \cdots < m_{[L_m]}$ and L_t distinct calendar times $t_{[1]} < \cdots < t_{[L_t]}$ such that there is at least one default event observed at the corresponding time. One may find the MLE of the coefficients $\boldsymbol{\alpha}, \boldsymbol{\beta}$ in (4.17) by maximizing (4.6) or (4.8). Rather than the crude optimization, we give an iterative *DtBreslow estimator* based on Breslow (1972), by treating (4.16) as the product of a baseline hazard function and a relative-risk multiplier.

Given the dual-time-to-default data (4.1) and any fixed $\lambda_g = \widehat{\lambda}_g$, the Breslow estimator of $\lambda_f(m)$ is given by

$$\widehat{\lambda}_f(m) \leftarrow \frac{\sum_{j=1}^J \mathsf{nevent}_j(m)}{\sum_{j=1}^J \widehat{\lambda}_g(V_j + m) \cdot \mathsf{nrisk}_j(m)}, \quad m = m_{[1]}, \ldots, m_{[L_m]} \tag{4.18}$$

and similarly given any fixed $\lambda_f = \widehat{\lambda}_f$,

$$\widehat{\lambda}_g(t) \leftarrow \frac{\sum_{j=1}^J \mathsf{nevent}_j(t - V_j)}{\sum_{j=1}^J \widehat{\lambda}_f(t - V_j) \cdot \mathsf{nrisk}_j(t - V_j)}, \quad t = t_{[1]}, \ldots, t_{[L_t]}. \tag{4.19}$$

Both marginal Breslow estimators are known to be the nonparametric MLE of the baseline hazards conditional on the relative-risk terms. To take into the identifiability constraint $\mathbb{E} \log_g(t) = 0$, we may update $\lambda_g(t)$ by

$$\lambda_g(t) \leftarrow \exp\Big\{ \log \widehat{\lambda}_g(t) - \mathsf{mean}\big(\log \widehat{\lambda}_g\big) \Big\}. \tag{4.20}$$

By iterating (4.18) to (4.20) until convergence, we obtain the DtBreslow estimator. It could be viewed as a natural extension of the one-way nonparametric estimators

(4.14) and (4.15). In (4.18), it is usually appropriate to set the initial exogenous multipliers $\widehat{\lambda}_g \equiv 1$, then the algorithm could converge in a few iterations.

We carry out a simulation study to assess the performance of DtBreslow estimator. Using the simulation data by (4.2), the two-way estimations of $\lambda_f(m)$ and $\lambda_g(t)$ are shown in Figure 4.2 (top). Also plotted are the one-way nonparametric estimation and the underlying truth. For this rectangular diagram of survival data, the one-way and DtBreslow estimations have comparable performance. However, if we use the non-rectangular Lexis diagram (e.g. either pre-2005 vintages or post-2005 vintages), the DtBreslow estimator would demonstrate the significant improvement over the one-way estimation; see Figure 4.2 (middle and bottom). It is demonstrated that the one-way nonparametric methods suffer the potential bias in both cases. By comparing (4.16) to (4.14) and (4.15), it is clear that the lifting power of DtBreslow estimator (e.g. in calibrating $\lambda_f(m)$) depends on not only the exogenous scaling factor $\widehat{\lambda}_g(t)$, but also the vintage-specific weights $\mathsf{nrisk}_j(m)$.

4.2.3 MEV Decomposition

Lastly, we consider a three-way nonparametric procedure that is consistent with the MEV (maturation-exogenous-vintage) decomposition framework we have developed for vintage data analysis in Chatper III. For each vintage with accounts $(V_j, U_{ji}, M_{ji}, \Delta_{ji})$ for $i = 1, \ldots, n_j$, assume the underlying hazards to be

$$\lambda_{ji}(m,t) = \lambda_f(m)\lambda_g(t)\lambda_h(V_j) = \exp\Big\{f(m) + g(t) + h(V_j)\Big\}. \tag{4.21}$$

For identifiability, let $\mathbb{E}g = \mathbb{E}h = 0$. Such MEV hazards model can be viewed as the extension of dual-time nonparametric model (4.16) with the vintage heterogeneity.

Due to the hyperplane nature of the vintage diagram such that $t \equiv V_j + m$ in (4.21), the linear trends of f, g, h are not identifiable; see Lemma 3.1. Among other

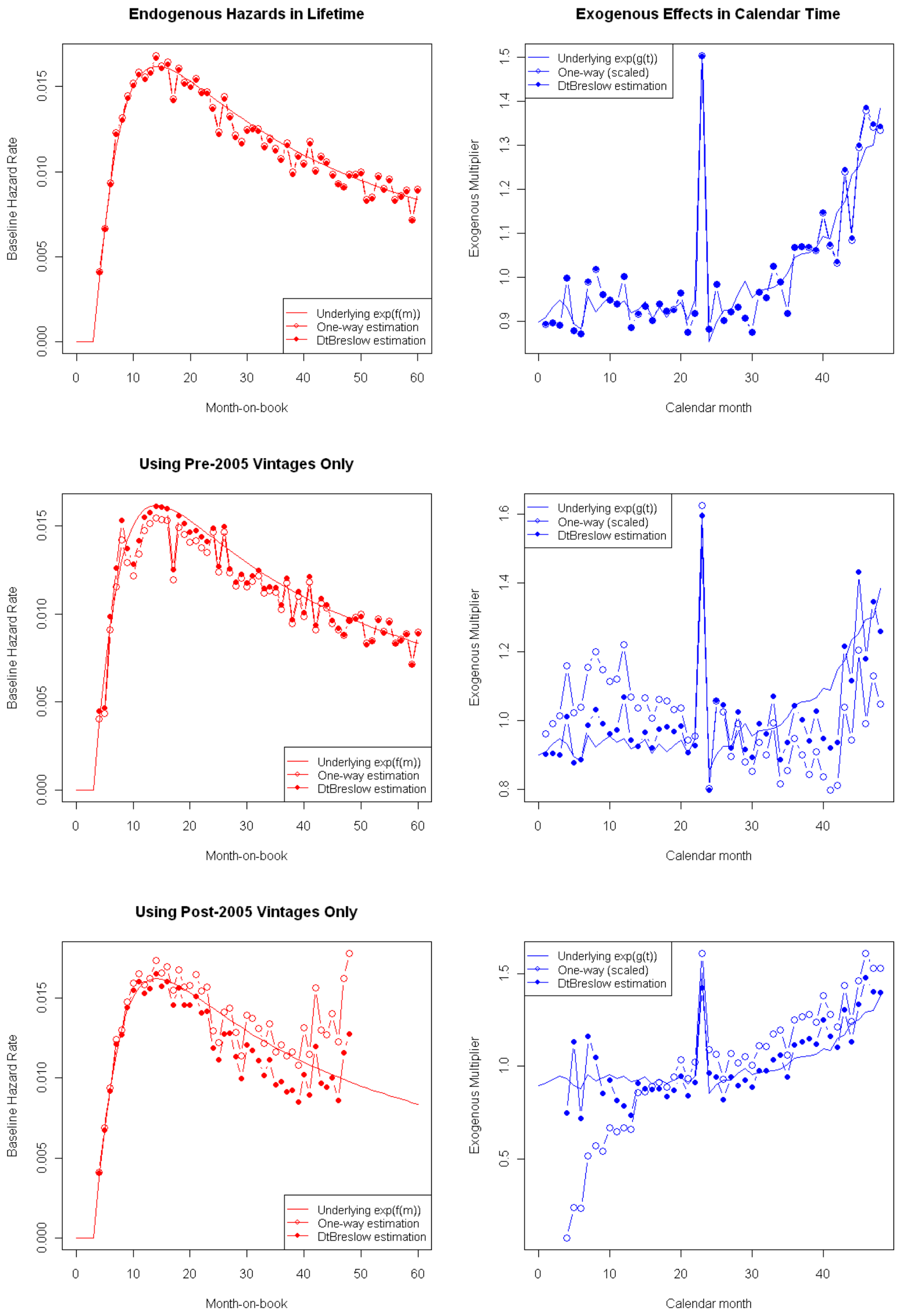


Figure 4.2: DtBreslow estimator vs one-way nonparametric estimator using the data of (top) both pre-2005 and post-2005 vintages; (middle) pre-2005 vintages only; (bottom) post-2005 vintages only.

approaches to break up the linear dependency, we consider

1. Binning in vintage, lifetime or calendar time for especially the marginal regions with sparse observations;
2. Removing the marginal trend in the vintage dimension $h(\cdot)$.

Note that the second strategy assumes the zero trend in vintage heterogeneity, which is an *ad hoc* approach when we have no prior knowledge about the trend of f, g, h. Later in Section 4.4 we discuss also a frailty type of vintage effect.

The nonparametric MLE for the MEV hazard model (4.21) follows the DtBreslow estimator discussed above. Specifically, we may estimate the three conditional marginal effects iteratively by:

a) **Maturation effect** for fixed $\widehat{\lambda}_g(t)$ and $\widehat{\lambda}_h(v)$

$$\widehat{\lambda}_f(m) \leftarrow \frac{\sum_{j=1}^{J} \mathsf{nevent}_j(m)}{\sum_{j=1}^{J} \widehat{\lambda}_h(V_j)\widehat{\lambda}_g(V_j+m)\cdot \mathsf{nrisk}_j(m)} \tag{4.22}$$

for $m = m_{[1]}, \ldots, m_{[L_m]}$.

b) **Exogenous effect** for fixed $\widehat{\lambda}_f(m)$ and $\widehat{\lambda}_h(v)$

$$\begin{cases} \widehat{\lambda}_g(t) \leftarrow \dfrac{\sum_{j=1}^{J} \mathsf{nevent}_j(t-V_j)}{\sum_{j=1}^{J} \widehat{\lambda}_h(V_j)\widehat{\lambda}_f(t-V_j)\cdot \mathsf{nrisk}_j(t-V_j)} \\ \widehat{\lambda}_g(\cdot) \leftarrow \exp\left\{\widehat{g} - \mathsf{mean}\big(\widehat{g}\big)\right\}, \quad \text{with } \widehat{g} = \log \widehat{\lambda}_g \end{cases} \tag{4.23}$$

for $t = t_{[1]}, \ldots, t_{[L_t]}$.

c) **Vintage effect** for fixed $\widehat{\lambda}_f(m)$ and $\widehat{\lambda}_g(t)$

$$\begin{cases} \widehat{\lambda}_h(V_j) \leftarrow \dfrac{\sum_{i=1}^{n_j} I(\Delta_{ji}=1)}{\sum_{l=1}^{L_m} \widehat{\lambda}_f(m_{[l]})\widehat{\lambda}_g(V_j+m_{[l]})\cdot \mathsf{nrisk}_j(m_{[l]})} \\ \widehat{\lambda}_h(\cdot) \leftarrow \exp\left\{\widehat{h} - \mathsf{mean}\big(\widehat{h}\big)\right\}, \quad \text{with } \widehat{h} = \log \widehat{\lambda}_h \end{cases} \tag{4.24}$$

for $j = 1, \ldots, J$ (upon necessary binning). In (4.24), one may also perform the trend-removal $\widehat{\lambda}_h(\cdot) \leftarrow \exp\{\widehat{h} - \mathsf{mean}\oplus\mathsf{trend}(\widehat{h})\}$ if zero-trend can be assumed for the vintage heterogeneity.

Setting the initial values $\widehat{\lambda}_g \equiv 1$ and $\widehat{\lambda}_h \equiv 1$, then iterating the above three steps until convergence would give us the nonparametric estimation of the three-way marginal effects. Consider again the simulation data with both pre-2005 and post-2005 vintages observed on the rectangular Lexis diagram Figure 4.1. Since there are limited sample sizes for very old and very new vintages, the vintages with $V_j \leq -36$ are merged and the vintages with $V_j \geq 40$ are cut off. Besides, given the prior knowledge that pre-2005 vintages have relatively flat heterogeneity, binning is performed in every half year. Then, running the above MEV iterative procedure, we obtain the estimation of maturation hazard $\lambda_f(m)$, exogenous hazard $\lambda_g(t)$ and the vintage heterogeneity $\lambda_h(v)$, which are plotted in Figure 4.3 (top). They are consistent with the underlying true hazard functions we have pre-specified in (4.3).

Furthermore, we have also tested the MEV nonparametric estimation for the dual-time data with either pre-205 vintages only (trapezoidal diagram) or the post-2005 vintages only (triangular diagram). The fitting results are also presented in Figure 4.3, which implies that our MEV decomposition is quite robust to the irregularity of the underlying Lexis diagram.

Remarks: before ending this section, we make several remarks about the nonparametric methods for dual-time survival analysis:

1. First of all, the dual-time-to-default data have essential difference from the usual bivariate lifetime data that are associated with either two separate failure modes per subject or a single mode for a pair of subjects. For this reason, it is questionable whether the nonparametric methods developed for bivariate survival are applicable to Lexis diagram; see Keiding (1990). Our developments

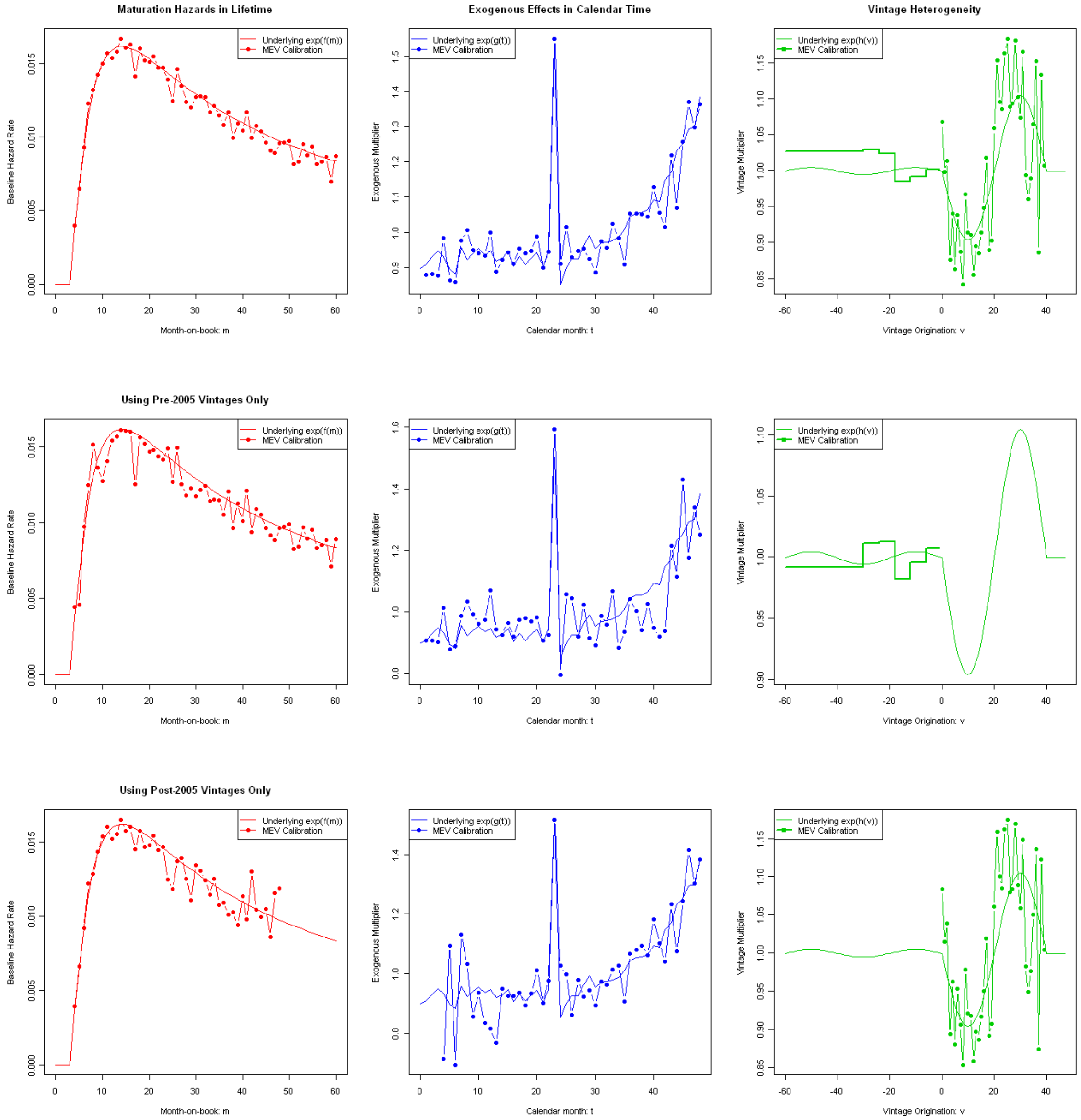


Figure 4.3: MEV modeling of empirical hazard rates, based on the dual-time-to-default data of (top) both pre-2005 and post-2005 vintages; (middle) pre-2005 vintages only; (bottom) post-2005 vintages only.

of DtBreslow and MEV nonparametric estimators for default intensities correspond to the age-period or age-period-cohort model in cohort or vintage data analysis of loss rates; see (3.4) in the previous chapter. For either default intensities or loss rates on Lexis diagram, one needs to be careful about the intrinsic identification problem, for which we have discussed the practical solutions in both of these chapters.

