

## ABSTRACT

Title of dissertation: Semiparametric Regression and  
Mortality Rate Prediction

Anastasia Voulgaraki, Doctor of Philosophy, 2011

Dissertation directed by: Professor Benjamin Kadem  
Department of Mathematics

This dissertation is divided into two parts. In the first part we consider the general multivariate multiple sample semiparametric density ratio model. In this model one distribution serves as a reference or baseline, and all other distributions are weighted tilts of the reference. The weights are considered known up to a parameter. All the parameters in the model, as well as the reference distribution, are estimated from the combined data from all samples. A kernel-based density estimator can be constructed based on the semiparametric model. In this dissertation we discuss the asymptotic theory and convergence properties for the semiparametric kernel density estimator. The estimator is shown to be not only consistent, but also more efficient than the general kernel density estimator. Several ways for selecting the bandwidth are also discussed. This opens the door to regression analysis with random covariates from a semiparametric perspective where information is combined from multiple multivariate sources. Accordingly, each multivariate distribution and a corresponding conditional expectation (or regression) of interest is then estimated from the combined data from all sources. Graphical and quantitative diagnostic

tools are suggested to assess model validity. The method is applied to real and simulated data. Comparisons are made with multiple regression, generalized additive models (GAM) and nonparametric kernel regression.

In the second part we study mortality rate prediction. The National Center for Health Statistics (NCHS) uses observed mortality data to publish race-gender specific life tables for individual states decennially. At ages over 85 years, the reliability of death rates based on these data is compromised to some extent by age misreporting. The eight-parameter Heligman-Pollard parametric model is then used to smooth the data and obtain estimates/extrapolation of mortality rates for advanced ages. In States with small sub-populations the observed mortality rates are often zero, particularly among young ages. The presence of zero death rates makes the fitting of the Heligman-Pollard model difficult and at times outright impossible. In addition, since death rates are reported on a log scale, zero mortality rates are problematic. To overcome observed zero death rates, appropriate probability models are used. Using these models, observed zero mortality rates are replaced by the corresponding expected values. This enables using logarithmic transformations, and the fitting of the Heligman-Pollard model to produce mortality estimates for ages 0 – 130 years.

Semiparametric Regression  
&  
Mortality Rate Prediction

by

Anastasia Voulgaraki

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  
2011

Advisory Committee:  
Professor Benjamin Kedem, Chair/Advisor  
Professor Doron Levy  
Professor Prakash Narayan  
Professor Paul Smith  
Professor Peter Wolfe

© Copyright by  
Anastasia Voulgaraki  
2011

# Dedication

To Kostas

## Acknowledgments

It is customary the first person to thank is someone's advisor. However, the truth is that this work would not have been realized if not for Dr. Kedem. I am thankful to him for giving me two very unique and challenging problems and thus enabling me to work on different areas of statistics. Dr. Kedem has deep knowledge of statistics, great intuition, a lot of patience and truly cares for his students. It was an honor to work with such a great Statistician. Through his advice and guidance I became not only a better scientist, but a better person as well.

I would also like to thank Dr. Doron Levy, Dr. Prakash Narayan, Dr. Paul Smith and Dr. Peter Wolfe for taking time out of their busy schedules to serve in my committee and provide several remarks and corrections that have greatly improved this work. Furthermore, I am grateful to Dr. Abram Kagan, Dr. Paul Smith and Dr. Eric Slud. I have taken a lot of courses from you and I have learned so much. Thank you for keeping your doors always open for us students and making us feel so welcome.

Many thanks go to my office mates, Katya Sotiris, Carolina Franco and Dr. Ritaja Sur. It has been both a pleasure and an honor to share an office with you for so many years. Thank you for all the times you stopped your work to listen to me complain, and for providing advice and encouragement!

A person that deserves a special mention is Kleoniki Vlachou. She stood by me at a very difficult point in my life and she taught me the meaning of unconditional friendship. Since then she has always been there for me and I know that if I need

anything, I only have to ask. I only hope I will be able to reciprocate when the time comes.

My sister, Dr. Despina Voulgaraki, has specifically requested to be mentioned by name in the acknowledgements! So here it is big sis, the moment has finally come. Where would I be without my big sister's advice to guide me throughout all these years? I learned from your experiences and I can truly say you are one of the few people who have inspired me and influenced me at the same time. Also, I have to mention here my mother, Evangelia and my father Nikolaos. They have always tried to give only the best to their daughters. We would not have aimed this high, if you have not supported us so much. Thank you for being there for us, I only hope we have made you proud.

Lastly, I need to mention my fiancé, Dr. Konstantinos Spiliopoulos. I don't think I would have made it this far, if not for you. But you would never let me quit. You aim for the skies and I fly with you.

# Contents

List of Tables	vii
List of Figures	ix
1 Introduction	1
1.1 Semiparametric density ratio models . . . . .	1
1.2 The problem of zero death rates in U.S. states with small subpopulations	3
1.3 Outline of the thesis . . . . .	5
2 Introduction to semiparametric density ratio models	7
2.1 Introduction . . . . .	7
2.2 Inference . . . . .	11
2.3 Asymptotic theory for $\hat{\theta}$ and $\hat{\mu}$ . . . . .	16
3 Density estimation using the semiparametric density ratio model	18
3.1 Introduction . . . . .	18
3.2 The classical kernel estimator . . . . .	19
3.3 Combined semiparametric density estimators . . . . .	20
3.3.1 Asymptotic results for $\hat{g}_l$ . . . . .	21
3.3.2 Comparison of $\hat{g}_l$ and the traditional $\hat{f}$ . . . . .	37
3.3.3 Bandwidth selection for $\hat{g}_l$ . . . . .	43
4 Semiparametric regression	48
4.1 Introduction . . . . .	48
4.2 Statistical formulation . . . . .	49
4.3 Hypothesis testing . . . . .	53
4.4 Computing $E[y \mathbf{x}]$ using the density ratio model . . . . .	54
4.5 Other ways of computing $E[y \mathbf{x}]$ . . . . .	56
4.5.1 Multiple regression with random covariates . . . . .	56
4.5.2 The Nadaraya-Watson estimator . . . . .	57
4.5.3 Generalized additive models (GAM's) . . . . .	58
4.6 Diagnostic plots and measures of goodness-of-fit . . . . .	59
4.7 Some simulation results . . . . .	62
4.7.1 Comparison of the different measures of goodness-of-fit . . . . .	62
4.7.2 Bandwidth selection . . . . .	65
4.7.3 Comparison with Nadaraya-Watson, GAM's and multiple regression . . . . .	66
5 The Testicular Germ Cell Tumor data set	70
5.1 Introduction . . . . .	70
5.2 Bandwidth selection for the TGCT data set . . . . .	72
5.3 Data analysis . . . . .	74
5.4 Conclusion . . . . .	81



6	Estimation of death rates in U.S. States with small subpopulations	84
6.1	Introduction . . . . .	84
6.2	Models and methods . . . . .	85
6.2.1	Mixed distributions . . . . .	85
6.2.2	Probability models . . . . .	87
6.2.2.1	Mixed lognormal distribution . . . . .	87
6.2.2.2	Two-Part model . . . . .	90
6.2.2.3	Poisson regression . . . . .	92
6.2.2.4	Hurdle model . . . . .	93
6.2.2.5	Zero-inflated model . . . . .	95
6.2.3	The Heligman-Pollard model . . . . .	96
6.3	Data application . . . . .	97
6.3.1	Selection of the models . . . . .	100
6.3.2	Model comparison . . . . .	106
6.3.3	Fitting the Heligman-Pollard model . . . . .	110
6.4	Discussion and Conclusions . . . . .	113
A	Appendix	117
A.1	Computing $\mathbf{W}$ . . . . .	117
A.2	Computing $\mathbf{S}, \mathbf{V}$ . . . . .	123
	Bibliography	125

## List of Tables

4.1	Comparison of goodness of fit measures for case and control. . . . .	64
4.2	Comparison of $R_3^2$ and $R_{0.5,2}^2$ for 100 repetitions of case and control. . . . .	65
4.3	Bandwidth selection using (3.9). Results in bold indicate cases where the integrals did not converge. . . . .	66
4.4	Bandwidth selection using the cross validation method (3.10). . . . .	66
4.5	Bandwidth selection using the cross validation method (3.12) . . . . .	67
4.6	MAE and MSE Comparison of the semiparametric prediction, multiple regression, GAM and Nadaraya-Watson estimators for Simulations 1 and 2. $G_1, G_2$ signify case and control respectively. . . . .	67
5.1	Case-control summary statistics regarding height (cm) and weight (kg), and the correlation between them. . . . .	71
5.2	Bandwidth selection using (3.9). Mathematica failed to calculate the integrals in (3.9) for the 3D TGCT data set. . . . .	73
5.3	Bandwidth selection using the cross validation method (3.10). The method was not used to calculate the bandwidth in the 3D TGCT data set because it is not time efficient. . . . .	73
5.4	Bandwidth selection using the cross validation method (3.12) . . . . .	74
5.5	Some joint probabilities of height and weight in the case and control groups. . . . .	75
5.6	MAE and MSE comparison of the semiparametric prediction and multiple regression for 2D and 3D TGCT data. . . . .	79
5.7	Predicted control values of weight given height and age. . . . .	80
5.8	Predicted case values of weight given height and age. . . . .	80
5.9	Case-control weight and $\hat{E}[\text{weight} \text{height, age}]$ . Empty entries in the table correspond to subjects with the same height and age, but possibly different weights. . . . .	82
5.10	Case-control weight and $\hat{E}[\text{weight} \text{height, age}]$ continued. Empty entries in the table correspond to subjects with the same height and age, but possibly different weights. . . . .	83
6.1	Covariates in the fitted two-part, Poisson, and negative binomial models for the indicated states. . . . .	99
6.2	Covariates in the fitted hurdle and zero-inflated models for the indicated states. . . . .	99
6.3	Estimated expected values of log(death rates) provided by the different models for black females living in Nevada in 2000. . . . .	100
6.4	California RMSE and MAE. Black females, ages 1-84, period 1970-2002. . . . .	105
6.5	Iowa RMSE and MAE. Black females, ages 1-84, period 1970-2002. . . . .	105
6.6	Minnesota RMSE and MAE. Black females, ages 1-84, period 1970-2002. . . . .	106
6.7	Nebraska RMSE and MAE. Black females, ages 1-84, period 1970-2002. . . . .	106

6.8	New Mexico RMSE and MAE. Black females, age 1-84, period 1970-2002. . . . .	107
6.9	Nevada RMSE and MAE. Black females, age 1-84, period 1970-2002.	107
6.10	Oregon RMSE and MAE. Black females, age 1-84, period 1970-2002.	108
6.11	Rhode Island RMSE and MAE. Black females, age 1-84, period 1970-2002. . . . .	108

## List of Figures

4.1	Case-control plots of $\hat{G}_i$ vs. $\tilde{G}_i$ , $i = 1, 2$ , simulations (1) and (2) . . .	63
4.2	Case-control plots of $\hat{G}_i$ vs. $\tilde{G}_i$ , $i = 1, 2$ , simulations (3) and (4) . . .	63
4.3	Comparison of $E[Y X]$ for $G_1$ in simulation 1. . . . .	68
4.4	Comparison of $E[Y X]$ for $G_2$ in simulation 1. . . . .	68
4.5	Comparison of $E[Y X]$ for $G_1$ in simulation 2. . . . .	69
4.6	Comparison of $E[Y X]$ for $G_2$ in simulation 2. . . . .	69
5.1	2D problem: Plots of $\hat{G}_i$ versus $\tilde{G}_i$ , $i = 1, 2$ evaluated at (height,weight) pairs for the case and control groups from the TGCT data. . . . .	75
5.2	Comparison of $E[Y X]$ for $G_1$ in the 2D TGCT data set. . . . .	76
5.3	Comparison of $E[Y X]$ for $G_2$ in the 2D TGCT data set. . . . .	76
5.4	Residual plots for the semiparametric model in the 2D TGCT data set. . . . .	78
5.5	Case-control plots of $\hat{G}_i$ versus $\tilde{G}_i$ , $i = 1, 2$ for the 3D TGCT problem: the $\hat{G}_i, \tilde{G}_i$ are evaluated at selected (age,height,weight) triplets. . . . .	78
5.6	Residual plots for the semiparametric model in the 3D TGCT data set. . . . .	79
6.1	Two-part model: CA, 2000 . . . . .	102
6.2	Mixed lognormal: IA, 2000 . . . . .	102
6.3	Poisson model: MN, 2000 . . . . .	102
6.4	Hurdle model: NE, 2000 . . . . .	102
6.5	Poisson model: NM, 2000 . . . . .	103
6.6	Neg. binomial model: NV, 2000 . . . . .	103
6.7	Zero-inflated model: OR, 2000 . . . . .	103
6.8	Hurdle model: RI, 2000 . . . . .	103
6.9	H-P curve: CA, 2000 . . . . .	110
6.10	H-P curve: OR, 2000 . . . . .	110
6.11	Comparison of H-P curves: NE, 2000 . . . . .	111
6.12	Comparison of H-P curves: NM, 2000 . . . . .	111

## Chapter 1

### Introduction

This dissertation is divided into two parts. The first part evolves around the semiparametric density ratio model. In particular we extend the existing asymptotic results for the semiparametric kernel density estimator to the general multivariate multiple sample case. We propose a new estimator for  $E[y|\mathbf{x}]$  based on the semiparametric model, study its asymptotic properties and propose goodness of fit tests to check the validity of the model. Simulation results and real data applications are also considered. More details for the semiparametric density ratio models are given in Section 1.1. In the second part of the dissertation we are interested in States with small subpopulations where the observed mortality rates are often zero. The zero death rates pose difficulties in the construction of life tables which has resulted in the non publication of some life tables for one fifth of the States. We present a methodology that overcomes these difficulties. An introduction to this problem is given in Section 1.2. Section 1.3 gives an outline of this dissertation.

### 1.1 Semiparametric density ratio models

Assume that data from multiple related sources are available. Examples of such data are case-control data, numerous related time series, weather measurements from different instruments, data from factorial designs and data from many sensors

in a surveillance system. A question of interest is how to combine information from multiple sources in such a way so as to improve distribution inference and hypothesis testing. In this dissertation we focus on a system of distributions, representing multiple data sources, of which one serves as a reference distribution and the rest are distortions or deviations from the reference. We refer to this model as the *semiparametric density ratio model*. See equation (2.1) for a mathematical definition of the model. The model has the advantage it can accommodate both continuous and discrete distributions with minimal assumptions. It is semiparametric because it involves both finite parameters and infinite dimensional parameters. Different forms of model (2.1) have been studied and applied by many authors including Prentice and Pyke [57] (case-control studies), Qin and Zhang [60] (logistic model validation), Qin [61] (case-control studies), Gilbert et al [21] (AIDS vaccine trials), Zhang [80] (goodness of fit), Fokianos et al [18] (analysis of variance), Fokianos [19], Cheng and Chu [11], Qin and Zhang [62] (kernel density estimation), Phue et al [56] (microarrays evaluation), Kedem and Wen [35] (cluster detection), Kedem et al [36] (mortality rate forecasting). The picture which emerges from all this and related work is that, under the density ratio model, by combining all the samples we get both better estimates and more powerful tests. The increase in efficiency has been studied rigorously in Gilbert [22] and Fokianos [19]. We are particularly interested in how fusion of information from multiple sources can be used to create a more efficient kernel density estimator and how we can approach some well known statistical procedures (such as regression and analysis of variance) in a novel way bypassing linearity and the normal assumption.

## 1.2 The problem of zero death rates in U.S. states with small sub-populations

The National Center for Health Statistics (NCHS) publishes sex- and race-specific state decennial life tables for all U.S. states and the District of Columbia (DC) based on the Census of Population and mortality data since 1900. However, in one fifth of the states, life tables are not published due to their small population size [12]. Small populations raise concerns regarding reliability of their mortality rate estimates and fidelity of mortality patterns after smoothing.

The age-specific mortality rate is a key variable in life tables. As a biological feature of human populations, it is expected to have a smooth pattern as a function of age. Hence, based on observed mortality data, age-specific mortality rates are estimated and smoothed in the process of generating life tables. For populations in large states, the estimation/smoothing procedure based on current data is quite reliable because of sufficient data at each age and because the observed mortalities have relatively clear patterns. Usually a non-parametric smoothing procedure is sufficient for providing reliable mortality estimates [6]. However, in small states, observed mortality rates are often interrupted by gaps of zero death observations for some ages. In these cases, non-parametric smoothing is not sufficient in providing smoothed mortality curves. In contrast, by using parametric models the problem can be overcome [13].

Parametric models in mortality estimation have been studied extensively by Hartmann [25], Lambert [41], McCullagh and Nelder [45], Rosenwaike and Hill [66]

and Siller [68]. Almost all reported parametric models use data on a logarithmic scale. This leads to a difficulty with observations of zero deaths in some ages. Zero death rates cannot be ignored. In one pilot study, the Heligman-Pollard model was fitted to data in states with small populations, resulting in overestimated mortalities because the zeros were treated as missing values [76]. A sensible way to overcome the “zeros” problem is to estimate their corresponding (extremely low) death rates before model fitting. This can be done by appropriate probability models which take zero observations into account.

A parametric model is also necessary for extrapolating mortality rates of old age populations for which the reported ages are deemed not accurate [12]. For ages over 85, it has been demonstrated that the reliability of death rate is compromised to some extent by age misreporting. This is increasingly problematic as age increases [12], [17], [58], [66]. For ages between 85 and 100, data from the Medicare program were used to supplement vital statistics and census data in NCHS Life Tables’ estimation [5].

The reliability of the Medicare data deteriorates for ages over 100 [38],[39]. In these cases parametric models can be used to estimate mortality rates in advanced ages. There are several benefits that arise from such models, such as biological interpretation of human mortality, comparison of mortality rates across populations, and continuous interpolation of death rates between ages. In addition, the models assist in the study and forecasting of population mortality trends. For examples of these models see [24], [30], [43], [48], [68] and [71]. A pilot study [75], [76] found that the eight-parameter Heligman-Pollard model is a practical model which,



overall, captures well the age death patterns in US race-gender specific populations. However, in states with small population size, often convergence of the estimation algorithm is not achieved. A major problem in these states is insufficient mortality data; frequently the data are quite variable and do not track the patterns observed in states with large population sizes. Moreover, the occurrence of zero deaths, common for young ages, makes the problem even more complicated since the logarithmic transformation of zero deaths is problematic. The nature of the problem makes it necessary to obtain accurate estimated death numbers or death rates by resorting to sophisticated statistical methods prior to fitting parametric models to the data.

In states with small populations, observed zero deaths result from extremely low mortality rates at some ages and short data collection periods (e.g. 1 – 3 years). But as a biological feature of the human population, age specific death rates are in general continuous and non zero, without interruptions of zero death rates. It is sensible to think that if the data collection time is extended, at least one death will always be observed. Therefore, using data from an extended time period could improve the estimation of death rates by considering methods where time variations are taken into consideration. NCHS has well documented mortality data for over 30 years, which permits an application of the methodology proposed in this dissertation.

### 1.3 Outline of the thesis

This dissertation is organized as follows. Chapters 2 – 5 are devoted to the semiparametric density ratio model. In particular, in Chapter 2 we define the gen-

eral multidimensional semiparametric density ratio model, review the procedure for estimating the parameters of the model, and discuss the asymptotic behavior of the estimators. In Chapter 3 we define the combined (from many samples) semiparametric kernel density estimator and extend the work of Fokianos [19], Cheng and Chu [11], and Qin and Zhang [62] for the general multivariate multiple sample case. Chapter 3 contains the main mathematical results of this dissertation. In Chapter 4 we discuss how the semiparametric model can be used in regression with random covariates and analysis of variance problems. We propose a new way to estimate  $E[y|\mathbf{x}]$ , as well as various measures of goodness of fit to check the validity of the model. The new estimator may be viewed as a semiparametric extension of the Nadaraya-Watson nonparametric estimator. Chapter 4 ends with a simulation study. In Chapter 5 we apply the results presented in Chapters 3 and 4 to Testicular Germ Cell Tumor (TGCT) data. Chapter 6 discusses the problem of zero death rates in States with small subpopulations.

## Chapter 2

### Introduction to semiparametric density ratio models

In this Chapter we review some basic results for the general semiparametric density ratio model. In Section 2.1 we give the formal definition for the model and some examples of distributions that follow this particular model. In Sections 2.2-2.3 we describe the inference procedure and the asymptotic properties of the estimators. For more details on the methodology see for example [18], [19] and [42].

#### 2.1 Introduction

Suppose we have  $m = q + 1$  independent data sets or random samples of  $p$ -dimensional vectors  $\mathbf{x} = \mathbf{x}_{p \times 1} = (x_1, x_2, \dots, x_p)'$ . Let  $g_i(x_1, x_2, \dots, x_p)$  be the probability function corresponding to the  $i$ th sample. Assume that the  $i$ th sample size is  $n_i$  and  $n = \sum_{i=1}^m n_i$  is the total sample size. Thus, for  $i = 1, \dots, q, m$ ,  $j = 1, \dots, n_i$ , we have that

$$\mathbf{x}_{ij} = (x_{ij1}, x_{ij2}, \dots, x_{ijp}) \sim g_i(x_1, \dots, x_p)$$

and

$$\mathbf{x}_{i1}, \mathbf{x}_{i2}, \dots, \mathbf{x}_{in_i} \stackrel{iid}{\sim} g_i$$

where  $\mathbf{x}_{ij}, \mathbf{x}_{ij'}$  are independent for  $j \neq j'$  and  $\mathbf{x}_{ij}, \mathbf{x}_{i'k}$  are independent for  $i \neq i'$  and all  $j$  and  $k$ . We choose  $\mathbf{x}_{mj}$  as the reference sample. Then  $g \equiv g_m(\mathbf{x}) \equiv g_m(x_1, \dots, x_p)$  is called the reference or baseline probability density function (pdf).

We assume that  $g_i(\mathbf{x})$ ,  $i = 1, \dots, q$  satisfy the *density ratio model*:

$$\frac{g_i(\mathbf{x})}{g_m(\mathbf{x})} = w(\mathbf{x}, \boldsymbol{\theta}_i) \quad (2.1)$$

or equivalently

$$g_i(\mathbf{x}) = w(\mathbf{x}, \boldsymbol{\theta}_i)g_m(\mathbf{x}) \quad (2.2)$$

where  $g_i(\mathbf{x})$ ,  $g_m(\mathbf{x})$  are not specified,  $w$  is a known positive continuous function, and  $\boldsymbol{\theta}_i$  is an unknown vector of parameters with finite dimension equal to  $d$ . Rao [63] and Patil and Rao [55] refer to the  $g_i$  as weighted distributions. This construction has the advantage it can accommodate both continuous and discrete distributions, whereas at the same time it does not require normality or even symmetry of continuous distributions. Model (2.1) involves both finite dimensional parameters ( $\boldsymbol{\theta}$ 's) and infinite dimensional parameters in the form of probability density  $g_m$ , and hence a semiparametric approach is appropriate.

**Remark 2.1.** *Model (2.1) is identifiable if and only if for all  $\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}$  with  $\boldsymbol{\theta} \neq \tilde{\boldsymbol{\theta}}$ , there is an  $i \in \{1, \dots, q\}$  such that  $w(\mathbf{x}, \boldsymbol{\theta}_i)$  and  $w(\mathbf{x}, \tilde{\boldsymbol{\theta}}_i)$  are linearly independent as functions of  $\mathbf{x}$  (for more details see [21]). For the remainder of this dissertation we will assume there are no identifiability issues.*

**Example 2.1.** *K-Parameter exponential families.* Consider the general  $k$ -parameter

exponential family

$$g(x, \boldsymbol{\theta}) = d(\boldsymbol{\theta})S(x) \exp \left\{ \sum_{i=1}^k c_i(\boldsymbol{\theta})T_i(x) \right\}$$

where  $\boldsymbol{\theta} = (\theta_1, \dots, \theta_k)$ . For simplicity, consider only two distinct values of  $\boldsymbol{\theta}$ . Then, with  $\alpha = \log[d(\boldsymbol{\theta}_1)/d(\boldsymbol{\theta}_2)]$ ,  $\boldsymbol{\beta} = (c_1(\boldsymbol{\theta}_1) - c_1(\boldsymbol{\theta}_2), \dots, c_k(\boldsymbol{\theta}_1) - c_k(\boldsymbol{\theta}_2))'$ , and  $\mathbf{h}(x) = (T_1(x), \dots, T_k(x))'$ , we obtain the ratio

$$\frac{g_1(x)}{g_2(x)} \equiv \frac{g(x, \boldsymbol{\theta}_1)}{g(x, \boldsymbol{\theta}_2)} = \exp\{\alpha + \boldsymbol{\beta}'\mathbf{h}(x)\} \quad (2.3)$$

or

$$g_1(x) = \exp\{\alpha + \boldsymbol{\beta}'\mathbf{h}(x)\}g_2(x). \quad (2.4)$$

In the normal case with mean  $\mu$  and variance  $\sigma^2$ ,  $\boldsymbol{\theta} = (\mu, \sigma^2)$ , we have

$$g_1(x) = \exp\{\alpha_1 + \beta_1 x + \gamma_1 x^2\}g_2(x)$$

where  $\mathbf{h}(x) = (x, x^2)'$  and

$$\alpha_1 = \ln \left( \frac{\sigma_m}{\sigma_1} \right) + \frac{\mu_m^2}{2\sigma_m^2} - \frac{\mu_1^2}{2\sigma_1^2}, \quad \beta_1 = \frac{\mu_1\sigma_m^2 - \mu_m\sigma_1^2}{\sigma_1^2\sigma_m^2}, \quad \gamma_1 = \frac{\sigma_1^2 - \sigma_m^2}{2\sigma_1^2\sigma_m^2}$$

In the gamma case with shape parameter  $r$  and scale parameter  $\lambda$ ,  $\boldsymbol{\theta} = (r, \lambda)$ ,

$$\alpha = \log \frac{\lambda_1^{r_1}\Gamma(r_2)}{\lambda_2^{r_2}\Gamma(r_1)}, \quad \boldsymbol{\beta} = (\lambda_2 - \lambda_1, r_1 - r_2)', \quad \mathbf{h}(x) = (x, \log x)'$$

As for the Rayleigh distribution with scalar parameter  $\theta$ , (2.4) holds with

$$\alpha = \log \frac{\theta_2^2}{\theta_1^2}, \quad \beta = \frac{1}{2\theta_2^2} - \frac{1}{2\theta_1^2}, \quad h(x) = x^2.$$

**Example 2.2.** *Logistic regression.* Prentice and Pyke [57] studied logistic regression models in case-control studies. Let  $D = i$  denote the development of the  $i$ th disease during a defined accession period,  $i = 1, \dots, q$ , and let  $D = m$  indicate disease-free state at the end of the accession period. The probabilities  $\pi_i = P(D = i)$  satisfy  $\sum_{i=1}^m \pi_i = 1$ . Let  $P(D = i | \mathbf{x})$  denote the conditional probability that an individual with covariate vector  $\mathbf{x}$  has disease  $D = i$ , where  $\mathbf{x} \sim f(\mathbf{x})$ . In a prospective study, if  $\mathbf{x}$  is chosen in advance, we would sample directly from  $P(D = i | \mathbf{x})$ , whereas in case-control studies we sample directly from  $g_i(\mathbf{x}) = P(\mathbf{x} | D = i)$ ,  $i = 1, \dots, q, m$ . In case-control data we often assume that  $P(D | \mathbf{x})$  follows a logistic regression model:

$$P(D = i | \mathbf{x}) = \frac{\exp(\alpha_i^* + \boldsymbol{\beta}'_i \mathbf{x})}{1 + \sum_{i=1}^q \exp(\alpha_i^* + \boldsymbol{\beta}'_i \mathbf{x})}, \quad i = 1, \dots, q, m \quad (2.5)$$

where  $\alpha_m^* = 0$  and  $\boldsymbol{\beta}_m = \mathbf{0}$  for (2.5) to be well defined. Then, from Bayes Theorem,

$$\frac{g_i(\mathbf{x})}{g_m(\mathbf{x})} = \exp(\alpha_i + \boldsymbol{\beta}'_i \mathbf{x}), \quad i = 1, \dots, q \quad (2.6)$$

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

## 2.2 Inference

Let  $G(\mathbf{x}) \equiv G_m(\mathbf{x})$  denote the reference cdf. The problem is to estimate  $g_i$  and  $\boldsymbol{\theta}_i$  from the entire combined data, and not just from the corresponding samples  $\mathbf{x}_{ij}$  and  $\mathbf{x}_{mj}$ . The method of constrained empirical likelihood estimates  $p_{ij} = dG(\mathbf{x}_{ij}) = dG_m(\mathbf{x}_{ij})$  ([19], [37], [52], [59], [60], [72], [73]). The weight functions  $w(\mathbf{x}_{ij}, \boldsymbol{\theta}_i)$  are considered known up to a parameter. The empirical likelihood based on the pooled data  $\mathbf{x}_{ij}$ ,  $i = 1, \dots, m$ ,  $j = 1, \dots, n_i$ , is:

$$\begin{aligned} L(\boldsymbol{\theta}, G_m) &= \left[ \prod_{j=1}^{n_1} p_{1j} w(\mathbf{x}_{1j}, \boldsymbol{\theta}_1) \right] \left[ \prod_{j=1}^{n_2} p_{2j} w(\mathbf{x}_{2j}, \boldsymbol{\theta}_2) \right] \cdots \left[ \prod_{j=1}^{n_m} p_{mj} \right] \\ &= \left[ \prod_{i=1}^m \prod_{j=1}^{n_i} p_{ij} \right] \left[ \prod_{i=1}^m \prod_{j=1}^{n_i} w(\mathbf{x}_{ij}, \boldsymbol{\theta}_i) \right]. \end{aligned} \quad (2.7)$$

Let  $\boldsymbol{\theta} = (\boldsymbol{\theta}'_1, \dots, \boldsymbol{\theta}'_q)'$ , a vector of dimension of  $qd$ . The log-likelihood is given by:

$$l = \log L = \sum_{i=1}^m \sum_{j=1}^{n_i} \log(p_{ij}) + \sum_{i=1}^q \sum_{j=1}^{n_i} \log(w(\mathbf{x}_{ij}, \boldsymbol{\theta}_i)) \quad (2.8)$$

and is subject to the constraints:

$$p_{ij} \geq 0, \quad \sum_{i=1}^m \sum_{j=1}^{n_i} p_{ij} = 1, \quad \sum_{i=1}^m \sum_{j=1}^{n_i} p_{ij} w(\mathbf{x}_{ij}, \boldsymbol{\theta}_k) = 1 \text{ for } k = 1, \dots, q. \quad (2.9)$$

**Remark 2.2.** An equivalent form for the constraint  $\sum_{i=1}^m \sum_{j=1}^{n_i} p_{ij} w(\mathbf{x}_{ij}, \boldsymbol{\theta}_k) = 1$  is

$$\sum_{i=1}^m \sum_{j=1}^{n_i} p_{ij} (w(\mathbf{x}_{ij}, \boldsymbol{\theta}_k) - 1) = 0$$

In order to find estimators for  $p_{ij}$  and  $\boldsymbol{\theta}_i$  we follow a two step procedure: First we express  $p_{ij}$  as a function of  $\boldsymbol{\theta}_i$ , where the  $\boldsymbol{\theta}_i$  are treated as fixed. Then

we substitute the  $p_{ij}$ 's back in the log-likelihood to produce a function of the  $\boldsymbol{\theta}_i$  only. For the first step it suffices to maximize  $\sum_{i=1}^m \sum_{j=1}^{n_i} \log(p_{ij})$  subject to all the constraints. The Lagrangian up to a constant is:

$$\begin{aligned} \Phi(p_{ij}, \boldsymbol{\theta}) = & \sum_{i=1}^m \sum_{j=1}^{n_i} \log(p_{ij}) + \lambda_0 \left( \sum_{i=1}^m \sum_{j=1}^{n_i} p_{ij} - 1 \right) - \lambda_1 \left( \sum_{i=1}^m \sum_{j=1}^{n_i} p_{ij} (w(\mathbf{x}_{ij}, \boldsymbol{\theta}_1) - 1) \right) \\ & - \dots - \lambda_q \left( \sum_{i=1}^m \sum_{j=1}^{n_i} p_{ij} (w(\mathbf{x}_{ij}, \boldsymbol{\theta}_q) - 1) \right). \end{aligned}$$

If we differentiate the Lagrangian with respect to  $p_{ij}$  and set the derivative equal to 0:

$$\begin{aligned} \frac{\partial \Phi(p_{ij}, \boldsymbol{\theta})}{\partial p_{ij}} &= 0 \\ \Rightarrow \frac{1}{p_{ij}} + \lambda_0 - \lambda_1 (w(\mathbf{x}_{ij}, \boldsymbol{\theta}_1) - 1) - \dots - \lambda_q (w(\mathbf{x}_{ij}, \boldsymbol{\theta}_q) - 1) &= 0 \\ \Rightarrow 1 + \lambda_0 p_{ij} - \lambda_1 p_{ij} (w(\mathbf{x}_{ij}, \boldsymbol{\theta}_1) - 1) - \dots - \lambda_q p_{ij} (w(\mathbf{x}_{ij}, \boldsymbol{\theta}_q) - 1) &= 0 \quad (2.10) \\ \Rightarrow \sum_{i=1}^m \sum_{j=1}^{n_i} (1 + \lambda_0 p_{ij} - \lambda_1 p_{ij} (w(\mathbf{x}_{ij}, \boldsymbol{\theta}_1) - 1) - \dots - \lambda_q p_{ij} (w(\mathbf{x}_{ij}, \boldsymbol{\theta}_q) - 1)) &= 0 \\ \Rightarrow n + \lambda_0 = 0 \Rightarrow \lambda_0 = -n \end{aligned}$$

To express  $p_{ij}$  as a function of  $\boldsymbol{\theta}_i$  substitute  $\lambda_0 = -n$  in (2.10):

$$\begin{aligned} 1 - n p_{ij} - \lambda_1 p_{ij} (w(\mathbf{x}_{ij}, \boldsymbol{\theta}_1) - 1) - \dots - \lambda_q p_{ij} (w(\mathbf{x}_{ij}, \boldsymbol{\theta}_q) - 1) &= 0 \\ \Rightarrow p_{ij} = \frac{1}{n + \lambda_1 (w(\mathbf{x}_{ij}, \boldsymbol{\theta}_1) - 1) + \dots + \lambda_q (w(\mathbf{x}_{ij}, \boldsymbol{\theta}_q) - 1)} \end{aligned} \quad (2.11)$$



**Remark 2.3.** By substituting (2.11) back to the constraints we get:

$$\sum_{i=1}^m \sum_{j=1}^{n_i} p_{ij} = 1 \Rightarrow \sum_{i=1}^m \sum_{j=1}^{n_i} \frac{1}{n + \lambda_1 p_{ij}(w(\mathbf{x}_{ij}, \boldsymbol{\theta}_1) - 1) + \dots + \lambda_q p_{ij}(w(\mathbf{x}_{ij}, \boldsymbol{\theta}_q) - 1)} = 1$$

and for  $k = 1, \dots, q$

$$\begin{aligned} & \sum_{i=1}^m \sum_{j=1}^{n_i} p_{ij} w(\mathbf{x}_{ij}, \boldsymbol{\theta}_i) = 1 \\ \Rightarrow & \sum_{i=1}^m \sum_{j=1}^{n_i} \frac{w(\mathbf{x}_{ij}, \boldsymbol{\theta}_i)}{n + \lambda_1 (w(\mathbf{x}_{ij}, \boldsymbol{\theta}_1) - 1) + \dots + \lambda_q (w(\mathbf{x}_{ij}, \boldsymbol{\theta}_q) - 1)} = 1 \\ \Rightarrow & \sum_{i=1}^m \sum_{j=1}^{n_i} \frac{w(\mathbf{x}_{ij}, \boldsymbol{\theta}_i) - 1}{n + \lambda_1 (w(\mathbf{x}_{ij}, \boldsymbol{\theta}_1) - 1) + \dots + \lambda_q (w(\mathbf{x}_{ij}, \boldsymbol{\theta}_q) - 1)} = 0 \end{aligned}$$

For the second step we substitute (2.11) back in the log-likelihood (2.8):

$$\begin{aligned} l(\boldsymbol{\theta}, \lambda_1, \dots, \lambda_q) = & - \sum_{i=1}^m \sum_{j=1}^{n_i} \log(n + \lambda_1 (w(\mathbf{x}_{ij}, \boldsymbol{\theta}_1) - 1) + \dots + \lambda_q (w(\mathbf{x}_{ij}, \boldsymbol{\theta}_q) - 1)) \\ & + \sum_{i=1}^q \sum_{j=1}^{n_i} \log(w(\mathbf{x}_{ij}, \boldsymbol{\theta}_i)) \quad (2.12) \end{aligned}$$

Notice that the log-likelihood (2.12) depends only on the unknown Lagrange multipliers  $\lambda_1, \dots, \lambda_q$  and on  $\boldsymbol{\theta}$ . To express the Lagrange multipliers as a function of  $\boldsymbol{\theta}$

we differentiate (2.12) with respect to  $\lambda_1, \dots, \lambda_q$  and set equal to 0:

$$\begin{aligned} & \frac{\partial l}{\partial \lambda_k} = 0 \\ \Rightarrow & - \sum_{i=1}^m \sum_{j=1}^{n_i} \frac{w(\mathbf{x}_{ij}, \boldsymbol{\theta}_i) - 1}{n + \lambda_1 (w(\mathbf{x}_{ij}, \boldsymbol{\theta}_1) - 1) + \dots + \lambda_q (w(\mathbf{x}_{ij}, \boldsymbol{\theta}_q) - 1)} = 0 \end{aligned}$$

$$\begin{aligned}
&\Rightarrow -\frac{1}{n} \sum_{i=1}^m \sum_{j=1}^{n_i} \frac{w(\mathbf{x}_{ij}, \boldsymbol{\theta}_i) - 1}{1 + \mu_1(w(\mathbf{x}_{ij}, \boldsymbol{\theta}_1) - 1) + \dots + \mu_q(w(\mathbf{x}_{ij}, \boldsymbol{\theta}_q) - 1)} = 0 \\
&\Rightarrow \sum_{i=1}^m \sum_{j=1}^{n_i} \frac{w(\mathbf{x}_{ij}, \boldsymbol{\theta}_i) - 1}{1 + \mu_1(w(\mathbf{x}_{ij}, \boldsymbol{\theta}_1) - 1) + \dots + \mu_q(w(\mathbf{x}_{ij}, \boldsymbol{\theta}_q) - 1)} = 0
\end{aligned}$$

where  $\mu_k \equiv \lambda_k/n$  for  $k = 1, \dots, q$ . Denote by  $\boldsymbol{\mu}$  the vector  $(\mu_1, \dots, \mu_q)'$ . Using  $\mu_k$  instead of  $\lambda_k$ , (2.11) becomes:

$$p_{ij} = \frac{1}{n} \frac{1}{1 + \mu_1(w(\mathbf{x}_{ij}, \boldsymbol{\theta}_1) - 1) + \dots + \mu_q(w(\mathbf{x}_{ij}, \boldsymbol{\theta}_q) - 1)}. \quad (2.13)$$

The log-likelihood can be written in terms of  $\mu_1, \dots, \mu_q$  as follows:

$$\begin{aligned}
l(\boldsymbol{\theta}, \mu_1, \dots, \mu_q) = & - \sum_{i=1}^m \sum_{j=1}^{n_i} \log(1 + \mu_1(w(\mathbf{x}_{ij}, \boldsymbol{\theta}_1) - 1) + \dots + \mu_q(w(\mathbf{x}_{ij}, \boldsymbol{\theta}_q) - 1)) \\
& - n \log(n) + \sum_{i=1}^q \sum_{j=1}^{n_i} \log(w(\mathbf{x}_{ij}, \boldsymbol{\theta}_i)) \quad (2.14)
\end{aligned}$$

The following lemma implies the existence of the maximum empirical likelihood estimators. Let  $h(\mathbf{x}, \boldsymbol{\theta}) = (w(\mathbf{x}, \boldsymbol{\theta}_1) - 1, \dots, w(\mathbf{x}, \boldsymbol{\theta}_q) - 1)'$ . Fokianos ([19]) and Qin and Lawless ([59]) gave conditions guaranteeing that, with probability approaching 1, there is a maximum in a small neighborhood of the true parameter  $\boldsymbol{\theta}_0$ :

**Lemma 2.1.** *Assume that*

(a)  $E_m(h(\mathbf{x}, \boldsymbol{\theta}_0)h'(\mathbf{x}, \boldsymbol{\theta}_0))$  is positive definite,

(b)  $\partial h(\mathbf{x}, \boldsymbol{\theta})/\partial \boldsymbol{\theta}$  is continuous in a neighborhood of the true value  $\boldsymbol{\theta}_0$ ,

(c)  $\|\partial h(\mathbf{x}, \boldsymbol{\theta})/\partial \boldsymbol{\theta}\|$  and  $\|h(\mathbf{x}, \boldsymbol{\theta})\|^3$  are bounded by some integrable function  $H(x)$  with respect to  $G_m(x)$  in this neighborhood,

(d) the rank of  $E(\partial h(\mathbf{x}, \boldsymbol{\theta})/\partial \boldsymbol{\theta})$  is  $qd$ ,

where  $E_m(\cdot)$  and  $\text{Var}_m(\cdot)$  denote expectation and variance with respect to  $G_m$ . Then, as  $n \rightarrow \infty$ , the log-likelihood (2.14) attains its maximum value at some point  $\hat{\boldsymbol{\theta}}$  in the interior of the ball  $\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \leq n^{-1/3}$  and  $\hat{\boldsymbol{\theta}}$  and  $\hat{\mu} = \mu(\hat{\boldsymbol{\theta}})$  can be estimated from the score equations:

$$\begin{aligned} \frac{\partial l(\boldsymbol{\theta}, \mu_1, \dots, \mu_q)}{\partial \boldsymbol{\theta}_l} &= - \sum_{i=1}^m \sum_{j=1}^{n_i} \frac{\mu_l \partial w(\mathbf{x}_{ij}, \boldsymbol{\theta}_l) / \partial \boldsymbol{\theta}_l}{1 + \sum_{k=1}^q \mu_k (w(\mathbf{x}_{ij}, \boldsymbol{\theta}_k) - 1)} \\ &\quad + \sum_{j=1}^{n_l} \frac{1}{w(\mathbf{x}_{lj}, \boldsymbol{\theta}_l)} \frac{\partial w(\mathbf{x}_{lj}, \boldsymbol{\theta}_l)}{\partial \boldsymbol{\theta}_l} \end{aligned} \quad (2.15)$$

$$\frac{\partial l(\boldsymbol{\theta}, \mu_1, \dots, \mu_q)}{\partial \mu_l} = - \sum_{i=1}^m \sum_{j=1}^{n_i} \frac{w(\mathbf{x}_{ij}, \boldsymbol{\theta}_l) - 1}{1 + \sum_{k=1}^q \mu_k (w(\mathbf{x}_{ij}, \boldsymbol{\theta}_k) - 1)} \quad (2.16)$$

for  $l = 1, \dots, q$ .

