ABSTRACT

Title of dissertation: Asymptotic Theory for Multiple-Sample Semiparametric Density Ratio Models and Its Application To Mortality Forecasting

Guanhua Lu
Doctor of Philosophy, 2007

Dissertation directed by: Professor Benjamin Kedem
Department of Mathematics

A multiple-sample semiparametric density ratio model, which is equivalent to a generalized logistic regression model, can be constructed by multiplicative exponential distortions of a reference distribution. Distortion functions are assumed to be nonnegative and of a known finite-dimensional parametric form, and the reference distribution is left as nonparametric. The combined data from all the samples are used in the semiparametric large sample problem of estimating each distortion and the reference distribution. The large sample behavior for both the parameters and the unknown reference distribution are studied. The estimated reference cumulative distribution function is proved to converge weakly to a zero-mean Gaussian process, whose covariance structure provides confidence bands for the reference distribution function. A Kolmogorov-Smirnov type statistic for a goodness-of-fit test of the density ratio model is also studied.

In the second part, an approach to modeling and forecasting age-specific mor-

tality in the United States is provided. The approach is based on an extension of a class of semiparametric models to time series. The method combines information from several time series and estimates their predictive distributions conditional on past data. The conditional expectation, the most common predictor, is obtained as a by product from the first moment of the predictive distribution. The confidence band of the predictor is obtained by applying the asymptotic results of the semiparametric density ratio model. A comparison of short term prediction is made between the semiparametric method and the well known method of Lee and Carter [19]. Judging by the mean square error (MSE) of prediction for all ages, the semiparametric method reduces the overall MSE appreciably.

Keywords: Semiparametric; Density-ratio model; Distortion Function; Logistic Regression; Biased Sampling; Case-control Data; Strong Consistency; Profile Log Likelihood; Constrained maximum likelihood estimation; Weak Convergence of Stochastic Processes; Asymptotic Theory; Gaussian Process; Mortality Forecasting; Time Series; U.S. Mortality.

Asymptotic Theory for Multiple-Sample Semiparametric Density Ratio Models and Its Application to Mortality Forecasting

by

Guanhua Lu

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2007

Advisory Committee:

Professor Benjamin Kedem, Chair/Advisor
Professor Doron Avramov
Dr. Myron Katzoff
Professor Partha Lahiri
Professor Paul Smith

Dedication

This dissertation is dedicated to my parents, and to all my family!

Acknowledgments

The past five years of pursuing my Ph.D. degree was accompanied with bitterness, hardships, frustration, encouragement and trust and with so many people's kind help. When I looked back again after finishing writing this dissertation, I found that my graduate experience has been such a joyful and unforgettable one, and my graduate studies at the University of Maryland will be cherished forever. Though it will not be enough to express my gratitude in words to all those people who helped me, I would still like to give my many, many thanks to all these people.

First of all, I would like to gratefully and sincerely thank my advisor, Dr. Benjamin Kedem, for his guidance, understanding, patience, and most importantly, his friendship during my graduate studies at the University of Maryland. His mentorship was paramount in providing inspiration and consistent support for my research study. His wisdom, knowledge and commitment to the highest standards inspired and motivated me. It has always been a pleasure to work with and learn from him. Without his guidance and consistent support, I could not have finished my dissertation successfully. For everything you have done for me, Dr. kedem, thank you!

I would also like to thank for the support, both academically and financially, from the National Center for Health Statistics (NCHS), CDC. Special thanks are given to Dr. Rong Wei, with whom I had been working closely for more than two years, and Paul, D. Williams, my supervisor at NCHS. Without their support and help, the project "Forecasting US Mortality" would not be a possibility.

The members of my dissertation committee, Doron Avramov, Myron Katzoff, Partha Lahiri and Paul Smith, have generously given their time and expertise to better my work. I thank them for their contribution and their good-natured support.

I must acknowledge as well the many friends, colleagues, students and teachers who assisted, advised, and supported my research over the years. Especially, I need to express my gratitude and deep appreciation to Denise Sam and Shihua Wen whose friendship, hospitality, knowledge, and wisdom have supported, enlightened, and entertained me over the many years of our friendship.

Finally, and sure most importantly, I want to thank my family - my mother and father and all my family members, for their love. My parents have always supported and believed in me and guided me through my life. I thank my parents for letting me go, not just this last time, but so many times in my life, thank them for supporting me over the past five years studying and living abroad. I devoted this dissertation to them.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction to Semiparametric Density Ratio Models

1.1 Historical Review of Biased Sampling Models

1.1.1 Nonparametric Maximum Likelihood Estimation

Vardi[1982] studied a length-biased sampling model and developed the asymptotic theory for the corresponding nonparametric estimator. If the length of an object is distributed according to the cumulative distribution function (cdf) G, and if the selection probability for any particular object is proportional to its length, then the following model gives the distribution of the length of sampled objects

$$F(y) = 1/\mu \int_0^y x dG(x),\ y \geq 0, \tag{1.1}$$

where $\mu = \int_0^\infty x dG(x) < \infty$ is considered as a normalization constant, which depends on the distribution G. Here the cdf G is unknown and is to be estimated. The cdf F, the *length-biased* distribution corresponding to G, is a weighted version of G in terms of the weight function x.

Vardi[1985] generalized the two-sample biased sampling model (1.1) to allow

for $s+1$ different biased samples as follows:

$$F_i(y) = W_i(G)^{-1} \int_{-\infty}^{y} w_i(x) dG(x), \quad i = 1, \ldots, s \tag{1.2}$$

where w_i's are given nonnegative selection bias weight functions and $W_i(G) = \int_{-\infty}^{\infty} w_i(x) dG(x)$. Suppose we observe $s+1$ different independent samples

$$X_i = (X_{i1}, \ldots, X_{in_i}) \overset{i.i.d.}{\sim} F_i, \quad i = 0, 1, \ldots, s,$$

where we assume $G \equiv F_0$, which is usually referred as the distribution of the reference sample X_0. The problem is to estimate the underlying distribution G. A simple way to estimate G is on the basis of the reference sample X_0 only. But this ignores the other s samples. We want to find a *biased-corrected* estimator which corrects for the biasing involved in the distributions F_i. Vardi[1985] developed methodology for obtaining a nonparametric maximum likelihood estimate (NPMLE) by using all the $n = n_0 + n_1 + \cdots + n_s$ observations from the $s+1$ independent samples. Gill, Vardi and Wellner[1988] showed the consistency and asymptotic normality of Vardi's NPMLE.

In Vardi[1985] original treatment, the weight functions were assumed completely known. But there are many practical situations in which a complete specification of the weight functions is too restrictive and mostly unrealistic. One way to relax the assumption on the weight functions is to assume that the weight functions belong to a parametric family. Therefore, there are two components in the model that are to be estimated: the unknown reference distribution G and the parameters involved in the weight functions. These kinds of models are called semiparametric biased sampling (selection bias) models. The logistic regression model in

case-control studies, which will be introduced in the following section, is an example of semiparametric biased sampling models.

1.1.2 Logistic Regression Models in Case-control Studies

Another important class of biased sampling models was studied in Prentice and Pyke[1979]. A case-control study is a frequently used tool to study risk factors related to disease incidence. Suppose that m mutually exclusive and exhaustive disease groups are defined and let $D = i$ denote the development of the ith disease during the defined accession period. Let $D = 0$ indicate the disease-free state at the end of the accession period. Suppose that a regression vector $x = (x_1, \ldots, x_p)$ is to be related to disease incidence. Let $P(D = i \mid x)$ denote the probability that an individual with regression vector x develops disease $D = i$ in the defined accession period.

A prospective study in which initially disease-free individuals are followed throughout the whole period would involve direct sampling from $P(D \mid x)$. By comparison, case-control studies involve direct sampling from $P(x \mid D)$, and each sample is obtained from each disease category $D = 0, 1, \ldots, m$. Since $P(x \mid D)$ does not completely determine $P(D \mid x)$, the full prospective model can not be estimated from case-control data alone. However, the odds ratios defined below can be estimated from $P(x \mid D)$. The 'Odds' for disease $D = i$ for an individual with characteristics x, relative to that for an individual with some standard regression

vector x_0, is defined as

$$\frac{P(D=i \mid x)/P(D=0 \mid x)}{P(D=i \mid x_0)/P(D=0 \mid x_0)}, \quad i=1,\ldots,m. \tag{1.3}$$

By applying Bayes' rule

$$P(D \mid x) = P(x \mid D)P(D)/P(x),$$

the odds ratio can be written

$$\frac{P(x \mid D=i)/P(x_0 \mid D=i)}{P(x \mid D=0)/P(x_0 \mid D=0)}, \tag{1.4}$$

$i=1,\ldots,m$. It follows that odds ratio (1.3) can be estimated from case-control data.

Logistic regression models are commonly used in analyzing case-control data. The probability $P(D=i \mid x)$ that an individual develops disease $D=i$ can be specified in terms of the logistic regression model

$$P(D=i \mid x) = \exp(\alpha_i + \beta_i' x)/\sum_{j=0}^{m} \exp(\alpha_j + \beta_j' x), \quad i=0,1,\ldots,m, \tag{1.5}$$

where β_i is a $p \times 1$ vector parameter, with $\alpha_0 = 0, \beta_0 = 0$ for uniqueness. The odds ratios are easily calculated to be

$$\exp\{\beta_i'(x - x_0)\}, \quad i=1,\ldots,m. \tag{1.6}$$

The β_i's are usually called the odds ratio parameters.

Let $p(x)$ be the marginal distribution of x, and let $\pi_i = P(D=i)$. Then by Bayes rule, we have

$$P(x \mid D=i) = \frac{P(D=i \mid x)p(x)}{\pi_i}, \quad i=0,1,\ldots,m,$$

where $p(x)$ is the density function of x, and the π_i's satisfy $\sum_{i=0}^{m} \pi_i = 1$. Therefore,

$$\frac{P(x \mid D = i)}{P(x \mid D = 0)} = \frac{\pi_0}{\pi_i} \frac{P(D = i \mid x)}{P(D = 0 \mid x)}. \tag{1.7}$$

Substitute (1.5) into the previous formula, and notice that $\alpha_0 = 0, \beta_0 = 0$. Then the density ratio becomes

$$\frac{P(x \mid D = i)}{P(x \mid D = 0)} = \exp(\alpha_i^* + \beta_i' x),$$

where $\alpha_i^* = log(\pi_0/\pi_i) + \alpha_i$.

Let $g_i(x)$ denote the conditional density function of $P(x \mid D = i), i = 0, 1, \ldots, m$. We rewrite the previous formula as

$$\frac{g_i(x)}{g_0(x)} = \exp(\alpha_i^* + \beta_i' x). \tag{1.8}$$

This is a 'tilt density ratio model'. The exponential function $\exp(\alpha_i^* + \beta_i' x)$ is the weight function and x is called the distortion function. The function $g_0(x)$ is regarded as the density of the reference sample. Both, the parameters α_i, β_i, and the density g_0 are to be estimated.

Qin and Zhang[1997] considered the two-sample case of model (1.8), and studied the asymptotic theory for the estimates of both the parameters and the reference distribution. A Kolmogorov-Smirnov type statistic was constructed to test the goodness of fit of model (1.8). Later, Zhang[2000c] extended the study to two-sample multiplicative-intercept risk models based on case-control data by replacing the distortion function x with $r(x; \theta)$ in model (1.8), where $r(\cdot; \theta)$ has a known form. Fokianos et al[2001] studied model (1.8) based on multiple samples for the

one-way layout with $\beta_j' x$ replaced by $\beta_j' h(x)$, and developed a statistic for testing the homogeneity among the samples.

In this dissertation, we extend Qin and Zhang[1997] and Zhang[2000c] to the multiple-sample case, and obtain the corresponding asymptotic theory for the estimates. In addition, the density ratio model is applied in U.S. mortality forecasting using multiple short time series.

1.2 Semiparametric Density Ratio Models

Consider the following $m+1$ independent samples,

$$\begin{aligned} X_0 &= (x_{01}, \quad \ldots \quad, x_{0n_0})' \sim g(x) \\ X_1 &= (x_{11}, \quad \ldots \quad, x_{1n_1})' \sim g_1(x) \\ &\vdots \\ X_m &= (x_{m1}, \quad \ldots \quad, x_{mn_m})' \sim g_m(x), \end{aligned} \tag{1.9}$$

where $g_j(x)$ is the probability density of the jth sample. We consider X_0 as the reference sample, and assume its distribution $G(x)$ is unknown. We assume the *density ratio model* relative to the reference $g(x)$

$$\frac{g_j(x)}{g(x)} = \exp(\alpha_j + \beta_j' h(x)), \quad j = 1, \ldots, m. \tag{1.10}$$

This in turn gives the tilt model

$$g_j(x) = e^{\alpha_j + \beta_j' h(x)} g(x), \quad j = 1, ..., m, \tag{1.11}$$

where β_j is a $p \times 1$ vector parameters, and α_j is a scalar parameter, which renders $g_j(x)$ a probability density. In other words, it is a normalization constant. We

assume α_0, β_0 are equal to 0 for uniqueness and $h(x)$ is a $p \times 1$ vector valued distortion or tilt function. The "distorted" densities g_j, the reference g, as well as the α_j and β_j are all unknown, but the distortion function $h(x)$ is assumed known and its choice depends on the data.

Define the weight functions $w_j(x) = \exp(\alpha_j + \beta_j h(x)),\ j = 1, \ldots, m$, and make the following assumption:

Assumption: The first and second moments of $h(t)$ with respect to each distribution defined in (1.11) are finite,

$$\int h(t) w_j(t) dG(t) < \infty, \quad \int h(t) h'(t) w_j(t) dG(t) < \infty, \tag{1.12}$$

for $j = 0, 1, \ldots, m$.

Example 1.1 *Normal distributions.* An important special case of (1.10) is obtained in the normal case. Assume that $x_1 \sim N(\mu_1, \sigma_1^2)$ and $x_2 \sim N(\mu_2, \sigma_2^2)$ with densities g_1 and g_2, respectively. Then the density ratio (1.10) becomes

$$\frac{g_1(x)}{g_2(x)} = \exp\Big\{ \log\left(\frac{\sigma_2}{\sigma_1}\right) + \frac{\mu_2^2}{2\sigma_2^2} - \frac{\mu_1^2}{2\sigma_1^2} + \left(\frac{\mu_1}{\sigma_1^2} - \frac{\mu_2}{\sigma_2^2}, \quad \frac{1}{2\sigma_2^2} - \frac{1}{2\sigma_1^2}\right) \begin{pmatrix} x \\ x^2 \end{pmatrix} \Big\}, \tag{1.13}$$

with parameters

$$\begin{aligned} \alpha &= \log\left(\frac{\sigma_2}{\sigma_1}\right) + \frac{\mu_2^2}{2\sigma_2^2} - \frac{\mu_1^2}{2\sigma_1^2}, \\ \beta &= \left(\frac{\mu_1}{\sigma_1^2} - \frac{\mu_2}{\sigma_2^2}, \frac{1}{2\sigma_2^2} - \frac{1}{2\sigma_1^2}\right)', \end{aligned}$$

and distortion function

$$h(x) = (x, x^2)'.$$

Notice that $h(x)$ reduces to x^2 when $\mu_1 = \mu_2 = 0$, and (1.13) reduces to

$$g_1(x) = e^{\alpha+\beta x^2} g_2(x) \tag{1.14}$$

with scalars α, β. The tilt model (1.14) is useful when the distributions are centered at zero and are symmetric.

Example 1.2 *Gamma distributions.* Let $X_j \sim g_j(x) = \text{Gamma}(\alpha_{\gamma j}, \beta_\gamma)$, $j = 0, 1, \ldots, m$, with a common scale parameter β_γ. Then the weight function is given by

$$\begin{aligned} w_j(x|\alpha_j, \beta_j) &= g_j(x)/g_0(x) = \exp(\alpha_j + \beta_j \log(x)), \\ \alpha_j &= \log \frac{\Gamma(\alpha_{\gamma 0})}{\Gamma(\alpha_{\gamma j})} + (\Gamma(\alpha_{\gamma j}) - \Gamma(\alpha_{\gamma 0})) \log \beta_\gamma, \\ \beta_j &= \alpha_{\gamma j} - \alpha_{\gamma 0}, \end{aligned}$$

where $h(x) = \log(x)$ is the distortion function.

Example 1.3 *Lognormal distributions.* Let $X_j \sim g_j(x) = \text{LN}(\alpha_j, \sigma^2)$, $j = 0, 1, \ldots, m$ with a common σ^2 parameter. Then the weight function is given by

$$\begin{aligned} w_j(x|\alpha_j, \beta_j) &= g_j(x)/g_0(x) = \exp(\alpha_j + \beta_j \log(x)), \\ (\alpha_j, \beta_j) &= \left(\frac{\mu_0^2 - \mu_j^2}{2\sigma^2}, \frac{\mu_0 - \mu_j}{\sigma^2} \right), \end{aligned}$$

The distortion function $h(x) = \log(x)$ is the same as in Example 1.2.

Note that model (1.10) is a biased sampling model with weight function $\exp(\alpha_i + \beta_i' h(x))$ depending on the unknown parameters $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_m)'$ and $\boldsymbol{\beta} = (\beta_1', \ldots, \beta_m')'$. This is a more general model than that of case-control induced by the logistic regression model. Gill, Vardi and Wellner (1988) have discussed the biased sampling problem in the case where $G(x)$ belongs to a nonparametric family and the weight functions are completely specified. Here we leave $G(x)$ nonparametric, but assume that the weight function of the jth sample is $w_j(x, \theta_j) = \exp(\alpha_j + \beta_j' h(x))$, where $\theta_j = (\alpha_j, \beta_j')'$ is unknown.

Denote the combined data from the $m + 1$ samples by $\mathbf{t}$,

$$\mathbf{t} = (t_1, \ldots, t_n)' = (X_0', X_1', \ldots, X_m')', \tag{1.15}$$

where $n = n_0 + n_1 + \cdots + n_m$, the total sample size.

1.3 Estimation

In this section, we will give a profiling procedure to estimate the parameters $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_m)'$ and $\boldsymbol{\beta} = (\beta_1', \ldots, \beta_m')'$. Let $\boldsymbol{\theta} = (\boldsymbol{\alpha}', \boldsymbol{\beta}')'$.

A maximum likelihood estimator of $G(x)$ can be obtained by maximizing the likelihood over the class of step cumulative distribution functions with jumps at the observed values $t_1, \ldots, t_n$. Accordingly, if $p_i = dG(t_i)$ is the mass at t_i, for $i = 1, \ldots, n$, the likelihood becomes

$$\mathcal{L}(\boldsymbol{\theta}, G) = \prod_{i=1}^{n} p_i \prod_{j=1}^{n_1} \exp(\alpha_1 + \beta_1' h(x_{1j})) \cdots \prod_{j=1}^{n_m} \exp(\alpha_m + \beta_m' h(x_{mj})). \tag{1.16}$$

We follow a profiling procedure whereby first we express each p_i in terms of

$\boldsymbol{\theta}$ and then we substitute the p_i back into the likelihood to produce a function of $\boldsymbol{\theta}$ only. Now for fixed $\boldsymbol{\theta}$, (1.16) is maximized by maximizing only the product term $\prod_{i=1}^{n} p_i$, subject to the $m+1$ constraints

$$\sum_{i=1}^{n} p_i = 1, \ \sum_{i=1}^{n} p_i[w_1(t_i) - 1] = 0, \ldots, \sum_{i=1}^{n} p_i[w_m(t_i) - 1] = 0 \tag{1.17}$$

where

$$w_j(t) = \exp(\alpha_j + \beta_j' h(t)), \quad j = 1, \ldots, m.$$

The maximization employs the method of Lagrange multipliers. First, set up the objective function

$$\log \mathcal{L}(\boldsymbol{\theta}, G) - \lambda_0(1 - \sum_{i=1}^{n} p_i) - \lambda_1 \sum_{i=1}^{n} p_i[w_1(t_i) - 1] - \cdots - \lambda_m \sum_{i=1}^{n} p_i[w_m(t_i) - 1],$$

$i = 1, \ldots, n$, then differentiate the objective function with respect to p_i, and set the derivative equal 0,

$$\frac{1}{p_i} + \lambda_0 - \lambda_1[w_1(t_i) - 1] - \cdots - \lambda_m[w_m(t_i) - 1] = 0, \quad i = 1, \ldots, n.$$

Multiply both sides by p_i,

$$1 + \lambda_0 p_i - \lambda_1 p_i[w_1(t_i) - 1] - \cdots - \lambda_m p_i[w_m(t_i) - 1] = 0, \quad i = 1, \ldots, n. \tag{1.18}$$

Sum up over $i = 1, \ldots, n$, and apply the constraints (1.17). Then we have

$$n + \lambda_0 = 0.$$

Substitute $\lambda_0 = -n$ into (1.18), and solve for p_i,

$$p_i = \frac{1}{n + \lambda_1[w_1(t_i) - 1] + \cdots + \lambda_m[w_m(t_i) - 1]}, \quad i = 1, \ldots, n. \tag{1.19}$$

Next, substitute p_i back into the log likelihood function

$$\begin{aligned}\log \mathcal{L}(\boldsymbol{\theta}, G) &= \sum_{i=1}^{n} \log p_i + \sum_{j=1}^{n_1} \log w_1(x_{1j}) + \cdots + \sum_{j=1}^{n_m} \log w_m(x_{mj}) \\ &= -\sum_{i=1}^{n} \log[n + \lambda_1(e^{\alpha_1+\beta_1' h(t_i)} - 1) + \cdots + \lambda_m(e^{\alpha_m+\beta_m' h(t_i)} - 1)] \\ &\quad + \sum_{j=1}^{n_1}(\alpha_1 + \beta_1' h(x_{1j})) + \cdots + \sum_{j=1}^{n_m}(\alpha_m + \beta_m' h(x_{mj})) \qquad (1.20)\end{aligned}$$

To get $\lambda_1, \ldots, \lambda_m$, we differentiate $\log \mathcal{L}$ with respect to $\alpha_1, \ldots, \alpha_m$, respectively, notice (1.19) and apply the constraints (1.17):

$$\begin{aligned}&\frac{\partial \log \mathcal{L}}{\partial \alpha_1} \\ &= -\sum_{i=1}^{n} \frac{\frac{\partial \lambda_1}{\partial \alpha_1}(w_1(t_i) - 1) + \lambda_1 w_1(t_i) + \frac{\partial \lambda_2}{\partial \alpha_1}(w_2(t_i) - 1) + \cdots + \frac{\partial \lambda_m}{\partial \alpha_1}(w_m(t_i) - 1)}{n + \lambda_1[w_1(t_i) - 1] + \cdots + \lambda_m[w_m(t_i) - 1]} \\ &\quad + n_1 \\ &= -\lambda_1 \sum_{i=1}^{n} p_i w_1(t_i) + n_1 \\ &= -\lambda_1 + n_1. \qquad (1.21)\end{aligned}$$

Let $\frac{\partial \log \mathcal{L}}{\partial \alpha_1} = 0$, then $\lambda_1 = n_1$. $\lambda_j = n_j,\ j = 2, \ldots, m$ can be obtained similarly. Substitute all the λ_j's into (1.19). Then we have

$$\begin{aligned}p_i &= \frac{1}{n + n_1[w_1(t_i) - 1] + \cdots + n_m[w_m(t_i) - 1]} \\ &= \frac{1}{n_0 + n_1 w_1(t_i) + \cdots + n_m w_m(t_i)} \\ &= \frac{1}{n_0} \cdot \frac{1}{1 + \rho_1 w_1(t_i) + \cdots + \rho_m w_m(t_i)}, \quad i = 1, \ldots, n, \qquad (1.22)\end{aligned}$$

where $\rho_i = n_i/n_0,\ i = 1, \ldots, n$.

After substituting the p_i's back into (1.16), we obtain the profile log likelihood as a function of $\boldsymbol{\theta}$ only,

$$\ell(\boldsymbol{\theta}) = -n \log n_0 - \sum_{i=1}^{n} \log[1 + \rho_1 w_1(t_i) + \cdots + \rho_m w_m(t_i)]$$

$$+\sum_{j=1}^{n_1}(\alpha_1+\beta_1' h(x_{1j}))+\cdots+\sum_{j=1}^{n_m}(\alpha_m+\beta_m' h(x_{mj})). \qquad (1.23)$$

The score equations for $j=1,\ldots,m$ are therefore

$$\begin{aligned}\frac{\partial \ell}{\partial \alpha_j} &= -\sum_{i=1}^{n}\frac{\rho_j w_j(t_i)}{1+\rho_1 w_1(t_i)+\cdots+\rho_m w_m(t_i)}+n_j=0\\ \frac{\partial \ell}{\partial \beta_j} &= -\sum_{i=1}^{n}\frac{\rho_j w_j(t_i)h(t_i)}{1+\rho_1 w_1(t_i)+\cdots+\rho_m w_m(t_i)}+\sum_{i=1}^{n_j}h(x_{ji})=0. \qquad (1.24)\end{aligned}$$

The solution of the score equations gives the maximum likelihood estimators $\hat{\boldsymbol{\theta}}=(\hat{\boldsymbol{\alpha}}',\hat{\boldsymbol{\beta}}')'$, and consequently by substitution also

$$\hat{p_i}=\frac{1}{n_0}\cdot\frac{1}{1+\rho_1\exp(\hat{\alpha}_1+\hat{\beta}_1 h(t_i))+\cdots+\rho_m\exp(\hat{\alpha}_m+\hat{\beta}_m h(t_i))} \qquad (1.25)$$

and therefore

$$\hat{G}(t)=\frac{1}{n_0}\cdot\sum_{i=1}^{n}\frac{I(t_i\le t)}{1+\rho_1\exp(\hat{\alpha}_1+\hat{\beta}_1 h(t_i))+\cdots+\rho_m\exp(\hat{\alpha}_m+\hat{\beta}_m h(t_i))}. \qquad (1.26)$$

In summary, by following a profiling procedure, we obtained a nonparametric estimator (1.26) for the reference cdf G, and estimating equations (1.24) for the parameters $\boldsymbol{\theta}$.

Chapter 2

Asymptotic Theory for $\hat{\boldsymbol{\theta}}$

In this chapter, we study the asymptotic properties of the estimator $\hat{\boldsymbol{\theta}}$ for the parameter $\boldsymbol{\theta}$. Let $\boldsymbol{\theta}_0 = (\boldsymbol{\alpha}_0', \boldsymbol{\beta}_0')'$ be the true values of $\boldsymbol{\theta} = (\boldsymbol{\alpha}', \boldsymbol{\beta}')'$ under the model (1.11). Throughout this chapter, we assume that the sample size ratio $\rho_j = n_j/n_0$ is positive and finite, and remains fixed as the total sample size $n = \sum_{j=0}^{m} n_j \to \infty$.

A first-order Taylor expansion of $\partial\ell(\hat{\boldsymbol{\theta}})/\partial\boldsymbol{\theta}$ around the true $\boldsymbol{\theta}_0$ gives

$$0 = \frac{\partial\ell(\hat{\boldsymbol{\theta}})}{\partial\boldsymbol{\theta}} = \frac{\partial\ell(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}} + \frac{\partial^2\ell(\boldsymbol{\theta}^*)}{\partial\boldsymbol{\theta}^2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0), \tag{2.1}$$

where $\boldsymbol{\theta}^*$ is between $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_0$. For arbitrary $\boldsymbol{\theta}$, $S_n(\boldsymbol{\theta}) = \partial^2\ell(\boldsymbol{\theta})/\partial\boldsymbol{\theta}^2$ is positive-definite if the model (1.11) is not degenerate. Expression (2.1) can be rewritten as

$$n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = -nS_n(\boldsymbol{\theta}^*)^{-1}n^{-1/2}\frac{\partial\ell(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}}.$$

In classical results, Fisher information $\mathrm{E}(\partial\ell(\boldsymbol{\theta})/\partial\boldsymbol{\theta})^2$ is equal to $-\mathrm{E}S_n(\boldsymbol{\theta})$, where E denotes expectation under $\boldsymbol{\theta}_0$. However, under the density ratio model (1.11), ℓ is a profile log likelihood function, and the contributions to the score statistic $\partial\ell(\boldsymbol{\theta}_0)/\partial\boldsymbol{\theta}$ from individual samples do not in general have mean zero.

Consequently, the variance matrix for $\partial\ell(\boldsymbol{\theta}_0)/\partial\boldsymbol{\theta}$ is not $-\mathrm{E}S_n(\boldsymbol{\theta})$. Assuming that $n_j/n_0 \to \rho_j,\ j = 1, \ldots, m$, as $n \to \infty$, we can derive an asymptotic normal distribution for $n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ by applying the central limit theorem to $\ell(\boldsymbol{\theta}_0)/\partial\boldsymbol{\theta}$ and showing $S_n(\boldsymbol{\theta}^*)$ to be a consistent estimator of $\mathrm{E}S_n(\boldsymbol{\theta})$, which later will be denoted by S.

This chapter is organized as follows. The first section gives the structure of the limit S of matrix $-1/nS_n(\boldsymbol{\theta}_0)$. The second section gives the covariance matrix of the score statistic $\partial\ell(\boldsymbol{\theta}_0)/\partial\boldsymbol{\theta}$. In the last section, the large sample normality of $\hat{\boldsymbol{\theta}}$ and strong consistency as an estimator of $\boldsymbol{\theta}$ are given in Theorem 2.1.

2.1 The Structure of the Limit Matrix S

We start by calculating the second derivatives of the log likelihood function with respect to α_i's and β_i's respectively. It is easy to show that

$$\begin{aligned}
\frac{\partial^2\ell}{\partial\alpha_j^2} &= -\sum_{i=1}^{n} \frac{\left(1 + \sum_{\substack{k=1\\k\neq j}}^{m} \rho_k w_k(t_i)\right) \rho_j w_j(t_i)}{(1 + \sum_{k=1}^{m} \rho_k w_k(t_i))^2} \\
\frac{\partial^2\ell}{\partial\alpha_j\partial\alpha_{j'}} &= \sum_{i=1}^{n} \frac{\rho_j w_j(t_i)\rho_{j'} w_{j'}(t_i)}{(1 + \sum_{k=1}^{m} \rho_k w_k(t_i))^2} \\
\frac{\partial^2\ell}{\partial\alpha_j\partial\beta_j'} &= -\sum_{i=1}^{n} \frac{\left(1 + \sum_{\substack{k=1\\k\neq j}}^{m} \rho_k w_k(t_i)\right) \rho_j w_j(t_i) h'(t_i)}{(1 + \sum_{k=1}^{m} \rho_k w_k(t_i))^2} \\
\frac{\partial^2\ell}{\partial\alpha_j\partial\beta_{j'}'} &= \sum_{i=1}^{n} \frac{\rho_j w_j(t_i)\rho_{j'} w_{j'}(t_i) h'(t_i)}{(1 + \sum_{k=1}^{m} \rho_k w_k(t_i))^2} \\
\frac{\partial^2\ell}{\partial\beta_j\partial\beta_j'} &= -\sum_{i=1}^{n} \frac{\left(1 + \sum_{\substack{k=1\\k\neq j}}^{m} \rho_k w_k(t_i)\right) \rho_j w_j(t_i) h(t_i) h'(t_i)}{(1 + \sum_{k=1}^{m} \rho_k w_k(t_i))^2} \\
\frac{\partial^2\ell}{\partial\beta_j\partial\beta_{j'}'} &= \sum_{i=1}^{n} \frac{\rho_j w_j(t_i)\rho_{j'} w_{j'}(t_i) h(t_i) h'(t_i)}{(1 + \sum_{k=1}^{m} \rho_k w_k(t_i))^2}
\end{aligned} \tag{2.2}$$

where $j, j' = 1, \ldots, m$, and $\frac{\partial^2 \ell}{\partial \alpha_j \partial \beta'_j} = (\frac{\partial^2 \ell}{\partial \beta_j \partial \alpha_j})'$, $\frac{\partial^2 l}{\partial \alpha_j \partial \beta'_{j'}} = (\frac{\partial^2 \ell}{\partial \beta_j \partial \alpha_{j'}})'$. All the second derivatives form the following matrix

$$
\begin{aligned}
S_n = \frac{\partial^2 \ell(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}^2} &= \left. \frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^2} \right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \\
&= \left(\begin{array}{cccccc}
\frac{\partial^2 \ell}{\partial \alpha_1^2} & \cdots & \frac{\partial^2 \ell}{\partial \alpha_1 \partial \alpha_m} & \frac{\partial^2 \ell}{\partial \alpha_1 \partial \beta'_1} & \cdots & \frac{\partial^2 \ell}{\partial \alpha_1 \partial \beta'_m} \\
\vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
\frac{\partial^2 \ell}{\partial \alpha_m \partial \alpha_1} & \cdots & \frac{\partial^2 \ell}{\partial \alpha_m^2} & \frac{\partial^2 \ell}{\partial \alpha_m \partial \beta'_1} & \cdots & \frac{\partial^2 \ell}{\partial \alpha_m \partial \beta'_m} \\
\frac{\partial^2 \ell}{\partial \beta_1 \partial \alpha_1} & \cdots & \frac{\partial^2 \ell}{\partial \beta_1 \partial \alpha_m} & \frac{\partial^2 \ell}{\partial \beta_1 \partial \beta'_1} & \cdots & \frac{\partial^2 \ell}{\partial \beta_1 \partial \beta'_m} \\
\vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
\frac{\partial^2 \ell}{\partial \beta_m \partial \alpha_1} & \cdots & \frac{\partial^2 \ell}{\partial \beta_m \partial \alpha_m} & \frac{\partial^2 \ell}{\partial \beta_m \partial \beta'_1} & \cdots & \frac{\partial^2 \ell}{\partial \beta_m \partial \beta'_m}
\end{array} \right)_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}, \qquad (2.3)
\end{aligned}
$$

where S_n is an $(p+1)m \times (p+1)m$ matrix.

Notice that observation x_{uv} from the uth sample has the density function $w_u(x)g(x)$. By the strong law of large numbers, as $n \to \infty$, we have

$$
\begin{aligned}
-\frac{1}{n}\frac{\partial^2 \ell}{\partial \alpha_j^2} &= \frac{1}{n} \sum_{u=0}^{m} \sum_{v=1}^{n_u} \frac{\left(1 + \sum_{\substack{k=1 \\ k \neq j}}^{m} \rho_k w_k(x_{uv})\right) \rho_j w_j(x_{uv})}{(1 + \sum_{k=1}^{m} \rho_k w_k(x_{uv}))^2} \\
&= \rho_j \frac{n_0}{n} \sum_{u=0}^{m} \rho_u \frac{1}{n_u} \sum_{v=1}^{n_u} \frac{\left(1 + \sum_{\substack{k=1 \\ k \neq j}}^{m} \rho_k w_k(x_{uv})\right) w_j(x_{uv})}{(1 + \sum_{k=1}^{m} \rho_k w_k(x_{uv}))^2} \\
&\to \rho_j \frac{n_0}{n} \sum_{u=0}^{m} \rho_u \int \frac{\left(1 + \sum_{\substack{k=1 \\ k \neq j}}^{m} \rho_k w_k(t)\right) w_j(t)}{(1 + \sum_{k=1}^{m} \rho_k w_k(t))^2} w_u(t) dG(t) \\
&= \rho_j \frac{n_0}{n} \int \frac{(1 + \sum_{\substack{k=1 \\ k \neq j}}^{m} \rho_k w_k(t)) w_j(t)}{(1 + \sum_{k=1}^{m} \rho_k w_k(t))^2} \sum_{u=0}^{m} \rho_u w_u(t) dG(t) \\
&= \frac{\rho_j}{1 + \sum_{k=1}^{m} \rho_k} \int \frac{(1 + \sum_{\substack{k=1 \\ k \neq j}}^{m} \rho_k w_k(t)) w_j(t)}{1 + \sum_{k=1}^{m} \rho_k w_k(t)} dG(t). \qquad (2.4)
\end{aligned}
$$

Similarly,

$$
-\frac{1}{n}\frac{\partial^2 \ell}{\partial \alpha_j \partial \alpha_{j'}} \to \frac{-\rho_j \rho_{j'}}{1 + \sum_{k=1}^{m} \rho_k} \int \frac{w_j(t) w_{j'}(t)}{1 + \sum_{k=1}^{m} \rho_k w_k(t)} dG(t)
$$

$$
\begin{aligned}
-\frac{1}{n}\frac{\partial^2\ell}{\partial\alpha_j\partial\beta_j'} &\rightarrow \frac{\rho_j}{1+\sum_{k=1}^m \rho_k}\int \frac{(1+\sum_{\substack{k=1\\k\neq j}}^m \rho_k w_k(t))w_j(t)h'(t)}{1+\sum_{k=1}^m \rho_k w_k(t)}dG(t)\\
-\frac{1}{n}\frac{\partial^2\ell}{\partial\alpha_j\partial\beta_{j'}'} &\rightarrow \frac{-\rho_j\rho_{j'}}{1+\sum_{k=1}^m \rho_k}\int \frac{w_j(t)w_{j'}(t)h'(t)}{1+\sum_{k=1}^m \rho_k w_k(t)}dG(t)\\
-\frac{1}{n}\frac{\partial^2\ell}{\partial\beta_j\partial\beta_{j'}'} &\rightarrow \frac{-\rho_j\rho_{j'}}{1+\sum_{k=1}^m \rho_k}\int \frac{w_j(t)w_{j'}(t)h(t)h'(t)}{1+\sum_{k=1}^m \rho_k w_k(t)}dG(t)\\
-\frac{1}{n}\frac{\partial^2\ell}{\partial\beta_j\partial\beta_j'} &\rightarrow \frac{\rho_j}{1+\sum_{k=1}^m \rho_k}\int \frac{(1+\sum_{\substack{k=1\\k\neq j}}^m \rho_k w_k(t))w_j(t)h(t)h'(t)}{1+\sum_{k=1}^m \rho_k w_k(t)}dG(t). \quad (2.5)
\end{aligned}
$$

Let S be a matrix consisting of the limits of the components of $-(1/n)S_n$. Then we have

$$-\frac{1}{n}S_n \xrightarrow{a.s.} S,$$

as $n \rightarrow \infty$.