2. The nonparametric models (4.16) and (4.21) underlying the DtBreslow and MEV estimators are special cases of the dual-time Cox models in Section 4.4. In the finite-sample (or discrete) case, we have assumed in (4.17) the endogenous and exogenous hazard rates to be pointwise, or equivalently assumed the cumulative hazards to be step functions. In this case, statistical properties of DtBreslow estimates $\widehat{\lambda}_f(m)$ and $\widehat{\lambda}_g(t)$ can be studied by the standard likelihood theory for the *finite-dimensional* parameters $\boldsymbol{\alpha}$ or $\boldsymbol{\beta}$. However on the continuous domain, the large-sample properties of DtBreslow estimators are worthy of future study when the dimensions of both $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ increase to infinity at the essentially same speed.

3. Smoothing techniques can be applied to one-way, two-way or three-way nonparametric estimates, where the smoothness assumptions follow our discussion in the previous chapter; see (3.2). Consider e.g. the DtBreslow estimator with output of two-way empirical hazard rates $\widehat{\lambda}_f(m)$ and $\widehat{\lambda}_g(t)$, i.e. Nelson-Aalen increments on discrete time points; see (4.12). One may *post-process* them by kernel smoothing

$$\lambda_f(m) = \frac{1}{b_f}\sum_{l=1}^{L_m} K_f\Big(\frac{m-m_{[l]}}{b_f}\Big)\widehat{\lambda}_f(m_{[l]}), \quad \lambda_g(t) = \frac{1}{b_g}\sum_{l=1}^{L_t} K_g\Big(\frac{t-t_{[l]}}{b_g}\Big)\widehat{\lambda}_g(t_{[l]})$$

for m, t on the continuous domains, where $K_f(\cdot)$ and $K_g(\cdot)$ together with bandwidths b_f and b_g are the kernel functions that are bounded and vanish outside

$[-1,1]$; see Ramlau-Hansen (1983). Alternatively, one may use the penalized likelihood formulation via smoothing spline; see Gu (2002; §7). The challenging issues in these smoothing methods are known to be the selection of bandwidths or smoothing parameters, as well as the correction of boundary effects. More details can be referred to Wang (2005) and references therein.

4. An alternative way of dual-time hazard smoothing can be based on the vintage data analysis (VDA) developed in the previous chapter. For large-sample observations, one may first calculate the empirical hazard rates $\widehat{\lambda}_j(m)$ per vintage by (4.11), use them as the *pseudo-responses* of vintage data structure as in (3.1), then perform Gaussian process modeling discussed in Section 3.3. In this case, the variance of $\widehat{\lambda}_j(m)$ can be estimated by a tie-corrected formula

$$\widehat{\sigma}_j^2(m) = \frac{\big(\mathsf{nrisk}_j(m) - \mathsf{nevent}_j(m)\big) \cdot \mathsf{nevent}_j(m)}{\mathsf{nrisk}_j^3(m)}$$

and $\widehat{\lambda}_j(m)$ are uncorrelated in m; see e.g. Aalen, et al. (2008; p.85). Then, we may write $\widehat{\lambda}_j(m) = \lambda_j(m)(1 + \varepsilon_j(m))$ with $\mathbb{E}\varepsilon_j(m) = 0$ and

$$\mathbb{E}\varepsilon_j^2(m) \approx \frac{\widehat{\sigma}_j^2(m)}{\widehat{\lambda}_j^2(m)} = \frac{1}{\mathsf{nevent}_j(m)} - \frac{1}{\mathsf{nrisk}_j(m)} \to 0, \quad \text{as } n_j \to \infty$$

on the discrete time domain; by Taylor expansion, $\log \widehat{\lambda}_j(m) \approx \log \lambda_j(m) + \varepsilon_j(m)$, corresponding to the MEV model (3.15) with non-stationary noise.

5. An important message delivered by Figure 4.2 (middle and bottom) is that the nonparametric estimation of $\lambda_f(m)$ in conventional lifetime-only survival analysis might be misleading when there exists exogenous hazard $\lambda_g(t)$ in the calendar dimension. Both the DtBreslow and MEV nonparametric methods may correct the bias of lifetime-only estimation, as demonstrated in Figure 4.2 and Figure 4.3 with either pre-2005 only or post-2005 only vintages.

4.3 Structural Models

We have introduced in Chapter 1.2.1 the structural approach to credit risk modeling, which is popular in the sense that the default events can be triggered by a latent stochastic process when it hits the default boundary. For a quick review, one may refer to (1.6) for the latent distance-to-default process w.r.t. zero boundary, (1.9) for the notion of first-passage-time, and (1.13) for the concrete expression of hazard rate based on the inverse Gaussian lifetime distribution. In this section, we consider the extension of structural approach to dual-time-to-default analysis.

4.3.1 First-passage-time Parameterization

Consider the Lexis diagram of dual-time default events for vintages $j = 1, \ldots, J$. Let $X_j(m)$ be the *endogenous* distant-to-defalut (DD) process in lifetime and assume it follows a drifted Wiener process

$$X_j(m) = c_j + b_j m + W_j(m), \quad m \geq 0 \tag{4.25}$$

where $X_j(0) \equiv c_j \in \mathbb{R}_+$ denotes the parameter of *initial distance-to-default*, $b_j \in \mathbb{R}$ denotes the parameter of *trend of credit deterioration* and $W_j(t)$ is the Wiener process. Without getting complicated, let us assume that W_j and $W_{j'}$ are uncorrelated whenever $j \neq j'$. Recall the hazard rate (or, default intensity) of the first-passage-time of $X_j(m)$ crossing zero:

$$\lambda_j(m; c_j, b_j) = \frac{\dfrac{c_j}{\sqrt{2\pi m^3}} \exp\left\{-\dfrac{(c_j + b_j m)^2}{2m}\right\}}{\Phi\left(\dfrac{c_j + b_j m}{\sqrt{m}}\right) - e^{-2b_j c_j}\Phi\left(\dfrac{-c_j + b_j m}{\sqrt{m}}\right)}, \tag{4.26}$$

whose numerator and denominator correspond to the density and survival functions; see (1.10)–(1.13). By (4.6), the MLE of (c_j, b_j) can be obtained by maximizing the log-likelihood of the j-th vintage observations

$$\ell_j = \sum_{i=1}^{n_j} \Delta_{ji} \log \lambda_j(M_{ji}; c_j, b_j) + \log S_j(M_{ji}; c_j, b_j) - \log S_j(U_{ji}; c_j, b_j), \qquad (4.27)$$

where the maximization can be carried out by nonlinear optimization. We supplement the gradients of the log-likelihood function at the end of the chapter.

As their literal names indicate, both the initial distance-to-default c_j and the trend of credit deterioration b_j have interesting structural interpretations. With reference to Aalen and Gjessing (2001) or Aalen, et al. (2008; §10):

1. The c-parameter represents the initial value of DD process. When c becomes smaller, a default event tends to happen sooner, hence moving the concentration of the defaults towards left. In theory, the hazard function (4.26) for $m \in (0, \infty)$ is always first-increasing-then-decreasing, with maximum at m_* such that the derivative $\lambda'(m_*) = 0$. In practice, let us focus on the bounded interval $m \in [\epsilon, m_{\max}]$ for some $\epsilon > 0$. Then, when c gradually approaches zero, the hazard rate would be shaped from (a) monotonic increasing to (b) first-increasing-then-decreasing and finally to (c) monotonic decreasing. See Figure 1.2 for the illustration with fixed b and varying c values.

2. The b-parameter represents the trend of DD process, and it determines the level of stabilized hazard rate as $m \to \infty$. In Aalen, et al. (2008; §10), let $\phi_m(x)$ be the conditional density of $X(m)$ given that $\min_{0<s\le m} X(s) > 0$ and it is said to be *quasi-stationary* if $\phi_m(x)$ is m-invariant. In our case, the quasi-stationary distribution is $\phi(x) = b^2 x e^{-bx}$ for $x \ge 0$ and the limiting hazard rate is given by

$$\lim_{m\to\infty} \lambda(m) = \frac{1}{2} \frac{\partial \phi(x)}{\partial x}\Big|_{x=0} = \frac{b^2}{2},$$

which does not depend on the c-parameter. In Figure 1.2 with $b = -0.02$, all the hazard rate curves will converge to the constant 0.0002 after a long run.

The above first-passage-time parameterization is only on the endogenous or maturation hazard in lifetime. Next, we extend the first-passage-time structure taking into account the exogenous effect in calendar time. To achieve that, we model the dual-time DD process for each vintage V_j by a subordinated process

$$X_j^*(m) = X_j(\Psi_j(m)), \quad \text{with } \Psi_j(m) = \int_0^m \psi(V_j + s) ds \tag{4.28}$$

where $X_j(m)$ is the endogenous DD process (4.25) and $\Psi_j(m)$ is an increasing time-transformation function such that all vintages share the same integrand $\psi(t) > 0$ in calendar time. We set the constraint $\mathbb{E} \log \psi = 0$ on the exogenous effect, in the sense of letting the 'average' DD process be captured endogenously. Such time-transformation approach can be viewed as a natural generalization of the log-location-scale transformation (1.16) with correspondence $\sigma_i = 1$ and $\psi(\cdot) = e^{-\mu_i}$. It has been used to extend the accelerated failure time (AFT) models with time-varying covariates; see e.g. Lawless (2003; §6.4.3).

The exogenous function $\psi(t)$ rescales the original lifetime increment Δm to be $\psi(V_j + m)\Delta m$, and therefore transforms the diffusion $dX_j(m) = b_j dm + dW_j(m)$ to

$$dX_j^*(m) = b_j \psi(V_j + m) dm + \sqrt{\psi(V_j + m)} dW_j(m) \tag{4.29}$$

w.r.t. the zero default boundary. Alternatively, one may rewrite it as

$$dX_j^*(m) = \sqrt{\psi(V_j + m)} dX_j(m) - b_j \left(\sqrt{\psi(V_j + m)} - \psi(V_j + m) \right) dm. \tag{4.30}$$

Let us introduce a new process $\widetilde{X}_j(m)$ such that $d\widetilde{X}_j(m) = \sqrt{\psi(V_j + m)} dX_j(m)$, and introduce a time-varying default boundary

$$\widetilde{D}_j(m) = b_j \int_0^m \left(\sqrt{\psi(V_j + s)} - \psi(V_j + s) \right) ds.$$

Then, by (4.30), it is clear that the time-to-default $\tau_j^* = \inf\{m \geq 0 : X_j^*(m) \leq 0\} = \inf\{m \geq 0 : \widetilde{X}(m) \leq \widetilde{D}_j(m)\}$.

By time-transformation (4.28), it is straightforward to obtain the survival function of the first-passage-time τ_j^* by

$$\begin{aligned} S_j^*(m) &= \mathbb{P}\big(X_j^*(s) > 0 : s \in [0, m]\big) = \mathbb{P}\big(X_j(s) > 0 : s \in [0, \Psi_j(m)]\big) \\ &= S_j(\Psi_j(m)) = S_j\left(\int_0^m \psi(V_j + s)ds\right) \end{aligned} \tag{4.31}$$

where the benchmark survival $S_j(\cdot)$ is given by the denominator of (4.26). Then, the hazard rate of τ_j^* is obtained by

$$\begin{aligned} \lambda_j^*(m) &= \frac{-\log S_j^*(m)}{dm} = \frac{-\log S_j(\Psi_j(m))}{d\Psi_j(m)} \frac{d\Psi_j(m)}{dm} = \lambda_j(\Psi_j(m))\Psi_j'(m) \\ &= \lambda_j\left(\int_0^m \psi(V_j + s)ds\right)\psi(V_j + m) \end{aligned} \tag{4.32}$$

with $\lambda_j(\cdot)$ given in (4.26). From (4.32), the exogenous function $\psi(t)$ not only maps the benchmark hazard rate to the rescaled time $\Psi_j(m)$, but also scales it by $\psi(t)$. This is the same as the AFT regression model (1.21).

Now consider the parameter estimation of both vintage-specific structural parameters (c_j, b_j) and the exogenous time-scaling function $\psi(t)$ from the dual-time-to-default data (4.1). By (4.6), the joint log-likelihood is given by $\ell = \sum_{j=1}^J \ell_j$, with ℓ_j given by (4.27) based on $\lambda_j^*(m; c_j, b_j)$ and $S_j^*(m; c_j, b_j)$. The complication lies in the arbitrary formation of $\psi(t)$ involved in the time-transformation function. Nonparametric techniques can be used if extra properties like the smoothness can be imposed onto the function $\psi(t)$. In our case the exogenous time-scaling effects are often dynamic and volatile, for which we recommend a piecewise estimation procedure based on the discretized Lexis diagram.

1. Discretize the Lexis diagram with $\Delta m = \Delta t$ such that $t = t_{\min} + l\Delta t$ for $l =$

$0, 1, \ldots, L$, where $t_{\min}$ is set to be the left boundary and $t_{\min} + L\Delta t$ corresponds to the right boundary. Specify $\psi(t)$ to be piecewise constant

$$\psi(t) = \psi_l, \quad \text{for } t - t_{\min} \in \big((l-1)\Delta t,\ l\Delta t\big],\ l = 1, 2, \ldots, L$$

and $\psi(t) \equiv 1$ for $t \leq t_{\min}$. Then, the time-transformation for the vintage with origination time V_j can be expressed as

$$\widetilde{\Psi}_j(m) = \sum_{l=(V_j - t_{\min})/\Delta t}^{(V_j+m-t_{\min})/\Delta t} \psi(t_{\min} + l\Delta t)\Delta t, \quad \text{for } V_j + m \geq t_{\min}$$

which reduces to $t_{\min} - V_j + \sum_{l=1}^{(V_j+m-t_{\min})/\Delta t} \psi_l \Delta t$ in the case of left truncation.