If we replace  $\mu_k$  and  $\boldsymbol{\theta}_k$  with their estimators from the score equations to (2.13), we can estimate  $p_{ij}$  by:

$$\hat{p}_{ij} = \frac{1}{n} \frac{1}{1 + \sum_{k=1}^q \hat{\mu}_k [w(\mathbf{x}_{ij}, \hat{\boldsymbol{\theta}}_k) - 1]}. \quad (2.17)$$

The maximum likelihood estimator for the reference cdf  $G_m$  is:

$$\hat{G}_m(\mathbf{x}) = \sum_{i=1}^m \sum_{j=1}^{n_i} \hat{p}_{ij} I(\mathbf{x}_{ij} \leq \mathbf{x}) = \frac{1}{n} \sum_{i=1}^m \sum_{j=1}^{n_i} \frac{I(\mathbf{x}_{ij} \leq \mathbf{x})}{1 + \sum_{k=1}^q \hat{\mu}_k [w(\mathbf{x}_{ij}, \hat{\boldsymbol{\theta}}_k) - 1]} \quad (2.18)$$

where  $I(\mathbf{x}_{ij} \leq \mathbf{x})$  is defined componentwise and  $I(B)$  is the indicator of the event  $B$ . More generally, for  $l = 1, \dots, m$  and  $w(\mathbf{x}_{ij}, \hat{\boldsymbol{\theta}}_m) \equiv 1$ :

$$\begin{aligned} \hat{G}_l(\mathbf{x}) &= \sum_{i=1}^m \sum_{j=1}^{n_i} \hat{p}_{ij} w(\mathbf{x}_{ij}, \hat{\boldsymbol{\theta}}_l) I(\mathbf{x}_{ij} \leq \mathbf{x}) \\ &= \frac{1}{n} \sum_{i=1}^m \sum_{j=1}^{n_i} \frac{w(\mathbf{x}_{ij}, \hat{\boldsymbol{\theta}}_l)}{1 + \sum_{k=1}^q \hat{\mu}_k [w(\mathbf{x}_{ij}, \hat{\boldsymbol{\theta}}_k) - 1]} I(\mathbf{x}_{ij} \leq \mathbf{x}) \quad (2.19) \end{aligned}$$

Summarizing, using the method of empirical likelihood, one can obtain score estimating equations for  $\boldsymbol{\theta}$  (2.15) and  $\mu_k$  (2.16) and a semiparametric estimator (2.18) for the cdf  $G_m$ .

### 2.3 Asymptotic theory for $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\mu}}$

In this section, we study the asymptotic properties of  $\hat{\boldsymbol{\theta}}$  and  $\hat{\boldsymbol{\mu}}$ . Let  $\boldsymbol{\theta}_0$  be the true value of  $\boldsymbol{\theta}$  under model (2.1). Define the sample size ratio  $\rho_i = n_i/n_m$  and set  $w(\mathbf{x}, \hat{\boldsymbol{\theta}}_i) = w_i(\mathbf{x})$  for  $i = 1, \dots, m$ . Then  $\rho_m \equiv 1$ ,  $w_m(\mathbf{x}) \equiv 1$ . We assume that  $\rho_i$  is positive and finite and remains fixed as  $n \rightarrow \infty$ . Let  $\boldsymbol{\zeta}$  denote the true value of  $\boldsymbol{\mu}$ . Set  $\boldsymbol{\zeta}_n = (\zeta_{1n}, \dots, \zeta_{qn})$  and  $\zeta_{ln} = n_l/n$  for  $l = 1, \dots, q$ . As  $n \rightarrow \infty$ , assume that  $\zeta_{ln} \rightarrow \zeta_l$ . Then  $\boldsymbol{\zeta}_n \rightarrow \boldsymbol{\zeta}$ .

**Remark 2.4.** Notice that:

$$\begin{aligned} 1 + \sum_{k=1}^q \zeta_{kn} (w(\mathbf{x}_{ij}, \boldsymbol{\theta}_k) - 1) &= 1 + \sum_{k=1}^q \zeta_{kn} w(\mathbf{x}_{ij}, \boldsymbol{\theta}_k) - \sum_{k=1}^q \zeta_{kn} \\ &= 1 + \sum_{k=1}^q \zeta_{kn} w(\mathbf{x}_{ij}, \boldsymbol{\theta}_k) - \frac{n - n_m}{n} = \sum_{k=1}^q \zeta_{kn} w(\mathbf{x}_{ij}, \boldsymbol{\theta}_k) + \frac{n_m}{n} = \sum_{k=1}^m \zeta_{kn} w(\mathbf{x}_{ij}, \boldsymbol{\theta}_k). \end{aligned}$$

Fokianos [19] proved the following theorem:

**Theorem 2.1.** *Assume that the conditions of Lemma (2.1) hold. In addition assume that*

(a)  $\partial^2 h(\mathbf{x}, \boldsymbol{\theta}) / \partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'$  is continuous in a neighborhood of the true parameter,

(b) there is a function  $\dot{H}(\mathbf{x})$  which is integrable with respect to  $G_m$  and which bounds

$$\|\partial^2 h(\mathbf{x}, \boldsymbol{\theta}) / \partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'\|.$$

Then

$$\sqrt{n} \begin{pmatrix} \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 \\ \hat{\boldsymbol{\mu}} - \boldsymbol{\zeta} \end{pmatrix} \xrightarrow{D} N(\mathbf{0}, \mathbf{W}) \quad (2.20)$$

as  $n \rightarrow \infty$ . The matrix  $\mathbf{W}$  can be found in the Appendix.

## Chapter 3

### Density estimation using the semiparametric density ratio model

#### 3.1 Introduction

Fokianos [19], Cheng and Chu [11], and Qin and Zhang [62] constructed a kernel-based density estimator by smoothing the increments of  $\hat{G}_i$ ,  $i = 1, \dots, m$ . In [19], Fokianos studied the statistical properties of the proposed kernel density estimator (mean, variance) and showed that combining data leads to more efficient kernel density estimators when the univariate case of the general model (2.1) was considered. Qin and Zhang [62] considered the univariate version of model (2.1) with  $w(x, \alpha, \beta) = \exp(\alpha + r(x)\beta)$ . For this special case they studied some statistical properties and the convergence in distribution of the estimator. Cheng and Chu [11] studied the same special case as Qin and Zhang [62] but they used a different approach. They also used the new estimate to define a procedure for testing the goodness of fit of the density ratio model. In all three papers, the authors discussed the problem of bandwidth selection and proposed different methods, all for the univariate case.

In this chapter we aim to extend their results for the general multivariate multiple-sample case model (2.1) and to study the corresponding asymptotic theory and convergence properties of the proposed kernel density estimator. The estimator is shown to be not only consistent, but also more efficient than the traditional

kernel density estimator. In addition, several methods for calculating the optimal bandwidth are discussed.

This chapter is organized as follows. In Section 3.2 we review the classical kernel estimator and in Section 3.3 we introduce the semiparametric kernel density estimator  $\hat{g}_l$ . Specifically, in Section 3.3.1 we examine the asymptotic behavior of  $\hat{g}_l$  and in Section 3.3.2 we compare it with the classical kernel estimator. In Section 3.3.3 we discuss the advantages and disadvantages of several methods for selecting the bandwidth for  $\hat{g}_l$ .

## 3.2 The classical kernel estimator

The traditional kernel density estimator is a convolution of the jumps in the empirical distribution function obtained from a single sample of size  $n$  and a kernel function taken as a symmetric probability density function parameterized by a bandwidth parameter ([54]). Specifically, the kernel density estimator of a probability density  $f(\mathbf{x})$  is given by

$$\hat{f}(\mathbf{x}) = \frac{1}{nh_n^p} \sum_{i=1}^n K\left(\frac{\mathbf{x} - \mathbf{x}_i}{h_n}\right) \quad (3.1)$$

where  $h_n$  is a sequence of bandwidths such that  $h_n \rightarrow 0$  and  $nh_n^p \rightarrow \infty$  as  $n \rightarrow \infty$ . The kernel function  $K(\mathbf{x})$  is defined for  $p$ -dimensional  $\mathbf{x}$ . It is nonnegative, symmetric around  $\mathbf{0}$  and satisfies  $\int_{\mathbf{R}^p} K(\mathbf{x})d\mathbf{x} = 1$ . The standard multivariate normal density is a convenient choice for the kernel  $K(\mathbf{x})$ . Under certain conditions,  $\hat{f}(\mathbf{x})$  is a consistent estimator of  $f(\mathbf{x})$  and is asymptotically normal ([54], [67]). As such,

the traditional kernel density estimator is a “single sample” estimator. Improved estimators can be obtained when multiple data sources are available.

### 3.3 Combined semiparametric density estimators

Using a similar idea to (3.1), we may use the expressions for the probabilities in (2.17) as the basis to form kernel estimates for the probability distributions  $g_l(\mathbf{x})$ :

$$\hat{g}_l(\mathbf{x}) = \frac{1}{h_n^p} \sum_{i=1}^m \sum_{j=1}^{n_i} \hat{p}_{ij} \hat{w}_l(\mathbf{x}_{ij}) K\left(\frac{\mathbf{x} - \mathbf{x}_{ij}}{h_n}\right) \quad (3.2)$$

where  $h_n$  is a sequence of bandwidths such that  $h_n \rightarrow 0$  and  $nh_n^p \rightarrow \infty$  as  $n \rightarrow \infty$ ,  $w_l(\mathbf{x}) \equiv w(\mathbf{x}, \boldsymbol{\theta}_l)$ ,  $\hat{w}_l(\mathbf{x}) \equiv w(\mathbf{x}, \hat{\boldsymbol{\theta}}_l)$ , and  $K$  is a nonnegative kernel function that satisfies the following requirements:

1.  $\int K(\mathbf{x})d\mathbf{x} = 1$  and  $\int |K(\mathbf{x})|d\mathbf{x} < \infty$ ;
2.  $\int \mathbf{x}K(\mathbf{x})d\mathbf{x} = \mathbf{0}$  and  $\int |\mathbf{x}K(\mathbf{x})|d\mathbf{x} < \infty$ ;
3.  $\int \mathbf{x}'\mathbf{x}K(\mathbf{x})d\mathbf{x} = k_2$  and  $\int |\mathbf{x}'\mathbf{x}K(\mathbf{x})|d\mathbf{x} < \infty$ .

Notice that  $\hat{g}_l$  depends on both the unknown reference distribution function and the parameter  $\hat{\boldsymbol{\theta}}_l$  of the model, and is therefore a semiparametric density estimator. Moreover, it is easy to verify that  $\hat{g}_l$  is a proper probability function. Indeed from (2.9):



$$\begin{aligned}
\int \hat{g}_l(\mathbf{x}) d\mathbf{x} &= \int \frac{1}{h_n^p} \sum_{i=1}^m \sum_{j=1}^{n_i} \hat{p}_{ij} \hat{w}_l(\mathbf{x}_{ij}) K\left(\frac{\mathbf{x} - \mathbf{x}_{ij}}{h_n}\right) d\mathbf{x} \\
&\stackrel{\mathbf{u} = \frac{\mathbf{x} - \mathbf{x}_{ij}}{h_n}}{=} \sum_{i=1}^m \sum_{j=1}^{n_i} \hat{p}_{ij} \hat{w}_l(\mathbf{x}_{ij}) \int K(\mathbf{u}) d\mathbf{u} = \sum_{i=1}^m \sum_{j=1}^{n_i} \hat{p}_{ij} \hat{w}_l(\mathbf{x}_{ij}) = 1
\end{aligned}$$

Therefore (3.2) defines a proper probability density function.

### 3.3.1 Asymptotic results for $\hat{g}_l$

In this section we will prove some asymptotic results for  $\hat{g}_l$ . To facilitate the study of  $\hat{g}_l$ , it is convenient to define first  $\tilde{g}_l(\mathbf{x})$ :

$$\tilde{g}_l(\mathbf{x}) = \frac{1}{h_n^p} \sum_{i=1}^m \sum_{j=1}^{n_i} p_{ij} w_l(\mathbf{x}_{ij}) K\left(\frac{\mathbf{x} - \mathbf{x}_{ij}}{h_n}\right). \quad (3.3)$$

**Lemma 3.1.** *Assume  $\boldsymbol{\theta}$  and  $p_{ij}$  are known for  $i = 1, \dots, q, m, j = 1, \dots, n_i$ . Then*

$$\tilde{g}_l(\mathbf{x}) = \frac{1}{nh_n^p} \sum_{i=1}^m \sum_{j=1}^{n_i} \frac{w_l(\mathbf{x}_{ij})}{\sum_{k=1}^m \zeta_k w_k(\mathbf{x}_{ij})} K\left(\frac{\mathbf{x} - \mathbf{x}_{ij}}{h_n}\right) \quad (3.4)$$

where  $\zeta_k = n_k/n$ .

*Proof.*

$$\begin{aligned}
\tilde{g}_l(\mathbf{x}) &= \frac{1}{h_n^p} \sum_{i=1}^m \sum_{j=1}^{n_i} p_{ij} w_l(\mathbf{x}_{ij}) K\left(\frac{\mathbf{x} - \mathbf{x}_{ij}}{h_n}\right) \\
&= \frac{1}{h_n^p} \sum_{i=1}^m \sum_{j=1}^{n_i} \frac{1}{n} \frac{w_l(\mathbf{x}_{ij})}{1 + \sum_{k=1}^q \zeta_k [w_k(\mathbf{x}_{ij}) - 1]} K\left(\frac{\mathbf{x} - \mathbf{x}_{ij}}{h_n}\right)
\end{aligned}$$

$$\begin{aligned}
&= \frac{1}{nh_n^p} \sum_{i=1}^m \sum_{j=1}^{n_i} \frac{w_l(\mathbf{x}_{ij})}{1 + \sum_{k=1}^q n_k/n [w_k(\mathbf{x}_{ij}) - 1]} K\left(\frac{\mathbf{x} - \mathbf{x}_{ij}}{h_n}\right) \\
&= \frac{1}{nh_n^p} \sum_{i=1}^m \sum_{j=1}^{n_i} \frac{w_l(\mathbf{x}_{ij})}{1 + \sum_{k=1}^q (n_k/n)w_k(\mathbf{x}_{ij}) - 1 + n_m/n} K\left(\frac{\mathbf{x} - \mathbf{x}_{ij}}{h_n}\right) \\
&= \frac{1}{nh_n^p} \sum_{i=1}^m \sum_{j=1}^{n_i} \frac{w_l(\mathbf{x}_{ij})}{\sum_{k=1}^m \zeta_k w_k(\mathbf{x}_{ij})} K\left(\frac{\mathbf{x} - \mathbf{x}_{ij}}{h_n}\right).
\end{aligned}$$

□

**Lemma 3.2.** *Assume that  $K(\cdot)$  is a nonnegative bounded symmetric function with  $\int K(\mathbf{x})d\mathbf{x} = 1$ ,  $\int \mathbf{x}K(\mathbf{x})d\mathbf{x} = 0$ ,  $\int \mathbf{x}'\mathbf{x}K(\mathbf{x})d\mathbf{x} = k_2 > 0$ ,  $\int K^2(\mathbf{x})d\mathbf{x} < \infty$ . Assume that  $g_l$  is continuous and bounded at  $\mathbf{x}$ . Then*

(a) *As  $n \rightarrow \infty$  and  $h_n \rightarrow 0$ ,*

$$E\tilde{g}_l(\mathbf{x}) = \frac{1}{h_n^p} \int K\left(\frac{\mathbf{x} - \mathbf{y}}{h_n}\right) g_l(\mathbf{y})d\mathbf{y} = g_l(\mathbf{x}) + o(1).$$

(b) *If  $g_l$  is twice continuously differentiable in a neighborhood of  $\mathbf{x}$ , then as  $n \rightarrow \infty$  and  $h_n \rightarrow 0$ ,*

$$E\tilde{g}_l(\mathbf{x}) = g_l(\mathbf{x}) + \frac{1}{2}h_n^2 \int \mathbf{u}' \frac{\partial^2}{\partial \mathbf{x} \partial \mathbf{x}'} g_l(\mathbf{x}) \mathbf{u} K(\mathbf{u}) d\mathbf{u} + o(h_n^2).$$

(c) *As  $n \rightarrow \infty$ ,  $h_n \rightarrow 0$  and  $nh_n^p \rightarrow \infty$ ,*

$$\begin{aligned}
\text{Var}(\tilde{g}_l(\mathbf{x})) &= \frac{1}{n(h_n)^{2p}} \int \frac{w_l^2(\mathbf{y})}{\sum_{k=1}^m \zeta_k w_k(\mathbf{y})} K^2\left(\frac{\mathbf{x} - \mathbf{y}}{h_n}\right) g(\mathbf{y})d\mathbf{y} \\
&\quad - \frac{1}{n} \sum_{i=1}^m \zeta_i \left[ \int \frac{1}{h_n^p} \frac{w_l(\mathbf{y})}{\sum_{k=1}^m \zeta_k w_k(\mathbf{y})} K\left(\frac{\mathbf{x} - \mathbf{y}}{h_n}\right) w_i(\mathbf{y})g(\mathbf{y})d\mathbf{y} \right]^2
\end{aligned}$$

$$\begin{aligned}
&= \frac{1}{nh_n^p} \frac{w_l(\mathbf{x})g_l(\mathbf{x})}{\sum_{k=1}^m \zeta_k w_k(\mathbf{x})} \int K^2(\mathbf{u})d\mathbf{u} + o\left(\frac{1}{nh_n^p}\right) \\
&= \frac{1}{nh_n^p} \sigma^2(\mathbf{x}) + o\left(\frac{1}{nh_n^p}\right)
\end{aligned}$$

with

$$\sigma^2(\mathbf{x}) = \frac{w_l(\mathbf{x})g_l(\mathbf{x})}{\sum_{k=1}^m \zeta_k w_k(\mathbf{x})} \int K^2(\mathbf{u})d\mathbf{u}.$$

*Proof.* (a)

$$\begin{aligned}
E\tilde{g}_l(\mathbf{x}) &= E\left[\frac{1}{nh_n^p} \sum_{i=1}^m \sum_{j=1}^{n_i} \frac{w_l(\mathbf{x}_{ij})}{\sum_{k=1}^m \zeta_k w_k(\mathbf{x}_{ij})} K\left(\frac{\mathbf{x} - \mathbf{x}_{ij}}{h_n}\right)\right] \\
&= \frac{1}{nh_n^p} \sum_{i=1}^m n_i E_i \left[ \frac{w_l(\mathbf{y})}{\sum_{k=1}^m \zeta_k w_k(\mathbf{y})} K\left(\frac{\mathbf{x} - \mathbf{y}}{h_n}\right) \right] \\
&= \frac{1}{h_n^p} \sum_{i=1}^m \zeta_i \int \frac{w_l(\mathbf{y})}{\sum_{k=1}^m \zeta_k w_k(\mathbf{y})} K\left(\frac{\mathbf{x} - \mathbf{y}}{h_n}\right) w_i(\mathbf{y}) dG(\mathbf{y}) \\
&= \frac{1}{h_n^p} \int \frac{\sum_{i=1}^m \zeta_i w_i(\mathbf{y})}{\sum_{k=1}^m \zeta_k w_k(\mathbf{y})} w_l(\mathbf{y}) K\left(\frac{\mathbf{x} - \mathbf{y}}{h_n}\right) dG(\mathbf{y}) \\
&= \frac{1}{h_n^p} \int K\left(\frac{\mathbf{x} - \mathbf{y}}{h_n}\right) g_l(\mathbf{y}) d\mathbf{y} \stackrel{\mathbf{u}=\frac{\mathbf{x}-\mathbf{y}}{h_n}}{=} \int K(\mathbf{u})g_l(\mathbf{x} - h_n\mathbf{u})d\mathbf{u}
\end{aligned}$$

where  $E_i(\mathbf{x})$  is the expected value of  $\mathbf{x}$  with respect to sample  $i$ . Next, fix  $\varepsilon$ ,  $\varepsilon_1$ ,  $\varepsilon_2$ . By continuity, for  $\varepsilon > 0$ , there exists  $\delta > 0$  such that if

$$|(\mathbf{x} - h_n\mathbf{u}) - \mathbf{x}| < \delta \Leftrightarrow |h_n\mathbf{u}| < \delta \Leftrightarrow |\mathbf{u}| < \delta/h_n$$

then  $|g_l(\mathbf{x} - h_n \mathbf{u}) - g_l(\mathbf{x})| < \varepsilon$ .

For  $\varepsilon_1 > 0$  and for  $\delta > 0$ , there exists  $h_1 > 0$  such that for every  $h \leq h_1$ :

$$\int_{-\infty}^{-\delta/h} K(\mathbf{u}) \cdot |g_l(\mathbf{x} - h_n \mathbf{u}) - g_l(\mathbf{x})| d\mathbf{u} < \varepsilon_1$$

because  $g_l$  is bounded and  $K(\mathbf{x})$  is integrable. Similarly for  $\varepsilon_2 > 0$  and for  $\delta > 0$ ,

there exists  $h_2 > 0$  such that for every  $h \leq h_2$ :

$$\int_{\delta/h}^{\infty} K(\mathbf{u}) \cdot |g_l(\mathbf{x} - h_n \mathbf{u}) - g_l(\mathbf{x})| d\mathbf{u} < \varepsilon_2.$$

Set  $h_n \leq \min(h_1, h_2)$  and notice that:

$$\begin{aligned} |E\tilde{g}_l(\mathbf{x}) - g_l(\mathbf{x})| &= \left| E\tilde{g}_l(\mathbf{x}) - g_l(\mathbf{x}) \int K(\mathbf{u}) d\mathbf{u} \right| \\ &= \left| \int K(\mathbf{u}) g_l(\mathbf{x} - h_n \mathbf{u}) d\mathbf{u} - \int g_l(\mathbf{x}) K(\mathbf{u}) d\mathbf{u} \right| \\ &\leq \int K(\mathbf{u}) \cdot |g_l(\mathbf{x} - h_n \mathbf{u}) - g_l(\mathbf{x})| d\mathbf{u} \\ &= \int_{-\infty}^{-\delta/h_n} K(\mathbf{u}) \cdot |g_l(\mathbf{x} - h_n \mathbf{u}) - g_l(\mathbf{x})| d\mathbf{u} \\ &\quad + \int_{-\delta/h_n}^{\delta/h_n} K(\mathbf{u}) \cdot |g_l(\mathbf{x} - h_n \mathbf{u}) - g_l(\mathbf{x})| d\mathbf{u} \\ &\quad + \int_{\delta/h_n}^{\infty} K(\mathbf{u}) \cdot |g_l(\mathbf{x} - h_n \mathbf{u}) - g_l(\mathbf{x})| d\mathbf{u} \rightarrow 0 \quad (3.5) \end{aligned}$$

pointwise by the discussion above. In conclusion, as  $h_n \rightarrow 0$ ,  $n \rightarrow \infty$ ,

$$E\tilde{g}_l(\mathbf{x}) = g_l(\mathbf{x}) \int K(\mathbf{u}) d\mathbf{u} + o(1) = g_l(\mathbf{x}) + o(1),$$

where  $o(1) \rightarrow 0$  as  $h_n \rightarrow 0$ ,  $n \rightarrow \infty$ .

(b)

$$\begin{aligned}
E\tilde{g}_l(\mathbf{x}) &\stackrel{\text{part (a)}}{=} \int K(\mathbf{u})g_l(\mathbf{x} - h_n\mathbf{u})d\mathbf{u} \\
&\stackrel{\text{2nd Taylor exp.}}{=} \int K(\mathbf{u}) \left[ g_l(\mathbf{x}) - h_n\mathbf{u}'\frac{\partial g_l(\mathbf{x})}{\partial \mathbf{x}} + \frac{h_n^2}{2}\mathbf{u}'\frac{\partial^2}{\partial \mathbf{x}\partial \mathbf{x}'}g_l(\mathbf{x})\mathbf{u} + o(h_n^2) \right] d\mathbf{u} \\
&= g_l(\mathbf{x}) + \frac{h_n^2}{2} \int \mathbf{u}'\frac{\partial^2}{\partial \mathbf{x}\partial \mathbf{x}'}g_l(\mathbf{x})\mathbf{u}K(\mathbf{u})d\mathbf{u} + o(h_n^2).
\end{aligned}$$

(c)

$$\begin{aligned}
\text{Var}(\tilde{g}_l(\mathbf{x})) &= \text{Var} \left[ \frac{1}{nh_n^p} \sum_{i=1}^m \sum_{j=1}^{n_i} \frac{w_l(\mathbf{x}_{ij})}{\sum_{k=1}^m \zeta_k w_k(\mathbf{x}_{ij})} K \left( \frac{\mathbf{x} - \mathbf{x}_{ij}}{h_n} \right) \right] \\
&= \frac{1}{n^2(h_n)^{2p}} \text{Var} \left[ \sum_{i=1}^m \sum_{j=1}^{n_i} \frac{w_l(\mathbf{x}_{ij})}{\sum_{k=1}^m \zeta_k w_k(\mathbf{x}_{ij})} K \left( \frac{\mathbf{x} - \mathbf{x}_{ij}}{h_n} \right) \right] \\
&= \frac{1}{n^2(h_n)^{2p}} \sum_{i=1}^m n_i \text{Var} \left[ \frac{w_l(\mathbf{x}_{i1})}{\sum_{k=1}^m \zeta_k w_k(\mathbf{x}_{i1})} K \left( \frac{\mathbf{x} - \mathbf{x}_{i1}}{h_n} \right) \right] \\
&= \frac{1}{n^2(h_n)^{2p}} \sum_{i=1}^m n_i \left\{ E_i \left[ \frac{w_l^2(\mathbf{x}_{i1})}{(\sum_{k=1}^m \zeta_k w_k(\mathbf{x}_{i1}))^2} K^2 \left( \frac{\mathbf{x} - \mathbf{x}_{i1}}{h_n} \right) \right] \right. \\
&\quad \left. - E_i^2 \left[ \frac{w_l(\mathbf{x}_{i1})}{\sum_{k=1}^m \zeta_k w_k(\mathbf{x}_{i1})} K \left( \frac{\mathbf{x} - \mathbf{x}_{i1}}{h_n} \right) \right] \right\} \\
&= \frac{1}{n(h_n)^{2p}} \sum_{i=1}^m \zeta_i \left[ \int \frac{w_l^2(\mathbf{y})}{(\sum_{k=1}^m \zeta_k w_k(\mathbf{y}))^2} K^2 \left( \frac{\mathbf{x} - \mathbf{y}}{h_n} \right) w_i(\mathbf{y})g(\mathbf{y})d\mathbf{y} \right. \\
&\quad \left. - \frac{1}{n(h_n)^{2p}} \sum_{i=1}^m \zeta_i \left[ \int \frac{w_l(\mathbf{y})}{\sum_{k=1}^m \zeta_k w_k(\mathbf{y})} K \left( \frac{\mathbf{x} - \mathbf{y}}{h_n} \right) w_i(\mathbf{y})g(\mathbf{y})d\mathbf{y} \right]^2 \right] \\
&= \frac{1}{nh_n^p} \left\{ \frac{1}{h_n^p} \int \frac{w_l^2(\mathbf{y})}{\sum_{k=1}^m \zeta_k w_k(\mathbf{y})} K^2 \left( \frac{\mathbf{x} - \mathbf{y}}{h_n} \right) g(\mathbf{y})d\mathbf{y} \right. \\
&\quad \left. - \frac{1}{h_n^p} \sum_{i=1}^m \zeta_i \left[ \int \frac{w_l(\mathbf{y})}{\sum_{k=1}^m \zeta_k w_k(\mathbf{y})} K \left( \frac{\mathbf{x} - \mathbf{y}}{h_n} \right) w_i(\mathbf{y})g(\mathbf{y})d\mathbf{y} \right]^2 \right\} (3.6)
\end{aligned}$$

We will examine each term in (3.6) separately. Notice the following:

- The term

$$\frac{w_l^2(\mathbf{x} - h_n \mathbf{u})}{\sum_{k=1}^m \zeta_k w_k(\mathbf{x} - h_n \mathbf{u})} g(\mathbf{x} - h_n \mathbf{u})$$

is bounded. Indeed:

$$\begin{aligned} \frac{w_l^2(\mathbf{x} - h_n \mathbf{u})}{\sum_{k=1}^m \zeta_k w_k(\mathbf{x} - h_n \mathbf{u})} g(\mathbf{x} - h_n \mathbf{u}) &= \frac{1}{\zeta_l} \frac{\zeta_l w_l(\mathbf{x} - h_n \mathbf{u})}{\sum_{k=1}^m \zeta_k w_k(\mathbf{x} - h_n \mathbf{u})} g_l(\mathbf{x} - h_n \mathbf{u}) \\ &\leq \frac{1}{\zeta_l} \sup_{\mathbf{x} \in \mathbb{R}^p} |g_l(\mathbf{x})|. \end{aligned}$$

Then

$$\begin{aligned} \frac{1}{h_n^p} \int \frac{w_l^2(\mathbf{y})}{\sum_{k=1}^m \zeta_k w_k(\mathbf{y})} K^2\left(\frac{\mathbf{x} - \mathbf{y}}{h_n}\right) g(\mathbf{y}) d\mathbf{y} \\ \stackrel{\mathbf{u} = \frac{\mathbf{x} - \mathbf{y}}{h_n}}{=} \int \frac{w_l^2(\mathbf{x} - h_n \mathbf{u})}{\sum_{k=1}^m \zeta_k w_k(\mathbf{x} - h_n \mathbf{u})} K^2(\mathbf{u}) g(\mathbf{x} - h_n \mathbf{u}) d\mathbf{u} \\ \stackrel{\text{By Dom. Conv. Thm}}{\underset{\text{as } h_n \rightarrow 0}{\underset{=}{}}} \frac{w_l^2(\mathbf{x}) g(\mathbf{x})}{\sum_{k=1}^m \zeta_k w_k(\mathbf{x})} \int K^2(\mathbf{u}) d\mathbf{u} + o(1). \end{aligned}$$

- The terms

$$\frac{\zeta_i w_i(\mathbf{x} - h_n \mathbf{u})}{\sum_{k=1}^m \zeta_k w_k(\mathbf{x} - h_n \mathbf{u})}$$

and

$$w_l(\mathbf{x} - h_n \mathbf{u}) g(\mathbf{x} - h_n \mathbf{u}) = g_l(\mathbf{x} - h_n \mathbf{u})$$

are bounded. Then

$$\begin{aligned}
& \frac{1}{h_n^p} \zeta_i \left[ \int \frac{w_l(\mathbf{y})}{\sum_{k=1}^m \zeta_k w_k(\mathbf{y})} K\left(\frac{\mathbf{x}-\mathbf{y}}{h_n}\right) w_i(\mathbf{y}) g(\mathbf{y}) d\mathbf{y} \right]^2 \\
& \stackrel{\mathbf{u}=\frac{\mathbf{x}-\mathbf{y}}{h_n}}{=} h_n^p \zeta_i \left[ \int \frac{w_l(\mathbf{x}-h_n \mathbf{u})}{\sum_{k=1}^m \zeta_k w_k(\mathbf{x}-h_n \mathbf{u})} K(\mathbf{u}) w_i(\mathbf{x}-h_n \mathbf{u}) g(\mathbf{x}-h_n \mathbf{u}) d\mathbf{u} \right]^2 \\
& \stackrel{\text{By Dom. Conv. Thm}}{\underset{\text{as } h_n \rightarrow 0}}{=} h_n^p \zeta_i \left[ \left( \frac{w_l(\mathbf{x}) g_i(\mathbf{x})}{\sum_{k=1}^m \zeta_k w_k(\mathbf{x})} \right)^2 \left( \int K(\mathbf{u}) d\mathbf{u} \right)^2 + o(1) \right] \\
& = h_n^p \zeta_i \left( \frac{w_l(\mathbf{x}) g_i(\mathbf{x})}{\sum_{k=1}^m \zeta_k w_k(\mathbf{x})} \right)^2 + o(h_n^p).
\end{aligned}$$

Therefore

$$\begin{aligned}
\text{Var}(\tilde{g}_l(\mathbf{x})) &= \frac{1}{nh_n^p} \left[ \frac{w_l^2(\mathbf{x}) g(\mathbf{x})}{\sum_{k=1}^m \zeta_k w_k(\mathbf{x})} \int K^2(\mathbf{u}) d\mathbf{u} + o(1) \right. \\
&\quad \left. - \sum_{i=1}^m h_n^p \zeta_i \left( \frac{w_l(\mathbf{x}) g_i(\mathbf{x})}{\sum_{k=1}^m \zeta_k w_k(\mathbf{x})} \right)^2 + o(h_n^p) \right] \\
&= \frac{1}{nh_n^p} \frac{w_l^2(\mathbf{x}) g(\mathbf{x})}{\sum_{k=1}^m \zeta_k w_k(\mathbf{x})} \int K^2(\mathbf{u}) d\mathbf{u} + o\left(\frac{1}{nh_n^p}\right) + o(n^{-1}) \\
&= \frac{1}{nh_n^p} \frac{w_l(\mathbf{x}) g_l(\mathbf{x})}{\sum_{k=1}^m \zeta_k w_k(\mathbf{x})} \int K^2(\mathbf{u}) d\mathbf{u} + o\left(\frac{1}{nh_n^p}\right)
\end{aligned}$$

since the term  $\sum_{i=1}^m h_n^p \zeta_i (w_l(\mathbf{x}) g_i(\mathbf{x}) / \sum_{k=1}^m \zeta_k w_k(\mathbf{x}))^2$  is finite.

□

**Lemma 3.3.** *Assume that  $K(\cdot)$  is a nonnegative bounded symmetric function with  $\int K(\mathbf{x}) d\mathbf{x} = 1$ ,  $\int \mathbf{x}' \mathbf{x} K(\mathbf{x}) d\mathbf{x} = k_2 > 0$ . Assume that  $g_l$  is continuous at  $\mathbf{x}$  and bounded, and that the conditions of Lemma 2.1 hold. If the quantity*

$$\mathbf{I}_r(\boldsymbol{\theta}, \mu) = \frac{\partial}{\partial \boldsymbol{\theta}_r} \frac{w(\mathbf{x}_{ij}, \boldsymbol{\theta}_l)}{1 + \sum_{k=1}^q \mu_k [w(\mathbf{x}_{ij}, \boldsymbol{\theta}_k) - 1]}$$

is bounded, then  $\hat{g}_l(\mathbf{x}) = \tilde{g}_l(\mathbf{x}) + O_p(n^{-1/2})$  as  $n \rightarrow \infty$  and  $h_n \rightarrow 0$ .

*Proof.*

$$\begin{aligned}
\hat{g}_l(\mathbf{x}) - \tilde{g}_l(\mathbf{x}) &= \frac{1}{h_n^p} \sum_{i=1}^m \sum_{j=1}^{n_i} \hat{p}_{ij} \hat{w}_l(\mathbf{x}_{ij}) K\left(\frac{\mathbf{x} - \mathbf{x}_{ij}}{h_n}\right) \\
&\quad - \frac{1}{nh_n^p} \sum_{i=1}^m \sum_{j=1}^{n_i} \frac{w_l(\mathbf{x}_{ij})}{\sum_{k=1}^m \zeta_k w_k(\mathbf{x}_{ij})} K\left(\frac{\mathbf{x} - \mathbf{x}_{ij}}{h_n}\right) \\
&= \frac{1}{nh_n^p} \sum_{i=1}^m \sum_{j=1}^{n_i} \left[ \frac{\hat{w}_l(\mathbf{x}_{ij})}{1 + \sum_{k=1}^q \hat{\mu}_k (\hat{w}_k(\mathbf{x}_{ij}) - 1)} - \frac{w_l(\mathbf{x}_{ij})}{\sum_{k=1}^m \zeta_k w_k(\mathbf{x}_{ij})} \right] \\
&\quad \times K\left(\frac{\mathbf{x} - \mathbf{x}_{ij}}{h_n}\right) \\
&= R_n(\mathbf{x}_{ij}) \\
&\stackrel{\text{1st Order Taylor}}{=} \sum_{r=1}^q R_{1nr} (\hat{\mu}_r - \zeta_r) + \sum_{r=1}^q \mathbf{R}_{2nr} (\hat{\boldsymbol{\theta}}_r - \boldsymbol{\theta}_r)
\end{aligned}$$

where

$$\begin{aligned}
\mathbf{I}_r(\boldsymbol{\theta}^*, \mu^*) &= \frac{\partial}{\partial \boldsymbol{\theta}_r} \frac{w(\mathbf{x}_{ij}, \boldsymbol{\theta}_l)}{1 + \sum_{k=1}^q \mu_k [w(\mathbf{x}_{ij}, \boldsymbol{\theta}_k) - 1]} \Bigg|_{\substack{\mu_k = \mu_k^* \\ \boldsymbol{\theta}_k = \boldsymbol{\theta}_k^*}} \\
&= \frac{\frac{\partial w(\mathbf{x}_{ij}, \boldsymbol{\theta}_r^*)}{\partial \boldsymbol{\theta}_r} [1 + \sum_{k=1}^q \mu_k^* (w(\mathbf{x}_{ij}, \boldsymbol{\theta}_k^*) - 1)] I(r=l) - w(\mathbf{x}_{ij}, \boldsymbol{\theta}_l^*) \mu_r^* \frac{\partial w(\mathbf{x}_{ij}, \boldsymbol{\theta}_r^*)}{\partial \boldsymbol{\theta}_r}}{(1 + \sum_{k=1}^q \mu_k^* [w(\mathbf{x}_{ij}, \boldsymbol{\theta}_k^*) - 1])^2}
\end{aligned}$$

and

$$\begin{aligned}
R_{1nr} &= -\frac{1}{nh_n^p} \sum_{i=1}^m \sum_{j=1}^{n_i} \frac{w(\mathbf{x}_{ij}, \boldsymbol{\theta}_l^*) [w(\mathbf{x}_{ij}, \boldsymbol{\theta}_r^*) - 1]}{(1 + \sum_{k=1}^q \mu_k^* [w(\mathbf{x}_{ij}, \boldsymbol{\theta}_k^*) - 1])^2} K\left(\frac{\mathbf{x} - \mathbf{x}_{ij}}{h_n}\right) \\
\mathbf{R}_{2nr} &= \frac{1}{nh_n^p} \sum_{i=1}^m \sum_{j=1}^{n_i} \mathbf{I}_r(\boldsymbol{\theta}^*, \mu^*) K\left(\frac{\mathbf{x} - \mathbf{x}_{ij}}{h_n}\right)
\end{aligned}$$



and  $(\mu_k^*, \boldsymbol{\theta}_k^*)$  is on the line segment between  $(\hat{\mu}_k, \hat{\boldsymbol{\theta}}_k)$  and  $(\zeta_k, \boldsymbol{\theta}_k)$ . Define

$$p_{ij}^* = \frac{1}{n} \frac{1}{1 + \sum_{k=1}^q \mu_k^* [w(\mathbf{x}_{ij}, \boldsymbol{\theta}_k^*) - 1]},$$

$$g_l^*(\mathbf{x}) = \frac{1}{h_n^p} \sum_{i=1}^m \sum_{j=1}^{n_i} p_{ij}^* w(\mathbf{x}_{ij}, \boldsymbol{\theta}_l^*) K\left(\frac{\mathbf{x} - \mathbf{x}_{ij}}{h_n}\right).$$

Since  $(\mu_k^*, \boldsymbol{\theta}_k^*)$  is on the line segment between  $(\hat{\mu}_k, \hat{\boldsymbol{\theta}}_k)$  and  $(\zeta_k, \boldsymbol{\theta}_k)$ , we may assume that  $\mu_k^* > 0$  and that  $p_{ij}^*$  is close to  $\hat{p}_{ij}$  or  $p_{ij}$ , and therefore, it is a positive bounded quantity. A similar statement holds for  $g_l^*(\mathbf{x})$ . Then

$$\begin{aligned} |R_{1nr}(\mathbf{x}_{ij})| &= \left| -\frac{1}{nh_n^p} \sum_{i=1}^m \sum_{j=1}^{n_i} \frac{w(\mathbf{x}_{ij}, \boldsymbol{\theta}_l^*) [w(\mathbf{x}_{ij}, \boldsymbol{\theta}_r^*) - 1]}{(1 + \sum_{k=1}^q \mu_k^* [w(\mathbf{x}_{ij}, \boldsymbol{\theta}_k^*) - 1])^2} K\left(\frac{\mathbf{x} - \mathbf{x}_{ij}}{h_n}\right) \right| \\ &\leq \frac{1}{h_n^p} \sum_{i=1}^m \sum_{j=1}^{n_i} p_{ij}^* w(\mathbf{x}_{ij}, \boldsymbol{\theta}_l^*) \left| \frac{[w(\mathbf{x}_{ij}, \boldsymbol{\theta}_r^*) - 1]}{1 + \sum_{k=1}^q \mu_k^* [w(\mathbf{x}_{ij}, \boldsymbol{\theta}_k^*) - 1]} \right| K\left(\frac{\mathbf{x} - \mathbf{x}_{ij}}{h_n}\right) \\ &= \frac{1}{h_n^p} \sum_{i=1}^m \sum_{j=1}^{n_i} p_{ij}^* w(\mathbf{x}_{ij}, \boldsymbol{\theta}_l^*) \left| \frac{1}{\mu_r^*} \frac{\mu_r^* [w(\mathbf{x}_{ij}, \boldsymbol{\theta}_r^*) - 1]}{1 + \sum_{k=1}^q \mu_k^* [w(\mathbf{x}_{ij}, \boldsymbol{\theta}_k^*) - 1]} \right| K\left(\frac{\mathbf{x} - \mathbf{x}_{ij}}{h_n}\right) \\ &\leq \frac{1}{\mu_r^* h_n^p} \sum_{i=1}^m \sum_{j=1}^{n_i} p_{ij}^* w(\mathbf{x}_{ij}, \boldsymbol{\theta}_l^*) K\left(\frac{\mathbf{x} - \mathbf{x}_{ij}}{h_n}\right) = \frac{1}{\mu_r^*} g_l^*(\mathbf{x}) = O_p(1) \end{aligned}$$

It remains to show why

$$\frac{\mu_r^* [w(\mathbf{x}_{ij}, \boldsymbol{\theta}_r^*) - 1]}{1 + \sum_{k=1}^q \mu_k^* [w(\mathbf{x}_{ij}, \boldsymbol{\theta}_k^*) - 1]} \leq 1.$$

Notice the following:

$$\begin{aligned} \frac{\mu_r^* [w(\mathbf{x}_{ij}, \boldsymbol{\theta}_r^*) - 1]}{1 + \sum_{k=1}^q \mu_k^* [w(\mathbf{x}_{ij}, \boldsymbol{\theta}_k^*) - 1]} &\leq 1 \\ \Leftrightarrow \mu_r^* [w(\mathbf{x}_{ij}, \boldsymbol{\theta}_r^*) - 1] &\leq 1 + \sum_{k=1}^q \mu_k^* [w(\mathbf{x}_{ij}, \boldsymbol{\theta}_k^*) - 1] \end{aligned}$$

$$\Leftrightarrow 1 + \sum_{\substack{k=1 \\ k \neq r}}^q \mu_k^* w(\mathbf{x}_{ij}, \boldsymbol{\theta}_k^*) - \sum_{\substack{k=1 \\ k \neq r}}^q \mu_k^* > 0,$$

which holds since  $\mu_k^*$  is close to  $\hat{\mu}_k$ ,  $\zeta_k = n_k/n$  and  $1 - \sum_{\substack{k=1 \\ k \neq r}}^q \zeta_k \geq 0$ . Similarly:

$$\begin{aligned} \|\mathbf{R}_{2nr}(\mathbf{x}_{ij})\| &= \left\| \frac{1}{nh_n^p} \sum_{i=1}^m \sum_{j=1}^{n_i} \mathbf{I}_r(\boldsymbol{\theta}^*, \mu^*) K\left(\frac{\mathbf{x} - \mathbf{x}_{ij}}{h_n}\right) \right\| \\ &= \left\| \frac{1}{nh_n^p} \sum_{i=1}^m \sum_{j=1}^{n_i} \frac{\partial}{\partial \boldsymbol{\theta}_r} \frac{w(\mathbf{x}_{ij}, \boldsymbol{\theta}_l)}{1 + \sum_{k=1}^q \mu_k [w(\mathbf{x}_{ij}, \boldsymbol{\theta}_k) - 1]} \Big|_{\substack{\mu_k = \mu_k^* \\ \boldsymbol{\theta}_r = \boldsymbol{\theta}_r^*}} K\left(\frac{\mathbf{x} - \mathbf{x}_{ij}}{h_n}\right) \right\| \\ &\leq \frac{1}{nh_n^p} \sum_{i=1}^m \sum_{j=1}^{n_i} \left\| \frac{\partial}{\partial \boldsymbol{\theta}_r} \frac{w(\mathbf{x}_{ij}, \boldsymbol{\theta}_l)}{1 + \sum_{k=1}^q \mu_k [w(\mathbf{x}_{ij}, \boldsymbol{\theta}_k) - 1]} \Big|_{\substack{\mu_k = \mu_k^* \\ \boldsymbol{\theta}_r = \boldsymbol{\theta}_r^*}} \right\| K\left(\frac{\mathbf{x} - \mathbf{x}_{ij}}{h_n}\right) \\ &= O_p(1) \end{aligned}$$

The above holds because

$$\left\| \frac{\partial}{\partial \boldsymbol{\theta}_r} \frac{w(\mathbf{x}_{ij}, \boldsymbol{\theta}_l)}{1 + \sum_{k=1}^q \mu_k [w(\mathbf{x}_{ij}, \boldsymbol{\theta}_k) - 1]} \Big|_{\substack{\mu_k = \mu_k^* \\ \boldsymbol{\theta}_r = \boldsymbol{\theta}_r^*}} \right\|$$

is by assumption bounded.