2.2 Covariance Matrix of the Score Statistic

To show that $\mathrm{Var}(\partial\ell(\boldsymbol{\theta}_0)/\partial\boldsymbol{\theta}) = \mathrm{E}(\partial\ell(\boldsymbol{\theta}_0)/\partial\boldsymbol{\theta})^2$, we need to check that the expectation of the score statistic $\partial\ell(\boldsymbol{\theta}_0)/\partial\boldsymbol{\theta}$ equals 0. We can write

$$
\begin{aligned}
E\left\{\frac{\partial\ell}{\partial\alpha_j}\right\} &= -E\left\{\sum_{i=1}^n \frac{\rho_j w_j(t_i)}{1+\rho_k w_k(t_i)+\cdots+\rho_m w_m(t_i)} + n_j\right\}\\
&= -\sum_{u=0}^m \sum_{v=1}^{n_u} E\left\{\frac{\rho_j w_j(t_{uv})}{\sum_{k=0}^m \rho_k w_k(t_{uv})}\right\} + n_j\\
&= -\sum_{u=0}^m n_u \int \frac{\rho_j w_j(t)}{\sum_{k=0}^m \rho_k w_k(t)} w_u(t)dG(t) + n_j\\
&= -n_0\rho_j \int \frac{w_j(t)}{\sum_{k=0}^m \rho_k w_k(t)} \sum_{u=0}^m \rho_u w_u(t)dG(t) + n_j\\
&= -n_0\rho_j \int w_j(t)dG(t) + n_j. \quad (2.6)
\end{aligned}
$$

Since the integral in the last term equals 1, it follows that $E(\partial\ell/\partial\alpha_j) = 0$ from the definition of ρ_j. $E(\partial\ell/\partial\beta_j) = 0$ can be obtained from the same reasoning.

Observations from the same samples are i.i.d., and those from different samples are independent. Keeping this in mind, and noticing that the index j is fixed, the variance of $\partial\ell/\partial\alpha_j$ is as follows:

$$\begin{aligned}
\mathrm{Var}\left[\frac{1}{\sqrt{n}}\frac{\partial\ell}{\partial\alpha_j}\right] &= \frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}\left[\frac{\rho_j w_j(t_i)}{\sum_{k=0}^{m}\rho_k w_k(t_i)}\right] \\
&= \frac{1}{n}\sum_{u=0}^{m}\sum_{v=1}^{n_u}\mathrm{Var}\left[\frac{\rho_j w_j(x_{uv})}{\sum_{k=0}^{m}\rho_k w_k(x_{uv})}\right] \\
&= \frac{1}{n}\rho_j^2\sum_{u=0}^{m} n_u\left[\int\frac{w_j^2(t)w_u(t)}{(\sum_{k=0}^{m}\rho_k w_k(t))^2}dG(t) - \left(\int\frac{w_j(t)w_u(t)}{\sum_{k=0}^{m}\rho_k w_k(t)}dG(t)\right)^2\right] \\
&= \frac{n_0}{n}\rho_j^2\left[\int\frac{w_j^2(t)\sum_{u=0}^{m}\rho_u w_u(t)}{(\sum_{k=0}^{m}\rho_k w_k(t))^2}dG(t) - \sum_{u=0}^{m}\rho_u\left(\int\frac{w_j(t)w_u(t)}{\sum_{k=0}^{m}\rho_k w_k(t)}dG(t)\right)^2\right] \\
&= \frac{\rho_j^2}{\sum_{k=0}^{m}\rho_k}\left[\int\frac{w_j^2(t)}{\sum_{k=0}^{m}\rho_k w_k(t)}dG(t) - \sum_{u=0}^{m}\rho_u\left(\int\frac{w_j(t)w_u(t)}{\sum_{k=0}^{m}\rho_k w_k(t)}dG(t)\right)^2\right] \\
&= \frac{\rho_j^2}{\sum_{k=0}^{m}\rho_k}\left[\int\frac{w_j^2(t)}{\sum_{k=0}^{m}\rho_k w_k(t)}dG(t) - \sum_{u=1}^{m}\rho_u\left(\int\frac{w_j(t)w_u(t)}{\sum_{k=0}^{m}\rho_k w_k(t)}dG(t)\right)^2\right. \\
&\quad \left. - \left(1 - \sum_{u=1}^{m}\rho_u\int\frac{w_j(t)w_u(t)}{\sum_{k=0}^{m}\rho_k w_k(t)}dG(t)\right)^2\right]. \qquad (2.7)
\end{aligned}$$

The second equality is obtained by grouping observations by samples. The third equality comes from the i.i.d property of observations in the same sample, and independence among different samples. The fourth equality is obtained by summing integrals over sample index u. The last equality comes from the fact that, for $j = 1, \ldots, m$,

$$\begin{aligned}
\rho_0\int\frac{w_j(t)w_0(t)}{\sum_{k=0}^{m}\rho_k w_k(t)}dG(t) &= \int\frac{w_j(t)}{\sum_{k=0}^{m}\rho_k w_k(t)}dG(t) \\
&= 1 - \sum_{u=1}^{m}\rho_u\int\frac{w_j(t)w_u(t)}{\sum_{k=0}^{m}\rho_k w_k(t)}dG(t).
\end{aligned}$$

This fact will be used frequently in the future.

For $j \neq j'$, the covariance of the derivatives with respect to α_j and $\alpha_{j'}$ is

$$
\begin{aligned}
\mathrm{Cov}\left(\frac{1}{\sqrt{n}}\frac{\partial \ell}{\partial \alpha_j}, \frac{1}{\sqrt{n}}\frac{\partial \ell}{\partial \alpha_{j'}}\right) &= \frac{1}{n}\mathrm{Cov}\left(\sum_{i=1}^{n}\frac{\rho_j w_j(t_i)}{\sum_{k=0}^{m}\rho_k w_k(t_i)}, \sum_{i=1}^{n}\frac{\rho_{j'} w_{j'}(t_i)}{\sum_{k=0}^{m}\rho_k w_k(t_i)}\right) \\
&= \frac{\rho_j\rho_{j'}}{n}\sum_{i=1}^{n}\left(\mathrm{E}\frac{w_j(t)w_{j'}(t)}{(\sum_{k=0}^{m}\rho_k w_k(t_i))^2} - \mathrm{E}\frac{w_j(t_i)}{\sum_{k=0}^{m}\rho_k w_k(t_i)}\mathrm{E}\frac{w_{j'}(t_i)}{\sum_{k=0}^{m}\rho_k w_k(t_i)}\right) \\
&= \frac{\rho_j\rho_{j'}}{n}\sum_{u=0}^{m} n_u\left(\int\frac{w_j w_{j'} w_u dG}{(\sum_{k=0}^{m}\rho_k w_k)^2} - \int\frac{w_j w_u dG}{\sum_{k=0}^{m}\rho_k w_k}\int\frac{w_{j'} w_u dG}{\sum_{k=0}^{m}\rho_k w_k}\right) \\
&= \frac{n_0\rho_j\rho_{j'}}{n}\left(\int\frac{w_j w_{j'}\sum_{u=0}^{m}\rho_u w_u dG}{(\sum_{k=0}^{m}\rho_k w_k)^2} - \sum_{u=0}^{m}\rho_u\int\frac{w_j w_u dG}{\sum_{k=0}^{m}\rho_k w_k}\int\frac{w_{j'} w_u dG}{\sum_{k=0}^{m}\rho_k w_k}\right) \\
&= \frac{\rho_j\rho_{j'}}{\sum_{k=0}^{m}\rho_k}\left(\int\frac{w_j w_{j'} dG}{\sum_{k=0}^{m}\rho_k w_k} - \sum_{u=0}^{m}\rho_u\int\frac{w_j w_u dG}{\sum_{k=0}^{m}\rho_k w_k}\int\frac{w_{j'} w_u dG}{\sum_{k=0}^{m}\rho_k w_k}\right) \\
&= \frac{\rho_j\rho_{j'}}{\sum_{k=0}^{m}\rho_k}\left(\int\frac{w_j w_{j'} dG}{\sum_{k=0}^{m}\rho_k w_k} - \sum_{u=1}^{m}\rho_u\int\frac{w_j w_u dG}{\sum_{k=0}^{m}\rho_k w_k}\int\frac{w_{j'} w_u dG}{\sum_{k=0}^{m}\rho_k w_k}\right. \\
&\quad \left. -\left(1-\sum_{u=1}^{m}\rho_u\int\frac{w_j(t)w_u(t)}{\sum_{k=0}^{m}\rho_k w_k(t)}dG(t)\right)\left(1-\sum_{u=1}^{m}\rho_u\int\frac{w_{j'}(t)w_u(t)}{\sum_{k=0}^{m}\rho_k w_k(t)}dG(t)\right)\right).
\end{aligned} \tag{2.8}
$$

The covariance of the derivatives with respect to α_j and β_j is

$$
\begin{aligned}
&\mathrm{Cov}\left(\frac{1}{\sqrt{n}}\frac{\partial \ell}{\partial \alpha_j}, \frac{1}{\sqrt{n}}\frac{\partial \ell}{\partial \beta_j}\right) \\
&= \frac{1}{n}\mathrm{Cov}\left[-\sum_{i=1}^{n}\frac{\rho_j w_j(t_i)}{\sum_{k=0}^{m}\rho_k w_k(t_i)} + n_j, -\sum_{i=1}^{n}\frac{\rho_j w_j(t_i)h(t_i)}{\sum_{k=0}^{m}\rho_k w_k(t_i)} + \sum_{i=1}^{n_j}h(x_{ji})\right] \\
&= \frac{1}{n}\mathrm{Cov}\left[-\sum_{i=1}^{n}\frac{\rho_j w_j(t_i)}{\sum_{k=0}^{m}\rho_k w_k(t_i)}, -\sum_{i=1}^{n}\frac{\rho_j w_j(t_i)h(t_i)}{\sum_{k=0}^{m}\rho_k w_k(t_i)} + \sum_{i=1}^{n_j}h(x_{ji})\right] \\
&= \frac{\rho_j^2}{n}\sum_{u=0}^{m} n_u\left(\int\frac{w_j^2 h(t) w_u dG}{(\sum_{k=0}^{m}\rho_k w_k)^2} - \int\frac{w_j w_u dG}{\sum_{k=0}^{m}\rho_k w_k}\int\frac{w_j h(t) w_u dG}{\sum_{k=0}^{m}\rho_k w_k}\right) \\
&\quad -\frac{\rho_j}{n}n_j\left(\int\frac{w_j^2 h(t) dG}{\sum_{k=0}^{m}\rho_k w_k} - \int\frac{w_j^2(t) dG}{\sum_{k=0}^{m}\rho_k w_k}\int h(t)w_j(t)dG\right) \\
&= \frac{\rho_j^2}{\sum_{k=0}^{m}\rho_k}\left(\int\frac{w_j^2(t)dG}{\sum_{k=0}^{m}\rho_k w_k}\int w_j h(t)dG\right. \\
&\quad \left. -\sum_{u=0}^{m}\rho_u\int\frac{w_j w_u dG}{\sum_{k=0}^{m}\rho_k w_k}\int\frac{w_j w_u h(t) dG}{\sum_{k=0}^{m}\rho_k w_k}\right) \\
&= \frac{\rho_j^2}{\sum_{k=0}^{m}\rho_k}\left[\int\frac{w_j^2(t)dG}{\sum_{k=0}^{m}\rho_k w_k}\int w_j h(t)dG\right)
\end{aligned}
$$

$$
\begin{aligned}
&-\sum_{u=1}^{m}\rho_u\int\frac{w_jw_udG}{\sum_{k=0}^{m}\rho_kw_k}\int\frac{w_jw_uh(t)dG}{\sum_{k=0}^{m}\rho_kw_k}\\
&-\left(1-\sum_{u=1}^{m}\rho_u\int\frac{w_j(t)w_u(t)}{\sum_{k=0}^{m}\rho_kw_k(t)}dG(t)\right)\\
&\cdot\left(\int w_j(t)h(t)dG(t)-\sum_{u=1}^{m}\rho_u\int\frac{w_j(t)w_u(t)h(t)}{\sum_{k=0}^{m}\rho_kw_k(t)}dG(t)\right)\Bigg]. \qquad (2.9)
\end{aligned}
$$

In the preceding calculation, the second equality comes from the fact that $E(\frac{\partial\ell}{\partial\beta_j})=0$. After summing the first integral on the forth line over u, the first integrals on the fourth line and the fifth line get canceled. The product in the bracket of the last equality comes from the fact that

$$
\begin{aligned}
&\int\frac{w_j(t)h(t)}{\sum_{k=0}^{m}\rho_kw_k(t)}dG(t)=\\
&\qquad\int w_j(t)h(t)dG(t)-\sum_{u=1}^{m}\rho_u\int\frac{w_j(t)w_u(t)h(t)}{\sum_{k=0}^{m}\rho_kw_k(t)}dG(t). \qquad (2.10)
\end{aligned}
$$

For $j\neq j'$, The covariance of the derivatives with respect to α_j and $\beta_{j'}$ is

$$
\begin{aligned}
&\mathrm{Cov}\left(\frac{1}{\sqrt{n}}\frac{\partial\ell}{\partial\alpha_j},\frac{1}{\sqrt{n}}\frac{\partial\ell}{\partial\beta_{j'}}\right)\\
&=\frac{1}{n}\mathrm{Cov}\left(-\sum_{i=1}^{n}\frac{\rho_jw_j(t_i)}{\sum_{k=0}^{m}\rho_kw_k(t_i)},-\sum_{i=1}^{n}\frac{\rho_{j'}w_{j'}(t_i)h(t_i)}{\sum_{k=0}^{m}\rho_kw_k(t_i)}+\sum_{i=1}^{n_{j'}}h(x_{j'i})\right)\\
&=\frac{\rho_j\rho_{j'}}{n}\sum_{u=0}^{m}n_u\left(\int\frac{w_jw_{j'}h'(t)w_udG}{(\sum_{k=0}^{m}\rho_kw_k)^2}-\int\frac{w_jw_udG}{\sum_{k=0}^{m}\rho_kw_k}\int\frac{w_{j'}h'(t)w_udG}{\sum_{k=0}^{m}\rho_kw_k}\right)\\
&\quad-\frac{\rho_j}{n}n_{j'}\left(\int\frac{w_jw_{j'}h'(t)dG}{\sum_{k=0}^{m}\rho_kw_k}-\int\frac{w_jw_{j'}dG}{\sum_{k=0}^{m}\rho_kw_k}\int h'(t)w_{j'}dG\right)\\
&=\frac{\rho_j\rho_{j'}}{\sum_{k=0}^{m}\rho_k}\left[\int\frac{w_jw_{j'}dG}{\sum_{k=0}^{m}\rho_kw_k}\int h'(t)w_{j'}dG\right.\\
&\quad-\sum_{u=0}^{m}\rho_u\int\frac{w_jw_udG}{\sum_{k=0}^{m}\rho_kw_k}\int\frac{w_{j'}w_uh'(t)dG}{\sum_{k=0}^{m}\rho_kw_k}\\
&\quad-\left(1-\sum_{u=1}^{m}\rho_u\int\frac{w_jw_udG(t)}{\sum_{k=0}^{m}\rho_kw_k(t)}\right)\\
&\quad\left.\cdot\left(\int w_{j'}h\,dG-\sum_{u=1}^{m}\rho_u\int\frac{w_{j'}w_uhdG(t)}{\sum_{k=0}^{m}\rho_kw_k(t)}\right)\right]. \qquad (2.11)
\end{aligned}
$$

The variance of $\frac{\partial \ell}{\partial \beta_j}$ is

$$\begin{aligned}
\mathrm{Var}\left(\frac{1}{\sqrt{n}}\frac{\partial \ell}{\partial \beta_j}\right) &= \frac{1}{n}\mathrm{Var}\left(-\sum_{i=1}^{n}\frac{\rho_j w_j(t_i)h(t_i)}{\sum_{k=0}^{m}\rho_k w_k(t_i)}+\sum_{i=1}^{n_j}h(x_{ji})\right) \\
&= \frac{1}{n}\mathrm{Var}\left(\sum_{i=1}^{n}\frac{\rho_j w_j(t_i)h(t_i)}{\sum_{k=0}^{m}\rho_k w_k(t_i)}\right)+\frac{1}{n}\mathrm{Var}\left(\sum_{i=1}^{n_j}h(x_{ji})\right) \\
&\quad -2\frac{1}{n}\mathrm{Cov}\left(\sum_{i=1}^{n}\frac{\rho_j w_j(t_i)h(t_i)}{\sum_{k=0}^{m}\rho_k w_k(t_i)},\ \sum_{i=1}^{n_j}h(x_{ji})\right) \\
&= \mathrm{I}+\mathrm{II}-2\mathrm{III}. \qquad (2.12)
\end{aligned}$$

Next, we compute the three components separately.

$$\begin{aligned}
\mathrm{I} &= \frac{1}{n}\sum_{u=0}^{m}\sum_{v=1}^{n_u}\mathrm{Var}\left(\frac{\rho_j w_j(x_{uv})h(x_{uv})}{\sum_{k=0}^{m}\rho_k w_k(x_{uv})}\right) \\
&= \frac{\rho_j^2}{n}\sum_{u=0}^{m}n_u\left[\int\frac{w_j^2 h(t)h'(t)w_u}{(\sum_{k=0}^{m}\rho_k w_k(t))^2}dG(t)\right. \\
&\quad \left.-\int\frac{w_j w_u h(t)}{\sum_{k=0}^{m}\rho_k w_k(t)}dG(t)\int\frac{w_j w_u h'(t)}{\sum_{k=0}^{m}\rho_k w_k(t)}dG(t)\right] \\
&= \frac{n_0}{n}\rho_j^2\left[\int\frac{w_j^2(t)h(t)h'(t)\sum_{u=0}^{m}\rho_u w_u(t)}{(\sum_{k=0}^{m}\rho_k w_k(t))^2}dG(t)\right. \\
&\quad \left.-\sum_{u=0}^{m}\rho_u\int\frac{w_j w_u h(t)}{\sum_{k=0}^{m}\rho_k w_k(t)}dG(t)\int\frac{w_j w_u h'(t)}{\sum_{k=0}^{m}\rho_k w_k(t)}dG(t)\right] \\
&= \frac{\rho_j^2}{\sum_{k=0}^{m}\rho_k}\left[\int\frac{w_j^2(t)h(t)h'(t)}{\sum_{k=0}^{m}\rho_k w_k(t)}dG(t)\right. \\
&\quad \left.-\sum_{u=0}^{m}\rho_u\int\frac{w_j w_u h(t)}{\sum_{k=0}^{m}\rho_k w_k(t)}dG(t)\int\frac{w_j w_u h'(t)}{\sum_{k=0}^{m}\rho_k w_k(t)}dG(t)\right]. \qquad (2.13)
\end{aligned}$$

$$\begin{aligned}
\mathrm{II} &= \frac{1}{n}\mathrm{Var}\left[\sum_{i=1}^{n_j}h(x_{ji})\right] \\
&= \frac{n_j}{n}\left[\int h(t)h'(t)w_j(t)dG(t)-\int h(t)w_j(t)dG(t)\int h'(t)w_j(t)dG(t)\right] \qquad (2.14)
\end{aligned}$$

$$\mathrm{III} = \frac{1}{n}\mathrm{Cov}\left(\sum_{i=1}^{n}\frac{\rho_j w_j(t_i)h(t_i)}{\sum_{k=0}^{m}\rho_k w_k(t_i)},\ \sum_{i=1}^{n_j}h(x_{ji})\right)$$

$$
\begin{aligned}
&= \frac{1}{n}\mathrm{Cov}\left(\sum_{i=1}^{n_j}\frac{\rho_j w_j(x_{ji})h(x_{ji})}{\sum_{k=0}^m \rho_k w_k(x_{ji})},\ \sum_{i=1}^{n_j} h(x_{ji})\right) \qquad (2.15)\\
&= \frac{\rho_j}{n} n_j \left(\int \frac{w_j^2(t)h(t)h'(t)}{\sum_{k=0}^m \rho_k w_k(t)} dG(t) - \int \frac{w_j^2 h(t) dG(t)}{\sum_{k=0}^m \rho_k w_k(t)} \int h'(t) w_j(t) dG(t)\right).
\end{aligned}
$$

All the three components are $p \times p$ matrices. Therefore,

$$
\begin{aligned}
\mathrm{Var}\left(\frac{1}{\sqrt{n}}\frac{\partial \ell}{\partial \beta_j}\right) = \frac{\rho_j^2}{\sum_{k=0}^m \rho_k} \Bigg[&- \int \frac{w_j^2(t)h(t)h'(t)}{\sum_{k=0}^m \rho_k w_k(t)} dG(t)\\
&- \sum_{u=0}^m \rho_u \int \frac{w_j w_u h(t)}{\sum_{k=0}^m \rho_k w_k(t)} dG(t) \int \frac{w_j w_u h'(t)}{\sum_{k=0}^m \rho_k w_k(t)} dG(t)\\
&+ 2\int \frac{w_j^2(t)h(t)dG(t)}{\sum_{k=0}^m \rho_k w_k(t)} \int h'(t) w_j(t) dG(t)\Bigg]\\
&+ \frac{\rho_j}{\sum_{k=0}^m \rho_k}\left[\int h(t)h'(t)w_j dG(t) - \int h(t) w_j dG(t) \int h'(t) w_j dG(t)\right]. \qquad (2.16)
\end{aligned}
$$

The second term inside the bracket on the right hand side of the equality is equal to

$$
\begin{aligned}
&\sum_{u=1}^m \rho_u \int \frac{w_j w_u h(t)}{\sum_{k=0}^m \rho_k w_k(t)} dG(t) \int \frac{w_j w_u h'(t)}{\sum_{k=0}^m \rho_k w_k(t)} dG(t)\\
&\quad + \left(\int w_j h(t) dG(t) - \sum_{u=1}^m \rho_u \int \frac{w_j(t) w_u(t) h(t)}{\sum_{k=0}^m \rho_k w_k(t)} dG(t)\right)\\
&\quad \cdot \left(\int w_j h'(t) dG(t) - \sum_{u=1}^m \rho_u \int \frac{w_j(t) w_u(t) h'(t)}{\sum_{k=0}^m \rho_k w_k(t)} dG(t)\right).
\end{aligned}
$$

For $j \neq j'$, the covariance of derivatives with respect to β_j and $\beta_{j'}$ is

$$
\begin{aligned}
&\mathrm{Cov}\left(\frac{1}{\sqrt{n}}\frac{\partial \ell}{\partial \beta_j},\ \frac{1}{\sqrt{n}}\frac{\partial \ell}{\partial \beta_{j'}}\right)\\
&= \frac{1}{n}\mathrm{Cov}\left[-\sum_{i=1}^n \frac{\rho_j w_j(t_i)h(t_i)}{\sum_{k=0}^m \rho_k w_k(t_i)} + \sum_{i=1}^{n_j} h(x_{ji}),\ -\sum_{i=1}^n \frac{\rho_{j'} w_{j'}(t_i)h(t_i)}{\sum_{k=0}^m \rho_k w_k(t_i)} + \sum_{i=1}^{n_{j'}} h(x_{j'i})\right]\\
&= \frac{1}{n}\mathrm{Cov}\left(\sum_{i=1}^n \frac{\rho_j w_j(t_i)h(t_i)}{\sum_{k=0}^m \rho_k w_k(t_i)},\ \sum_{i=1}^n \frac{\rho_{j'} w_{j'}(t_i)h(t_i)}{\sum_{k=0}^m \rho_k w_k(t_i)}\right)\\
&\quad - \frac{1}{n}\mathrm{Cov}\left(\sum_{i=1}^n \frac{\rho_j w_j(t_i)h(t_i)}{\sum_{k=0}^m \rho_k w_k(t_i)},\ \sum_{i=1}^{n_{j'}} h(x_{j'i})\right)
\end{aligned}
$$

$$
\begin{aligned}
&-\frac{1}{n}\mathrm{Cov}\left(\sum_{i=1}^{n_j} h(x_{ji}),\ \sum_{i=1}^{n}\frac{\rho_{j'}w_{j'}(t_i)h(t_i)}{\sum_{k=0}^m \rho_k w_k(t_i)}\right)\\
&+\frac{1}{n}\mathrm{Cov}\left(\sum_{i=1}^{n_j} h(x_{ji}),\ \sum_{i=1}^{n_{j'}} h(x_{j'i})\right)\\
=\ & \mathrm{I}-\mathrm{II}-\mathrm{III}+\mathrm{IV}. \qquad (2.17)
\end{aligned}
$$

The four components are calculated separately,

$$
\begin{aligned}
\mathrm{I} \ =\ & \frac{\rho_j\rho_{j'}}{\sum_{k=0}^m \rho_k}\left(\int \frac{w_j w_{j'}h(t)h'(t)}{\sum_{k=0}^m \rho_k w_k(t)}dG(t)\right.\\
&\left.-\sum_{u=0}^m \rho_u \int \frac{w_j w_u h(t)}{\sum_{k=0}^m \rho_k w_k(t)}dG(t)\int \frac{w_{j'}w_u h'(t)}{\sum_{k=0}^m \rho_k w_k(t)}dG(t)\right).
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{II} \ =\ & \frac{1}{n}\mathrm{Cov}\left(\sum_{i=1}^{n_{j'}}\frac{\rho_j w_j(x_{j'i})h(x_{j'i})}{\sum_{k=0}^m \rho_k w_k(x_{j'i})},\ \sum_{i=1}^{n_{j'}} h(x_{j'i})\right)\\
=\ & \frac{\rho_j}{n}n_{j'}\left(\int \frac{w_j w_{j'}h(t)h'(t)}{\sum_{k=0}^m \rho_k w_k(t)}dG(t)-\int \frac{w_j w_{j'}h(t)dG(t)}{\sum_{k=0}^m \rho_k w_k(t)}\int h'(t)w_{j'}(t)dG(t)\right).
\end{aligned}
$$

Similarly,

$$
\mathrm{III}=\frac{\rho_{j'}}{n}n_j\left(\int \frac{w_j w_{j'}h(t)h'(t)}{\sum_{k=0}^m \rho_k w_k(t)}dG(t)-\int \frac{w_j w_{j'}h(t)dG(t)}{\sum_{k=0}^m \rho_k w_k(t)}\int h'(t)w_j(t)dG(t)\right).
$$

For (IV), since $j\neq j'$, that means observations come from different samples, so we have

$$
\mathrm{IV}=\frac{1}{n}\mathrm{Cov}\left(\sum_{i=1}^{n_j} h(x_{ji}),\ \sum_{i=1}^{n_{j'}} h(x_{j'i})\right)=0.
$$

Therefore,

$$
\begin{aligned}
\mathrm{Cov}\left(\frac{1}{\sqrt{n}}\frac{\partial \ell}{\partial \beta_j},\ \frac{1}{\sqrt{n}}\frac{\partial \ell}{\partial \beta_{j'}}\right)=&\frac{\rho_j\rho_{j'}}{\sum_{k=0}^m \rho_k}\left(-\int \frac{w_j w_{j'}h(t)h'(t)}{\sum_{k=0}^m \rho_k w_k(t)}dG(t)\right. \qquad (2.18)\\
&\left.-\sum_{u=0}^m \rho_u \int \frac{w_j w_u h(t)}{\sum_{k=0}^m \rho_k w_k(t)}dG(t)\int \frac{w_{j'}w_u h'(t)}{\sum_{k=0}^m \rho_k w_k(t)}dG(t)\right.
\end{aligned}
$$

$$+\int \frac{w_j w_{j'} h(t) dG(t)}{\sum_{k=0}^m \rho_k w_k(t)} \int h'(t) w_{j'} dG(t) + \int \frac{w_j w_{j'} h(t) dG(t)}{\sum_{k=0}^m \rho_k w_k(t)} \int h'(t) w_j dG(t) \Bigg).$$

The second term inside the parentheses is equal to

$$\begin{aligned} &-\sum_{u=1}^m \rho_u \int \frac{w_j w_u h(t)}{\sum_{k=0}^m \rho_k w_k(t)} dG(t) \int \frac{w_{j'} w_u h'(t)}{\sum_{k=0}^m \rho_k w_k(t)} dG(t) \\ &\quad -\Bigg(\int w_j h(t) dG(t) - \sum_{u=1}^m \rho_u \int \frac{w_j(t) w_u(t) h(t)}{\sum_{k=0}^m \rho_k w_k(t)} dG(t) \Bigg) \\ &\quad \cdot \Bigg(\int w_{j'} h'(t) dG(t) - \sum_{u=1}^m \rho_u \int \frac{w_{j'}(t) w_u(t) h'(t)}{\sum_{k=0}^m \rho_k w_k(t)} dG(t) \Bigg). \end{aligned}$$

For convenience we introduce the following notation,

$$\begin{aligned} A_{jj'} &= \int \frac{w_j(t) w_{j'}(t)}{\sum_{k=0}^m \rho_k w_k(t)} dG(t) \\ B_{jj'} &= \int \frac{w_j(t) w_{j'}(t) h(t)}{\sum_{k=0}^m \rho_k w_k(t)} dG(t) \\ C_{jj'} &= \int \frac{w_j(t) w_{j'}(t) h(t) h'(t)}{\sum_{k=0}^m \rho_k w_k(t)} dG(t) \\ E_j &= \mathrm{E}(h(x_{ji})) = \int w_j(t) h(t) dG(t) \quad \bar{E}_j = \int w_j(t) h(t) h'(t) dG(t) \\ V_j &= \mathrm{Var}(h(x_{ji})) \\ &= \int w_j h(t) h'(t) dG(t) - \int h(t) w_j dG(t) \int h'(t) w_j dG(t) \\ &= \bar{E}_j - E_j E_j', \end{aligned}$$

where $B_{jj'}$ and E_j are $p \times 1$ vectors, and $C_{jj'}, \bar{E}_j$ and V_j are all $p \times p$ matrices. The finitenes of $B_{jj'}, C_{jj'}, E_j$ and V_j follows from the boundedness of $w_j(t)/(\sum_{k=0}^m \rho_k w_k(t))$ for all j, along with the **Assumption** following the density ratio model (1.11).

We can now rewrite the components of the variance-covariance matrix as

$$\begin{aligned} \mathrm{Var}\left(\frac{1}{\sqrt{n}} \frac{\partial \ell}{\partial \alpha_j} \right) &= \frac{\rho_j^2}{\sum_{k=0}^m \rho_k} \left(A_{jj} - \sum_{u=1}^m \rho_u A_{ju}^2 - (1 - \sum_{u=1}^m \rho_u A_{ju})^2 \right) \\ \mathrm{Cov}\left(\frac{1}{\sqrt{n}} \frac{\partial \ell}{\partial \alpha_j}, \frac{1}{\sqrt{n}} \frac{\partial \ell}{\partial \alpha_{j'}} \right) &= \frac{\rho_j \rho_{j'}}{\sum_{k=0}^m \rho_k} \left(A_{jj'} - \sum_{u=1}^m \rho_u A_{ju} A_{j'u} \right. \end{aligned}$$

$$
\begin{aligned}
& -(1-\sum_{u=1}^{m}\rho_u A_{ju})(1-\sum_{u=1}^{m}\rho_u A_{j'u})\Bigg) \\
\mathrm{Cov}\left(\frac{1}{\sqrt{n}}\frac{\partial \ell}{\partial \alpha_j},\ \frac{1}{\sqrt{n}}\frac{\partial \ell}{\partial \beta_{j'}}\right) &= \frac{\rho_j\rho_{j'}}{\sum_{k=0}^{m}\rho_k}\Bigg(A_{jj'}E_{j'}-\sum_{u=1}^{m}\rho_u A_{ju}B_{j'u} \\
& -(1-\sum_{u=1}^{m}\rho_u A_{ju})(E'_{j'}-\sum_{u=1}^{m}\rho_u B'_{j'u})\Bigg) \\
\mathrm{Var}\left(\frac{1}{\sqrt{n}}\frac{\partial \ell}{\partial \beta_j}\right) &= \frac{\rho_j^2}{\sum_{k=0}^{m}\rho_k}\Bigg(-C_{jj}-\sum_{u=0}^{m}\rho_u B_{ju}B'_{ju}+2B_{jj}E'_j \\
& -(E_j-\sum_{u=1}^{m}\rho_u B_{ju})(E'_j-\sum_{u=1}^{m}\rho_u B'_{ju})\Bigg)+\frac{\rho_j}{\sum_{k=0}^{m}\rho_k}V_j \\
\mathrm{Cov}\left(\frac{1}{\sqrt{n}}\frac{\partial \ell}{\partial \beta_j},\ \frac{1}{\sqrt{n}}\frac{\partial \ell}{\partial \beta_{j'}}\right) &= \frac{\rho_j\rho_{j'}}{\sum_{k=0}^{m}\rho_k}\Bigg(-C_{jj'}-\sum_{u=0}^{m}\rho_u B_{ju}B'_{j'u}+B_{jj'}(E'_j+E'_{j'}) \\
& -(E_j-\sum_{u=1}^{m}\rho_u B_{ju})(E'_{j'}-\sum_{u=1}^{m}\rho_u B'_{j'u})\Bigg). \qquad (2.19)
\end{aligned}
$$

2.3 Consistency and Asymptotic Normality of $\hat{\boldsymbol{\theta}}$

In this section, first we represent the limit matrix S and the variance-covariance matrix in terms of matrices. The large sample properties are summarized in Theorem 2.1.

We define matrices,

$$
A=(A_{ij})_{m\times m},\ B=(B_{ij})_{mp\times m},\ C=(C_{ij})_{mp\times mp},\ \boldsymbol{\rho}=diag(\rho_1,\ldots,\rho_m)_{m\times m}
$$

$$
E=\begin{pmatrix} E_1 & \cdots & \tilde{0} \\ \vdots & \ddots & \vdots \\ \tilde{0} & \cdots & E_m \end{pmatrix}_{mp\times m} \qquad \bar{E}=\begin{pmatrix} \bar{E}_1 & \cdots & \hat{0} \\ \vdots & \ddots & \vdots \\ \hat{0} & \cdots & \bar{E}_m \end{pmatrix}_{mp\times mp}
$$

$$\mathbf{1}_m = \begin{pmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{pmatrix}_{m\times m} \qquad V = \begin{pmatrix} V_1 & \cdots & \hat{0} \\ \vdots & \ddots & \vdots \\ \hat{0} & \cdots & V_m \end{pmatrix}_{mp\times mp} \tag{2.20}$$

where $\tilde{0}$ is a $p \times 1$ vector of 0's, and $\hat{0}$ is a $p \times p$ matrix of 0's.