2. For each of $l = 1, 2, \ldots$, get the corresponding likelihood contribution:

$$\ell_{[l]} = \sum_{(j,i)\in\mathcal{G}_{[l]}} \Delta_{ji}\Big[\log\lambda_j(\widetilde{\Psi}_j(M_{ji})) + \log\psi_l\Big] + \log S_j(\widetilde{\Psi}_j(M_{ji})) - \log S_j(\widetilde{\Psi}_j(U_{ji}))$$

where $\mathcal{G}_{[l]} = \{(j,i) : V_j + M_{ji} = t_{\min} + l\Delta t\}$, and $\widetilde{\Psi}_j(U_{ji}) = \max\{0, t_{\min} - V_j\}$ if $t_{\min}$ is the systematic left truncation in calendar time.

3. Run nonlinear optimization for estimating (c_i, b_i) and ψ_l iteratively:

   (a) For fixed ψ and each fixed vintage $j = 1, \ldots, J$, obtain the MLE $(\widehat{c}_j, \widehat{b}_j)$ by maximizing (4.27) with M_{ji} replaced by $\widetilde{\Psi}_j(M_{ji})$.

   (b) For all fixed (c_j, b_j), obtain the MLE of $(\psi_1, \ldots, \psi_L)$ by maximizing $\sum_{l=1}^{L} \ell_{[l]}$ subject to the constraint $\sum_{l=1}^{L} \log\psi_l = 0$.

In Step (3.b), one may perform further binning for $(\psi_1, \ldots, \psi_L)$ for the purpose of reducing the dimension of parameter space in constrained optimization. One may also apply the multi-resolution wavelet bases to represent $\{\psi_1, \ldots, \psi_L)$ in a time-frequency perspective. Such ideas will be investigated in our future work.

4.3.2 Incorporation of Covariate Effects

It is straightforward to extend the above structural parameterization to incorporate the subject-specific covariates. For each account i with origination V_j, let $\tilde{\mathbf{z}}_{ji}$ be the static covariates observed upon origination (including the intercept and otherwise centered predictors) and $\mathbf{z}_{ji}(m)$ for $m \in (U_{ji}, M_{ji}]$ be the dynamic covariates observed since left truncation. Let us specify

$$\begin{aligned} c_{ji} &= \exp\{\eta^T \tilde{\mathbf{z}}_{ji}\} && (4.33) \\ \Psi_{ji}(m) &= U_{ji} - V_j + \int_{U_{ji}}^{m} \exp\{\theta^T \mathbf{z}_{ji}(s)\} ds, \quad m \geq U_{ji} && (4.34) \end{aligned}$$

subject to the unknown parameters (η, θ), and suppose the subject-specific DD process

$$\begin{cases} X^*_{ji}(m) = X_{ji}(\Psi_{ji}(m)), \\ X_{ji}(m) = c_{ji} + b_j m + W_{ji}(m) \end{cases} \qquad (4.35)$$

with an independent Wiener process $W_{ji}(m)$, the initial value c_{ji} and the vintage-specific trend of credit deterioration b_j. Then, the default occurs at $\tau^*_{ji} = \inf\{m > 0 : X^*_{ji}(m) \leq 0\}$. Note that when the covariates in (4.34) are not time-varying, we may specify $\Psi_{ji}(m) = m \exp\{\theta^T \mathbf{z}_{ji}\}$ and obtain the AFT regression model $\log \tau^*_{ji} = \theta^T \mathbf{z}_{ji} + \log \tau_{ji}$ with inverse Gaussian distributed τ_{ji}. In the presence of calendar-time state variables $\mathbf{x}(t)$ such that $\mathbf{z}_{ji}(m) = \mathbf{x}(V_j + m)$ for all (j, i), the time-transformation (4.34) becomes a parametric version of (4.28).

By the same arguments as in (4.31) and (4.32), we obtain the hazard rate $\lambda^*_{ji}(m) = \lambda_{ji}(\Psi_{ji}(m))\Psi'_{ji}(m)$, or

$$\lambda^*_{ji}(m) = \frac{\dfrac{c_{ji}}{\sqrt{2\pi\Psi^3_{ji}(m)}} \exp\left\{-\dfrac{\left(c_{ji} + b_j\Psi_{ji}(m)\right)^2}{2\Psi_{ji}(m)}\right\} \Psi'_{ji}(m)}{\Phi\left(\dfrac{c_{ji} + b_j\Psi_{ji}(m)}{\sqrt{\Psi_{ji}(m)}}\right) - e^{-2b_j c_{ji}}\Phi\left(\dfrac{-c_{ji} + b_j\Psi_{ji}(m)}{\sqrt{\Psi_{ji}(m)}}\right)} \qquad (4.36)$$

and the survival function $S^*_{ji}(m) = S_{ji}(\Psi_{ji}(m))$ given by the denominator in (4.36).

The parameters of interest include (η, θ) and b_j for $j = 1, \ldots, J$, and they can be estimated by MLE. Given the dual-time-to-default data (4.1), the log-likelihood function follows the generic form (4.6). Standard nonlinear optimization programs can be employed, e.g. the "optim" algorithm of R (stats-library) and the "fminunc" algorithm of MATLAB (Optimization Toolbox).

Remarks:

1. There is no unique way to incorporate the *static* covariates in the first-passage-time parameterization. For example, we considered including them by the initial distance-to-default parameter (4.33), while they may also be incorporated into other structural parameters; see Lee and Whitmore (2006). However, there has been limited discussion on incorporation of *dynamic* covariates in the first-passage-time approach, for which we make an attempt to use the time-transformation technique (4.28) under nonparametric setting and (4.34) under parameter setting.

2. The endogenous hazard rate (4.26) is based on the fixed effects of initial distance-to-default and trend of deterioration, while these structural parameters can be random effects due to the incomplete information available (Giesecke, 2006). One may refer to Aalen, et al. (2008; §10) about the random-effect extension. For example, when the trend parameter $b \sim N(\mu, \sigma^2)$ and the initial value c is fixed, the endogenous hazard rate is given by $p(m)/S(m)$ with

$$\begin{aligned} p(m) &= \frac{c}{\sqrt{2\pi(m^3 + \sigma^2 m^4)}} \exp\left\{-\frac{(c+\mu t)^2}{2(t+\sigma^2 t^2)}\right\} \\ S(m) &= \Phi\left(\frac{c+\mu m}{\sqrt{m+\sigma^2 m^2}}\right) - e^{-2\mu c + 2\sigma^2 c^2}\Phi\left(\frac{-c+\mu m - 2\sigma^2 cm}{\sqrt{m+\sigma^2 m^2}}\right). \end{aligned}$$

3. In the first-passage-time parameterization with inverse Gaussian lifetime distri-

bution, the log-likelihood functions (4.27) and (4.6) with the plugged-in (4.26) or (4.36) are rather messy, especially when we attempt to evaluate the derivatives of the log-likelihood function in order to run gradient search; see the supplementary material. Without gradient provided, one may rely on the crude nonlinear optimization algorithms, and the parameter estimation follows the standard MLE procedure. Our discussion in this section mainly illustrates the feasibility of dual-time structural approach.

4.4 Dual-time Cox Regression

The semiparametric Cox models are popular for their effectiveness of estimating the covariate effects. Suppose that we are given the dual-time-to-default data (4.1) together with the covariates $\mathbf{z}_{ji}(m)$ observed for $m \in (U_{ji}, M_{ji}]$, which may include the static covariates $\tilde{\mathbf{z}}_{ji}(m) \equiv \tilde{\mathbf{z}}_{ji}$. Using the traditional CoxPH modeling approach, one may either adopt an arbitrary baseline in lifetime

$$\lambda_{ji}(m) = \lambda_0(m) \exp\{\boldsymbol{\theta}^T \mathbf{z}_{ji}(m)\}, \tag{4.37}$$

or adopt the stratified Cox model with arbitrary vintage-specific baselines

$$\lambda_{ji}(m) = \lambda_{j0}(m) \exp\{\boldsymbol{\theta}^T \mathbf{z}_{ji}(m)\} \tag{4.38}$$

for $i = 1, \ldots, n_j$ and $j = 1, \ldots, L$. However, these two Cox models may not be suitable for (4.1) observed on the Lexis diagram in the following sense:

1. (4.37) is too stringent to capture the variation of vitnage-specific baselines;
2. (4.38) is too relaxed to capture the interrelation of vintage-specific baselines.

For example, we have simulated a Lexis diagram in Figure 4.1 whose vintage-specific hazards are interrelated not only in lifetime m but also in calendar time t.

4.4.1 Dual-time Cox Models

The dual-time Cox models considered in this section take the following multiplicative hazards form in general,

$$\lambda_{ji}(m,t) = \lambda_0(m,t)\exp\{\boldsymbol{\theta}^T\mathbf{z}_{ji}(m)\} \tag{4.39}$$

where $\lambda_0(m,t)$ is a bivariate base hazard function on Lexis diagram. Upon the specification of $\lambda_0(m,t)$, the traditional models (4.37) and (4.38) can be included as special cases. In order to balance between (4.37) and (4.38), we assume the MEV decomposition framework of the base hazard $\lambda_0(m,t) = \lambda_f(m)\lambda_g(t)\lambda_h(t-v)$, i.e. the nonparametric model (4.21). Thus, we restrict ourselves to the dual-time Cox models with MEV baselines and relative-risk multipliers

$$\begin{aligned}\lambda_{ji}(m,t) &= \lambda_f(m)\lambda_g(t)\lambda_h(V_j)\exp\{\boldsymbol{\theta}^T\mathbf{z}_{ji}(m)\} \\ &= \exp\{f(m)+g(t)+h(V_j)+\boldsymbol{\theta}^T\mathbf{z}_{ji}(m)\}\end{aligned} \tag{4.40}$$

where $\lambda_f(m) = \exp\{f(m)\}$ represents the arbitrary maturation/endogenous baseline, $\lambda_g(t) = \exp\{g(t)\}$ and $\lambda_h(v) = \exp\{h(v)\}$ represent the exogenous multiplier and the vintage heterogeneity subject to $\mathbb{E}g = \mathbb{E}h = 0$. Some necessary binning or trend-removal need to be imposed on g, h to ensure the model estimability; see Lemma 3.1.

Given the finite-sample observations (4.1) on Lexis diagram, one may always find L_m distinct lifetime points $m_{[1]} < \cdots < m_{[L_m]}$ and L distinct calendar time points $t_{[1]} < \cdots < t_{[L_t]}$ such that there exist at least one default event observed at the corresponding marginal time. Following the least-information assumption of Breslow (1972), let us treat the endogenous baseline hazard $\lambda_f(m)$ and the exogenous baseline hazard $\lambda_g(t)$ as pointwise functions of the form (4.17) subject to $\sum_{l=1}^{L_t}\beta_l = 0$. For the discrete set of vintage originations, the vintage heterogeneity effects can be regarded as the categorical covariates effects, essentially. We may either assume the pointwise function of the form $h(v) = \sum_{j=1}^{J} \gamma_j I(v = V_j)$ subject to constraints on both mean and trend (i.e. the *ad hoc* approach), or assume the piecewise-constant functional form upon appropriate binning in vintage originations

$$h(V_j) = \gamma_\kappa, \quad \text{if } j \in \mathcal{G}_\kappa, \quad \kappa = 1, \ldots, K \tag{4.41}$$

where the vintage buckets $\mathcal{G}_\kappa = \{j : V_j \in (\nu_{\kappa-1}, \nu_\kappa]\}$ are pre-specified based on $V_j^- \equiv \nu_0 < \cdots < \nu_K \equiv V_J$ (where $K < J$). Then, the semiparametric model (4.40) is *parametrized* with dummy variables to be

$$\begin{aligned}\lambda_{ji}(m, t) = \exp\Bigg\{ & \sum_{l=1}^{L_m} \alpha_l I(m = m_{[l]}) + \sum_{l=2}^{L_t} \beta_l I(t = t_{[l]}) \\ & + \sum_{\kappa=2}^{K} \gamma_\kappa I(j \in \mathcal{G}_\kappa) + \boldsymbol{\theta}^T \mathbf{z}_{ji}(m) \Bigg\}\end{aligned} \tag{4.42}$$

where $\beta_1 \equiv 0$ and $\gamma_1 \equiv 0$ are excluded from the model due to the identifiability constraints. Plugging it into the joint likelihood (4.6), one may then find the MLE of $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\theta})$. (Note that it is rather straightforward to convert the resulting parameter estimates $\widehat{\boldsymbol{\alpha}}$, $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\gamma}}$ to satisfy the zero sum constraints for $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, so is it for (4.45) to be studied.)

As the sample size tends to infinity, both endogenous and exogenous baseline hazards $\lambda_f(m)$ and $\lambda_g(t)$ defined on the continuous dual-time scales (m, t) tend to be infinite-dimensional, hence making intractable the above pointwise parameterization. The semiparametric estimation problem in this case is challenging, as the dual-time Cox models (4.40) involves two nonparametric components that need to be profiled out for the purpose of estimating the parametric effects. In what follows, we take an "intermediate approach" to semiparametric estimation via exogenous timescale discretization, to which the theory of partial likelihood applies.

4.4.2 Partial Likelihood Estimation

Consider an intermediate reformulation of the dual-time Cox models (4.40) by exogenous timescale discretization:

$$t_l = t_{\min} + l\Delta t, \quad \text{for } l = 0, 1, \ldots, L \tag{4.43}$$

for some time increment Δt bounded away from zero, where $t_{\min}$ corresponds to the left boundary and $t_{\min} + L\Delta t$ corresponds to the the right boundary of the calendar time under consideration. (If the Lexis diagram is discrete itself, Δt is automatically defined. The irregular time spacing is feasible, too.) On each sub-interval, assume the exogenous baseline $\lambda_g(t)$ is defined on the L cutting points (and zero otherwise)

$$g(t) = \sum_{l=1}^{L} \beta_l I(t = t_{\min} + l\Delta t), \tag{4.44}$$

subject to $\sum_{l=1}^{L} \beta_l = 0$. Then, consider the one-way CoxPH reformulation of (4.40):

$$\begin{aligned} \lambda_{ji}(m) = \lambda_f(m) \exp\Bigg\{ & \sum_{l=2}^{L} \beta_l I(V_j + m \in (t_{l-1}, t_l]) \\ & + \sum_{\kappa=2}^{K} \gamma_\kappa I(j \in \mathcal{G}_\kappa) + \boldsymbol{\theta}^T \mathbf{z}_{ji}(m) \Bigg\} \end{aligned} \tag{4.45}$$

where $\lambda_f(m)$ is an arbitrary baseline hazard, $g(t)$ for any $t \in (t_{l-1}, t_l]$ in (4.40) is shifted to $g(t_l)$ in (4.44), and the vintage effect is bucketed by (4.41). It is easy to see that if the underlying Lexis diagram is discrete, the reformulated CoxPH model (4.45) is equivalent to the fully parametrized version (4.42).