From [19] we have that  $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \rightarrow N(\mathbf{0}, \mathbf{W})$ , so

$$\begin{aligned} R_n(\mathbf{x}_{ij}) &= \sum_{r=1}^q R_{1nr}(\hat{\mu}_r - \zeta_r) + \sum_{r=1}^q \mathbf{R}_{2nr}(\hat{\boldsymbol{\theta}}_r - \boldsymbol{\theta}_r) \\ &= O_p(1)O_p(n^{-1/2}) + O_p(1)O_p(n^{-1/2}) = O_p(n^{-1/2}). \end{aligned}$$

□

**Remark 3.1.** The proof of Lemma 3.3 can be greatly simplified depending on the choice for  $w(\mathbf{x}, \boldsymbol{\theta}_i)$  in the general model (2.1). Qin and Zhang [62] take  $w(x, \alpha, \beta) =$

$\exp(\alpha + r(x)\beta)$ . In this case  $\mathbf{I}_r(\boldsymbol{\theta}, \mu)$  is clearly bounded.

The following lemma will be used in the proof of Theorem 3.1:

**Lemma 3.4.** Let  $E_i(\mathbf{x})$  denote the expected value of  $\mathbf{x}$  with respect to sample  $i$ .

Then

$$\begin{aligned} E_i \left| \frac{w_l(\mathbf{x}_{i1}) K\left(\frac{\mathbf{x}-\mathbf{x}_{i1}}{h_n}\right)}{\sum_{k=1}^m \zeta_k w_k(\mathbf{x}_{i1})} \right|^{2+\delta} &= O(h_n^p), \\ \left| \int \frac{w_l(\mathbf{y}) K\left(\frac{\mathbf{x}-\mathbf{y}}{h_n}\right)}{\sum_{k=1}^m \zeta_k w_k(\mathbf{y})} w_i(\mathbf{y}) dG(\mathbf{y}) \right|^{2+\delta} &= O(h_n^{p(2+\delta)}). \end{aligned}$$

*Proof.* Define

$$H_{\alpha\beta\gamma}(\mathbf{x}) = \frac{1}{nh_n^p} \sum_{i=1}^m \sum_{j=1}^{n_i} \frac{w_l^\alpha(\mathbf{x}_{ij}) D(\mathbf{x}_{ij})}{\left(\sum_{k=1}^m \zeta_k w_k(\mathbf{x}_{ij})\right)^\beta} K^\gamma\left(\frac{\mathbf{x}-\mathbf{x}_{ij}}{h_n}\right)$$

where  $D(\cdot)$  is some measurable function. Then

$$\begin{aligned} E(H_{\alpha\beta\gamma}(\mathbf{x})) &= \frac{1}{nh_n^p} \sum_{i=1}^m n_i E\left(\frac{w_l^\alpha(\mathbf{x}_{ij}) D(\mathbf{x}_{ij})}{\left(\sum_{k=1}^m \zeta_k w_k(\mathbf{x}_{ij})\right)^\beta} K^\gamma\left(\frac{\mathbf{x}-\mathbf{x}_{ij}}{h_n}\right)\right) \\ &= \frac{1}{nh_n^p} \sum_{i=1}^m n_i \int \frac{w_l^\alpha(\mathbf{y}) D(\mathbf{y})}{\left(\sum_{k=1}^m \zeta_k w_k(\mathbf{y})\right)^\beta} K^\gamma\left(\frac{\mathbf{x}-\mathbf{y}}{h_n}\right) w_i(\mathbf{y}) dG(\mathbf{y}) \\ &= \frac{1}{h_n^p} \sum_{i=1}^m \zeta_i \int \frac{w_i(\mathbf{y}) D(\mathbf{y})}{\left(\sum_{k=1}^m \zeta_k w_k(\mathbf{y})\right)^\beta} K^\gamma\left(\frac{\mathbf{x}-\mathbf{y}}{h_n}\right) w_l^\alpha(\mathbf{y}) dG(\mathbf{y}) \\ &= \frac{1}{h_n^p} \int \frac{w_l^\alpha(\mathbf{y}) D(\mathbf{y})}{\left(\sum_{k=1}^m \zeta_k w_k(\mathbf{y})\right)^{\beta-1}} K^\gamma\left(\frac{\mathbf{x}-\mathbf{y}}{h_n}\right) dG(\mathbf{y}), \end{aligned} \tag{3.7}$$

assuming the integral is finite. Set  $\alpha = \beta - 1 = \gamma = 2 + \delta$  and  $D(\mathbf{y}) = w_i(\mathbf{y})$  in

(3.7). Then

$$E_i \left| \frac{w_l(\mathbf{x}_{i1}) K\left(\frac{\mathbf{x}-\mathbf{x}_{i1}}{h_n}\right)}{\sum_{k=1}^m \zeta_k w_k(\mathbf{x}_{i1})} \right|^{2+\delta} = h_n^p E(H_{2+\delta, 3+\delta, 2+\delta}) = O(h_n^p).$$

Also, if we replace  $\alpha = \beta - 1 = \gamma = 1$  and  $D(\mathbf{y}) = w_i(\mathbf{y})$  in (3.7),

$$\left| \int \frac{w_l(\mathbf{y}) K\left(\frac{\mathbf{x}-\mathbf{y}}{h_n}\right)}{\sum_{k=1}^m \zeta_k w_k(\mathbf{y})} w_i(\mathbf{y}) dG(\mathbf{y}) \right|^{2+\delta} = \left| E(H_{1,2,1}) h_n^p \right|^{2+\delta} = O(h_n^{p(2+\delta)}).$$

□

**Theorem 3.1.** *Assume that  $K(\cdot)$  is a nonnegative bounded symmetric function with  $\int K(\mathbf{x}) d\mathbf{x} = 1$ ,  $\int \mathbf{x}'\mathbf{x}K(\mathbf{x}) d\mathbf{x} = k_2 > 0$ . Assume that  $g_l$  is continuous at  $\mathbf{x}$ . If  $\int [K(\mathbf{u})]^{2+\delta} d\mathbf{u} < \infty$  for some  $\delta > 0$ , then*

$$\sqrt{nh_n^p} (\hat{g}_l(\mathbf{x}) - E\tilde{g}_l(\mathbf{x})) \xrightarrow{D} N(\mathbf{0}, \sigma^2(\mathbf{x}))$$

as  $n \rightarrow \infty$ ,  $h_n \rightarrow 0$  and  $nh_n^p \rightarrow \infty$  with

$$\sigma^2(\mathbf{x}) = \frac{w_l(\mathbf{x})g_l(\mathbf{x})}{\sum_{k=1}^m \zeta_k w_k(\mathbf{x})} \int K^2(\mathbf{u}) d\mathbf{u}$$

for any fixed  $\mathbf{x}$ .

*Proof.* It suffices to show that

$$\sqrt{nh_n^p} (\tilde{g}_l(\mathbf{x}) - E\tilde{g}_l(\mathbf{x})) \xrightarrow{D} N(\mathbf{0}, \sigma^2(\mathbf{x}))$$

since, by Lemma 3.3, we have that

$$\sqrt{nh_n^p} (\hat{g}_l(\mathbf{x}) - E\tilde{g}_l(\mathbf{x})) = \sqrt{nh_n^p} (\tilde{g}_l(\mathbf{x}) - E\tilde{g}_l(\mathbf{x})) + O_p(\sqrt{h_n^p}).$$

Then

$$\begin{aligned} \sqrt{nh_n^p} (\tilde{g}_l(\mathbf{x}) - E\tilde{g}_l(\mathbf{x})) &= \sqrt{nh_n^p} \left[ \frac{1}{nh_n^p} \sum_{i=1}^m \sum_{j=1}^{n_i} \frac{w_l(\mathbf{x}_{ij})}{\sum_{k=1}^m \zeta_k w_k(\mathbf{x}_{ij})} K\left(\frac{\mathbf{x} - \mathbf{x}_{ij}}{h_n}\right) \right. \\ &\quad \left. - \frac{1}{h_n^p} \sum_{i=1}^m \zeta_i \int \frac{w_l(\mathbf{y})}{\sum_{k=1}^m \zeta_k w_k(\mathbf{y})} K\left(\frac{\mathbf{x} - \mathbf{y}}{h_n}\right) w_i(\mathbf{y}) dG(\mathbf{y}) \right] \\ &= \sum_{i=1}^m \left[ \frac{1}{\sqrt{nh_n^p}} \sum_{j=1}^{n_i} \frac{w_l(\mathbf{x}_{ij})}{\sum_{k=1}^m \zeta_k w_k(\mathbf{x}_{ij})} K\left(\frac{\mathbf{x} - \mathbf{x}_{ij}}{h_n}\right) \right. \\ &\quad \left. - \frac{\sqrt{n}}{\sqrt{h_n^p}} \zeta_i \int \frac{w_l(\mathbf{y})}{\sum_{k=1}^m \zeta_k w_k(\mathbf{y})} K\left(\frac{\mathbf{x} - \mathbf{y}}{h_n}\right) w_i(\mathbf{y}) dG(\mathbf{y}) \right] \\ &= \sum_{i=1}^m \sum_{j=1}^{n_i} \left[ \frac{1}{\sqrt{nh_n^p}} \frac{w_l(\mathbf{x}_{ij})}{\sum_{k=1}^m \zeta_k w_k(\mathbf{x}_{ij})} K\left(\frac{\mathbf{x} - \mathbf{x}_{ij}}{h_n}\right) \right. \\ &\quad \left. - \frac{1}{\sqrt{nh_n^p}} \int \frac{w_l(\mathbf{y})}{\sum_{k=1}^m \zeta_k w_k(\mathbf{y})} K\left(\frac{\mathbf{x} - \mathbf{y}}{h_n}\right) w_i(\mathbf{y}) dG(\mathbf{y}) \right] \\ &= \sum_{i=1}^m U_{ni}(\mathbf{x}) \end{aligned}$$

where

$$U_{ni}(\mathbf{x}) = \frac{1}{\sqrt{nh_n^p}} \sum_{j=1}^{n_i} \left[ \frac{w_l(\mathbf{x}_{ij})}{\sum_{k=1}^m \zeta_k w_k(\mathbf{x}_{ij})} K\left(\frac{\mathbf{x} - \mathbf{x}_{ij}}{h_n}\right) - \int \frac{w_l(\mathbf{y})}{\sum_{k=1}^m \zeta_k w_k(\mathbf{y})} K\left(\frac{\mathbf{x} - \mathbf{y}}{h_n}\right) w_i(\mathbf{y}) dG(\mathbf{y}) \right].$$

Notice that for  $i = 1, \dots, m$ :

$$\begin{aligned} EU_{ni}(\mathbf{x}) &= E \left[ \frac{1}{\sqrt{nh_n^p}} \sum_{j=1}^{n_i} \frac{w_l(\mathbf{x}_{ij})}{\sum_{k=1}^m \zeta_k w_k(\mathbf{x}_{ij})} K\left(\frac{\mathbf{x} - \mathbf{x}_{ij}}{h_n}\right) \right] \\ &\quad - \frac{1}{\sqrt{nh_n^p}} \sum_{j=1}^{n_i} \int \frac{w_l(\mathbf{y})}{\sum_{k=1}^m \zeta_k w_k(\mathbf{y})} K\left(\frac{\mathbf{x} - \mathbf{y}}{h_n}\right) w_i(\mathbf{y}) dG(\mathbf{y}) = 0. \end{aligned}$$

Clearly,

$$E \left[ \sqrt{nh_n^p} (\tilde{g}_l(\mathbf{x}) - E\tilde{g}_l(\mathbf{x})) \right] = 0.$$

Moreover, from Lemma 3.2:

- $\text{Var}(\sqrt{nh_n^p}(\tilde{g}_l(\mathbf{x}) - E\tilde{g}_l(\mathbf{x}))) = nh_n^p \text{Var}(\tilde{g}_l(\mathbf{x})) = \sigma^2(\mathbf{x}) + o(1)$
- $\text{Var}(\sqrt{nh_n^p}(\tilde{g}_l(\mathbf{x}) - E\tilde{g}_l(\mathbf{x}))) = \sum_{i=1}^m \text{Var}(U_{ni}(\mathbf{x})) = s_n^2(\mathbf{x})$

where  $\mathbf{x}$  is fixed. So  $s_n^2(\mathbf{x}) = \sigma^2(\mathbf{x}) + o(1)$ . Observe that:

$$\begin{aligned} \sqrt{nh_n^p}(\tilde{g}_l(\mathbf{x}) - E\tilde{g}_l(\mathbf{x})) &= \sum_{i=1}^m U_{ni}(\mathbf{x}) \\ &= \sum_{j=1}^{n_1} \frac{1}{\sqrt{nh_n^p}} \left( \frac{w_l(\mathbf{x}_{1j}) K\left(\frac{\mathbf{x}-\mathbf{x}_{1j}}{h_n}\right)}{\sum_{k=1}^m \zeta_k w_k(\mathbf{x}_{1j})} - \int \frac{w_l(\mathbf{y}) K\left(\frac{\mathbf{x}-\mathbf{y}}{h_n}\right)}{\sum_{k=1}^m \zeta_k w_k(\mathbf{y})} dG_1(\mathbf{y}) \right) \\ &\quad + \sum_{j=1}^{n_2} \frac{1}{\sqrt{nh_n^p}} \left( \frac{w_l(\mathbf{x}_{2j}) K\left(\frac{\mathbf{x}-\mathbf{x}_{2j}}{h_n}\right)}{\sum_{k=1}^m \zeta_k w_k(\mathbf{x}_{2j})} - \int \frac{w_l(\mathbf{y}) K\left(\frac{\mathbf{x}-\mathbf{y}}{h_n}\right)}{\sum_{k=1}^m \zeta_k w_k(\mathbf{y})} dG_2(\mathbf{y}) \right) \\ &\quad \vdots \\ &\quad + \sum_{j=1}^{n_m} \frac{1}{\sqrt{nh_n^p}} \left( \frac{w_l(\mathbf{x}_{mj}) K\left(\frac{\mathbf{x}-\mathbf{x}_{mj}}{h_n}\right)}{\sum_{k=1}^m \zeta_k w_k(\mathbf{x}_{mj})} - \int \frac{w_l(\mathbf{y}) K\left(\frac{\mathbf{x}-\mathbf{y}}{h_n}\right)}{\sum_{k=1}^m \zeta_k w_k(\mathbf{y})} dG_m(\mathbf{y}) \right). \end{aligned}$$

We will show that Lyapunov's Condition ([7], p. 362) holds using the  $c_r$ -inequality:

$$E|x + y|^r \leq c_r [E|x|^r + E|y|^r], \text{ where } c_r = \begin{cases} 1, & \text{if } 0 < r \leq 1 \\ 2^{r-1}, & \text{if } r > 1 \end{cases}$$

We have already showed that  $E(U_{ni})$ ,  $\text{Var}(U_{ni})$  are finite for  $i = 1, 2, \dots, m$ . For

some  $\delta > 0$ , the Lyapunov condition becomes:

$$\begin{aligned}
& \frac{1}{s_n^{2+\delta}} \left[ \sum_{j=1}^{n_1} E \left| \frac{1}{\sqrt{nh_n^p}} \left( \frac{w_l(\mathbf{x}_{1j}) K \left( \frac{\mathbf{x}-\mathbf{x}_{1j}}{h_n} \right)}{\sum_{k=1}^m \zeta_k w_k(\mathbf{x}_{1j})} - \int \frac{w_l(\mathbf{y}) K \left( \frac{\mathbf{x}-\mathbf{y}}{h_n} \right)}{\sum_{k=1}^m \zeta_k w_k(\mathbf{y})} w_1(\mathbf{y}) dG(\mathbf{y}) \right) \right|^{2+\delta} \right. \\
& + \sum_{j=1}^{n_2} E \left| \frac{1}{\sqrt{nh_n^p}} \left( \frac{w_l(\mathbf{x}_{2j}) K \left( \frac{\mathbf{x}-\mathbf{x}_{2j}}{h_n} \right)}{\sum_{k=1}^m \zeta_k w_k(\mathbf{x}_{2j})} - \int \frac{w_l(\mathbf{y}) K \left( \frac{\mathbf{x}-\mathbf{y}}{h_n} \right)}{\sum_{k=1}^m \zeta_k w_k(\mathbf{y})} w_2(\mathbf{y}) dG(\mathbf{y}) \right) \right|^{2+\delta} \\
& \vdots \\
& + \sum_{j=1}^{n_m} E \left| \frac{1}{\sqrt{nh_n^p}} \left( \frac{w_l(\mathbf{x}_{mj}) K \left( \frac{\mathbf{x}-\mathbf{x}_{mj}}{h_n} \right)}{\sum_{k=1}^m \zeta_k w_k(\mathbf{x}_{mj})} - \int \frac{w_l(\mathbf{y}) K \left( \frac{\mathbf{x}-\mathbf{y}}{h_n} \right)}{\sum_{k=1}^m \zeta_k w_k(\mathbf{y})} w_m(\mathbf{y}) dG(\mathbf{y}) \right) \right|^{2+\delta} \Big] \\
& \leq \frac{1}{s_n^{2+\delta}} \frac{1}{n^{1+\frac{\delta}{2}} h_n^{p(1+\frac{\delta}{2})}} \left[ n_1 2^{\delta+1} E_1 \left| \frac{w_l(\mathbf{x}_{11}) K \left( \frac{\mathbf{x}-\mathbf{x}_{11}}{h_n} \right)}{\sum_{k=1}^m \zeta_k w_k(\mathbf{x}_{11})} \right|^{2+\delta} \right. \\
& + n_1 2^{\delta+1} \left| \int \frac{w_l(\mathbf{y}) K \left( \frac{\mathbf{x}-\mathbf{y}}{h_n} \right)}{\sum_{k=1}^m \zeta_k w_k(\mathbf{y})} w_1(\mathbf{y}) dG(\mathbf{y}) \right|^{2+\delta} \\
& + n_2 2^{\delta+1} E_2 \left| \frac{w_l(\mathbf{x}_{21}) K \left( \frac{\mathbf{x}-\mathbf{x}_{21}}{h_n} \right)}{\sum_{k=1}^m \zeta_k w_k(\mathbf{x}_{21})} \right|^{2+\delta} + n_2 2^{\delta+1} \left| \int \frac{w_l(\mathbf{y}) K \left( \frac{\mathbf{x}-\mathbf{y}}{h_n} \right)}{\sum_{k=1}^m \zeta_k w_k(\mathbf{y})} w_2(\mathbf{y}) dG(\mathbf{y}) \right|^{2+\delta} \\
& \vdots \\
& + n_m 2^{\delta+1} E_m \left| \frac{w_l(\mathbf{x}_{m1}) K \left( \frac{\mathbf{x}-\mathbf{x}_{m1}}{h_n} \right)}{\sum_{k=1}^m \zeta_k w_k(\mathbf{x}_{m1})} \right|^{2+\delta} \\
& \left. + n_m 2^{\delta+1} \left| \int \frac{w_l(\mathbf{y}) K \left( \frac{\mathbf{x}-\mathbf{y}}{h_n} \right)}{\sum_{k=1}^m \zeta_k w_k(\mathbf{y})} w_m(\mathbf{y}) dG(\mathbf{y}) \right|^{2+\delta} \right] \\
& \stackrel{\text{Lemma 3.4}}{=} \frac{1}{s_n^{2+\delta}} \frac{1}{n^{1+\delta/2} h_n^{p(1+\delta/2)}} \left[ n_1 2^{\delta+1} O(h_n^p) + n_1 2^{\delta+1} O(h_n^{p(2+\delta)}) + \dots \right. \\
& \left. + n_m 2^{\delta+1} O(h_n^p) + n_m 2^{\delta+1} O(h_n^{p(2+\delta)}) \right] \\
& = O((nh_n^p)^{-\delta/2}),
\end{aligned}$$

where  $E_i$  is the expected value with respect to the  $i$  sample.

Therefore, as  $h_n \rightarrow 0$ ,  $n \rightarrow \infty$  and  $nh_n^p \rightarrow \infty$ , by the Lyapunov condition:

$$\frac{\sqrt{nh_n^p}(\tilde{g}_l(\mathbf{x}) - E\tilde{g}_l(\mathbf{x}))}{s_n(\mathbf{x})} \xrightarrow{D} N(\mathbf{0}, 1).$$

By Slutsky's theorem:

$$\frac{\sqrt{nh_n^p}(\tilde{g}_l(\mathbf{x}) - E\tilde{g}_l(\mathbf{x}))}{\sigma(\mathbf{x})} = \frac{\sqrt{nh_n^p}(\tilde{g}_l(\mathbf{x}) - E\tilde{g}_l(\mathbf{x}))}{s_n(\mathbf{x})} \frac{\sqrt{\sigma^2(\mathbf{x}) + o(1)}}{\sigma(\mathbf{x})} \xrightarrow{D} N(\mathbf{0}, 1).$$

□

**Corollary 3.1.** *Assume that  $K(\cdot)$  is a nonnegative bounded symmetric function with  $\int K(\mathbf{x})d\mathbf{x} = 1$ ,  $\int \mathbf{x}'\mathbf{x}K(\mathbf{x})d\mathbf{x} = k_2 > 0$ . Assume that  $g_l$  is continuous at  $\mathbf{x}$  and twice differentiable in a neighborhood of  $\mathbf{x}$ . If, as  $n \rightarrow \infty$ ,  $h_n = O(n^{-\frac{1}{4+p}})$ , then*

$$\sqrt{nh_n^p} \left( \hat{g}_l(\mathbf{x}) - g(\mathbf{x}) - \frac{1}{2}h_n^2 \int \mathbf{u}' \frac{\partial^2 g_l(\mathbf{x}^*)}{\partial \mathbf{x} \partial \mathbf{x}'} \mathbf{u} K(\mathbf{u}) d\mathbf{u} \right) \xrightarrow{D} N(\mathbf{0}, \sigma^2(\mathbf{x}))$$

as  $n \rightarrow \infty$

*Proof.* From Theorem 3.1 and Lemma 3.2 we have:

$$\sqrt{nh_n^p}(\hat{g}_l(\mathbf{x}) - E\hat{g}_l(\mathbf{x})) \xrightarrow{D} N(\mathbf{0}, \sigma^2(\mathbf{x})),$$

or equivalently

$$\sqrt{nh_n^p} \left( \hat{g}_l(\mathbf{x}) - g(\mathbf{x}) - \frac{1}{2}h_n^2 \int \mathbf{u}' \frac{\partial^2 g_l(\mathbf{x}^*)}{\partial \mathbf{x} \partial \mathbf{x}'} \mathbf{u} K(\mathbf{u}) d\mathbf{u} - o(h_n^2) \right) \xrightarrow{D} N(\mathbf{0}, \sigma^2(\mathbf{x})).$$



Therefore

$$\sqrt{nh_n^p} \left( \hat{g}_l(\mathbf{x}) - g(\mathbf{x}) - \frac{1}{2}h_n^2 \int \mathbf{u}' \frac{\partial^2 g_l(\mathbf{x}^*)}{\partial \mathbf{x} \partial \mathbf{x}'} \mathbf{u} K(\mathbf{u}) d\mathbf{u} \right) - o(\sqrt{nh_n^{4+p}}) \xrightarrow{D} N(\mathbf{0}, \sigma^2(\mathbf{x})).$$

Since, as  $n \rightarrow \infty$ ,  $h_n = O(n^{-\frac{1}{4+p}})$ , then also  $o(\sqrt{nh_n^{4+p}})$  goes to 0, as  $n \rightarrow \infty$ .  $\square$

### 3.3.2 Comparison of $\hat{g}_l$ and the traditional $\hat{f}$

**Definition 3.1.** *The mean integrated square error (MISE) is defined as:*

$$MISE(\hat{g}_l(\mathbf{x})) = E \left( \int |\hat{g}_l(\mathbf{x}) - g_l(\mathbf{x})|^2 d\mathbf{x} \right) \quad (3.8)$$

**Theorem 3.2.** *Assume that  $K(\cdot)$  is a nonnegative bounded symmetric function with  $\int K(\mathbf{x}) d\mathbf{x} = 1$ ,  $\int \mathbf{x}' \mathbf{x} K(\mathbf{x}) d\mathbf{x} = k_2 > 0$  and  $\int K^2(\mathbf{x}) d\mathbf{x} < \infty$ . If  $g_l$  is twice continuously differentiable at  $\mathbf{x}$  and the conditions in Lemma 2.1 hold, then*

(a) as  $n \rightarrow \infty$ ,  $h_n \rightarrow 0$  and  $nh_n^p \rightarrow \infty$

$$\begin{aligned} MISE(\hat{g}_l) &= \frac{1}{nh_n^p} \int \frac{w_l(\mathbf{x}) g_l(\mathbf{x})}{\sum_{k=1}^m \zeta_k w_k(\mathbf{x})} d\mathbf{x} \int K^2(\mathbf{u}) d\mathbf{u} \\ &\quad + \frac{h_n^4}{4} \int \left( \int \mathbf{u}' \frac{\partial^2 g_l(\mathbf{x})}{\partial \mathbf{x} \partial \mathbf{x}'} \mathbf{u} K(\mathbf{u}) d\mathbf{u} \right)^2 d\mathbf{x} + o\left(\frac{1}{nh_n^p}\right) + o(h_n^4), \end{aligned}$$

(b) by minimizing the sum of the two leading terms in (a) with respect to  $h_n$ , the

asymptotically optimal bandwidth is:

$$h_n^* = \left( \frac{(p/n) \int w_l(\mathbf{x})g_l(\mathbf{x}) / [\sum_{k=1}^m \zeta_k w_k(\mathbf{x})] d\mathbf{x} \int K^2(\mathbf{u}) d\mathbf{u}}{\int \left( \int \mathbf{u}' (\partial^2 g_l(\mathbf{x}) / \partial \mathbf{x} \partial \mathbf{x}') \mathbf{u} K(\mathbf{u}) d\mathbf{u} \right)^2 d\mathbf{x}} \right)^{\frac{1}{4+p}}. \quad (3.9)$$

The mean integrated square error of  $\hat{g}_l$  with optimal bandwidth  $h_n^*$  is:

$$\begin{aligned} MISE^*(\hat{g}_l) &= n^{-\frac{4}{4+p}} \left( p^{-\frac{p}{4+p}} + \frac{1}{4} p^{\frac{4}{4+p}} \right) \left( \int \frac{w_l(\mathbf{x})g_l(\mathbf{x})}{\sum_{k=1}^m \zeta_k w_k(\mathbf{x})} d\mathbf{x} \int K^2(\mathbf{u}) d\mathbf{u} \right)^{\frac{4}{4+p}} \\ &\quad \cdot \left( \int \left( \int \mathbf{u}' \frac{\partial^2 g_l(\mathbf{x})}{\partial \mathbf{x} \partial \mathbf{x}'} \mathbf{u} K(\mathbf{u}) d\mathbf{u} \right)^2 d\mathbf{x} \right)^{\frac{p}{4+p}} + o(n^{-\frac{4}{4+p}}). \end{aligned}$$

*Proof.* (a)

$$\begin{aligned} MISE(\hat{g}_l) &= E \int |\hat{g}_l(\mathbf{x}) - g_l(\mathbf{x})|^2 d\mathbf{x} \\ &= E \int |\hat{g}_l(\mathbf{x}) - \tilde{g}_l(\mathbf{x}) + \tilde{g}_l(\mathbf{x}) - g_l(\mathbf{x})|^2 d\mathbf{x} \\ &= E \left[ \int (\hat{g}_l(\mathbf{x}) - \tilde{g}_l(\mathbf{x}))^2 d\mathbf{x} + \int (\tilde{g}_l(\mathbf{x}) - g_l(\mathbf{x}))^2 d\mathbf{x} \right. \\ &\quad \left. + 2 \int (\hat{g}_l(\mathbf{x}) - \tilde{g}_l(\mathbf{x})) (\tilde{g}_l(\mathbf{x}) - g_l(\mathbf{x})) d\mathbf{x} \right] \\ &= \int \left[ E(\hat{g}_l(\mathbf{x}) - \tilde{g}_l(\mathbf{x}))^2 + E(\tilde{g}_l(\mathbf{x}) - g_l(\mathbf{x}))^2 \right. \\ &\quad \left. + 2E(\hat{g}_l(\mathbf{x}) - \tilde{g}_l(\mathbf{x})) (\tilde{g}_l(\mathbf{x}) - g_l(\mathbf{x})) \right] d\mathbf{x} \end{aligned}$$

However

- By construction:  $E(\hat{g}_l(\mathbf{x}) - \tilde{g}_l(\mathbf{x}))^2 = O(n^{-1})$  (see also [62])

- From Lemma 3.2:

$$\begin{aligned}
E(\tilde{g}_l(\mathbf{x}) - g_l(\mathbf{x}))^2 &= \text{Var}(\tilde{g}_l(\mathbf{x})) + (E\tilde{g}_l(\mathbf{x}) - g_l(\mathbf{x}))^2 \\
&= \frac{1}{nh_n^p} \sigma^2(\mathbf{x}) + o\left(\frac{1}{nh_n^p}\right) + \left(\frac{h_n^2}{2} \left( \int \mathbf{u}' \frac{\partial^2 g_l(\mathbf{x})}{\partial \mathbf{x} \partial \mathbf{x}'} \mathbf{u} K(\mathbf{u}) d\mathbf{u} + o(h_n^2) \right)\right)^2 \\
&= \frac{1}{nh_n^p} \sigma^2(\mathbf{x}) + \frac{h_n^4}{4} \left( \int \mathbf{u}' \frac{\partial^2 g_l(\mathbf{x})}{\partial \mathbf{x} \partial \mathbf{x}'} \mathbf{u} K(\mathbf{u}) d\mathbf{u} \right)^2 + o\left(\frac{1}{nh_n^p}\right) + o(h_n^4)
\end{aligned}$$

- Using Schwartz inequality

$$\begin{aligned}
&| E(\hat{g}_l(\mathbf{x}) - \tilde{g}_l(\mathbf{x}))(\tilde{g}_l(\mathbf{x}) - g_l(\mathbf{x})) | \\
&\leq \sqrt{E(\hat{g}_l(\mathbf{x}) - \tilde{g}_l(\mathbf{x}))^2 E(\tilde{g}_l(\mathbf{x}) - g_l(\mathbf{x}))^2} \\
&= \sqrt{O(n^{-1}) \left[ \frac{\sigma^2(\mathbf{x})}{nh_n^p} + \frac{h_n^4}{4} \left( \int \mathbf{u}' \frac{\partial^2 g_l(\mathbf{x})}{\partial \mathbf{x} \partial \mathbf{x}'} \mathbf{u} K(\mathbf{u}) d\mathbf{u} \right)^2 + o\left(\frac{1}{nh_n^p}\right) + o(h_n^4) \right]} \\
&= \sqrt{O\left(\frac{1}{n^2 h_n^p}\right)} = O((n^2 h_n^p)^{-1/2})
\end{aligned}$$

Therefore

$$\begin{aligned}
MISE(\hat{g}_l) &= \int \left[ \frac{1}{nh_n^p} \sigma^2(\mathbf{x}) + \frac{h_n^4}{4} \left( \int \mathbf{u}' \frac{\partial^2 g_l(\mathbf{x})}{\partial \mathbf{x} \partial \mathbf{x}'} \mathbf{u} K(\mathbf{u}) d\mathbf{u} \right)^2 \right] d\mathbf{x} + o\left(\frac{1}{nh_n^p}\right) + o(h_n^4) \\
&= \frac{1}{nh_n^p} \int \frac{w_l(\mathbf{x}) g_l(\mathbf{x})}{\sum_{k=1}^m \zeta_k w_k(\mathbf{x})} d\mathbf{x} \int K^2(\mathbf{u}) d\mathbf{u} \\
&\quad + \frac{h_n^4}{4} \int \left( \int \mathbf{u}' \frac{\partial^2 g_l(\mathbf{x})}{\partial \mathbf{x} \partial \mathbf{x}'} \mathbf{u} K(\mathbf{u}) d\mathbf{u} \right)^2 d\mathbf{x} + o\left(\frac{1}{nh_n^p}\right) + o(h_n^4)
\end{aligned}$$

(b) Differentiate  $MISE(\hat{g}_l)$  with respect to  $h_n$  and set equal to 0. We may ignore

the terms  $o((nh_n^p)^{-1})$ ,  $o(h_n^4)$ .

$$\begin{aligned}
& \frac{\partial}{\partial h_n} MISE(\hat{g}_l) = 0 \\
& \Rightarrow \frac{-p}{nh_n^{p+1}} \int \frac{w_l(\mathbf{x})g_l(\mathbf{x})}{\sum_{k=1}^m \zeta_k w_k(\mathbf{x})} d\mathbf{x} \int K^2(\mathbf{u})d\mathbf{u} + h_n^3 \int \left( \int \mathbf{u}' \frac{\partial^2 g_l(\mathbf{x})}{\partial \mathbf{x} \partial \mathbf{x}'} \mathbf{u} K(\mathbf{u}) d\mathbf{u} \right)^2 d\mathbf{x} = 0 \\
& \Rightarrow \frac{p}{n} \int \frac{w_l(\mathbf{x})g_l(\mathbf{x})}{\sum_{k=1}^m \zeta_k w_k(\mathbf{x})} d\mathbf{x} \int K^2(\mathbf{u})d\mathbf{u} = h_n^{4+p} \int \left( \int \mathbf{u}' \frac{\partial^2 g_l(\mathbf{x})}{\partial \mathbf{x} \partial \mathbf{x}'} \mathbf{u} K(\mathbf{u}) d\mathbf{u} \right)^2 d\mathbf{x} \\
& \Rightarrow h_n^* = \left( \frac{\frac{p}{n} \int \frac{w_l(\mathbf{x})g_l(\mathbf{x})}{\sum_{k=1}^m \zeta_k w_k(\mathbf{x})} d\mathbf{x} \int K^2(\mathbf{u})d\mathbf{u}}{\int \left( \int \mathbf{u}' \frac{\partial^2 g_l(\mathbf{x})}{\partial \mathbf{x} \partial \mathbf{x}'} \mathbf{u} K(\mathbf{u}) d\mathbf{u} \right)^2 d\mathbf{x}} \right)^{\frac{1}{4+p}}
\end{aligned}$$

Therefore the mean integrated square error of  $\hat{g}_l$  with optimal bandwidth  $h_n^*$  is:

$$\begin{aligned}
& MISE^*(\hat{g}_l) = \\
& = \frac{1}{n} \left( \frac{n \int \left( \int \mathbf{u}' \frac{\partial^2 g_l(\mathbf{x})}{\partial \mathbf{x} \partial \mathbf{x}'} \mathbf{u} K(\mathbf{u}) d\mathbf{u} \right)^2 d\mathbf{x}}{p \int \frac{w_l(\mathbf{x})g_l(\mathbf{x})}{\sum_{k=1}^m \zeta_k w_k(\mathbf{x})} d\mathbf{x} \int K^2(\mathbf{u})d\mathbf{u}} \right)^{\frac{p}{4+p}} \int \frac{w_l(\mathbf{x})g_l(\mathbf{x})}{\sum_{k=1}^m \zeta_k w_k(\mathbf{x})} d\mathbf{x} \int K^2(\mathbf{u})d\mathbf{u} \\
& + \frac{1}{4} \left( \frac{p \int \frac{w_l(\mathbf{x})g_l(\mathbf{x})}{\sum_{k=1}^m \zeta_k w_k(\mathbf{x})} d\mathbf{x} \int K^2(\mathbf{u})d\mathbf{u}}{n \int \left( \int \mathbf{u}' \frac{\partial^2 g_l(\mathbf{x})}{\partial \mathbf{x} \partial \mathbf{x}'} \mathbf{u} K(\mathbf{u}) d\mathbf{u} \right)^2 d\mathbf{x}} \right)^{\frac{4}{4+p}} \int \left( \int \mathbf{u}' \frac{\partial^2 g_l(\mathbf{x})}{\partial \mathbf{x} \partial \mathbf{x}'} \mathbf{u} K(\mathbf{u}) d\mathbf{u} \right)^2 d\mathbf{x} \\
& + o(n^{-\frac{4}{4+p}}) \\
& = \frac{\left( \int \left( \int \mathbf{u}' \frac{\partial^2 g_l(\mathbf{x})}{\partial \mathbf{x} \partial \mathbf{x}'} \mathbf{u} K(\mathbf{u}) d\mathbf{u} \right)^2 d\mathbf{x} \right)^{\frac{p}{4+p}} \left( \int \frac{w_l(\mathbf{x})g_l(\mathbf{x})}{\sum_{k=1}^m \zeta_k w_k(\mathbf{x})} d\mathbf{x} \int K^2(\mathbf{u})d\mathbf{u} \right)^{\frac{4}{4+p}}}{(n^4 p^p)^{\frac{1}{4+p}}} \\
& + \frac{1}{4} \left( \frac{p}{n} \int \frac{w_l(\mathbf{x})g_l(\mathbf{x})}{\sum_{k=1}^m \zeta_k w_k(\mathbf{x})} d\mathbf{x} \int K^2(\mathbf{u})d\mathbf{u} \right)^{\frac{4}{4+p}} \left( \int \left( \int \mathbf{u}' \frac{\partial^2 g_l(\mathbf{x})}{\partial \mathbf{x} \partial \mathbf{x}'} \mathbf{u} K(\mathbf{u}) d\mathbf{u} \right)^2 d\mathbf{x} \right)^{\frac{p}{4+p}} \\
& + o(n^{-\frac{4}{4+p}})
\end{aligned}$$

$$\begin{aligned}
&= n^{-\frac{4}{4+p}} \left( p^{-\frac{p}{4+p}} + \frac{1}{4} p^{\frac{4}{4+p}} \right) \left( \int \frac{w_l(\mathbf{x})g_l(\mathbf{x})}{\sum_{k=1}^m \zeta_k w_k(\mathbf{x})} d\mathbf{x} \int K^2(\mathbf{u}) d\mathbf{u} \right)^{\frac{4}{4+p}} \\
&\quad \cdot \left( \int \left( \int \mathbf{u}' \frac{\partial^2 g_l(\mathbf{x})}{\partial \mathbf{x} \partial \mathbf{x}'} \mathbf{u} K(\mathbf{u}) d\mathbf{u} \right)^2 d\mathbf{x} \right)^{\frac{p}{4+p}} + o(n^{-\frac{4}{4+p}})
\end{aligned}$$