By introducing the previous notation and (2.19), the variance-covariance matrix has the following structure,

$$\Lambda = \mathrm{Var}\left(\frac{1}{\sqrt{n}}\frac{\partial \ell}{\partial \boldsymbol{\theta}}\right) = \frac{1}{\sum_{k=0}^{m}\rho_k}\begin{pmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{pmatrix}, \tag{2.21}$$

where

$$\begin{aligned}
\Lambda_{11} &= \boldsymbol{\rho}A\boldsymbol{\rho} - \boldsymbol{\rho}A\boldsymbol{\rho}A\boldsymbol{\rho} - \boldsymbol{\rho}\mathbf{1}_m\boldsymbol{\rho} + \boldsymbol{\rho}A\boldsymbol{\rho}\mathbf{1}_m\boldsymbol{\rho} + \boldsymbol{\rho}\mathbf{1}_m\boldsymbol{\rho}A\boldsymbol{\rho} - \boldsymbol{\rho}A\boldsymbol{\rho}\mathbf{1}_m\boldsymbol{\rho}A\boldsymbol{\rho} \\
\Lambda_{12} &= \boldsymbol{\rho}AE'(\boldsymbol{\rho}\otimes I_p) - \boldsymbol{\rho}A\boldsymbol{\rho}B'(\boldsymbol{\rho}\otimes I_p) - \boldsymbol{\rho}\mathbf{1}_mE'(\boldsymbol{\rho}\otimes I_p) + \boldsymbol{\rho}A\boldsymbol{\rho}\mathbf{1}_mE'(\boldsymbol{\rho}\otimes I_p) \\
&\quad + \boldsymbol{\rho}\mathbf{1}_m\boldsymbol{\rho}B'(\boldsymbol{\rho}\otimes I_p) - \boldsymbol{\rho}A\boldsymbol{\rho}\mathbf{1}_mB'(\boldsymbol{\rho}\otimes I_p) \\
\Lambda_{21} &= \Lambda_{12}' = (\boldsymbol{\rho}\otimes I_p)EA\boldsymbol{\rho} - (\boldsymbol{\rho}\otimes I_p)B\boldsymbol{\rho}A\boldsymbol{\rho} - (\boldsymbol{\rho}\otimes I_p)E\mathbf{1}_m\boldsymbol{\rho} \\
&\quad + (\boldsymbol{\rho}\otimes I_p)E\mathbf{1}_m\boldsymbol{\rho}A\boldsymbol{\rho} + (\boldsymbol{\rho}\otimes I_p)B\boldsymbol{\rho}\mathbf{1}_m\boldsymbol{\rho} - (\boldsymbol{\rho}\otimes I_p)B\mathbf{1}_m\boldsymbol{\rho}A\boldsymbol{\rho} \\
\Lambda_{22} &= -(\boldsymbol{\rho}\otimes I_p)C(\boldsymbol{\rho}\otimes I_p) - (\boldsymbol{\rho}\otimes I_p)B\boldsymbol{\rho}B'(\boldsymbol{\rho}\otimes I_p) \\
&\quad + (\boldsymbol{\rho}\otimes I_p)BE'(\boldsymbol{\rho}\otimes I_p) + (\boldsymbol{\rho}\otimes I_p)EB'(\boldsymbol{\rho}\otimes I_p) + (\boldsymbol{\rho}\otimes I_p)V \\
&\quad - (\boldsymbol{\rho}\otimes I_p)E\mathbf{1}_mE'(\boldsymbol{\rho}\otimes I_p) + (\boldsymbol{\rho}\otimes I_p)B\boldsymbol{\rho}\mathbf{1}_mE'(\boldsymbol{\rho}\otimes I_p) \\
&\quad + (\boldsymbol{\rho}\otimes I_p)E\mathbf{1}_m\boldsymbol{\rho}B'(\boldsymbol{\rho}\otimes I_p) - (\boldsymbol{\rho}\otimes I_p)B\boldsymbol{\rho}\mathbf{1}_m\boldsymbol{\rho}B'(\boldsymbol{\rho}\otimes I_p),
\end{aligned} \tag{2.22}$$

where I_p is the $p \times p$ identity matrix, and $\otimes$ denotes the kronecker product.

We rewrite some equations in (2.5) as

$$-\frac{1}{n}\frac{\partial^2\ell}{\partial\alpha_j^2} \;\rightarrow\; \frac{\rho_j}{\sum_{k=0}^{m}\rho_k} - \frac{\rho_j^2}{\sum_{k=0}^{m}\rho_k}\int \frac{w_j^2(t)}{\sum_{k=0}^{m}\rho_k w_k(t)}dG(t).$$

$$
\begin{aligned}
-\frac{1}{n}\frac{\partial^2 \ell}{\partial \alpha_j \partial \beta_j'} &\rightarrow \frac{\rho_j}{\sum_{k=0}^m \rho_k}\int w_j(t)h'(t)dG(t) - \frac{\rho_j^2}{\sum_{k=0}^m \rho_k}\int \frac{w_j^2(t)h'(t)}{\sum_{k=0}^m \rho_k w_k(t)}dG(t) \\
-\frac{1}{n}\frac{\partial^2 \ell}{\partial \beta_j \partial \beta_j'} &\rightarrow \frac{\rho_j}{\sum_{k=0}^m \rho_k}\int w_j h(t)h'(t)dG(t) - \frac{\rho_j^2}{\sum_{k=0}^m \rho_k}\int \frac{w_j^2 h(t)h'(t)}{\sum_{k=0}^m \rho_k w_k}dG(t).
\end{aligned}
\tag{2.23}
$$

Using the previous notation, the limit matrix S from displays (2.5) and (2.23) can be written as

$$
S = \frac{1}{\sum_{k=0}^m \rho_k}\begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix}, \tag{2.24}
$$

where

$$
\begin{aligned}
S_{11} &= \boldsymbol{\rho} - \boldsymbol{\rho} A \boldsymbol{\rho} \\
S_{12} &= \boldsymbol{\rho} E' - \boldsymbol{\rho} B'(\boldsymbol{\rho} \otimes I_p) \\
S_{21} &= S_{12}' = E\boldsymbol{\rho} - (\boldsymbol{\rho} \otimes I_p) B \boldsymbol{\rho} \\
S_{22} &= (\boldsymbol{\rho} \otimes I_p)\bar{E} - (\boldsymbol{\rho} \otimes I_p) C (\boldsymbol{\rho} \otimes I_p).
\end{aligned}
\tag{2.25}
$$

The limit matrix S and covariance matrix Λ are connected by the following lemma.

Lemma 2.1 *The following relationship between S and Λ holds:*

$$
\begin{aligned}
&\Lambda_{11} = S_{11} - S_{11}(\mathbf{1}_m + \boldsymbol{\rho}^{-1})S_{11}, \quad \Lambda_{12} = S_{12} - S_{11}(\mathbf{1}_m + \boldsymbol{\rho}^{-1})S_{12} \\
&\Lambda_{21} = S_{21} - S_{21}(\mathbf{1}_m + \boldsymbol{\rho}^{-1})S_{11}, \quad \Lambda_{22} = S_{22} - S_{21}(\mathbf{1}_m + \boldsymbol{\rho}^{-1})S_{12}.
\end{aligned}
\tag{2.26}
$$

Therefore, we have

$$
\Sigma \stackrel{def}{=} S^{-1}\Lambda S^{-1} = S^{-1} - \sum_{k=0}^m \rho_k \begin{pmatrix} \mathbf{1}_m + \boldsymbol{\rho}^{-1} & 0_{m\times mp} \\ 0_{mp\times m} & 0_{mp\times mp} \end{pmatrix}. \tag{2.27}
$$

Proof. Noticing that $E\boldsymbol{\rho} = (\boldsymbol{\rho} \otimes I_p)E$, the following calculations follow from (2.22) and (2.25):

$$\begin{aligned}
S_{22} - S_{21}\boldsymbol{\rho}^{-1}S_{12} &= (\boldsymbol{\rho} \otimes I_p)\bar{E} - (\boldsymbol{\rho} \otimes I_p)C(\boldsymbol{\rho} \otimes I_p) \\
&\quad -\Big(E\boldsymbol{\rho} - (\boldsymbol{\rho} \otimes I_p)B\boldsymbol{\rho}\Big)\boldsymbol{\rho}^{-1}\Big((\boldsymbol{\rho}E' - \boldsymbol{\rho}B'(\boldsymbol{\rho} \otimes I_p)\Big) \\
&= (\boldsymbol{\rho} \otimes I_p)\bar{E} - (\boldsymbol{\rho} \otimes I_p)C(\boldsymbol{\rho} \otimes I_p) - E\boldsymbol{\rho}\boldsymbol{\rho}^{-1}\boldsymbol{\rho}E' + E\boldsymbol{\rho}\boldsymbol{\rho}^{-1}\boldsymbol{\rho}B'(\boldsymbol{\rho} \otimes I_p) \\
&\quad +(\boldsymbol{\rho} \otimes I_p)B\boldsymbol{\rho}\boldsymbol{\rho}^{-1}\boldsymbol{\rho}E' - (\boldsymbol{\rho} \otimes I_p)B\boldsymbol{\rho}\boldsymbol{\rho}^{-1}\boldsymbol{\rho}B'(\boldsymbol{\rho} \otimes I_p) \\
&= (\boldsymbol{\rho} \otimes I_p)\bar{E} - (\boldsymbol{\rho} \otimes I_p)C(\boldsymbol{\rho} \otimes I_p) - E\boldsymbol{\rho}E' + E\boldsymbol{\rho}B'(\boldsymbol{\rho} \otimes I_p) \\
&\quad +(\boldsymbol{\rho} \otimes I_p)B\boldsymbol{\rho}E' - (\boldsymbol{\rho} \otimes I_p)B\boldsymbol{\rho}B'(\boldsymbol{\rho} \otimes I_p) \\
&= -(\boldsymbol{\rho} \otimes I_p)C(\boldsymbol{\rho} \otimes I_p) - (\boldsymbol{\rho} \otimes I_p)B\boldsymbol{\rho}B'(\boldsymbol{\rho} \otimes I_p) \\
&\quad +(\boldsymbol{\rho} \otimes I_p)\bar{E} - (\boldsymbol{\rho} \otimes I_p)EE' + (\boldsymbol{\rho} \otimes I_p)BE'(\boldsymbol{\rho} \otimes I_p) + (\boldsymbol{\rho} \otimes I_p)EB'(\boldsymbol{\rho} \otimes I_p) \\
&= -(\boldsymbol{\rho} \otimes I_p)C(\boldsymbol{\rho} \otimes I_p) - (\boldsymbol{\rho} \otimes I_p)B\boldsymbol{\rho}B'(\boldsymbol{\rho} \otimes I_p) \\
&\quad +(\boldsymbol{\rho} \otimes I_p)V + (\boldsymbol{\rho} \otimes I_p)BE'(\boldsymbol{\rho} \otimes I_p) + (\boldsymbol{\rho} \otimes I_p)EB'(\boldsymbol{\rho} \otimes I_p)
\end{aligned}$$

and

$$\begin{aligned}
S_{21}\mathbf{1}_m S_{12} &= \Big(E\boldsymbol{\rho} - (\boldsymbol{\rho} \otimes I_p)B\boldsymbol{\rho}\Big)\mathbf{1}_m\Big(\boldsymbol{\rho}E' - \boldsymbol{\rho}B'(\boldsymbol{\rho} \otimes I_p)\Big) \\
&= E\boldsymbol{\rho}\mathbf{1}_m\boldsymbol{\rho}E' - E\boldsymbol{\rho}\mathbf{1}_m\boldsymbol{\rho}B'(\boldsymbol{\rho} \otimes I_p) - (\boldsymbol{\rho} \otimes I_p)B\boldsymbol{\rho}\mathbf{1}_m\boldsymbol{\rho}E' \\
&\quad +(\boldsymbol{\rho} \otimes I_p)B\boldsymbol{\rho}\mathbf{1}_m\boldsymbol{\rho}B'(\boldsymbol{\rho} \otimes I_p) \\
&= -(\boldsymbol{\rho} \otimes I_p)E\mathbf{1}_m E'(\boldsymbol{\rho} \otimes I_p) + (\boldsymbol{\rho} \otimes I_p)B\boldsymbol{\rho}\mathbf{1}_m E'(\boldsymbol{\rho} \otimes I_p) \\
&\quad +(\boldsymbol{\rho} \otimes I_p)E\mathbf{1}_m\boldsymbol{\rho}B'(\boldsymbol{\rho} \otimes I_p) - (\boldsymbol{\rho} \otimes I_p)B\boldsymbol{\rho}\mathbf{1}_m\boldsymbol{\rho}B'(\boldsymbol{\rho} \otimes I_p).
\end{aligned}$$

This completes the derivation of $\Lambda_{22} = S_{22} - S_{21}(\mathbf{1}_m + \boldsymbol{\rho}^{-1})S_{12}$. The rest is similar.

Now we can write Λ in terms of S,

$$\Lambda = S - \frac{1}{\sum_{k=0}^{m} \rho_k} \begin{pmatrix} S_{11} \\ S_{21} \end{pmatrix} (\mathbf{1}_m + \boldsymbol{\rho}^{-1}) \begin{pmatrix} S_{11} & S_{12} \end{pmatrix}.$$

Therefore,

$$\begin{aligned} & S^{-1}\Lambda S^{-1} \\ &= S^{-1} - \sum_{k=0}^{m} \rho_k \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix}^{-1} \begin{pmatrix} S_{11} \\ S_{21} \end{pmatrix} (\mathbf{1}_m + \boldsymbol{\rho}^{-1}) \begin{pmatrix} S_{11} & S_{12} \end{pmatrix} \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix}^{-1} \\ &= S^{-1} - \sum_{k=0}^{m} \rho_k \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} (\mathbf{1}_m + \boldsymbol{\rho}^{-1}) \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \\ &= S^{-1} - \sum_{k=0}^{m} \rho_k \begin{pmatrix} \mathbf{1}_m + \boldsymbol{\rho}^{-1} & 0 \\ 0 & 0 \end{pmatrix}. \end{aligned}$$

This completes the proof of (2.27). □

The following theorem provides the strong consistency and asymptotic theory of the estimator for $\boldsymbol{\theta}_0$.

Theorem 2.1 *Suppose that the model (1.11) and* ***Assumption*** *1.12 hold and that S is positive definite.*

(*a*) *The solution $\hat{\boldsymbol{\theta}}$ to the score equation system (1.24) is a strongly consistent estimator for $\boldsymbol{\theta}_0$.*

(*b*) *As $n \to \infty$,*

$$\sqrt{n} \begin{pmatrix} \hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0 \\ \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 \end{pmatrix} \xrightarrow{d} N_{(p+1)m}(\mathbf{0}, \Sigma), \tag{2.28}$$

where $\Sigma = S^{-1}\Lambda S^{-1}$.

Before proving the consistency and asymptotic normality of the estimator $\hat{\boldsymbol{\theta}}$, we first need to formulate some preliminary tools. The profile log likelihood given by (1.23) can be decomposed into two parts plus a constant,

$$\begin{aligned}\ell(\boldsymbol{\theta}) &= -n\log n_0 - \sum_{i=1}^{n}\log[1+\rho_1 w_1(t_i)+\cdots+\rho_m w_m(t_i)] \\ &\quad + \sum_{j=1}^{n_1}(\alpha_1+\beta_1' h(x_{1j}))+\cdots+\sum_{j=1}^{n_m}(\alpha_m+\beta_m' h(x_{mj})) \\ &\equiv -n\log n_0 - \ell_1 + \ell_2,\end{aligned}$$

where

$$\begin{aligned}\ell_1 &= \sum_{i=1}^{n}\log[1+\rho_1 w_1(t_i)+\cdots+\rho_m w_m(t_i)], \\ \ell_2 &= \sum_{j=1}^{n_1}(\alpha_1+\beta_1' h(x_{1j}))+\cdots+\sum_{j=1}^{n_m}(\alpha_m+\beta_m' h(x_{mj})).\end{aligned}$$

Then we compare the first derivatives of ℓ and ℓ_1.

$$\frac{\partial\ell_1}{\partial\alpha_j}=\sum_{i=1}^{n}\frac{\rho_j w_j(t_i)}{\sum_{k=0}^{m}\rho_k w_k(t_i)},\quad \frac{\partial\ell_1}{\partial\beta_j}=\sum_{i=1}^{n}\frac{\rho_j w_j(t_i)h(t_i)}{\sum_{k=0}^{m}\rho_k w_k(t_i)}.$$

Recall the first derivative of ℓ (also referred as the score statistic) given in (1.24). Then we have

$$\frac{\partial\ell}{\partial\alpha_j}=-\frac{\partial\ell_1}{\partial\alpha_j}+n_j,\quad \frac{\partial\ell}{\partial\beta_j}=-\frac{\partial\ell_1}{\partial\beta_j}+\sum_{i=1}^{n_j}h(x_{ji}). \tag{2.29}$$

This shows that the sum of the derivatives of ℓ and $-\ell_1$ are totally independent of the parameter $(\boldsymbol{\alpha},\boldsymbol{\beta})$. Also, $\partial^2\ell/\partial\boldsymbol{\theta}^2=-\partial\ell_1^2/\partial\boldsymbol{\theta}^2$, and $\partial\ell_1^2/\partial\boldsymbol{\theta}^2$ is positive-definite provided that the model (1.11) is not degenerate.

As we showed before, the expectation of the score statistic $\mathrm{E}\partial\ell/\partial\boldsymbol{\theta}=0$. From (2.29) we have

$$\mathrm{E}(\frac{\partial\ell_1}{\partial\alpha_j}) = n_j,$$

$$\mathrm{E}(\frac{\partial \ell_1}{\partial \beta_j}) \;=\; \mathrm{E}\sum_{i=1}^{n_j} h(x_{ji}) = n_j \mathrm{E}^j h(t),$$

where E^j is the expectation with respect to the jth sample under the true parameter $\boldsymbol{\theta} = \boldsymbol{\theta}_0$; that is, $\mathrm{E}^j f = \int f(t) w_j(t) dG(t)$.

The strong law of large numbers applied to each of the $m+1$ samples, along with the convergence $n_i/n \to n_0 \rho_i$, implies the almost sure convergence

$$\begin{aligned}
\frac{1}{n}\frac{\partial \ell_1}{\partial \alpha_j} &= \frac{1}{n}\sum_{i=1}^{n} \frac{\rho_j w_k(t_i)}{\sum_{k=0}^m \rho_k w_k(t_i)} \\
&= \frac{n_0}{n}\sum_{u=0}^{m} \frac{n_u}{n_0}\frac{1}{n_u}\sum_{v=1}^{n_u} \frac{\rho_j w_j(x_{uv})}{\sum_{k=0}^m \rho_k w_k(x_{uv})} \\
&\to \frac{1}{\sum_{k=0}^m \rho_k}\sum_{u=1}^{m} \rho_u \mathrm{E}^u \frac{\rho_j w_j(t)}{\sum_{k=0}^m \rho_k w_k(t)} \\
&= \mathrm{E}\left[\frac{1}{n}\frac{\partial \ell_1}{\partial \alpha_j}\right] = \frac{\rho_j}{\sum_{k=0}^m \rho_k}, \\
\frac{1}{n}\frac{\partial \ell_1}{\partial \beta_j} &= \frac{1}{n}\sum_{i=1}^{n} \frac{\rho_j w_k(t_i) h(t_i)}{\sum_{k=0}^m \rho_k w_k(t_i)} \\
&\to \frac{1}{\sum_{k=0}^m \rho_k}\sum_{u=1}^{m} \rho_u \mathrm{E}^u \frac{\rho_j w_j(t) h(t)}{\sum_{k=0}^m \rho_k w_k(t)} \\
&= \mathrm{E}\left[\frac{1}{n}\frac{\partial \ell_1}{\partial \beta_j}\right] = \frac{\rho_j \mathrm{E}^j h(t)}{\sum_{k=0}^m \rho_k}.
\end{aligned}$$

Therefore,

$$\frac{1}{n}\frac{\partial \ell_1(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} \xrightarrow{a.s.} \frac{1}{\sum_{k=0}^m \rho_k}(\rho_1, \ldots, \rho_m, \rho_1 \mathrm{E}^1 h'(t), \ldots, \rho_m \mathrm{E}^m h'(t))'. \tag{2.30}$$

Notice that

$$\begin{aligned}
\ell_2 &= \sum_{j=1}^{n_1}(\alpha_1 + \beta_1' h(x_{1j})) + \cdots + \sum_{j=1}^{n_m}(\alpha_m + \beta_m' h(x_{mj})) \\
&= \sum_{i=1}^{m} n_i \alpha_i + \sum_{j=1}^{n_1} \beta_1' h(x_{1j}) + \cdots + \sum_{j=1}^{n_m} \beta_m' h(x_{mj}),
\end{aligned}$$

and, therefore,

$$\frac{\partial \ell_2(\boldsymbol{\alpha}, \boldsymbol{\beta})}{\partial(\boldsymbol{\alpha}, \boldsymbol{\beta})} = (n_1, \ldots, n_m, \sum_{j=1}^{n_1} h'(x_{1j}), \ldots, \sum_{j=1}^{n_m} h'(x_{mj}))', \tag{2.31}$$

which is a vector independent of $(\boldsymbol{\alpha}, \boldsymbol{\beta})$. Also, ℓ_2 is a linear combination of the parameter vectors $(\boldsymbol{\alpha}, \boldsymbol{\beta})$. Therefore,

$$\ell_2 = (\boldsymbol{\alpha}', \boldsymbol{\beta}') \frac{\partial \ell_2(\boldsymbol{\alpha}, \boldsymbol{\beta})}{\partial(\boldsymbol{\alpha}, \boldsymbol{\beta})}. \tag{2.32}$$

The strong law of large numbers and the convergence $n_i/n \to n_0 \rho_i$ imply the almost sure convergence

$$\frac{1}{n} \frac{\partial \ell_2(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0)}{\partial(\boldsymbol{\alpha}, \boldsymbol{\beta})} \to \frac{1}{\sum_{k=0}^{m} \rho_k} (\rho_1, \ldots, \rho_m, \rho_1 \mathrm{E}^1 h'(t), \ldots, \rho_m \mathrm{E}^m h'(t))' \tag{2.33}$$

for the true parameter $(\boldsymbol{\alpha}, \boldsymbol{\beta}) = (\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0)$.

Therefore, from (2.30) and (2.33),

$$\frac{1}{n} \left| \frac{\partial \ell_2(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0)}{\partial(\boldsymbol{\alpha}, \boldsymbol{\beta})} - \frac{\partial \ell_1(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0)}{\partial(\boldsymbol{\alpha}, \boldsymbol{\beta})} \right| \xrightarrow{a.s.} 0. \tag{2.34}$$

Next we prove the strong consistency of $\hat{\boldsymbol{\theta}}$ as an estimator of $\boldsymbol{\theta}_0$. Since $\ell(\boldsymbol{\theta})$ is continuous and differentiable, it takes a maximum on the closed sphere $|\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0| \leq \varepsilon$ around $\boldsymbol{\theta}_0$ for any $\varepsilon > 0$. If we can show that this maximum does not occur on the boundary with probability arbitrarily close to one, then we have a local maximum $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}_n$, within the sphere, for which $\partial \ell(\hat{\boldsymbol{\theta}}_n)/\partial(\boldsymbol{\theta}) = 0$. Let $\varepsilon \to 0$ as n goes to infinity. Then we have a consistent sequence $\hat{\boldsymbol{\theta}}_n$, which converges to $\boldsymbol{\theta}_0$.

First apply Taylor expansion to $n^{-1}\ell_1(\tilde{\boldsymbol{\theta}})$ at $\boldsymbol{\theta}_0$:

$$\frac{1}{n}\ell_1(\tilde{\boldsymbol{\theta}}) = \frac{1}{n}\ell_1(\boldsymbol{\theta}_0) + \left(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)' \frac{1}{n} \frac{\partial \ell_1(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} + \left(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)' \frac{1}{n} \frac{\partial^2 \ell_1(\boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}^2} \left(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right),$$

where $\boldsymbol{\theta}^*$ lies between $\boldsymbol{\theta}_0$ and $\tilde{\boldsymbol{\theta}}$. Noticing that $\partial^2\ell_1(\boldsymbol{\theta}^*)/\partial\boldsymbol{\theta}^2$ is positive definite, it follows that

$$\left(\boldsymbol{\theta}_0'\frac{1}{n}\frac{\partial\ell_1(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}} - \frac{1}{n}\ell_1(\boldsymbol{\theta}_0)\right) - \left(\tilde{\boldsymbol{\theta}}'\frac{1}{n}\frac{\partial\ell_1(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}} - \frac{1}{n}\ell_1(\tilde{\boldsymbol{\theta}})\right) > 0. \tag{2.35}$$

Expression (2.34) shows that for sufficiently large n, $n^{-1}\partial\ell_2(\boldsymbol{\theta}_0)/\partial\boldsymbol{\theta}$ is close to $n^{-1}\partial\ell_1(\boldsymbol{\theta}_0)/\partial\boldsymbol{\theta}$ almost surely. Thus, we can replace $n^{-1}\partial\ell_1(\boldsymbol{\theta}_0)/\partial\boldsymbol{\theta}$ in (2.35) by $n^{-1}\partial\ell_2(\boldsymbol{\theta}_0)/\partial\boldsymbol{\theta}$. Equation (2.31) indicates that $\partial\ell_2/\partial\boldsymbol{\theta}$ is independent of $\boldsymbol{\theta}$. Therefore, we have, for sufficiently large n,

$$\begin{aligned}
0 &< \left(\boldsymbol{\theta}_0'\frac{1}{n}\frac{\partial\ell_2(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}} - \frac{1}{n}\ell_1(\boldsymbol{\theta}_0)\right) - \left(\tilde{\boldsymbol{\theta}}'\frac{1}{n}\frac{\partial\ell_2(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}} - \frac{1}{n}\ell_1(\tilde{\boldsymbol{\theta}})\right) \\
&= \left(\frac{1}{n}\boldsymbol{\theta}_0'\frac{\partial\ell_2(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}} - \frac{1}{n}\ell_1(\boldsymbol{\theta}_0)\right) - \left(\frac{1}{n}\tilde{\boldsymbol{\theta}}'\frac{\partial\ell_2(\tilde{\boldsymbol{\theta}})}{\partial\boldsymbol{\theta}} - \frac{1}{n}\ell_1(\tilde{\boldsymbol{\theta}})\right) \\
&= \left(\frac{1}{n}\ell_2(\boldsymbol{\theta}_0) - \frac{1}{n}\ell_1(\boldsymbol{\theta}_0)\right) - \left(\frac{1}{n}\ell_2(\tilde{\boldsymbol{\theta}}) - \frac{1}{n}\ell_1(\tilde{\boldsymbol{\theta}})\right) \\
&= \frac{1}{n}\ell(\boldsymbol{\theta}_0) - \frac{1}{n}\ell(\tilde{\boldsymbol{\theta}})
\end{aligned}$$

with probability one. The equality on the third line is obtained from (2.32), which shows that ℓ_2 is a linear functional of $\partial\ell_2(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$. Therefore, the local maximum of ℓ does not occur on the boundary with probability arbitrarily close to one. This completes the proof of consistency.

For part (b), since $\hat{\boldsymbol{\theta}}$ is strongly consistent as an estimator of $\boldsymbol{\theta}_0$, we expand $\partial\ell(\hat{\boldsymbol{\theta}})/\partial\boldsymbol{\theta}$ at $\boldsymbol{\theta}_0$,

$$0 = \frac{\partial\ell(\hat{\boldsymbol{\theta}})}{\partial\boldsymbol{\theta}} = \frac{\partial\ell(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}} + \frac{\partial^2\ell(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}^2}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) + o_p(\delta_n),$$

where $\delta_n = |\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0| \to 0$, as $n \to \infty$. Since for sufficiently large n, $-\frac{1}{n}S_n = S + o_p(1)$ by the law of large numbers, and $n^{-1/2}\partial\ell(\boldsymbol{\theta}_0)/\partial\boldsymbol{\theta} = O_p(1)$ by the central limit

theorem, it follows that

$$\begin{aligned}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) &= -S_n^{-1}\frac{\partial \ell(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} + -S_n^{-1}o_p(\delta_n) \\ &= (\frac{1}{n}S^{-1} + o_p(1/n))\frac{\partial \ell(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} + (\frac{1}{n}S^{-1} + o_p(1/n))o_p(\delta_n) \\ &= \frac{1}{n}S^{-1}\frac{\partial \ell(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} + o_p(1/\sqrt{n}),\end{aligned}$$

From (2.21), we have

$$\Lambda = \mathrm{Var}\left(\frac{1}{\sqrt{n}}\frac{\partial \ell(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}}\right).$$

By the central limit theorem and the fact that $\mathrm{E}\left(\partial \ell(\boldsymbol{\theta}_0)/\partial \boldsymbol{\theta}\right) = 0$,

$$\frac{1}{\sqrt{n}}\frac{\partial \ell(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} \xrightarrow{d} N_{(m+1)p}(\mathbf{0}, \Lambda). \tag{2.36}$$

The fact that $-(1/n)S_n \to S$, along with Slutsky's theorem and Lemma 2.1, gives

$$\begin{aligned}\sqrt{n}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) &= S^{-1}\frac{1}{\sqrt{n}}\frac{\partial \ell(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} + o_p(1) \\ &\xrightarrow{d} N_{(p+1)m}(\mathbf{0}, S^{-1}\Lambda S^{-1}) \\ &= N_{(p+1)m}(\mathbf{0}, \Sigma).\end{aligned}$$

This completes the proof of Theorem 2.1. □

The result of this theorem indicates that the asymptotic distribution $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ has mean zero and variance matrix S_{22}^{-1}. This gives a very convenient procedure for investigating $\hat{\boldsymbol{\beta}}$, and for testing hypothesis about $\boldsymbol{\beta}_0$.

Chapter 3

Asymptotic Theory for $\hat{G}$

The multi-sample semiparametric density ratio model given in (1.11) is constructed by multiplicative exponential distortions of the reference distribution. Distortion functions are assumed to be nonnegative and of a known finite-dimensional parametric form, and the reference distribution G is left unknown. In the preceding chapter, we have obtained a strongly consistent constrained maximum likelihood estimator for $(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0)$, and established the corresponding asymptotic normality theory. In this chapter, we will investigate the large sample behavior of $\hat{G}$ given by (1.26) as an estimator for G.

This chapter provides the proof of the weak convergence of the stochastic process $\sqrt{n}(\hat{G}(t) - G(t))$ to a Gaussian process. First express $\sqrt{n}(\hat{G}(t) - G(t))$ as a sum of two components $\sqrt{n}(\hat{G}(t) - \tilde{G}(t)) + \sqrt{n}(\tilde{G}(t) - G(t))$, where $\tilde{G}(t)$ is the empirical distribution of the reference sample X_0 only. The asymptotic properties of $\sqrt{n}(\tilde{G}(t) - G(t))$ are well-known from empirical process considerations. Therefore, the goal is to prove the weak convergence of $\sqrt{n}(\hat{G}(t) - \tilde{G}(t))$. By the strong consistency of $\hat{\boldsymbol{\theta}}$ from Theorem 2.1, a Taylor expansion of $\hat{G}(t)$ at the true parameter

$\boldsymbol{\theta}_0$ approximates $\hat{G}(t)$ uniformly in t. The approximation $H_1(t) - H_2(t)$ is given in Lemma 3.4. Hence the asymptotic properties of $\sqrt{n}(\hat{G}(t) - \tilde{G}(t))$ are equivalent to those of $\sqrt{n}(H_1(t) - H_2(t) - \tilde{G}(t))$, which involves only the true parameter $\boldsymbol{\theta}_0$.

The first step is to prove the joint weak convergence of the finite-dimensional distributions of $\sqrt{n}(H_1(t) - H_2(t) - \tilde{G}(t))$. This can be easily achieved by the multivariate central limit theorem after obtaining the variance-covariance structure. The finite-dimensional convergence is provided in Lemma 3.5, and the variance-covariance structure is given by (3.22).

The second step is to prove the tightness of $\sqrt{n}(H_1(t) - H_2(t) - \tilde{G}(t))$. Calculations show that both $\sqrt{n}(H_1(t) - \tilde{G}(t))$ and $\sqrt{n}H_2(t)$ can be decomposed into sums of empirical processes. Each empirical process is evaluated at a function $f(\cdot)$ in a Donsker class; that is, $P_n f = n^{-1} \sum_{i=1}^{n} f(T_i)$, where $P_n = n^{-1} \sum_{i=1}^{n} \delta_{T_i}$ is an empirical measure defined on i.i.d. observations $T_1, \ldots, T_n$. Therefore, the weak convergence of each empirical process follows from the classical Donsker theory. The tightness of $\sqrt{n}(H_1(t) - H_2(t) - \tilde{G}(t))$ also follows from Lemma 3.6 and Lemma 3.7, and the weak convergence of $\sqrt{n}(\hat{G}(t) - G(t))$ is stated in Theorem 3.9.

This chapter is organized as follows. The first section introduces some basic concepts in weak convergence. The second section reviews the classical results from empirical processes. Both of these two sections provide important results which will be applied in the proof of weak convergence of $\sqrt{n}(\hat{G}(t) - G(t))$. The last section gives the complete proof.

3.1 Weak Convergence of Stochastic Processes

In this section, we introduce the basic concepts behind weak convergence that will be applied later to prove asymptotic results. Before we proceed to weak convergence, we need some topological concepts to specify the structure of metric or topological spaces.

Compactness: A subset of a topological space is *compact* if each of its open covers has a finite subcover. A subset of Euclidean space R^p is called compact if it is closed and bounded. In Euclidean space R^p, every sequence in a compact set has a convergent subsequence, the limit point of which belongs to the set. A *relatively compact* set is a subset whose closure is compact. In metric spaces, every sequence in a relatively compact subset has a convergent subsequence but the limit may not be in the set.

Separability: A topological space is called *separable* if it contains a countable dense subset; that is, a set with a countable number of elements whose closure is the entire space. Obviously, the real line R is separable since the rational numbers form a countable dense subset. More generally, Euclidean space R^p is separable, as the set of all points with rational coordinates is dense. Another example is the space $C[0,1]$ of continuous functions on the unit interval $[0,1]$ with the supremum metric, which has a dense subset of polynomials with rational coefficients (this is the Weierstrass approximation theorem).

Completeness: A metric space $(\mathbf{S}, d)$ is said to be *complete* (or Cauchy) if

every Cauchy sequence of points in $\mathbf{S}$ has a limit that is also in $\mathbf{S}$. It is easy to see that every compact metric space is complete. In fact, a metric space is compact if and only if it is complete and totally bounded.

Let $(\mathbf{S}, d)$ be a metric space with the metric d, equipped with topology $\mathcal{S}$, where $\mathcal{S}$ is the Borel σ-field on $\mathbf{S}$, the smallest σ-field containing the open sets, and let P_n and P be Borel probability measures on $(\mathbf{S}, \mathcal{S})$. We say the sequence P_n converges weakly to P, if and only if

$$\int_{\mathcal{S}} f dP_n \to \int_{\mathbf{S}} f dP, \quad \text{for all } f \in C_b(\mathbf{S}),$$

where $C_b(\mathbf{S})$ denotes the set of all bounded, continuous, real functions on $\mathbf{S}$. Weak convergence is denoted by

$$P_n \xrightarrow{d} P.$$

Equivalently, if X_n and X are $\mathbf{S}$-valued random variables with distributions P_n and P respectively, then X_n weakly converges to X if and only if

$$\mathrm{E}f(X_n) \to \mathrm{E}f(X), \quad \text{for all } f \in C_b(\mathbf{S}). \tag{3.1}$$

A thorough investigation of the classical theory of weak convergence based on these definitions can be found in Billingsley (1999).