We consider the partial likelihood estimation of the parameters $(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\theta}) \equiv \boldsymbol{\xi}$. For convenience of discussion, write for each (j, i, m) the $(L-1)$-vector of $\{I(V_j + m \in (t_{l-1}, t_l])\}_{l=2}^{L}$, the $(K-1)$-vector of $\{I(V_j \in (\nu_{\kappa-1}, \nu_\kappa])\}_{\kappa=2}^{K}$, and $\mathbf{z}_{ji}(m)$ as a long vector $\mathbf{x}_{ji}(m)$. Then, by Cox (1972, 1975), the parameter $\boldsymbol{\xi}$ can be estimated by maximizing the partial likelihood

$$PL(\boldsymbol{\xi}) = \prod_{j=1}^{J}\prod_{i=1}^{n_j}\left[\frac{\exp\{\boldsymbol{\xi}^T\mathbf{x}_{ji}(M_{ji})\}}{\sum_{(j',i')\in\mathcal{R}_{ji}}\exp\left\{\boldsymbol{\xi}^T\mathbf{x}_{j'i'}(M_{ji})\right\}}\right]^{\Delta_{ji}} \tag{4.46}$$

where $\mathcal{R}_{ji} = \{(j',i') : U_{j'i'} < M_{ji} \leq M_{j'i'}\}$ is the at-risk set at each observed M_{ji}. Given the maximum partial likelihood estimator $\widehat{\boldsymbol{\xi}}$, the endogenous cumulative baseline hazard $\Lambda_f(m) = \int_0^m \lambda_f(s)ds$ can be estimated by the Breslow estimator

$$\widehat{\Lambda}_f(m) = \sum_{j=1}^{J}\sum_{i=1}^{n_j}\frac{I(m \geq M_{ji})\Delta_{ji}}{\sum_{(j',i')\in\mathcal{R}_{ji}}\exp\left\{\widehat{\boldsymbol{\xi}}^T\mathbf{x}_{j'i'}(M_{ji})\right\}} \tag{4.47}$$

which is a step function with increments at $m_{[1]} < \cdots < m_{[L_m]}$. It is well-known in the CoxPH context that both estimators $\widehat{\boldsymbol{\xi}}$ and $\widehat{\Lambda}_f(\cdot)$ are consistent and asymptotically normal; see Andersen and Gill (1982) based on counting process martingale theory. Specifically, $\sqrt{N}(\widehat{\boldsymbol{\xi}} - \boldsymbol{\xi})$ converges to a zero-mean multivariate normal distribution with covariance matrix that can be consistently estimated by $\{\mathcal{I}(\widehat{\boldsymbol{\xi}})/N\}^{-1}$, where $\mathcal{I}(\boldsymbol{\xi}) = -\partial^2\log PL(\boldsymbol{\xi})/\partial\boldsymbol{\xi}\boldsymbol{\xi}^T$. For the nonparametric baseline, $\sqrt{N}\big(\widehat{\Lambda}_f(m) - \Lambda_f(m)\big)$ converges to a zero-mean Gaussian process; see also Andersen, et al. (1993).

The DtBreslow and MEV estimators we have discussed in Section 4.2 can be studied by the theory of partial likelihood above, since the underlying nonparametric models are special cases of (4.45) with $\boldsymbol{\theta} \equiv 0$. Without the account-level covariates, the hazard rates $\lambda_{ji}(m)$ roll up to $\lambda_j(m)$, under which the partial likelihood function (4.46) can be expressed through the counting notations introduced in (4.9),

$$PL(g,h) = \prod_{j=1}^{J}\prod_{m}\left[\frac{\exp\{g(V_j+m)+h(V_j)\}}{\sum_{j'=1}^{J}\mathsf{nrisk}_{j'}(m)\exp\{g(V_{j'}+m)+h(V_{j'})\}}\right]^{\mathsf{nevent}_j(m)} \tag{4.48}$$

where $g(t)$ is parameterized by (4.44) and $h(v)$ is by (4.41). Let $\widehat{\lambda}_g(t) = \exp\{\widehat{g}(t)\}$ and $\widehat{\lambda}_h(v) = \exp\{\widehat{h}(v)\}$ upon maximization of $PL(g,h)$ and centering in the sense of

$\mathbb{E}g = \mathbb{E}h = 0$. Then, by (4.47), we have the Breslow estimator of the cumulative version of endogenous baseline hazards

$$\widehat{\Lambda}_f(m) = \int_0^m \frac{\sum_{j=1}^J \mathsf{nevent}_j(s)ds}{\sum_{j=1}^J \mathsf{nrisk}_j(s)\widehat{\lambda}_g(V_j+s)\widehat{\lambda}_h(V_j)}. \tag{4.49}$$

It is clear that the increments of $\widehat{\Lambda}_f(m)$ correspond to the maturation effect (4.22) of the MEV estimator. They also reduce to (4.18) of DtBreslow estimator in the absence of vintage heterogeneity.

The implementation of dual-time Cox modeling based on (4.45) can be carried out by maximizing the partial likelihood (4.46) to find $\widehat{\boldsymbol{\theta}}$ and using (4.47) to estimate the baseline hazard. Standard softwares with survival package or toolbox can be used to perform the parameter estimation. In the case when the coxph procedure in these survival packages (e.g. R:survival) requires the time-indepenent covariates, we may convert (4.1) to an enlarged counting process format with piecewise-constant approximation of time-varying covariates $\mathbf{x}_{ji}(m)$; see the supplementary material in the last section of the chapter.

For example, we have tested the partial likelihood estimation using the coxph of R:survival package and the simulation data in Figure 4.1, where the raw data (4.1) are converted to the counting process format upon the same binning preprocessing as in Section 4.2.3. The numerical results of $\widehat{\lambda}_g, \widehat{\lambda}_h$ by maximizing (4.48) and $\widehat{\lambda}_f$ by (4.49) are nearly identical to the plots shown in Figure 4.3.

4.4.3 Frailty-type Vintage Effects

Recall the frailty models introduced in Section 1.3 about their *double* roles in charactering both the odd effects of unobserved heterogeneity and the dependence of correlated defaults. They provide an alternative approach to model the vintage effects as the random block effects. Let Z_κ for $\kappa = 1, \ldots, K$ be a random sample from some frailty distribution, where Z_κ represents the shared frailty level of the vintage bucket $\mathcal{G}_\kappa$ according to (4.41). Consider the dual-time Cox frailty models

$$\lambda_{ji}(m, t|Z_\kappa) = \lambda_f(m)\lambda_g(t)Z_\kappa \exp\{\boldsymbol{\theta}^T \mathbf{z}_{ji}(m)\}, \quad \text{if } j \in \mathcal{G}_\kappa \tag{4.50}$$

for $i = 1, \ldots, n_j$ and $j \in \mathcal{G}_\kappa$ for $\kappa = 1, \ldots, K$. We also write $\lambda_{ji}(m, t|Z_\kappa)$ as $\lambda_{ji}(m|Z_\kappa)$ with $t \equiv V_j + m$. It is assumed that given the frailty Z_κ, the observations for all vintages in $\mathcal{G}_\kappa$ are independent. In this frailty model setting, the between-bucket heterogeneity is captured by realizations of the random variate Z, while the within-bucket dependence is captured by the same realization Z_κ.

Similar to the intermediate approach taken by the partial likelihood (4.46) under exogenously discretized dual-time model (4.45), let us write $\tilde{\boldsymbol{\xi}} = (\boldsymbol{\beta}, \boldsymbol{\theta})$ and let $\tilde{\mathbf{x}}_{ji}(m)$ join the vectors of $\{I(V_j + m \in (t_{l-1}, t_l])\}_{l=2}^L$ and $\mathbf{z}_{ji}(m)$. Then, we may write (4.50) as $\lambda_{ji}(m|Z_\kappa) = \lambda_f(m)Z_\kappa \exp\left\{\tilde{\boldsymbol{\xi}}^T \tilde{\mathbf{x}}_{ji}(m)\right\}$. The log-likelihood based on the dual-time-to-default data (4.1) is given by

$$\ell = \sum_{j=1}^{J}\sum_{i=1}^{n_j} \Delta_{ji}\left(\log \lambda_f(M_{ji}) + \tilde{\boldsymbol{\xi}}^T \tilde{\mathbf{x}}_{ji}(m)\right) + \sum_{\kappa=1}^{K} \log\left\{(-1)^{d_\kappa} \mathscr{L}^{(d_\kappa)}(A_\kappa)\right\}. \tag{4.51}$$

where $\mathscr{L}(x) = \mathbb{E}_Z[\exp\{-xZ\}$ is the Laplace transform of Z, $\mathscr{L}^{(d)}(x)$ is its d-th derivative, $d_\kappa = \sum_{j\in\mathcal{G}_\kappa}\sum_{i=1}^{n_j} \Delta_{ji}$ and $A_\kappa = \sum_{j\in\mathcal{G}_\kappa}\sum_{i=1}^{n_j} \int_{U_{ji}}^{M_{ji}} \lambda_f(s) \exp\left\{\tilde{\boldsymbol{\xi}}^T \tilde{\mathbf{x}}_{ji}(s)\right\} ds$; see the supplementary material at the end of the chapter.

The gamma distribution $\mathsf{Gamma}(\delta^{-1}, \delta)$ with mean 1 and variance δ is the most frequently used frailty due to its simplicity. It has the Laplace transform $\mathscr{L}(x) = (1+\delta x)^{-1/\delta}$ whose d-th derivative is given by $\mathscr{L}^{(d)}(x) = (-\delta)^d(1+\delta x)^{-1/\delta-d}\prod_{i=1}^d(1/\delta + i - 1)$. For a fixed δ, often used is the EM (expectation-maximization) algorithm for estimating $(\lambda_f(\cdot), \widetilde{\boldsymbol{\xi}})$; see Nielsen, et al. (1992) or the supplementary material.

Alternatively, the gamma frailty models can be estimated by the penalized likelihood approach, by treating the second term of (4.51) as a kind of regularization.

Therneau and Grambsch (2000; §9.6) justified that for each fixed δ, the EM algorithmic solution coincides with the maximizer of the following penalized log-likelihood

$$\sum_{j=1}^{J}\sum_{i=1}^{n_j}\Delta_{ji}\left(\log\lambda_f(M_{ji})+\boldsymbol{\xi}^T\mathbf{x}_{ji}(m)\right)-\frac{1}{\delta}\sum_{\kappa=1}^{K}\Big[\gamma_\kappa-\exp\{\gamma_\kappa\}\Big] \tag{4.52}$$

where $\boldsymbol{\xi}=(\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\theta})$ and $\mathbf{x}_{ji}(m)$ are the same as in (4.46). They also provide the programs (in R:survival package) with Newton-Raphson iterative solution as an inner loop and selection of δ in an outer loop. Fortunately, these programs can be used to fit the dual-time Cox frailty models (4.50) up to certain modifications.

Remarks:

1. The dual-time Cox regression considered in this section is tricked to the standard CoxPH problems upon exogenous timescale discretization, such that the existing theory of partial likelihood and the asymptotic results can be applied. We have not studied the dual-time asymptotic behavior when both lifetime and calendar time are considered to be absolutely continuous. In practice with finite-sample observations, the time-discretization trick works perfectly.

2. The dual-time Cox model (4.40) includes only the endogenous covariates $\mathbf{z}_{ji}(m)$ that vary across individuals. In practice, it is also interesting to see the effects of exogenous covariates $\widetilde{\mathbf{z}}(t)$ that are invariant to all the individuals at the same calendar time. Such $\widetilde{\mathbf{z}}(t)$ might be entertained by the traditional lifetime Cox models (4.37) upon the time shift, however, they are not estimable if included as a log-linear term by (4.40), since $\exp\{\boldsymbol{\theta}^T\widetilde{\mathbf{z}}(t)\}$ would be absorbed by the exogenous baseline $\lambda_g(t)$. Therefore, in order to investigate the exogenous covariate effects of $\widetilde{\mathbf{z}}(t)$, we recommend a two-stage procedure with the first stage fitting the dual-time Cox models (4.40), and the second stage modeling the correlation between $\widetilde{\mathbf{z}}(t)$ and the estimated $\widehat{\lambda}_g(t)$ based on the time series techniques.

4.5 Applications in Retail Credit Risk Modeling

The retail credit risk modeling has become of increasing importance in the recent years; see among others the special issues edited by Berlin and Mester (*Journal of Banking and Finance*, 2004) and Wallace (*Real Estate Economics*, 2005). In quantitative understanding of the risk determinants of retail credit portfolios, there have been large gaps among academic researchers, banks' practitioners and governmental regulators. It turns out that many existing models for corporate credit risk (e.g. the Merton model) are not directly applicable to retail credit.

In this section we demonstrate the application of dual-time survival analysis to credit card and mortgage portfolios in retail banking. Based on our real data-analytic experiences, some tweaked samples are considered here, only for the purpose of methodological illustration. Specifically, we employ the pooled credit card data for illustrating the dual-time nonparametric methods, and use the mortgage loan-level data to illustrate the dual-time Cox regression modeling. Before applying these DtSA techniques, we have assessed them under the simulation mechanism (4.2); see Figure 4.2 and Figure 4.3.

4.5.1 Credit Card Portfolios

As we have introduced in Section 1.4, the credit card portfolios range from product types, acquisition channels, geographical regions and the like. These attributes may be either treated as the endogenous covariates, or used to define the multiple segments. Here, we demonstrate the credit card risk modeling with both segmentation and covariates. For simplicity, we consider two generic segments and a categorical covariate with three levels, where the categorical variable is defined by the credit score buckets upon loan origination: Low if FICO < 640, Medium if $640 \leq$ FICO < 740 and High if FICO ≥ 740. Other variables like the credit line and utilization rate are not considered here.

Table 4.1: Illustrative data format of pooled credit card loans

SegID	FICO	V	U	M	D	W
1	Low	200501	0	3	0	988
1	Low	200501	0	3	1	2
1	Low	200501	3	4	0	993
1	Low	200501	3	4	1	5
1	Low	200501	4	5	0	979
1	Low	200501	4	5	1	11
1	Low	200501	⋮	⋮	⋮	⋮
1	Low	200502	⋮	⋮	⋮	⋮
1	Medium	⋮	⋮	⋮	⋮	⋮
2	⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮	⋮	⋮

V: vintage. [U,M]: lifetime interval. D: status. W: counts.

In this simple case, one may aggregate the observations at the vintage level for each FICO bucket in each segment. See Table 4.1 for an illustration of the counting format of the input data, where each single vintage at each lifetime interval $[U, M]$ has two records of weight W, namely the number of survived loans ($D = 0$) and the number of defaulted loans ($D = 1$). Note that the last column of Table 4.1 also illustrates that in the lifetime interval $[3, 4]$, the 200501 vintage has 3 attrition accounts (i.e. $993 - 979 - 11$) which are examples of being randomly censored.

Let us consider the card vintages originated from year 2001 to 2008 with observations (a) left truncated at the beginning of 2005, (b) right censored by the end of 2008, and (c) up censored by 48 months-on-book maximum. Refer to Figure 1.3 about the vintage diagram of dual-time data collection, which resembles the rectangular Lexis diagram of Figure 4.1. First of all, we explore the empirical hazards of the two hypothetical segments, regardless of the FICO variable.