□

**Theorem 3.3.** *If  $\hat{f}(\mathbf{x}) = \frac{1}{n_l h_n^p} \sum_{i=1}^{n_l} K\left(\frac{\mathbf{x}-\mathbf{x}_i}{h_n}\right)$  is the classic multivariate kernel density estimator of  $g_l$  and AMISE is the asymptotic mean integrated square error, then*

(a) *As  $n \rightarrow \infty$ ,  $h_n \rightarrow 0$  and  $nh_n^p \rightarrow \infty$*

$$AMISE(\hat{g}_l) \leq AMISE(\hat{f})$$

(b) *Using optimal bandwidths, the proposed semiparametric density estimator  $\hat{g}_l(\mathbf{x})$  is more efficient than  $\hat{f}(\mathbf{x})$ , i.e for every  $l$*

$$eff(\hat{f}, \hat{g}_l) \equiv \frac{AMISE^*(\hat{g}_l)}{AMISE^*(\hat{f})} \leq 1$$

where  $AMISE^*$  is the optimal AMISE

*Proof.* (a) According to [10], if  $\hat{f}(\mathbf{x}) = \frac{1}{n_l h_n^p} \sum_{i=1}^{n_l} K\left(\frac{\mathbf{x}-\mathbf{x}_i}{h_n}\right)$  is the classic multivariate kernel density estimator of  $g_l$ , then, as  $n \rightarrow \infty$ ,  $h_n \rightarrow 0$  and  $nh_n^p \rightarrow \infty$ :

$$AMISE(\hat{f}) = \frac{1}{4} h_n^4 \int \left( \int \mathbf{u}' \frac{\partial^2 g_l(\mathbf{x})}{\partial \mathbf{x} \partial \mathbf{x}'} \mathbf{u} K(\mathbf{u}) d\mathbf{u} \right)^2 d\mathbf{x} + \frac{1}{n_l h_n^p} \int K^2(\mathbf{x}) d\mathbf{x}$$

In Theorem 3.2 we proved that:

$$AMISE(\hat{g}_l) = \frac{1}{4}h_n^4 \int \left( \int \mathbf{u}' \frac{\partial^2 g_l(\mathbf{x})}{\partial \mathbf{x} \partial \mathbf{x}'} \mathbf{u} K(\mathbf{u}) d\mathbf{u} \right)^2 d\mathbf{x} \\ + \frac{1}{n_l h_n^p} \int \frac{\zeta_l w_l(\mathbf{x}) g_l(\mathbf{x})}{\sum_{k=1}^m \zeta_k w_k(\mathbf{x})} d\mathbf{x} \int K^2(\mathbf{x}) d\mathbf{x}$$

Since for every  $l$ ,  $\int \frac{\zeta_l w_l(\mathbf{x}) g_l(\mathbf{x})}{\sum_{k=1}^m \zeta_k w_k(\mathbf{x})} d\mathbf{x} \leq 1$ , it follows that

$$AMISE(\hat{g}_l) \leq AMISE(\hat{f}).$$

(b)  $AMISE(\hat{f})$  is optimized for  $h_n = \left[ \frac{p}{n_l} \frac{\int K^2(\mathbf{x}) d\mathbf{x}}{\int \left( \int \mathbf{u}' \frac{\partial^2 g_l(\mathbf{x})}{\partial \mathbf{x} \partial \mathbf{x}'} \mathbf{u} K(\mathbf{u}) d\mathbf{u} \right)^2 d\mathbf{x}} \right]^{\frac{1}{p+4}}$ . Indeed, if we differentiate  $AMISE(\hat{f})$  and set equal to 0:

$$h_n^3 \int \left( \int \mathbf{u}' \frac{\partial^2 g_l(\mathbf{x})}{\partial \mathbf{x} \partial \mathbf{x}'} \mathbf{u} K(\mathbf{u}) d\mathbf{u} \right)^2 d\mathbf{x} = \frac{p}{n_l h_n^{p+1}} \int K^2(\mathbf{x}) d\mathbf{x} \\ \Rightarrow h_n^{p+4} = \frac{p}{n_l} \frac{\int K^2(\mathbf{x}) d\mathbf{x}}{\int \left( \int \mathbf{u}' \frac{\partial^2 g_l(\mathbf{x})}{\partial \mathbf{x} \partial \mathbf{x}'} \mathbf{u} K(\mathbf{u}) d\mathbf{u} \right)^2 d\mathbf{x}}.$$

The optimal  $AMISE^*(\hat{f})$  is therefore,

$$AMISE^*(\hat{f}) = \frac{1}{4} \left[ \frac{p}{n_l} \frac{\int K^2(\mathbf{x}) d\mathbf{x}}{\int \left( \int \mathbf{u}' \frac{\partial^2 g_l(\mathbf{x})}{\partial \mathbf{x} \partial \mathbf{x}'} \mathbf{u} K(\mathbf{u}) d\mathbf{u} \right)^2 d\mathbf{x}} \right]^{\frac{4}{p+4}} \int \left( \int \mathbf{u}' \frac{\partial^2 g_l(\mathbf{x})}{\partial \mathbf{x} \partial \mathbf{x}'} \mathbf{u} K(\mathbf{u}) d\mathbf{u} \right)^2 d\mathbf{x} \\ + \frac{1}{n_l} \left[ \frac{n_l \int \left( \int \mathbf{u}' \frac{\partial^2 g_l(\mathbf{x})}{\partial \mathbf{x} \partial \mathbf{x}'} \mathbf{u} K(\mathbf{u}) d\mathbf{u} \right)^2 d\mathbf{x}}{p \int K^2(\mathbf{x}) d\mathbf{x}} \right]^{\frac{p}{p+4}} \int K^2(\mathbf{x}) d\mathbf{x} \\ = \frac{1}{4} \left[ \frac{p}{n_l} \int K^2(\mathbf{x}) d\mathbf{x} \right]^{\frac{4}{p+4}} \left[ \int \left( \int \mathbf{u}' \frac{\partial^2 g_l(\mathbf{x})}{\partial \mathbf{x} \partial \mathbf{x}'} \mathbf{u} K(\mathbf{u}) d\mathbf{u} \right)^2 d\mathbf{x} \right]^{\frac{p}{p+4}} \\ + \left[ \frac{1}{n_l} \int K^2(\mathbf{x}) d\mathbf{x} \right]^{\frac{4}{p+4}} \left[ \frac{1}{p} \int \left( \int \mathbf{u}' \frac{\partial^2 g_l(\mathbf{x})}{\partial \mathbf{x} \partial \mathbf{x}'} \mathbf{u} K(\mathbf{u}) d\mathbf{u} \right)^2 d\mathbf{x} \right]^{\frac{p}{p+4}}$$

$$\begin{aligned}
&= \left[ (p^p n_l^4)^{-\frac{1}{p+4}} + \frac{1}{4} (p n_l^{-1})^{\frac{4}{p+4}} \right] \left[ \int \left( \int \mathbf{u}' \frac{\partial^2 g_l(\mathbf{x})}{\partial \mathbf{x} \partial \mathbf{x}'} \mathbf{u} K(\mathbf{u}) d\mathbf{u} \right)^2 d\mathbf{x} \right]^{\frac{p}{p+4}} \\
&\quad \cdot \left( \int K^2(\mathbf{x}) d\mathbf{x} \right)^{\frac{4}{p+4}}.
\end{aligned}$$

Therefore the asymptotic relative efficiency of  $\hat{f}$  with respect to  $\hat{g}_l$  is given by:

$$\begin{aligned}
eff(\hat{f}, \hat{g}_l) &\equiv \frac{AMISE^*(\hat{g}_l)}{AMISE^*(\hat{f})} \\
&= \frac{n^{-\frac{4}{4+p}} \left[ p^{-\frac{p}{4+p}} + \frac{1}{4} p^{\frac{4}{4+p}} \right] \left[ \int \frac{w_l(\mathbf{x}) g_l(\mathbf{x})}{\sum_{k=1}^m \zeta_k w_k(\mathbf{x})} d\mathbf{x} \right]^{\frac{4}{4+p}}}{n_l^{-\frac{4}{4+p}} \left[ p^{-\frac{p}{4+p}} + \frac{1}{4} p^{\frac{4}{4+p}} \right]} \\
&= \left[ \int \frac{\zeta_l w_l(\mathbf{x}) g_l(\mathbf{x})}{\sum_{k=1}^m \zeta_k w_k(\mathbf{x})} d\mathbf{x} \right]^{\frac{4}{4+p}} \leq 1 \text{ for every } l.
\end{aligned}$$

Thus, unless only the  $l$ th sample is available, the proposed semiparametric density estimator is more efficient than the traditional kernel density estimator.

□

### 3.3.3 Bandwidth selection for $\hat{g}_l$

In section 3.3.1 it was shown that, as is the case with the traditional single sample estimator, the pooled estimator  $\hat{g}_l$  also suffers from a similar bias-variance trade-off problem where a smaller  $h_n$  reduces the bias at the expense of the variance, whereas a larger  $h_n$  increases the bias but reduces the variance.

From equation (3.9), we have a formula for the asymptotically optimal bandwidth  $h_n^*$ . In practice though, it is difficult to use it since  $g_l$  is not known. In the one dimensional case Silverman [69] proposes to either use the normal density  $N(\mu, \Sigma)$ ,

where  $\mu$  and  $\Sigma$  are estimated from the data, or  $\hat{f}$  to approximate  $g_l$ . Following Silverman [69], Fokianos [19] and Qin and Zhang [62] both use  $\hat{g}_l$  to approximate  $g_l$ . However things become more complicated in the multidimensional setting. The computational burden is heavier and, as Silverman [69] remarks, it is somewhat hazardous to estimate  $\partial^2 g_l(\mathbf{x})/\partial \mathbf{x} \partial \mathbf{x}'$  by  $\partial^2 \hat{g}_l(\mathbf{x})/\partial \mathbf{x} \partial \mathbf{x}'$  unless very large samples are available.

Another way to select the bandwidth is using *cross validation*. Cross validation minimizes with respect to  $h_n$  an estimate for the integrated squared error (ISE):

$$ISE(h_n) = \int (\hat{g}_l(\mathbf{x}) - g_l(\mathbf{x}))^2 d\mathbf{x} = \int \hat{g}_l^2(\mathbf{x}) d\mathbf{x} - 2 \int \hat{g}_l(\mathbf{x}) g_l(\mathbf{x}) d\mathbf{x} + \int g_l^2(\mathbf{x}) d\mathbf{x}$$

The last term does not depend on  $h_n$ , so we may drop it in the minimization of ISE. To minimize ISE we need to rewrite the first and second term as a function of  $h_n$  and the data. Denote by  $\mathbf{t} = [\mathbf{x}'_{11}, \dots, \mathbf{x}'_{1n_1}, \dots, \mathbf{x}'_{m1}, \dots, \mathbf{x}'_{mn_m}]'_{n \times 1} = (\mathbf{t}'_1, \dots, \mathbf{t}'_n)'$  the combined data. So  $\mathbf{t}$  has  $n$  rows. The first term can be written:

$$\begin{aligned} \int \hat{g}_l^2(\mathbf{x}) d\mathbf{x} &= \int \left[ \frac{1}{h_n^p} \sum_{i=1}^m \sum_{j=1}^{n_i} \hat{p}_{ij} \hat{w}_l(\mathbf{x}_{ij}) K \left( \frac{\mathbf{x} - \mathbf{x}_{ij}}{h_n} \right) \right]^2 d\mathbf{x} \\ &= \frac{1}{h_n^{2p}} \int \sum_{i=1}^m \sum_{j=1}^{n_i} \sum_{i'=1}^m \sum_{j'=1}^{n_{i'}} \hat{p}_{ij} \hat{w}_l(\mathbf{x}_{ij}) K \left( \frac{\mathbf{x} - \mathbf{x}_{ij}}{h_n} \right) \hat{p}_{i'j'} \hat{w}_l(\mathbf{x}_{i'j'}) K \left( \frac{\mathbf{x} - \mathbf{x}_{i'j'}}{h_n} \right) d\mathbf{x} \\ &= \frac{1}{h_n^{2p}} \sum_{i=1}^m \sum_{j=1}^{n_i} \sum_{i'=1}^m \sum_{j'=1}^{n_{i'}} \int \hat{p}_{ij} \hat{w}_l(\mathbf{x}_{ij}) K \left( \frac{\mathbf{x} - \mathbf{x}_{ij}}{h_n} \right) \hat{p}_{i'j'} \hat{w}_l(\mathbf{x}_{i'j'}) K \left( \frac{\mathbf{x} - \mathbf{x}_{i'j'}}{h_n} \right) d\mathbf{x} \\ &\stackrel{\mathbf{z} = \frac{\mathbf{x} - \mathbf{x}_{ij}}{h_n}}{=} h_n^{-p} \sum_{i=1}^m \sum_{j=1}^{n_i} \sum_{i'=1}^m \sum_{j'=1}^{n_{i'}} \int \hat{p}_{ij} \hat{w}_l(\mathbf{x}_{ij}) \hat{p}_{i'j'} \hat{w}_l(\mathbf{x}_{i'j'}) K(\mathbf{z}) K \left( \mathbf{z} + \frac{\mathbf{x}_{ij} - \mathbf{x}_{i'j'}}{h_n} \right) d\mathbf{z} \\ &= h_n^{-p} \sum_{i=1}^m \sum_{j=1}^{n_i} \sum_{i'=1}^m \sum_{j'=1}^{n_{i'}} \hat{p}_{ij} \hat{w}_l(\mathbf{x}_{ij}) \hat{p}_{i'j'} \hat{w}_l(\mathbf{x}_{i'j'}) \int K(\mathbf{z}) K \left( \mathbf{z} + \frac{\mathbf{x}_{ij} - \mathbf{x}_{i'j'}}{h_n} \right) d\mathbf{z} \end{aligned}$$



$$= h_n^{-p} \sum_{i=1}^n \sum_{i'=1}^n \hat{p}(\mathbf{t}_i) \hat{w}_l(\mathbf{t}_i) \hat{p}(\mathbf{t}_{i'}) \hat{w}_l(\mathbf{t}_{i'}) \int K(\mathbf{z}) K\left(\mathbf{z} + \frac{\mathbf{t}_i - \mathbf{t}_{i'}}{h_n}\right) d\mathbf{z}$$

For the second term notice that  $\int \hat{g}_l(\mathbf{x}) g_l(\mathbf{x}) d\mathbf{x} = E\hat{g}_l(\mathbf{x})$ . Following Silverman [69] and Cheng and Chu [11] we can estimate  $E\hat{g}_l(\mathbf{x})$  using the leave one out estimator:

$$\widehat{E\hat{g}_l(\mathbf{x})} = \frac{1}{n_l} \sum_{i=n_1+\dots+n_{l-1}+1}^{n_l} \hat{g}_{l,i}(\mathbf{t}_i)$$

where  $\hat{g}_{l,i}(\mathbf{t}_i)$  is  $\hat{g}_l(\mathbf{t}_i)$  with  $\mathbf{t}_i$  dropped from the combined data. Therefore, in order to find a value for the bandwidth  $h_n$ , it suffices to minimize

$$h_n^{-p} \sum_{i=1}^n \sum_{i'=1}^n \hat{p}(\mathbf{t}_i) \hat{w}_l(\mathbf{t}_i) \hat{p}(\mathbf{t}_{i'}) \hat{w}_l(\mathbf{t}_{i'}) \int K(\mathbf{z}) K\left(\mathbf{z} + \frac{\mathbf{t}_i - \mathbf{t}_{i'}}{h_n}\right) d\mathbf{z} - \frac{2}{n_l} \sum_{i=n_1+\dots+n_{l-1}+1}^{n_l} \hat{g}_{l,i}(\mathbf{t}_i) \quad (3.10)$$

Equation (3.10) can have many local minima so it is better to use grid methods rather than Newton-Raphson methods for the minimization. The above procedure should be used for each  $l$ ,  $l = 1, \dots, m$  to determine the optimal bandwidth. For the special case of the reference distribution  $\hat{g}_m$ , where by assumption  $\hat{w}_m \equiv 1$ , we choose  $h_n$  by minimizing:

$$h_n^{-p} \sum_{i=1}^n \sum_{i'=1}^n \hat{p}(\mathbf{t}_i) \hat{p}(\mathbf{t}_{i'}) \int K(\mathbf{z}) K\left(\mathbf{z} + \frac{\mathbf{t}_i - \mathbf{t}_{i'}}{h_n}\right) d\mathbf{z} - \frac{2}{n_m} \sum_{i=n_1+\dots+n_q+1}^{n_m} \hat{g}_{m,i}(\mathbf{t}_i) \quad (3.11)$$

In general, cross validation using the leave one out estimator is computationally inefficient. However, for sufficiently large samples and  $l = 1, \dots, q, m$  notice the

following heuristic argument:

$$\begin{aligned}\int \hat{g}_l(\mathbf{x})g_l(\mathbf{x})d\mathbf{x} &= \int (\tilde{g}_l(\mathbf{x}) + O(n^{-1/2}))g_l(\mathbf{x})d\mathbf{x} \\ &= \int \tilde{g}_l(\mathbf{x})g_l(\mathbf{x})d\mathbf{x} + \int O(n^{-1/2})g_l(\mathbf{x})d\mathbf{x} \xrightarrow{n \rightarrow \infty} \int \tilde{g}_l(\mathbf{x})g_l(\mathbf{x})d\mathbf{x}\end{aligned}$$

where  $\int \tilde{g}_l(\mathbf{x})g_l(\mathbf{x})d\mathbf{x} = \int \frac{1}{nh_n^p} \sum_{i=1}^m \sum_{j=1}^{n_i} \frac{w_l(\mathbf{x}_{ij})}{\sum_{k=1}^m \zeta_k w_k(\mathbf{x}_{ij})} K\left(\frac{\mathbf{x}-\mathbf{x}_{ij}}{h_n}\right) g_l(\mathbf{x})d\mathbf{x}$ . Moreover:

$$\begin{aligned}E\left[\int \tilde{g}_l(\mathbf{x})g_l(\mathbf{x})d\mathbf{x}\right] &= E\left[\int \frac{1}{nh_n^p} \sum_{i=1}^m \sum_{j=1}^{n_i} \frac{w_l(\mathbf{x}_{ij})}{\sum_{k=1}^m \zeta_k w_k(\mathbf{x}_{ij})} K\left(\frac{\mathbf{x}-\mathbf{x}_{ij}}{h_n}\right) g_l(\mathbf{x})d\mathbf{x}\right] \\ &= \frac{1}{nh_n^p} \sum_{i=1}^m \int \int \frac{n_i w_l(\mathbf{y})}{\sum_{k=1}^m \zeta_k w_k(\mathbf{y})} K\left(\frac{\mathbf{x}-\mathbf{y}}{h_n}\right) g_l(\mathbf{x})w_i(\mathbf{y})g(\mathbf{y})d\mathbf{x}d\mathbf{y} \\ &= h_n^{-p} \int \int \frac{\sum_{i=1}^m \zeta_i w_i(\mathbf{y})}{\sum_{k=1}^m \zeta_k w_k(\mathbf{y})} K\left(\frac{\mathbf{x}-\mathbf{y}}{h_n}\right) g_l(\mathbf{x})g_l(\mathbf{y})d\mathbf{x}d\mathbf{y} \\ &= h_n^{-p} \int \int K\left(\frac{\mathbf{x}-\mathbf{y}}{h_n}\right) g_l(\mathbf{x})g_l(\mathbf{y})d\mathbf{x}d\mathbf{y} \\ &= E\left[h_n^{-p} K\left(\frac{\mathbf{x}-\mathbf{y}}{h_n}\right)\right] \\ &= E\left[\frac{1}{n_l(n_l-1)h_n^p} \sum_{i \neq j} K\left(\frac{\mathbf{x}_{li}-\mathbf{x}_{lj}}{h_n}\right)\right]\end{aligned}$$

Thus, for sufficient large  $n$ , an unbiased estimator for  $\int \tilde{g}_l(\mathbf{x})g_l(\mathbf{x})d\mathbf{x}$  is

$$\frac{1}{n_l(n_l-1)h_n^p} \sum_{i \neq j} K\left(\frac{\mathbf{x}_{li}-\mathbf{x}_{lj}}{h_n}\right).$$

Therefore, an alternative way to find  $h_n$  is by minimizing

$$\begin{aligned}h_n^{-p} \sum_{i=1}^n \sum_{i'=1}^n \hat{p}(\mathbf{t}_i)\hat{w}_l(\mathbf{t}_i)\hat{p}(\mathbf{t}_{i'})\hat{w}_l(\mathbf{t}_{i'}) \int K(\mathbf{z})K\left(\mathbf{z} + \frac{\mathbf{t}_i-\mathbf{t}_{i'}}{h_n}\right) d\mathbf{z} \\ - \frac{2}{n_l(n_l-1)h_n^p} \sum_{i \neq j} K\left(\frac{\mathbf{x}_{li}-\mathbf{x}_{lj}}{h_n}\right) \quad (3.12)\end{aligned}$$

Cross validation has the advantage that equations (3.10), (3.11) and (3.12) can easily be modified if we wish to use different bandwidths  $h_1, \dots, h_p$  to smooth each variable. All the results presented in this Chapter still hold in this case.

## Chapter 4

### Semiparametric regression

#### 4.1 Introduction

In this Chapter we discuss a novel approach to regression analysis with random covariates from a semiparametric perspective where information is combined from multiple multivariate sources. The approach assumes a semiparametric density ratio model where multivariate distributions are “regressed” on a reference distribution. Each multivariate distribution and a corresponding conditional expectation/regression of interest is then estimated from the combined data from all sources. An advantage of the method is that we avoid making any explicit distributional assumptions and that all quantities are estimated from the combined data. Graphical and quantitative diagnostic tools are suggested to assess model validity. Comparisons are made with multiple regression, generalized additive models (GAM) and nonparametric kernel regression. Some of the results of this Chapter were first discussed for the two-dimensional case in [37].

This Chapter is organized as follows: In Sections 4.2 and 4.3 we introduce the model we are considering, which is a special case of the general model (2.1), provide the score equations for the parameters and discuss hypothesis testing. Section 4.4 discusses the estimation of the conditional expectation based on the semiparametric model, whereas Section 4.5 gives a literature review for other ways of estimating the

conditional expectation. In Section 4.6 we propose a coefficient of determination  $R^2$  to assess the goodness of fit of the semiparametric model. Finally in Section 4.7 we conduct a simulation study where we apply the results of Chapters 3 and 4.

## 4.2 Statistical formulation

Suppose we have  $m = q + 1$  data sets or samples of  $p$ -dimensional vectors, where each vector consists of  $p - 1$  covariates and one response, and assume that the  $i$ th sample size is  $n_i$ . Thus, for  $i = 1, \dots, q, m$ ,  $j = 1, \dots, n_i$  we have

$$(x_{ij1}, x_{ij2}, \dots, x_{ij(p-1)}, y_{ij}) \sim g_i(x_1, \dots, x_{(p-1)}, y).$$

We choose  $g \equiv g_m(x_1, \dots, x_{(p-1)}, y)$  as a reference or baseline probability density function (pdf), and let each  $g_i(x_1, \dots, x_{(p-1)}, y)$ ,  $i = 1, \dots, q$  be an exponential distortion or tilt of the reference distribution,

$$\frac{g_i(\mathbf{x})}{g(\mathbf{x})} = \exp(\alpha_i + \boldsymbol{\beta}'_i \mathbf{x}), \quad i = 1, \dots, q \quad (4.1)$$

where  $\mathbf{x} = (x_1, \dots, x_{(p-1)}, y)'$  and  $\boldsymbol{\beta}_i = (\beta_{i1}, \dots, \beta_{ip})'$ . Since the  $g_i(\mathbf{x})$ ,  $i = 1, \dots, q, m$  are probability densities,  $\boldsymbol{\beta}_i = \mathbf{0}$  implies  $\alpha_i = 0$ ,  $j = 1, \dots, q$ . It follows that the hypothesis  $H_0 : \boldsymbol{\beta}_1 = \dots = \boldsymbol{\beta}_q = \mathbf{0}$  implies equidistribution: all the  $g_i$  are equal.

**Remark 4.1.** *Model (4.1) is a special case of model (2.1) with  $w(\mathbf{x}, \boldsymbol{\theta}_i) = w(\mathbf{x}, \alpha_i, \boldsymbol{\beta}_i) \equiv \exp(\alpha_i + \boldsymbol{\beta}'_i \mathbf{x})$ .*

**Example 4.1.** *Two-dimensional normal distributions.* Suppose we have  $m = q + 1$

two-dimensional data sets,

$$(x_{j1}, y_{j1}), (x_{j2}, y_{j2}), \dots, (x_{jn_j}, y_{jn_j}) \sim g_j(x, y), \quad j = 1, \dots, q, m$$

where  $g_j(x, y)$  is the probability density of  $N(\boldsymbol{\mu}_j, \boldsymbol{\Sigma})$ , with

$$\boldsymbol{\mu}_j = \begin{pmatrix} \mu_{jx} \\ \mu_{jy} \end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{xx} & \sigma_{xy} \\ \sigma_{xy} & \sigma_{yy} \end{pmatrix}, \quad j = 1, \dots, m.$$

Then, choosing  $g_m(x, y)$  as a reference density we have

$$\frac{g_j(x, y)}{g_m(x, y)} = \exp[(\boldsymbol{\mu}_j - \boldsymbol{\mu}_m)' \boldsymbol{\Sigma}^{-1} \boldsymbol{x} - \frac{1}{2}(\boldsymbol{\mu}_j' \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_j - \boldsymbol{\mu}_m' \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_m)], \quad (4.2)$$

where  $\boldsymbol{x} = (x, y)'$ . Notice that (4.2) is a special case of model (4.1) where

$$\alpha_j = -\frac{1}{2}(\boldsymbol{\mu}_j' \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_j - \boldsymbol{\mu}_m' \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_m)$$

$$\boldsymbol{\beta}_j = \begin{pmatrix} \beta_{j1} \\ \beta_{j2} \end{pmatrix} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_j - \boldsymbol{\mu}_m)$$

□

To estimate the parameters and the reference density  $g$ , or equivalently the reference distribution function  $G$ , we follow the same procedure described in Section

2.2. First the data are combined in a single vector  $\mathbf{t}$  of length  $n = n_1 + n_2 + \dots + n_m$ ,

$$\begin{aligned}\mathbf{t} &= ((x_{ij1}, x_{ij2}, \dots, x_{ij(p-1)}, y_{ij}) : i = 1, \dots, q, m, j = 1, \dots, n_i)' \\ &= (\mathbf{t}'_1, \mathbf{t}'_2, \dots, \mathbf{t}'_n)'\end{aligned}\quad (4.3)$$

where  $\mathbf{t}_i \equiv (t_{ix_1}, \dots, t_{ix_{p-1}}, t_{iy})'$ . The idea is to approximate the reference distribution function by a step function  $G$  with jumps  $p_i$  at all the observed points ([72], [73]). For the two dimensional case the  $p_i$ 's can be defined as:

$$dG(\mathbf{t}_i) = p_i = G(t_{ix}, t_{iy}) - G(t_{i-1,x_1}, t_{iy}) - G(t_{ix}, t_{i-1,y}) + G(t_{i-1,x}, t_{i-1,y}), \quad i = 1, \dots, n.$$

whereas for the three dimensional case:

$$\begin{aligned}dG(\mathbf{t}_i) = p_i &= G(t_{ix_1}, t_{ix_2}, t_{iy}) - G(t_{i-1,x_1}, t_{ix_2}, t_{iy}) - G(t_{ix_1}, t_{i-1,x_2}, t_{iy}) \\ &\quad - G(t_{ix_1}, t_{ix_2}, t_{i-1,y}) + G(t_{i-1,x_1}, t_{i-1,x_2}, t_{iy}) + G(t_{i-1,x_1}, t_{ix_2}, t_{i-1,y}) \\ &\quad + G(t_{ix_1}, t_{i-1,x_2}, t_{i-1,y}) - G(t_{i-1,x_1}, t_{i-1,x_2}, t_{i-1,y}), \quad i = 1, \dots, n.\end{aligned}$$

Generally, the  $p_i$  are the jumps in the  $p$ -dimensional step function  $G$  at  $\mathbf{t}_1, \dots, \mathbf{t}_n$ .

The empirical likelihood is a function of  $p_i$ ,  $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_q)'$  and  $\boldsymbol{\beta} = (\boldsymbol{\beta}'_1, \dots, \boldsymbol{\beta}'_q)'$ :

$$\begin{aligned}L(\boldsymbol{\alpha}, \boldsymbol{\beta}, G) &= \prod_{i=1}^n p_i \prod_{k=1}^{n_1} \exp(\alpha_1 + \beta_{11}x_{1k1} + \dots + \beta_{1(p-1)}x_{1k(p-1)} + \beta_{1p}y_{1k}) \\ &\quad \dots \prod_{k=1}^{n_q} \exp(\alpha_q + \beta_{q1}x_{qk1} + \dots + \beta_{q(p-1)}x_{qk(p-1)} + \beta_{qp}y_{qk})\end{aligned}\quad (4.4)$$

subject to the constraints

$$\sum_{i=1}^n p_i = 1, \quad \sum_{i=1}^n w_1(\mathbf{t}_i)p_i = 1, \dots, \quad \sum_{i=1}^n w_q(\mathbf{t}_i)p_i = 1 \quad (4.5)$$

where  $w_j(\mathbf{t}_i) = \exp(\alpha_j + \boldsymbol{\beta}'_j \mathbf{t}_i)$ ,  $j = 1, \dots, q$ .

Estimates for  $\hat{\alpha}_j$  and  $\hat{\boldsymbol{\beta}}_j$  are obtained by solving the score equations:

$$\frac{\partial l}{\partial \alpha_j} = - \sum_{i=1}^n \frac{\rho_j w_j(\mathbf{t}_i)}{1 + \rho_1 w_1(\mathbf{t}_i) + \dots + \rho_q w_q(\mathbf{t}_i)} + n_j = 0 \quad (4.6)$$

$$\frac{\partial l}{\partial \boldsymbol{\beta}_j} = - \sum_{i=1}^n \frac{\rho_j w_j(\mathbf{t}_i) \mathbf{t}_i}{1 + \rho_1 w_1(\mathbf{t}_i) + \dots + \rho_q w_q(\mathbf{t}_i)} + \sum_{i=1}^{n_j} (x_{ji1}, \dots, y_{ji})' = 0 \quad (4.7)$$

for  $j = 1, \dots, q$  and  $\rho_j = n_j/n_m$ . Then

$$\hat{p}_i = \frac{1}{n_m} \cdot \frac{1}{1 + \rho_1 \hat{w}_1(\mathbf{t}_i) + \dots + \rho_q \hat{w}_q(\mathbf{t}_i)} \quad (4.8)$$

$$\hat{G}(\mathbf{t}) = \frac{1}{n_m} \cdot \sum_{i=1}^n \frac{I(\mathbf{t}_i \leq \mathbf{t})}{1 + \rho_1 \hat{w}_1(\mathbf{t}_i) + \dots + \rho_q \hat{w}_q(\mathbf{t}_i)} \quad (4.9)$$

where  $(\mathbf{t}_i \leq \mathbf{t})$  is defined componentwise,  $\hat{w}_j(\mathbf{t}_i) = \exp(\hat{\alpha}_j + \hat{\boldsymbol{\beta}}'_j \mathbf{t}_i)$ , and  $I(B)$  is the indicator of the event  $B$ .

**Theorem 4.1.** *As  $n \rightarrow \infty$ , the estimators  $\hat{\boldsymbol{\theta}} = (\hat{\alpha}_1, \dots, \hat{\alpha}_q, \hat{\boldsymbol{\beta}}_1, \dots, \hat{\boldsymbol{\beta}}_q)'$  are asymptotically normal*

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{D} N(\mathbf{0}, \boldsymbol{\Sigma}), \quad (4.10)$$

where  $\boldsymbol{\theta}_0$  denotes the true parameters and  $\boldsymbol{\Sigma} = \mathbf{S}^{-1} \mathbf{V} \mathbf{S}^{-1}$  is defined in the appendix.

*Proof.* For a detailed proof see Lu [42]. □



**Remark 4.2.**  $G_j(\mathbf{t}_i)$  and  $g_j(\mathbf{t}_i)$ ,  $j = 1, \dots, q$  can be estimated as exponential tilts of  $\hat{G}$ ,  $\hat{p}_i$  as in (2.19).

### 4.3 Hypothesis testing

There are several ways to test the hypothesis  $H_0 : \beta_1 = \beta_2 = \dots = \beta_q = \mathbf{0}$  for the model (4.1). One way is to use the likelihood ratio test:

$$\begin{aligned}
LR &\equiv -2[l(0, 0) - l(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})] \\
&= -2 \sum_{i=1}^n \log[1 + \rho_1 \hat{w}_1(\mathbf{t}_i) + \dots + \rho_q \hat{w}_q(\mathbf{t}_i)] \\
&\quad + 2 \sum_{i=1}^q \sum_{j=1}^{n_i} [\alpha_i + \beta_{i1} x_{ik1} + \dots + \beta_{i(p-1)} x_{ik(p-1)} + \beta_{ip} y_{ik}] \\
&\quad + 2n \log[1 + \sum_{i=1}^q \rho_i] \tag{4.11}
\end{aligned}$$

Under  $H_0$ , the likelihood ratio is asymptotically approximately distributed as  $\chi^2$  with  $qp$  degrees of freedom, and  $H_0$  is rejected for large values. Power considerations of (4.11) have been studied in [35] and [77]. Another test that can be used is based on the  $\mathcal{X}_1$  statistic. For more details see [35].

#### 4.4 Computing $E[y|\mathbf{x}]$ using the density ratio model

Under the  $p$ -dimensional density ratio model we can predict the response  $y$  given the covariate information  $x_1, x_2, \dots, x_{(p-1)}$  for any of the  $m$  data sets as follows:

$$\hat{E}_j(y | x_1, \dots, x_{(p-1)}) = \sum_i^{n_j} y_i \frac{\hat{g}_j(x_1, \dots, x_{(p-1)}, y_i)}{\sum_{y_i} \hat{g}_j(x_1, \dots, x_{(p-1)}, y_i)}, \quad j = 1, \dots, q, m. \quad (4.12)$$

The  $\hat{g}_j$  in (4.12) are the semiparametric kernel density estimates described in Section 3.3,

$$\hat{g}_j(\mathbf{z}_0) = \frac{1}{h^p} \sum_{i=1}^n \hat{p}_i \hat{w}_j(\mathbf{t}_i) K((\mathbf{t}_i - \mathbf{z}_0)/h), \quad j = 1, \dots, m. \quad (4.13)$$

where  $\mathbf{z}_0$  is  $p$ -dimensional.

**Theorem 4.2.** *Assume that the data are bounded. Then:*

(a) *As  $n \rightarrow \infty, h \rightarrow 0$  and  $nh^p \rightarrow \infty$ ,*

$$\int |\hat{E}(y|\mathbf{x}) - E(y|\mathbf{x})| g(\mathbf{x}) d\mathbf{x} \rightarrow 0$$

*in the mean square sense.*

(b) *If, in addition  $0 < A < g(\mathbf{x})$ , then*

$$\int |\hat{E}(y|\mathbf{x}) - E(y|\mathbf{x})| d\mathbf{x} \rightarrow 0$$

*in the mean square sense.*

*Proof.* (a) Let  $\mathbf{x}$  be a vector of size  $k = p-1$  of bounded covariates, and  $y$  a bounded response. We wish to prove  $\hat{E}(y|\mathbf{x}) \xrightarrow{p} E(y|\mathbf{x})$ . We have:

$$\hat{E}(y|\mathbf{x}) - E(y|\mathbf{x}) = \frac{\frac{1}{n} \sum_{i=1}^n y_i \hat{g}(\mathbf{x}, y_i)}{\frac{1}{n} \sum_{i=1}^n \hat{g}(\mathbf{x}, y_i)} - \frac{\int y g(\mathbf{x}, y) dy}{g(\mathbf{x})}$$

Or for sufficiently large  $n$

$$\hat{E}(y|\mathbf{x}) - E(y|\mathbf{x}) = \frac{\int y \hat{g}(\mathbf{x}, y) dy}{\hat{g}(\mathbf{x})} - \frac{\int y g(\mathbf{x}, y) dy}{g(\mathbf{x})} \sim \frac{\int y [\hat{g}(\mathbf{x}, y) - g(\mathbf{x}, y)] dy}{g(\mathbf{x})}$$

Thus by Cauchy-Schwarz,

$$\begin{aligned} \left[ \int |\hat{E}(y|\mathbf{x}) - E(y|\mathbf{x})| g(\mathbf{x}) d\mathbf{x} \right]^2 &\sim \left[ \int \left| \int y [\hat{g}(\mathbf{x}, y) - g(\mathbf{x}, y)] dy \right| d\mathbf{x} \right]^2 \\ &\leq \left[ \int \int |y| |[\hat{g}(\mathbf{x}, y) - g(\mathbf{x}, y)]| dy d\mathbf{x} \right]^2 \\ &\leq \int \int |y|^2 dy d\mathbf{x} \int \int |[\hat{g}(\mathbf{x}, y) - g(\mathbf{x}, y)]|^2 dy d\mathbf{x} \\ &\leq C \int \int |[\hat{g}(\mathbf{x}, y) - g(\mathbf{x}, y)]|^2 dy d\mathbf{x} \end{aligned}$$

Therefore  $E \left[ \int |\hat{E}(y|\mathbf{x}) - E(y|\mathbf{x})| g(\mathbf{x}) d\mathbf{x} \right]^2 \leq C \cdot MISE(\hat{g}_l)$ . But by Theorem 3.2,  $MISE(\hat{g}_l)$  converges to 0 as  $n \rightarrow \infty, h \rightarrow 0$  and  $nh^p \rightarrow \infty$ , so:

$$\int |\hat{E}(y|\mathbf{x}) - E(y|\mathbf{x})| g(\mathbf{x}) d\mathbf{x} \rightarrow 0$$

in mean square.

(b) It follows directly from (a) since  $g(\mathbf{x})$  is uniformly bounded away from 0.

□

## 4.5 Other ways of computing $E[y|\mathbf{x}]$

In this section we will briefly give an introduction to three other ways that could be used to estimate  $E[y|\mathbf{x}]$ . A comparison of the four different methods can be found in Section 4.7.3 and in Section 5.3.

### 4.5.1 Multiple regression with random covariates

In multiple regression we assume there is a linear relationship between the response variable  $y$  and the random covariates  $x_1, \dots, x_{(p-1)}$ . The model we are trying to fit is the following:

$$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_{p-1} X_{p-1} + \text{error}. \quad (4.14)$$

It is further assumed that the data are independent and uncorrelated; however,  $x_1, \dots, x_{(p-1)}$  and  $y$  are correlated. For this section only, set  $\mathbf{x} = (x_1, \dots, x_{(p-1)})$ . Although we don't need any distributional assumptions to estimate the parameters  $\beta_0, \dots, \beta_{p-1}$ , however, for reasons of convenience, we often assume that  $(\mathbf{x}, y)$  follows a multivariate normal distribution  $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$  with  $\boldsymbol{\mu} = (\mu_{\mathbf{x}}, \mu_y)'$  and  $\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{\mathbf{xx}} & \sigma'_{\mathbf{xy}} \\ \sigma_{\mathbf{xy}} & \sigma_{yy} \end{pmatrix}$ . Then it can be easily shown (see for example [65]) that  $y|\mathbf{x}$  is normal

with

$$E(y | \mathbf{x}) = \mu_y + \sigma'_{xy} \Sigma_{xx}^{-1} (\mathbf{x} - \boldsymbol{\mu}_x) \quad (4.15)$$

$$= \beta_0 + \boldsymbol{\beta}_1 \mathbf{x} \quad (4.16)$$

where

$$\beta_0 = \mu_y - \boldsymbol{\sigma}'_{xy} \Sigma_{xx}^{-1} \boldsymbol{\mu}_x$$

$$\boldsymbol{\beta}_1 = \Sigma_{xx}^{-1} \boldsymbol{\sigma}_{xy}$$

Moreover  $\text{Var}(y|\mathbf{x}) = \sigma_{yy} - \boldsymbol{\sigma}'_{yx} \Sigma_{xx} \boldsymbol{\sigma}_{yx}$ . Notice that the mean  $E(y|\mathbf{x})$  is a linear function of  $\mathbf{x}$  but the variance  $\text{Var}(y|\mathbf{x})$  is constant. Therefore, under the multivariate normal assumption, model (4.14) has constant variance, as in the fixed  $\mathbf{x}$ -case. The mean is linear in the  $\mathbf{x}$ 's and in the  $\beta$ 's and thus, it does not allow curvature.