The classical theory requires that P_n be defined on the Borel σ-field $\mathcal{S}$ for each n, or equivalently, that X_n is a Borel measure map for each n. Provided that $(\Omega_n, \mathcal{A}_n, P_n)$ and $(\Omega, \mathcal{A}, P)$ are the underlying probability spaces respectively, this means that $X_n^{-1}(D) \in \mathcal{A}_n$ for every Borel set $D \in \mathcal{S}$. This required measurability

usually holds when $\mathbf{S}$ is a separable metric space such as R^p or $C[0,1]$ with supremum metric. However, this requirement can easily fail when the metric space $\mathbf{S}$ is not separable. For example, this happens when $\mathbf{S}$ is the *Skorohod* space $D[0,1]$ of all the right continuous functions on $[0,1]$ with left limits, equipped with the metric induced by the supremum norm $\|P(f)\|_{D[0,1]} = \sup_{f \in D[0,1]} \|P(f)\|$.

This difficulty arises in empirical processes. Suppose that i.i.d. random variables $X_1, \ldots, X_n$ are defined as the coordinate projections on the product probability space $([0,1], \mathcal{A}, \lambda)^n$, where λ denotes the Lebesgue measure on $[0,1]$ and $\mathcal{A}$ the Borel σ-field. The empirical distribution function F_n is the random function

$$F_n(t) = \frac{1}{n} \sum_{i=1}^{n} I_{[0,t]}(X_i),\ 0 \leq t \leq 1,$$

and the uniform emiprical process is

$$G_n = \sqrt{n}(F_n(t) - t),\ 0 \leq t \leq 1.$$

Both F_n and G_n are maps from $[0,1]^n$ into $D[0,1]$, but neither of them is a Borel measurable map if $D[0,1]$ is endowed with the supremum norm.

To deal with this problem, Skorohod (1956) and Billingsley (1999) endowed $D[0,1]$ with the Skorohod metric under which $D[0,1]$ is separable and the classical theory could be applied without difficulty. Another idea is to drop the requirement of Borel measurability of X_n and meanwhile requiring (3.1), where the expectations are to be interpreted as *outer expectations* (Van der Vaart and Wellner (1996) Section 1.2). This idea is used to formulate Donsker's theorem in Van der Vaart and Wellner (1996). However, in this section, we still follow the procedure in Billingsley (1999) and Ethier and Kurtz (1985) by endowing $D[0,1]$ with the Skorohod metric

to formulate the criteria for weak convergence.

Let $\mathcal{P}(\mathbf{S})$ denote the family of Borel probability measures on $\mathbf{S}$. We topologize $\mathcal{P}(\mathbf{S})$ with the Prohorov metric

$$d_p(P,Q) = \inf\{\varepsilon > 0 : P(F) \leq Q(F^\varepsilon) + \varepsilon; \quad \text{for all } F \in \mathcal{C}\},$$

where $\mathcal{C}$ is a collection of closed subsets of $\mathcal{S}$ and

$$F^\varepsilon = \{x \in \mathbf{S} : \inf_{y \in F} d(x,y) < \varepsilon\}.$$

In Ethier and Kurtz(1985) Section 3.1 it is proved that d_p is a metric. Provided that $\mathbf{S}$ is separable, $d_p(P_n, P) \to 0$ is equivalent to $P_n \xrightarrow{d} P$.

Theorem 3.1 *(Ethier and Kurtz(1985), Theorem 1.7) If* $\mathbf{S}$ *is separable, then* $\mathcal{P}(\mathbf{S})$ *is separable. If in addition* $(\mathbf{S}, d)$ *is complete, then* $(\mathcal{P}(\mathbf{S}), d_p)$ *is complete.*

A probability measure P is said to be ***tight*** if for each $\varepsilon > 0$ there exists a compact set $K \subset \mathcal{S}$ such that $P(K) \geq 1 - \varepsilon$. A family of probability measures $\mathcal{M} \subset \mathcal{P}(\mathbf{S})$ is *tight* if for each $\varepsilon > 0$ there exists a compact set $K \subset \mathcal{S}$ such that $\inf_{P \in \mathcal{M}} P(K) \geq 1 - \varepsilon$.

Lemma 3.1 *(Ethier and Kurtz(1985), Lemma 2.1) If* $(\mathbf{S}, d)$ *is complete and separable, then each* $P \in \mathcal{P}(\mathbf{S})$ *is tight.*

The following theorem gives the connection between relative compactness and tightness, which is usually called Prohorov's theorem.

Theorem 3.2 *(Ethier and Kurtz(1985) Theorem 2.2) Let* $(\mathbf{S}, d)$ *be complete and separable, and let* $\mathcal{M} \subset \mathcal{P}(\mathbf{S})$*. Then the following are equivalent:*

(a) $\mathcal{M}$ *is tight.*

(b) $\mathcal{M}$ *is relatively compact.*

From Theorem 3.2, we have the following corollary.

Corollary 3.1 *Let* $(\mathbf{S}, d)$ *be complete and separable, and suppose that* $\{P_n\}$ *and* P *belong to* $\mathcal{P}(\mathbf{S})$. *If* $P_n \xrightarrow{d} P$, *then* $\{P_n\}$ *is tight.*

Proof. Since $d_p(P_n, P) \to 0$ is equivalent to $P_n \xrightarrow{d} P$. Then $P_n \xrightarrow{d} P$ implies $d_p(P_n, P) \to 0$. Therefore, $\{P_n\}$ is relatively compact. By Theorem 3.2, it follows that $\{P_n\}$ is tight. □

Now we consider the space $D[0, \infty)$ of real functions f on $[0, \infty)$ that are right continuous and have left limits: for $t > 0$,

$$\begin{aligned} f(t+) &= \lim_{s \downarrow t} f(s) \text{ exists and } f(t+) = f(t), \\ f(t-) &= \lim_{s \uparrow t} f(s) \text{ exists.} \end{aligned}$$

Functions having these two properties are called *cadlag* functions. Most stochastic processes arising in applications have the property that almost every sample path is a *cadlag* function. For example, classical empirical processes and Poisson processes have sample paths which are *cadlag* functions.

Theorem 3.3 *(Ethier and Kurtz(1985), Theorem5.6) The space* $D[0, \infty)$ *endowed with the Skorohod metric* d_s *is complete and separable.*

Skorohod metric is introduced both in Section 3.5 of Ethier and Kurtz(1985) and Section 12 of Billingsley(1999). Replacing $(\mathbf{S}, d)$ by $(D[0, \infty), d_s)$, Theorems 3.1 and 3.2 still hold.

The next theorem provides an important criterion for judging weak convergence. It is equivalent to the Theorem 7.8 in Ethier and Kurtz (1985) and Theorem 13.1 in Billingsley (1999).

Theorem 3.4 *Let $Y_1, \ldots, Y_n$ and Y be stochastic processes with sample paths in $D[0,\infty)$. If $\{Y_n\}$ is tight and there exists a dense set $D \subset [0,\infty)$ such that*

$$(Y_n(t_1), \ldots, Y_n(t_k)) \xrightarrow{d} (Y(t_1), \ldots, Y(t_k)) \tag{3.2}$$

for every finite set $\{t_1, \ldots, t_k\} \subset D$, then Y_n converges weakly to Y.

The finite dimensional convergence (3.2) gives the structure of the limit process, and the tightness guarantee the existence of a convergent subsequence. Given the sequence of processes $\{Y_n\}$, usually the finite dimensional convergence can be checked directly. To prove tightness, a criterion listed in Chapter 3 of Billingsley (1999) may be applied. But we mainly use the the result from Corollary 3.1.

3.2 A Brief Review of Empirical Process Theory

Before we proceed to describe the asymptotic properties of $\hat{G}$, we first introduce the classical asymptotic results for empirical processes. Let $T_1, \ldots, T_n$ be a real-valued random sample from a distribution function F. The empirical distribution function is defined as

$$F_n(t) = \frac{1}{n} \sum_{i=1}^{n} I_{[T_i<t]}. \tag{3.3}$$

This is a natural estimator for the underlying distribution F. Because $nF_n(t)$ is binomially distributed with mean $nF(t)$, this estimator is unbiased. Furthermore,

it is also consistent by the law of large numbers:

$$F_n(t) \stackrel{a.s.}{\to} F(t), \quad \text{for every } t.$$

By the central limit theorem, it is asymptotically normally distributed with mean 0 and variance $F(t)(1-F(t))$:

$$\sqrt{n}(F_n(t) - F(t)) \stackrel{d}{\to} N(0, F(t)(1-F(t))).$$

These results are obtained by considering $F_n(t)$ as an estimator for each t separately. On the other hand, the empirical distribution of a random sample is the uniform discrete measure on the observations. To study the convergence of $\sqrt{n}(F_n(t) - F(t))$, we have to consider $F_n(t)$ as a random function $F_n(t, \omega)$, and $F_n(\cdot, \omega)$ is a sample path. This leads to the law of large numbers and central limit theorems that are uniform in classes of functions.

We define the following uniform distance

$$|F_n - F|_\infty = \sup_t |F_n(t) - F(t)|,$$

which is known as the *Kolmogorov-Smirnov statistic*. The following *Glivenko-Cantelli theorem* extends the law of large numbers to random functions and gives uniform convergence.

Theorem 3.5 (*Glivenko − Cantelli*) *If $T_1, \ldots, T_n$ are i.i.d. random variables with distribution F, then $\| F_n - F \|_\infty \xrightarrow{a.s.} 0$.*

This means that the sample paths of $F_n(t)$ get uniformly closer to F as n increases; hence F_n, which we observe, approximates F better and better as we collect more observations.

By the multivariate central limit theorem, we have the weak convergence of the joint finite dimensional distributions:

$$\sqrt{n}(F_n(T_{i_1}) - F(T_{i_1}), \ldots, F_n(T_{i_m}) - F(T_{i_m})) \xrightarrow{d} (\mathcal{G}_F(T_{i_1}), \ldots, \mathcal{G}_F(T_{i_m})),$$

where $(i_1, \ldots, i_m) \subset (1, \ldots, n)$, and the vector on the right hand side has a multivariate normal distribution with mean 0 and covariances

$$\mathrm{E}\mathcal{G}_F(T_{i_u})\mathcal{G}_F(T_{i_v}) = F(T_{i_u} \wedge T_{i_v}) - F(T_{i_u})F(T_{i_v}). \tag{3.4}$$

This suggest that the sequence of empirical processes $\sqrt{n}(F_n - F)$ converges in distribution to a Gaussian process $\mathcal{G}_F$ with mean 0 and covariances (3.4). According to an extension of *Donsker's theorem*, this is true in the sense of weak convergence of these processes in the Skorohod space $D[-\infty, \infty)$ equipped with the uniform norm.

Theorem 3.6 (*Donsker*) *If $T_1, \ldots, T_n$ are i.i.d. random variables with distribution F, then the sequence of empirical processes converges weakly*

$$\sqrt{n}(F_n - F) \xrightarrow{d} \mathcal{G}_F$$

in the Skorohod space $D[-\infty, \infty]$. The Gaussian process $\mathcal{G}_F$ has mean 0 and covariances (3.4).

The limit process $\mathcal{G}_F$ is known as the *F-Brownian bridge* process, and it is a standard Brownian bridge if F is the uniform distribution on $[0, 1]$. Obviously, the F-Brownian bridge can be obtained as $\mathcal{G}_0 \circ F$ from a standard Brownian bridge $\mathcal{G}_0$. The name "bridge" results from the fact that the sample paths of the process are pinned at zero at the endpoints $-\infty$ and ∞. Therefore, the Brownian bridge is not

a process with independent increments, which is true for Brownian motion.

Modern empirical process theory views the empirical measure as a stochastic process indexed by a large class of functions $\mathcal{F}$. To avoid problems of measurability, the following convergence in distribution is defined using *outer expectation* as in Van der Vaart and Wellner (1996). Suppose independent and identically distributed random variables $T_1, \ldots, T_n$ are from a probability distribution P on a measurable space $(\mathcal{X}, \mathcal{A})$. We define the empirical measure of a sample of random elements $T_1, \ldots, T_n$ as

$$P_n = n^{-1} \sum_{i=1}^{n} \delta_{T_i},$$

where δ_x is the Dirac measure at x. Given a measurable function $f : \mathcal{X} \mapsto \mathcal{R}$, let $P_n f$ be the expectation of f under the empirical measure P_n, and Pf be the expectation of f under P. Thus

$$P_n f = \frac{1}{n} \sum_{i=1}^{n} f(T_i), \quad Pf = \int f dP. \tag{3.5}$$

Based on this definition, the resulting stochastic process, as $f \in \mathcal{F}$ varies, is just $\{P_n f : f \in \mathcal{F}\}$. Actually, (3.3) is a realization of (3.5) when f is the indicator function $I_{(-\infty,t]}$.

Let $\mathcal{F}$ be a class of measurable functions $f : \mathcal{X} \mapsto \mathcal{R}$. Under the supremum norm $\|P\|_{\mathcal{F}} = \sup_{f \in \mathcal{F}} \|Pf\|$, if

$$\|P_n - P\|_{\mathcal{F}} \to 0,$$

where the convergence is in outer probability almost surely, then we call $\mathcal{F}$ a

Glivenko-Cantelli *class*, or also a *P-Glivenko-Cantelli* class, to point out the dependence on the underlying measure P.

Assume

$$\sup_{f\in\mathcal{F}} \|f(x) - Pf\| < \infty, \quad \text{for every } x.$$

Under this condition, the empirical process $\{P_n f : f \in \mathcal{F}\}$ can be viewed as a map into $\ell^\infty(\mathcal{F})$, where $\ell^\infty(\mathcal{F})$ is a set of uniformly bounded real functions on $\mathcal{F}$. Under this setup, we wish to investigate the convergence

$$G_n = \sqrt{n}(P_n - P) \xrightarrow{d} G, \quad \text{in } \ell^\infty(\mathcal{F}), \tag{3.6}$$

where G is a tight Borel measurable element in $\ell^\infty(\mathcal{F})$. A class $\mathcal{F}$ is called *P-**Donsker*** if (3.6) holds. It is known that the collection of all indicator functions of lower rectangles $\{I_{(-\infty,t]} : t \in \bar{\mathcal{R}}^p\}$ is Donsker for any underlying law of i.i.d. random variables $T_1, \ldots, T_n$ in $\bar{\mathcal{R}}^p$, where $\bar{\mathcal{R}}$ is the extended real line $[-\infty, \infty]$.

The structure of the limit process G follows from the convergence of the finite dimensional distributions. The multivariate central limit theorem gives that for any finite set $f_1, \ldots, f_k$,

$$(G_n f_1, \ldots, G_n f_k) \xrightarrow{d} N_k(0, \Sigma),$$

where the (i, j)th element of the covariance matrix Σ is $P(f_i - Pf_i)(f_j - Pf_j)$. It follows that the limit process $Gf : f \in \mathcal{F}$ must be a zero-mean Gaussian process with covariance function

$$\mathrm{E}Gf_1Gf_2 = Pf_1f_2 - Pf_1Pf_2.$$

It is also called the P-Brownian bridge. The result in Theorem 3.6 is a special case

when $\mathcal{F}$ is a collection of the indicator functions $I_{(-\infty,t]}$.

Whether a given class $\mathcal{F}$ is a Glivenko-Cantelli or Donsker class depends on the size of the class. A relatively simple way to measure the size of a class is to use *entropy numbers*. Consider $\mathcal{F}$ as a subset of a metric space (such as $(L_r(P), \|\cdot\|_{P,r})$, the space with L_r-norm). The *covering number* $N(\varepsilon, \mathcal{F}, \|\cdot\|)$ is the minimal number of balls $\{g : \|g - f\| < \varepsilon\}$ of radius ε needed to cover the set $\mathcal{F}$. The *entropy (without bracketing)* is the logarithm of the covering number. Given two functions l and u, the *bracket* $[l, u]$ is the set of all functions f with $l \leq f \leq u$. An *ε-bracket* is a bracket $[l, u]$ with $\|u - l\| < \varepsilon$. The *bracketing number* $N_{[]}(\varepsilon, \mathcal{F}, \|\cdot\|)$ is the minimum number of ε-brackets needed to cover $\mathcal{F}$. The *entropy with bracketing* is the logarithm of the bracketing number.

Under the $L_r(P)$-norm

$$\|f\|_{P,r} = (\int |f|^r dP)^{1/r},$$

the covering and bracketing numbers are related by

$$N(\varepsilon, \mathcal{F}, L_r(P)) \leq N_{[]}(2\varepsilon, \mathcal{F}, L_r(P)).$$

An *envelope function* of a class $\mathcal{F}$ is any function $x \mapsto F(x)$ such that $|f(x)| \leq F(x)$, for every x and f. The *uniform entropy numbers* (relative to L_r) are defined as

$$\sup_P \log N(\varepsilon \|F\|_{P,r}, \mathcal{F}, L_r(P)),$$

where the supremum is over all probability measures P on $(\mathcal{X}, \mathcal{A})$ with $0 < PF^r < \infty$.

Now a heuristic summary of the two types of limit theorems below is as follows:

SLLN or Glivenko-Cantelli theorem: Let $\mathcal{F}$ be a class of measurable functions such that $N_{[]}(\varepsilon, \mathcal{F}, L_1(P)) < \infty$ for every $\varepsilon > 0$. Then $\mathcal{F}$ is a Glivenko-Cantelli class.

CLT or Donsker theorem: Let $\mathcal{F}$ be a class of measurable functions that satisfies the *uniform entropy bound*

$$\int_0^\infty \sup_P \sqrt{\log N(\varepsilon \|F\|_{P,2}, \mathcal{F}, L_2(P))} d\varepsilon < \infty. \tag{3.7}$$

If $PF^2 < \infty$, then $\mathcal{F}$ is P-Donsker.

To verify the hypothesis on uniform entropy numbers and bracketing numbers, we introduce the notion of *Vapnik-Červonenkis class* of sets, or simply *VC-class*. A collection $\mathcal{C}$ of subsets of the sample space $\mathcal{X}$ *picks out* a certain subset of the finite set $\{x_1, \ldots, x_n\} \subset \mathcal{X}$ if it can be written as $\{x_1, \ldots, x_n\} \cap C$ for some $C \in \mathcal{C}$. The collection $\mathcal{C}$ is said to *shatter* $\{x_1, \ldots, x_n\}$ if $\mathcal{C}$ picks out each of its 2^n subsets. The *VC-index* $V(\mathcal{C})$ of $\mathcal{C}$ is the smallest n for which no set of size n is shattered by $\mathcal{C}$. A collection $\mathcal{C}$ of measurable sets is called a *VC-class* if its index $V(\mathcal{C})$ is finite.

The *subgraph* of a function $f : x \mapsto \mathcal{R}$ is the subset of $\mathcal{X} \times \mathcal{R}$ given by

$$\{(x, t) : t < f(x)\}.$$

A collection $\mathcal{F}$ of measurable functions on a sample space is called a *VC-subgraph class*, if the collection of all subgraphs of the functions in $\mathcal{F}$ forms a VC-class of sets in $\mathcal{X} \times \mathcal{R}$. Let $V(\mathcal{F})$ be the VC-index of the set of subgraphs of functions in $\mathcal{F}$. We have the following theorem (Van der Vaart and Wellner (1996), Theorem 2.6.7).

Theorem 3.7 *For a VC-subgraph class of functions with measurable envelope function F and $r \geq 1$ one has for any probability measure P with $\|F\|_{P,r} > 0$,*

$$N(\varepsilon \|F\|_{P,r}, \mathcal{F}, L_r(P)) \leq KV(\mathcal{F})(16e)^{V(\mathcal{F})}(1/\varepsilon)^{r(V(\mathcal{F})-1)}, \tag{3.8}$$

for a universal constant K and $0 < \varepsilon < 1$.

This theorem shows that a VC-subgraph class satisfies the uniform entropy condition (3.7). This indicates that a collection $\mathcal{F}$ of measurable functions is P-Donsker for any underlying measure P if the envelope function is square integrable and $\mathcal{F}$ is a VC-subgraph class. The next lemma gives a basic method for generating VC-subgraph classes (Van der Vaart and Wellner(1996), Lemma 2.6.15).

Lemma 3.2 *Any finite dimensional vector space $\mathcal{F}$ of measurable functions is a VC-subgraph class of index smaller than or equal to $dim(\mathcal{F}) + 2$.*

The next lemma is useful for proving almost surely uniform convergence.

Lemma 3.3 *Let $T_1, \ldots, T_n$ be i.i.d. random variables with probability measure P and corresponding distribution function F. Suppose that f is a Borel measurable function satisfying $\int |f| dP < \infty$. Let $G_n(t) = 1/n \sum_{i=1}^{n} f(T_i) I_{[T_i \leq t]}$ and $G(t) = \int f(y) I_{[y \leq t]} dF(y)$. Then $\sup_{-\infty \leq t \leq \infty} |G_n(t) - G(t)| \overset{a.s.}{\to} 0$.*

Proof. Let $P_n = (1/n) \sum_{i=1}^{n} \delta_{T_i}$, and let $\mathcal{F}$ be the collection of all indicator functions of the form $I_{(-\infty,t]}$. It is known that $\mathcal{F}$ is P-Donsker. By Example 2.10.10 in Van der Vaart and Wellner(1996), $\mathcal{F} \cdot f$ is also P-Donsker, and consequently P-Glivenko-Cantelli as well. Noticing that $G_n(t) = P_n(f \cdot I_{(-\infty,t]})$ and $G(t) =$

$P(f \cdot I_{(-\infty,t]})$, it follows by the Glivenko-Cantelli theorem that

$$\sup_{-\infty \le t \le \infty} |G_n(t) - G(t)| = \sup_{g \in \mathcal{F} \cdot f} |P_n g - Pg| \stackrel{a.s.}{\to} 0.$$

□

3.3 Asymptotic Distribution for Semiparametric Model

The estimator $\hat{G}$ is a weighted discrete measure on observations. Motivated by the empirical distribution, we are going to develop the asymptotic distribution for $\sqrt{n}(\hat{G} - G)$ in a semiparametric model in a way similar to the nonparametric case. However, unlike the nonparametric case, the estimator $\hat{G}$ in the semiparametric model is related to the parameters $(\boldsymbol{\alpha}, \boldsymbol{\beta})$, and has to be estimated in terms of the estimators of the parameters. This makes the covariance structure of the limit processes more complicated than that in the nonparametric case.

3.3.1 An Approximation of $\hat{G}(t)$

Let $\tilde{G}(t)$ denote the empirical distribution based on the reference sample $X_0 = (x_{01}, \ldots, x_{0n_0})$:

$$\tilde{G}(t) = \frac{1}{n_0} \sum_{i=1}^{n_0} I_{[x_{0i} < t]}. \tag{3.9}$$

The semiparametric estimator for G is

$$\hat{G}(t) = \frac{1}{n_0} \cdot \sum_{i=1}^{n} \frac{I(t_i \le t)}{\sum_{k=0}^{m} \rho_k \exp(\hat{\alpha}_k + \hat{\beta}_k h(t_i))} = \frac{1}{n_0} \cdot \sum_{i=1}^{n} \frac{I(t_i \le t)}{\sum_{k=0}^{m} \rho_k w_k(t_i; \hat{\alpha}_k, \hat{\beta}_k)},$$

where $w_k(t_i;\hat{\alpha}_k,\hat{\beta}_k)=\exp(\hat{\alpha}_k+\hat{\beta}_k h(t_i))$. Define

$$H_1(t;\boldsymbol{\alpha},\boldsymbol{\beta})=\frac{1}{n_0}\cdot\sum_{i=1}^{n}\frac{I(t_i\le t)}{\sum_{k=0}^{m}\rho_k w_k(t_i;\alpha_k,\beta_k)}.$$

For convenience, without further notice we write $H_1(t)=H_1(t;\boldsymbol{\alpha}_0,\boldsymbol{\beta}_0)$. Apparently, $\hat{G}(t)$ is a realization of $H_1(t;\boldsymbol{\alpha},\boldsymbol{\beta})$ at $(\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\beta}})$.

Differentiate $H_1(t;\boldsymbol{\alpha},\boldsymbol{\beta})$ with respect to $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ respectively. Then we have

$$\begin{aligned}\frac{\partial H_1(t;\boldsymbol{\alpha},\boldsymbol{\beta})}{\partial\alpha_j}&=-\frac{1}{n_0}\sum_{i=1}^{n}\frac{\rho_j w_j(t_i)I(t_i\le t)}{(\sum_{k=0}^{m}\rho_k w_k(t_i;\alpha_k,\beta_k))^2}\\ \frac{\partial H_1(t;\boldsymbol{\alpha},\boldsymbol{\beta})}{\partial\beta_j}&=-\frac{1}{n_0}\sum_{i=1}^{n}\frac{\rho_j w_j(t_i)h(t_i)I(t_i\le t)}{(\sum_{k=0}^{m}\rho_k w_k(t_i;\alpha_k,\beta_k))^2},\end{aligned}\tag{3.10}$$

where $j=1,\ldots,m$. Next, take the expectation of the derivatives evaluated at $(\boldsymbol{\alpha}_0,\boldsymbol{\beta}_0)$

$$\begin{aligned}\mathrm{E}\left(\frac{\partial H_1(t;\boldsymbol{\alpha}_0,\boldsymbol{\beta}_0)}{\partial\alpha_j}\right)&=-\frac{1}{n_0}\mathrm{E}\left(\sum_{i=1}^{n}\frac{\rho_j w_j(t_i)I(t_i\le t)}{(\sum_{k=0}^{m}\rho_k w_k(t_i))^2}\right)\\&=-\sum_{i=0}^{m}\frac{n_i}{n_0}\int\frac{\rho_j w_j(y)w_i(y)I(y\le t)}{(\sum_{k=0}^{m}\rho_k w_k(y))^2}dG(y)\\&=-\rho_j\int\frac{w_j(y)I(y\le t)}{\sum_{k=0}^{m}\rho_k w_k(y)}dG(y)\\&=-\rho_j A_j(t),\end{aligned}$$

where $A_j(t)=\int\frac{w_j(y)I(y\le t)}{\sum_{k=0}^{m}\rho_k w_k(y)}dG(y)$.

Similarly,

$$\begin{aligned}\mathrm{E}\left(\frac{\partial H_1(t;\boldsymbol{\alpha}_0,\boldsymbol{\beta}_0)}{\partial\beta_j}\right)&=-\frac{1}{n_0}\mathrm{E}\left(\sum_{i=1}^{n}\frac{\rho_j w_j(t_i)h(t_i)I(t_i\le t)}{(\sum_{k=0}^{m}\rho_k w_k(t_i))^2}\right)\\&=-\rho_j\int\frac{w_j(y)h(y)I(y\le t)}{\sum_{k=0}^{m}\rho_k w_k(y)}dG(y)\\&=-\rho_j B_j(t),\end{aligned}$$

where $B_j(t)=\int\frac{w_j(y)h(y)I(y\le t)}{\sum_{k=0}^{m}\rho_k w_k(y)}dG(y)$.

By the law of large numbers, we have

$$\begin{aligned}\frac{\partial H_1(t;\boldsymbol{\alpha}_0,\boldsymbol{\beta}_0)}{\partial \alpha_j} &\xrightarrow{a.s.} -\rho_j A_j(t)\\ \frac{\partial H_1(t;\boldsymbol{\alpha}_0,\boldsymbol{\beta}_0)}{\partial \beta_j} &\xrightarrow{a.s.} -\rho_j B_j(t), \quad \text{as } n\to\infty.\end{aligned}$$

In fact, the almost sure convergence holds uniformly in t. This can be seen from the boundedness of $w_j(t)/(\sum_{k=0}^m \rho_k w_k(t))$ and the assumption that the second moments of h are bounded with respect to all the sample distributions. We have

$$\begin{aligned}&-\frac{\partial H_1(t;\boldsymbol{\alpha}_0,\boldsymbol{\beta}_0)}{\partial \beta_j} - \rho_j B_j(t)\\ &= \frac{1}{n_0}\sum_{i=1}^{n}\frac{\rho_j w_j(t_i)h(t_i)I(t_i\le t)}{(\sum_{k=0}^m \rho_k w_k(t_i))^2} - \sum_{i=0}^{m}\frac{n_i}{n_0}\int \frac{\rho_j w_j(y)w_i(y)h(y)I(y\le t)}{(\sum_{k=0}^m \rho_k w_k(y))^2}dG(y)\\ &= \rho_j\sum_{i=0}^{m}\left(\frac{1}{n_i}\sum_{l=1}^{n_i}\frac{\rho_i w_j(x_{il})h(x_{il})I(x_{il}\le t)}{(\sum_{k=0}^m \rho_k w_k(x_{il}))^2}\right.\\ &\qquad\left. - \int \frac{\rho_i w_j(y)w_i(y)h(y)I(y\le t)}{(\sum_{k=0}^m \rho_k w_k(y))^2}dG(y)\right). \qquad (3.11)\end{aligned}$$

By applying Lemma 3.3 to the inside of the parentheses of (3.11), we have

$$\begin{aligned}\sup_{-\infty<t<\infty}\left|\frac{\partial H_1(t;\boldsymbol{\alpha}_0,\boldsymbol{\beta}_0)}{\partial \alpha_j} - (-\rho_j A_j(t))\right| &\xrightarrow{a.s.} 0\\ \sup_{-\infty<t<\infty}\left\|\frac{\partial H_1(t;\boldsymbol{\alpha}_0,\boldsymbol{\beta}_0)}{\partial \beta_j} - (-\rho_j B_j(t))\right\| &\xrightarrow{a.s.} 0, \quad j=1,\ldots,m. \qquad (3.12)\end{aligned}$$

Denote

$$\bar{A}(t) = (A_1(t),\ldots,A_m(t))', \quad \bar{B}(t) = (B_1'(t),\ldots,B_m'(t))'.$$

Lemma 3.4 *The function $\hat{G}(t)$ has an approximation uniformly in t,*

$$\hat{G}(t) = H_1(t) - H_2(t) + R_n(t),$$

where $H_1(t)$ is defined as before and

$$H_2(t) = \frac{1}{n}\left(\bar{A}'(t)\boldsymbol{\rho}, \bar{B}'(t)(\boldsymbol{\rho}\otimes I_p)\right)S^{-1}\begin{pmatrix}\frac{\partial \ell(\boldsymbol{\alpha}_0,\boldsymbol{\beta}_0)}{\partial \boldsymbol{\alpha}}\\ \\ \frac{\partial \ell(\boldsymbol{\alpha}_0,\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}}\end{pmatrix}, \qquad (3.13)$$

and the remainder term $R_n(t)$ *satisfies* $\sup_{-\infty<t<\infty}|R_n(t)| = o_p(n^{-1/2})$.

Proof: By the strong consistency of the estimator $(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})$, the Taylor expansion of $\hat{G}(t)$ at $(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0)$ gives

$$\begin{aligned}
\hat{G}(t) &= \frac{1}{n_0}\cdot\sum_{i=1}^{n}\frac{I(t_i\le t)}{\sum_{k=0}^{m}\rho_k w_k(t_i;\hat{\alpha}_k,\hat{\beta}_k)} \\
&= H_1(t;\boldsymbol{\alpha}_0,\boldsymbol{\beta}_0) + \begin{pmatrix} \frac{\partial H_1(t;\boldsymbol{\alpha}_0,\boldsymbol{\beta}_0)}{\partial\boldsymbol{\alpha}} \\ \frac{\partial H_1(t;\boldsymbol{\alpha}_0,\boldsymbol{\beta}_0)}{\partial\boldsymbol{\beta}} \end{pmatrix}' \begin{pmatrix} \hat{\boldsymbol{\alpha}}-\boldsymbol{\alpha}_0 \\ \hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0 \end{pmatrix} + o_p(\delta_n) \\
&= H_1(t) + \begin{pmatrix} \mathrm{E}\frac{\partial H_1(t;\boldsymbol{\alpha}_0,\boldsymbol{\beta}_0)}{\partial\boldsymbol{\alpha}} \\ \mathrm{E}\frac{\partial H_1(t;\boldsymbol{\alpha}_0,\boldsymbol{\beta}_0)}{\partial\boldsymbol{\beta}} \end{pmatrix}' \begin{pmatrix} \hat{\boldsymbol{\alpha}}-\boldsymbol{\alpha}_0 \\ \hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0 \end{pmatrix} + R_{n1}(t) + o_p(\delta_n) \\
&= H_1(t) - \Big(\rho_1 A_1(t),\ldots,\rho_m A_m(t),\rho_1 B_1'(t),\ldots,\rho_m B_m'(t)\Big)\begin{pmatrix} \hat{\boldsymbol{\alpha}}-\boldsymbol{\alpha}_0 \\ \hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0 \end{pmatrix} \\
&\quad + R_{n1}(t) + o_p(\delta_n) \\
&= H_1(t) - \Big(\bar{A}'(t)\boldsymbol{\rho}, \bar{B}'(t)(\boldsymbol{\rho}\otimes I_p)\Big)\begin{pmatrix} \hat{\boldsymbol{\alpha}}-\boldsymbol{\alpha}_0 \\ \hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0 \end{pmatrix} + R_{n1}(t) + o_p(\delta_n) \\
&= H_1(t) - \frac{1}{n}\Big(\bar{A}'(t)\boldsymbol{\rho}, \bar{B}'(t)(\boldsymbol{\rho}\otimes I_p)\Big)S^{-1}\begin{pmatrix} \frac{\partial \ell(\boldsymbol{\alpha}_0,\boldsymbol{\beta}_0)}{\partial\boldsymbol{\alpha}} \\ \frac{\partial \ell(\boldsymbol{\alpha}_0,\boldsymbol{\beta}_0)}{\partial\boldsymbol{\beta}} \end{pmatrix} \\
&\quad + o_p(n^{-1/2}) + R_{n1}(t) + o_p(\delta_n) \\
&= H_1(t) - H_2(t) + R_n(t) \qquad (3.14)
\end{aligned}$$

where $\delta_n = \|(\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\beta}}) - (\boldsymbol{\alpha}_0,\boldsymbol{\beta}_0)\|$ and

$$R_{n1} = \begin{pmatrix} \frac{\partial H_1(t;\boldsymbol{\alpha}_0,\boldsymbol{\beta}_0)}{\partial\boldsymbol{\alpha}} - \mathrm{E}\frac{\partial H_1(t;\boldsymbol{\alpha}_0,\boldsymbol{\beta}_0)}{\partial\boldsymbol{\alpha}} \\ \frac{\partial H_1(t;\boldsymbol{\alpha}_0,\boldsymbol{\beta}_0)}{\partial\boldsymbol{\beta}} - \mathrm{E}\frac{\partial H_1(t;\boldsymbol{\alpha}_0,\boldsymbol{\beta}_0)}{\partial\boldsymbol{\beta}} \end{pmatrix}' \begin{pmatrix} \hat{\boldsymbol{\alpha}}-\boldsymbol{\alpha}_0 \\ \hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0 \end{pmatrix},$$

$$R_n(t) = o_p(n^{-1/2}) + R_{n1}(t) + o_p(\delta_n).$$

The sixth equality of (3.14) follows from Theorem 2.1. By Theorem 2.1 again, it follows that $\delta_n = o_p(n^{-1/2})$. The uniformly almost sure convergence of (3.12) and the part (b) of Theorem 2.1 imply that $\sup_{-\infty<t<\infty} |R_{n1}(t)| = o_p(n^{-1/2})$. As a result, $\sup_{-\infty<t<\infty} |R_n(t)| = o_p(n^{-1/2})$. The proof is complete. □

Therefore, $H_1(t) - H_2(t)$ is an approximation of $\hat{G}$ uniformly in t. In order to derive the asymptotic distribution of $\sqrt{n}(\hat{G} - G)$, we start from $\sqrt{n}(\hat{G} - \tilde{G})$ in that

$$\sqrt{n}(\hat{G} - G) = \sqrt{n}(\hat{G} - \tilde{G}) - \sqrt{n}(\tilde{G} - G),$$

where $\tilde{G}$ is the empirical distribution of the reference sample, and $\sqrt{n}(\tilde{G} - G)$ is the corresponding classical empirical processes. By Lemma 3.4, $\sqrt{n}(\hat{G} - \tilde{G})$ can be approximated by $\sqrt{n}(H_1(t) - H_2(t) - \tilde{G})$.