We applied the one-way, two-way and three-way nonparametric methods developed in Section 4.2 for each segment of dual-time-to-default data. Note that given the pooled data with weights W for $D = 0, 1$, it is rather straightforward to modify

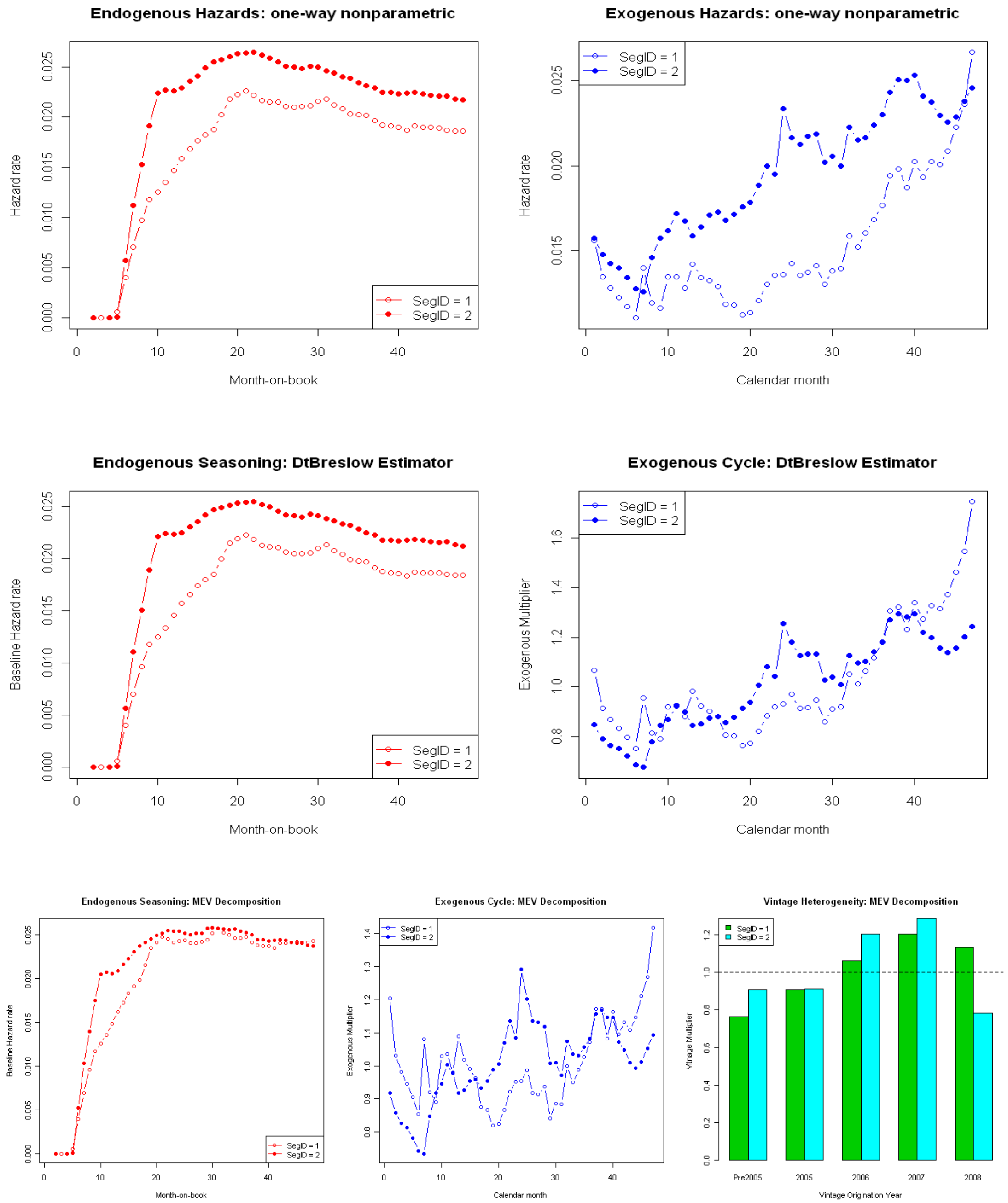


Figure 4.4: Nonparametric analysis of credit card risk: (top) one-way empirical hazards, (middle) DtBrewlow estimation, (bottom) MEV decomposition.

the iterative nonparametric estimators in (4.14) – (4.24), and one may also use the weighted partial likelihood estimation based on our discussion in Section 4.4. The numeric results for both segments are shown in Figure 4.4, where we may compare the performances of two segments as follows.

1. The top panel shows the empirical hazard rates in both lifetime and calendar time. By such one-way estimates, Segment 1 has shown better lifetime performance (i.e. lower hazards) than Segment 2, and it has also better calendar-time performance than Segment 2 except for the latest couple of months.

2. The middle panel shows the results of DtBreslow estimator that simultaneously calibrates the endogenous seasoning $\lambda_f(m)$ and the exogenous cycle effects $\lambda_g(t)$ under the two-way hazard model (4.16). Compared to the one-way calibration above, the seasoning effects are similar, but the exogenous cycle effects show dramatic difference. Conditional on the endogenous baseline, the credit cycle of Segment 2 grows relatively steady, while Segment 1 has a sharp exogenous trend since $t = 30$ (middle 2007) and exceeds Segment 1 since about $t = 40$.

3. The bottom panel takes into the vintage heterogeneity, where we consider only the yearly vintages for simplicity. The estimated of $\lambda_f(m)$ and $\lambda_g(t)$ are both affected after adding the vintage effects, since the more heterogeneity effects are explained by the vintage originations, the less variations are attributed to the endogenous and exogenous baselines. Specifically, the vintage performance in both segments have deteriorated from pre-2005 to 2007, and made a turn when entering 2008. Note that Segment 2 has a dramatic improvement of vintage originations in 2008.

Next, consider the three FICO buckets for each segment, and we still take the nonparametric approach for each of 2×3 segments. The MEV decomposition of empirical hazards are shown in Figure 4.5. It is found that in lifetime the low, medium and high FICO buckets have decreasing levels of endogenous baseline hazard rates, where for Segment 2 the hazard rates are evidently non-proportional. The exogenous cycles of low, medium and high FICO buckets co-move (more or less) within each segment. As for the vintage heterogeneity, it is interesting to note that for Segment 1 the high FICO vintages keep deteriorating from pre-2005 to 2008, which is a risky signal. The vintage heterogeneity in other FICO buckets are mostly consistent with the overall pattern in Figure 4.4.

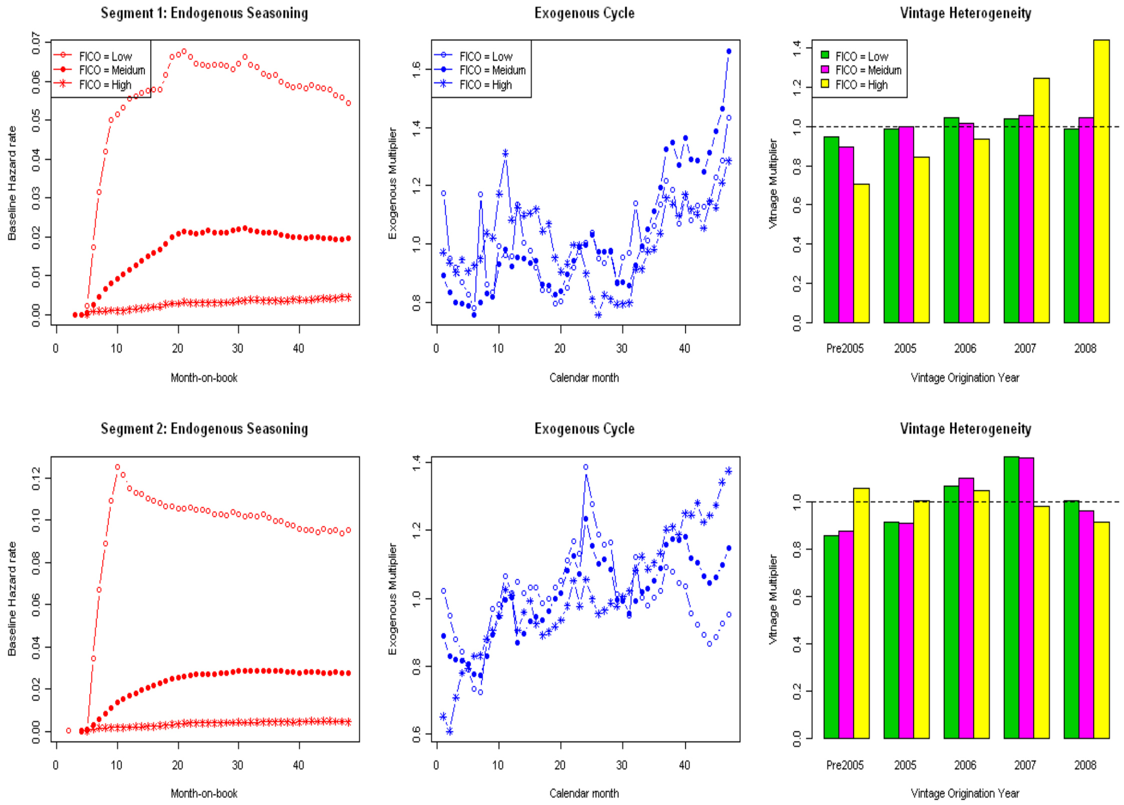


Figure 4.5: MEV decomposition of credit card risk for low, medium and high FICO buckets: Segment 1 (top panel); Segment 2 (bottom panel).

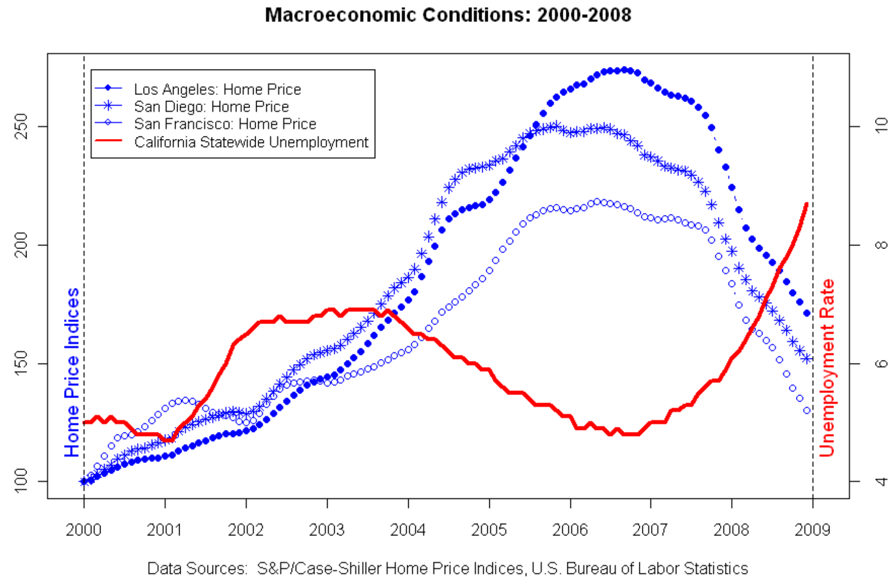
This example mainly illustrates the effectiveness of dual-time nonparametric methods when there are limited or no covariates provided. By the next example of mortgage loans, we shall demonstrate the dual-time Cox regression modeling with various covariate effects on competing risks.

4.5.2 Mortgage Competing Risks

The U.S. residential mortgage market is huge. As of March 2008, the first-lien mortgage debt is estimated to be about \$10 trillion (for 54.7 million loans) with a distribution of \$1.2 trillion subprime mortgages, \$8.2 trillion prime and near-prime combined, and the remaining being government guaranteed; see U.S. Federal Researve report by Frame, Lehnert and Prescott (2008). In brief, the prime mortgages generally target the borrowers with good credit histories and normal down payments, while the subprime mortgages are made to borrowers with poor credit and high leverage.

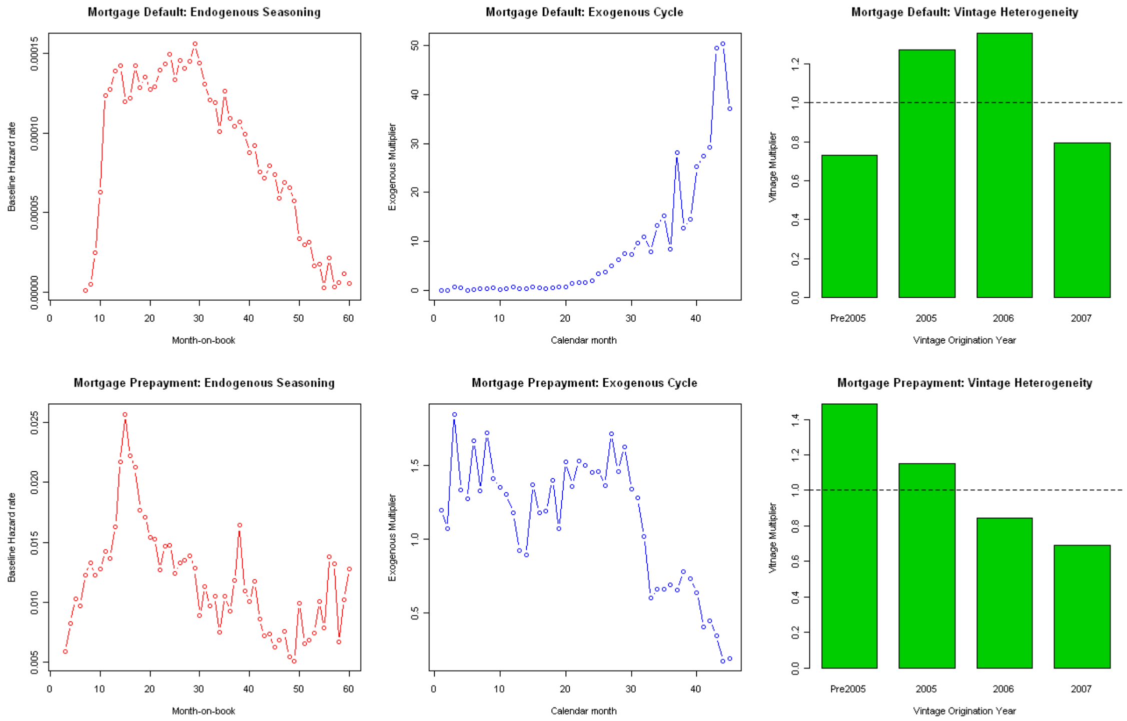
The rise in mortgage defaults is unprecedented. Mayer, Pence and Sherlund (2009) described the rising defaults in nonprime mortgages in terms of underwriting standards and macroeconomic factors. Quantitatively, Sherlund (2008) considered subprime mortgages by proportional hazards models (4.37) with presumed fixed lifetime baselines (therefore, a parametric setting). One may refer to Gerardi, et al. (2008) and Goodman, et al. (2008) for some further details of the subprime issue. In retrospect, the 2007-08 collapse of mortgage market has rather complicated causes, including the loose underwriting standards, the increased leverage, the originate-then-securitize incentives (a.k.a. moral hazard), the fall of home prices, and the worsening unemployment rates.

We make an attempt via dual-time survival analysis to understand the risk determinants of mortgage default and prepayment. Considered here is a tweaked sample of mortgages loans originated from year 2001 to 2007 with Lexis diagram of observations (a) left truncated at the beginning of 2005, (b) right censored by October 2008, and (c) up censored by 60 months-on-book maximum. These loans resemble the nonconforming mortgages in California, where the regional macroeconomic conditions have become worse and worse since middle 2006; see Figure 4.6 for the plots of 2000-08 home price indices and unemployment rate, where we use the up-to-date data sources from S&P/Case-Shiller and U.S. Bureau of Labor Statistics.