Although the model (4.14) has been used extensively, it is overly restrictive in its assumptions of a linear relationship between  $\mathbf{x}$  and  $y$  and the multivariate normal joint distribution and it is easy to run into misspecification problems.

## 4.5.2 The Nadaraya-Watson estimator

The Nadaraya-Watson estimator ([51], [74]) is a nonparametric estimator of the conditional expectation of  $Y$  relative to a vector of covariates  $\mathbf{X}$ ,  $E[Y|\mathbf{X}]$ . The idea is to estimate  $E[Y|\mathbf{X}]$  as a locally weighted average using a kernel as a weighting

function:

$$\hat{E}[Y|\mathbf{X}] = \frac{\sum_{i=1}^n K(H^{-1}(\mathbf{X}_i - \mathbf{x}))y_i}{\sum_{i=1}^n K(H^{-1}(\mathbf{X}_i - \mathbf{x}))} \quad (4.17)$$

where  $K(\mathbf{x})$  is the kernel function and  $H$  is the bandwidth matrix for  $\mathbf{x}$  which is positive and symmetric. The motivation for equation (4.17) is the same as for the semiparametric estimator (4.12): In the equation:

$$E[Y|\mathbf{X} = \mathbf{x}] = \frac{\int yf(\mathbf{x}, y)dy}{f(\mathbf{x})} \quad (4.18)$$

replace  $f(\mathbf{x}, y)$  and  $f(\mathbf{x})$  with their kernel estimators  $\hat{f}(\mathbf{x}, y) = \frac{1}{n|H|h_y} \sum_{i=1}^n K(H^{-1}(\mathbf{X}_i - \mathbf{x}))K\left(\frac{y_i - y}{h_y}\right)$  and  $\hat{f}(\mathbf{x}) = \int \hat{f}(\mathbf{x}, y)dy = \frac{1}{n|H|} \sum_{i=1}^n K(H^{-1}(\mathbf{X}_i - \mathbf{x}))$  respectively.

**Remark 4.3.** *The Nadaraya-Watson kernel estimate and the estimated conditional expectation (4.12) are both of the form  $\sum_i w_i y_i$ , where the  $w_i$  are positive weights which sum to 1, except that in (4.12) the  $w_i$  also depend on the  $y_i$ .*

### 4.5.3 Generalized additive models (GAM's)

Generalized additive models (GAM's) were first developed by Hastie and Tibshirani ([26], [27], [28]). They are blending properties of generalized linear models with additive models. The model has a structure of the form

$$g(\mu_i) = \mathbf{X}_i^* \boldsymbol{\theta} + f_1(x_{1i}) + f_2(x_{2i}) + f_3(x_{3i}, x_{4i}) + \dots \quad (4.19)$$

where  $\mu_i \equiv E[Y_i]$  and  $Y_i$  is the response variable which follows some exponential family distribution,  $g$  is a link function relating the expected value of the distribu-

tion to the predictors,  $\mathbf{X}_i$  is a row of the model matrix for any strictly parametric components,  $\boldsymbol{\theta}$  is the corresponding parameter vector, and  $f_j$  are smooth functions of the covariates  $x_k$ . In order for these kinds of models to be estimable, it is necessary to impose some identifiability conditions. A general method, which can deal with any choice of bases for the smooths, is described in [78]. By allowing nonparametric fits, GAM's allow good fits to the data. GAM's can be represented in various ways: using penalized regression splines, thin plate splines, tensor product smooths etc. The appropriate degree of smoothness for the  $f_j$  can be estimated using cross validation. GAM's are described in detail in [29] and [78]. For the simulation studies in Section 4.7 and the data analysis in Chapter 5 we fitted GAM's using the library `mgcv` in R [78].

## 4.6 Diagnostic plots and measures of goodness-of-fit

The density ratio model motivates graphical and quantitative diagnostic tools for measuring both goodness-of-fit of the model and the quality of the regression (4.12). Goodness-of-fit tests have been proposed by Gilbert [23], Qin and Zhang [60], and Zhang ([79], [81], [82]), where the appropriateness of the model is judged by the closeness of the estimated reference distribution to the corresponding empirical distribution. Bondell [8] suggests a reformulation of this in terms of the corresponding kernel density estimates. We suggest data analytic tools to measure discrepancies stemming from all case *and* control (reference) groups.

Graphical evidence of goodness-of-fit can be obtained from the plots of  $\hat{G}_i$  ver-

sus the corresponding empirical multivariate distribution function  $\tilde{G}_i$ ,  $i = 1, \dots, q, m$ , evaluated at some selected  $p$ -dimensional points as to obtain two dimensional plots. Figures 4.1 and 4.2 in the next section are examples of this. We refer to these plots as diagnostic plots.

We found the following measure of goodness-of-fit useful. Consider the  $i$ th sample of size  $n_i$ . The variance of the empirical cdf  $\hat{G}_i$  is  $G_i(1 - G_i)/n_i$  which can be estimated by  $\hat{G}_i(1 - \hat{G}_i)/n_i$ . Let  $x_\alpha$  be the number of times the estimated semiparametric cdf falls in the estimated  $1 - \alpha$  confidence interval obtained from the corresponding empirical cdf, both evaluated at the sample points. Define

$$R_{\alpha,k}^2 = 1 - \exp \left\{ - \left( \frac{x_\alpha}{n_i - x_\alpha} \right)^k \right\} \quad (4.20)$$

where  $k > 0$ , and  $k$  and  $\alpha$  are free parameters, which can be set by the user. Observe that:

- $R_{\alpha,k}^2$  takes values between 0 and 1, being close to 1 when  $x_\alpha$  approaches  $n_i$  and close to 0 when  $x_\alpha$  is close to 0.
- $R_{\alpha,k}^2$  is a flexible criterion that can be adjusted by changing the parameters  $\alpha$  and  $k$ . Larger  $\alpha$  means smaller confidence interval bounds.
- Computing  $R_{\alpha,k}^2$  is both simple and fast.

We now describe three natural alternatives to  $R_{\alpha,k}^2$ . First, as in multiple regression, goodness-of-fit may be approached by residual analysis. In this vein,



consider the decomposition

$$E[y - E(y)]^2 = E[y - E(y|\mathbf{x})]^2 + E[E(y|\mathbf{x}) - E(y)]^2. \quad (4.21)$$

Therefore, replacing  $\bar{y} \approx E(y)$  and  $\hat{y} \equiv E(y|\mathbf{x})$  in (4.21),

$$\frac{1}{n} \sum (y_i - \bar{y})^2 \approx \frac{1}{n} \sum (y_i - \hat{y}_i)^2 + \frac{1}{n} \sum (\hat{y}_i - \bar{y})^2$$

we define “ $R^2$ ” as in linear regression:

$$R_1^2 = \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2} \quad (4.22)$$

Next, define

$$R_2^2 = \text{corr}(y, \hat{y})^2 \quad (4.23)$$

Lastly, following Qin and Zhang [60], define

$$R_3^2 = \exp(-\sqrt{n} \cdot \max |\tilde{G}_i - \hat{G}_i|) \quad (4.24)$$

Clearly,  $R_3^2$  takes values between 0 and 1. Alternatives to  $R_3^2$  are  $\exp(-\sqrt{n} \cdot \text{median}|\tilde{G}_i - \hat{G}_i|)$  or  $\exp(-\frac{1}{n} \sum |\tilde{G}_i - \hat{G}_i|^2)$ .

The following simulation study suggests that  $R_{\alpha,k}^2$  is a more useful indicator of goodness-of-fit compared to  $R_1^2$ ,  $R_2^2$ , and  $R_3^2$ . An interesting problem would be to study the convergence of  $R_{\alpha,k}^2$ .

## 4.7 Some simulation results

In the present simulation study  $m = 2$ , and  $g_2$  denotes the reference distribution. To conform to the real data analysis in Chapter 5 we use the terms “case” and “control”, referring to the reference data as control. We considered the following bivariate simulation cases (runs):

1.  $g_1$  from  $N((0, 0)', \Sigma)$  and  $g_2$  from  $N((0, 0)', \Sigma)$  with  $\Sigma = \begin{pmatrix} 2 & 0.5 \\ 0.5 & 1 \end{pmatrix}$ . The corresponding sample sizes were  $n_1 = 90$  and  $n_2 = 70$ .
2.  $g_1$  from  $N((0, 0)', \Sigma_1)$  and  $g_2$  from  $N((1, 1)', \Sigma_2)$  with  $\Sigma_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$  and  $\Sigma_2 = \begin{pmatrix} 3 & 1 \\ 1 & 2 \end{pmatrix}$ . The corresponding sample sizes were  $n_1 = 200$  and  $n_2 = 200$ .
3.  $g_1$  from standard two dimensional multivariate Cauchy and  $g_2$  from two dimensional Multivariate Cauchy with  $\boldsymbol{\mu} = (1, 1)'$  and  $\mathbf{V} = \begin{pmatrix} 5 & 5 \\ 5 & 10 \end{pmatrix}$ . The corresponding sample sizes were  $n_1 = 200$  and  $n_2 = 200$ .
4.  $g_1$  from standard two dimensional multivariate Cauchy and  $g_2$  from uniform distribution on the triangle  $(0, 0), (6, 0), (-3, 4)$ . The corresponding sample sizes were  $n_1 = 200$  and  $n_2 = 200$ .

### 4.7.1 Comparison of the different measures of goodness-of-fit

The normal distribution follows the density ratio model, but this is not true for the Cauchy and the uniform distributions. Hence we expect to see straight lines

in the diagnostic plots and high  $R^2$ 's, as defined above, in cases (1) and (2). On the other hand, we expect to see deviations from straight lines in the diagnostic plots and lower  $R^2$ 's in cases (3) and (4).

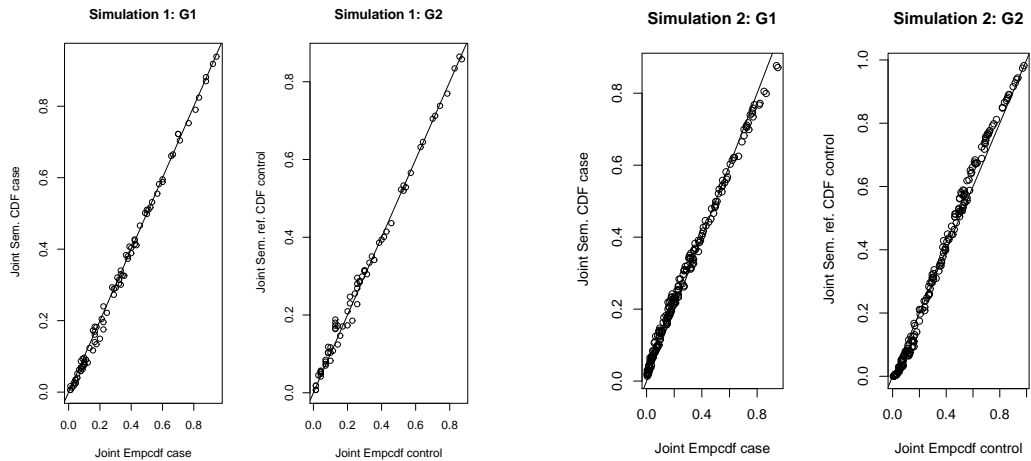


Figure 4.1: Case-control plots of  $\hat{G}_i$  vs.  $\tilde{G}_i$ ,  $i = 1, 2$ , simulations (1) and (2)

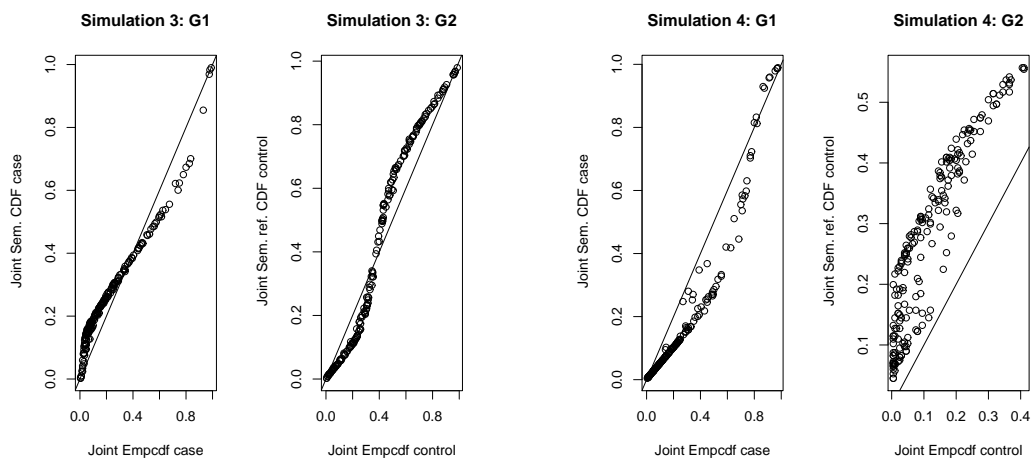


Figure 4.2: Case-control plots of  $\hat{G}_i$  vs.  $\tilde{G}_i$ ,  $i = 1, 2$ , simulations (3) and (4)

Figures 4.1-4.2 show the estimated  $\hat{G}_1$  and  $\hat{G}_2$  (where  $\hat{G}_1$  is the exponential tilt of  $\hat{G}_2$  defined in (4.9)) versus the empirical cdf  $\tilde{G}_1$  and  $\tilde{G}_2$ , respectively, all obtained

from one run of the simulated case-control data, and evaluated at selected points in  $\mathbb{R}^2$ . As expected, in cases (1),(2), there is almost a perfect agreement between  $\hat{G}_i$  versus  $\tilde{G}_i$ ,  $i = 1, 2$ , whereas Figure 4.2 shows clearly that the density ratio model is not appropriate for the data from cases (3) and (4).

A comparison of the different measures of goodness of fit for one run is given in Table 4.1. Apparently here  $R_1^2$  and  $R_2^2$  are misleading as measures of goodness of fit. They are erroneously higher at the cases where the simulated distributions do not follow the density ratio model. It seems that  $R_3^2$  is more appropriate than both  $R_1^2$  and  $R_2^2$  but it is sensitive to outliers and can give low values even for data that follow the density ratio model (e.g. case 2). On the other hand, the proposed measure  $R_{\alpha,k}^2$  classifies correctly the four cases, giving high values for simulations (1) and (2) and low values for (3) and (4). The values of  $R_{\alpha,k}^2$  in Table 4.1 were calculated with  $k = 2$  and  $1 - \alpha = 90\%$ . In general,  $R_{\alpha,k}^2$  gets closer to  $R_3^2$  by lowering  $1 - \alpha$ .

Table 4.1: Comparison of goodness of fit measures for case and control.

Run	Group	$R_1^2$	$R_2^2$	$R_3^2$	$R_{.10,2}^2$
(1)	Case	0.0098	0.1556	0.6193	1
	Control	0.0462	0.0761	0.6056	1
(2)	Case	0.0290	0.0470	0.3281	0.9998
	Control	0.1214	0.2356	0.3651	0.9999
(3)	Case	0.6948	0.8441	0.1390	0.1469
	Control	0.6792	0.7537	0.1294	0.1219
(4)	Case	0.4978	0.5662	0.0340	0.0999
	Control	0.4277	0.4372	0.0305	0.0001

We also run 100 repetitions of the four simulations and for each repetition we evaluated  $R_3^2$  and  $R_{\alpha,k}^2$  with  $k = 2$  and  $1 - \alpha = 90\%$  for case and control. Table 4.2

shows the mean value of  $R_3^2$  and  $R_{.95,2}^2$ . The results agree with the results presented in table 4.1.

Table 4.2: Comparison of  $R_3^2$  and  $R_{.05,2}^2$  for 100 repetitions of case and control.

Run	Group	$R_3^2$	$R_{.05,2}^2$
(1)	Case	0.6307	1
	Control	0.5976	1
(2)	Case	0.3912	0.9353
	Control	0.3766	0.9718
(3)	Case	0.1080	0.3342
	Control	0.1129	0.3324
(4)	Case	0.0507	0.3361
	Control	0.0495	0.0033

## 4.7.2 Bandwidth selection

In Chapter 3 we discussed several ways for selecting the bandwidth  $h$ . One way is to use equation (3.9) for the asymptotically optimal bandwidth and replace  $g_l$  with the normal density  $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ , where  $\boldsymbol{\mu}$  and  $\boldsymbol{\Sigma}$  are estimated from the data. The other option is to use cross validation and minimize either (3.10) or (3.12). Cross validation has the advantage that it allows us to use different bandwidths  $h_1, \dots, h_p$  to smooth each variable. Tables (4.3)-(4.5) summarize the results for the estimated  $h$  using equations (3.9), (3.10) and (3.12) for one run of the simulations. The integrals in (3.9) were calculated using *Mathematica*. However, certain integrals failed to converge and the results obtained are not trustworthy (see table (4.3)).

In Simulation 1 we used bandwidth 0.46 for case and 0.47 for control. In Simulation 2 we used bandwidth 0.33 for case and 0.51 for control. The bandwidth was selected after comparing visually the results for the fitted  $E[y|\mathbf{x}]$  and comparing

Table 4.3: Bandwidth selection using (3.9). Results in bold indicate cases where the integrals did not converge.

	Case BW	Control BW
Simulation 1	0.46	0.47
Simulation 2	0.33	0.51
Simulation 3	<b>6.23</b>	3.84
Simulation 4	<b>7.37</b>	0.31

Table 4.4: Bandwidth selection using the cross validation method (3.10).

	Case			Control		
	Same BW	Diff. BW's		Same BW	Diff. BW's	
	h	$h_1$	$h_2$	h	$h_1$	$h_2$
Simulation 1	0.61	0.90	0.40	0.59	0.31	0.61
Simulation 2	0.38	0.50	0.20	0.61	0.36	0.71
Simulation 3	0.34	0.40	0.30	2.11	2.51	0.96
Simulation 4	0.60	0.30	1.10	0.32	0.61	0.06

the MSE and MAE results. We also tried using different bandwidths for the variables but there wasn't any significant difference in the results.

### 4.7.3 Comparison with Nadaraya-Watson, GAM's and multiple regression

Using the semiparametric model, the standard normal distribution for kernel and (4.12), we estimated  $E[Y|X]$  for a single predictor. Table (4.6) provides a comparison of the MSE and MAE between the different methods for the first two simulations. In this table S.P. stands for Semiparametric Regression, M.R. for multiple regression, GAM for generalized additive model, and NW for Nadaraya-Watson. We did not estimate  $E[Y|X]$  for simulations 3 and 4 because the semiparametric

Table 4.5: Bandwidth selection using the cross validation method (3.12)

	Case			Control		
	Same BW	Diff. BW's		Same BW	Diff. BW's	
	h	$h_1$	$h_2$	h	$h_1$	$h_2$
Simulation 1	0.64	0.90	0.50	0.63	0.21	0.71
Simulation 2	0.30	0.40	0.20	0.74	0.11	0.96
Simulation 3	0.30	0.30	0.30	3.33	4.76	0.96
Simulation 4	0.30	0.30	0.30	0.15	0.36	0.06

model is not applicable in these cases (and was rejected as we saw from the  $R^2$  comparison). In simulations 1 - 2, for both case and control, we fitted a thin plate regression spline GAM assuming normal distribution and identity link. Another possible choice could have been a tensor product, but the results were almost identical. In simulation 1 the GAM line was almost exactly the same as the multiple regression line.

Table 4.6: MAE and MSE Comparison of the semiparametric prediction, multiple regression, GAM and Nadaraya-Watson estimators for Simulations 1 and 2.  $G_1, G_2$  signify case and control respectively.

		MSE				MAE			
		S.P.	M.R.	GAM	NW	S.P.	M.R.	GAM	NW
Simulation 1	$G_1$	0.913	0.834	0.834	0.851	0.752	0.741	0.741	0.736
	$G_2$	0.856	0.892	0.892	0.849	0.750	0.786	0.786	0.740
Simulation 2	$G_1$	0.820	0.841	0.799	0.792	0.723	0.730	0.709	0.704
	$G_2$	1.740	1.482	1.429	1.388	1.001	0.992	0.958	0.946

Figures 4.3-4.6 show the estimated  $E[Y|X]$  using equation (4.12). The prediction line is apparently influenced by the endpoints but otherwise it is a smooth curve. Superimposed are the lines obtained from multiple regression, GAM and the Nadaraya-Watson estimator and the true  $E[Y|X]$  line calculated from the theoretical distributions. From Table 4.6, we see that the semiparametric estimator

performs comparably with the other estimators in terms of the MSE and MAE.

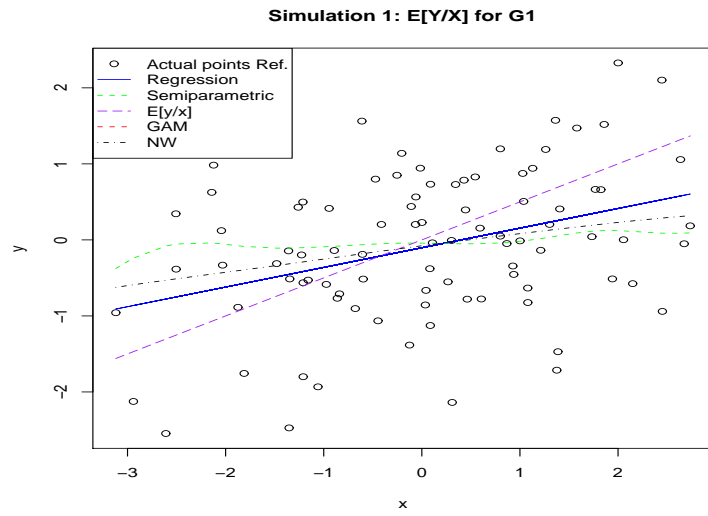


Figure 4.3: Comparison of  $E[Y|X]$  for  $G_1$  in simulation 1.

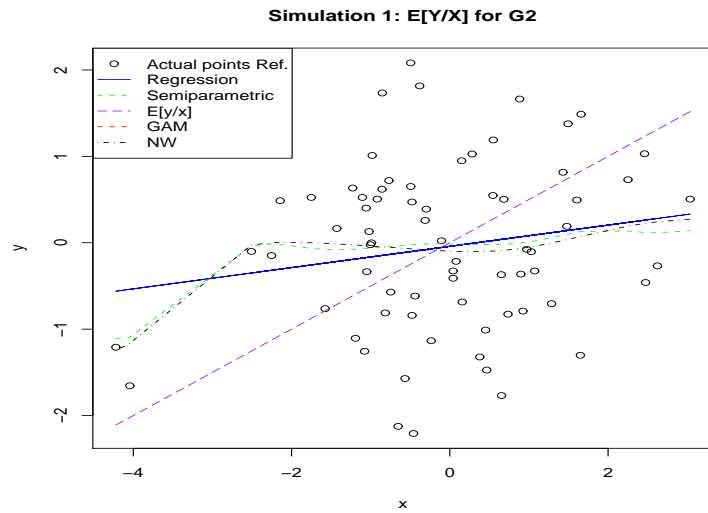


Figure 4.4: Comparison of  $E[Y|X]$  for  $G_2$  in simulation 1.



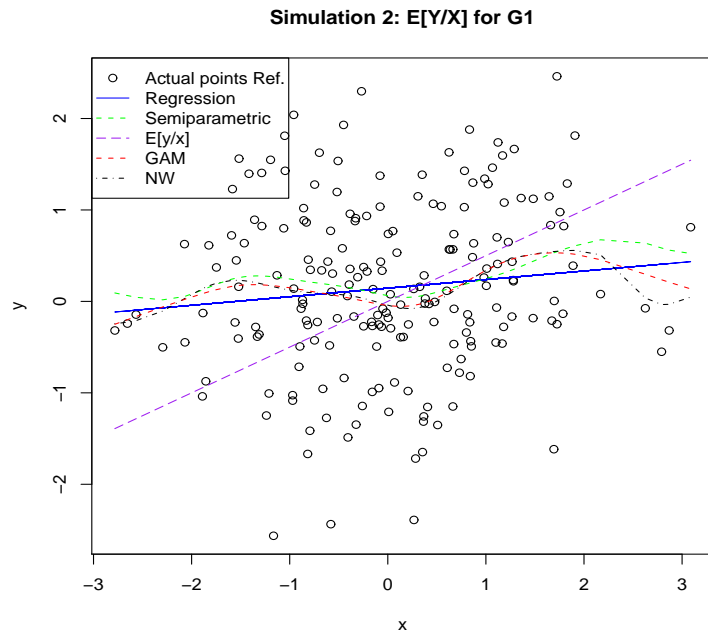


Figure 4.5: Comparison of  $E[Y|X]$  for  $G_1$  in simulation 2.

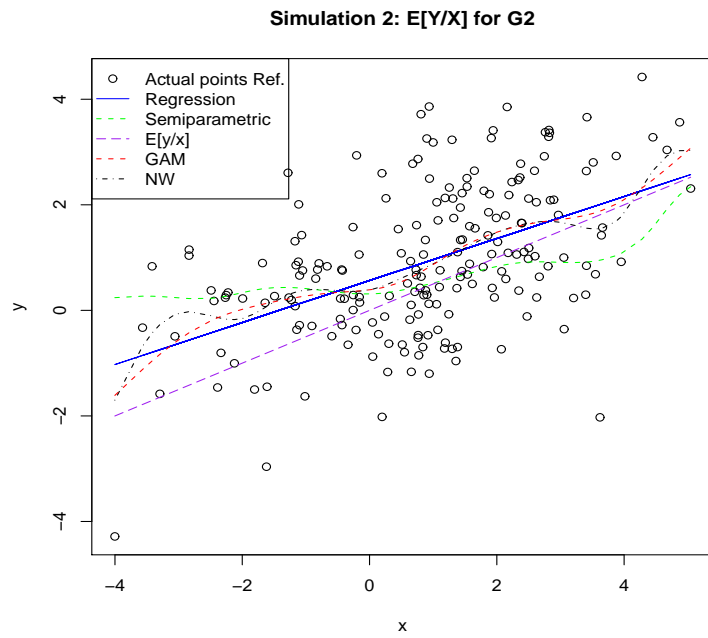


Figure 4.6: Comparison of  $E[Y|X]$  for  $G_2$  in simulation 2.

## Chapter 5

### The Testicular Germ Cell Tumor data set

#### 5.1 Introduction

In this chapter we apply many of the results and methods discussed in Chapters 3 and 4 to testicular germ cell tumor (TGCT) data set. Testicular cancer is the most common solid malignancy affecting mainly Caucasian men between the ages 15 and 35. It is rare among men of African or Asian descent. The cure rate is more than 90%, approaching 100% if it has not metastasized. Even for the relatively few cases in which malignant cancer has spread widely, chemotherapy offers a cure rate of at least 85%. A major risk factor for the development of testicular cancer is cryptorchidism (undescended testicles). Other risk factors include seminoma, prior history of TGCT, family history of TGCT. Physical activity is associated with decreased risk, whereas sedentary lifestyle and early onset of male characteristics is associated with increased risk. Other possible risk factors include body size, dairy consumption, and age at puberty [46], [47].

The TGCT data set consists of 763 cases and 928 controls enrolled in the Servicemen's Testicular Tumor Environmental and Endocrine Determinants Study (2002 – 2005). In [47], McGlynn *et al* determined that increased height was significantly related to risk (odds ratio (OR) = 1.83, 95% confidence interval (CI): 1.36, 2.45), where this OR is for men with height greater than 182.88 cm compared to

those with less than or equal to 172.72 cm. On the other hand, body mass index (BMI: weight in kilograms divided by height in meters squared) was not significant (OR = 1.06, 95% CI: 0.66, 1.69), where this OR is for men with body mass index greater than or equal to 30 compared to those with less than 18.5. Furthermore, there was no association found for age at puberty (based on ages at first shaving), voice changing, nocturnal emissions, and dairy consumption at any age between birth and 12th grade.

The original TGCT data set had several variables [47] but the portion made available to us consisted of the following eight variables: subject ID, age at reference date, an indicator for case or control group (0=case, 1–4=control), the participant’s height in cm, participant’s weight in kg, participant’s BMI (kg/m<sup>2</sup>), family history of testicular cancer (0=no, 1=yes) and race/ethnicity (1=white, 2=black, 3=other). We focused on three variables: height, weight and age. Out of these, height and weight had undergone some kind of discretization: there were 21 unique values for height and 89 unique values for weight. Table (5.1) gives summary statistics for height and weight for both groups. We notice that the variance-covariance structure in the two groups is quite similar.

Table 5.1: Case-control summary statistics regarding height (cm) and weight (kg), and the correlation between them.

	Height				Weight				corr
	min	max	ave	sd	min	max	ave	sd	
Case	160.0	203.2	179.6	7.0	50.8	131.5	81.4	11.7	0.521
Control	152.4	215.9	178.3	7.1	38.6	127.0	80.1	11.1	0.505

This Chapter is organized as follows: In Section 5.2 we discuss the problems encountered in the bandwidth selection process for both the 2D and the 3D TGCT data set. Section 5.3 presents the results of the data analysis.

## 5.2 Bandwidth selection for the TGCT data set

We used equations (3.9), (3.10), (3.12) with kernel  $K = N(\mathbf{0}, \mathbf{1})$  and  $w(\mathbf{x}, \boldsymbol{\theta}_i) \equiv \exp(\alpha_i + \boldsymbol{\beta}'_i \mathbf{x})$  from Chapter 3 to calculate the bandwidth in the same way as we did in Section 4.7.2. We considered two cases: the 2D TGCT data set with variables height and weight and the 3D TGCT data set with variables height, weight and age. The integrals in (3.9) were calculated using *Mathematica*. However, *Mathematica* failed to calculate the integrals for the TGCT data set when using all three variables age, height and weight. When applied to both the 2D and the 3D TGCT data set, equations (3.10)-(3.12) were strictly increasing, which means they lead to the degenerate choice of  $h = 0$  of smoothing parameter. An obvious solution is to add a little bit of noise to the data, enough to break the ties in the data without changing the data too much. The results in tables (5.3) and (5.4) were obtained by adding noise generated by  $N(0, 0.1^2)$  to age,  $N(0, 0.7^2)$  to height,  $N(0.06^2)$  to weight. This resulted in a change of about  $\pm 0.3$  for age,  $\pm 2$  cm for height and about  $\pm 1.8$  kg for weight. However it should be noted that if we used noise generated from normal distribution with smaller variance the results changed. Generally the smaller the variance used, the smaller was the optimal bandwidth obtained from minimizing equations (3.10)-(3.12). However, by looking at the nature of the data and the

distance between the discretized values of the variables, we would not recommend adding less noise. For the 3D TGCT data set the optimal bandwidth was calculated using only equation (3.12) since equations (3.10) and (3.11) are not time efficient. For the 2D TGCT data set, for case, we decided to smooth the data using 1.01 and 3.51 for height and weight respectively, whereas, for control, we used 2.02 and 1.01. The results would be similar if we had used bandwidths 2.06 and 1.61 for case and control respectively. For the 3D TGCT the best results were produced when using bandwidth 2.24 for control and 2.5 for case.

Table 5.2: Bandwidth selection using (3.9). Mathematica failed to calculate the integrals in (3.9) for the 3D TGCT data set.

	Case BW	Control BW
2D TGCT (height, weight)	4.60	2.11
3D TGCT (age, height, weight)	-	-

Table 5.3: Bandwidth selection using the cross validation method (3.10). The method was not used to calculate the bandwidth in the 3D TGCT data set because it is not time efficient.

	Case				Control			
	Same BW h	$h_1$	Diff. BW's $h_2$ $h_3$		Same BW h	$h_1$	Diff. BW's $h_2$ $h_3$	
2D TGCT	2.06	1.01	3.51	NA	1.61	2.01	1.01	NA
3D TGCT	-	-	-	-	-	-	-	-

Table 5.4: Bandwidth selection using the cross validation method (3.12)

	Case				Control			
	Same BW	Diff. BW's			Same BW	Diff. BW's		
	h	$h_1$	$h_2$	$h_3$	h	$h_1$	$h_2$	$h_3$
2D TGCT	2.85	1.01	4.26	NA	1.53	2.01	1.01	NA
3D TGCT	2.24	0.1	4.3	4.5	2.5	0.1	4.1	6.85

### 5.3 Data analysis

In [37] the 2D TGCT data set was analyzed with variables height and weight. Assume the density ratio model (4.1) which in the present case can be written as,

$$\frac{g_1(x, y)}{g_2(x, y)} = \exp(\alpha_1 + \beta'_1 \mathbf{x}) \quad (5.1)$$

where  $g_1$  is the distribution of the case group, and  $g_2$  is the reference distribution of the control group. From the score equations (4.6)-(4.7) we obtain:

$$(\hat{\alpha}, \hat{\beta}_{11}, \hat{\beta}_{12}) = (-4.676, 0.025, 0.002) \quad (5.2)$$

with respective standard errors (0.914, 0.006, 0.004), indicating dissimilarity between the two groups. We can test the hypothesis  $H_0 : \beta_1 = 0$  using the likelihood ratio test (4.11). In this case, the likelihood ratio (4.11) is equal to 15.108, and with 2 degrees of freedom the corresponding  $p$ -value is 0.0005. Thus, when height and weight are considered jointly, we reject the null hypothesis  $H_0 : \beta_1 = 0$  of equidistribution quite conclusively. This means that *jointly* height and weight are significant risk factors. An advantage of the method is that we can find estimates for the joint

probabilities of height and weight in both the case and the control group as in table 5.5. The table shows there are differences between the two groups, thus verifying the likelihood ratio test.

Table 5.5: Some joint probabilities of height and weight in the case and control groups.

Probability	Case	Control
$\Pr(H \leq 152.40, W \leq 58.967)$	0.000374	0.000770
$\Pr(H \leq 165.10, W \leq 58.967)$	0.005829	0.009216
$\Pr(H \leq 177.80, W \leq 65.317)$	0.052177	0.067014
$\Pr(H \leq 185.42, W \leq 70.307)$	0.185290	0.218345
$\Pr(H \leq 180.34, W \leq 79.832)$	0.376876	0.434745
$\Pr(H \leq 180.34, W \leq 89.811)$	0.558730	0.627897
$\Pr(H \leq 187.96, W \leq 94.801)$	0.819068	0.857814
$\Pr(H \leq 200.66, W \leq 99.790)$	0.945452	0.958643
$\Pr(H \leq 203.20, W \leq 117.934)$	0.995568	0.997178

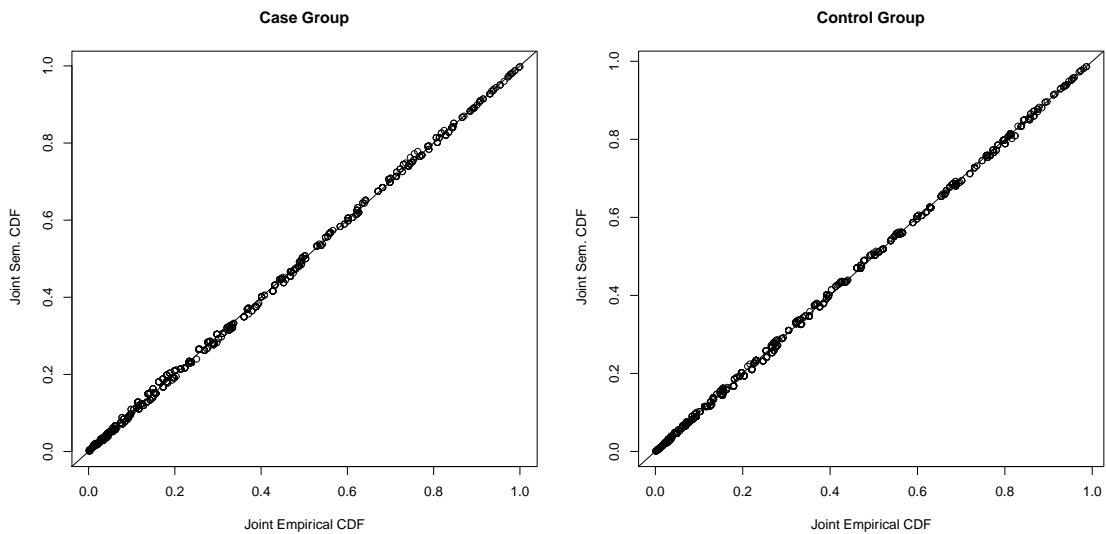


Figure 5.1: 2D problem: Plots of  $\hat{G}_i$  versus  $\tilde{G}_i$ ,  $i = 1, 2$  evaluated at (height, weight) pairs for the case and control groups from the TGCT data.

Before applying the three-dimensional density ratio model to the TGCT data, it is interesting to apply the two-dimensional model to get a prediction of weight

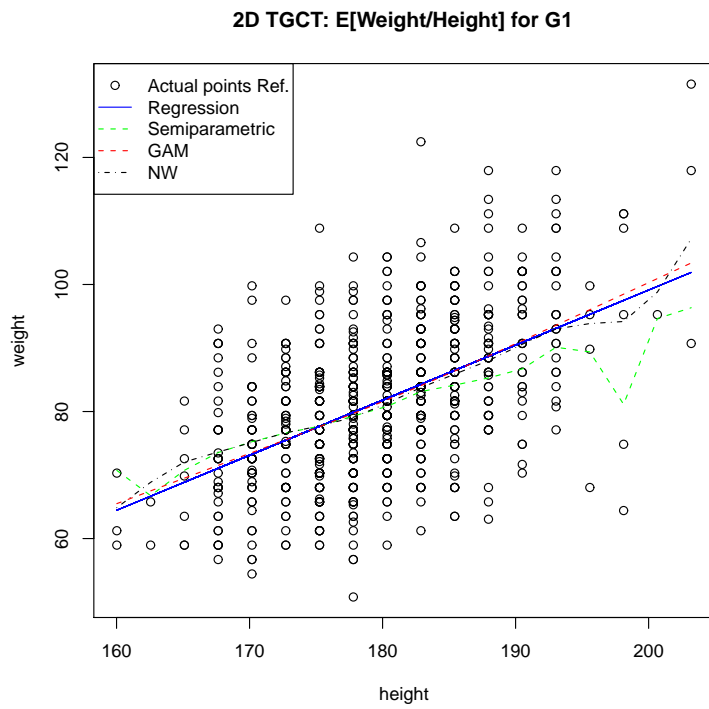


Figure 5.2: Comparison of  $E[Y|X]$  for  $G_1$  in the 2D TGCT data set.

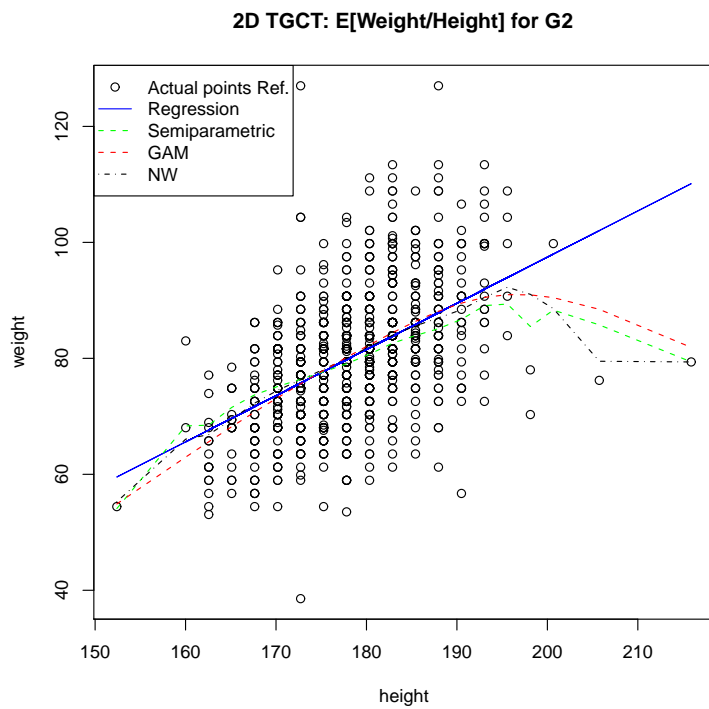


Figure 5.3: Comparison of  $E[Y|X]$  for  $G_2$  in the 2D TGCT data set.



given height only. As Figure 5.1 shows, the density ratio model is a suitable model for the TGCT data: there is almost a perfect agreement between the plots of the estimated semiparametric  $\hat{G}_i$  and the corresponding empirical  $\tilde{G}_i$ ,  $i = 1, 2$ . The value of  $R_{20,1}^2$  is 1 for both case and control. Figures 5.2-5.3 show the estimated  $E[Y|X]$  using equation (4.12) for the case and control groups, where in the 2D TGCT data set  $Y$  is the variable weight and  $X$  is the height. If we had used the bandwidth given by equation (3.9), then the conditional expectation lines would have been smoother but the MSE and MAE would have been higher. Superimposed is the regression line obtained from linear regression under the normal assumption, the GAM line and the Nadaraya-Watson regression line. For the 2D TGCT data, assuming normal distribution and identity link, we fitted a tensor product GAM, although there were not a lot of differences between the different kinds of splines. We notice that all models give similar results. The residual plots for the semiparametric model in Figure 5.4 are centered around zero.