3.3.2 Variance-covariance Structure

It is well-known that weak convergence of stochastic processes is determined by the convergence of finite-dimensional distributions and tightness. The behavior of the limit of the finite dimensional distributions determines the law of the limit process. Next, we will start investigating the structure of the finite-dimensional distributions of $\sqrt{n}(H_1(t) - H_2(t) - \tilde{G})$.

First we show that $\mathrm{E}\Big(\sqrt{n}(H_1(t) - H_2(t) - \tilde{G})\Big) = 0$. That $\mathrm{E}(H_2(t)) = 0$ follows from $\mathrm{E}\partial\ell(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0)/\partial(\boldsymbol{\alpha}, \boldsymbol{\beta}) = 0$, and $\mathrm{E}(H_1(t) - \tilde{G}) = 0$ can be seen from the following derivation:

$$\mathrm{E}(H_1(t)) \;=\; \frac{1}{n_0} \cdot \mathrm{E} \sum_{i=1}^{n} \frac{I(t_i \le t)}{\sum_{k=0}^{m} \rho_k w_k(t_i)} = \frac{1}{n_0} \sum_{j=0}^{m} n_j \int \frac{w_j I(y \le t)}{\sum_{k=0}^{m} \rho_k w_k(y)} dG(y)$$

$$
\begin{aligned}
&= \sum_{j=0}^{m} \rho_j \int \frac{w_j I(y \le t)}{\sum_{k=0}^{m} \rho_k w_k(y)} dG(y) = G(t), \\
\mathrm{E}(\tilde{G}(t)) &= \mathrm{E}\left(\frac{1}{n_0} \sum_{j=1}^{n_0} I_{[x_{0j} < t]} \right) = G(t).
\end{aligned}
$$

Then we have

$$
\begin{aligned}
&\mathrm{Cov}\left(\sqrt{n}(H_1(t) - \tilde{G}(t) - H_2(t)), \sqrt{n}(H_1(s) - \tilde{G}(s) - H_2(s)) \right) \\
&= n \left[\mathrm{E}\left((H_1(t) - \tilde{G}(t))(H_1(s) - \tilde{G}(s)) \right) - \mathrm{E}\left((H_1(t) - \tilde{G}(t)) H_2(s) \right) \right. \\
&\quad \left. - \mathrm{E}\left(H_2(t)(H_1(s) - \tilde{G}(s)) \right) + \mathrm{E}\left(H_2(t) H_2'(s) \right) \right].
\end{aligned}
$$

Notice that

$$
\begin{aligned}
&H_1(t) - \tilde{G}(t) \\
&= \frac{1}{n_0} \sum_{j=1}^{m} \sum_{i=1}^{n_j} \frac{I(x_{ji} \le t)}{\sum_{k=0}^{m} \rho_k w_k(x_{ji})} - \frac{1}{n_0} \sum_{i=1}^{n_0} \frac{\sum_{k=1}^{m} \rho_k w_k(x_{0i}) I(x_{0i} \le t)}{\sum_{k=0}^{m} \rho_k w_k(x_{0i})}, \quad (3.15)
\end{aligned}
$$

then

$$
\begin{aligned}
&\mathrm{E}\left((H_1(t) - \tilde{G}(t))(H_1(s) - \tilde{G}(s)) \right) \\
&= \frac{1}{n_0^2} \left\{ \mathrm{E}\left[\sum_{j=1}^{m} \sum_{i=1}^{n_j} \frac{I(x_{ji} \le t)}{\sum_{k=0}^{m} \rho_k w_k(x_{ji})} \cdot \sum_{j=1}^{m} \sum_{i=1}^{n_j} \frac{I(x_{ji} \le s)}{\sum_{k=0}^{m} \rho_k w_k(x_{ji})} \right] \right. \\
&\quad - \mathrm{E}\left[\sum_{j=1}^{m} \sum_{i=1}^{n_j} \frac{I(x_{ji} \le t)}{\sum_{k=0}^{m} \rho_k w_k(x_{ji})} \cdot \sum_{i=1}^{n_0} \frac{\sum_{k=1}^{m} \rho_k w_k(x_{0i}) I(x_{0i} \le s)}{\sum_{k=0}^{m} \rho_k w_k(x_{0i})} \right] \\
&\quad - \mathrm{E}\left[\sum_{i=1}^{n_0} \frac{\sum_{k=1}^{m} \rho_k w_k(x_{0i}) I(x_{0i} \le t)}{\sum_{k=0}^{m} \rho_k w_k(x_{0i})} \cdot \sum_{j=1}^{m} \sum_{i=1}^{n_j} \frac{I(x_{ji} \le s)}{\sum_{k=0}^{m} \rho_k w_k(x_{ji})} \right] \\
&\quad \left. + \mathrm{E}\left[\sum_{i=1}^{n_0} \frac{\sum_{k=1}^{m} \rho_k w_k(x_{0i}) I(x_{0i} \le t)}{\sum_{k=0}^{m} \rho_k w_k(x_{0i})} \cdot \sum_{i=1}^{n_0} \frac{\sum_{k=1}^{m} \rho_k w_k(x_{0i}) I(x_{0i} \le s)}{\sum_{k=0}^{m} \rho_k w_k(x_{0i})} \right] \right\} \\
&\equiv \frac{1}{n_0^2} \{ I - II - III + IV \}.
\end{aligned}
$$

Since observations within the same sample are i.i.d., and observations from different

samples are independent, it follows that

$$\begin{aligned} I &= \sum_{j=1}^{m} n_j \int \frac{w_j(t)I(y \le t \wedge s)}{(\sum_{k=0}^{m} \rho_k w_k(y))^2} dG(y) \\ &\quad + \sum_{j,j'=1}^{m} n_j n_{j'} \int \frac{w_j(t)I(y \le t)}{\sum_{k=0}^{m} \rho_k w_k(y)} dG(y) \int \frac{w_{j'}(t)I(y \le s)}{\sum_{k=0}^{m} \rho_k w_k(y)} dG(y) \\ &\quad - \sum_{j=1}^{m} n_j \int \frac{w_j(t)I(y \le t)}{\sum_{k=0}^{m} \rho_k w_k(y)} dG(y) \int \frac{w_j(t)I(y \le s)}{\sum_{k=0}^{m} \rho_k w_k(y)} dG(y) \\ &= n_0 \int \frac{\sum_{j=1}^{m} \rho_j w_j(t)I(y \le t \wedge s)}{(\sum_{k=0}^{m} \rho_k w_k(y))^2} dG(y) \\ &\quad + n_0^2 \sum_{j,j'=1}^{m} \rho_j \rho_{j'} \int \frac{w_j(t)I(y \le t)}{\sum_{k=0}^{m} \rho_k w_k(y)} dG(y) \int \frac{w_{j'}(t)I(y \le s)}{\sum_{k=0}^{m} \rho_k w_k(y)} dG(y) \\ &\quad - n_0 \sum_{j=1}^{m} \rho_j \int \frac{w_j(t)I(y \le t)}{\sum_{k=0}^{m} \rho_k w_k(y)} dG(y) \int \frac{w_j(t)I(y \le s)}{\sum_{k=0}^{m} \rho_k w_k(y)} dG(y). \end{aligned}$$

The first term on the right hand side of the first equality gives the expectation of the product of the pair of same observations. The second term gives the products of the expectations of the pairs of all the observations. The third term gives the products of the expectations of the pairs of same observations. The difference of the last two terms gives the products of the expectations of the pairs of different observations.

The second term on the right hand side of the second equality can be rewritten as

$$n_0^2 \int \frac{\sum_{j=1}^{m} \rho_j w_j(t)I(y \le t)}{\sum_{k=0}^{m} \rho_k w_k(y)} dG(y) \int \frac{\sum_{j'=1}^{m} \rho_{j'} w_{j'}(t)I(y \le s)}{\sum_{k=0}^{m} \rho_k w_k(y)} dG(y).$$

Expressions II and III involve only the product of the pairs of different observations. Therefore, there are only products of expectations of the pairs of different observations in the calculation:

$$II = \sum_{j=1}^{m} n_j n_0 \int \frac{w_j(t)I(y \le t)}{\sum_{k=0}^{m} \rho_k w_k(y)} dG(y) \int \frac{\sum_{k=1}^{m} \rho_k w_k(y)I(y \le s)}{\sum_{k=0}^{m} \rho_k w_k(y)} dG(y)$$

$$
\begin{aligned}
&= n_0^2 \int \frac{\sum_{j=1}^m \rho_j w_j(t) I(y \le t)}{\sum_{k=0}^m \rho_k w_k(y)} dG(y) \int \frac{\sum_{k=1}^m \rho_k w_k(y) I(y \le s)}{\sum_{k=0}^m \rho_k w_k(y)} dG(y), \\
III &= n_0^2 \int \frac{\sum_{j=1}^m \rho_j w_j(t) I(y \le s)}{\sum_{k=0}^m \rho_k w_k(y)} dG(y) \int \frac{\sum_{k=1}^m \rho_k w_k(y) I(y \le t)}{\sum_{k=0}^m \rho_k w_k(y)} dG(y) \\
&= II.
\end{aligned}
$$

Similar to the calculation of I, we have

$$
\begin{aligned}
IV = n_0 \int & \frac{(\sum_{j=1}^m \rho_j w_j(t))^2 I(y \le t \wedge s)}{(\sum_{k=0}^m \rho_k w_k(y))^2} dG(y) \\
& + n_0(n_0 - 1) \int \frac{\sum_{j=1}^m \rho_j w_j(t) I(y \le t)}{\sum_{k=0}^m \rho_k w_k(y)} dG(y) \int \frac{\sum_{k=1}^m \rho_k w_k(y) I(y \le s)}{\sum_{k=0}^m \rho_k w_k(y)} dG(y).
\end{aligned}
$$

Notice that the sum of the first term of I and the first term of IV equals

$$
n_0 \int \frac{\sum_{j=1}^m \rho_j w_j(t) I(y \le t \wedge s)}{\sum_{k=0}^m \rho_k w_k(y)} dG(y).
$$

After collecting terms, then we have

$$
\begin{aligned}
&\mathrm{E}\Big((H_1(t) - \tilde{G}(t))(H_1(s) - \tilde{G}(s)) \Big) \\
&= \frac{1}{n_0^2} \Bigg\{ n_0 \int \frac{\sum_{j=1}^m \rho_j w_j(t) I(y \le t \wedge s)}{\sum_{k=0}^m \rho_k w_k(y)} dG(y) \\
&\quad - n_0 \sum_{j=1}^m \rho_j \int \frac{w_j(t) I(y \le t)}{\sum_{k=0}^m \rho_k w_k(y)} dG(y) \int \frac{w_j(t) I(y \le s)}{\sum_{k=0}^m \rho_k w_k(y)} dG(y) \\
&\quad - n_0 \int \frac{\sum_{j=1}^m \rho_j w_j(t) I(y \le t)}{\sum_{k=0}^m \rho_k w_k(y)} dG(y) \int \frac{\sum_{k=1}^m \rho_k w_k(y) I(y \le s)}{\sum_{k=0}^m \rho_k w_k(y)} dG(y) \Bigg\} \\
&= \frac{1}{n_0} \left\{ \sum_{j=1}^m \rho_j A_j(t \wedge s) - \sum_{j=1}^m \rho_j A_j(t) A_j(s) - \sum_{j=1}^m \rho_j A_j(t) \sum_{j=1}^m \rho_j A_j(s) \right\}.
\end{aligned}
\tag{3.16}
$$

Recall the definition of $H_2(t)$ from Lemma 3.4. Then we have

$$
\mathrm{E}\Big((H_1(t) - \tilde{G}(t)) H_2(s) \Big)
$$

$$= \frac{1}{n}\Big(\bar{A}'(s)\boldsymbol{\rho}, \bar{B}'(s)(\boldsymbol{\rho}\otimes I_p)\Big)S^{-1}\begin{pmatrix} \mathrm{E}\Big[(H_1(t)-\tilde{G}(t))\frac{\partial \ell(\boldsymbol{\alpha}_0,\boldsymbol{\beta}_0)}{\partial \boldsymbol{\alpha}}\Big] \\ \mathrm{E}\Big[(H_1(t)-\tilde{G}(t))\frac{\partial \ell(\boldsymbol{\alpha}_0,\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}}\Big] \end{pmatrix}.$$

Notice that the index j is fixed in the following calculation.

$$\begin{aligned}
&\mathrm{E}\Big[(H_1(t)-\tilde{G}(t))\frac{\partial \ell(\boldsymbol{\alpha}_0,\boldsymbol{\beta}_0)}{\partial \alpha_j}\Big] \\
&= \mathrm{E}\Bigg\{\Bigg[\frac{1}{n_0}\sum_{u=1}^{m}\sum_{i=1}^{n_u}\frac{I(x_{ui}\le t)}{\sum_{k=0}^m \rho_k w_k(x_{ui})} - \frac{1}{n_0}\sum_{i=1}^{n_0}\frac{\sum_{k=1}^m \rho_k w_k(x_{0i}) I(x_{0i}\le t)}{\sum_{k=0}^m \rho_k w_k(x_{0i})}\Bigg] \\
&\quad \cdot\Bigg[-\sum_{i=1}^{n}\frac{\rho_j w_j(t_i)}{\sum_{k=0}^m \rho_k w_k(t_i)} + n_j\Bigg]\Bigg\} \\
&= \frac{1}{n_0}\Bigg\{-\mathrm{E}\Bigg[\sum_{u=1}^{m}\sum_{i=1}^{n_u}\frac{I(x_{ui}\le t)}{\sum_{k=0}^m \rho_k w_k(x_{ui})}\sum_{i=1}^{n}\frac{\rho_j w_j(t_i)}{\sum_{k=0}^m \rho_k w_k(t_i)}\Bigg] \\
&\quad +\mathrm{E}\Bigg[\sum_{i=1}^{n_0}\frac{\sum_{k=1}^m \rho_k w_k(x_{0i}) I(x_{0i}\le t)}{\sum_{k=0}^m \rho_k w_k(x_{0i})}\sum_{i=1}^{n}\frac{\rho_j w_j(t_i)}{\sum_{k=0}^m \rho_k w_k(t_i)}\Bigg]\Bigg\} \\
&\equiv \frac{1}{n_0}\{-I + II\}.
\end{aligned}$$

Similar to the calculation in (3.16), we have

$$\begin{aligned}
I &= \mathrm{E}\Bigg[\sum_{u=1}^{m}\sum_{i=1}^{n_u}\frac{I(x_{ui}\le t)}{\sum_{k=0}^m \rho_k w_k(x_{ui})}\cdot\sum_{u=1}^{m}\sum_{i=1}^{n_u}\frac{\rho_j w_j(x_{ui})}{\sum_{k=0}^m \rho_k w_k(x_{ui})} \\
&\quad +\sum_{u=1}^{m}\sum_{i=1}^{n_u}\frac{I(x_{ui}\le t)}{\sum_{k=0}^m \rho_k w_k(x_{ui})}\cdot\sum_{i=1}^{n_0}\frac{\rho_j w_j(x_{0i})}{\sum_{k=0}^m \rho_k w_k(x_{0i})}\Bigg] \\
&= \sum_{u=1}^{m} n_u\int\frac{\rho_j w_j(y) w_u(y) I(y\le t)}{(\sum_{k=0}^m \rho_k w_k(y))^2}dG(y) \\
&\quad +\sum_{u,u'=1}^{m} n_u n_{u'}\int\frac{w_u(t) I(y\le t)}{\sum_{k=0}^m \rho_k w_k(y)}dG(y)\int\frac{\rho_j w_j(t) w_{u'}(y)}{\sum_{k=0}^m \rho_k w_k(y)}dG(y) \\
&\quad -\sum_{u=1}^{m} n_u\int\frac{w_u(y) I(y\le t)}{\sum_{k=0}^m \rho_k w_k(y)}dG(y)\int\frac{\rho_j w_j(y) w_u(y)}{\sum_{k=0}^m \rho_k w_k(y)}dG(y) \\
&\quad +\sum_{u=1}^{m} n_u n_0\int\frac{w_u(y) I(y\le t)}{\sum_{k=0}^m \rho_k w_k(y)}dG(y)\int\frac{\rho_j w_j(y)}{\sum_{k=0}^m \rho_k w_k(y)}dG(y) \\
&= n_0\int\frac{\rho_j w_j(y)\sum_{u=1}^m \rho_u w_u(y) I(y\le t)}{(\sum_{k=0}^m \rho_k w_k(y))^2}dG(y) \\
&\quad +n_0^2\int\frac{\sum_{u=1}^m \rho_u w_u(y) I(y\le t)}{\sum_{k=0}^m \rho_k w_k(y)}dG(y)\int\frac{\rho_j w_j(y)\sum_{u'=1}^m \rho_{u'} w_{u'}(y)}{\sum_{k=0}^m \rho_k w_k(y)}dG(y)
\end{aligned}$$

$$
\begin{aligned}
&-n_0 \sum_{u=1}^{m} \rho_u \int \frac{w_u(y) I(y \le t)}{\sum_{k=0}^{m} \rho_k w_k(y)} dG(y) \int \frac{\rho_j w_j(y) w_u(y)}{\sum_{k=0}^{m} \rho_k w_k(y)} dG(y) \\
&+n_0^2 \int \frac{\sum_{u=1}^{m} \rho_u w_u(y) I(y \le t)}{\sum_{k=0}^{m} \rho_k w_k(y)} dG(y) \int \frac{\rho_j w_j(y)}{\sum_{k=0}^{m} \rho_k w_k(y)} dG(y).
\end{aligned}
$$

and

$$
\begin{aligned}
II &= \mathrm{E}\left[\sum_{i=1}^{n_0} \frac{\sum_{k=1}^{m} \rho_k w_k(x_{0i}) I(x_{0i} \le t)}{\sum_{k=0}^{m} \rho_k w_k(x_{0i})} \cdot \sum_{u=1}^{m} \sum_{i=1}^{n_u} \frac{\rho_j w_j(x_{ui})}{\sum_{k=0}^{m} \rho_k w_k(x_{ui})} \right. \\
&\quad \left. + \sum_{i=1}^{n_0} \frac{\sum_{k=1}^{m} \rho_k w_k(x_{0i}) I(x_{0i} \le t)}{\sum_{k=0}^{m} \rho_k w_k(x_{0i})} \cdot \sum_{i=1}^{n_0} \frac{\rho_j w_j(x_{0i})}{\sum_{k=0}^{m} \rho_k w_k(x_{0i})} \right] \\
&= \sum_{u=1}^{m} n_u n_0 \int \frac{\sum_{k=1}^{m} \rho_k w_k(y) I(y \le t)}{\sum_{k=0}^{m} \rho_k w_k(y)} dG(y) \int \frac{\rho_j w_j(y) w_u(y)}{\sum_{k=0}^{m} \rho_k w_k(y)} dG(y) \\
&\quad + n_0 \int \frac{\rho_j w_j(y) \sum_{k=1}^{m} \rho_k w_k(y) I(y \le t)}{(\sum_{k=0}^{m} \rho_k w_k(y))^2} dG(y) \\
&\quad + n_0(n_0 - 1) \int \frac{\sum_{k=1}^{m} \rho_k w_k(y) I(y \le t)}{\sum_{k=0}^{m} \rho_k w_k(y)} dG(y) \int \frac{\rho_j w_j(y)}{\sum_{k=0}^{m} \rho_k w_k(y)} dG(y) \\
&= n_0^2 \int \frac{\sum_{k=1}^{m} \rho_k w_k(y) I(y \le t)}{\sum_{k=0}^{m} \rho_k w_k(y)} dG(y) \int \frac{\rho_j w_j(y) \sum_{u=1}^{m} \rho_u w_u(y)}{\sum_{k=0}^{m} \rho_k w_k(y)} dG(y) \\
&\quad + n_0 \int \frac{\rho_j w_j(y) \sum_{k=1}^{m} \rho_k w_k(y) I(y \le t)}{(\sum_{k=0}^{m} \rho_k w_k(y))^2} dG(y) \\
&\quad + (n_0^2 - n_0) \int \frac{\sum_{k=1}^{m} \rho_k w_k(y) I(y \le t)}{\sum_{k=0}^{m} \rho_k w_k(y)} dG(y) \int \frac{\rho_j w_j(y)}{\sum_{k=0}^{m} \rho_k w_k(y)} dG(y).
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
&\mathrm{E}\left[(H_1(t) - \tilde{G}(t)) \frac{\partial \ell(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0)}{\partial \alpha_j} \right] \\
&= \sum_{u=1}^{m} \rho_u \int \frac{w_u(y) I(y \le t)}{\sum_{k=0}^{m} \rho_k w_k(y)} dG(y) \int \frac{\rho_j w_j(y) w_u(y)}{\sum_{k=0}^{m} \rho_k w_k(y)} dG(y) \\
&\quad - \int \frac{\sum_{k=1}^{m} \rho_k w_k(y) I(y \le t)}{\sum_{k=0}^{m} \rho_k w_k(y)} dG(y) \int \frac{\rho_j w_j(y)}{\sum_{k=0}^{m} \rho_k w_k(y)} dG(y) \\
&= \rho_j \sum_{u=1}^{m} \rho_u A_u(t) A_{uj} - \rho_j A_j \sum_{u=1}^{m} \rho_u A_u(t) \\
&= \rho_j \sum_{u=1}^{m} \rho_u A_u(t) A_{uj} - \rho_j (1 - \sum_{u=1}^{m} \rho_u A_{uj}) \sum_{u=1}^{m} \rho_u A_u(t), \qquad (3.17)
\end{aligned}
$$

where

$$A_j = \int \frac{w_j(y)}{\sum_{k=0}^m \rho_k w_k(y)} dG(y),$$

and the last equality is due to $A_j = 1 - \sum_{u=1}^m \rho_u A_{uj}$.

We have

$$\begin{aligned}
&\mathrm{E}\left[(H_1(t) - \tilde{G}(t))\frac{\partial \ell(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0)}{\partial \beta_j}\right] \\
&= \mathrm{E}\left\{\left[\frac{1}{n_0}\sum_{u=1}^m\sum_{i=1}^{n_u}\frac{I(x_{ui}\le t)}{\sum_{k=0}^m \rho_k w_k(x_{ui})} - \frac{1}{n_0}\sum_{i=1}^{n_0}\frac{\sum_{k=1}^m \rho_k w_k(x_{0i}) I(x_{0i}\le t)}{\sum_{k=0}^m \rho_k w_k(x_{0i})}\right]\right. \\
&\quad \left.\cdot\left[-\sum_{i=1}^{n}\frac{\rho_j w_j(t_i)h(t_i)}{\sum_{k=0}^m \rho_k w_k(t_i)} + \sum_{i=1}^{n_j} h(x_{ji})\right]\right\} \\
&= \frac{1}{n_0}\left\{-\mathrm{E}\left[\sum_{u=1}^m\sum_{i=1}^{n_u}\frac{I(x_{ui}\le t)}{\sum_{k=0}^m \rho_k w_k(x_{ui})}\sum_{i=1}^{n}\frac{\rho_j w_j(t_i)h(t_i)}{\sum_{k=0}^m \rho_k w_k(t_i)}\right]\right. \\
&\quad +\mathrm{E}\left[\sum_{i=1}^{n_0}\frac{\sum_{k=1}^m \rho_k w_k(x_{0i}) I(x_{0i}\le t)}{\sum_{k=0}^m \rho_k w_k(x_{0i})}\sum_{i=1}^{n}\frac{\rho_j w_j(t_i)h(t_i)}{\sum_{k=0}^m \rho_k w_k(t_i)}\right] \\
&\quad +\mathrm{E}\left[\sum_{u=1}^m\sum_{i=1}^{n_u}\frac{I(x_{ui}\le t)}{\sum_{k=0}^m \rho_k w_k(x_{ui})}\sum_{i=1}^{n_j} h(x_{ji})\right] \\
&\quad \left.-\mathrm{E}\left[\sum_{i=1}^{n_0}\frac{\sum_{k=1}^m \rho_k w_k(x_{0i}) I(x_{0i}\le t)}{\sum_{k=0}^m \rho_k w_k(x_{0i})}\sum_{i=1}^{n_j} h(x_{ji})\right]\right\} \\
&= \frac{1}{n_0}\left\{-I + II + III - IV\right\}.
\end{aligned}$$

We skip the details of the calculation here since the reasoning is similar to what we did before:

$$\begin{aligned}
I &= n_0\int \frac{\rho_j w_j(y)h(y)\sum_{u=1}^m \rho_u w_u(y) I(y\le t)}{(\sum_{k=0}^m \rho_k w_k(y))^2} dG(y) \\
&\quad + n_0^2 \int \frac{\sum_{u=1}^m \rho_u w_u(y) I(y\le t)}{\sum_{k=0}^m \rho_k w_k(y)} dG(y) \int \frac{\rho_j w_j(y)h(y)\sum_{u'=1}^m \rho_{u'} w_{u'}(y)}{\sum_{k=0}^m \rho_k w_k(y)} dG(y) \\
&\quad - n_0 \sum_{u=1}^m \rho_u \int \frac{w_u(y) I(y\le t)}{\sum_{k=0}^m \rho_k w_k(y)} dG(y) \int \frac{\rho_j w_j(y) w_u(y) h(y)}{\sum_{k=0}^m \rho_k w_k(y)} dG(y) \\
&\quad + n_0^2 \int \frac{\sum_{u=1}^m \rho_u w_u(y) I(y\le t)}{\sum_{k=0}^m \rho_k w_k(y)} dG(y) \int \frac{\rho_j w_j(y) h(y)}{\sum_{k=0}^m \rho_k w_k(y)} dG(y);
\end{aligned}$$

$$
\begin{aligned}
II &= n_0^2 \int \frac{\sum_{u=1}^m \rho_u w_u(y) I(y \le t)}{\sum_{k=0}^m \rho_k w_k(y)} dG(y) \int \frac{\rho_j w_j(y) h(y) \sum_{u'=1}^m \rho_{u'} w_{u'}(y)}{\sum_{k=0}^m \rho_k w_k(y)} dG(y) \\
&\quad + n_0 \int \frac{\rho_j w_j(y) h(y) \sum_{u=1}^m \rho_u w_u(y) I(y \le t)}{(\sum_{k=0}^m \rho_k w_k(y))^2} dG(y) \\
&\quad + (n_0^2 - n_0) \int \frac{\sum_{u=1}^m \rho_u w_u(y) I(y \le t)}{\sum_{k=0}^m \rho_k w_k(y)} dG(y) \int \frac{\rho_j w_j(y) h(y)}{\sum_{k=0}^m \rho_k w_k(y)} dG(y);
\end{aligned}
$$

$$
\begin{aligned}
III &= n_0 \int \frac{\rho_j w_j(y) h(y) I(y \le t)}{\sum_{k=0}^m \rho_k w_k(y)} dG(y) \\
&\quad + n_0^2 \int \frac{\sum_{u=1}^m \rho_u w_u(y) I(y \le t)}{\sum_{k=0}^m \rho_k w_k(y)} dG(y) \int \rho_j w_j(y) h(y) dG(y) \\
&\quad - n_0 \int \frac{\rho_j w_j(y) I(y \le t)}{\sum_{k=0}^m \rho_k w_k(y)} dG(y) \int w_j(y) h(y) dG(y);
\end{aligned}
$$

$$
IV = n_0^2 \int \frac{\sum_{u=1}^m \rho_u w_u(y) I(y \le t)}{\sum_{k=0}^m \rho_k w_k(y)} dG(y) \int \rho_j w_j(y) h(y) dG(y).
$$

Therefore,

$$
\begin{aligned}
&\mathrm{E}\left[(H_1(t) - \tilde{G}(t)) \frac{\partial \ell(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0)}{\partial \beta_j} \right] \\
&= \int \frac{\rho_j w_j(y) h(y) I(y \le t)}{\sum_{k=0}^m \rho_k w_k(y)} dG(y) - \int \frac{\rho_j w_j(y) I(y \le t)}{\sum_{k=0}^m \rho_k w_k(y)} dG(y) \int w_j(y) h(y) dG(y) \\
&\quad + \sum_{u=1}^m \rho_u \int \frac{w_u(y) I(y \le t)}{\sum_{k=0}^m \rho_k w_k(y)} dG(y) \int \frac{\rho_j w_j(y) w_u(y) h(y)}{\sum_{k=0}^m \rho_k w_k(y)} dG(y) \\
&\quad - \int \frac{\sum_{u=1}^m \rho_u w_u(y) I(y \le t)}{\sum_{k=0}^m \rho_k w_k(y)} dG(y) \int \frac{\rho_j w_j(y) h(y)}{\sum_{k=0}^m \rho_k w_k(y)} dG(y) \\
&= \rho_j \left[B_j(t) - A_j(t) E_j \right] + \sum_{u=1}^m \rho_u \rho_j A_u(t) B_{uj} - \rho_j B_j \sum_{u=1}^m \rho_u A_u(t) \qquad (3.18) \\
&= \rho_j \left[B_j(t) - A_j(t) E_j \right] + \sum_{u=1}^m \rho_u \rho_j A_u(t) B_{uj} - \rho_j (E_j - \sum_{u=1}^m \rho_u B_{uj}) \sum_{u=1}^m \rho_u A_u(t),
\end{aligned}
$$

where

$$
B_j(t) = \int \frac{w_j(y) h(y) I(y \le t)}{\sum_{k=0}^m \rho_k w_k(y)} dG(y),
$$

and the last equality is due to $B_j = E_j - \sum_{u=1}^{m} \rho_u B_{uj}$.

By expressing the results from (3.17) and (3.18) in matrix, we have

$$\begin{aligned}
&\begin{pmatrix} \mathrm{E}\left[(H_1(t) - \tilde{G}(t))\frac{\partial \ell(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0)}{\partial \boldsymbol{\alpha}}\right] \\ \mathrm{E}\left[(H_1(t) - \tilde{G}(t))\frac{\partial \ell(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}}\right] \end{pmatrix} \\
&= \begin{pmatrix} \boldsymbol{\rho} A & 0 \\ (\boldsymbol{\rho} \otimes I_p)B - E & I_{mp} \end{pmatrix} \begin{pmatrix} \boldsymbol{\rho}\bar{A}(t) \\ (\boldsymbol{\rho} \otimes I_p)\bar{B}(t) \end{pmatrix} \\
&\quad - \begin{pmatrix} \boldsymbol{\rho}\mathbf{1}_m - \boldsymbol{\rho} A \boldsymbol{\rho}\mathbf{1}_m & 0 \\ E\boldsymbol{\rho}\mathbf{1}_m - (\boldsymbol{\rho} \otimes I_p)B\boldsymbol{\rho}\mathbf{1}_m & 0 \end{pmatrix} \begin{pmatrix} \boldsymbol{\rho}\bar{A}(t) \\ (\boldsymbol{\rho} \otimes I_p)\bar{B}(t) \end{pmatrix} \\
&= S\Sigma \begin{pmatrix} \boldsymbol{\rho}\bar{A}(t) \\ (\boldsymbol{\rho} \otimes I_p)\bar{B}(t) \end{pmatrix},
\end{aligned}$$

where S and Σ are from Theorem 2.1.

Therefore, the covariance of $H_1(t) - \tilde{G}(t)$ and $H_2(s)$ is

$$\begin{aligned}
&\mathrm{E}\Big((H_1(t) - \tilde{G}(t))H_2(s)\Big) \\
&= \frac{1}{n}\Big(\bar{A}'(s)\boldsymbol{\rho}, \bar{B}'(s)(\boldsymbol{\rho} \otimes I_p)\Big) S^{-1} S\Sigma \begin{pmatrix} \boldsymbol{\rho}\bar{A}(t) \\ (\boldsymbol{\rho} \otimes I_p)\bar{B}(t) \end{pmatrix} \\
&= \frac{1}{n}\Big(\bar{A}'(s)\boldsymbol{\rho}, \bar{B}'(s)(\boldsymbol{\rho} \otimes I_p)\Big) \Sigma \begin{pmatrix} \boldsymbol{\rho}\bar{A}(t) \\ (\boldsymbol{\rho} \otimes I_p)\bar{B}(t) \end{pmatrix}. \qquad (3.19)
\end{aligned}$$

Next, we calculate the covariance of $H_2(t)$ and $H_2(s)$. Since we already know that the variance of $\partial \ell(\boldsymbol{\theta}_0)/\partial \boldsymbol{\theta}$ is $n \cdot \Lambda$, it follows that

$$\mathrm{E}\Big(H_2(t)H_2'(s)\Big)$$

$$
\begin{aligned}
&= \frac{1}{n^2}\Big(\bar{A}'(t)\boldsymbol{\rho}, \bar{B}'(t)(\boldsymbol{\rho}\otimes I_p)\Big) S^{-1}\mathrm{E}\Big(\frac{\partial \ell(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}}\frac{\partial \ell(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}}'\Big) S^{-1}\begin{pmatrix} \boldsymbol{\rho}\bar{A}(s) \\ (\boldsymbol{\rho}\otimes I_p)\bar{B}(s) \end{pmatrix} \\
&= \frac{1}{n}\Big(\bar{A}'(t)\boldsymbol{\rho}, \bar{B}'(t)(\boldsymbol{\rho}\otimes I_p)\Big) S^{-1}\Lambda S^{-1}\begin{pmatrix} \boldsymbol{\rho}\bar{A}(s) \\ (\boldsymbol{\rho}\otimes I_p)\bar{B}(s) \end{pmatrix} \\
&= \frac{1}{n}\Big(\bar{A}'(t)\boldsymbol{\rho}, \bar{B}'(t)(\boldsymbol{\rho}\otimes I_p)\Big) \Sigma\begin{pmatrix} \boldsymbol{\rho}\bar{A}(s) \\ (\boldsymbol{\rho}\otimes I_p)\bar{B}(s) \end{pmatrix}. \qquad (3.20)
\end{aligned}
$$

Furthermore, by the representation of Σ from Theorem 2.1,

$$
\begin{aligned}
&\Big(\bar{A}'(t)\boldsymbol{\rho}, \bar{B}'(t)(\boldsymbol{\rho}\otimes I_p)\Big) \Sigma\begin{pmatrix} \boldsymbol{\rho}\bar{A}(s) \\ (\boldsymbol{\rho}\otimes I_p)\bar{B}(s) \end{pmatrix} \\
&\quad = \Big(\bar{A}'(t)\boldsymbol{\rho}, \bar{B}'(t)(\boldsymbol{\rho}\otimes I_p)\Big) S^{-1}\begin{pmatrix} \boldsymbol{\rho}\bar{A}(s) \\ (\boldsymbol{\rho}\otimes I_p)\bar{B}(s) \end{pmatrix} \\
&\qquad -\sum_{k=0}^{m}\rho_k \bar{A}'(t)\boldsymbol{\rho}\bar{A}(s) - \sum_{k=0}^{m}\rho_k \bar{A}'(t)\boldsymbol{\rho}\mathbf{1}_m\boldsymbol{\rho}\bar{A}(s) \\
&\quad = \Big(\bar{A}'(t)\boldsymbol{\rho}, \bar{B}'(t)(\boldsymbol{\rho}\otimes I_p)\Big) S^{-1}\begin{pmatrix} \boldsymbol{\rho}\bar{A}(s) \\ (\boldsymbol{\rho}\otimes I_p)\bar{B}(s) \end{pmatrix} \\
&\qquad -\sum_{k=0}^{m}\rho_k \sum_{j=1}^{m}\rho_j A_j(t)A_j(s) - \sum_{k=0}^{m}\rho_k \sum_{j=1}^{m}\rho_j A_j(t)\sum_{j=1}^{m}\rho_j A_j(s). \qquad (3.21)
\end{aligned}
$$

Therefore, the results from (3.16), (3.19) and (3.20), along with (3.21) and the fact $n/n_0 = \sum_{k=0}^{m}\rho_k$ give the variance-covariance structure of the process $\sqrt{n}(H_1(t) - \tilde{G}(t) - H_2(t))$,

$$
\begin{aligned}
&\mathrm{Cov}\Big(\sqrt{n}(H_1(t) - \tilde{G}(t) - H_2(t)), \sqrt{n}(H_1(s) - \tilde{G}(s) - H_2(s))\Big) \\
&\quad = n\Big[\mathrm{E}\Big((H_1(t) - \tilde{G}(t))(H_1(s) - \tilde{G}(s))\Big) - \mathrm{E}\Big((H_1(t) - \tilde{G}(t))H_2(s)\Big) \\
&\qquad -\mathrm{E}\Big(H_2(t)(H_1(s) - \tilde{G}(s))\Big) + \mathrm{E}\Big(H_2(t)H_2'(s)\Big)\Big]
\end{aligned}
$$

$$= \sum_{k=0}^{m} \rho_k \sum_{j=1}^{m} \rho_j A_j(t \wedge s)$$

$$- \left(\bar{A}'(t)\boldsymbol{\rho}, \bar{B}'(t)(\boldsymbol{\rho} \otimes I_p) \right) S^{-1} \begin{pmatrix} \boldsymbol{\rho}\bar{A}(s) \\ (\boldsymbol{\rho} \otimes I_p)\bar{B}(s) \end{pmatrix}. \tag{3.22}$$

By the multivariate central limit theorem, it is easy to show that the finite dimensional distributions of $\sqrt{n}(H_1(t) - \tilde{G} - H_2(t))$ converges to a mean zero multivariate normal distribution with covariance structure determined by (3.22). Consequently, we have the following lemma.