Macroeconomic Conditions: 2000-2008

Data Sources: S&P/Case-Shiller Home Price Indices, U.S. Bureau of Labor Statistics

Figure 4.6: Home prices and unemployment rate of California state: 2000-2008.


Mortgage Default: Endogenous Seasoning

Mortgage Default: Exogenous Cycle

Mortgage Default: Vintage Heterogeneity

Mortgage Prepayment: Endogenous Seasoning

Mortgage Prepayment: Exogenous Cycle

Mortgage Prepayment: Vintage Heterogeneity

Figure 4.7: MEV decomposition of mortgage hazards: default vs. prepayment

For a mortgage loan with competing risks caused by default and prepayment, the observed event time corresponds to time to the first event or the censoring time. One may refer to Lawless (2003; §9) for the modeling issues on survival analysis of competing risks. One practically simple approach is to reduce the multiple events to individual survival analysis by treating alternative events as censored. Then, we may fit two marginal hazard rate models for default and prepayment, respectively. One may refer to Therneau and Grambsch (2000; §8) about robust variance estimation for the marginal Cox models of competing risks.

We begin with nonparametric exploration of endogenous and exogenous hazards, as well as the vintage heterogeneity effects. Similar to the credit card risk modeling above, the monthly vintages are grouped yearly to be pre-2005, 2005 2006, 2007 originations. The fitting results of MEV hazard decomposition are shown in Figure 4.7 for both default and prepayment. Conditional on the estimated endogenous baseline hazards of default or prepayment, we find that

1. The exogenous cycle of the default risk has been low (with exogenous multiplier less than 1) prior to $t = 20$ (August 2006), then become rising sharply up to about 50 times more risker at about $t = 43$ (July 2008). In contrast, the exogenous cycle of the prepayment risk has shown a modest town-turn trend during the same calendar time period.

2. The vintage heterogeneity plot implies that the 2005-06 originations have much higher probability of default than the mortgage originations in other years. In contrast, the vintage heterogeneity of prepayment keeps decreasing from pre-2005 to 2007. These findings match the empirical evidences presented lately by Gerardi, et al. (2008).

Compare the estimated exogenous effects on default hazards to the macroeconomic conditions in Figure 4.6. It is obvious that the rise in mortgage defaults follows closely the fall of home prices (positively correlated) and the worsening unemployment rate (negatively correlated). However, based on our demonstration data, it is not possible to quantify what percentage of exogenous effects should be attributed to either house prices or unemployment, because (a) the house prices and unemployment in California are themselves highly correlated from 2005 to 2008, and (b) the exogenous effects on default hazards could be also caused by other macroeconomic indicators. One may of course perform multivariate time series analysis in order to study the correlation between exogenous hazards and macroeconomic indicators, which is beyond the scope of the present thesis.

Next, we try to decode the vintage effect of unobserved heterogeneity by adding the loan-level covariates. For demonstration purpose, we consider only the few mortgage covariates listed in Table 4.2, where CLTV stands for the combined loan-to-value leverage ratio, FICO is the same measurement of credit score as in the credit card risk, and other underwriting covariates upon loan origination can be referred to Goodman, et al. (2008; §1) or Wikipedia. We also include the mortgage note rate, either fixed or floating (i.e., adjustable), to demonstrate the capability of Cox modeling with time-varying covariates. Note that the categorical variable settings are simplified in Table 4.2, and they could be way more complicated in real situations. The mean, standard deviation and range statistics presented for the continuous covariates CLTV, FICO and NoteRate are calculated from our tweaked sample.

The dual-time Cox regression models considered here take the form of (4.4) with both endogenous and exogenous baselines. After incorporating the underwriting and dynamic covariates in Table 4.2, we have

$$\begin{aligned}\lambda_{ji}^{(q)}(m,t) = \lambda_f^{(q)}(m)\lambda_g^{(q)}(t)\exp\Big\{&\theta_1^{(q)}\mathrm{CLTV}_{ji} + \theta_2^{(q)}\mathrm{FICO}_{ji} + \theta_3^{(q)}\mathrm{NoteRate}_{ji}(m) \\ &+ \theta_4^{(q)}\mathrm{DocFull}_{ji} + \theta_5^{(q)}\mathrm{IO}_{ji} + \theta_{6.p}^{(q)}\mathrm{Purpose.purchase}_{ji} + \theta_{6.r}^{(q)}\mathrm{Purpose.refi}_{ji}\Big\}, \quad (4.53)\end{aligned}$$

Table 4.2: Loan-level covariates considered in mortgage credit risk modeling, where NoteRate could be dynamic and others are static upon origination.

CLTV	$\mu = 78, \sigma = 13$ and range $[10, 125]$
FICO	$\mu = 706, \sigma = 35$ and range $[530, 840]$
NoteRate	$\mu = 6.5,\ \sigma = 1.1$ and range $[1, 11]$
Documentation	*full* or *limited*
IO Indicator	*Interest-Only* or not
Loan Purpose	*purchase*, *refi* or *cash-out refi*

Table 4.3: Maximum partial likelihood estimation of mortgage covariate effects in dual-time Cox regression models.

Covariate	Default	Prepayment
CLTV	2.841	−0.285
FICO	−0.500	0.385
NoteRate	1.944	0.2862
DocFull	−1.432	−0.091
IO	1.202	0.141
Purpose.p	−1.284	0.185
Purpose.r	−0.656	−0.307

where the superscript (q) indicates the competing-risk hazards of default or prepayment, the continuous type of covariates FICO, CLTV and NoteRate are all scaled to have zero mean and unit variance, and the 3-level Purpose covariate is broken into 2 dummy variables (Purpose.p and Purpose.r against the standard level of *cash-out refi*). For simplicity, we have assumed the log-linear effects of all 3 continuous covariates on both default and prepayment hazards, while in practice one may need strive to find the appropriate functional forms or break them into multiple buckets.

By the partial likelihood estimation discussed in Section 4.4, we obtain the numerical results tabulated in Table 4.3, where the coefficient estimates are all statistically significant based on our demonstration data (some other non-significant covariates were actually screened out and not considered here). From Table 4.3, we find that conditional on the nonparametric dual-time baselines,

1. The default risk is driven up by *high* CLTV, *low* FICO, *high* NoteRate, *limited* documentation, *Interest-Only*, and *cash-out refi* Purpose.

2. The prepayment risk is driven up by *low* CLTV, *high* FICO, *high* NoteRate, *limited* documentation, *Interest-Only*, and *purchase* Purpose.

3. The covariates CLTV, FICO and Purpose reveals the competing nature of default and prepayment risks. It is actually reasonable to claim that a homeowner with good credit and low leverage would more likely prepay rather than choose to default.

4.6 Summary

This chapter is devoted completely to the survival analysis on Lexis diagram with coverage of nonparametric estimators, structural parameterization and semiparametric Cox regression. For the structural approach based on inverse Gaussian first-passage-time distribution, we have discussed the feasibility of its dual-time extension and the calibration procedure, while its implementation is open to future investigation. Computational results are provided for other dual-time methods developed in this chapter, including assessment of the methodology through the simulation study. Finally, we have demonstrated the application of dual-time survival analysis to the retail credit risk modeling. Some interesting findings in credit card and mortgage risk analysis are presented, which may shed some light on understanding the ongoing credit crisis from a new dual-time perspective.

To end this thesis, we conclude that statistical methods play a key role in credit risk modeling. Developed in this thesis is mainly the *dual-time analytics*, including VDA (vintage data analysis) and DtSA (dual-time survival analysis), which can be applied to model both corporate and retail credit. We have also remarked throughout the chapters the possible extensions and other open problems of interest. So, the end is also a new beginning . . .

4.7 Supplementary Materials

(A) Calibration of Inverse Gaussian Hazard Function

Consider the calibration of the two-parameter inverse Gaussian hazard rate function $\lambda(m; c, b)$ of the form (4.26) from the n observations (U_i, M_i, Δ_i), $i = 1, \ldots, n$ each subject to left truncation and right censoring. Knowing the parametric likelihood given by (4.27), the MLE requires nonlinear optimization that can be carried out by gradient search (e.g. Newton-Raphason algorithm). Provided below are some calculations of the derivatives for the gradient search purpose.

Denote by $\boldsymbol{\eta} = (c, b)^T$ the structural parameters of inverse Gaussian hazard. The log-likelihood function is given by

$$\ell = \sum_{i=1}^{n} \Delta_i \log \lambda(M_i; \boldsymbol{\eta}) - \int_{U_i}^{M_i} \lambda(s; \boldsymbol{\eta}) ds \tag{4.54}$$

Taking the first and second order derivatives, we have the score and Hessian functions

$$\begin{aligned} \frac{\partial \ell}{\partial \boldsymbol{\eta}} &= \sum_{i=1}^{n} \frac{\Delta_i}{\lambda(M_i; \boldsymbol{\eta})} - \int_{U_i}^{M_i} \lambda'(s; \boldsymbol{\eta}) ds \\ \frac{\partial^2 \ell}{\partial \boldsymbol{\eta} \partial \boldsymbol{\eta}^T} &= \sum_{i=1}^{n} \Delta_i \left(\frac{\lambda''(M_i; \boldsymbol{\eta})}{\lambda(M_i; \boldsymbol{\eta})} - \frac{(\lambda'(M_i; \boldsymbol{\eta}))^{\otimes 2}}{\lambda^2(M_i; \boldsymbol{\eta})} \right) - \int_{U_i}^{M_i} \lambda''(s; \boldsymbol{\eta}) ds \end{aligned}$$

where $\lambda'(m; \boldsymbol{\eta}) = \partial \lambda(m; \boldsymbol{\eta}) / \partial \boldsymbol{\eta}$, $\lambda''(m; \boldsymbol{\eta}) = \partial^2 \lambda(t; \boldsymbol{\eta}) / \partial \boldsymbol{\eta} \partial \boldsymbol{\eta}^T$, and $\mathbf{a}^{\otimes 2} = \mathbf{a}\mathbf{a}^T$. By supplying the gradients, the optimization algorithms in most of standard softwares are faster and more reliable. Note that the evaluation of the integral terms above can be carried out by numerical integration methods like the composite trapezoidal or Simpon's rule; see e.g. Press, et al. (2007). Alternatively, one may replace the second term of (4.54) by $\log S(M_i; \boldsymbol{\eta}) - \log S(U_i; \boldsymbol{\eta})$ and take their derivatives based on the explicit expressions below.

The complications lie in the evaluations of $\partial \lambda(m; \boldsymbol{\eta}) / \partial \boldsymbol{\eta}$ and $\partial^2 \lambda(m; \boldsymbol{\eta}) / \partial \boldsymbol{\eta} \partial \boldsymbol{\eta}^T$

based on the inverse Gaussian baseline hazard (4.26), which are, messy though, provided below. Denoting

$$z_1 = \frac{\eta_1 + \eta_2 m}{\sqrt{m}}, \quad z_2 = \frac{-\eta_1 + \eta_2 m}{\sqrt{m}}, \quad \phi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2},$$

evaluate first the derivatives of the survival function (i.e. the denominator of (4.26)):

$$\begin{aligned}
\frac{\partial S(m;\boldsymbol{\eta})}{\partial \eta_1} &= 2\eta_2 e^{-2\eta_1\eta_2}\Phi(z_2) + \frac{1}{\sqrt{m}}\Big[\phi(z_1) + e^{-2\eta_1\eta_2}\phi(z_2)\Big] \\
\frac{\partial S(m;\boldsymbol{\eta})}{\partial \eta_2} &= 2\eta_1 e^{-2\eta_1\eta_2}\Phi(z_2) + \sqrt{m}\Big[\phi(z_1) - e^{-2\eta_1\eta_2}\phi(z_2)\Big] \\
\frac{\partial^2 S(m;\boldsymbol{\eta})}{\partial \eta_1^2} &= -4\eta_2^2 e^{-2\eta_1\eta_2}\Phi(z_2) - \frac{4\eta_2}{\sqrt{m}} e^{-2\eta_1\eta_2}\phi(z_2) - \frac{1}{m}\Big[\phi(z_1)z_1 - e^{-2\eta_1\eta_2}\phi(z_2)z_2\Big] \\
\frac{\partial^2 S(m;\boldsymbol{\eta})}{\partial \eta_1 \partial \eta_2} &= (2 - 4\eta_1\eta_2) e^{-2\eta_1\eta_2}\Phi(z_2) - \Big[\phi(z_1)z_1 - e^{-2\eta_1\eta_2}\phi(z_2)z_2\Big] \\
\frac{\partial^2 S(m;\boldsymbol{\eta})}{\partial \eta_2^2} &= -4\eta_1^2 e^{-2\eta_1\eta_2}\Phi(z_2) + 4\eta_1\sqrt{m} e^{-2\eta_1\eta_2}\phi(z_2) - m\Big[\phi(z_1)z_1 - e^{-2\eta_1\eta_2}\phi(z_2)z_2\Big].
\end{aligned}$$

Then, the partial derivatives of $\lambda(m;\boldsymbol{\eta})$ can be evaluated explicitly by

$$\begin{aligned}
\frac{\partial \lambda(m;\boldsymbol{\eta})}{\partial \eta_1} &= -\lambda_0(m;\boldsymbol{\eta})\left(\frac{\eta_1}{m} + \eta_2 - \frac{1}{\eta_1}\right) - \frac{\lambda_0(m;\boldsymbol{\eta})}{S(m;\boldsymbol{\eta})}\frac{\partial S(m;\boldsymbol{\eta})}{\partial \eta_1} \\
\frac{\partial \lambda(m;\boldsymbol{\eta})}{\partial \eta_2} &= -\lambda(m;\boldsymbol{\eta})\left(\eta_1 + \eta_2 m\right) - \frac{\lambda(m;\boldsymbol{\eta})}{S(m;\boldsymbol{\eta})}\frac{\partial S(m;\boldsymbol{\eta})}{\partial \eta_2} \\
\frac{\partial^2 \lambda(m;\boldsymbol{\eta})}{\partial \eta_1^2} &= -\frac{\partial \lambda(m;\boldsymbol{\eta})}{\partial \eta_1}\left(\frac{\eta_1}{m} + \eta_2 - \frac{1}{\eta_1}\right) - \lambda(m;\boldsymbol{\eta})\left(\frac{1}{\eta_1^2} + \frac{1}{m}\right) \\
&\quad - \frac{\lambda(m;\boldsymbol{\eta})}{S(m;\boldsymbol{\eta})}\frac{\partial^2 S(m;\boldsymbol{\eta})}{\partial \eta_1^2} - \frac{1}{S(m;\boldsymbol{\eta})}\frac{\partial \lambda(m;\boldsymbol{\eta})}{\partial \eta_1}\frac{\partial S(m;\boldsymbol{\eta})}{\partial \eta_1} + \frac{\lambda(m;\boldsymbol{\eta})}{S^2(m;\boldsymbol{\eta})}\left(\frac{\partial S(m;\boldsymbol{\eta})}{\partial \eta_1}\right)^2 \\
\frac{\partial^2 \lambda(m;\boldsymbol{\eta})}{\partial \eta_2^2} &= -\frac{\partial \lambda(m;\boldsymbol{\eta})}{\partial \eta_2}\left(\eta_1 + \eta_2 m\right) - \lambda(m;\boldsymbol{\eta})m \\
&\quad - \frac{\lambda(m;\boldsymbol{\eta})}{S(m;\boldsymbol{\eta})}\frac{\partial^2 S(m;\boldsymbol{\eta})}{\partial \eta_2^2} - \frac{1}{S(m;\boldsymbol{\eta})}\frac{\partial \lambda(m;\boldsymbol{\eta})}{\partial \eta_2}\frac{\partial S(m;\boldsymbol{\eta})}{\partial \eta_2} + \frac{\lambda(m;\boldsymbol{\eta})}{S^2(m;\boldsymbol{\eta})}\left(\frac{\partial S(m;\boldsymbol{\eta})}{\partial \eta_2}\right)^2 \\
\frac{\partial^2 \lambda(m;\boldsymbol{\eta})}{\partial \eta_1 \partial \eta_2} &= -\frac{\partial \lambda(m;\boldsymbol{\eta})}{\partial \eta_1}\left(\eta_1 + \eta_2 m\right) - \lambda(m;\boldsymbol{\eta}) \\
&\quad - \frac{\lambda(m;\boldsymbol{\eta})}{S(m;\boldsymbol{\eta})}\frac{\partial^2 S(m;\boldsymbol{\eta})}{\partial \eta_1 \partial \eta_2} - \frac{1}{S(m;\boldsymbol{\eta})}\frac{\partial \lambda(m;\boldsymbol{\eta})}{\partial \eta_1}\frac{\partial S(m;\boldsymbol{\eta})}{\partial \eta_2} + \frac{\lambda(m;\boldsymbol{\eta})}{S^2(m;\boldsymbol{\eta})}\frac{\partial S(m;\boldsymbol{\eta})}{\partial \eta_1}\frac{\partial S(t;\boldsymbol{\eta})}{\partial \eta_2}.
\end{aligned}$$

(B) Implementation of Cox Modeling with Time-varying Covariates

Consider the Cox regression model with time-varying covariates $\mathbf{x}(m) \in \mathbb{R}^p$:

$$\lambda(m; \mathbf{x}(m)) = \lambda_0(m) \exp\{\boldsymbol{\theta}^T \mathbf{x}(m)\} \tag{4.55}$$

where $\lambda_0(m)$ is an arbitrary baseline hazard function and $\boldsymbol{\theta}$ is the p-vector of regression coefficients. Provided below is the implementation of maximum likelihood estimation of $\boldsymbol{\theta}$ and $\lambda_0(m)$ by tricking it to be a standard CoxPH modeling with time-indepdent covariates.