Next we fitted the 3D TGCT data with variables age, height and weight. The semiparametric model is an appropriate model for this data set as Figure 5.5 shows. The value of  $R_{20,1}^2$  is 1 for both case and control. Again we used equation (4.12) to calculate  $E[Y | \mathbf{X}]$  for the case and control groups, where in the 3D TGCT data set  $Y$  is the weight and  $\mathbf{X}$  represents jointly height and age. Figure 5.6 shows the residual plots for the semiparametric model. Table 5.6 gives the MSE and MAE comparison between the different estimators for the 2D and the 3D TGCT data. For the 3D TGCT data, assuming normal distribution and identity link, we fitted a thin plate regression spline GAM because it produced better looking residual

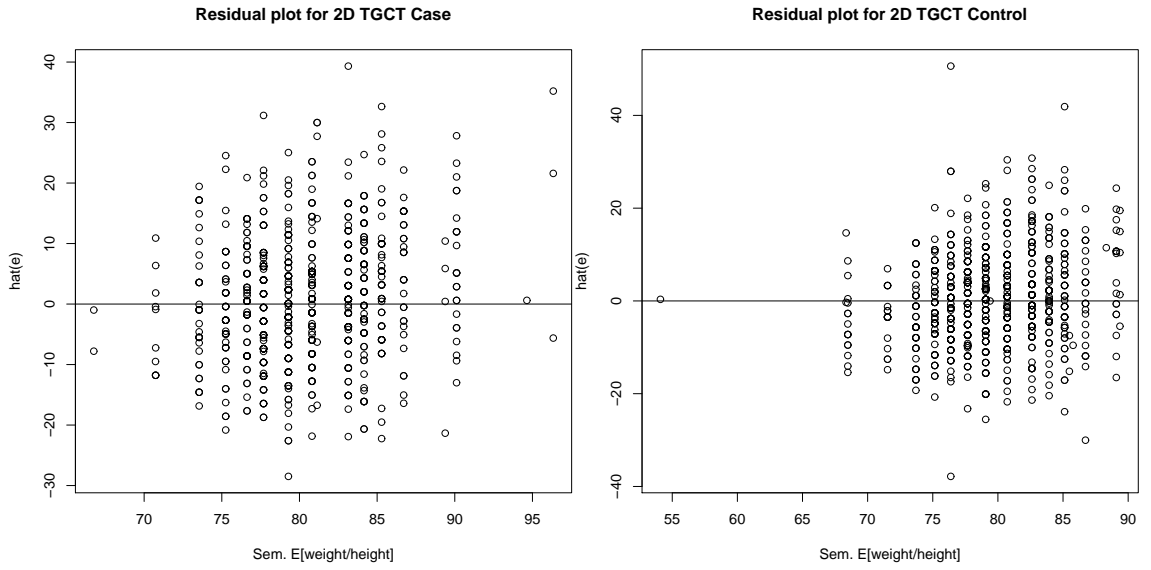


Figure 5.4: Residual plots for the semiparametric model in the 2D TGCT data set.

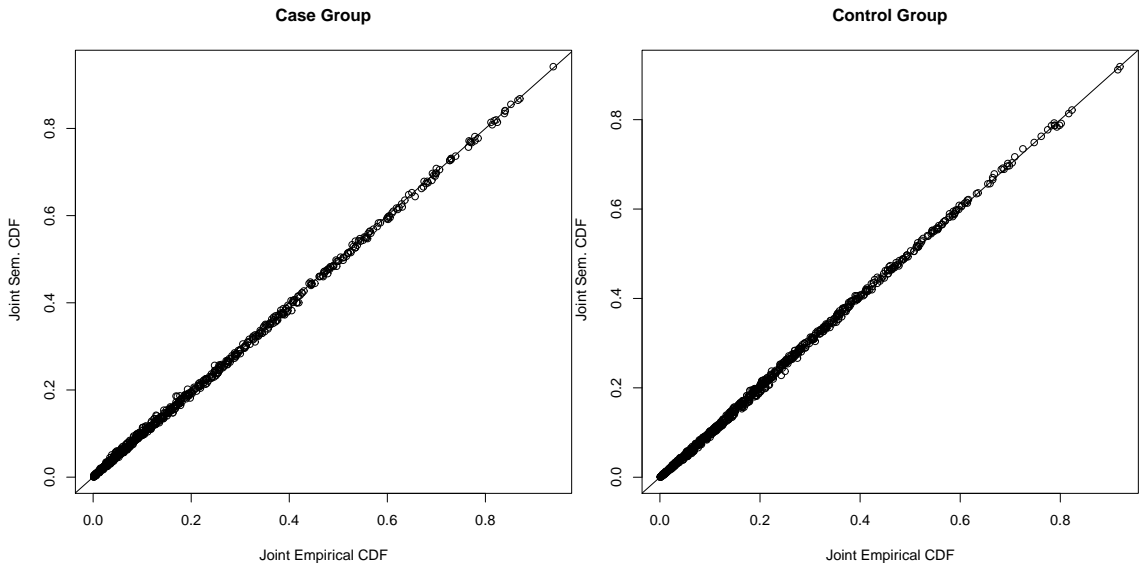


Figure 5.5: Case-control plots of  $\hat{G}_i$  versus  $\tilde{G}_i$ ,  $i = 1, 2$  for the 3D TGCT problem: the  $\hat{G}_i, \tilde{G}_i$  are evaluated at selected (age,height,weight) triplets.

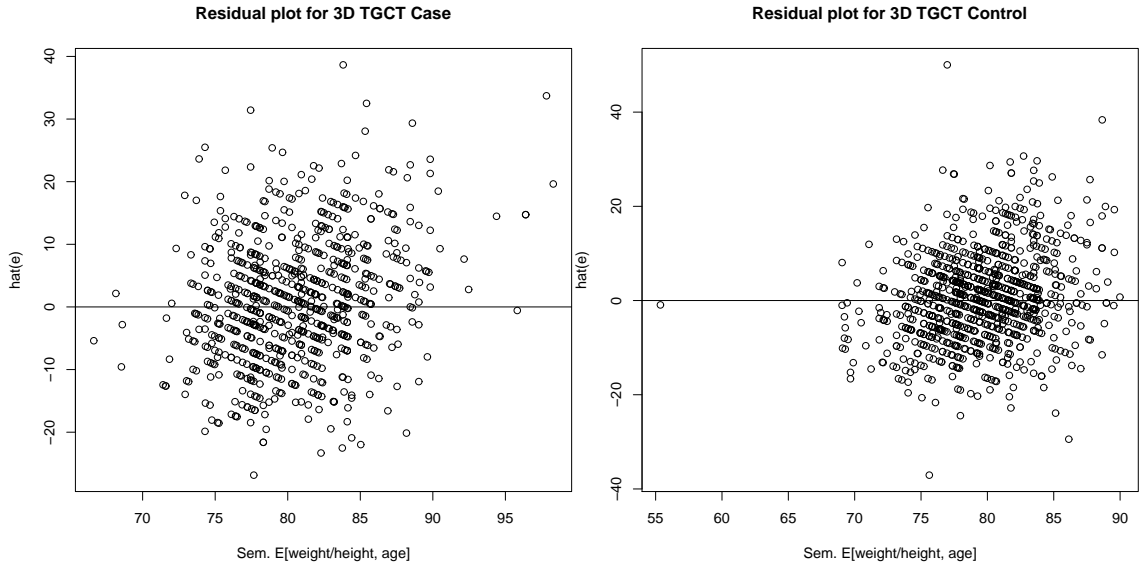


Figure 5.6: Residual plots for the semiparametric model in the 3D TGCT data set.

and Q-Q plots. The semiparametric estimator performs comparably with the other estimators, although somewhat worse. These results can be explained under the light that our method consists of an extra step of density estimation. However we have the extra advantage that we also calculate the joint probabilities of the variables without making any distributional assumptions like multiple regression and GAM's.

Table 5.6: MAE and MSE comparison of the semiparametric prediction and multiple regression for 2D and 3D TGCT data.

		MSE				MAE			
		S.P.	M.R.	GAM	NW	S.P.	M.R.	GAM	NW
2D TGCT	$G_1$	104.003	99.510	99.250	98.648	7.947	7.784	7.770	7.774
	$G_2$	93.010	92.264	90.284	90.332	7.347	7.296	7.246	7.241
3D TGCT	$G_1$	98.283	96.367	96.091	89.124	7.770	7.679	7.672	7.390
	$G_2$	91.643	90.291	88.147	86.932	7.280	7.244	7.173	7.139

Tables 5.7 and 5.8 give some predicted values for weight given age and height for the two models.

Table 5.7: Predicted control values of weight given height and age.

Case						
Age	Height	Weight	S.P.	M.R.	GAM	NW
26	193.04	102.058	89.81775	92.47554	92.80697	95.96000
24	167.64	72.575	73.59282	70.00329	70.68805	71.90371
29	180.34	65.771	81.41551	82.42360	82.17237	81.60395
38	185.42	81.647	86.29762	89.46406	89.50287	89.70666
34	195.58	89.811	89.03635	97.03194	98.08814	92.45555
27	162.56	58.967	68.53652	66.51540	67.76775	65.18988

Table 5.8: Predicted case values of weight given height and age.

Control						
Age	Height	Weight	S.P.	M.R.	GAM	NW
29	180.34	90.718	81.11841	82.06293	83.06542	82.35544
39	175.26	77.111	79.40282	80.36549	79.78087	80.05940
19	172.72	63.503	74.76493	73.58821	72.76199	73.40060
33	177.80	83.915	80.51759	80.97707	81.4916	81.14195
31	190.50	102.058	86.0598	90.67494	90.69862	87.47080
25	165.10	58.967	72.08147	68.90777	68.0279	69.49050

We end this section by noting that, as expected,  $\hat{E}(y|\mathbf{x})$  in (4.12) *tends to be close to the average of  $y$ 's which correspond to the same  $\mathbf{x}$* . This is demonstrated in Tables 5.9 and 5.10 which give the case-control weight predictions (4.12) and the actual weights. Empty entries in the table correspond to subjects with the same height and age (i.e. same  $\mathbf{x}$ ), but possibly different weights. The averaging property can be seen by averaging the run of weights in the “empty cells” and the run upper bound. Thus, for example, the control-weights corresponding to age 22 and height 175.26 average to 74.3894 and the prediction is 76.62195.

## 5.4 Conclusion

In this Chapter we have demonstrated that the multidimensional density ratio model has several advantages. It provides estimates of the joint probabilities in both the case and the control groups. All the parameters and the reference cdf are estimated from the combined data, and not just from the reference sample, leading to more precise estimates. The process of fitting the model and obtaining estimates for the parameters is very simple, straightforward and quick. Moreover, the semi-parametric model provides a way for determining the difference between two or more multivariate distributions and for testing multivariate equidistribution. Going one step further it can be used in estimating the conditional expectation of a response variable given random covariates when multiple data sources are available without making any distributional assumptions. In addition the suggested graphs and quantitative validation measures are useful in assessing the suitability of the method. The method works best for a small number of covariates since technical difficulties can arise in the bandwidth computation as the number of variables increases.

The approach offers a way of understanding how multivariate distributions representing many different data sources are related to each other. This leads to a ramification of the notion of regression where the objective is to model relationships between distributions. Relationships between response variables and their covariates, corresponding to the data sources, are byproducts.

Table 5.9: Case-control weight and  $\hat{E}[\text{weight}|\text{height, age}]$ . Empty entries in the table correspond to subjects with the same height and age, but possibly different weights.

Age	Height	Control		Case	
		Weight	$\hat{E}[W   H, A]$	Weight	$\hat{E}[W   H, A]$
27	162.56	58.967	69.08335	58.967	68.53652
28	162.56	77.111	69.05132	65.771	68.59858
		68.039			
30	165.10	68.039	72.20524	72.575	72.0028
37	165.10	69.40	72.42138	63.503	71.8504
25	167.64	86.183	73.68129	72.575	73.69978
				90.718	
				63.503	
30	167.64	72.575	74.81333	88.451	74.93543
18	170.18	61.235	73.67032	72.575	73.67518
32	170.18	70.307	76.53351	81.647	76.64543
		63.503			
37	172.72	74.843	77.88598	88.451	77.9417
40	172.72	70.307	77.97789	90.718	78.0441
		77.111			
22	175.26	77.111	76.62195	86.183	76.70862
		65.771		65.771	
		79.379		86.183	
		83.915			
		65.771			
25	175.26	68.039	77.14234	79.379	77.21755
		83.915		72.575	
		74.843		83.915	
		83.915		74.843	
		79.379		72.575	
		86.183		74.843	
				61.235	
				61.235	
				65.771	
				79.379	

Table 5.10: Case-control weight and  $\hat{E}[\text{weight}|\text{height, age}]$  continued. Empty entries in the table correspond to subjects with the same height and age, but possibly different weights.

Age	Height	Control		Case	
		Weight	$\hat{E}[W   H, A]$	Weight	$\hat{E}[W   H, A]$
26	177.80	79.379	78.74752	77.111	78.92705
		81.647		104.326	
		58.967		77.111	
		81.647			
		79.379			
		74.843			
		88.451			
		68.039			
42	177.80	70.307	80.50100	91.626	80.67493
20	180.34	79.832	79.17623	84.368	79.35688
		65.771		68.039	
		77.111		79.379	
		79.379		81.647	
				72.575	
33	180.34	79.379	81.92536	77.111	82.17689
				81.647	
18	182.88	77.111	80.23013	68.039	80.29011
41	182.88	79.379	83.65558	86.183	84.06475
19	185.42	63.503	81.45580	68.039	82.09186
				94.347	
				68.039	
21	185.42	86.183	82.46773	79.379	82.78140
		72.575		77.111	
		102.058		97.522	
22	190.50	97.522	85.23493	86.183	85.64845
		95.254		71.668	
31	190.50	102.058	86.05980	104.326	86.27744
				74.843	
22	193.04	86.183	86.73352	102.058	87.18440
				80.739	
24	193.04	99.337	87.50020	108.862	88.23938
		86.183			
		99.790			
		108.862			
34	193.04	113.398	87.72937	88.451	88.58960
				117.934	
34	195.58	83.915	88.81524	89.811	89.036535

## Chapter 6

### Estimation of death rates in U.S. States with small subpopulations

#### 6.1 Introduction

In this Chapter we use historical death rate data from states with small populations to fit appropriate probability models supported discretely at zero in order to replace zero death rate observations with estimates of their expected values. Since expected mortality is positive, its logarithmic transformation is not a problem. Then, we use a combination of the actual and the estimated points to fit the eight parameter Heligman-Pollard model [30] to smooth the data and get estimated values of mortality for older ages. In some cases this procedure is useful in relaxing the minimum sample size criteria for the publication of state-race-sex specific life tables.

In Section 6.2 we give a detailed description of the probability models that were used to estimate the expected number of deaths, or the corresponding expected death rate, when the observed value was zero, and of the procedures used to fit the Heligman-Pollard model. In Section 6.3 we present the results of our analysis, including a comparison of the performance of the probability models relative to the percentage of observed zeros in the data set, and figures of mortality curves fitted for each data set.



## 6.2 Models and methods

This Section consists of three parts: In the first part we describe the concept of the mixed distribution, on which most of the probability models are based. In the second part we present the probability models, and in the third part we give a brief description of the Heligman-Pollard model.

### 6.2.1 Mixed distributions

The response variable in our data is either the number of deaths or the corresponding mortality rate (death rate) for a given year and age. Because we address *zero – valued* observations, almost all our probability distribution models are mixtures of a discrete component defined at 0, and a second component defined for values greater than 0. The latter can be either discrete or continuous depending on the response variable. To introduce this basic feature of the present work, it suffices to consider a continuous second component only. The discrete component is a spike at zero whose magnitude is equal to the probability of admitting the value zero. Some useful references on mixed models include [2], [3], [31], [33] and [53].

Let  $Y$  be the variable of interest representing mortality rate. A natural model for  $Y$  is a mixed distribution probability model:  $Y = 0$  with probability  $1 - p$ , where  $0 < p < 1$ , but otherwise, for positive rate,  $Y$  follows a continuous distribution with cumulative distribution function (cdf)  $F(y, \boldsymbol{\theta}_1)$ . Here,  $\boldsymbol{\theta}_1$  represents a vector of parameters. Then, the distribution of  $Y$  is a mixture of discrete and continuous

components,

$$P(Y \leq y) \equiv G_m(y; p, \boldsymbol{\theta}_1) = (1 - p)H(y) + pF(y; \boldsymbol{\theta}_1) \quad (6.1)$$

where  $H(y)$  is a step function:

$$H(y) = \begin{cases} 0, & y < 0 \\ 1, & y \geq 0. \end{cases}$$

The corresponding generalized probability density is:

$$g_m(y; p, \boldsymbol{\theta}_1) = (1 - p)^{1 - I[y > 0]} [pf(y; \boldsymbol{\theta}_1)]^{I[y > 0]}, \quad y \geq 0 \quad (6.2)$$

where  $f(y; \boldsymbol{\theta}_1)$  is a probability density function conditional on  $Y > 0$  corresponding to  $F(y; \boldsymbol{\theta}_1)$ , and  $I[A]$  is the indicator of the event  $A$ . That is,  $I[A] = 1$  if  $A$  occurs, and  $I[A] = 0$  if  $A$  does not occur.

In this setup, the goal is to estimate the mean of the mixed distribution,

$$E(Y) = pE(Y | Y > 0) \equiv p\alpha \quad (6.3)$$

which is a function of  $\boldsymbol{\theta} \equiv (p, \boldsymbol{\theta}_1)$ . Observe that (6.3) is a product of two factors,  $p$  and  $\alpha \equiv E(Y | Y > 0)$ , corresponding to the two distribution components. We can estimate (6.3) using the maximum likelihood estimates of  $p$  and  $\alpha$ , as in the mixed lognormal distribution, or by regressing each of the two factors on covariates and

then taking the product of the two regressions, as in the so called two-part, hurdle, and zero-inflated models. All these models are discussed next.

## 6.2.2 Probability models

In the models below the response variable  $Y$  can be either the number of deaths for a given age and year, or the corresponding death rate. Thus, in the mixed lognormal distribution and the two-part model we model death rate, whereas the Poisson, Hurdle and the Zero-inflated models address the closely related number of deaths, from which expected death rate can be estimated. In the regression models we used, the independent variables are continuous, binary, or a mixture of the two.

### 6.2.2.1 Mixed lognormal distribution

Let  $Y$  denote death rate. Referring to the general mixed distribution (6.1), a useful model is the mixed lognormal distribution where the continuous part of the distribution of death rate is lognormal  $LN(\mu, \sigma^2)$ , with density,

$$f(y; \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma y}} \exp\{-(\log y - \mu)^2/(2\sigma^2)\}, \quad y > 0 \quad (6.4)$$

Let  $\boldsymbol{\theta} = (p, \mu, \sigma)$ . Then the mean of  $Y$  is a function, say  $g$ , of  $\boldsymbol{\theta}$ ,

$$g(\boldsymbol{\theta}) = E(Y) = p \exp\{\mu + \sigma^2/2\}, \quad (6.5)$$

and we have to estimate  $\theta = (p, \mu, \sigma)$ . Assuming that the data are independent and identically distributed (iid), the maximum likelihood estimators for  $p, \mu, \sigma$  are:

$$\hat{p} = \frac{\sum_i I[y_i > 0]}{n}, \hat{\mu} = \frac{\sum_i \ln(y_i) I[y_i > 0]}{\sum_i I[y_i > 0]}, \hat{\sigma} = \sqrt{\frac{\sum_i (\ln(y_i) I[y_i > 0] - \hat{\mu})^2}{\sum_i I[y_i > 0]}}$$

where  $n$  is the sample size and  $-\infty \cdot 0 \equiv 0$ . For the sake of meaningful discussion, we rule out the case that  $\sum_i I[y_i > 0] = 0$ . The estimated mean death rate is then

$$\hat{E}(Y) = \hat{p} \exp\{\hat{\mu} + 0.5\hat{\sigma}^2\}. \quad (6.6)$$

Estimation of the parameters in mixed distributions is discussed in detail in [3], [31], [33], [53].

We note that the data are not iid since death rates are decreasing moderately as a function of time. Hence there is dependence in annual rates [36]. We therefore must view our maximum likelihood estimates in a partial sense [32]. Our data analysis indicates that despite of this difficulty, the prediction results are quite satisfactory.

In order to overcome the time and age trends, the mixed lognormal model is applied to non-overlapping windows of either 10 or 11 years and two consecutive ages. For this short window of time×age it is reasonable to assume that death rates are approximately equidistributed. Each sample consists of all the observations for the length of the window in time and for two consecutive ages, for example, ages

1 and 2, 3 and 4, and so on over 10 or 11 years. This gives 20 to 22 observations per sample. For every sample the maximum likelihood estimators for  $p, \mu, \sigma$  are computed and then used in the estimation of (6.6).

The continuous part of the distribution of death rate can also be modeled by the Gamma or Beta distributions. However, since there is no noticeable significant improvement, the mixed lognormal model is preferable for reasons of computational efficiency.

The precision of the maximum likelihood estimates is obtained from the diagonal of the Fisher information matrix per observation,

$$\mathbf{I}_f = -\mathbf{E} \left( \frac{\partial^2 \log g(y, \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \right). \quad (6.7)$$

In particular, for the mixed lognormal distribution the Fisher information matrix is

$$\mathbf{I}_f = \begin{bmatrix} \frac{1}{p(1-p)} & 0 & 0 \\ 0 & \frac{p}{\sigma^2} & 0 \\ 0 & 0 & \frac{2p}{\sigma^2} \end{bmatrix}$$

and

$$\sqrt{n}\{(\hat{p}, \hat{\mu}, \hat{\sigma})' - (p, \mu, \sigma)'\} \rightarrow N(\mathbf{0}, \mathbf{I}_f^{-1}).$$

Using the delta method [64] the variance of the mean estimator (6.6) is

$$\text{Var}(\hat{\mathbf{E}}(Y)) \approx \frac{1}{n} \exp(2\mu + \sigma^2)[p(1-p) + p\sigma^2 + p\sigma^4/2]. \quad (6.8)$$

Equation (6.8) can be estimated by replacing  $(p, \mu, \sigma)$  by their ML estimates. Then approximate 95% confidence intervals can be calculated as

$$\hat{E}(Y) \pm 1.96 \cdot \sqrt{\widehat{\text{Var}}(\hat{E}(Y))}.$$

### 6.2.2.2 Two-Part model

Let  $Y$  denote death rate. In the two-part model the regression of  $p$  on covariates is referred to as the “hurdle component,” and the regression of  $\alpha$  on covariates is referred to as the “levels component” [4]. The hurdle component is different from the “hurdle model”. In the general case when both numerical variables and factors are available, the hurdle component can be modeled by logistic regression. An example (using the logit link) is given by

$$P(Y_{\text{age, year}} > 0) = \frac{\exp(\eta)}{1 + \exp(\eta)}$$

where

$$\eta = \mu + \alpha_{\text{age}} + \gamma_{\text{year}} + \beta \log(\text{population size}_{\text{age, year}}),$$

or, in matrix notation we write,

$$P(\mathbf{Y} > \mathbf{0}) = \frac{\exp(\mathbf{Z}\boldsymbol{\alpha} + \mathbf{X}\boldsymbol{\beta})}{1 + \exp(\mathbf{Z}\boldsymbol{\alpha} + \mathbf{X}\boldsymbol{\beta})}, \tag{6.9}$$

where  $\mathbf{Z}$  is a matrix consisting of 0’s and 1’s, and  $\mathbf{X}$  contains the covariate values. Model (6.9) is a general hurdle component in matrix form. The levels component

is an exponential model mimicking the lognormal mean. In matrix notation the general form of the levels component is,

$$E(\mathbf{Y} \mid \mathbf{Y} > \mathbf{0}) = \exp(\tilde{\mathbf{Z}}\boldsymbol{\alpha} + \tilde{\mathbf{X}}\boldsymbol{\beta}) \cdot \lambda, \quad (6.10)$$

where  $\tilde{\mathbf{Z}}$  is a matrix containing 0's and 1's,  $\tilde{\mathbf{X}}$  contains the covariate values and  $\lambda$  is the smearing factor estimated by the average of the exponentiated residuals  $\hat{\boldsymbol{\varepsilon}} = \log(\mathbf{Y}) - (\tilde{\mathbf{Z}}\hat{\boldsymbol{\alpha}} + \tilde{\mathbf{X}}\hat{\boldsymbol{\beta}})$  ([15],[44]).

Given the year and age, the expected death rate is the product of the two components (6.9) and (6.10). Zero death rates then are replaced by the estimates of the corresponding products.

Let

$$\begin{aligned} p &= P(Y_{\text{age, year}} > 0) \\ \alpha &\equiv E(Y_{\text{age, year}} \mid Y_{\text{age, year}} > 0) \end{aligned}$$

and let  $\hat{p}$  and  $\hat{\alpha}$  be their corresponding estimates. Then

$$\hat{E}(Y) = \hat{p} \cdot \hat{\alpha} = (\hat{p} - p + p)(\hat{\alpha} - \alpha + \alpha) \Rightarrow$$

$$\hat{p} \cdot \hat{\alpha} - p \cdot \alpha = p(\hat{\alpha} - \alpha) + \alpha(\hat{p} - p) + (\hat{p} - p)(\hat{\alpha} - \alpha)$$

For large samples  $\hat{p} \rightarrow p$  and  $\hat{\alpha} \rightarrow \alpha$  so the magnitude of  $\text{Var}((\hat{p} - p)(\hat{\alpha} - \alpha))$  is much smaller than that of  $\text{Var}(\hat{p} - p)$  and  $\text{Var}(\hat{\alpha} - \alpha)$ . If  $\hat{p}$ ,  $\hat{\alpha}$  independent, then

asymptotically:

$$\hat{\text{Var}}(\hat{p} \cdot \hat{\alpha}) \approx \hat{p}^2 \text{Var}(\hat{\alpha}) + \hat{\alpha}^2 \text{Var}(\hat{p}) \quad (6.11)$$

Approximate 95% confidence intervals can be calculated as

$$\hat{\text{E}}(Y) \pm 1.96 \sqrt{\hat{p}^2 \text{Var}(\hat{\alpha}) + \hat{\alpha}^2 \text{Var}(\hat{p})}.$$

In our data analysis, the smearing factor was very close to 1, and thus had no bearing either on  $\hat{\text{E}}(Y)$  or on the confidence intervals. It is interesting to compare this interval with that obtained formally from the mixed lognormal case using (6.8), replacing  $p, \mu, \sigma^2$  with their estimates under the two-part model.

### 6.2.2.3 Poisson regression

We assume that  $Y$ , the number of deaths, follows a Poisson distribution with some mean  $\mu$ :

$$f(y; \mu) = \frac{\exp(-\mu)\mu^y}{y!}, \quad y = 0, 1, 2, \dots$$

Using the Poisson GLM we model

$$g(\boldsymbol{\mu}) = \mathbf{Z}\boldsymbol{\alpha} + \mathbf{X}\boldsymbol{\beta},$$

where  $\mathbf{Z}$ ,  $\boldsymbol{\alpha}$ ,  $\mathbf{X}$  and  $\boldsymbol{\beta}$  are defined as in the two-part model, and  $\boldsymbol{\alpha}$  and  $\boldsymbol{\beta}$  are estimated using maximum likelihood. Observe that under the Poisson distribution the mean is equal to the variance. However count data often show greater variability



than is possible under the Poisson model. In this case a negative binomial GLM model is more appropriate. Then,

$$f(y; \mu, \theta) = \frac{\Gamma(y + \theta)}{\Gamma(\theta)y!} \frac{\mu^y \theta^\theta}{(\mu + \theta)^{y+\theta}}, \quad y = 0, 1, 2, \dots$$

which gives  $E(Y) = \mu$ ,  $\text{Var}(Y) = \mu + \mu^2/\theta$  and we let  $g(\boldsymbol{\mu}) = \mathbf{Z}\boldsymbol{\alpha} + \mathbf{X}\boldsymbol{\beta}$ . Regression analysis of count data is discussed in detail in [1],[9],[32], [45].

Standard errors and confidence intervals for the mean number of deaths can be obtained using the estimated Fisher information matrix. Confidence intervals for the mean death rate are obtained from the previous expressions by dividing the variances by the square of the population size for a given age and year.

#### 6.2.2.4 Hurdle model

Let  $Y$  denote the number of deaths. Hurdle models are two-component models, where a truncated count component is employed for positive counts, and a hurdle component models the zeros. Formally,

$$f_{hurdle}(y; \mu, \gamma) = \begin{cases} f_{zero}(0, \gamma), & y = 0 \\ \frac{(1-f_{zero}(0, \gamma)) \cdot f_{count}(y, \mu)}{(1-f_{count}(0, \mu))}, & y > 0 \end{cases} \quad (6.12)$$

Ordinarily, the count component is modeled as a left truncated Poisson distribution truncated at 0 and defined for  $y > 0$ , or, if there is overdispersion in the data, as a left truncated negative binomial distribution, again truncated at 0. Then, if  $\mu$  is the mean of the distribution, we use the GLM model  $g(\boldsymbol{\mu}) = \mathbf{Z}\boldsymbol{\alpha} + \mathbf{X}\boldsymbol{\beta}$  to

estimate  $\mu$ . The hurdle component is naturally modeled using a binomial GLM. The model parameters are estimated by maximum likelihood. Since the likelihood for the general hurdle model is the product of the likelihood of the zero component and the likelihood of the count component, the two components can be maximized separately using GLM theory [50].

Suppose we model the hurdle component using binomial GLM, such that  $1 - f_{zero} = p$ , and  $f_{count}$  is the Poisson distribution with mean  $\mu$ . Then

$$E(Y) = \frac{p\mu}{1 - \exp(-\mu)}.$$

If  $f_{count}$  is the negative binomial distribution with mean  $\mu$  and dispersion parameter  $\theta$ , then

$$E(Y) = \frac{p\mu}{1 - (\frac{\theta}{\mu+\theta})^\theta}.$$

From (6.1) we obtain the useful formula for the variance,

$$\text{Var}(Y) = P(Y > 0)\text{Var}(Y | Y > 0) + P(Y > 0)(1 - P(Y > 0))(E(Y | Y > 0))^2. \quad (6.13)$$

Let  $f_c(0) = f_{count}(0, \mu)$  and let  $\sigma^2$  be the variance of  $f_{count}(y, \mu)$ . Then, in the general hurdle model (6.12):

$$\text{Var}(Y) = p \left[ \frac{\sigma^2 + \mu^2}{1 - f_c(0)} - \frac{\mu^2}{(1 - f_c(0))^2} \right] + p(1 - p) \left( \frac{\mu}{1 - f_c(0)} \right)^2. \quad (6.14)$$

### 6.2.2.5 Zero-inflated model

Zero inflated models were first suggested by Lambert [41]. As before, they are also two-component models, except that a point mass at zero is now combined with a count distribution such as Poisson or negative binomial, which are supported at zero as well. Therefore, the probability of a zero is constructed from two sources: the point mass at zero and the count distribution. Let  $\pi \equiv 1 - p$  be the unobserved probability of belonging to the point mass component. Then the distribution of the number of deaths  $Y$  is modeled as

$$f_{zeroinfl}(y; \mu) = \pi I_{\{0\}}(y) + (1 - \pi)f_{count}(y; \mu), \quad y = 0, 1, 2, \dots \quad (6.15)$$

The probability  $\pi$  can be modeled using binomial GLM. Let  $\mu$  be the mean of the count distribution. Then we can use the GLM model  $g(\boldsymbol{\mu}) = \mathbf{Z}\boldsymbol{\alpha} + \mathbf{X}\boldsymbol{\beta}$  to estimate  $\mu$ . However it is difficult to estimate the model parameters directly from the log likelihood. For this reason Lambert proposed to use the EM algorithm [41], [83]. Assume  $Z_i = 1$  when  $Y_i$  is from the point mass component and  $Z_i = 0$  when  $Y_i$  is from the count component. Then at the E step we estimate  $Z_i$  by its expected value under the current estimates of the model parameters and at the M step, with  $Z_i$ 's fixed, we maximize the complete log likelihood. The process is iterated until both the estimates for the model parameters and for the  $Z_i$ 's converge. These are the MLE estimates for the zero-inflated model. In the case that the count component is Poisson distribution with mean  $\mu$ , the mean of  $Y$  is  $E(Y) = p\mu$ .

It should be noted that if there are only a few positive counts and  $\pi$  and  $\mu$

are not related then only simple models should be considered for the count component. The parameters of a zero-inflated model can be estimated only if the observed information matrix is nonsingular [41].

If the count component is Poisson distribution with mean  $\mu$  then directly from (6.15) the approximate 95% confidence interval for the mean number of deaths is

$$\hat{Y} \pm 1.96\sqrt{\hat{p}\hat{\mu} + \hat{p}(1 - \hat{p})\hat{\mu}^2}. \quad (6.16)$$

### 6.2.3 The Heligman-Pollard model

In 1980 Heligman and Pollard [30] proposed modeling the graduation of the age pattern of mortality using the eight parameter curve:

$$\frac{q_x}{1 - q_x} = A^{(x+B)^C} + De^{-E(\log x - \log F)^2} + GH^x, \quad (6.17)$$

where  $q_x$  is the probability of dying within one year for a person aged  $x$  exactly. All the parameters in (6.17) have demographic interpretations and are therefore considered to take non-negative values [25]: A, B, C reflect early age mortalities, D, E, F reflect mid-life mortality components, and G, H are late age mortality components. Let  $q_x = \frac{d_x}{n_x}$  be the death rate for a person aged  $x$ , where  $d_x$  is the number of people aged  $x$  who died at a certain year and  $n_x$  is the total number of people aged  $x$  who were alive at the beginning of the year. Then if we plot  $\log q_x$  as a function of age  $x$  for a certain year, we observe that the mortality curve has the

three components that Heligman and Pollard described in their paper and can be modeled using (6.17). Wei et al. [75] showed that the Heligman-Pollard model can be used to smooth the observed mortalities in race-sex U.S. sub-populations and provide estimates for the mortality in large age groups.

Heligman and Pollard suggested estimating the values of the parameters in (6.17) by least squares using Gauss-Newton iteration. We estimated the parameters using the following objective functions:

$$S^2 = \sum_x w_x (q_x - \dot{q}_x)^2, \text{ with } w_x = \frac{1}{\dot{q}_x^2} \quad (6.18a)$$

$$S^2 = \sum_x \left( \frac{q_x}{\dot{q}_x} - 1 \right)^2, \quad (6.18b)$$

$$S^2 = \sum_x (\log(q_x) - \log(\dot{q}_x))^2 \quad (6.18c)$$

where  $q_x$  is the fitted value at age  $x$  and  $\dot{q}_x$  is the observed mortality rate. In (6.18a) the weights are updated in each iteration. Since a large number of parameters has to be estimated, and the Gauss-Newton algorithm is sensitive to the starting values, we try minimizing (6.18a)-(6.18c) using a series of different starting values. The objective function (6.18a) is easy to fit, however, the estimated values for the parameters B, D, E are sometimes negative.

### 6.3 Data application

For our analysis, we used data from the NCHS public-use mortality files from 1970 to 2002. The models described above were applied to the population of black

females in California (CA), Iowa (IA), Minnesota (MN), Nevada (NV), New Mexico (NM), Nebraska (NE), Oregon (OR), and Rhode Island (RI) for the period 1970 – 2002. With the exception of California, life tables for 1989–91 were not published for the black population in these states. The population of black females in California contains only a small number of zeros and is used to check the model quality, whereas the other data sets contain a medium to large proportion of zeros, even for older ages. The largest proportions of zeros are observed in New Mexico and Nevada. For each state, the race-sex specific population size is available annually only for ages in five year intervals e.g. 0 – 4, 5 – 9 etc. The exact annual population size for every age and for race-sex specific groups is available at the national level only. Using interpolation we can obtain estimates of the race-sex specific sub-population size for every age at the state level. Therefore, for each state available are the population size, number of deaths, and death rate, corresponding to each age (1 – 84) and year (1970 – 2002) combination. Data for age 0 and for ages 85+ are also available but are excluded from the present analysis because they exhibit a much different behavior. In ANOVA type regressions, the “as factors” main effects are represented below by capital letters: Age, Year.

AIC, BIC as well as the root mean square error criterion’s were used to select the most appropriate models within the families of Poisson, hurdle, and zero-inflated models, for each state separately. For the family of two-part models the selection was based on root mean square error (RMSE) and mean absolute error (MAE). The mixed lognormal distribution was fitted to non overlapping windows as described in section 6.2.2.1. The levels component of the two-part models was fitted to non

Table 6.1: Covariates in the fitted two-part, Poisson, and negative binomial models for the indicated states.

State	Two-Part Model		Poisson Model	Neg. Binomial Model
	Hurdle Comp.	Levels Comp.		
CA	log(pop)	Age, Year	NA	Age, Year, log(pop)
IA	Age, Year	Age, Year	Age, Year	NA
MN	Age, Year, pop	Age, Year	Age, Year, log(pop)	Age, Year, log(pop)
NE	Age, Year, log(pop)	Age, Year	Age, Year, log(pop)	NA
NM	Age, Year, log(pop)	Age, Year	Age, Year, log(pop)	NA
NV	Age, Year, log(pop)	Age, Year	Age, Year, log(pop)	Age, Year, log(pop)
OR	Age, Year, log(pop)	Age, Year	Age, Year, offset=log(pop)	Age, Year, log(pop)
RI	Age, Year, log(pop)	Age, Year	Age, Year, log(pop)	Age, Year, log(pop)

Table 6.2: Covariates in the fitted hurdle and zero-inflated models for the indicated states. The count distribution was Poisson in all cases except for the hurdle model under California, where negative binomial was used.

State	Hurdle Model		Zero-Inflated Model	
	Hurdle Comp.	Count Comp.	Hurdle Comp.	Count Comp.
CA	log(pop)	Age, Year, log(pop)	NA	NA
IA	Age, Year, log(pop)	Age, Year, log(pop)	Year, log(pop)	Age, Year, log(pop)
MN	Age, Year, pop	Age, Year, log(pop)	constant	Age, Year, log(pop)
NE	Age, Year, log(pop)	Age, Year, log(pop)	Year	Age, Year, log(pop)
NM	Age, Year, log(pop)	Year, log(pop)	Age, Year	Year, log(pop)
NV	Age, Year, log(pop)	Age, Year, log(pop)	constant	Age, Year, log(pop)
OR	Age, Year, log(pop)	Age, Year, log(pop)	log(pop)	Age, Year, log(pop)
RI	Age, Year, log(pop)	Age, Year, log(pop)	log(pop)	Age, Year, log(pop)

zero observations only, and then the fitted model was applied in the estimation of the mean death rate in the cases where the observed value was zero. The hurdle component was modeled as a binomial GLM with logit link fitted to the whole data set. The expected death rate was then the product of the levels and the hurdle components. Poisson, hurdle, and zero-inflated models were fitted to the whole data set. In Poisson and negative binomial GLM's the log link was used. The hurdle component for the hurdle models was always modeled as binomial GLM with logit link. The count component for zero-inflated models was taken as Poisson with log link, and for the hurdle models it was either Poisson or negative binomial with log link. A zero-inflated model was not fitted to the California mortality data because the data displayed a very small number of zeros.

Table 6.3: Estimated expected values of  $\log(\text{death rates})$  provided by the different models for black females living in Nevada in 2000.

Age	Two-Part	Mixed Log.	Poisson	Neg. Bin.	Hurdle	Zero-Infl.
2	-6.94314	-7.24543	-7.08009	-7.08116	-6.94178	-7.07965
5	-7.34094	-8.14984	-7.92234	-7.92250	-7.59846	-7.92195
6	-8.21745	-8.14984	-8.69531	-8.69691	-8.36856	-8.69501
7	-7.82242	-8.96718	-8.70695	-8.70730	-8.04079	-8.70665
8	-7.98267	-8.96718	-8.69763	-8.69820	-8.21543	-8.69740
10	-8.74442	-9.06976	-9.58980	-9.59023	-9.18880	-9.58959
11	-8.83071	-9.96421	-9.55804	-9.55884	-9.20170	-9.55774
12	-7.74899	-9.96421	-8.42722	-8.42850	-8.24954	-8.42683
13	-7.67346	-7.89281	-8.10648	-8.10796	-7.99588	-8.10606
14	-7.87022	-7.89281	-8.24168	-8.24430	-8.12569	-8.24123
15	-7.67390	-8.12793	-7.97005	-7.97250	-7.88955	-7.96961
17	-7.20033	-7.17614	-7.40813	-7.41028	-7.26214	-7.40760
19	-7.18161	-7.29072	-7.30322	-7.30589	-7.21920	-7.30263
21	-6.88942	-6.86785	-6.93396	-6.93748	-6.90148	-6.93335
23	-7.11821	-7.18543	-7.14420	-7.14751	-7.15275	-7.14357
24	-7.19266	-7.18543	-7.34725	-7.35005	-7.25640	-7.34666
27	-6.88141	-6.95996	-6.91894	-6.92108	-6.92245	-6.91838

### 6.3.1 Selection of the models

In Section 6.3.2 it will be shown that simpler models perform sufficiently well. Tables 6.1 - 6.2 show the parameters which were used to fit each model for the indicated states. In Nevada, Oregon and Rhode Island, there was some evidence of overdispersion for the older age groups. For these states, we fitted hurdle models with negative binomial count distribution. However, the fitted points were almost identical to the points produced from hurdle models with Poisson count distribution.

Table 6.3 provides a comparison of the estimated expected values of  $\log(\text{death rates})$  obtained from the different models for black females living in Nevada in 2000, when the observed number of deaths is zero. The estimates show small differences.

Figures 6.1-6.8 show examples of some of the models fitted and the appropriate confidence intervals for the year 2000. The points on the  $x$ -axis correspond to the



ages where the observed death rates were 0. Despite the zeros, the age pattern of mortality is captured well. Mixed lognormal was applied to non overlapping windows of 11 years and two consecutive ages. The models give similar estimates in all states and capture the pointed hook pattern for lower ages. In California there are few zero death rates and the estimated points follow the actual points closely apart from ages 35 – 40 where there is some overestimation. In New Mexico during the period 1970-2002 the observed death rate for age 12 was always zero, i.e., there were no deaths documented for black females during this period. The mixed lognormal and two-part models could not produce expected values for age 12 and the hurdle and zero-inflated models could not be fitted with the factor Age in the count part because the observed information matrix was nonsingular. The expected values for age 12 produced by Poisson, hurdle and zero-inflated model are indeed much lower than the other data points. For this data set the simplest model to fit, the Poisson model, performs very well, as seen from figure 6.5.