Lemma 3.5 *For any finite set $(t_1, \ldots, t_k)$ of points on the real line, let G_n denote $\sqrt{n}(H_1(t) - \tilde{G} - H_2(t))$. Then we have*

$$(G_n(t_1), \ldots, G_n(t_k)) \xrightarrow{d} N_k(0, \Delta),$$

where N_k is a mean-zero k-dimensional multivariate normal distribution with covariance matrix Δ, of which the (i, j)th element is determined by (3.22).

3.3.3 Tightness

Next, we prove that the process $\sqrt{n}(H_1(t) - \tilde{G} - H_2(t))$ is tight. We prove the tightness of $\sqrt{n}(H_1(t) - \tilde{G})$ and $\sqrt{n}(H_2(t))$ separately, and the results are presented in the following two lemmas. The basic idea is to decompose each of these two processes into several simple processes which converge to Gaussian processes. Recall from (3.15),

$$H_1(t) - \tilde{G}(t) = \frac{1}{n_0} \sum_{j=1}^{m} \sum_{i=1}^{n_j} \frac{I(x_{ji} \leq t)}{\sum_{k=0}^{m} \rho_k w_k(x_{ji})} - \frac{1}{n_0} \sum_{i=1}^{n_0} \frac{\sum_{k=1}^{m} \rho_k w_k(x_{0i}) I(x_{0i} \leq t)}{\sum_{k=0}^{m} \rho_k w_k(x_{0i})}.$$

Let

$$H_{1j}(t) = \frac{1}{n_0}\sum_{i=1}^{n_j}\frac{I(x_{ji}\le t)}{\sum_{k=0}^m \rho_k w_k(x_{ji})}, \quad H_{10}(t) = \frac{1}{n_0}\sum_{i=1}^{n_0}\frac{\sum_{k=1}^m \rho_k w_k(x_{0i}) I(x_{0i}\le t)}{\sum_{k=0}^m \rho_k w_k(x_{0i})}.$$

It is easy to check that

$$\mathrm{E}\Big(H_{1j}(t)\Big) = \rho_j A_j(t), \quad \mathrm{E}\Big(H_{10}(t)\Big) = \sum_{j=1}^m \rho_j A_j(t).$$

Therefore,

$$\begin{aligned} H_1(t) - \tilde{G}(t) &= \sum_{j=1}^m H_{1j}(t) - H_{10}(t) \\ &= \sum_{j=1}^m \Big(H_{1j}(t) - \rho_j A_j(t)\Big) - \Big(H_{10}(t) - \sum_{j=1}^m \rho_j A_j(t)\Big). \quad (3.23) \end{aligned}$$

Lemma 3.6 *Process $\sqrt{n}(H_1(t) - \tilde{G}(t))$ is tight in $D[-\infty,\infty]$.*

Proof. Let $\mathcal{F}$ be the collection of all indicator functions of the form $I_{(-\infty,t]}$. Obviously, $\mathcal{F}$ is a P_{X_j}-Donsker class, $j = 0, 1, \ldots, m$, where P_{X_j} is the law of X_j, the jth sample, $j = 0, 1, \ldots, m$. Let

$$f_0(y) = \frac{\sum_{k=1}^m \rho_k w_k(y)}{\sum_{k=0}^m \rho_k w_k(y)}, \quad f_j(y) = \frac{\rho_j}{\sum_{k=0}^m \rho_k w_k(y)}, \quad j = 1, \ldots, m.$$

Since all the functions f_j, $j = 0, 1, \ldots, m$ are uniformly bounded, according to Example 2.10.10 of Van der Vaart and Wellner (1996, p.192), $\mathcal{F}\cdot f_j$ is a P_{X_j}-Donsker class, $j = 0, 1, \ldots, m$.

Let $P_{nj} = (1/n_j)\sum_{i=1}^{n_j}\delta_{x_{ji}}$ be the empirical measure of the jth sample. Then we have

$$\begin{aligned} &\sqrt{n_j}(P_{nj} - P_{X_j})(I_{(-\infty,t]} f_j) \\ &= \sqrt{n_j}\left(\frac{1}{n_j}\sum_{i=1}^{n_j}\frac{\rho_j I(x_{ji}\le t)}{\sum_{k=0}^m \rho_k w_k(x_{ji})} - \rho_j \int \frac{w_j(y) I(y\le t)}{\sum_{k=0}^m \rho_k w_k(y)} dG(y)\right) \end{aligned}$$

$$= \left(\rho_j / \sum_{k=0}^{m} \rho_k\right)^{1/2} \sqrt{n}(H_{1j} - \rho_j A_j(t)), \quad j = 1, \ldots, m,$$

and, similarly,

$$\sqrt{n_0}(P_{n0} - P_{X_0})(I_{(-\infty,t]} f_0) = (1/\sum_{k=0}^{m} \rho_k)^{1/2} \sqrt{n}\left(H_{10}(t) - \sum_{j=1}^{m} \rho_j A_j(t)\right),$$

By Donsker's Theorem,

$$\sqrt{n_j}(P_{nj} - P_{X_j})(I_{(-\infty,t]} f_j) \xrightarrow{d} W_j \quad \text{in } D[-\infty, \infty], \quad j = 0, 1, \ldots, m,$$

where the W_j are zero-mean Gaussian processes. It follows by Corollary 3.1, that $\sqrt{n}(H_{1j} - \rho_j A_j(t))$, $j = 0, 1, \ldots, m$, are tight in $D(-\infty, \infty)$. From (3.23), $\sqrt{n}(H_1(t) - \tilde{G}(t))$ is tight in $D[-\infty, \infty]$. □

Let

$$U_l(y) = \frac{\rho_l w_l(y)}{\sum_{k=0}^{m} \rho_k w_k(y)}, \quad V_l(y) = \frac{\rho_l w_l(y) h(y)}{\sum_{k=0}^{m} \rho_k w_k(y)},$$

where $l = 0, 1, \ldots, m$.

From the score equations (1.24), we have

$$\begin{aligned}
\frac{1}{n_0}\frac{\partial \ell}{\partial \alpha_j} &= \frac{1}{n_0} \sum_{i=1}^{n_j} \frac{\sum_{\substack{l=0 \\ l \neq j}}^{m} \rho_l w_l(x_{ji})}{\sum_{k=0}^{m} \rho_k w_k(x_{ji})} - \frac{1}{n_0} \sum_{\substack{l=0 \\ l \neq j}}^{m} \sum_{i=1}^{n_l} \frac{\rho_j w_j(x_{li})}{\sum_{k=0}^{m} \rho_k w_k(x_{li})} \\
&= P_{n_j}\left(\sum_{\substack{l=0 \\ l \neq j}}^{m} \rho_j U_l\right) - \sum_{\substack{l=0 \\ l \neq j}}^{m} P_{n_l}(\rho_l U_j) \\
\frac{1}{n_0}\frac{\partial \ell}{\partial \beta_j} &= \frac{1}{n_0} \sum_{i=1}^{n_j} \frac{\sum_{\substack{l=0 \\ l \neq j}}^{m} \rho_l w_l(x_{ji}) h(x_{ji})}{\sum_{k=0}^{m} \rho_k w_k(x_{ji})} - \frac{1}{n_0} \sum_{\substack{l=0 \\ l \neq j}}^{m} \sum_{i=1}^{n_l} \frac{\rho_j w_j(x_{li}) h(x_{li})}{\sum_{k=0}^{m} \rho_k w_k(x_{li})} \\
&= P_{n_j}\left(\sum_{\substack{l=0 \\ l \neq j}}^{m} \rho_j V_l\right) - \sum_{\substack{l=0 \\ l \neq j}}^{m} P_{n_l}(\rho_l V_j). \qquad (3.24)
\end{aligned}$$

Notice that

$$P_{X_j}\left(\sum_{\substack{l=0\\l\neq j}}^{m}\rho_j U_l\right)=\sum_{\substack{l=0\\l\neq j}}^{m}P_{X_l}(\rho_l U_j)=\sum_{\substack{l=0\\l\neq j}}^{m}\int\frac{\rho_j\rho_l w_j(y)w_l(y)}{\sum_{k=0}^{m}\rho_k w_k(y)}dG(y),$$

and

$$P_{X_j}\left(\sum_{\substack{l=0\\l\neq j}}^{m}\rho_j V_l\right)=\sum_{\substack{l=0\\l\neq j}}^{m}P_{X_l}(\rho_l V_j)=\sum_{\substack{l=0\\l\neq j}}^{m}\int\frac{\rho_j\rho_l w_j(y)w_l(y)h(y)}{\sum_{k=0}^{m}\rho_k w_k(y)}dG(y).$$

Therefore,

$$\begin{aligned}\frac{1}{n_0}\frac{\partial\ell}{\partial\alpha_j} &= (P_{n_j}-P_{X_j})\left(\sum_{\substack{l=0\\l\neq j}}^{m}\rho_j U_l\right)-\sum_{\substack{l=0\\l\neq j}}^{m}(P_{n_l}-P_{X_l})(\rho_l U_j)\\ \frac{1}{n_0}\frac{\partial\ell}{\partial\beta_j} &= (P_{n_j}-P_{X_j})\left(\sum_{\substack{l=0\\l\neq j}}^{m}\rho_j V_l\right)-\sum_{\substack{l=0\\l\neq j}}^{m}(P_{n_l}-P_{X_l})(\rho_l V_j).\end{aligned}\tag{3.25}$$

Let $(a_1(t),\ldots,a_m(t),b_1'(t),\ldots,b_m'(t))$ denote the product $(\bar{A}'(t)\boldsymbol{\rho},\bar{B}'(t)(\boldsymbol{\rho}\otimes I_p))S^{-1}$, where $a_i(t)$'s are scalars and $b_i(t)$'s are $p\times 1$ vectors. From (3.13) and (3.25), we can rewrite $H_2(t)$ as

$$\begin{aligned}\sqrt{n}H_2(t) &= \frac{\sqrt{n}}{n}\left(a_1(t),\ldots,a_m(t),b_1'(t),\ldots,b_m'(t)\right)\begin{pmatrix}\frac{\partial\ell}{\partial\alpha_1}\\ \vdots\\ \frac{\partial\ell}{\partial\alpha_m}\\ \frac{\partial\ell}{\partial\beta_1}\\ \vdots\\ \frac{\partial\ell}{\partial\beta_m}\end{pmatrix}\\ &= \frac{\sqrt{n}}{n}\left(\sum_{j=1}^{m}a_j(t)\frac{\partial\ell}{\partial\alpha_j}+\sum_{j=1}^{m}b_j'(t)\frac{\partial\ell}{\partial\beta_j}\right)\\ &= \frac{\sqrt{n}n_0}{n}\left(\sum_{j=1}^{m}a_j(t)(P_{n_j}-P_{X_j})\left(\sum_{\substack{l=0\\l\neq j}}^{m}\rho_j U_l\right)-\sum_{j=1}^{m}a_j(t)\sum_{\substack{l=0\\l\neq j}}^{m}(P_{n_l}-P_{X_l})(\rho_l U_j)\right.\end{aligned}$$

$$
\begin{aligned}
&+\sum_{j=1}^{m} b_j'(t)(P_{n_j}-P_{X_j})\Bigg(\sum_{\substack{l=0\\ l\neq j}}^{m}\rho_j V_l\Bigg)-\sum_{j=1}^{m} b_j'(t)\sum_{\substack{l=0\\ l\neq j}}^{m}(P_{n_l}-P_{X_l})(\rho_l V_j)\Bigg)\\
=\;& \frac{n_0}{n}\Bigg(\sum_{j=1}^{m}\frac{\sum_{k=0}^{m}\rho_k}{\rho_j}\sqrt{n_j}(P_{n_j}-P_{X_j})\Bigg(\sum_{\substack{l=0\\ l\neq j}}^{m}\rho_j a_j(t)U_l\Bigg)\\
&-\sum_{j=1}^{m}\sum_{\substack{l=0\\ l\neq j}}^{m}\frac{\sum_{k=0}^{m}\rho_k}{\rho_l}\sqrt{n_l}(P_{n_l}-P_{X_l})\Bigg(\rho_l a_j(t)U_j\Bigg)\\
&+\sum_{j=1}^{m}\frac{\sum_{k=0}^{m}\rho_k}{\rho_j}\sqrt{n_j}(P_{n_j}-P_{X_j})\Bigg(\sum_{\substack{l=0\\ l\neq j}}^{m}\rho_j b_j'(t)V_l\Bigg)\\
&-\sum_{j=1}^{m}\sum_{\substack{l=0\\ l\neq j}}^{m}\frac{\sum_{k=0}^{m}\rho_k}{\rho_l}\sqrt{n_l}(P_{n_l}-P_{X_l})\Bigg(\rho_l b_j'(t)V_j\Bigg)\Bigg). \qquad (3.26)
\end{aligned}
$$

Equation (3.26) gives the decomposition of the process $\sqrt{n}H_2(t)$. If we can show that the collection of functions on which empirical measures are operated is P-Donsker, the tightness of $\sqrt{n}H_2(t)$ is followed immediately as in the proof of Lemma 3.6.

Lemma 3.7 *Process $\sqrt{n}H_2(t)$ is tight in $D[-\infty,\infty]$.*

Proof. Let $\mathcal{U}$ be a collection of linear combinations of functions $\{U_k : k = 0,1,\ldots,m\}$, and $\mathcal{V}$ be a collection of linear combinations of functions $\{V_k : k = 0,1,\ldots,m\}$. Coefficients are chosen from $\{\rho_k : k = 0,1,\ldots,m\}$ and $\{a_k(t), b_k(t) : k = 0,1,\ldots,m\}$. By Lemma 3.2, both $\mathcal{U}$ and $\mathcal{V}$ are VC-subgraph classes. According to Theorem 3.7, the covering numbers of $\mathcal{U}$ and $\mathcal{V}$ are bounded by a polynomial in $1/\varepsilon$. Therefore, the uniform entropy bound (3.7) is satisfied. Provided the assumption that the second moments of $h(x)$ are bounded with respect to $P_{X_j},\ j = 0,1,\ldots,m$ respectively, it is easy to check that both the envelope functions

of $\mathcal{U}$ and $\mathcal{V}$ are square integrable with respect to $P_{X_j},\ j = 0, 1, \ldots, m$ respectively. Then we can conclude that both $\mathcal{U}$ and $\mathcal{V}$ are P_{X_j}-Donsker classes, $j = 0, 1, \ldots, m$. Therefore, processes $\sqrt{n_j}(P_{n_j} - P_{X_j})\left(\sum_{\substack{l=0 \\ l\neq j}}^{m} \rho_j a_j(t) U_l\right)$, $\sqrt{n_l}(P_{n_l} - P_{X_l})(\rho_l a_j(t) U_j)$, $\sqrt{n_j}(P_{n_j} - P_{X_j})\left(\sum_{\substack{l=0 \\ l\neq j}}^{m} \rho_j b'_j(t) V_l\right)$ and $\sqrt{n_l}(P_{n_l} - P_{X_l})(\rho_l b'_j(t) V_j)$ in equation (3.26) converge to zero-mean Gaussian processes in $D[-\infty, \infty]$. It follows, by (3.26), that $\sqrt{n}H_2(t)$ is tight in $D[-\infty, \infty]$. □

The tightness derived from Lemmas 3.6 and 3.7, along with the finite dimensional convergence from Lemma 3.5 give the weak convergence of process $\sqrt{n}(\hat{G} - \tilde{G})$.

Theorem 3.8 *The process $\sqrt{n}(\hat{G} - \tilde{G})$ converges weakly to a zero-mean Gaussian process W with continuous sample paths in $D[-\infty, \infty]$, and the covariance matrix is determined by*

$$EW(t)W(s) = \sum_{k=0}^{m} \rho_k \sum_{j=1}^{m} \rho_j A_j(t \wedge s) - \left(\bar{A}'(t)\boldsymbol{\rho}, \bar{B}'(t)(\boldsymbol{\rho} \otimes I_p)\right) S^{-1} \begin{pmatrix} \boldsymbol{\rho}\bar{A}(s) \\ (\boldsymbol{\rho} \otimes I_p)\bar{B}(s) \end{pmatrix}.$$

3.3.4 Weak Convergence of $\sqrt{n}(\hat{G}(t) - G(t))$

As we mentioned before,

$$\sqrt{n}(\hat{G}(t) - G(t)) = \sqrt{n}(\hat{G}(t) - \tilde{G}(t)) + \sqrt{n}(\tilde{G}(t) - G(t)),$$

and $\sqrt{n}(H_1(t) - \tilde{G}(t) - H_2(T))$ is an approximation of $\sqrt{n}(\hat{G}(t) - \tilde{G}(t))$. The covariance structure of the process $\sqrt{n}(\hat{G}(t) - G(t))$ can be expressed as

$$\begin{aligned}
&\mathrm{Cov}\{\sqrt{n}(\hat{G}(t) - G(t)), \sqrt{n}(\hat{G}(s) - G(s))\} = \\
&\quad \mathrm{E}\{\sqrt{n}(\hat{G}(t) - \tilde{G}(t))\sqrt{n}(\hat{G}(s) - \tilde{G}(s))\} + \mathrm{E}\{\sqrt{n}(\hat{G}(t) - \tilde{G}(t))\sqrt{n}(\tilde{G}(s) - G(s))\} \\
&\quad +\mathrm{E}\{\sqrt{n}(\tilde{G}(t) - G(t))\sqrt{n}(\hat{G}(s) - \tilde{G}(s))\} \\
&\quad +\mathrm{E}\{\sqrt{n}(\tilde{G}(t) - G(t))\sqrt{n}(\tilde{G}(s) - G(s))\}.
\end{aligned}$$

The asymptotic covariance structure of $\sqrt{n}(\hat{G}(t) - \tilde{G}(t))$ is given by (3.22). And it is easy to calculate that the covariance matrix of the process $\sqrt{n}(\tilde{G}(t) - G(t))$, the empirical process of the reference sample, is

$$\mathrm{E}\{\sqrt{n}(\tilde{G}(t) - G(t))\sqrt{n}(\tilde{G}(s) - G(s))\} = \sum_{k=0}^{m} \rho_k \Big(G(t \wedge s) - G(t)G(s) \Big). \quad (3.27)$$

The asymptotic covariance $\mathrm{Cov}\{\sqrt{n}(\hat{G}(t) - \tilde{G}(t)), \sqrt{n}(\tilde{G}(s) - G(s))\}$ is equivalent to $\mathrm{E}\{\sqrt{n}(H_1(t) - \tilde{G}(t) - H_2(t))\sqrt{n}(\tilde{G}(s) - G(s))\}$. Since $G(t)$ is non-random, and the expectation of $\sqrt{n}(H_1(t) - \tilde{G}(t) - H_2(t))$ is zero, the expectation can be simplified as $\mathrm{E}\{\sqrt{n}(H_1(t) - \tilde{G}(t) - H_2(t))\sqrt{n}\tilde{G}(s)\}$.

We first start with $\mathrm{E}\{\sqrt{n}(H_1(t) - \tilde{G}(t))\sqrt{n}\tilde{G}(s)\}$. From (3.15) and that $\tilde{G}$ is the empirical process of the reference sample, we have

$$\begin{aligned}
&\mathrm{E}\{\sqrt{n}(H_1(t) - \tilde{G}(t))\sqrt{n}\tilde{G}(s)\} = \\
&\qquad \frac{n}{n_0^2} \left\{ \sum_{j=1}^{m} \sum_{i=1}^{n_j} \frac{I(x_{ji} \le t)}{\sum_{k=0}^{m} \rho_k w_k(x_{ji})} \sum_{i=1}^{n_0} I_{[x_{0i} < s]} \right. \\
&\qquad \left. - \sum_{i=1}^{n_0} \frac{\sum_{k=1}^{m} \rho_k w_k(x_{0i}) I(x_{0i} \le t)}{\sum_{k=0}^{m} \rho_k w_k(x_{0i})} \sum_{i=1}^{n_0} I_{[x_{0i} < s]} \right\} \\
&= \quad \frac{n}{n_0^2} \left\{ n_0^2 \int \frac{\sum_{j=1}^{m} \rho_j w_j(y) I(y \le t)}{\sum_{k=0}^{m} \rho_k w_k(y)} dG(y) \cdot G(s) \right.
\end{aligned}$$

$$
\begin{aligned}
& -n_0 \int \frac{\sum_{k=1}^m \rho_k w_k(y) I(y \le t \wedge s)}{\sum_{k=0}^m \rho_k w_k(y)} dG(y) \\
& -n_0(n_0-1) \int \frac{\sum_{k=1}^m \rho_k w_k(y) I(y \le t)}{\sum_{k=0}^m \rho_k w_k(y)} dG(y) \cdot G(s) \Bigg\} \\
= \; & \frac{n}{n_0} \left\{ \int \frac{\sum_{j=1}^m \rho_j w_j(y) I(y \le t)}{\sum_{k=0}^m \rho_k w_k(y)} dG(y) \cdot G(s) \right. \\
& \left. - \int \frac{\sum_{k=1}^m \rho_k w_k(y) I(y \le t \wedge s)}{\sum_{k=0}^m \rho_k w_k(y)} dG(y) \right\} \\
= \; & \sum_{k=0}^m \rho_k \left\{ \sum_{j=1}^m \rho_j A_j(t) G(s) - \sum_{j=1}^m \rho_j A_j(t \wedge s) \right\}. \qquad (3.28)
\end{aligned}
$$

The covariance of $\sqrt{n}H_2(t)$ and $\sqrt{n}\tilde{G}(s)$ is

$$
\mathrm{E}\left(\sqrt{n}H_2(t)\sqrt{n}\tilde{G}(s) \right) = \left(\bar{A}'(s)\boldsymbol{\rho}, \bar{B}'(s)(\boldsymbol{\rho} \otimes I_p) \right) S^{-1} \begin{pmatrix} \mathrm{E}\left[\tilde{G}(s) \frac{\partial \ell(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0)}{\partial \boldsymbol{\alpha}} \right] \\ \mathrm{E}\left[\tilde{G}(s) \frac{\partial \ell(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}} \right] \end{pmatrix}.
$$

$$
\begin{aligned}
\mathrm{E}\left[\tilde{G}(s) \frac{\partial \ell(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0)}{\partial \alpha_j} \right] &= \frac{1}{n_0} \mathrm{E} \left\{ \sum_{i=1}^{n_0} I_{[x_{0i} < s]} \left[n_j - \sum_{i=1}^n \frac{\rho_j w_j(t_i)}{\sum_{k=0}^m \rho_k w_k(t_i)} \right] \right\} \\
&= \frac{1}{n_0} \left\{ n_0 n_j G(s) - n_0 \sum_{i=0}^m n_i \int \frac{\rho_j w_j(y) w_i(y)}{\sum_{k=0}^m \rho_k w_k(y)} dG(y) \cdot G(s) \right. \\
&\quad \left. - n_0 \int \frac{\rho_j w_j(y) I_{[y<s]}}{\sum_{k=0}^m \rho_k w_k(y)} dG(y) + n_0 \int \frac{\rho_j w_j(y)}{\sum_{k=0}^m \rho_k w_k(y)} dG(y) \cdot G(s) \right\} \\
&= - \int \frac{\rho_j w_j(y) I_{[y<s]}}{\sum_{k=0}^m \rho_k w_k(y)} dG(y) + \int \frac{\rho_j w_j(y)}{\sum_{k=0}^m \rho_k w_k(y)} dG(y) \cdot G(s) \\
&= -\rho_j A_j(s) + \rho_j A_j G(s) \\
&= -\rho_j A_j(s) + \rho_j (1 - \sum_{i=1}^m \rho_i A_{ij}) G(s).
\end{aligned}
$$

The second equality is resulted from the cancelation of the first two terms on the right side of the first equality. Similarly,

$$
\begin{aligned}
\mathrm{E}\left[\tilde{G}(s) \frac{\partial \ell(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0)}{\partial \beta_j} \right] &= \frac{1}{n_0} \mathrm{E} \left\{ \sum_{i=1}^{n_0} I_{[x_{0i} < s]} \left[\sum_{i=1}^{n_j} h(x_{ji}) - \sum_{i=1}^n \frac{\rho_j w_j(t_i) h(t_i)}{\sum_{k=0}^m \rho_k w_k(t_i)} \right] \right\} \\
&= \frac{1}{n_0} \left\{ n_0 n_j \int w_j(y) h(y) dG(y) \cdot G(s) \right.
\end{aligned}
$$

$$
\begin{aligned}
& -n_0 \sum_{i=0}^{m} n_i \int \frac{\rho_j w_j(y) w_i(y) h(y)}{\sum_{k=0}^{m} \rho_k w_k(y)} dG(y) \cdot G(s) \\
& -n_0 \int \frac{\rho_j w_j(y) h(y) I_{[y<s]}}{\sum_{k=0}^{m} \rho_k w_k(y)} dG(y) + n_0 \int \frac{\rho_j w_j(y) h(y)}{\sum_{k=0}^{m} \rho_k w_k(y)} dG(y) \cdot G(s) \Bigg\} \\
= & -\int \frac{\rho_j w_j(y) h(y) I_{[y<s]}}{\sum_{k=0}^{m} \rho_k w_k(y)} dG(y) + \int \frac{\rho_j w_j(y) h(y)}{\sum_{k=0}^{m} \rho_k w_k(y)} dG(y) \cdot G(s) \\
= & -\rho_j B_j(s) + \rho_j B_j G(s) \\
= & -\rho_j B_j(s) + \rho_j (E_j - \sum_{i=1}^{m} \rho_i B_{ij}) G(s).
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
& \mathrm{E}\Big(\sqrt{n} H_2(t) \sqrt{n} \tilde{G}(s) \Big) \\
= & -\Big(\bar{A}'(s)\boldsymbol{\rho}, \bar{B}'(s)(\boldsymbol{\rho} \otimes I_p) \Big) S^{-1} \begin{pmatrix} \boldsymbol{\rho} \bar{A}(t) \\ (\boldsymbol{\rho} \otimes I_p) \bar{B}(t) \end{pmatrix} \\
& + \Big(\bar{A}'(s)\boldsymbol{\rho}, \bar{B}'(s)(\boldsymbol{\rho} \otimes I_p) \Big) S^{-1} \begin{pmatrix} \boldsymbol{\rho} \bar{\mathbf{1}}_m - \boldsymbol{\rho} A \boldsymbol{\rho} \bar{\mathbf{1}}_m \\ E \boldsymbol{\rho} \bar{\mathbf{1}}_m - (\boldsymbol{\rho} \otimes I_p) B \boldsymbol{\rho} \bar{\mathbf{1}}_m \end{pmatrix} G(s) \\
= & -\Big(\bar{A}'(s)\boldsymbol{\rho}, \bar{B}'(s)(\boldsymbol{\rho} \otimes I_p) \Big) S^{-1} \begin{pmatrix} \boldsymbol{\rho} \bar{A}(t) \\ (\boldsymbol{\rho} \otimes I_p) \bar{B}(t) \end{pmatrix} \\
& + \Big(\bar{A}'(s)\boldsymbol{\rho}, \bar{B}'(s)(\boldsymbol{\rho} \otimes I_p) \Big) S^{-1} \begin{pmatrix} S_{11} \\ S_{21} \end{pmatrix} \bar{\mathbf{1}}_m G(s) \\
= & -\Big(\bar{A}'(s)\boldsymbol{\rho}, \bar{B}'(s)(\boldsymbol{\rho} \otimes I_p) \Big) S^{-1} \begin{pmatrix} \boldsymbol{\rho} \bar{A}(t) \\ (\boldsymbol{\rho} \otimes I_p) \bar{B}(t) \end{pmatrix} \\
& + \sum_{k=0}^{m} \rho_k \Big(\bar{A}'(s)\boldsymbol{\rho}, \bar{B}'(s)(\boldsymbol{\rho} \otimes I_p) \Big) \begin{pmatrix} \bar{\mathbf{1}}_m \\ \bar{0}_{mp} \end{pmatrix} G(s) \\
= & \sum_{k=0}^{m} \rho_k \sum_{j=1}^{m} \rho_j A_j(t) G(s)
\end{aligned}
$$

$$-\Big(\bar{A}'(s)\boldsymbol{\rho}, \bar{B}'(s)(\boldsymbol{\rho}\otimes I_p)\Big)S^{-1}\begin{pmatrix}\boldsymbol{\rho}\bar{A}(t)\\ \\ (\boldsymbol{\rho}\otimes I_p)\bar{B}(t)\end{pmatrix}, \tag{3.29}$$

where $\bar{\mathbf{1}}_m$ is a $p\times 1$ vector of 1's, and $\bar{0}_{mp}$ is a $mp\times 1$ vector of 0's.

Equations (3.28) and (3.29) together give

$$\begin{aligned}&\mathrm{E}\{\sqrt{n}(H_1(t)-\tilde{G}(t)-H_2(t))\sqrt{n}\tilde{G}(s)\}=\\ &\Big(\bar{A}'(s)\boldsymbol{\rho}, \bar{B}'(s)(\boldsymbol{\rho}\otimes I_p)\Big)S^{-1}\begin{pmatrix}\boldsymbol{\rho}\bar{A}(t)\\ \\ (\boldsymbol{\rho}\otimes I_p)\bar{B}(t)\end{pmatrix}-\sum_{k=0}^{m}\rho_k\sum_{j=1}^{m}\rho_j A_j(t\wedge s).\end{aligned} \tag{3.30}$$

The asymptotic covariance structure of $\sqrt{n}(\hat{G}(t)-G(t))$ can be obtained from (3.22), (3.27) and (3.30).

$$\begin{aligned}&\mathrm{Cov}\{\sqrt{n}(\hat{G}(t)-G(t)),\sqrt{n}(\hat{G}(s)-G(s))\}=\\ &\quad\Big(\sum_{k=0}^{m}\rho_k\Big)\Big(G(t\wedge s)-G(t)G(s)-\sum_{j=1}^{m}\rho_j A_j(t\wedge s)\Big)\\ &\quad+\Big(\bar{A}'(s)\boldsymbol{\rho}, \bar{B}'(s)(\boldsymbol{\rho}\otimes I_p)\Big)S^{-1}\begin{pmatrix}\boldsymbol{\rho}\bar{A}(t)\\ \\ (\boldsymbol{\rho}\otimes I_p)\bar{B}(t)\end{pmatrix}.\end{aligned} \tag{3.31}$$

Obviously, the covariance structure of the limit of the finite dimensional distribution of $\sqrt{n}(\hat{G}(t)-G(t))$ is given by (3.31). The corresponding tightness is followed by the tightness of processes $\sqrt{n}(\hat{G}(t)-\tilde{G}(t))$ and $\sqrt{n}(\tilde{G}(t)-G(t))$, the first is already given by Theorem 3.8, and the second is well known. Therefore, we have the following theorem.

Theorem 3.9 *The process $\sqrt{n}(\hat{G}(t)-G(t))$ converges weakly to a zero-mean Gaussian process in $D[-\infty,\infty]$, with covariance matrix given by (3.31).*

By now we completed the asymptotic theory for the semiparametric estimator $\hat{G}$. The main results in this chapter can be reduced to those of the two-sample case in Qin and Zhang[1997] and Zhang[2000c].

Chapter 4

Simulation Studies

In this chapter we present simulation studies conducted to illustrate the results from previous chapters based on the semiparametric density ratio model (1.11).

4.1 Simulation Studies for the Estimation of Parameters

We generated random samples X_0, X_1 and X_2 from $X_0 \sim N(0,1)$, $X_1 \sim N(0,2)$ and $X_2 \sim N(0,4)$ with density functions $g(x)$, $g_1(x)$ and $g_2(x)$ respectively. From example 1.14, the semiparametric density ratio model (1.11) holds:

$$\begin{aligned} g_1(x) &= g(x)\exp(\alpha_1 + \beta_1 x^2), \\ g_2(x) &= g(x)\exp(\alpha_2 + \beta_2 x^2) \end{aligned} \tag{4.1}$$

with true parameters $(\alpha_1, \alpha_2, \beta_1, \beta_2) = (-0.34657, -0.69315, 0.25000, 0.37500)$.

In our simulation, we considered five different combinations of sample sizes, $(n_0, n_1, n_2) = (50,50,50), (50,50,100), (50,100,50), (100,50,50), (200,200,200)$. For each combination, we generated 1000 independent combined random samples from

X_0, X_1 and X_2. In each case, the average bias and sample variance are obtained from the 1000 samples. Simulation results are summarized in Tables 4.1, 4.2 and 4.3. From the simulation results of bias and variance in Tables 4.1 and 4.2, we can see the estimation accuracy improves with the increase of the total sample size n. The 95% confidence intervals in Table 4.3 are obtained from the covariance matrix (2.27), and they become narrower as the total sample size n increases.

Table 4.1: Bias of parameter estimates from the semiparametric density ratio model.

Sample Size	$\boldsymbol{\rho}$	Bias($\hat{\alpha_1}$)	Bias($\hat{\beta}_1$)	Bias($\hat{\alpha_2}$)	Bias($\hat{\beta}_2$)
(50, 50, 50)	(1, 1)	-0.01663	0.02337	-0.03752	0.03508
(50, 50, 100)	(1, 2)	-0.00022	0.00856	-0.02041	0.02142
(50, 100, 50)	(2, 1)	-0.01865	0.02550	-0.03797	0.03338
(100, 50, 50)	(0.5, 0.5)	-0.00326	0.00511	-0.02925	0.01811
(200, 200, 200)	(1, 1)	-0.00017	0.00217	-0.00303	0.00439

4.2 Goodness of Fit Test

The discrepancy between the semiparametric estimator $\hat{G}$ and the empirical estimator $\tilde{G}$ from the reference sample X_0 only allows us to assess the validity of model (1.11). We define the difference between $\hat{G}$ and $\tilde{G}$ as

$$\Delta_n(t) = \sqrt{n}\,|\hat{G} - \tilde{G}|, \quad \Delta_n = \sup_{-\infty \le t \le \infty} \Delta_n(t). \tag{4.2}$$

Table 4.2: Monte Carlo Variance of parameter estimates from the semiparametric density ratio model.

Sample Size	$\boldsymbol{\rho}$	$\text{Var}(\hat{\alpha_1})$	$\text{Var}(\hat{\beta}_1)$	$\text{Var}(\hat{\alpha_2})$	$\text{Var}(\hat{\beta}_2)$
(50, 50, 50)	(1, 1)	0.02425	0.01731	0.03362	0.01672
(50, 50, 100)	(1, 2)	0.02085	0.01497	0.02276	0.01421
(50, 100, 50)	(2, 1)	0.01961	0.01623	0.03168	0.01658
(100, 50, 50)	(0.5, 0.5)	0.01773	0.00929	0.02674	0.00828
(200, 200, 200)	(1, 1)	0.00611	0.00391	0.00837	0.00374

Δ_n can be used to measure the departure from the assumption of the semiparametric density ratio model (1.11).