Let τ_i be the event time for each account $i = 1, \ldots, n$ that is subject to left truncation U_i and right censoring C_i. Given the observations

$$(U_i, M_i, \Delta_i, \mathbf{x}_i(m)), \quad m \in [U_i, M_i], \quad i = 1, \ldots, n, \tag{4.56}$$

in which $M_i = \min\{\tau_i, C_i\}$, $\Delta_i = I(\tau_i < C_i)$, the complete likelihood is given by

$$L = \prod_{i=1}^n \left[\frac{f_i(M_i)}{S_i(U_i)}\right]^{\Delta_i} \left[\frac{S_i(M_i+)}{S_i(U_i)}\right]^{1-\Delta_i} = \prod_{i=1}^n [\lambda_i(M_i)]^{\Delta_i} S_i(M_i+) S_i^{-1}(U_i). \tag{4.57}$$

Under the dynamic Cox model (4.55), the likelihood function is

$$L(\boldsymbol{\theta}, \lambda_0) = \prod_{i=1}^n \Big[\lambda_0(M_i) \exp\{\boldsymbol{\theta}^T \mathbf{x}_i(M_i)\}\Big]^{\Delta_i} \exp\left\{-\int_{U_i}^{M_i} \lambda_0(s) \exp\{\boldsymbol{\beta}^T \mathbf{x}_i(s)\} ds\right\}. \tag{4.58}$$

Taking λ_0 as the nuisance parameter, one may estimate $\boldsymbol{\beta}$ by the partial likelihood; see Cox (1972, 1975) or Section 4.4.

The model estimation with time-varying covariates can be implemented by the standard software package that handles only the time-independent covariates by default. The trick is to construct little time segments such that $\mathbf{x}_i(m)$ can be viewed constant within each segment. See Figure 4.8 for an illustration with the construction

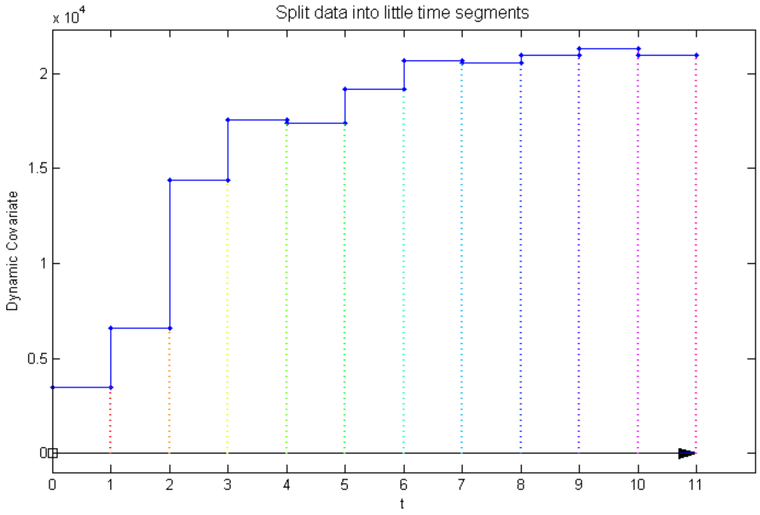


Figure 4.8: Split time-varying covariates into little time segments, illustrated.

of monthly segments. For each account i, let us partition $m \in [U_i, T_i]$ to be

$$U_i \equiv U_{i,0} < U_{i,1} < \cdots < U_{i,L_i} \equiv M_i$$

such that

$$\mathbf{x}_i(m) = \mathbf{x}_{i,l}, \quad m \in (U_{i,l-1}, U_{i,l}]. \tag{4.59}$$

Thus, we have converted the n-sample survival data (4.56) to the following $\sum_{i=1}^{n} L_i$ *independent* observations each with the constant covariates

$$(U_{i,l-1}, U_{i,l}, \Delta_{i,l}, \mathbf{x}_{i,l}), \quad l = 1, \ldots, L_i, \quad i = 1, \ldots, n \tag{4.60}$$

where $\Delta_{i,l} = \Delta_i$ for $l = L_i$ and 0 otherwise.

It is straightforward to show that the likelihood (4.57) can be reformulated by

$$L = \prod_{i=1}^{n} [\lambda_i(M_i)]^{\Delta_i} \prod_{l=1}^{L_i} S_i(U_{i,l}) S_i^{-1}(U_{i,l-1}) = \prod_{i=1}^{n} \prod_{l=1}^{L_i} [\lambda_i(U_{i,l})]^{\Delta_{i,l}} S_i(U_{i,l}) S_i^{-1}(U_{i,l-1}), \tag{4.61}$$

which becomes the likelihood function based on the enlarged data (4.60). Therefore, under the assumption (4.59), the parameter estimation by maximizing (4.58) can be equivalently handled by maximizing

$$L(\boldsymbol{\theta}, \lambda_0) = \prod_{i=1}^{n} \prod_{l=1}^{L_i} \Big[\lambda_0(U_{i,l}) \exp\{\boldsymbol{\theta}^T \mathbf{x}_{i,l}\}\Big]^{\Delta_{i,l}} \exp\left\{ - \exp\{\boldsymbol{\theta}^T \mathbf{x}_{i,l}\} \int_{U_{i,l-1}}^{U_{i,l}} \lambda_0(s) ds \right\}. \tag{4.62}$$

In R, one may perform the MLE directly by applying the CoxPH procedure (R: survival package) to the transformed data (4.60); see Therneau and Grambsch (2000) for computational details.

Furthermore, one may trick the implementation of MLE by logistic regression, provided that the underlying hazard rate function (4.55) admits the linear approximation up to the logit link:

$$\mathsf{logit}(\lambda_i(m)) \approx \boldsymbol{\eta}^T \boldsymbol{\phi}(m) + \boldsymbol{\theta}^T \mathbf{x}_i(m), \quad \text{for } i = 1, \ldots, n \tag{4.63}$$

where $\mathsf{logit}(x) = \log(x/1 - x)$ and $\boldsymbol{\phi}(m)$ is the vector of pre-specified basis functions (e.g. splines). Given the survival observations (4.60) with piecewise-constant covariates, it is easy to show that the discrete version of joint likelihood we have formulated in Section 4.2 can be rewritten as

$$\begin{aligned} L &= \prod_{i=1}^{n} \Big[\lambda_i(M_i)\Big]^{\Delta_i} [1 - \lambda_i(M_i)]^{1-\Delta_i} \prod_{m \in (U_i, M_i)} [1 - \lambda_i(m)] \\ &= \prod_{i=1}^{n} \prod_{l=1}^{L_i} \Big[\lambda_i(M_i)\Big]^{\Delta_{i,l}} [1 - \lambda_i(M_i)]^{1-\Delta_{i,l}} \end{aligned} \tag{4.64}$$

where $\Delta_{i,l}$ are the same as in (4.60). This becomes clearly the likelihood function of the independent binary data $(\Delta_{il}, \mathbf{x}_{i,l})$ for $l = 1, \ldots, L_i$ and $i = 1, \ldots, n$, where $\mathbf{x}_{i,l}$ are constructed by (4.59). Therefore, under the *prameterized* logistic model (4.63), one may perform the MLE for the parameters $(\boldsymbol{\eta}, \boldsymbol{\theta})$ by a standard GLM procedure; see e.g. Faraway (2006).

(C) Frailty-type Vintage Effects in Section 4.4.3

The likelihood function. By the independence of $Z_1, \ldots, Z_k$, it is easy to write down the likelihood as

$$L = \prod_{\kappa=1}^{K} \mathbb{E}_{Z_\kappa} \left[\prod_{j \in \mathcal{G}_\kappa} \prod_{i=1}^{n_j} \left[\lambda_{ji}(M_{ji}|Z_\kappa) \right]^{\Delta_{ji}} \exp\left\{ -\int_{U_{ji}}^{M_{ji}} \lambda_{ji}(s|Z_\kappa) ds \right\} \right]. \tag{4.65}$$

Substituting $\lambda_{ji}(m|Z_\kappa) = \lambda_f(m) Z_\kappa \exp\left\{ \tilde{\boldsymbol{\xi}}^T \tilde{\mathbf{x}}_{ji}(m) \right\}$, we obtain for each $\mathcal{G}_\kappa$ the likelihood contribution

$$L_\kappa = \prod_{j \in \mathcal{G}_\kappa} \prod_{i=1}^{n_j} \left[\lambda_f(M_{ji}) \exp\left\{ \tilde{\boldsymbol{\xi}}^T \tilde{\mathbf{x}}_{ji}(m) \right\} \right]^{\Delta_{ji}} \mathbb{E}_Z \left[Z^{d_\kappa} \exp\{-Z A_\kappa\} \right] \tag{4.66}$$

in which $d_\kappa = \sum_{j \in \mathcal{G}_\kappa} \sum_{i=1}^{n_j} \Delta_{ji}$ and $A_\kappa = \sum_{j \in \mathcal{G}_\kappa} \sum_{i=1}^{n_j} \int_{U_{ji}}^{M_{ji}} \lambda_f(s) \exp\left\{ \tilde{\boldsymbol{\xi}}^T \tilde{\mathbf{x}}_{ji}(s) \right\} ds$. The expectation term in (4.66) can be derived from the d_k-th derivative of Laplace transform of Z which satisfies that $\mathbb{E}_Z \left[Z^{d_\kappa} \exp\{-Z A_\kappa\} \right] = (-1)^{d_\kappa} \mathscr{L}^{(d_\kappa)}(A_\kappa)$; see e.g. Aalen, et al. (2008; §7). Thus, we obtain the joint likelihood

$$L = \prod_{\kappa=1}^{K} \prod_{j \in \mathcal{G}_\kappa} \prod_{i=1}^{n_j} \left[\lambda_f(M_{ji}) \exp\left\{ \tilde{\boldsymbol{\xi}}^T \tilde{\mathbf{x}}_{ji}(m) \right\} \right]^{\Delta_{ji}} (-1)^{d_\kappa} \mathscr{L}^{(d_\kappa)}(A_\kappa), \tag{4.67}$$

or the log-likelihood function of the form (4.51).

EM algorithm for fraitly model estimation.

1. E-step: given the fixed values of $(\widetilde{\boldsymbol{\xi}}, \lambda_f)$, estimate the individual frailty levels $\{Z_\kappa\}_{\kappa=1}^K$ by the conditional expectation

$$\widehat{Z}_\kappa = \mathbb{E}[Z_\kappa | \mathcal{G}_\kappa] = -\frac{\mathscr{L}^{(d_\kappa+1)}(A_\kappa)}{\mathscr{L}^{(d_\kappa)}(A_\kappa)} \tag{4.68}$$

   which corresponds to the empirical Bayes estimate; see Aalen, et al. (2008; §7.2.3). For the gamma frailty $\mathsf{Gamma}(\delta^{-1}, \delta)$ with $\mathscr{L}^{(d)}(x) = (-\delta)^d (1 +$

$\delta x)^{-1/\delta-d}\prod_{i=1}^{d}(1/\delta+i-1)$, the frailty level estimates are given by

$$\widehat{Z}_\kappa = \frac{1+\delta d_\kappa}{1+\delta A_\kappa}, \quad \kappa = 1,\ldots,K. \tag{4.69}$$

2. M-step: given the frailty levels $\{Z_\kappa\}_{\kappa=1}^{K}$, estimate $(\widetilde{\boldsymbol{\xi}},\lambda_f)$ by partial likelihood maximization and Breslow estimator with appropriate modification based on the fixed frailty multipliers

$$PL(\widetilde{\boldsymbol{\xi}}) = \prod_{j=1}^{J}\prod_{i=1}^{n_j}\left[\frac{\exp\{\widetilde{\boldsymbol{\xi}}^T\mathbf{x}_{ji}(M_{ji})\}}{\sum_{(j',i')\in\mathcal{R}_{ji}} Z_{\kappa(j')}\exp\left\{\widetilde{\boldsymbol{\xi}}^T\mathbf{x}_{j'i'}(M_{ji})\right\}}\right]^{\Delta_{ji}} \tag{4.70}$$

$$\widehat{\Lambda}_f(m) = \sum_{j=1}^{J}\sum_{i=1}^{n_j}\frac{I(m\geq M_{ji})\Delta_{ji}}{\sum_{(j',i')\in\mathcal{R}_{ji}} Z_{\kappa(j')}\exp\left\{\widehat{\boldsymbol{\xi}}^T\mathbf{x}_{j'i'}(M_{ji})\right\}}, \tag{4.71}$$

where $\kappa(j)$ indicates the bucket $\mathcal{G}_\kappa$ to which V_j belongs. In R, the partial likelihood maximization can be implemented by the standard CoxPH procedure based on the offset type of log-frailty predictors as provided; details omitted.

To estimate the frailty parameter δ, one may run EM algorithm for δ on a grid to obtain $(\widetilde{\boldsymbol{\xi}}^*(\delta),\lambda_f^*(\cdot;\delta))$, then find the maximizer of the profiled log-likelihood (4.51).

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] Aalen, O. Borgan, O. and Gjessing, H. K. (2008). *Survival and Event History Analysis: A Process Point of View*. Springer, New York.