In some states, the observed nonzero death rates for young ages (up to 25 years) are located higher than the predicted points. These are inflated points caused by the small size of the population and the fact that the observed number of deaths can be either zero or an integer number, not a fraction.

In Section 6.2 it was shown how to obtain asymptotic or approximate confidence intervals for the models. These intervals are in general quite wide, and since the death rate (or the corresponding number of deaths) can take only nonnegative values, the lower 95% confidence interval is often zero. For this reason we also use parametric bootstrap to obtain confidence intervals as follows:

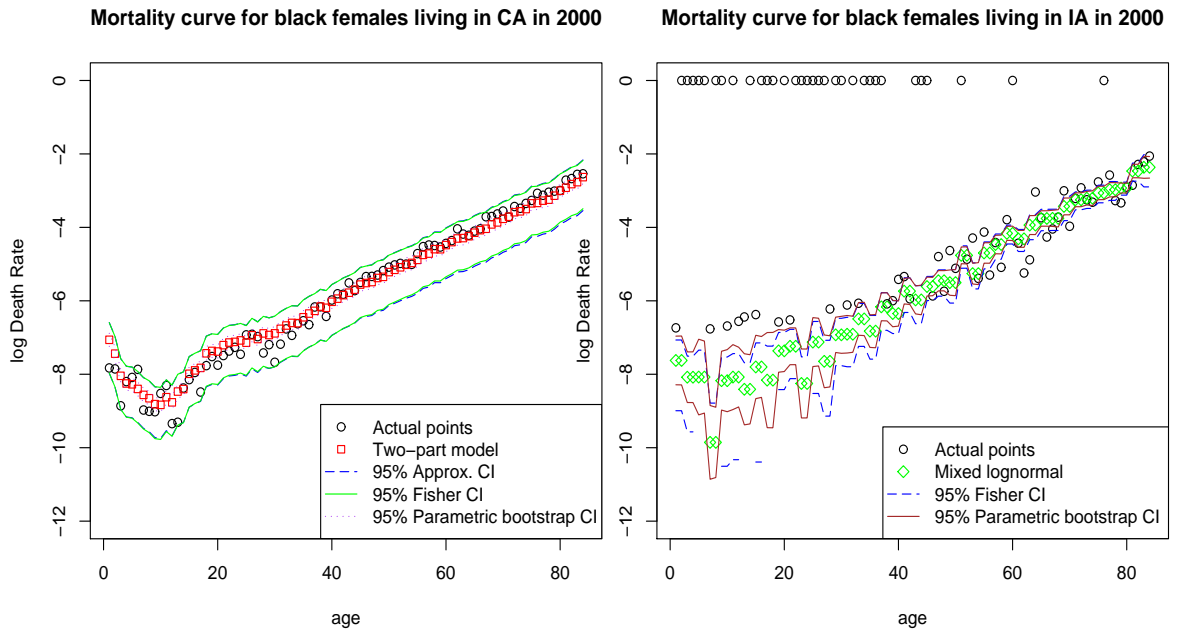


Figure 6.1: Two-part model: CA, 2000    Figure 6.2: Mixed lognormal: IA, 2000

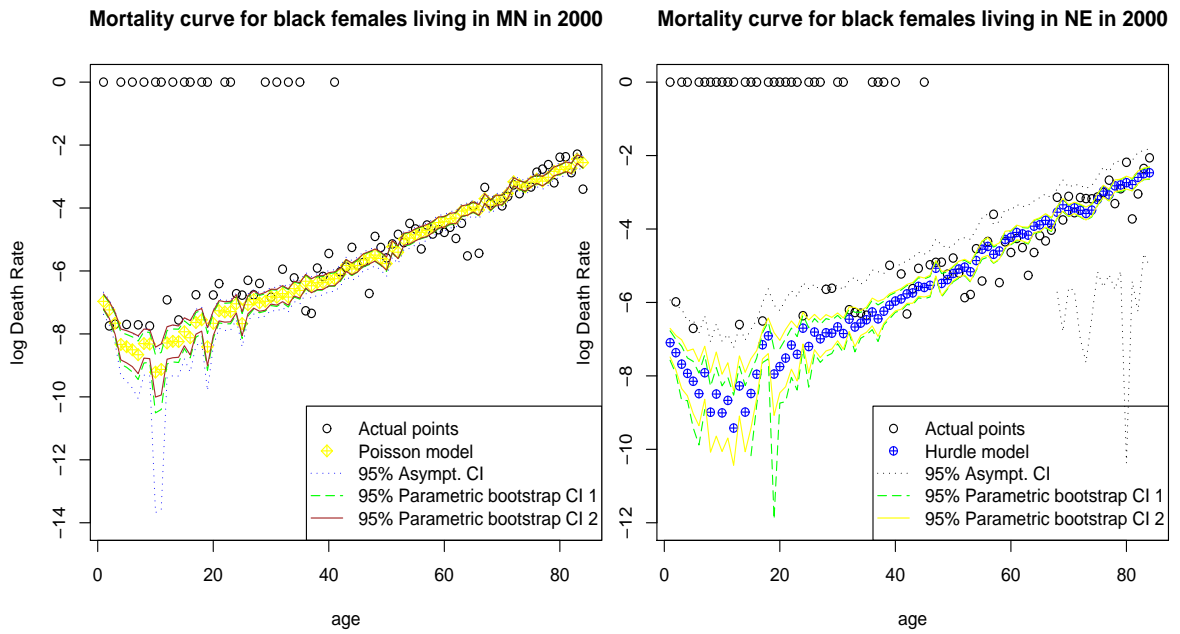


Figure 6.3: Poisson model: MN, 2000    Figure 6.4: Hurdle model: NE, 2000

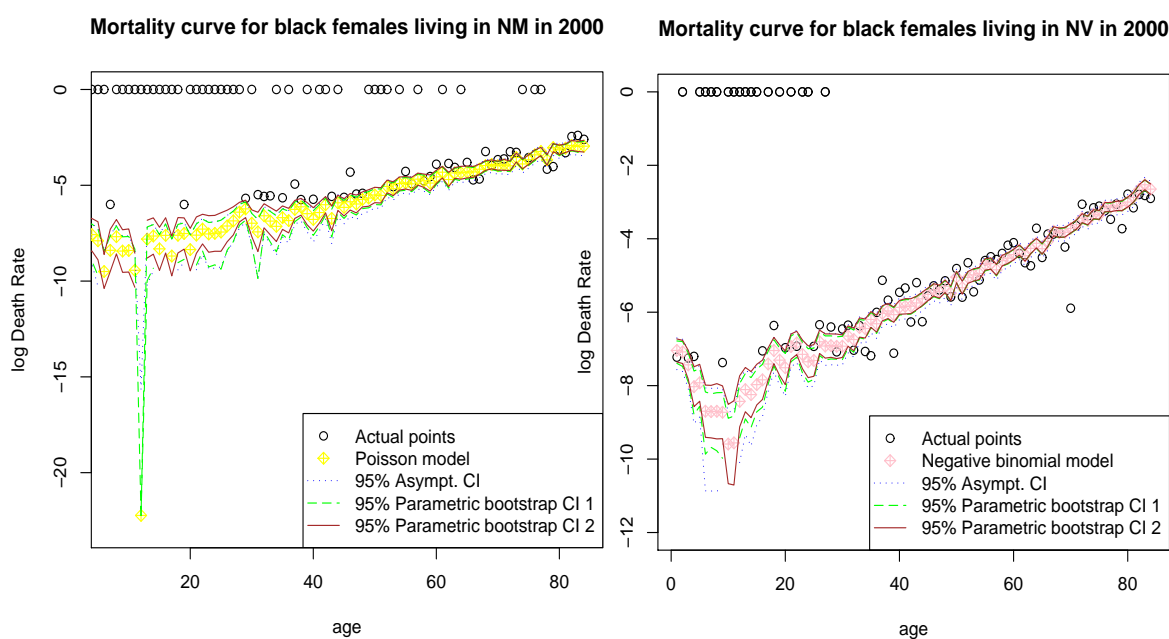


Figure 6.5: Poisson model: NM, 2000    Figure 6.6: Neg. binomial model: NV, 2000

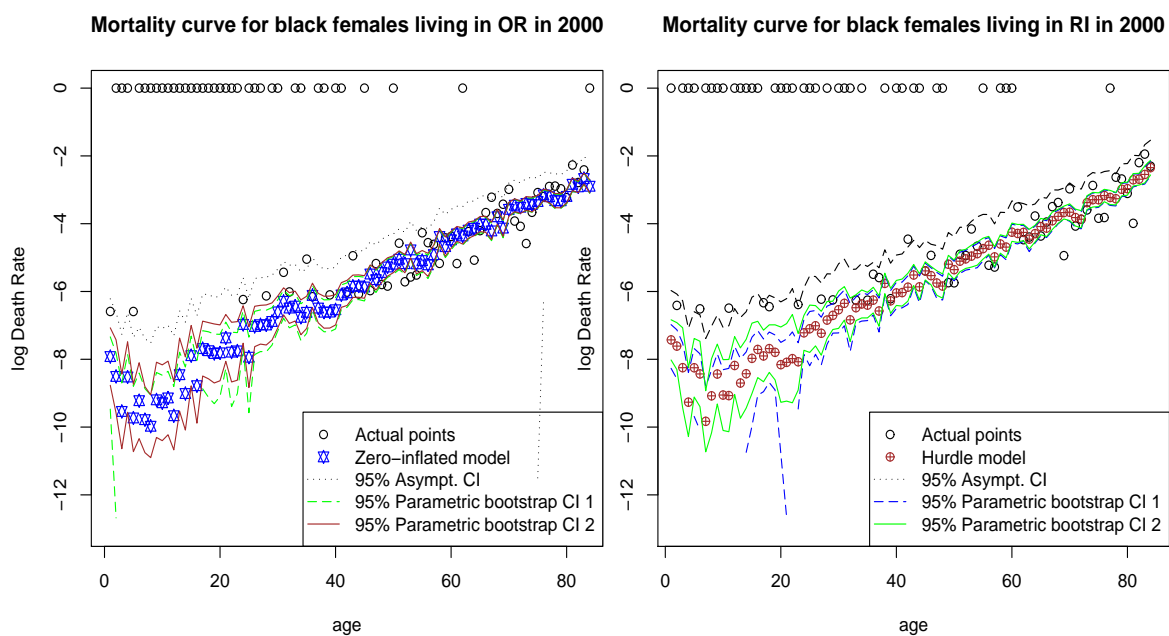


Figure 6.7: Zero-inflated model: OR, 2000    Figure 6.8: Hurdle model: RI, 2000

Suppose that the observations  $Y$  follow distribution  $f_\theta(y)$  and  $\hat{\theta}$  is the ML estimator of  $\theta$ . We draw  $B$  samples of size  $n$  from the density  $f_{\hat{\theta}}$ . For each sample we calculate the maximum likelihood estimate of  $\theta$  and we estimate  $E(Y)$ . The sample variance of these  $B$  values of  $\hat{E}(y)$  estimates the variance of  $\hat{E}(y)$ . A standard reference for the bootstrap is [16]. In practice we construct parametric bootstrap confidence intervals by selecting 1000 samples of size  $n = 33$ .

For the two-part models and the mixed lognormal distribution, parametric bootstrap is used to calculate 95% confidence intervals for the  $\log(\text{death rate})$ . For the Poisson, hurdle and zero inflated models, parametric bootstrap is performed on both the mean number of deaths and also on the corresponding  $\log(\text{death rate})$ . By performing bootstrap on  $\log(\text{death rate})$  directly we manage to calculate a lower 95% confidence interval. Obtaining bootstrap confidence intervals for  $\log(\text{death rates})$  from bootstrap confidence intervals for the mean number of deaths proves to be problematic as the lower confidence limit is often  $-\infty$ , especially for young ages. In Figures 6.1-6.8 parametric bootstrap 1 refers to the case where confidence intervals were computed on the mean number of deaths and then using suitable transformations we managed to calculate confidence intervals for  $\log(\text{death rate})$ . Parametric bootstrap 2 refers to the case where confidence intervals were computed directly on  $\log(\text{death rate})$ . An indication that the values selected for  $B$  and  $n$  work well comes from the fact that the upper Fisher and the upper parametric bootstrap confidence interval in mixed lognormal almost coincide. This is also the case for the upper asymptotic and the upper parametric bootstrap confidence interval in Poisson. In Poisson, hurdle, and zero-inflated models the two parametric bootstrap

Table 6.4: California RMSE and MAE. Black females, ages 1-84, period 1970-2002.

CA	RMSE				MAE			
	Total	1-30 yrs	31-50 yrs	51-84 yrs	Total	1-30 yrs	31-50 yrs	51-84 yrs
Mixed log.	3.401	0.242	0.700	5.314	1.493	0.180	0.516	3.226
Two-part	4.449	0.249	0.686	6.969	1.912	0.180	0.503	4.269
Neg. Bin.	3.520	0.270	0.643	5.505	1.530	0.195	0.474	3.328
Hurdle	3.519	0.270	0.643	5.503	1.529	0.195	0.473	3.328

Table 6.5: Iowa RMSE and MAE. Black females, ages 1-84, period 1970-2002.

IA	RMSE				MAE			
	Total	1-30 yrs	31-50 yrs	51-84 yrs	Total	1-30 yrs	31-50 yrs	51-84 yrs
Mixed log.	16.696	1.251	3.962	26.040	7.713	0.838	2.901	16.609
Two-part	16.656	1.276	4.023	25.971	7.681	0.877	2.967	16.459
Poisson	15.410	1.277	4.002	23.996	7.357	0.869	2.965	15.665
Hurdle	15.543	1.275	4.018	24.207	7.365	0.858	2.952	15.702
Zero-infl.	15.613	1.275	4.011	24.318	7.380	0.859	2.964	15.733

confidence intervals are very similar, but the second parametric bootstrap produces lower confidence intervals even for young ages.

For the two part model, equation (6.11) gives a crude approximation for the variance when  $p$  and  $\alpha$  are independent. We can sidestep the assumption of independence by constructing parametric bootstrap confidence intervals. In retrospect, (6.11) gives confidence intervals very close to the Fisher confidence intervals, but somewhat larger than the confidence intervals obtained from bootstrap. The Fisher confidence interval is wide and contains most of the actual points (observations). This generally holds for the asymptotic and approximate confidence intervals produced from all the models. In the hurdle model the lower 95% approximate confidence interval is almost always missing, whereas the upper 95% approximate confidence interval is quite large for some ages. In the zero inflated model the lower 95% approximate confidence interval is also missing and the upper 95% approximate confidence interval is large for some ages, but not as wide as in the hurdle model.

Table 6.6: Minnesota RMSE and MAE. Black females, ages 1-84, period 1970-2002.

MN	RMSE				MAE			
	Total	1-30 yrs	31-50 yrs	51-84 yrs	Total	1-30 yrs	31-50 yrs	51-84 yrs
Mixed log.	14.203	0.944	3.111	22.179	6.614	0.684	2.286	14.394
Two-part	14.407	0.944	3.135	22.553	6.698	0.653	2.281	14.631
Poisson	13.630	0.937	3.134	21.270	6.386	0.661	2.301	13.840
Neg. Bin.	13.630	0.937	3.134	21.270	6.386	0.661	2.301	13.840
Hurdle	13.643	0.930	3.143	21.291	6.350	0.650	2.299	13.762
Zero-infl.	13.629	0.937	3.134	21.269	6.386	0.661	2.301	13.840

Table 6.7: Nebraska RMSE and MAE. Black females, ages 1-84, period 1970-2002.

NE	RMSE				MAE			
	Total	1-30 yrs	31-50 yrs	51-84 yrs	Total	1-30 yrs	31-50 yrs	51-84 yrs
Mixed log.	<i>NA</i>	<i>NA</i>	3.477	24.266	<i>NA</i>	<i>NA</i>	2.590	15.647
Two-part	15.357	1.135	3.485	23.966	7.243	0.844	2.576	15.635
Poisson	14.961	1.133	3.492	23.339	7.112	0.845	2.596	15.298
Hurdle	14.985	1.130	3.484	23.378	7.090	0.832	2.590	15.258
Zero-infl.	14.963	1.133	3.491	23.341	7.110	0.845	2.595	15.293

When the amount of information is substantial (few points with zero death rate) as in California, the confidence intervals contain most of the actual points.

### 6.3.2 Model comparison

Tables 6.4 to 6.11 provide model comparison in terms of root mean-square error (RMSE) and mean absolute error (MAE) computed for death rate and broken down by period, state, model, and age group 1 – 30, 31 – 50 and 51 – 84 years. In tables 6.9- 6.11, NA in the mixed lognormal or the two-part model refers to cases where the RMSE and MAE were not calculated. In these cases, some of the samples consisted only of zeros and therefore no estimators could be produced by these models for these ages. Since the models are based on ramifications of a basic mixed distribution with an atom at zero, there are small differences between the models as judged by overall RMSE and MAE for the indicated age grouping and period. Thus, we cannot point to a clear winner among the models. However, for

Table 6.8: New Mexico RMSE and MAE. Black females, age 1-84, period 1970-2002.

NM	RMSE				MAE			
	Total	1-30 yrs	31-50 yrs	51-84 yrs	Total	1-30 yrs	31-50 yrs	51-84 yrs
Mixed log.	<i>NA</i>	<i>NA</i>	4.495	31.459	<i>NA</i>	<i>NA</i>	3.431	19.883
Two-part	<i>NA</i>	<i>NA</i>	4.515	32.408	<i>NA</i>	<i>NA</i>	3.372	20.197
Poisson	20.474	1.831	4.503	31.949	9.315	1.116	3.427	20.013
Hurdle	21.920	1.840	4.523	34.235	9.698	1.161	3.511	20.871
Zero-infl.	22.628	1.832	4.677	35.343	9.926	1.173	3.688	21.319

Table 6.9: Nevada RMSE and MAE. Black females, age 1-84, period 1970-2002.

NV	RMSE				MAE			
	Total	1-30 yrs	31-50 yrs	51-84 yrs	Total	1-30 yrs	31-50 yrs	51-84 yrs
Mixed log.	<i>NA</i>	<i>NA</i>	3.582	31.442	<i>NA</i>	<i>NA</i>	2.559	18.615
Two-part	20.055	1.142	3.505	31.390	8.351	0.792	2.510	18.455
Poisson	20.165	1.144	3.512	31.563	8.350	0.800	2.535	18.432
Neg. Bin.	20.153	1.143	3.511	31.543	8.348	0.800	2.535	18.427
Hurdle	19.900	1.139	3.488	31.146	8.234	0.785	2.502	18.177
Zero-infl.	19.898	1.139	3.488	31.142	8.233	0.785	2.502	18.177

some individual years there can be appreciable differences in the estimated mean death rates from the different models. To simplify the presentation of the numerical results, all entries in the tables are multiples of  $10^{-3}$ .

It is interesting to observe that the mixed lognormal model, which requires no covariates, performs well when there is a large proportion of non-zero observations. It is the best model in California, a data set with few zeros, and it performs consistently well in the windows 31 – 50 and 51 – 84 for most of the other data sets, where the percentage of zeros decreases rapidly. On the other hand, the two-part model, although very similar to mixed lognormal and equally easy to fit, performs consistently somewhat worse (“below average” as it were) than the rest of the models. When the number of zeros is not too high, the Poisson GLM is quite adequate. It has the advantage that it is very easy to fit and obtain confidence intervals for the mean number of deaths. The hurdle model often offers a small improvement, especially for data sets which contain many zero observations, but computation-

Table 6.10: Oregon RMSE and MAE. Black females, age 1-84, period 1970-2002.

OR	RMSE				MAE			
	Total	1-30 yrs	31-50 yrs	51-84 yrs	Total	1-30 yrs	31-50 yrs	51-84 yrs
Mixed log.	<i>NA</i>	<i>NA</i>	4.178	25.522	<i>NA</i>	<i>NA</i>	3.163	16.358
Two-part	16.837	1.286	4.226	26.238	7.784	0.877	3.071	16.651
Poisson	16.698	1.298	4.199	26.019	7.755	0.870	3.131	16.550
Neg. Bin.	16.704	1.299	4.201	26.028	7.759	0.871	3.137	16.557
Hurdle	16.682	1.283	4.193	25.994	7.702	0.856	3.073	16.467
Zero-infl.	16.665	1.294	4.202	25.967	7.749	0.879	3.146	16.519

Table 6.11: Rhode Island RMSE and MAE. Black females, age 1-84, period 1970-2002.

RI	RMSE				MAE			
	Total	1-30 yrs	31-50 yrs	51-84 yrs	Total	1-30 yrs	31-50 yrs	51-84 yrs
Mixed log.	<i>NA</i>	<i>NA</i>	4.724	27.252	<i>NA</i>	<i>NA</i>	3.313	17.380
Two-part	17.916	1.403	4.691	27.899	8.184	0.898	3.193	17.549
Poisson	17.496	1.402	4.683	27.233	8.149	0.893	3.305	17.401
Neg. Bin.	17.499	1.402	4.683	27.238	8.150	0.893	3.304	17.403
Hurdle	17.577	1.397	4.693	27.361	8.115	0.877	3.208	17.388
Zero-infl.	17.447	1.400	4.681	27.155	8.129	0.901	3.308	17.342

ally it is time consuming. Moreover, in the presence of a large number of zeros, the variance covariance matrix of the estimates has often missing values. In this case confidence intervals could only be calculated using approximate methods or parametric bootstrap. Zero-inflated models are appropriate when there is an excess number of zeros, except that computational difficulties are encountered when the observed information matrix is singular or close to being one, as was often the case with our data. Our experience indicates that in general no appreciable improvement is achieved over the Poisson or the hurdle models.

It should be noted that, fitting the models required a certain amount of experimentation, and that in the tables we report the results corresponding to successful fits.



**California:** From Table 6.4, in terms of RMSE and MAE, apparently all the models perform quite similarly for the California data set. However, in some age categories the two-part model gives slightly higher RMSE and MAE.

**Iowa:** From Table 6.5, notwithstanding the overall slight advantage of the Poisson and the hurdle models, and the slight advantage of the mixed lognormal model in the age categories 1-30 and 31-50, the models perform similarly.

**Minnesota:** From Table 6.6, the MAE points to the hurdle model as a slightly better model, however, the models' performance is practically the same.

**Nebraska:** From Table 6.7, the Poisson and the hurdle models perform well in this data set. The NA's in the mixed lognormal refer to cases where RMSE/MAE could not be calculated because one or more of the samples consisted entirely of zero observations.

**New Mexico:** From Table 6.8, the Poisson model fits the data quite well. Another possible model is mixed lognormal for the window 51-84 yrs.

**Nevada:** From Table 6.9, hurdle and zero-inflated models have smaller RMSE and MAE than the rest of the models. In the 31-50 and 51-84 window the two-part model performs also well.

**Oregon:** From Table 6.10, we observe that the Poisson, hurdle, and zero-inflated models perform well for young ages. For older ages, where more nonzero observations

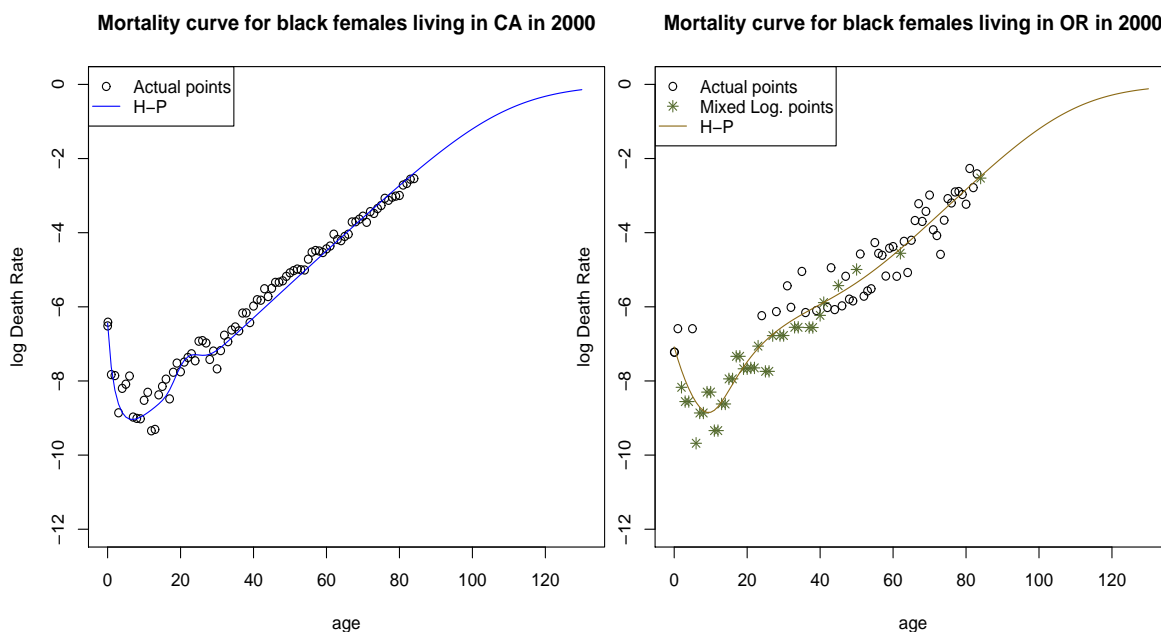


Figure 6.9: H-P curve: CA, 2000

Figure 6.10: H-P curve: OR, 2000

are available, the mixed lognormal model is quite adequate.

**Rhode Island:** As before, from Table 6.11, the Poisson, hurdle, and zero-inflated models perform reasonably well.

### 6.3.3 Fitting the Heligman-Pollard model

The next step was to fit the Heligman-Pollard (H-P) model to the death rates for black females living in California, Iowa, Minnesota, Nebraska, New Mexico, Nevada, Oregon and Rhode Island for the year 2000. The points  $\hat{q}_x$  used to fit the curve were the actual non zero observed death rates and, when the observed death rates were zero, the predicted death rates from one of the probability models mentioned above for ages 1 week, 1 month and 1-84 years. There was no restriction

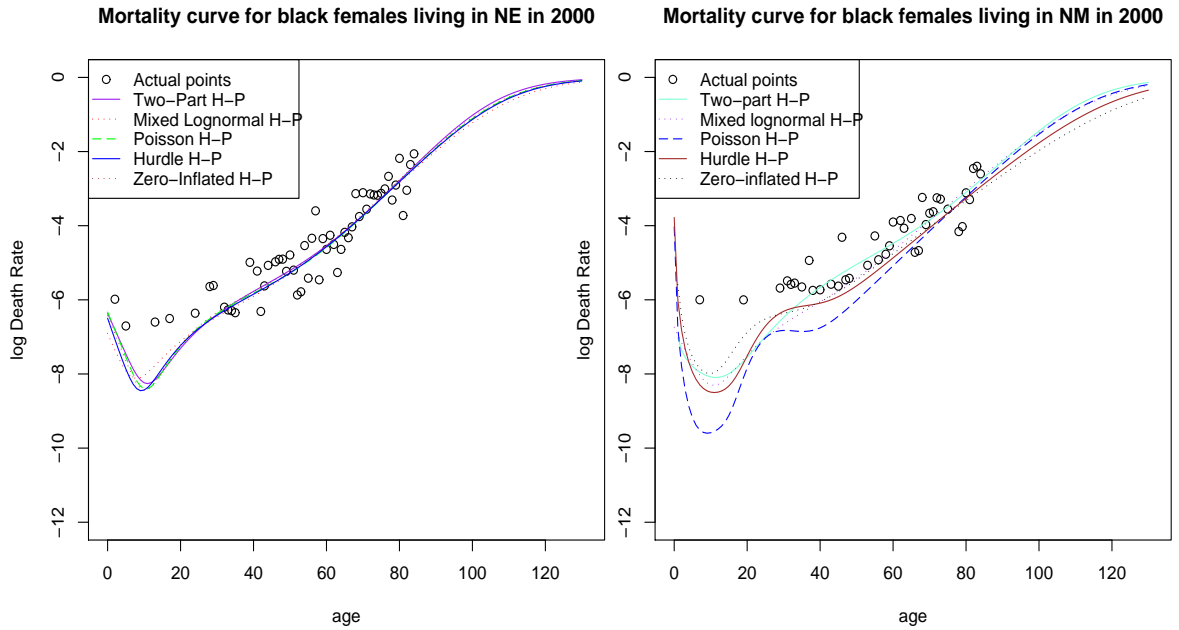


Figure 6.11: Comparison of H-P curves: NE, Figure 6.12: Comparison of H-P curves: NM, 2000

on the number of iterations but the minimum step-size factor allowed on any step in the iteration was restricted to be at least  $1/10,000$ . In the vast majority of cases the iteration stopped before it converged. In each state a total of five or six curves were fitted, depending on whether a negative binomial model was fitted to the data. We tried to minimize the objective functions (6.18a)-(6.18c). Since each objective function gives different estimates for the parameters  $A$ ,  $B$ ,  $C$ ,  $D$ ,  $E$ ,  $F$ ,  $G$  and  $H$ , we selected the set of parameters that had the lowest MSE (calculated on the death rates). The set of initial values plays a significant role in the fitting of the H-P model because the number of parameters is large and the parameters are dependent on each other. The initial values used were provided by Wei et al (2003) [75]. We also tried to use the final parameter values obtained by (6.18a) as initial values

in minimizing (6.18b), (6.18c). We did the same for the final values obtained by (6.18b).

The parameters obtained by minimizing (6.18b) lead consistently to a curve that doesn't fit as well as the curves obtained by minimizing (6.18a), (6.18c). The algorithm also makes fewer iterations. The best curve is most often given by minimizing (6.18c). Minimizing the objective function (6.18c) has the additional advantage that the estimated parameter values were almost always positive.

Figures 6.9-6.10 show two examples of the curves fitted. In 2000 in California, where the population of black females is large, deaths were observed for all ages so only the actual points were used to fit the H-P curve. The mortality points are more concentrated and there is not much variation. The estimated H-P curve follows the points closely and capture the pattern of mortality during early childhood and at older ages, as well as the "accident hump". On the other hand, in 2000, in Oregon there were very few deaths observed for the ages 0 – 30. The population of black females for this year ranged approximately from under one hundred up to eight hundred. We observe that the mixed lognormal managed to "reconstruct" all the missing part, and then the fitted H-P curve smoothed the data.

Since there are small differences between the probability models, the different H-P curves fitted for the same state are very close. The most important differences are observed in the drop in mortality during early childhood (e.g. in Oregon, New Mexico, Minnesota and Rhode Island). For example, consider figures (6.11), (6.12). New Mexico is a state where the population of black females ranges between 7 – 547 and therefore only a very small number of deaths is observed. During the period

1970-2002, only one death for age 11 and no deaths for age 12 are observed. The probability models gave different predictions for these ages, which affected in turn the shape of the curve. However all the curves did capture the distinct components of mortality.

In Nebraska, New Mexico and Rhode Island the mixed lognormal model failed to produce estimates for some ages. In this case we still use the estimates obtained for other ages and fitted the H-P model. When suitable estimates of the parameters are found, the H-P model is used to produce estimates for all ages between 0 and 130.

#### 6.4 Discussion and Conclusions

We have applied a variety of probability models to mortality data from eight states with small populations to compensate for zero observed death rates. Most of the models are based on mixed distributions. In general it is difficult to select a superior model. A model may be best for a certain age group only to be outperformed by another model at another age group. Mixed lognormal models with non-overlapping windows use subsets of the data to produce estimates, whereas the rest of the models are fitted using all available data. Generally, the mixed lognormal model is easy to fit, but it performs well only when there is sufficient non-zero data available, whereas the hurdle and zero-inflated models “thrive” when there are many zero observations. The Poisson model was found useful in all cases. All this leads to the practical conclusion that, whenever possible, it is sensible to apply routinely

all the above models.

The size of the population can be an indication of which model would be more appropriate to apply to the data. In states with extremely small subpopulations (less than 1,000 for fixed year and age) there is an abundance of zero observations; so this is an indication that more complicated models (hurdle, zero-inflated) would fit the data better. However, if we are interested in estimating death rates for older ages (51–84 yrs), a mixed lognormal model is easier to fit and gives equally good results. For larger subpopulations, provided that the samples in the non-overlapping windows contain both zero and nonzero observations, it is worth considering the mixed lognormal model as a plausible model. Regardless of the size of the subpopulation, the two-part model is usually outperformed by the other models, whereas the Poisson model is robust and can be used to compare results with the other models.

For each model confidence intervals were constructed using asymptotic methods and parametric bootstrap. From the figures, and in particular from the tables, it is seen that for the most part the application of the different models yielded very similar results, except for some individual years where the models produced very different estimates.

From the figures, our probability models capture well both the time and the age trends, and provide results consistent with the three basic characteristics of mortality curves. Therefore, in some cases these models can help to relax the minimum sample size requirements in the publication of reliable state race-sex specific life tables.

The probability models have to capture three distinct components of mortality: the fall in mortality during early childhood, the “accident hump” between ages

15 and 30 which is a distinct hump reflecting accident mortality, and the gradual rise in mortality at older ages. The eight-parameter, non-linear Heligman-Pollard equation can be used to describe the age pattern of mortality. In our application, the Heligman-Pollard model was used to smooth the data and predict/extrapolate mortality rates for older age groups. The curves were fitted to the observed nonzero death rates and (in case of zero mortality) the estimated expected death rates for each state. Usually the fitted curves were very similar regardless of the probability models used.

Frequently, the algorithm for the estimation of the parameters in the Heligman-Pollard equation did not converge. In some cases, the fitted curves did not capture the vast drop in mortality in young ages, as it is seen by the estimated points. These problems were caused by the fluctuation and the inflation of the observed death rates. The fluctuation in the observed death rates is inversely related to the size of the subpopulation. For small subpopulations there is noticeable variation in the observed death rates, whereas the observed nonzero death rates for young ages are often inflated.

Life tables are one of the most important products of the National Center for Health Statistics (NCHS). They summarize mortality patterns and characteristics, and as such, they have numerous actuarial applications. They also are widely used in the formulation of public health policies. Their publication on state and national levels is therefore crucial. In the previous methodology used to generate State subpopulation tables, a subjective and labor intensive procedure limited the reliability of death rate estimation, resulting in the non-publication of one fifth of

the subpopulations with total deaths fewer than 700. Based on the work presented in this article, we recommend a two-stage estimating/smoothing procedure: Firstly apply a suitable probability model on the data to get an estimate of the mortality rate for all ages. Secondly apply the Heligman-Pollard equation on the estimated data to obtain parameter estimates and smooth the mortality curve, covering the whole life span. In addition, the confidence interval from both stages could be used to establish a criterion for publication of final life tables. This new methodology will not only raise the reliability of estimation, but will also permit more efficient, repeatable, and comparable results in generating US life tables.



## Chapter A

### Appendix

#### A.1 Computing $\mathbf{W}$

Fokianos [19] utilized a Taylor expansion and the central limit theorem to show Theorem 2.1. Then,  $\mathbf{W} = \mathbf{D}^{-1}\mathbf{S}\mathbf{D}'^{-1}$ , where  $S = \text{var} \left( \frac{1}{\sqrt{n}} \begin{pmatrix} \frac{\partial l(\boldsymbol{\theta}_0, \boldsymbol{\zeta})}{\partial \boldsymbol{\theta}} \\ \frac{\partial l(\boldsymbol{\theta}_0, \boldsymbol{\zeta})}{\partial \boldsymbol{\mu}} \end{pmatrix} \right) = \begin{pmatrix} S_{11} & S_{12} \\ S'_{12} & S_{22} \end{pmatrix}$  is a  $(d+q) \times (d+q)$  matrix with  $S_{11} = \text{var} \left( \frac{1}{\sqrt{n}} \frac{\partial l(\boldsymbol{\theta}_0, \boldsymbol{\zeta})}{\partial \boldsymbol{\theta}} \right)$ ,  $S_{12} = \text{Cov} \left( \frac{1}{\sqrt{n}} \frac{\partial l(\boldsymbol{\theta}_0, \boldsymbol{\zeta})}{\partial \boldsymbol{\theta}}, \frac{1}{\sqrt{n}} \frac{\partial l(\boldsymbol{\theta}_0, \boldsymbol{\zeta})}{\partial \boldsymbol{\mu}} \right)$  and  $S_{22} = \text{var} \left( \frac{1}{\sqrt{n}} \frac{\partial l(\boldsymbol{\theta}_0, \boldsymbol{\zeta})}{\partial \boldsymbol{\mu}} \right)$ . For  $l = 1, \dots, q$ , the elements of  $S_{11}$  are calculated as follows:

- for  $l = l'$

$$\begin{aligned}
 \text{var} \left( \frac{1}{\sqrt{n}} \frac{\partial l(\boldsymbol{\theta}_0, \boldsymbol{\zeta})}{\partial \boldsymbol{\theta}_l} \right) &= \frac{1}{n} \text{var} \left( - \sum_{i=1}^m \sum_{j=1}^{n_i} \frac{\zeta_l \frac{\partial w(\mathbf{x}_{ij}, \boldsymbol{\theta}_{0l})}{\partial \boldsymbol{\theta}_l}}{\sum_{k=1}^m \zeta_k w(\mathbf{x}_{ij}, \boldsymbol{\theta}_{0k})} + \sum_{j=1}^{n_l} \frac{\partial w(\mathbf{x}_{lj}, \boldsymbol{\theta}_{0l})}{\partial \boldsymbol{\theta}_l} \right) \\
 &= \frac{1}{n} \sum_{i=1}^m n_i \text{var}_i \left( \frac{\zeta_l \frac{\partial w(\mathbf{x}_{ij}, \boldsymbol{\theta}_{0l})}{\partial \boldsymbol{\theta}_l}}{\sum_{k=1}^m \zeta_k w(\mathbf{x}_{ij}, \boldsymbol{\theta}_{0k})} \right) + \frac{n_l}{n} \text{var}_l \left( \frac{\partial w(\mathbf{x}_{lj}, \boldsymbol{\theta}_{0l})}{\partial \boldsymbol{\theta}_l} w(\mathbf{x}_{lj}, \boldsymbol{\theta}_{0l}) \right) \\
 &\quad - \frac{2n_l}{n} \text{Cov}_l \left( \frac{\zeta_l \frac{\partial w(\mathbf{x}_{ij}, \boldsymbol{\theta}_{0l})}{\partial \boldsymbol{\theta}_l}}{\sum_{k=1}^m \zeta_k w(\mathbf{x}_{ij}, \boldsymbol{\theta}_{0k})}, \frac{\partial w(\mathbf{x}_{lj}, \boldsymbol{\theta}_{0l})}{\partial \boldsymbol{\theta}_l} \right) \\
 &= \sum_{i=1}^m \zeta_i \zeta_l^2 \text{var}_i \left( \frac{\frac{\partial w(\mathbf{x}_{ij}, \boldsymbol{\theta}_{0l})}{\partial \boldsymbol{\theta}_l}}{\sum_{k=1}^m \zeta_k w(\mathbf{x}_{ij}, \boldsymbol{\theta}_{0k})} \right) + \zeta_l \text{var}_l \left( \frac{\partial w(\mathbf{x}_{lj}, \boldsymbol{\theta}_{0l})}{\partial \boldsymbol{\theta}_l} w(\mathbf{x}_{lj}, \boldsymbol{\theta}_{0l}) \right) \\
 &\quad - 2\zeta_l^2 \text{Cov}_l \left( \frac{\zeta_l \frac{\partial w(\mathbf{x}_{ij}, \boldsymbol{\theta}_{0l})}{\partial \boldsymbol{\theta}_l}}{\sum_{k=1}^m \zeta_k w(\mathbf{x}_{ij}, \boldsymbol{\theta}_{0k})}, \frac{\partial w(\mathbf{x}_{lj}, \boldsymbol{\theta}_{0l})}{\partial \boldsymbol{\theta}_l} \right)
 \end{aligned}$$