We showed that $\sqrt{n}(\hat{G}-\tilde{G})$ converges weakly to a Gaussian process W defined in Theorem 3.8. Let w_α denote the α-quantile of the distribution of $\sup_{-\infty\le t\le\infty}|W(t)|$. By Theorem 3.8,

$$\begin{aligned}\lim_{n\to\infty} P(\Delta_n \ge w_{1-\alpha}) &= \lim_{n\to\infty} P(\sup_{-\infty\le t\le\infty}\sqrt{n}\,|\hat{G}-\tilde{G}| \ge w_{1-\alpha}) \\ &= P(\sup_{-\infty\le t\le\infty}\sqrt{n}\,|W(t)| \ge w_{1-\alpha}) = \alpha.\end{aligned}$$

Therefore, we reject model (1.11) at level α if

$$\Delta_n > w_{1-\alpha}.$$

Since there are no analytic expressions available for the distribution of the supremum of a Gaussian process $W(t)$ and for the corresponding quantile function, we applied a bootstrap procedure to simulate the distribution of $\sup_{-\infty\le t\le\infty}|W(t)|$ and its quantiles.

Table 4.3: 95% Confidence intervals for parameters from the semiparametric density ratio model.

Sample Size	$\boldsymbol{\rho}$	α_1	α_2
(50, 50, 50)	(1, 1)	(-0.64010, -0.05305)	(-1.04119, -0.34511)
(50, 50, 100)	(1, 2)	(-0.63505, -0.05810)	(-0.98391, -0.40238)
(50, 100, 50)	(2, 1)	(-0.60031, -0.09284)	(-1.02833, -0.35796)
(100, 50, 50)	(0.5, 0.5)	(-0.60319, -0.08996)	(-1.00900, -0.37730)
(200, 200, 200)	(1, 1)	(-0.49333, -0.19981)	(-0.86717, -0.51913)

Sample Size	$\boldsymbol{\rho}$	β_1	β_2
(50, 50, 50)	(1, 1)	(0.02029, 0.47971)	(0.14742, 0.60258)
(50, 50, 100)	(1, 2)	(0.02349, 0.47651)	(0.15889, 0.59111)
(50, 100, 50)	(2, 1)	(0.03402, 0.46598)	(0.15327, 0.59673)
(100, 50, 50)	(0.5, 0.5)	(0.06846, 0.43154)	(0.19693, 0.55307)
(200, 200, 200)	(1, 1)	(0.13515, 0.36485)	(0.26121, 0.48879)

Let $X_0^*, X_1^*, \ldots, X_m^*$ be random samples generated from $\hat{G}, \hat{G}_1, \ldots, \hat{G}_m$ respectively, where $\hat{G}, \hat{G}_1, \ldots, \hat{G}_m$ are estimated from the combined sample $(X_0, X_1, \ldots, X_m)$. Let $(\hat{\boldsymbol{\alpha}}^*, \hat{\boldsymbol{\beta}}^*)$ and let $\hat{G}^*$ be the estimates for the parameters and the reference distribution obtained from $(X_0^*, X_1^*, \ldots, X_m^*)$ as we described before, and $\tilde{G}^*$ be the empirical distribution from X_0^* only. Then the corresponding bootstrap version of the test statistic Δ_n is given by

$$\Delta_n^*(t) = \sqrt{n}\,|\hat{G}^* - \tilde{G}^*|, \quad \Delta_n^* = \sup_{-\infty \le t \le \infty} \Delta_n^*(t).$$

It turns out that that as $n \to \infty$,

$$\sqrt{n}(\hat{G}^* - \tilde{G}^*) \xrightarrow{d} W$$

in $D[-\infty, \infty]$, where W is the Gaussian process defined in Theorem 3.8 (the proof is similar to that in Zhang[2000c]). This shows that the limit process of $\sqrt{n}(\hat{G}^* - \tilde{G}^*)$ agrees with that of $\sqrt{n}(\hat{G} - \tilde{G})$. It follows that $\Delta_n^* = \sup_{-\infty \le t \le \infty} \sqrt{n}\,|\hat{G}^* - \tilde{G}^*|$ has the same limiting behavior as does $\Delta_n = \sup_{-\infty \le t \le \infty} \sqrt{n}\,|\hat{G} - \tilde{G}|$ under model (1.11). Thus we can approximate the quantiles of Δ_n by those of Δ_n^*.

We now have the following decision rule: reject model (1.11) at level α if

$$\Delta_n > w_{1-\alpha}^n,$$

where $w_{1-\alpha}^n$ is the $(1-\alpha)$-quantile obtained from the bootstrap distribution of Δ_n^*.

We now apply the proposed goodness-of-fit test procedure to the data simulated from the previous model (4.1). First we simulated samples $X_0 \sim N(0,1), X_1 \sim N(0,2)$ and $X_2 \sim N(0,4)$ with sample sizes $(n_0, n_1, n_2) = (50, 60, 80)$. Then we obtained $(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})$ for parameters and $\hat{G}, \hat{G}_1$ and $\hat{G}_2$ for distributions from X_0, X_1, X_2

respectively under the correct model (4.1). To check the validity of model (4.1), we generated samples X_0^*, X_1^*, X_2^* from $\hat{G}, \hat{G}_1$ and $\hat{G}_2$, respectively, with the same sizes as X_0, X_1, X_2. 1000 bootstrap replications of Δ_n^* are calculated based on (X_0^*, X_1^*, X_2^*).

The estimated parameters are $(\hat{\alpha}_1, \hat{\alpha}_2, \hat{\beta}_1, \hat{\beta}_2) = (-0.576, -0.84, 0.436, 0.535)$. The value of the proposed test statistic $\Delta_n = 1.05$, and the observed p-value is $P(\Delta_n^* > 1.05) = 0.904$ based on 1000 bootstrap replications of Δ_n^*. This strongly suggests to accept model (4.1). The comparison of estimated distributions from different samples are plotted in Figure 4.1. From the plot we can see that the estimated distribution $\hat{G}$ is close to the empirical distribution $\tilde{G}$ from X_0 only, which supports our conclusion.

Next, we still use the same data generated from $X_0 \sim N(0,1), X_1 \sim N(0,2)$ and $X_2 \sim N(0,4)$, but we intentionally misspecified the model by replacing the distortion function x^2 with x, that is

$$\begin{aligned} g_1(x) &= g(x)\exp(\alpha_1 + \beta_1 x), \\ g_2(x) &= g(x)\exp(\alpha_2 + \beta_2 x). \end{aligned} \tag{4.3}$$

The estimated parameters are

$$(\hat{\alpha}_1, \hat{\alpha}_2, \hat{\beta}_1, \hat{\beta}_2) = (-0.00072, -0.03, -0.0015, 0.032).$$

The value of the proposed test statistic $\Delta_n = 2.31$, and the observed p-value is $P(\Delta_n^* > 2.31) = 0.007$ based on 1000 bootstrap replications of Δ_n^*. This implies a significant difference between the estimated distribution and the empirical distribution from the reference sample X_0, and suggests to reject model (4.3). The

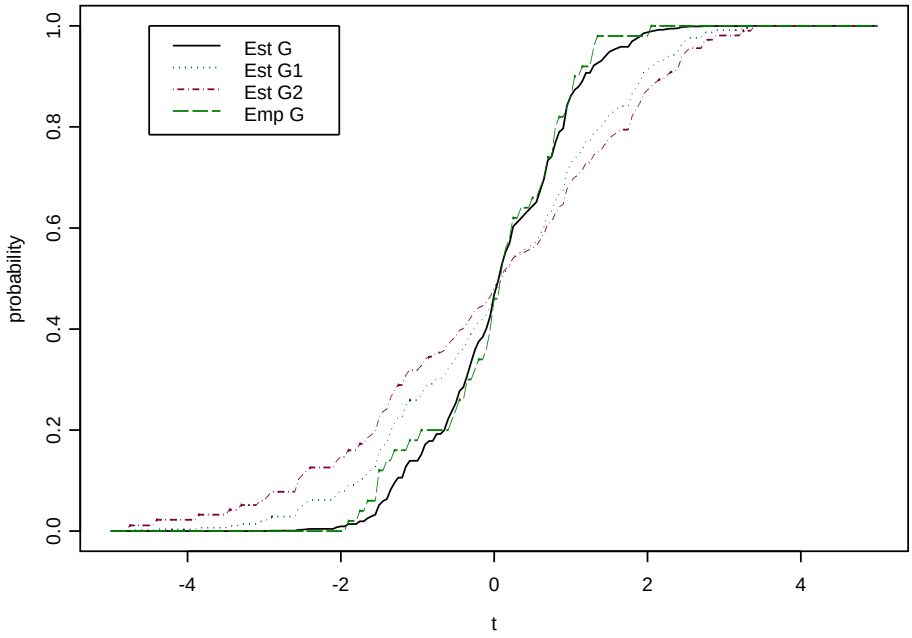

Figure 4.1: Comparison of estimated $\hat{G}$ and empirical distribution $\tilde{G}$ from X_0 only. Estimated distributions from X_0 ($\hat{G}$, solid curve), X_1 ($\hat{G}_1$, blue dotted curve), X_2 ($\hat{G}_2$, red dash-dot curve), empirical distribution $\tilde{G}$ (green dashed curve).

plot in Figure 4.2 shows the obvious difference between $\hat{G}$ and $\tilde{G}$. Moreover, Figure 4.2 indicates that the estimated distributions $\hat{G}_1$ and $\hat{G}_2$ are close to $\hat{G}$. But from the original random samples we generated, the difference among them should be significant.

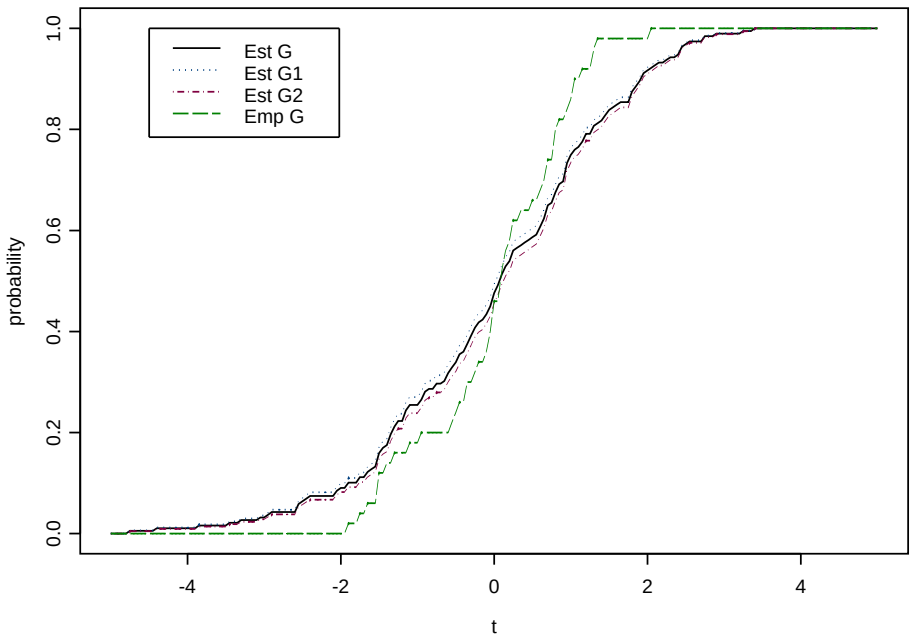

Figure 4.2: Comparison of estimated $\hat{G}$ from a misspecified model and empirical distribution $\tilde{G}$ from X_0 only. Estimated distributions from X_0 ($\hat{G}$, solid curve), X_1 ($\hat{G}_1$, blue dotted curve), X_2 ($\hat{G}_2$, red dash-dot curve), empirical distribution $\tilde{G}$ (green dashed curve).

Since the density ratio model (1.11) requires that the distortion function $h(x)$ to be specified first, it is possible that the model could be misspecified as shown in (4.3). To reduce the chance of misspecifying a model, specifying a more general

distortion function would be helpful. Let's consider again the data we simulated earlier for model (4.1). We assume that $h(x) = \exp(\alpha + \beta x + \gamma x^2)$. Then the tilted model is

$$\begin{aligned} g_1(x) &= g(x)\exp(\alpha_1 + \gamma_1 x + \beta_1 x^2), \\ g_2(x) &= g(x)\exp(\alpha_2 + \gamma_2 x + \beta_2 x^2). \end{aligned} \tag{4.4}$$

The estimated parameters are

$$(\hat{\alpha}_1, \hat{\alpha}_2, \hat{\gamma}_1, \hat{\gamma}_2, \hat{\beta}_1, \hat{\beta}_2) = (-0.562, -0.860, 0.023, 0.139, 0.427, 0.539).$$

Both $\hat{\gamma}_1$ and $\hat{\gamma}_2$ are close to 0, and the rest are close to the true parameters. The proposed test statistic $\Delta_n = 0.92$, and the observed p-value is $P(\Delta_n^* > 0.92) = 0.89$ based on 1000 bootstrap replications of Δ_n^*. This says that model (4.4) is acceptable.

4.3 Confidence Bands for G

We obtained the limiting process for $\hat{G}$ in Theorem 3.8 and Theorem 3.9, and we will demonstrate that the results can be used to construct confidence bands for G under model (1.11).

We can show that the stochastic process $\sqrt{n}(\hat{G}^* - \hat{G})$ converges weakly to the Gaussian process given in Theorem 3.9. The proof is similar to that in Zhang (2000). Therefore, the limit of $\Delta_n = \sup_{-\infty \le t \le \infty} \sqrt{n}\,|\hat{G}(t) - G(t)|$ agrees with the limit of its bootstrap counterpart $\Delta_n^* = \sup_{-\infty \le t \le \infty} \sqrt{n}\,|\hat{G}^*(t) - \hat{G}(t)|$ almost surely.

As a result, we can approximate the quantiles of Δ_n by those of the distribution of Δ_n^*.

For $\alpha \in (0,1)$, let

$$w_{1-\alpha}^n = \inf\{y | P^*(\Delta_n^* \leq y) \geq 1-\alpha\},$$

then a $1-\alpha$ level bootstrap confidence band for G is given by

$$\left(\hat{G}(\cdot) - w_{1-\alpha}^n/\sqrt{n},\ \ \hat{G}(\cdot) + w_{1-\alpha}^n/\sqrt{n}\right). \tag{4.5}$$

The confidence bands in (4.5) are forced to be symmetric and have the same width at all points regardless of the amount of data-support.

A pointwise symmetric confidence interval can be calculated by estimating the covariance matrix (3.31). Let $V(t_i)$ be the estimated variance for $\hat{G}(t_i)$ at each point t_i from (3.31). Then a $1-\alpha$ level pointwise confidence interval for $\hat{G}(t_i)$ is given by

$$\left(\hat{G}(t_i) - z_{1-\alpha/2}\sqrt{V(t_i)},\ \ \hat{G}(t_i) + z_{1-\alpha/2}\sqrt{V(t_i)}\right), \tag{4.6}$$

where $z_{1-\alpha/2}$ satisfies $P(Z \leq z_{1-\alpha/2}) = 1-\alpha/2$ with $Z \sim N(0,1)$. These confidence intervals in (4.6) do not hold simultaneously for all the data points. An alternative is the $1-\alpha$ Bonferroni simultaneous confidence intervals given by

$$\left(\hat{G}(t_i) - t_{1-\alpha/2n}^{n-1}\sqrt{V(t_i)},\ \ \hat{G}(t_i) + t_{1-\alpha/2n}^{n-1}\sqrt{V(t_i)}\right), \tag{4.7}$$

where $t_{1-\alpha/2n}^{n-1}$ is the $(1-\frac{\alpha}{2n})$ percent cutoff point of the t_{n-1} distribution with degree of freedom $n-1$. A plot of the bootstrap confidence bands, pointwise confidence intervals and Bonferroni simultaneous confidence intervals is shown in Figure 4.3.

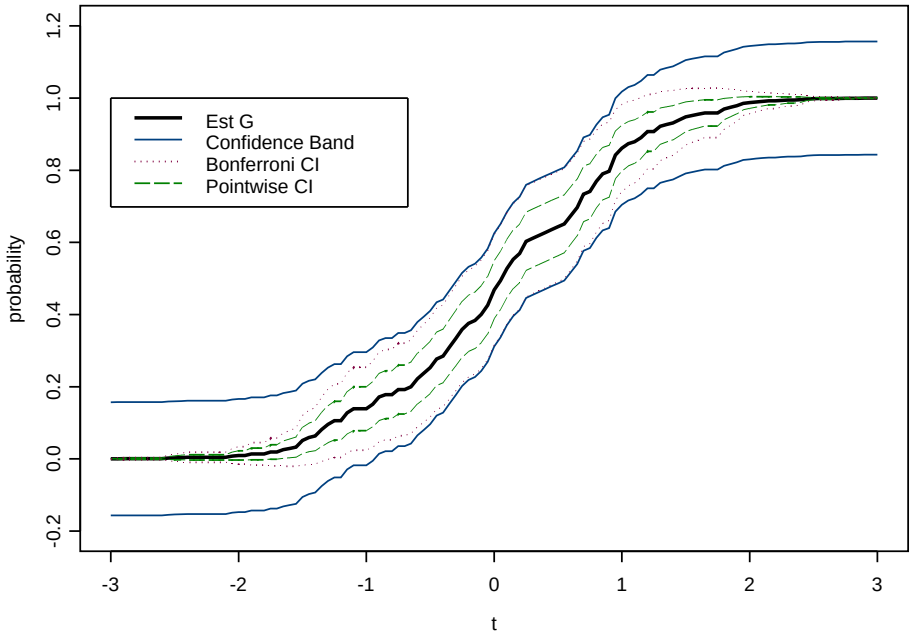

Figure 4.3: Estimated cdf $\hat{G}$ (black thick curve), 95% confidence band (blue curve), 95% Bonferroni simultaneous confidence intervals (red dotted curve), 95% pointwise confidence intervals(green dashed curve).

Chapter 5

Application of Density Ratio Models to Mortality Forecasting

5.1 Introduction

The US Government has been collecting mortality data from death registration records assembled by state vital statistics offices since 1900. The data are broken down mainly by state, race, gender, and age, and are published in the form of death rates and life expectancies decennially and/or annually for over 100 years. The process of collecting and recording these data evolved over time. Thus, before 1933, mortality data were not from the US as a whole but only from death registration areas, and electronically documented mortality data are available only from the late 1960's on. In this study we shall use well documented mortality time series from 1970 to 2002. This gives us relatively short annual time series consisting of a little over 30 observations for each given age, stratified by factors such as state, gender and race. Prediction of future annual death rates based on these time series must take into account their short length.

The National Center for Health Statistics (NCHS), a US Government agency, compiles statistical information important for public health and health policy. An important goal of NCHS is forecasting annual age specific mortality patterns. The objective of the present study is to address this problem, using relatively short historical time records, by following a two-stage procedure whereby each short series is modeled as a first order autoregressive process, and then the resulting residuals are combined or merged in some fashion to provide estimates of future predictive distributions.

Addressing short time records in a forecasting problem, in this paper we apply a semiparametric forecasting method advanced recently in [17]. The method compensates for short individual records by combining them via a *density ratio model* as described in Section 5.2. Accordingly, the residuals from several different fitted models are combined in this way in order to estimate the entire future conditional distributions of interest. From this we obtain future conditional probabilities as well as the conditional expectation of future values given past information, the most common predictor. We focus primarily on the prediction of centered annual age-specific log death-rates for the entire US using data from 1970 to 2002.

The plot of US age-specific annual mortality as a function of age resembles a pointed hook with a rather long handle, surprisingly similar to a dentist probe, as seen from Figure 5.1. Wei [30] fitted to these data the eight parameter model of Heligman and Pollard (HP) [15], demonstrating that the HP model captures well the pointed hook pattern of mortality versus age. Figure 5.1 also shows that the hook pattern repeats itself year after year persistently, and that in general annual

death-rates decline, again quite persistently. The decline in death-rate for fixed age as a function of time is shown in Figure 5.2 on a log-scale for several ages. These time series appear almost as parallel straight lines, but when drawn separately they are much more oscillatory.

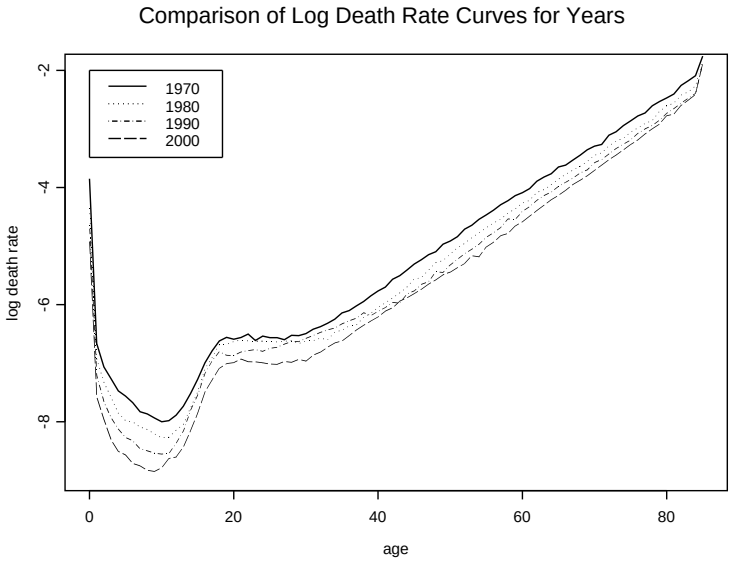

Figure 5.1: Log death-rate as a function of age

Let $m(x,t)$ denote the log death-rate matrix, where x and t are indices for age and time (by year), respectively, and let $d(x,t)$ denote the centered log death-rate matrix, $d(x,t) = m(x,t) - \sum_t m(x,t)/n$. We model $d(x,t)$ instead of $m(x,t)$ in order to compare our method with that of Lee and Carter (LC) [19] who also use centered data. Plots of $d(x,t)$ are shown in Figure 5.3 as a function of x for some fixed t, and also as a function of t for various fixed ages x. From the plots we see that neighboring time series $d(x,t)$ and $d(x',t)$, where x and x' are close, e.g. ages 60 and 61, behave quite similarly. To compensate for short time records,

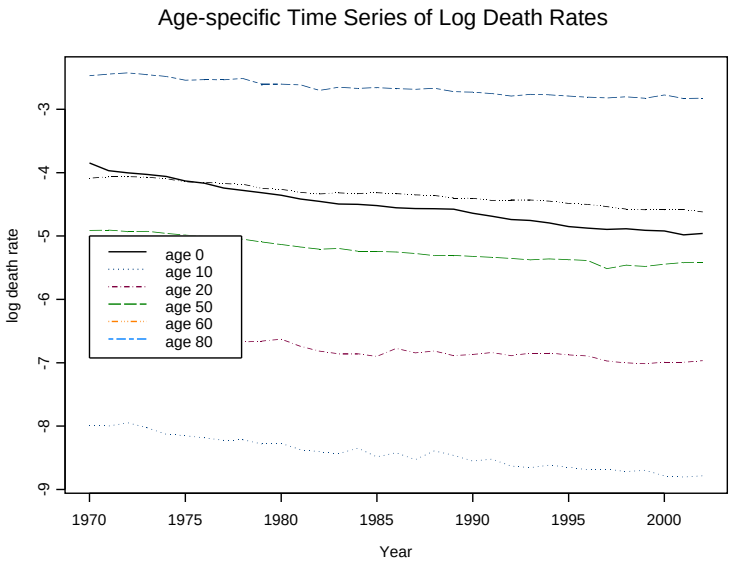

Figure 5.2: Age-specific time series

the semiparametric method combines information from several age-wise neighboring time series.

5.1.1 The Lee-Carter Model

The model proposed by Lee and Carter [1992] is used by the U.S. Census Bureau as a benchmark for their population forecasts, and its use has been recommended by the two most recent U.S. Social Security Technical Advisory Panels. It also appears to be the dominant method in the academic literature and is used widely by scholars forecasting all-cause and cause-specific mortality around the world. See [18],[20]. The LC model is based on principal components. If n denotes the number of mortality time series, each corresponding to a specific age, the LC model searches for the first principal component in n dimensional time series data, and solves for

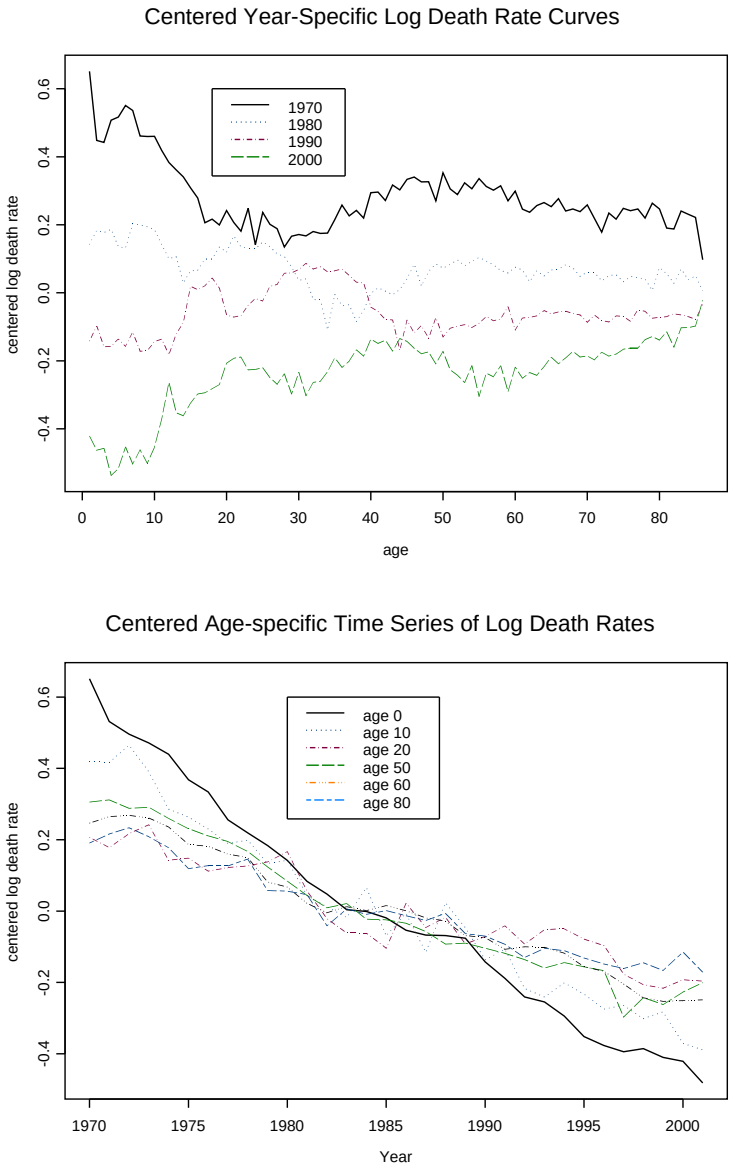

Figure 5.3: Plots of centered log death-rates $d(x,t)$ as a function of x (top) for fixed t and vice versa (bottom).

the age and time parameters by using singular value decomposition. A careful examination of the LC model was undertaken by Girosi and King [14]. Using USA mortality data they found that the LC model explains approximately 93% of the variance of death data from all causes, 90% from cardiovascular, and only 63% of the variance in stomach cancer death. They bring examples from other countries where the first principal component accounts for a much smaller percent of the variance. This suggests that the first principal component may be insufficient when explaining the variance in high-dimensional data, and that there are applications where it is beneficial to add principal components beyond the first one. The method presented in this paper is very different and seems appropriate for short range forecasting. Both methods, however, are extrapolative in the sense that future mortality rates are estimated from past rates. Lee and Carter employed tabulated mortality data available from 1900 to 1987. However we shall compare the two methods using the systematically collected annual data from 1970 to 2002.

5.2 An Approach to Semiparametric Time Series Forecasting

Our approach for tackling the problem of short time series is based on the semiparametric density ratio model (1.11).

5.2.1 The Density Ratio Model in Time Series Regressions

Consider the following $m = q + 1$ time series regressions,

$$\begin{aligned} x_{1t} &= f_1(z_{1,t-1}) + \epsilon_{1t}, \quad t = 1, ..., n_1 \\ &\vdots \\ x_{qt} &= f_q(z_{q,t-1}) + \epsilon_{qt}, \quad t = 1, ..., n_q \\ x_{mt} &= f_m(z_{m,t-1}) + \epsilon_{mt}, \quad t = 1, ..., n_m \end{aligned} \tag{5.1}$$

where the n_i are assumed to be small numbers, the vectors $z_{i,t-1}$ contain past values of covariate time series, and the ϵ_{kt} are independent noise components. In practice, the f_i's are fitted first using the data and the ϵ_{kt}'s are the residuals. Suppose the ϵ_{kt} have probability densities,

$$\begin{aligned} \epsilon_{1t} &\sim g_1(x), \quad t = 1, ..., n_1 \\ &\vdots \\ \epsilon_{qt} &\sim g_q(x), \quad t = 1, ..., n_q \\ \epsilon_{mt} &\sim g_m(x), \quad t = 1, ..., n_m \end{aligned} \tag{5.2}$$

Define the *reference* density $g(x) \equiv g_m(x)$ with $G(x) \equiv G_m(x)$ the corresponding cdf. Then we shall assume the *density ratio model* relative to the reference $g(x)$ as given by (1.11),

$$g_j(x) = e^{\alpha_j + \beta_j h(x)} g(x), \quad j = 1, ..., q. \tag{5.3}$$

with scalar α_j, vector β_j, and a vector valued distortion or tilt function $h(x)$. The "distorted" densities g_j, the reference g, as well as the α_j and β_j are all unknown,

but the distortion function $h(x)$ is assumed known and its choice depends on the data.

We combine all the residuals from the $q+1$ regressions into a single vector of length $n = n_1 + \cdots + n_q + n_m$,

$$\mathbf{t} = (t_1, \ldots, t_n)' = \{(\epsilon_{1t}, \ldots, \epsilon_{1n_1}), \ldots, (\epsilon_{qt}, \ldots, \epsilon_{qn_q}), (\epsilon_{mt}, \ldots, \epsilon_{mn_m})\}'. \qquad (5.4)$$

Maximum likelihood estimates for the α_j, β_j, and $G(x)$ can be obtained from the entire vector of residuals $\mathbf{t} = (t_1, \ldots, t_n)'$ by maximizing the likelihood over a class of step cdf's with jumps at the values $t_1, \ldots, t_n$ as in section 1.3.

The main point of the semiparametric paradigm discussed here is that the reference cdf $G(x)$ is estimated from many samples giving an improved estimate as compared with the empirical cdf which is obtained from the reference sample only. This fact has been proved from the asymptotic theory for both $\hat{\boldsymbol{\theta}}$ and $\hat{G}$ in Chapters 3 and 4. Likewise, we have also shown that quantile estimates obtained by the semiparametric method from both case and control samples are more efficient than estimates that are based on the control sample only, ignoring the case information. More recently Fokianos [2004] showed that by merging information following the semiparametric paradigm we obtain improved kernel density estimates with the same bias as the traditional kernel density estimates but with smaller asymptotic variance. Our data analysis below supports this fact. Moreover, merging information in this way can result in more powerful tests for distribution equality [9],[10]. Regarding the uncertainty in $\hat{G}$, we showed that $\sqrt{n}(\hat{G} - G)$ converges to a Gaussian process with mean zero and a rather complex covariance structure given by (3.31).

We shall apply the semiparametric paradigm in forecasting US mortality rates by combining information, or borrowing strength, from several short US mortality time series.

5.2.2 Forecasting

The preceding discussion motivates the following time series forecasting method [17]. Since $x_{m,t+1} = f_m(z_{m,t}) + \epsilon_{m,t+1}$, and $\epsilon_{m,t+1} \sim G$, where G is the reference distribution estimated semiparametrically by $\hat{G}$ as in (1.26), we have an approximation for the predictive distribution at time $t+1$ conditional on past data $z_{m,t}$,

$$\begin{aligned} P(x_{m,t+1} \leq x \mid z_{m,t}) &= P(x_{m,t+1} - f_m(z_{m,t}) \leq x - f_m(z_{m,t}) \mid z_{m,t}) \\ &= P(\epsilon_{m,t+1} \leq x - f_m(z_{m,t}) \mid z_{m,t}) \\ &= G(x - f_m(z_{m,t})) \\ &\approx \hat{G}(x - f_m(z_{m,t})) \end{aligned} \tag{5.5}$$

All sorts of point predictors can be obtained from (5.5). In particular, a one-step ahead predictor for $x_{m,t+1}$ given the past can be approximated by calculating the (conditional) mean of the shifted distribution $\hat{G}(x - f_m(z_{m,t}))$. Approximate prediction intervals can also be obtained from the estimated distribution.

5.3 One Year Ahead Prediction of US mortality

5.3.1 A Two Stage Procedure

Define $a_k = \sum_t m(k,t)/n$. As mentioned above, we analyze the centered log death-rate matrix $d(k,t)$, $d(k,t) = m(k,t) - a_k$. For each fixed age k, consider the annual time series of centered log death-rates from 1970 to 2001. Thus $t = 1, 2, ..., 32$.

Write $x_{kt} = d(k,t)$. First, to each such time series we fit the first order autoregressive model with drift c_k,

$$x_{kt} = b_k x_{k,t-1} + c_k + \epsilon_{kt}, \quad k = 1, \ldots, q, m. \tag{5.6}$$

The drift parameter is added in order to capture a downward trend observed in the age-specific centered log death-rate time series as exemplified in Figure 5.3. The coefficient b_k and the drift c_k are estimated by least squares, and in our application the ϵ_{kt} are replaced by the residuals derived from the model (5.6). Accordingly, the functions f_k in the system (5.1) are given by $f_k(x_{k,t-1}) = b_k x_{k,t-1} + c_k$.

Next, we choose a density ratio model for the residuals. Figure 5.4 shows the centered log death-rate curves, fitted time series, and histograms of the residuals. We can see that the residuals are centered around zero and that their histograms resemble those obtained from small normal samples. This motivates the distortion model (1.14) with $h(x) = x^2$.