[2] Aalen, O. and and Gjessing, H. K. (2001). Understanding the shape of the hazard rate: a process point of view (with discussion). *Statistical Science*, **16**, 1–22.

[3] Abramovich, F. and Steinberg, D. M. (1996). Improved inference in nonparametric regression using L_k-smoothing splines. *Journal of Statistical Planning and Inference*, **49**, 327–341.

[4] Abramowitz, M. and Stegun, I. A. (1972). *Handbook of Mathematical Functions* (10th printing). New York: Dover.

[5] Andersen, P. K., Borgan, O., Gill, R. D. and Keiding, N. (1993). *Statistical Models Based on Couting Processes*. Springer, New York.

[6] Andersen, P. K. and Gill, R. D. (1982). Cox's regression model for counting processes: a large sample study. *Annals of Statistics*, **10**, 1100–1120.

[7] Arjas, E. (1986). Stanford heart transplantation data revisited: a real time approach. In: *Modern Statistical Methods in Chronic Disease Epidemiology* (Moolgavkar, S. H. and Prentice, R. L., eds.), pp 65–81. Wiley, New York.

[8] Berlin, M. and Mester, L. J. (eds.) (2004). Special issue of "Retail credit risk management and measurement". *Journal of Banking and Finance*, **28**, 721–899.

[9] Bielecki, T. R. and Rutkowski, M. (2004). *Credit Risk: Modeling, Valuation and Hedging*. Springer-Verlag, Berlin.

[10] Bilias, Y., Gu, M. and Ying, Z. (1997). Towards a general asymptotic theory for Cox model with staggered entry. *Annals of Statistics*, **25**, 662–682.

[11] Black, F. and Cox, J. C. (1976). Valuing corporate securities: Some effects of bond indenture provisions. *Journal of Finance*, **31**, 351–367.

[12] Bluhm, C. and Overbeck, L. (2007) *Structured Credit Portfolio Analysis, Baskets and CDOs*. Boca Raton: Chapman and Hall.

[13] Breeden, J. L. (2007). Modeling data with multiple time dimensions. *Computational Statistics & Data Analysis*, **51**, 4761–4785.

[14] Breslow, N. E. (1972). Discussion of "Regression models and life-tables" by Cox, D. R. *Journal of the Royal Statistical Society: Ser. B*, **25**, 662–682.

[15] Chen, L., Lesmond, D. A. and Wei, J. (2007). Corporate yield spreads and bond liquidity. *Journal of Finance*, **62**, 119–149.

[16] Chhikara, R. S. and Folks, J. L. (1989). *The Inverse Gaussian Distribution: Theory, Methodology, and Applications*. Dekker, New York.

[17] Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B*, **34**, 187–220.

[18] Cox, D. R. (1975). Partial likelihood. *Biometrika*, **62**, 269–276.

[19] Cox, J. C., Ingersoll, J. E. and Ross, S. A. (1985). A theory of the term structure of interest rates. *Econometrika*, **53**, 385–407.

[20] Cressie, N. A. (1993). *Statistics for Spatial Data*. New York: Wiley.

[21] Das, S. R., Duffie, D., Kapadia, N. and Saita, L. (2007). Common failings: how corporate defaults are correlated. *Journal of Finance*, **62**, 93–117.

[22] Deng, Y., Quigley, J. M. and Van Order, R. (2000). Mortgage terminations, heterogeneity and the exercise of mortgage options. *Econometrica*, **68**, 275–307.

[23] Donoho, D. L. and Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, **81** (3), 425-455.

[24] Duchesne, T. and Lawless, J. F. (2000). Alternative time scales and failure time models. *Lifetime Data Analysis*, **6**, 157–179.

[25] Duffie, D., Pan, J. and Singleton, K. (2000). Transform analysis and option pricing for affine jump-diffusions. *Econometrica*, **68**, 1343–1376.

[26] Duffie, D. and Singleton, K. J. (2003). *Credit Risk: Pricing, Measurement, and Management*. Princeton University Press.

[27] Duffie, D., Saita, L. and Wang, K. (2007). Multi-period corporate default prediction with stochastic covariates. *Journal of Financial Economics*, **83**, 635–665.

[28] Duffie, D., Eckner, A., Horel, G. and Saita, L. (2008). Frailty correlated default. *Journal of Finance*, to appear.

[29] Efron, B. (2002). The two-way proportional hazards model. *Journal of the Royal Statistical Society: Ser. B*, **34**, 216–217.

[30] Esper, J., Cook, E. R. and Schweingruber (2002). Low-frequency signals in long tree-ring chronologies for reconstruction past temperature variability. *Science*, **295**, 2250–2253.

[31] Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications.* Chapman & Hall, Boca Raton.

[32] Fan, J., Huang, T. and Li, R. (2007). Analysis of longitudinal data with semiparametric estimation of covariance function. *Journal of the American Statistical Association*, **102**, 632–641.

[33] Fan, J. and Yao, Q. (2003). *Nonlinear Time Series: Nonparametric and Parametric Methods.* Springer, New York.

[34] Fang, K. T., Li, R. and Sudjianto, A. (2006). *Design and Modeling for Compute Experiments.* Baca Raton: Chapman and Hall.

[35] Faraway, J. J. (2006). *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models.* Baca Raton: Chapman and Hall.

[36] Farewell, V. T. and Cox, D. R. (1979). A note on multiple time scales in life testing. *Appl. Statist.*, **28**, 73–75.

[37] Frame, S., Lehnert, A. and Prescott, N. (2008). A snapshot of mortgage conditions with an emphasis on subprime mortgage performance. *U.S. Federal Reserve*, August 2008.

[80] Friedman, J. H. (1991). Multivariate adaptive regression splines (with discussion). *Annals of Statistics*, **19**, 1–141.

[39] Fu, W. J. (2000). Ridge estimator in singular design with applications to age-period-cohort analysis of disease rates. *Communications in Statistics – Theory and Method*, **29**, 263–278.

[40] Fu, W. J. (2008). A smoothing cohort model in age-period-cohort analysis with applications to homicide arrest rates and lung cancer mortality rates. *Sociological Methods and Research*, **36**, 327–361.

[41] Gerardi, K., Sherlund, S. M., Lehnert, A. and Willen, P. (2008). Making sense of the subprime crisis. *Brookings Papers on Economic Activity*, 2008 (2), 69–159.

[42] Giesecke, K. (2006). Default and information. *Journal of Economic Dynamics and Control*, **30**, 2281–2303.

[43] Golub, G. H., Heath, M. and Wahba G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, **21**, 215–223.

[44] Goodman, L. S., Li, S., Lucas, D. J., Zimmerman, T. A. and Fabozzi, F. J. (2008). *Subprime Mortgage Credit Derivatives.* Wiley, Hoboken NJ.

[45] Gramacy, R. B. and Lee, H. K. H. (2008). Bayesian treed Gaussian process models with an application to computer modeling. *Journal of the American Statistical Association*, **103**, 1119–1130.

[46] Green, P. G. and Silverman, B. W. (1994). *Nonparametric Regression and Generalized Linear Models*. Chapman & Hall, London.

[47] Gu, C. (2002). *Smoothing Spline ANOVA Models*. Springer, New York.

[48] Gu, M. G. and Lai, T. L. (1991). Weak convergence of time-sequential censored rank statistics with applications to sequential testing in clinical trials. *Ann. Statist.*, **19**, 1403–1433.

[49] Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models*. New York: Chapman and Hall.

[50] Hougaard, P. (2000). *Analysis of Multivariate Survival Data*. Springer, New York.

[51] Jarrow, R. and Turnbull, S. (1995). Pricing derivatives on financial securities subject to credit risk. *Journal of Finance*, **50**, 53–85.

[52] Keiding, N. (1990). Statistical inference in the Lexis diagram. *Phil. Trans. R. Soc. Lond. A*, **332**, 487–509.

[53] Kimeldorf, G. and Wahba, G. (1970). A correspondence between Bayesian estimation of stochastic processes and smoothing by splines. *Ann. Math. Stat.*, **41**, 495–502.

[54] Kimeldorf, G. and Wahba, G. (1971). Some results on Tchebycheffian spline functions. *J. Math. Anal. Appl.*, **33**, 82–95.

[55] Kupper, L. L., Janis, J. M., Karmous, A. and Greenberg, B. G. (1985). Statistical age-period-cohort analysis: a review and critique. *Journal of Chronic Disease*, **38**, 811–830.

[56] Lando, D. (1998). On Cox processes and credit-risky securities. *Review of Derivatives Research* **2**, 99–120.

[57] Lando, D. (2004). *Credit Risk Modeling: Theory and Applications*. Princeton University Press.

[58] Lawless, J. F. (2003). *Statistical Models and Methods for Lifetime Data*. Wiley, Hoboken NJ.

[59] Lee, M.-L. T. and Whitmore, G. A. (2006). Threshold regression for survival analysis: modeling event times by a stochastic process reaching a boundary. *Statistical Science*, **21**, 501–513.

[60] Lexis, W. (1875). *Einleitung in die Theorie der Bevölkerungsstatistik*. Strassburg: Trübner. Pages 5–7 translated to English by Keyfitz, N. and printed in *Mathematical Demography* (ed. Smith, D. and N. Keyfitz, N., 1977), Springer, Berlin.

[61] Luo, Z. and Wahba, G. (1997). Hybrid adaptive splines. *Journal of the American Statistical Association*, **92**, 107–116.

[62] Madan, D. and Unal, H. (1998). Pricing the risks of default. *Review of Derivatives Research*, **2**, 121–160.

[63] Mason, K. O., Mason, W. H., Winsborough, H. H. and Poole, K. (1973). Some methodological issues in cohort analysis of archival data. *American Sociological Review*, **38**, 242–258.

[64] Mason, W. H. and Wolfinger, N. H. (2002). Cohort analysis. In: *International Encyclopedia of the Social and Behavioral Sciences*, 2189–2194. New York: Elsevier.

[65] Marshall, A. W. and Olkin, I. (2007). *Life Distributions: Structural of Nonparametric, Semiparametric, and Parametric Families*. Springer.

[66] Mayer, C., Pence, K. and Sherlund, S. M. (2009). The rise in mortgage defaults. *Journal of Economic Perspectives*, **23**, 27–50.

[67] McCulloch, C. E. and Nelder, J. A. (1989). *Generalized Linear Models* (2nd ed.). London: Chapman and Hall.

[68] Merton, R. C. (1974). On the pricing of corporate debt: The risk structure of interest rates. *Journal of Finance*, **49**, 1213–1252.

[69] Moody's Special Comment: Corporate Default and Recovery Rates, 1920-2008. Data release: February 2009.

[70] Müller, H. G. and Stadtmüller, U. (1987). Variable bandwidth kernel estimators of regression curves. *Annals of Statistics*, **15**, 282–01.

[71] Nelsen, R. B. (2006). *An introduction to copulas* (2nd ed.). Springer, New York.

[72] Nielsen, G. G., Gill, R. D., Andersen, P. K. and Sørensen, T. I. A. (1992). A counting process approach to maximum likelihood estimation in frailty models *Scan. J. Statist.*, **19**, 25–43.

[73] Nychka, D. W. (1988). Bayesian confidence intervals for a smoothing spline. *Journal of The American Statistical Association*, **83**, 1134–1143.

[74] Oakes, D. (1995). Multiple Time Scales in Survival Analysis. *Lifetime Data Analysis*, **1**, 7–18.

[75] Pintore, A., Speckman, P. and Holmes, C. C. (2006). Spatially adaptive smoothing splines. *Biometrika*, **93**, 113–125.

[76] Press, W. H., Teukolsky, S. A., Vetterling, W. T. and Flannery, B. P. (2007) *Numerical Recipes: The Art of Scientific Computing* (3rd, ed.). Cambridge University Press.

[77] Ramlau-Hansen, H. (1983). Smoothing counting process intensities by means of kernel functions. *Ann. Statist.*, **11**, 453–466.

[78] Ramsay, J. O. and Li, X. (1998). Curve registration. *Journal of the Royal Statistical Society: Ser. B*, **60**, 351–363.

[79] Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning.* The MIT Press.

[80] Rice, J. and Rosenblatt, M. (1983). Smoothing splines: regression, derivatives and deconvolution. *Annals of Statistics*, **11**, 141–156.

[81] Ruppert, D. and Carroll, R. J. (2000). Spatially-adaptive penalties for spline fitting. *Aust. N. Z. J. Statist.*, **42**(2), 205–223.

[82] Schrödinger, E. (1915). Zür theorie der fall- und steigversuche an teilchen mit Brownscher bewegung. *Physikalische Zeitschrift*, **16**, pp. 289–295.

[83] Selke, T. and Siegmund, D. (1983). Sequential analysis of the proportional hazards model. *Biometrika*, **70**, 315–326.

[84] Shepp, L. A. (1966). Radon-Nikodynm derivatives of Gaussian measures. *Annals of Mathematical Statistics*, **37**, 321–354.

[85] Sherlund, S. M. (2008). The past, present, and future of subprime mortgages. *Finance and Economics Discussion Series*, 2008-63, U.S. Federal Reserve Board.

[86] Silverman, B. W. (1985). Some aspects of the spline smoothing approch to nonparametric curve fitting (with discussion). *Journal of the Royal Statistical Society: Ser. B*, **47**, 1–52.

[87] Slud, E. V. (1984). Sequential linear rank tests for two-sample censored survival data. *Ann. Statist.*, **12**, 551–571.

[88] Stein, M. L. (1999). *Interpolation of Spatial Data: Some Theory for Kriging.* New York: Springer-Verlag.

[89] Sujianto, A., Li, R. and Zhang, A. (2006). GAMEED documentation. *Technical report.* Enterprise Quantitative Risk Management, Bank of America, N.A.

[90] Therneau, R. M. and Grambsch, P. M. (2000). *Modeling Survival Data: Extending the Cox Model.* Springer, New York.

[91] Tweedie, M. C. K. (1957). Statistical Properties of Inverse Gaussian Distributions. I and II. *Annals of Mathematical Statistics*, **28**, 362–377 and 696–705.

[92] Vaupel, J. W., Manton, K. G. and Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography* **16**, 439–454.

[93] Vasicek, O. (1977). An equilibrium characterization of the term structure. *Journal of Financial Economics*, **5**, 177–188.

[94] Vidakovic, B. (1999). *Statistical Modeling by Wavelets*. Wiley, New York.

[95] Wahba, G. (1983). Bayesian confidence intervals for the cross-validated smoothing spline. *Journal of the Royal Statistical Society: Ser. B*, **45**, 133–150.

[96] Wahba, G. (1990). *Spline Models for Observational Data*, SSBMS-NSF Regional Conference Series in Applied Mathematics. Philadelphia: SIAM.

[97] Wahba, G. (1995). Discussion of "Wavelet shrinkage: asymptopia" by Donoho, et al. *Journal of The Royal Statistical Society, Ser. B*, **57**, 360-361.

[98] Wallace, N. (ed.) (2005). Special issue of "Innovations in mortgage modeling". *Real Estate Economics*, **33**, 587–782.

[99] Wand, M. P. and Jones, M. C. (1995). *Kernel Smoothing*. Chapman & Hall, London.

[100] Wang, J.-L. (2005). Smoothing hazard rate. In Arbmitage, P. and Colton, T. (Eds.): *Encyclopedia of Biostatistics* (2nd ed.), Volume 7, pp. 4986–4997. Wiley.

[101] Wikipedia, The Free Encylopedia. URL: http://en.wikipedia.org