We consider the mth "sample" $(\epsilon_{m1}, \ldots, \epsilon_{mn_m})$ as the reference with distribution function G and density g. Similarly, assume $(\epsilon_{k1}, \ldots, \epsilon_{kn_k})$ has distribution G_k and density g_k, $k = 1, \ldots, q$. Following the above semiparametric paradigm, and

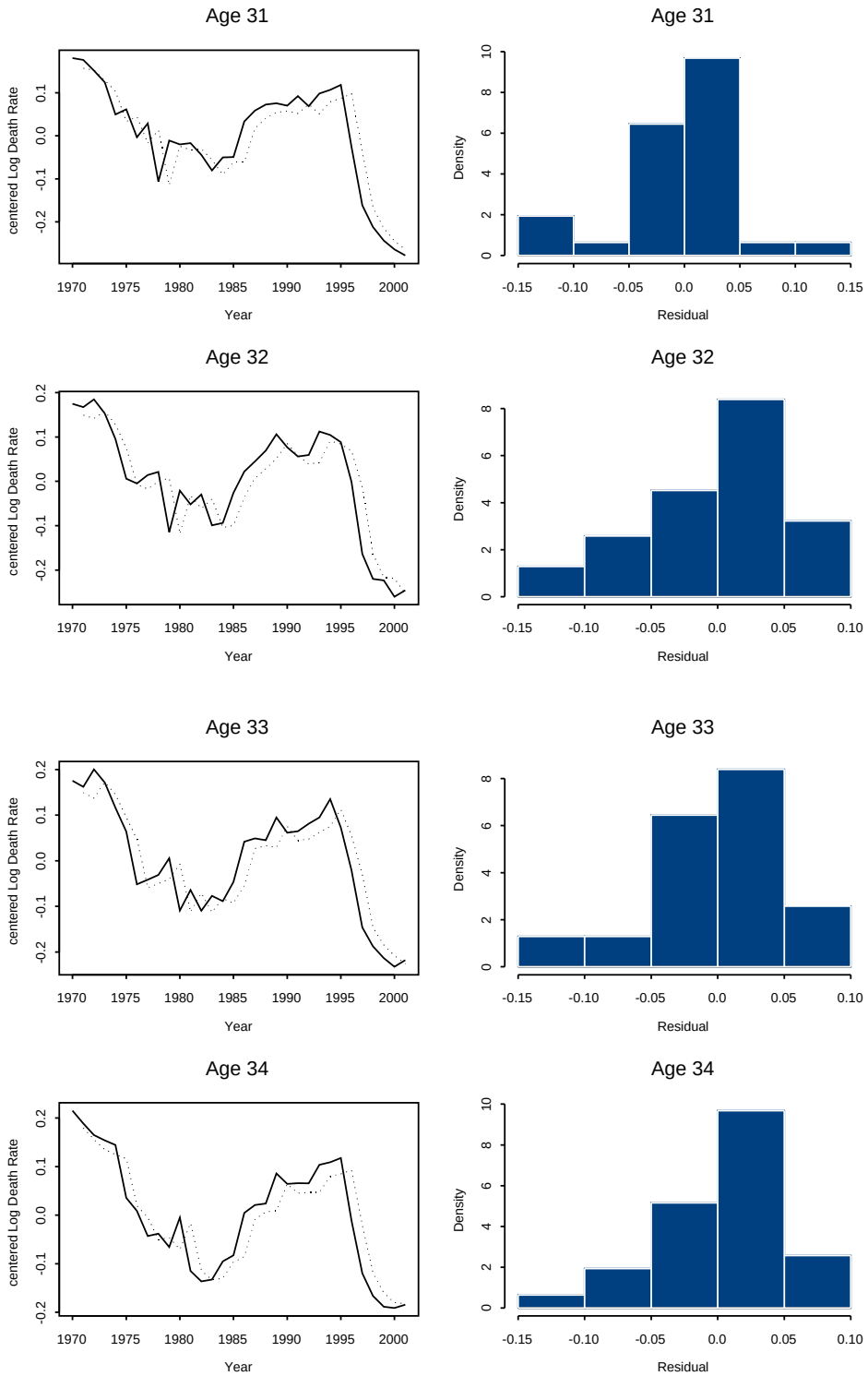

Figure 5.4: Plots of age-specific time series (solid line), fitted values (dotted line), and histograms of the resulting residuals

combining it with the insight gained from the histograms in Figure 5.4, we assume the density ratio relationship,

$$g_k(x) = e^{\alpha_k + \beta_k x^2} g(x), \quad k = 1, ..., q. \tag{5.7}$$

An application of the semiparametric procedure to the combined data $\mathbf{t}$ defined by (5.4) gives the semiparametric estimate $\hat{G}$ for the reference distribution. Similarly, from (1.26) and (5.7) we obtain the estimated distribution function of the kth sample $\hat{G}_k$ from which the predictive distribution is computed by

$$P(x_{k,t+1} \leq x \mid z_{kt}) \approx \hat{G}_k(x - b_k x_{kt} - c_k). \tag{5.8}$$

5.3.2 Data Analysis

We consider 85 age-specific time series of log death-rates (all-cause) for ages $1, ..., 85$, where the age category 85 includes ages 85 and older. To simplify the analysis, this grouping or lumping of ages 85 and older had to be done at some point and we chose, somewhat arbitrarily, age 85 as a threshold. However, the data file does have the specificity to subdivide this category to obtain a more detailed mortality prediction. Mortality at age 0 is not considered in the present analysis due to its behavior which is very different from that at other ages. See Figure 5.2.

From the previous discussion the assumption that the density ratio model (5.7) holds for time series groups corresponding to neighboring ages seems reasonable. Indeed, in retrospect our data analysis lends credence to this assumption. In our analysis, therefore, we apply the semiparametric method by combining information from each of the age groups, consisting of five ages each and dubbed "5-age", $1 -$

$5, 6-10, ..., 81-85$, a total of 17 groups, where the time series "in the middle" of each group is taken as the reference. For example, in the group $1-5$, the time series of age 3 is taken as the reference, meaning that the relevant distribution from this time series serves as the reference distribution for the group. We applied the semiparametric model separately to each group to estimate the reference distribution and the corresponding distorted distributions to obtain predicted mortality curves.

For illustration, consider the age group $31-35$ from 1970 to 2001. The fitted AR(1) curves and histograms of the residuals are plotted in Figure 5.4. As mentioned before, we chose a quadratic distortion function $h(x) = x^2$ due to the rough symmetry of the residuals around zero, resembling the behavior of normal residuals. There are altogether 5 residual samples, and the sample of residuals from age 33 is considered as the reference. The actual conditional point predictions of log death-rate in 2002 for the age group $31-35$ are obtained from (5.8) by computing the first moments of the shifted predictive distributions $\hat{G}_k, k = 31, ..., 35$, respectively, with $\hat{G_{33}}$ as the reference. This analysis is repeated for all 17 groups. The 2002 prediction results for all ages are compared with the true 2002 log-rates in the tables below.

It is also interesting to compare the results from the 3-age group $32-34$ with the 5-age group $31-35$. Figure 5.5 shows the histograms and overlaid estimated reference density of age 33 obtained from the combined data $\mathbf{t}$ for the 3-age and 5-age groups. Since we combined more information in the 5-age group there is a noticeable improvement in the density fit.

For age group $31-35$, the estimated tilted cdf's $\hat{G}_k(x)$ from (5.7), each estimated from $5 \times 32 = 160$ residuals, and the corresponding empirical cdf's, each from

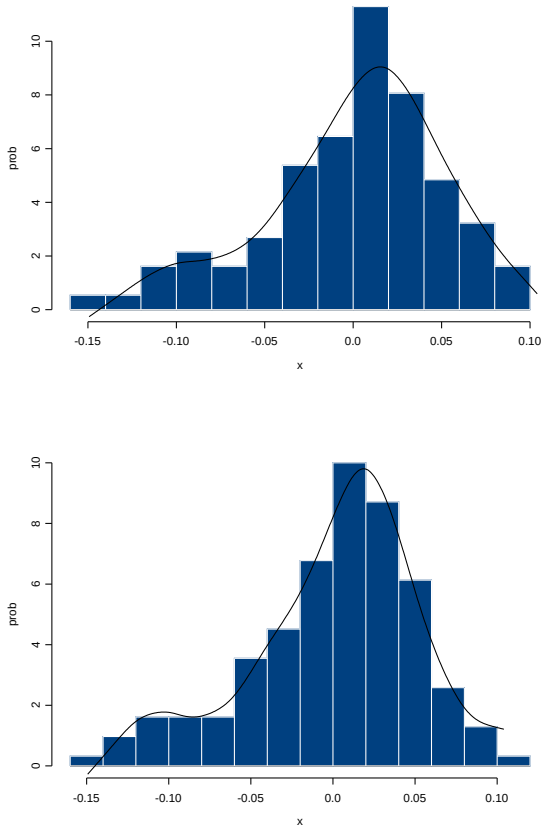

Figure 5.5: Estimated reference pdf of age 33 from the combined data $\mathbf{t}$ for the 3-age group 32-34 (above), and the 5-age group 31-35 (below), respectively.

32 residuals, are shown in Figures 5.6 for ages $31, 32, 33, 34$ (the cdf for age 35 is not plotted). Since more information is used (or combined) in deriving the $\hat{G}_k(x)$ than used in obtaining the empirical distributions, the $\hat{G}_k(x)$ are smoother as is evident from the figure. So, in some sense, the semiparametric cdf's are smooth versions of the corresponding empirical distributions.

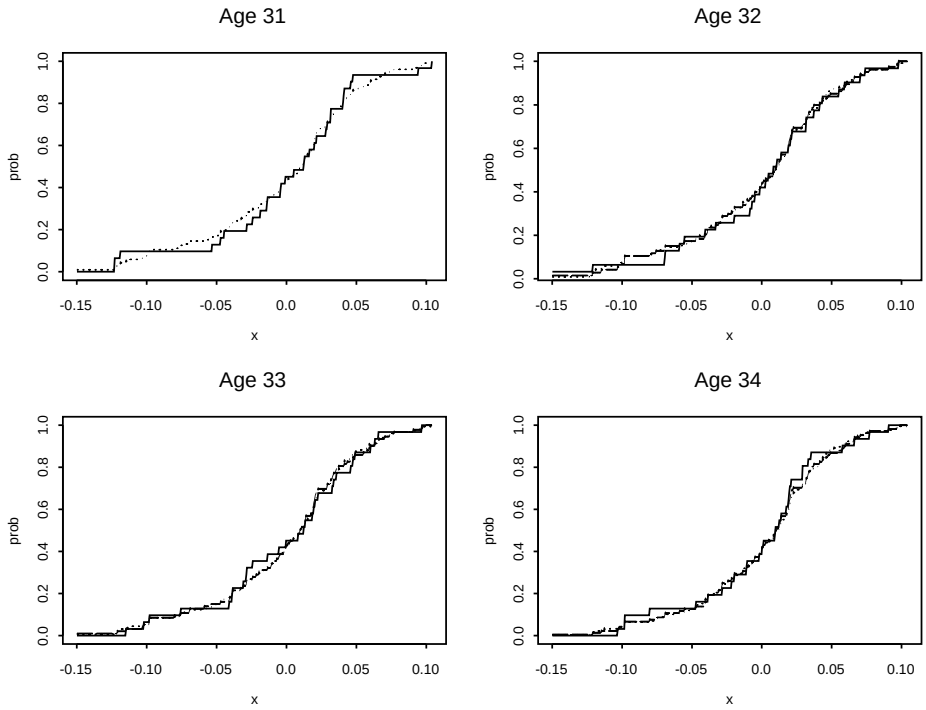

Figure 5.6: Comparison of the empirical (solid line) and estimated (dotted line) cdf's for the indicated ages. The estimated cdf for age 35 is not shown.

The corresponding 5-age smoothed pdf's $\hat{g}_k(x)$ (solid lines) and their related histograms are shown in Figure 5.7 for ages $31, 32, 33, 34$. For the sake of comparison, for ages $32, 33, 34$ the figure also depicts the 3-age smoothed $\hat{g}_k(x)$ (dotted lines). The estimated pdf for age 35 is not shown. The plots point to the consistency of

the method in the sense that the 3-age and 5-age estimates are not far apart.

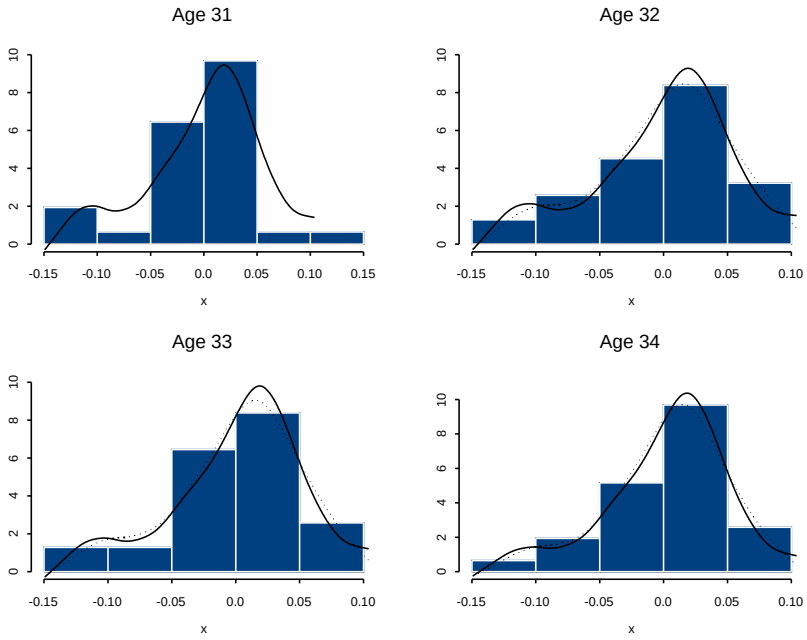

Figure 5.7: Histograms and overlaid estimated pdf's for 3-age group $32-34$ (dotted line) and 5-age group $31-35$ (solid line). The estimated pdf for age 35 is not shown.

Once the the estimated distributions $\hat{G}_k(x)$ are obtained, we apply (5.8) to approximate the probability distribution of the one-year-ahead log death-rate in 2002 for for the age group 31-35. The predictive distributions are shown in Figure 5.8. The corresponding estimated probability densities are shown in Figure 5.9.

Moreover, for each year, future conditional probabilities that the death rate is less than a given value can also be computed from the estimated predictive distributions as shown in Figure 5.10 for the age group 51-55.

As a point predictor we use the mean of the predictive distribution, that is, the conditional expectation. The corresponding 95% confidence interval is also derived

from the estimated predictive distribution.

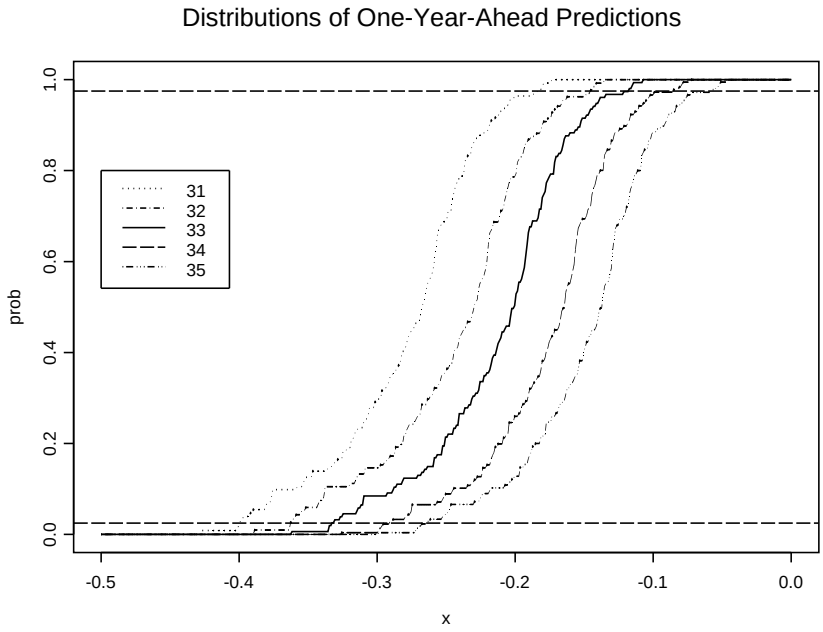

Figure 5.8: Estimated 2002 predictive distribution functions for the age group 31-35. The 95% PI's are bounded by two horizontal dashed lines

The semiparametric predictions from all the age groups for the year 2002 and the corresponding 95% prediction intervals are shown in Figures 5.11 and 5.12. The figures also show the 2002 predicted mortality curve from the Lee-Carter model [19]. Figure 5.11 shows the predicted mortality curves for all ages, and the four plots in Figure 5.12 are magnifications of sections of Figure 5.11. Table 5.1 is the numerical counterpart of Figure 5.11. For each age it gives the semiparametric prediction for 2002 and the corresponding prediction interval (PI), as well as the Lee-Carter prediction. Comparison by mean square error (MSE) between the two methods is given in Table 5.2. Generally speaking, compared with the Lee-Carter

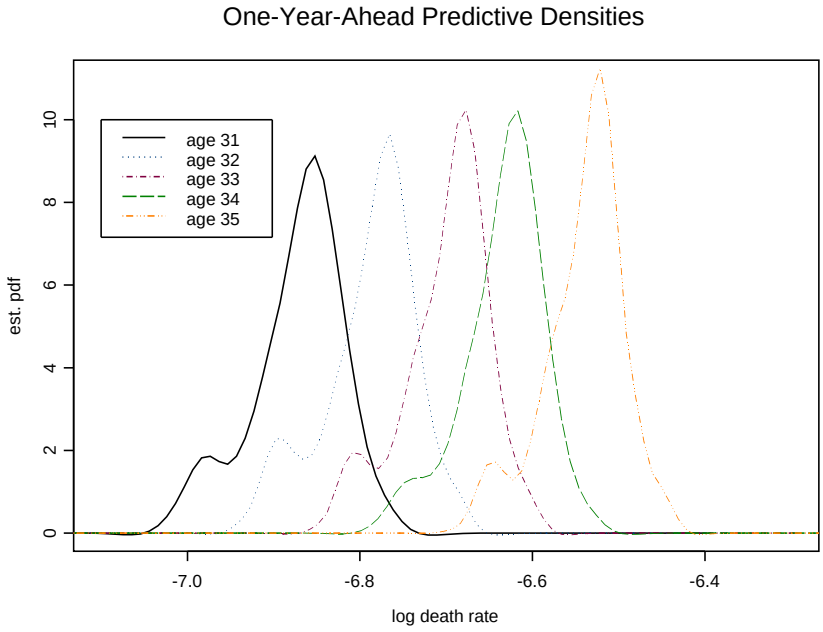

Figure 5.9: Estimated 2002 predictive probability densities for the age group 31-35.

method, the semiparametric method improves the prediction as measured by MSE. As seen from Figure Figure 5.11 and Table 5.1, almost all the true values fall within their respective 95% prediction intervals. The improvement of the semiparametric method is more noticeable for age groups which display more steady and gradual change of death-rate as in age groups 31-50 and 71-85. From Table 5.2, the overall prediction MSE from the semiparametric method is 0.104 compared to 0.297 from the Lee-Carter method. The most significant improvement is for the age groups 31-50 and 71-85, whereas both methods perform quite similarly for all other groups as we see from Table 5.2.

In the above data analysis we combined information from non-overlapping 5-age groups. The analysis was repeated by using a sliding window of over-

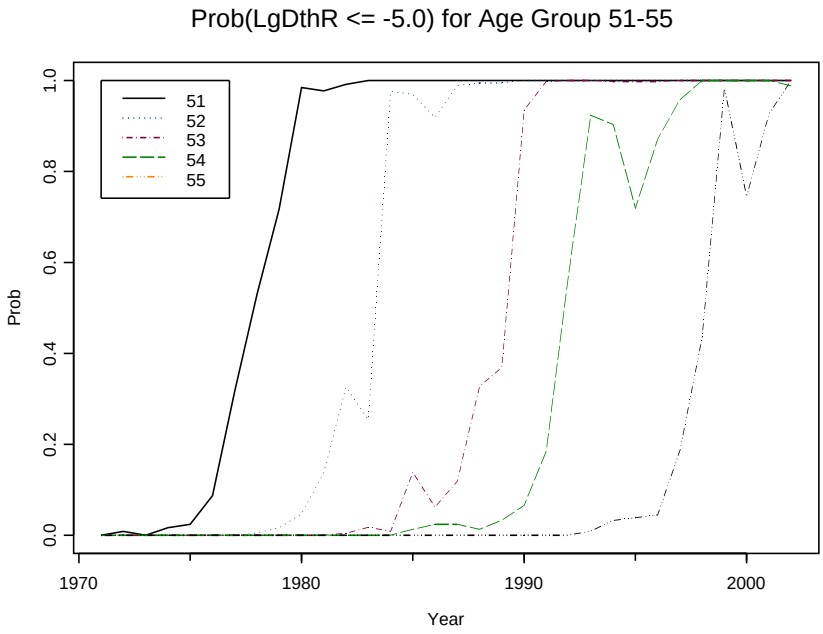

Figure 5.10: Estimated future conditional probability that log death-rate is less than -5 for age group 51-55

lapping 5-age groups, each time moving up by a single year. Interestingly, the MSE results were very close to those reported in Table 5.2, replacing the SP row (0.104,0.050,0.015,0.030,0.009) by (0.105, 0.051, 0.014, 0.031, 0.008). This suggests that the choice of the reference time series within an age group may be arbitrary.

We also consider prediction for gender (all-cause female only) and race-specific (all-cause white female only) mortality rate. Comparison by MSE between the semiparametric and Lee-Carter methods is given in Tables 5.3 and 5.4. Again, the overall prediction MSE from the semiparametric method is appreciably smaller than that from the Lee-Carter method.

Interestingly, we see that the MSE from the semiparametric method is lower in Table 5.2 than in both Tables 5.3 and 5.4. This is not surprising since more data are available from the total population, whereas in the other two cases we deal with subpopulations. The fact that in the three Tables 5.2,5.3, and 5.4, age group 1-30 has a larger MSE than that from the other groups is due to the large variation of the data associated with that age group.

Since our mortality data are truncated at age 85, we cannot calculate traditional life tables from the death rate forecasts. Instead we provide in Table 5.5 a comparison between the true and predicted (by our method) number of survivors by age and sex out of $100,000$ live births. The true values and their forecasts are close.

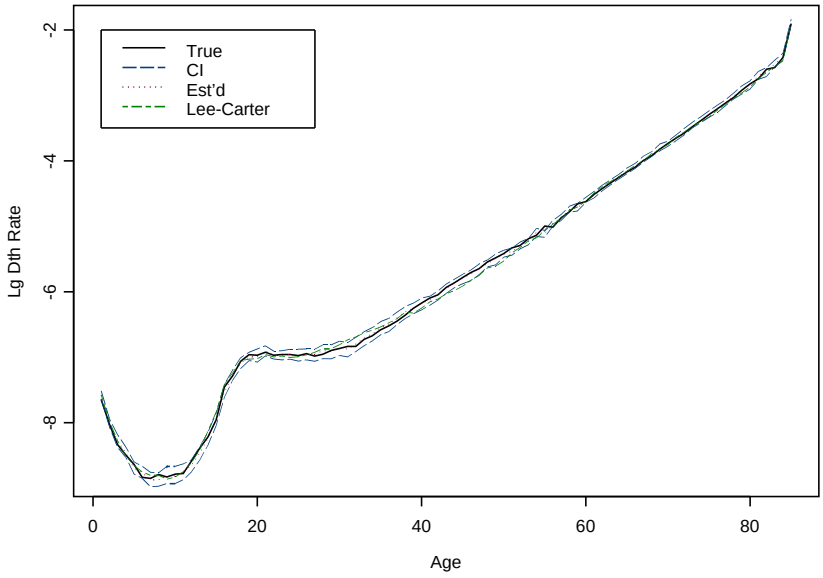

Figure 5.11: Predicted mortality curves from the Lee-Carter model (dash-dot) and the semiparametric method (dot), and 95% CI bounds (dash) for 2002

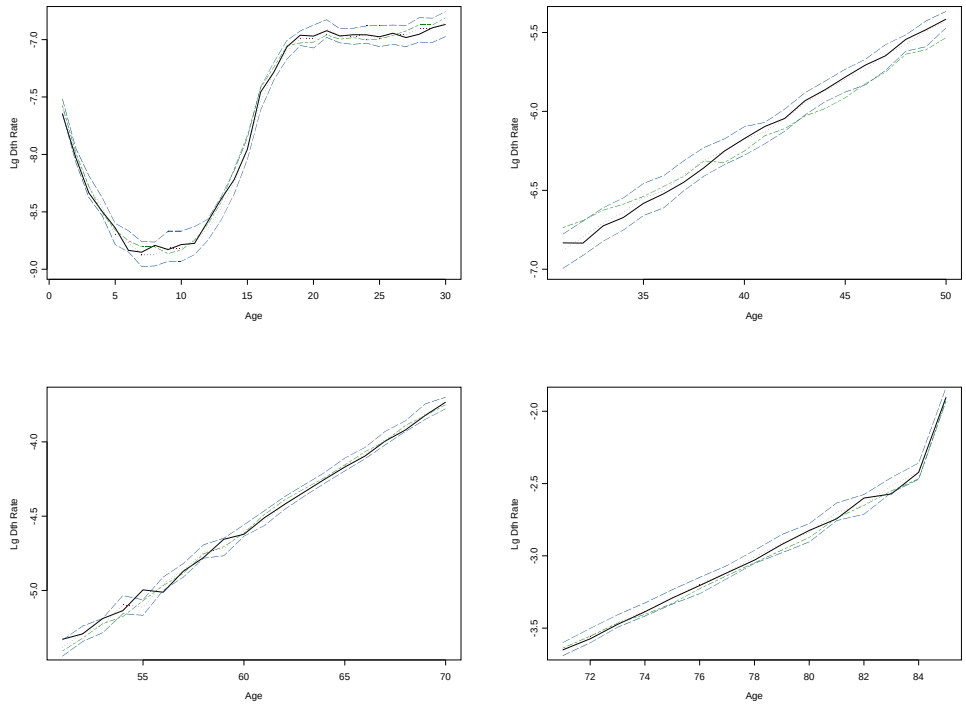

Figure 5.12: Predicted mortality curves (by part) from the Lee-Carter model (dash-dot) and the semiparametric method (dot), and 95% CI bounds (dash) for 2002

Table 5.1: Prediction comparison between the semiparametric and Lee-Carter methods for 2002. The first two rows give the 95% PI bounds for the semiparametric forecasts, and the rest are the prediction from the semiparametric method (SP), true values in 2002, and the prediction from the Lee-Carter (LC) method.

Age	1	5	10	15	20	25	30	35	40
Lower	-7.637	-8.781	-8.933	-8.038	-7.072	-7.059	-6.977	-6.653	-6.276
Upper	-7.516	-8.599	-8.671	-7.856	-6.870	-6.877	-6.755	-6.451	-6.094
SP	-7.583	-8.699	-8.819	-7.946	-6.997	-6.991	-6.858	-6.551	-6.178
True	-7.646	-8.639	-8.785	-7.955	-6.970	-6.975	-6.868	-6.583	-6.172
LC	-7.576	-8.661	-8.835	-7.834	-7.023	-6.996	-6.810	-6.540	-6.252
Age	45	50	55	60	65	70	75	80	85
Lower	-5.872	-5.473	-5.165	-4.628	-4.189	-3.776	-3.328	-2.905	-1.926
Upper	-5.731	-5.362	-5.064	-4.557	-4.109	-3.695	-3.227	-2.774	-1.835
SP	-5.799	-5.431	-5.115	-4.601	-4.166	-3.749	-3.293	-2.842	-1.897
True	-5.783	-5.416	-4.996	-4.622	-4.169	-3.733	-3.291	-2.824	-1.903
LC	-5.915	-5.534	-5.071	-4.615	-4.157	-3.752	-3.331	-2.874	-1.919

Table 5.2: Mean square error of (all-cause) prediction from the semiparametric (SP) and Lee-Carter (LC) methods.

Mean Square Error					
Age group	1-85	1-30	31-50	51-70	71-85
SP model	0.104	0.050	0.015	0.030	0.009
LC model	0.297	0.078	0.180	0.026	0.013

Table 5.3: Mean square error of prediction from the semiparametric (SP) and Lee-Carter (LC) methods for female.

Mean Square Error					
Age group	1-85	1-30	31-50	51-70	71-85
SP model	0.187	0.121	0.026	0.032	0.008
LC model	0.619	0.226	0.341	0.027	0.025

Table 5.4: Mean square error of prediction from the semiparametric (SP) and Lee-Carter (LC) methods for white female.

Mean Square Error					
Age group	1-85	1-30	31-50	51-70	71-85
SP model	0.249	0.176	0.031	0.033	0.007
LC model	0.645	0.257	0.329	0.041	0.019

Table 5.5: Number of survivors by age and sex, out of 100,000 born alive, from both SP forecasts and true values in 2002

	forecast			true		
Age	total	male	female	total	male	female
0	100000	100000	100000	100000	100000	100000
1	99311	99231	99371	99298	99217	99360
5	99182	99085	99256	99174	99076	99252
10	99107	99004	99186	99098	98992	99184
15	99013	98890	99108	99000	98875	99104
20	98685	98425	98914	98662	98400	98902
25	98219	97717	98679	98192	97691	98662
30	97746	97031	98394	97722	97006	98384
35	97189	96275	98012	97171	96249	98009
40	96384	95221	97425	96386	95228	97422
45	95231	93739	96550	95216	93733	96520
50	93558	91581	95293	93515	91553	95220
55	91205	88625	93446	91128	88521	93381
60	87762	84429	90632	87629	84211	90570
65	82616	78258	86313	82484	77986	86328
70	75218	69571	79978	75148	69339	80074
75	65081	57967	71052	65014	57710	71164
80	51665	43306	58630	51680	43142	58758
85	35348	26897	42244	35442	26938	42330

5.3.3 Two-Year Ahead Forecasting

So far we discussed one-year ahead prediction. However, our two-step procedure can be extended to multi-year ahead forecasting. One way to proceed is to use the predicted values from previous steps when making long terms predictions. Thus in two-year ahead forecasting we use the previous one-year ahead forecasts, and proceed as above. The prediction error may get amplified through each additional step even if minor deviation of prediction from true values occur. The results from this procedure are reported in Table 5.6. Again, the overall MSE is lower for the semiparametric method as compared with the Lee-Carter method.

A second procedure to forecast j-years ahead is to apply the one-step ahead forecasting method to residuals resulting from time series regression models where the present is regressed on the j previous values. Thus in the present case, to get two-year ahead mortality forecasts we use (5.6) with the modification that x_{kt} is regressed on $x_{k,t-2}$. The MSE from this method is reported in Table 5.7. Once more, the overall MSE is lower for the semiparametric method as compared with the Lee-Carter method. The disadvantage of this procedure is that some data are lost due to the larger lag.

5.4 Concluding Remarks

We have used a two-stage forecasting semiparametric procedure suitable for short time series to obtain forecasts of US age-specific mortality rates. To estimate conditional predictive distributions, the method combines short time series by ap-

Table 5.6: Prediction MSE from the semiparametric (SP) and Lee-Carter (LC) methods for two-year ahead forecasting: Predicted one-year ahead forecasts are used.

	Mean Square Error				
Age group	1-85	1-31	31-50	51-70	71-85
SP model	0.180	0.128	0.019	0.026	0.007
LC model	0.389	0.088	0.246	0.033	0.021

Table 5.7: Prediction MSE from the semiparametric (SP) and Lee-Carter (LC) methods for two-year ahead forecasting: Autoregression lagged by 2.

	Mean Square Error				
Age group	1-85	1-31	31-50	51-70	71-85
SP model	0.211	0.132	0.048	0.025	0.005
LC model	0.389	0.088	0.246	0.033	0.021

pealing to a density ratio model. Point predictors as well as future probabilities can be obtained from the estimated conditional distributions. A comparison with the well known Lee-Carter singular value decomposition method points to the potential of the semiparametric method. In general the semiparametric method provides more precise short term prediction as compared with the Lee-Carter procedure.

The method we used is non-Bayesian. Bayesian methods for forecasting in short time series are available, a useful special case of which is discussed [3],[16]. In general Bayesian methods result in relatively large prediction intervals.

Death rates drop rapidly from infants to children, thus, combining data from this age with other ages to form an age group is less appealing. It seems preferable to employ methods suitable for univariate time series to forecast the annual mortality for age zero. When monthly infant death rates are available, the semiparametric method can be applied to these data to cover age 0 separately.

For convenience, we chose to fit to the mortality rate time series the AR(1) model (5.6). This of course is only one possibility and there are other choices. For example, we could set the coefficient b_k in (5.6) to be 1, or use an AR(2) model. A model which provides a better fit could reduce the prediction error.

BIBLIOGRAPHY

[1] Billingsley, P. (1999). *Convergence of Probability Measures*, New York: Wiley.

[2] Billingsley, P. (1995). *Probability and Measure*, New York: Wiley, third edition.

[3] De Oliveira, V., Kedem, B., and Short, D. (1997). Bayesian prediction of transformed Gaussian random fields. *Jour. Americ. Statist. Assoc.*, Vol. 92, 1422-1433.

[4] Donsker, M.D. (1952). Justification and extension of Doob's heuristic approach to the Kolmogorov-Smirnov theorems, *Annals of Mathematical Statistics*, Vol. 23, 277-281.

[5] Dudley, R.M. (1978). Central limit theorems for empirical measures, *Annals of Probability*, Vol. 6(6), 899-929.

[6] Dudley, R.M. (1999). Uniform Central Limit Theorems, *Cambridge Studies in Advanced Mathematics*, 63, Cambridge University Press, Cambridge, UK.

[7] Ethier, S. N. and Kurtz, T. G. (1985). *Markov Processes: Characterization and Convergence*, New York: Wiley.

[8] Fokianos, K. (2004). Merging information for semiparametric density estimation. *J. R. Statist. Soc.* B, Vol. 66, 941-958.

[9] Fokianos, K., Kedem, B., Qin, J., and Short, D.A. (2001). A semiparametric approach to the one-way layout. *Technometrics*, Vol. 43, 56-65.

[10] Gagnon, R.E. (2005). Certain Computational Aspects of Power Efficiency and State Space Models. Ph.D. Dissertation, University of Maryland, College Park.

[11] Gilbert, P.B. (2000). Large sample theory of maximum likelihood estimates in semiparametric biased sampling models. *Annals of Statistics*, Vol. 28, 151-194.

[12] Gilbert, P.B., Lele, S.R. and Vardi, Y. (1999). Maximum likelihood estimation in semiparametric selection bias models with application to AIDS vaccine trials. *Biometrika*, Vol. 86, 27-43.

[13] Gill, R. D., Vardi, Y. and Wellner, J. A. (1988). Large sample theory of empirical distribution in biased sampling models. *Annals of Statistics*, Vol. 16, 1069-1112.

[14] Girosi, F., and King, G. (2003). Demographic Forecasting. Unpublished book.

[15] Heligman, L. and Pollard, J.H. (1980). The Age Pattern of Mortality. *Journal of the Institute of Actuaries.*, Vol. 107, Part I, 659-671.

[16] Kedem, B. and Fokianos, K. (2002). *Regression Models for Time Series Analysis*, New York: Wiley.

[17] Kedem, B., Gagnon, R.E., and Guo, H. (2005). Time Series Prediction Via Density Ratio Modeling. Unpublished.

[18] Lee, R. (2000). The Lee-Carter method for forecasting mortality, with various extensions and applications. *North American Actuarial Journal*, Vol. 4, 80-93.

[19] Lee, R.D. and Carter, L.R. (1992). Modeling and Forecasting U.S. Mortality. *Journal of American Statistical Association.*, Vol. 87, 659-671.

[20] Lee, R. and Miller, T. (2001). Evaluating the performance of the Lee-Carter approach to modeling and forecasting mortality. *Demography*, Vol 38, 537-549.

[21] Prentice, R. L. and Pyke, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika*, Vol. 66, 403-411.

[22] Qin, J. (1993). Empirical likelihood in biased sample problems. *Annals of Statistics*, Vol. 21, 1182-1196.

[23] Qin, J. (1998). Inferences for case-control and semiparametric two-sample density ratio models. *Biometrika*, Vol. 85 , 619-630

[24] Qin, J. and Lawless, J. F. (1994). Empirical likelihood and general estimating equations. *Annals of Statistics*, Vol. 22, 300-325.

[25] Qin, J. and Zhang, B. (1997). A Goodness of Fit Test for Logistic Regression Models Based on Case-Control Data. *Biometrika*, Vol 84, 609-618.

[26] Van der Vaart, A. W. (1998). *Asymptotic Statistics*, Cambridge University Press.

[27] Van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes with Applications to Statistics*, New York: Springer.

[28] Vardi, Y. (1982). Nonparametric estimation in the presence of length bias. *Annals of Statistics*, Vol. 10, 616-20.

[29] Vardi, Y. (1985). Empirical distribution in selection bias models. *Annals of Statistics*, Vol. 13, 178-203.

[30] Wei, R., Curtin, R.L. and Anderson, R. (2003) U.S. mortality data for building Life Tables and further studies. 2003 Joint Statistical Meeting proceedings: (Biometrics Section) 4458-64.

[31] Wellner, J. A. (1992). Empirical Processes in Action: A Review. *International statistical Review*, Vol 60, 247-269.

[32] Zhang, B. (2000a). M-estimation under a two sample semiparametric model. *Scand. Jour. of Statistics*, 27, 263-280.

[33] Zhang, B. (2000b). Quantile estimation under a two-sample semi-parametric model. *Bernoulli*, 6, 491-511.

[34] Zhang, B. (2000c). A goodness of fit test for multiplicative-intercept risk models based on case-control data. *Statistica Sinica*, 10, 839-865.