- for  $l \neq l'$

$$\begin{aligned}
& \text{Cov} \left( \frac{1}{\sqrt{n}} \frac{\partial l(\boldsymbol{\theta}_0, \boldsymbol{\zeta})}{\partial \boldsymbol{\theta}_l}, \frac{1}{\sqrt{n}} \frac{\partial l(\boldsymbol{\theta}_0, \boldsymbol{\zeta})}{\partial \boldsymbol{\theta}_{l'}} \right) = \\
&= \frac{1}{n} \text{Cov} \left( - \sum_{i=1}^m \sum_{j=1}^{n_i} \frac{\zeta_l \frac{\partial w(\mathbf{x}_{ij}, \boldsymbol{\theta}_{0l})}{\partial \boldsymbol{\theta}_l}}{\sum_{k=1}^m \zeta_k w(\mathbf{x}_{ij}, \boldsymbol{\theta}_{0k})} + \sum_{j=1}^{n_l} \frac{\frac{\partial w(\mathbf{x}_{lj}, \boldsymbol{\theta}_{0l})}{\partial \boldsymbol{\theta}_l}}{w(\mathbf{x}_{lj}, \boldsymbol{\theta}_{0l})}, \right. \\
&\quad \left. - \sum_{i=1}^m \sum_{j=1}^{n_i} \frac{\zeta_{l'} \frac{\partial w(\mathbf{x}_{ij}, \boldsymbol{\theta}_{0l'})}{\partial \boldsymbol{\theta}_{l'}}}{\sum_{k=1}^m \zeta_k w(\mathbf{x}_{ij}, \boldsymbol{\theta}_{0k})} + \sum_{j=1}^{n_{l'}} \frac{\frac{\partial w(\mathbf{x}_{l'j}, \boldsymbol{\theta}_{0l'})}{\partial \boldsymbol{\theta}_{l'}}}{w(\mathbf{x}_{l'j}, \boldsymbol{\theta}_{0l'})} \right) \\
&= \frac{1}{n} \sum_{i=1}^m n_i \text{Cov}_i \left( \frac{\zeta_l \frac{\partial w(\mathbf{x}, \boldsymbol{\theta}_{0l})}{\partial \boldsymbol{\theta}_l}}{\sum_{k=1}^m \zeta_k w(\mathbf{x}, \boldsymbol{\theta}_{0k})}, \frac{\zeta_{l'} \frac{\partial w(\mathbf{x}, \boldsymbol{\theta}_{0l'})}{\partial \boldsymbol{\theta}_{l'}}}{\sum_{k=1}^m \zeta_k w(\mathbf{x}, \boldsymbol{\theta}_{0k})} \right) \\
&\quad - \frac{n_{l'}}{n} \text{Cov}_{l'} \left( \frac{\zeta_l \frac{\partial w(\mathbf{x}, \boldsymbol{\theta}_{0l})}{\partial \boldsymbol{\theta}_l}}{\sum_{k=1}^m \zeta_k w(\mathbf{x}, \boldsymbol{\theta}_{0k})}, \frac{\frac{\partial w(\mathbf{x}, \boldsymbol{\theta}_{0l'})}{\partial \boldsymbol{\theta}_{l'}}}{w(\mathbf{x}, \boldsymbol{\theta}_{0l'})} \right) \\
&\quad - \frac{n_l}{n} \text{Cov}_l \left( \frac{\frac{\partial w(\mathbf{x}, \boldsymbol{\theta}_{0l})}{\partial \boldsymbol{\theta}_l}}{w(\mathbf{x}, \boldsymbol{\theta}_{0l})}, \frac{\zeta_{l'} \frac{\partial w(\mathbf{x}, \boldsymbol{\theta}_{0l'})}{\partial \boldsymbol{\theta}_{l'}}}{\sum_{k=1}^m \zeta_k w(\mathbf{x}, \boldsymbol{\theta}_{0k})} \right) \\
&= \sum_{i=1}^m \zeta_i \zeta_l \zeta_{l'} \text{Cov}_i \left( \frac{\frac{\partial w(\mathbf{x}, \boldsymbol{\theta}_{0l})}{\partial \boldsymbol{\theta}_l}}{\sum_{k=1}^m \zeta_k w(\mathbf{x}, \boldsymbol{\theta}_{0k})}, \frac{\frac{\partial w(\mathbf{x}, \boldsymbol{\theta}_{0l'})}{\partial \boldsymbol{\theta}_{l'}}}{\sum_{k=1}^m \zeta_k w(\mathbf{x}, \boldsymbol{\theta}_{0k})} \right) \\
&\quad - \zeta_l \zeta_{l'} \text{Cov}_{l'} \left( \frac{\frac{\partial w(\mathbf{x}, \boldsymbol{\theta}_{0l})}{\partial \boldsymbol{\theta}_l}}{\sum_{k=1}^m \zeta_k w(\mathbf{x}, \boldsymbol{\theta}_{0k})}, \frac{\frac{\partial w(\mathbf{x}, \boldsymbol{\theta}_{0l'})}{\partial \boldsymbol{\theta}_{l'}}}{w(\mathbf{x}, \boldsymbol{\theta}_{0l'})} \right) \\
&\quad - \zeta_l \zeta_{l'} \text{Cov}_l \left( \frac{\frac{\partial w(\mathbf{x}, \boldsymbol{\theta}_{0l})}{\partial \boldsymbol{\theta}_l}}{w(\mathbf{x}, \boldsymbol{\theta}_{0l})}, \frac{\frac{\partial w(\mathbf{x}, \boldsymbol{\theta}_{0l'})}{\partial \boldsymbol{\theta}_{l'}}}{\sum_{k=1}^m \zeta_k w(\mathbf{x}, \boldsymbol{\theta}_{0k})} \right)
\end{aligned}$$

For  $l = 1, \dots, q$ , the elements of  $S_{22}$  are calculated as follows:

- for  $l = l'$

$$\text{var} \left( \frac{1}{\sqrt{n}} \frac{\partial l(\boldsymbol{\theta}_0, \boldsymbol{\zeta})}{\partial \boldsymbol{\mu}_l} \right) = \frac{1}{n} \text{var} \left( \frac{\sum_{i=1}^m \sum_{j=1}^{n_i} (w(\mathbf{x}_{ij}, \boldsymbol{\theta}_{0l}) - 1)}{\sum_{k=1}^m \zeta_k w(\mathbf{x}_{ij}, \boldsymbol{\theta}_{0k})} \right)$$

$$= \frac{1}{n} \sum_{i=1}^m n_i \text{var}_i \left( \frac{w(\mathbf{x}, \boldsymbol{\theta}_{0l}) - 1}{\sum_{k=1}^m \zeta_k w(\mathbf{x}, \boldsymbol{\theta}_{0k})} \right) = \sum_{i=1}^m \zeta_i \text{var}_i \left( \frac{w(\mathbf{x}, \boldsymbol{\theta}_{0l}) - 1}{\sum_{k=1}^m \zeta_k w(\mathbf{x}, \boldsymbol{\theta}_{0k})} \right)$$

- for  $l \neq l'$

$$\begin{aligned} & \text{Cov} \left( \frac{1}{\sqrt{n}} \frac{\partial l(\boldsymbol{\theta}_0, \boldsymbol{\zeta})}{\partial \boldsymbol{\mu}_l}, \frac{1}{\sqrt{n}} \frac{\partial l(\boldsymbol{\theta}_0, \boldsymbol{\zeta})}{\partial \boldsymbol{\mu}_{l'}} \right) = \\ &= \frac{1}{n} \text{Cov} \left( \frac{\sum_{i=1}^m \sum_{j=1}^{n_i} (w(\mathbf{x}_{ij}, \boldsymbol{\theta}_{0l}) - 1)}{\sum_{k=1}^m \zeta_k w(\mathbf{x}_{ij}, \boldsymbol{\theta}_{0k})}, \frac{\sum_{i=1}^m \sum_{j=1}^{n_i} (w(\mathbf{x}_{ij}, \boldsymbol{\theta}_{0l'}) - 1)}{\sum_{k=1}^m \zeta_k w(\mathbf{x}_{ij}, \boldsymbol{\theta}_{0k})} \right) \\ &= \frac{1}{n} \sum_{i=1}^m n_i \text{Cov}_i \left( \frac{w(\mathbf{x}, \boldsymbol{\theta}_{0l}) - 1}{\sum_{k=1}^m \zeta_k w(\mathbf{x}, \boldsymbol{\theta}_{0k})}, \frac{w(\mathbf{x}, \boldsymbol{\theta}_{0l'}) - 1}{\sum_{k=1}^m \zeta_k w(\mathbf{x}, \boldsymbol{\theta}_{0k})} \right) \\ &= \sum_{i=1}^m \zeta_i \text{Cov}_i \left( \frac{w(\mathbf{x}, \boldsymbol{\theta}_{0l}) - 1}{\sum_{k=1}^m \zeta_k w(\mathbf{x}, \boldsymbol{\theta}_{0k})}, \frac{w(\mathbf{x}, \boldsymbol{\theta}_{0l'}) - 1}{\sum_{k=1}^m \zeta_k w(\mathbf{x}, \boldsymbol{\theta}_{0k})} \right) \end{aligned}$$

For  $l = 1, \dots, q$ , the elements of  $S_{12}$  are calculated as follows:

- for  $l = l'$

$$\begin{aligned} & \text{Cov} \left( \frac{1}{\sqrt{n}} \frac{\partial l(\boldsymbol{\theta}_0, \boldsymbol{\zeta})}{\partial \boldsymbol{\theta}_l}, \frac{1}{\sqrt{n}} \frac{\partial l(\boldsymbol{\theta}_0, \boldsymbol{\zeta})}{\partial \boldsymbol{\mu}_l} \right) \\ &= \frac{1}{n} \text{Cov} \left( - \sum_{i=1}^m \sum_{j=1}^{n_i} \frac{\zeta_l \frac{\partial w(\mathbf{x}_{ij}, \boldsymbol{\theta}_{0l})}{\partial \boldsymbol{\theta}_l}}{\sum_{k=1}^m \zeta_k w(\mathbf{x}_{ij}, \boldsymbol{\theta}_{0k})} + \sum_{j=1}^{n_l} \frac{\frac{\partial w(\mathbf{x}_{lj}, \boldsymbol{\theta}_{0l})}{\partial \boldsymbol{\theta}_l}}{w(\mathbf{x}_{lj}, \boldsymbol{\theta}_{0l})}, \right. \\ & \quad \left. - \frac{\sum_{i=1}^m \sum_{j=1}^{n_i} (w(\mathbf{x}_{ij}, \boldsymbol{\theta}_{0l}) - 1)}{\sum_{k=1}^m \zeta_k w(\mathbf{x}_{ij}, \boldsymbol{\theta}_{0k})} \right) \\ &= \frac{1}{n} \text{Cov} \left( \sum_{i=1}^m \sum_{j=1}^{n_i} \frac{\zeta_l \frac{\partial w(\mathbf{x}_{ij}, \boldsymbol{\theta}_{0l})}{\partial \boldsymbol{\theta}_l}}{\sum_{k=1}^m \zeta_k w(\mathbf{x}_{ij}, \boldsymbol{\theta}_{0k})}, \frac{\sum_{i=1}^m \sum_{j=1}^{n_i} (w(\mathbf{x}_{ij}, \boldsymbol{\theta}_{0l}) - 1)}{\sum_{k=1}^m \zeta_k w(\mathbf{x}_{ij}, \boldsymbol{\theta}_{0k})} \right) \\ & \quad - \frac{1}{n} \text{Cov} \left( \sum_{j=1}^{n_l} \frac{\frac{\partial w(\mathbf{x}_{lj}, \boldsymbol{\theta}_{0l})}{\partial \boldsymbol{\theta}_l}}{w(\mathbf{x}_{lj}, \boldsymbol{\theta}_{0l})}, \frac{\sum_{i=1}^m \sum_{j=1}^{n_i} (w(\mathbf{x}_{ij}, \boldsymbol{\theta}_{0l}) - 1)}{\sum_{k=1}^m \zeta_k w(\mathbf{x}_{ij}, \boldsymbol{\theta}_{0k})} \right) \\ &= \sum_{i=1}^m \zeta_i \zeta_l \text{Cov}_i \left( \frac{\frac{\partial w(\mathbf{x}, \boldsymbol{\theta}_{0l})}{\partial \boldsymbol{\theta}_l}}{\sum_{k=1}^m \zeta_k w(\mathbf{x}, \boldsymbol{\theta}_{0k})}, \frac{\sum_{i=1}^m \sum_{j=1}^{n_i} (w(\mathbf{x}, \boldsymbol{\theta}_{0l}) - 1)}{\sum_{k=1}^m \zeta_k w(\mathbf{x}, \boldsymbol{\theta}_{0k})} \right) \end{aligned}$$

$$-\zeta_l \text{Cov}_l \left( \frac{\frac{\partial w(\mathbf{x}, \boldsymbol{\theta}_{0l})}{\partial \boldsymbol{\theta}_l}}{w(\mathbf{x}, \boldsymbol{\theta}_{0l})}, \frac{w(\mathbf{x}, \boldsymbol{\theta}_{0l}) - 1}{\sum_{k=1}^m \zeta_k w(\mathbf{x}, \boldsymbol{\theta}_{0k})} \right)$$

- for  $l \neq l'$

$$\begin{aligned} & \text{Cov} \left( \frac{1}{\sqrt{n}} \frac{\partial l(\boldsymbol{\theta}_0, \boldsymbol{\zeta})}{\partial \boldsymbol{\theta}_l}, \frac{1}{\sqrt{n}} \frac{\partial l(\boldsymbol{\theta}_0, \boldsymbol{\zeta})}{\partial \boldsymbol{\mu}_{l'}} \right) \\ &= \frac{1}{n} \text{Cov} \left( - \sum_{i=1}^m \sum_{j=1}^{n_i} \frac{\zeta_l \frac{\partial w(\mathbf{x}_{ij}, \boldsymbol{\theta}_{0l})}{\partial \boldsymbol{\theta}_l}}{\sum_{k=1}^m \zeta_k w(\mathbf{x}_{ij}, \boldsymbol{\theta}_{0k})} + \sum_{j=1}^{n_l} \frac{\frac{\partial w(\mathbf{x}_{lj}, \boldsymbol{\theta}_{0l})}{\partial \boldsymbol{\theta}_l}}{w(\mathbf{x}_{lj}, \boldsymbol{\theta}_{0l})}, \right. \\ & \quad \left. - \frac{\sum_{i=1}^m \sum_{j=1}^{n_i} (w(\mathbf{x}_{ij}, \boldsymbol{\theta}_{0l'}) - 1)}{\sum_{k=1}^m \zeta_k w(\mathbf{x}_{ij}, \boldsymbol{\theta}_{0k})} \right) \\ &= \sum_{i=1}^m \zeta_i \zeta_l \text{Cov}_i \left( \frac{\frac{\partial w(\mathbf{x}, \boldsymbol{\theta}_{0l})}{\partial \boldsymbol{\theta}_l}}{\sum_{k=1}^m \zeta_k w(\mathbf{x}, \boldsymbol{\theta}_{0k})}, \frac{\sum_{i=1}^m \sum_{j=1}^{n_i} (w(\mathbf{x}, \boldsymbol{\theta}_{0l'}) - 1)}{\sum_{k=1}^m \zeta_k w(\mathbf{x}, \boldsymbol{\theta}_{0k})} \right) \\ & \quad - \zeta_l \text{Cov}_l \left( \frac{\frac{\partial w(\mathbf{x}, \boldsymbol{\theta}_{0l})}{\partial \boldsymbol{\theta}_l}}{w(\mathbf{x}, \boldsymbol{\theta}_{0l})}, \frac{w(\mathbf{x}, \boldsymbol{\theta}_{0l'}) - 1}{\sum_{k=1}^m \zeta_k w(\mathbf{x}, \boldsymbol{\theta}_{0k})} \right) \end{aligned}$$

Matrix  $\mathbf{D}$  is the limit, as  $n \rightarrow \infty$ , of  $-\frac{1}{n} \left( \frac{\partial^2 l(\boldsymbol{\theta}_0, \boldsymbol{\zeta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\mu}} \right) = \begin{pmatrix} -\frac{1}{n} \frac{\partial^2 l(\boldsymbol{\theta}_0, \boldsymbol{\zeta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} & -\frac{1}{n} \frac{\partial^2 l(\boldsymbol{\theta}_0, \boldsymbol{\zeta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\mu}'} \\ -\frac{1}{n} \frac{\partial^2 l(\boldsymbol{\theta}_0, \boldsymbol{\zeta})}{\partial \boldsymbol{\mu} \partial \boldsymbol{\theta}'} & -\frac{1}{n} \frac{\partial^2 l(\boldsymbol{\theta}_0, \boldsymbol{\zeta})}{\partial \boldsymbol{\mu} \partial \boldsymbol{\mu}'} \end{pmatrix}$

It is easy to show that:

- For  $l = l'$

$$\begin{aligned} \frac{\partial^2 l(\boldsymbol{\theta}_0, \boldsymbol{\zeta})}{\partial \boldsymbol{\theta}_l^2} &= - \sum_{i=1}^m \sum_{j=1}^{n_i} \frac{\sum_{k=1}^m \zeta_k w(\mathbf{x}_{ij}, \boldsymbol{\theta}_k) \zeta_l \frac{\partial^2 w(\mathbf{x}_{ij}, \boldsymbol{\theta}_l)}{\partial \boldsymbol{\theta}_l^2} - \left( \zeta_l \frac{\partial w(\mathbf{x}_{ij}, \boldsymbol{\theta}_l)}{\partial \boldsymbol{\theta}_l} \right)^2}{\left( \sum_{k=1}^m \zeta_k w(\mathbf{x}_{ij}, \boldsymbol{\theta}_k) \right)^2} \\ & \quad + \sum_{j=1}^{n_l} \frac{\partial^2 \log w(\mathbf{x}_{lj}, \boldsymbol{\theta}_l)}{\partial \boldsymbol{\theta}_l^2} \\ \frac{\partial^2 l(\boldsymbol{\theta}_0, \boldsymbol{\zeta})}{\partial \boldsymbol{\mu}_l^2} &= \sum_{i=1}^m \sum_{j=1}^{n_i} \frac{(w(\mathbf{x}_{ij}, \boldsymbol{\theta}_l) - 1)^2}{\left( \sum_{k=1}^m \zeta_k w(\mathbf{x}_{ij}, \boldsymbol{\theta}_k) \right)^2} \\ \frac{\partial^2 l(\boldsymbol{\theta}_0, \boldsymbol{\zeta})}{\partial \boldsymbol{\theta}_l \partial \boldsymbol{\mu}_l} &= - \sum_{i=1}^m \sum_{j=1}^{n_i} \frac{\partial w(\mathbf{x}_{ij}, \boldsymbol{\theta}_l)}{\partial \boldsymbol{\theta}_l} \frac{\zeta_l + \sum_{k \neq l}^m \zeta_k w(\mathbf{x}_{ij}, \boldsymbol{\theta}_k)}{\left( \sum_{k=1}^m \zeta_k w(\mathbf{x}_{ij}, \boldsymbol{\theta}_k) \right)^2} \end{aligned}$$

- For  $l \neq l'$

$$\begin{aligned}\frac{\partial^2 l(\boldsymbol{\theta}_0, \boldsymbol{\zeta})}{\partial \boldsymbol{\theta}_l \partial \boldsymbol{\theta}_{l'}} &= \sum_{i=1}^m \sum_{j=1}^{n_i} \frac{\zeta_l \frac{\partial w(\mathbf{x}_{ij}, \boldsymbol{\theta}_l)}{\partial \boldsymbol{\theta}_l} \zeta_{l'} \frac{\partial w(\mathbf{x}_{ij}, \boldsymbol{\theta}_{l'})}{\partial \boldsymbol{\theta}_{l'}}}{\left(\sum_{k=1}^m \zeta_k w(\mathbf{x}_{ij}, \boldsymbol{\theta}_k)\right)^2} \\ \frac{\partial^2 l(\boldsymbol{\theta}_0, \boldsymbol{\zeta})}{\partial \boldsymbol{\mu}_l \partial \boldsymbol{\mu}_{l'}} &= \sum_{i=1}^m \sum_{j=1}^{n_i} \frac{(w(\mathbf{x}_{ij}, \boldsymbol{\theta}_l) - 1)(w(\mathbf{x}_{ij}, \boldsymbol{\theta}_{l'}) - 1)}{\left(\sum_{k=1}^m \zeta_k w(\mathbf{x}_{ij}, \boldsymbol{\theta}_k)\right)^2} \\ \frac{\partial^2 l(\boldsymbol{\theta}_0, \boldsymbol{\zeta})}{\partial \boldsymbol{\theta}_l \boldsymbol{\mu}_l} &= \sum_{i=1}^m \sum_{j=1}^{n_i} \zeta_l \frac{\partial w(\mathbf{x}_{ij}, \boldsymbol{\theta}_l)}{\partial \boldsymbol{\theta}_l} \frac{w(\mathbf{x}_{ij}, \boldsymbol{\theta}_l) - 1}{\left(\sum_{k=1}^m \zeta_k w(\mathbf{x}_{ij}, \boldsymbol{\theta}_k)\right)^2}\end{aligned}$$

Notice that since  $\int w(\mathbf{x}, \boldsymbol{\theta}_i) dG_m(\mathbf{x}) = 1$ , for fixed  $i$ , then by differentiating twice we have that  $\int \frac{\partial^2 w(\mathbf{x}, \boldsymbol{\theta}_i)}{\partial \boldsymbol{\theta}_i^2} dG_m(\mathbf{x}) = 0$ .

For  $l = 1, \dots, q$ , the elements of  $\mathbf{D}$  are calculated as follows:

- For  $l = l'$

$$\begin{aligned}-\frac{1}{n} \frac{\partial^2 l(\boldsymbol{\theta}_0, \boldsymbol{\zeta})}{\partial \boldsymbol{\theta}_l^2} &= \sum_{i=1}^m \frac{n_i}{n} \frac{1}{n_i} \sum_{j=1}^{n_i} \frac{\sum_{k=1}^m \zeta_k w(\mathbf{x}_{ij}, \boldsymbol{\theta}_k) \zeta_l \frac{\partial^2 w(\mathbf{x}_{ij}, \boldsymbol{\theta}_l)}{\partial \boldsymbol{\theta}_l^2} - \left(\zeta_l \frac{\partial w(\mathbf{x}_{ij}, \boldsymbol{\theta}_l)}{\partial \boldsymbol{\theta}_l}\right)^2}{\left(\sum_{k=1}^m \zeta_k w(\mathbf{x}_{ij}, \boldsymbol{\theta}_k)\right)^2} \\ &\quad - \frac{n_l}{n} \frac{1}{n_l} \sum_{j=1}^{n_l} \frac{\partial^2 \log w(\mathbf{x}_{ij}, \boldsymbol{\theta}_l)}{\partial \boldsymbol{\theta}_l^2} \\ &\stackrel{\text{WLLN}}{\underset{n_i \rightarrow \infty}{\rightarrow}} \sum_{i=1}^m \zeta_i \int \frac{\sum_{k=1}^m \zeta_k w(\mathbf{x}, \boldsymbol{\theta}_k) \zeta_l \frac{\partial^2 w(\mathbf{x}, \boldsymbol{\theta}_l)}{\partial \boldsymbol{\theta}_l^2} - \left(\zeta_l \frac{\partial w(\mathbf{x}, \boldsymbol{\theta}_l)}{\partial \boldsymbol{\theta}_l}\right)^2}{\left(\sum_{k=1}^m \zeta_k w(\mathbf{x}, \boldsymbol{\theta}_k)\right)^2} dG_i(\mathbf{x}) \\ &\quad - \zeta_l \int \frac{\partial^2 \log w(\mathbf{x}, \boldsymbol{\theta}_l)}{\partial \boldsymbol{\theta}_l^2} w(\mathbf{x}, \boldsymbol{\theta}_l) dG_m(\mathbf{x}) \\ &= E_m \left[ \frac{\sum_{k=1}^m \zeta_k w(\mathbf{x}, \boldsymbol{\theta}_k) \zeta_l \frac{\partial^2 w(\mathbf{x}, \boldsymbol{\theta}_l)}{\partial \boldsymbol{\theta}_l^2} - \left(\zeta_l \frac{\partial w(\mathbf{x}, \boldsymbol{\theta}_l)}{\partial \boldsymbol{\theta}_l}\right)^2}{\sum_{k=1}^m \zeta_k w(\mathbf{x}, \boldsymbol{\theta}_k)} \right] \\ &\quad - \zeta_l \int \frac{\partial^2 \log w(\mathbf{x}, \boldsymbol{\theta}_l)}{\partial \boldsymbol{\theta}_l^2} w(\mathbf{x}, \boldsymbol{\theta}_l) dG_m(\mathbf{x}) \\ &= \zeta_l \int \left( \frac{\partial^2 w(\mathbf{x}, \boldsymbol{\theta}_l)}{\partial \boldsymbol{\theta}_l^2} \right)^2 \left( \frac{1}{w(\mathbf{x}, \boldsymbol{\theta}_l)} - \frac{\zeta_l}{\sum_{k=1}^m \zeta_k w(\mathbf{x}, \boldsymbol{\theta}_k)} \right) dG_m(\mathbf{x})\end{aligned}$$

$$\begin{aligned}
&= \zeta_l \int \left( \frac{\partial^2 w(\mathbf{x}, \boldsymbol{\theta}_l)}{\partial \boldsymbol{\theta}_l^2} \right)^2 \frac{\sum_{k \neq l}^m \zeta_k w(\mathbf{x}, \boldsymbol{\theta}_k)}{w(\mathbf{x}, \boldsymbol{\theta}_l) \sum_{k=1}^m \zeta_k w(\mathbf{x}, \boldsymbol{\theta}_k)} dG_m(\mathbf{x}) = \mathbf{D}_{11} \\
-\frac{1}{n} \frac{\partial^2 l(\boldsymbol{\theta}_0, \boldsymbol{\zeta})}{\partial \boldsymbol{\mu}_l^2} &= -\sum_{i=1}^m \frac{n_i}{n} \frac{1}{n_i} \sum_{j=1}^{n_i} \frac{(w(\mathbf{x}_{ij}, \boldsymbol{\theta}_l) - 1)^2}{\left( \sum_{k=1}^m \zeta_k w(\mathbf{x}_{ij}, \boldsymbol{\theta}_k) \right)^2} \\
&\xrightarrow[n_i \rightarrow \infty]{\text{WLLN}} -\sum_{i=1}^m \zeta_i \int \frac{(w(\mathbf{x}, \boldsymbol{\theta}_l) - 1)^2}{\left( \sum_{k=1}^m \zeta_k w(\mathbf{x}, \boldsymbol{\theta}_k) \right)^2} w(\mathbf{x}, \boldsymbol{\theta}_i) dG_m(\mathbf{x}) \\
&= -E_m \left[ \frac{(w(\mathbf{x}, \boldsymbol{\theta}_l) - 1)^2}{\sum_{k=1}^m \zeta_k w(\mathbf{x}, \boldsymbol{\theta}_k)} \right] = \mathbf{D}_{22} \\
-\frac{1}{n} \frac{\partial^2 l(\boldsymbol{\theta}_0, \boldsymbol{\zeta})}{\partial \boldsymbol{\theta}_l \partial \boldsymbol{\mu}_l} &= \sum_{i=1}^m \frac{n_i}{n} \frac{1}{n_i} \sum_{j=1}^{n_i} \frac{\partial w(\mathbf{x}_{ij}, \boldsymbol{\theta}_l)}{\partial \boldsymbol{\theta}_l} \frac{\zeta_l + \sum_{k \neq l}^m \zeta_k w(\mathbf{x}_{ij}, \boldsymbol{\theta}_k)}{\left( \sum_{k=1}^m \zeta_k w(\mathbf{x}_{ij}, \boldsymbol{\theta}_k) \right)^2} \\
&\xrightarrow[n_i \rightarrow \infty]{\text{WLLN}} \sum_{i=1}^m \zeta_i \int \frac{\partial w(\mathbf{x}, \boldsymbol{\theta}_l)}{\partial \boldsymbol{\theta}_l} \frac{\zeta_l + \sum_{k \neq l}^m \zeta_k w(\mathbf{x}, \boldsymbol{\theta}_k)}{\left( \sum_{k=1}^m \zeta_k w(\mathbf{x}, \boldsymbol{\theta}_k) \right)^2} w(\mathbf{x}, \boldsymbol{\theta}_i) dG_m(\mathbf{x}) \\
&= \int \frac{\partial w(\mathbf{x}, \boldsymbol{\theta}_l)}{\partial \boldsymbol{\theta}_l} \frac{\zeta_l + \sum_{k \neq l}^m \zeta_k w(\mathbf{x}, \boldsymbol{\theta}_k)}{\sum_{k=1}^m \zeta_k w(\mathbf{x}, \boldsymbol{\theta}_k)} dG_m(\mathbf{x}) \\
&= E_m \left[ \frac{\partial w(\mathbf{x}, \boldsymbol{\theta}_l)}{\partial \boldsymbol{\theta}_l} \frac{\zeta_l + \sum_{k \neq l}^m \zeta_k w(\mathbf{x}, \boldsymbol{\theta}_k)}{\sum_{k=1}^m \zeta_k w(\mathbf{x}, \boldsymbol{\theta}_k)} \right] = \mathbf{D}_{12}
\end{aligned}$$

- For  $l \neq l'$

$$\begin{aligned}
-\frac{1}{n} \frac{\partial^2 l(\boldsymbol{\theta}_0, \boldsymbol{\zeta})}{\partial \boldsymbol{\theta}_l \partial \boldsymbol{\theta}_{l'}} &= -\zeta_l \zeta_{l'} \sum_{i=1}^m \frac{n_i}{n} \frac{1}{n_i} \sum_{j=1}^{n_i} \frac{\frac{\partial w(\mathbf{x}_{ij}, \boldsymbol{\theta}_l)}{\partial \boldsymbol{\theta}_l} \frac{\partial w(\mathbf{x}_{ij}, \boldsymbol{\theta}_{l'})}{\partial \boldsymbol{\theta}_{l'}}}{\left( \sum_{k=1}^m \zeta_k w(\mathbf{x}_{ij}, \boldsymbol{\theta}_k) \right)^2} \\
&\xrightarrow[n_i \rightarrow \infty]{\text{WLLN}} -\zeta_l \zeta_{l'} \sum_{i=1}^m \zeta_i \int \frac{\frac{\partial w(\mathbf{x}, \boldsymbol{\theta}_l)}{\partial \boldsymbol{\theta}_l} \frac{\partial w(\mathbf{x}, \boldsymbol{\theta}_{l'})}{\partial \boldsymbol{\theta}_{l'}}}{\left( \sum_{k=1}^m \zeta_k w(\mathbf{x}, \boldsymbol{\theta}_k) \right)^2} w(\mathbf{x}, \boldsymbol{\theta}_i) dG_m(\mathbf{x}) \\
&= -\zeta_l \zeta_{l'} E_m \left[ \frac{\frac{\partial w(\mathbf{x}, \boldsymbol{\theta}_l)}{\partial \boldsymbol{\theta}_l} \frac{\partial w(\mathbf{x}, \boldsymbol{\theta}_{l'})}{\partial \boldsymbol{\theta}_{l'}}}{\sum_{k=1}^m \zeta_k w(\mathbf{x}, \boldsymbol{\theta}_k)} \right] = \mathbf{D}_{11} \\
-\frac{1}{n} \frac{\partial^2 l(\boldsymbol{\theta}_0, \boldsymbol{\zeta})}{\partial \boldsymbol{\mu}_l \partial \boldsymbol{\mu}_{l'}} &= -\sum_{i=1}^m \frac{n_i}{n} \frac{1}{n_i} \sum_{j=1}^{n_i} \frac{(w(\mathbf{x}_{ij}, \boldsymbol{\theta}_l) - 1)(w(\mathbf{x}_{ij}, \boldsymbol{\theta}_{l'}) - 1)}{\left( \sum_{k=1}^m \zeta_k w(\mathbf{x}_{ij}, \boldsymbol{\theta}_k) \right)^2} \\
&\xrightarrow[n_i \rightarrow \infty]{\text{WLLN}} -\sum_{i=1}^m \zeta_i \int \frac{(w(\mathbf{x}, \boldsymbol{\theta}_l) - 1)(w(\mathbf{x}, \boldsymbol{\theta}_{l'}) - 1)}{\left( \sum_{k=1}^m \zeta_k w(\mathbf{x}, \boldsymbol{\theta}_k) \right)^2} w(\mathbf{x}, \boldsymbol{\theta}_i) dG_m(\mathbf{x}) \\
&= -E_m \left[ \frac{(w(\mathbf{x}, \boldsymbol{\theta}_l) - 1)(w(\mathbf{x}, \boldsymbol{\theta}_{l'}) - 1)}{\sum_{k=1}^m \zeta_k w(\mathbf{x}, \boldsymbol{\theta}_k)} \right] = \mathbf{D}_{22}
\end{aligned}$$

$$\begin{aligned}
-\frac{1}{n} \frac{\partial^2 l(\boldsymbol{\theta}_0, \boldsymbol{\zeta})}{\partial \boldsymbol{\theta}_l \mu_l} &= -\sum_{i=1}^m \frac{n_i}{n} \frac{1}{n_i} \sum_{j=1}^{n_i} \zeta_l \frac{\partial w(\mathbf{x}_{ij}, \boldsymbol{\theta}_l)}{\partial \boldsymbol{\theta}_l} \frac{w(\mathbf{x}_{ij}, \boldsymbol{\theta}_l) - 1}{(\sum_{k=1}^m \zeta_k w(\mathbf{x}_{ij}, \boldsymbol{\theta}_k))^2} \\
&\xrightarrow[n_i \rightarrow \infty]{\text{WLLN}} -\sum_{i=1}^m \zeta_l \int \zeta_l \frac{\partial w(\mathbf{x}, \boldsymbol{\theta}_l)}{\partial \boldsymbol{\theta}_l} \frac{w(\mathbf{x}, \boldsymbol{\theta}_l) - 1}{(\sum_{k=1}^m \zeta_k w(\mathbf{x}, \boldsymbol{\theta}_k))^2} w(\mathbf{x}, \boldsymbol{\theta}_l) dG_m(\mathbf{x}) \\
&= -E_m \left[ \zeta_l \frac{\partial w(\mathbf{x}, \boldsymbol{\theta}_l)}{\partial \boldsymbol{\theta}_l} \frac{w(\mathbf{x}, \boldsymbol{\theta}_l) - 1}{\sum_{k=1}^m \zeta_k w(\mathbf{x}, \boldsymbol{\theta}_k)} \right] = \mathbf{D}_{12}
\end{aligned}$$

## A.2 Computing $\mathbf{S}, \mathbf{V}$

Define

$$\nabla \equiv \left( \frac{\partial}{\partial \alpha_1}, \dots, \frac{\partial}{\partial \alpha_q}, \frac{\partial}{\partial \boldsymbol{\beta}_1}, \dots, \frac{\partial}{\partial \boldsymbol{\beta}_q} \right)'$$

Then  $E[\nabla l(\boldsymbol{\theta})] = E[\nabla l(\alpha_1, \dots, \alpha_q, \boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_q)] = \mathbf{0}$ . Let

$$\begin{aligned}
E_j(\mathbf{t}) &\equiv \int \mathbf{t} w_j(\mathbf{t}) dG(\mathbf{t}) \\
A_0(j, r) &\equiv \int \frac{w_j(\mathbf{t}) w_r(\mathbf{t})}{1 + \sum_{k=1}^q \rho_k w_k(\mathbf{t})} dG(\mathbf{t}) \\
\mathbf{A}_1(j, j') &\equiv \int \frac{w_j(\mathbf{t}) w_{j'}(\mathbf{t}) \mathbf{t}}{1 + \sum_{k=1}^q \rho_k w_k(\mathbf{t})} dG(\mathbf{t}) \\
\mathbf{A}_2(j, j') &\equiv \int \frac{w_j(\mathbf{t}) w_{j'}(\mathbf{t}) \mathbf{t} \mathbf{t}'}{1 + \sum_{k=1}^q \rho_k w_k(\mathbf{t})} dG(\mathbf{t})
\end{aligned}$$

for  $j, j' = 1, \dots, q$ . The entries in

$$\mathbf{V} \equiv \text{var} \left[ \frac{1}{\sqrt{n}} \nabla l(\alpha_1, \dots, \alpha_q, \boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_q) \right] \quad (\text{A.1})$$

are

$$\frac{1}{n} \text{var} \left( \frac{\partial l}{\partial \alpha_j} \right) = \frac{\rho_j^2}{1 + \sum_{k=1}^q \rho_k} \left[ A_0(j, j) - \sum_{r=1}^m \rho_r A_0^2(j, r) \right]$$

$$\begin{aligned}
\frac{1}{n}\text{Cov}\left(\frac{\partial l}{\partial \alpha_j}, \frac{\partial l}{\partial \alpha_{j'}}\right) &= \frac{\rho_j \rho_{j'}}{1 + \sum_{k=1}^q \rho_k} [A_0(j, j') - \sum_{r=1}^m \rho_r A_0(j, r) A_0(j', r)] \\
\frac{1}{n}\text{Cov}\left(\frac{\partial l}{\partial \alpha_j}, \frac{\partial l}{\partial \beta_{j'}}\right) &= \frac{\rho_j^2}{1 + \sum_{k=1}^q \rho_k} [A_0(j, j) E_j(\mathbf{t}') - \sum_{r=1}^m \rho_r A_0(j, r) \mathbf{A}'_1(j, r)] \\
\frac{1}{n}\text{Cov}\left(\frac{\partial l}{\partial \alpha_j}, \frac{\partial l}{\partial \beta_j}\right) &= \frac{\rho_j \rho_{j'}}{1 + \sum_{k=1}^q \rho_k} [A_0(j, j') E_{j'}(\mathbf{t}') - \sum_{r=1}^m \rho_r A_0(j, r) \mathbf{A}'_1(j', r)] \\
\frac{1}{n}\text{Cov}\left(\frac{\partial l}{\partial \beta_j}, \frac{\partial l}{\partial \beta_{j'}}\right) &= \frac{\rho_j \rho_{j'}}{1 + \sum_{k=1}^q \rho_k} [-A_2(j, j') + E_j(\mathbf{t}) \mathbf{A}'_1(j, j') \\
&\quad + E_{j'}(\mathbf{t}) \mathbf{A}_1(j, j') - \sum_{r=1}^m \rho_r \mathbf{A}_1(j, r) \mathbf{A}'_1(j', r)] \\
&\quad + \frac{1}{n} \sum_{i=1}^{n_j} \sum_{i=1}^{n_{j'}} \text{Cov}[(x_{ji1}, \dots, y_{ji}), (x_{j'k1}, \dots, y_{j'k})']
\end{aligned}$$

The last term is zero for  $j \neq j'$ . As  $n \rightarrow \infty$ ,

$$-\frac{1}{n} \nabla \nabla' l(\alpha_1, \dots, \alpha_q, \beta_1, \dots, \beta_q) \rightarrow \mathbf{S} \quad (\text{A.2})$$

where  $\mathbf{S}$  is a  $q(1+p) \times q(1+p)$  matrix with entries corresponding to  $j, j' = 1, \dots, q$

$$\begin{aligned}
-\frac{1}{n} \frac{\partial l^2}{\partial \alpha_j^2} &\rightarrow \frac{\rho_j}{1 + \sum_{k=1}^q \rho_k} \int \frac{w_j(\mathbf{t}) [1 + \sum_{k \neq j}^q \rho_k w_k(\mathbf{t})]}{1 + \sum_{k=1}^q \rho_k w_k(\mathbf{t})} dG(\mathbf{t}) \\
-\frac{1}{n} \frac{\partial l^2}{\partial \alpha_j \partial \alpha_{j'}} &\rightarrow \frac{-\rho_j \rho_{j'}}{1 + \sum_{k=1}^q \rho_k} \int \frac{w_j(\mathbf{t}) w_{j'}(\mathbf{t})}{1 + \sum_{k=1}^q \rho_k w_k(\mathbf{t})} dG(\mathbf{t}) \\
-\frac{1}{n} \frac{\partial l^2}{\partial \alpha_j \partial \beta_{j'}} &\rightarrow \frac{\rho_j}{1 + \sum_{k=1}^q \rho_k} \int \frac{w_j(\mathbf{t}) \mathbf{t}' [1 + \sum_{k \neq j}^q \rho_k w_k(\mathbf{t})]}{1 + \sum_{k=1}^q \rho_k w_k(\mathbf{t})} dG(\mathbf{t}) \\
-\frac{1}{n} \frac{\partial l^2}{\partial \alpha_j \partial \beta_j} &\rightarrow \frac{-\rho_j \rho_{j'}}{1 + \sum_{k=1}^q \rho_k} \int \frac{w_j(\mathbf{t}) w_{j'}(\mathbf{t}) \mathbf{t}'}{1 + \sum_{k=1}^q \rho_k w_k(\mathbf{t})} dG(\mathbf{t}) \\
-\frac{1}{n} \frac{\partial l^2}{\partial \beta_j \partial \beta_{j'}} &\rightarrow \frac{\rho_j}{1 + \sum_{k=1}^q \rho_k} \int \frac{w_j(\mathbf{t}) \mathbf{t} \mathbf{t}' [1 + \sum_{k \neq j}^q \rho_k w_k(\mathbf{t})]}{1 + \sum_{k=1}^q \rho_k w_k(\mathbf{t})} dG(\mathbf{t}) \\
-\frac{1}{n} \frac{\partial l^2}{\partial \beta_j \partial \beta_{j'}} &\rightarrow \frac{-\rho_j \rho_{j'}}{1 + \sum_{k=1}^q \rho_k} \int \frac{w_j(\mathbf{t}) w_{j'}(\mathbf{t}) \mathbf{t} \mathbf{t}'}{1 + \sum_{k=1}^q \rho_k w_k(\mathbf{t})} dG(\mathbf{t})
\end{aligned}$$



## Bibliography

- [1] Agresti, A. (2002), *Categorical Data Analysis*, John Wiley & Sons, Inc., Hoboken, New Jersey.
- [2] Aitchison, J. (1955), "On the distribution of a positive random variable having discrete probability mass at the origin," *Journal of the American Statistical Association*, 50, 901-908.
- [3] Aitchison, J., and Brown, J.A.C. (1963), *The Lognormal Distribution*, Cambridge University Press, Cambridge, UK.
- [4] Andersen, C.K., Andersen, K., and Karghe-Sorensen (2000), "Cost function estimation: The choice of a model to apply to dementia," *Health Economics*, 9, 397-409.
- [5] Armstrong, R.J. (1997), *U.S. Decennial Life Tables for 1989-91*, National Center for Health Statistics, Hyattsville, MD.
- [6] Beers, H. S. (1945). Reply to the discussion of "Six-term formulas for routine interpolation". *Record of the American Institute of Actuaries*, 34, 52-61.
- [7] Billingsley, P. (1995), *Probability and Measure*, Wiley-Interscience, New York.
- [8] Bondell, H.D. (2007), "Testing goodness-of-fit in logistic case-control studies," *Biometrika*, 94, 487-495.
- [9] Cameron, C.A., and Trivedi, P.K. (1998), *Regression Analysis of Count Data*, Cambridge University Press, UK.
- [10] Cacoullos, T.(1966), "Estimation of a multivariate density," *Annals of the Institute of Statistical Mathematics (Tokyo)*, 18(2), 179-189.
- [11] Cheng, K.F., and Chu, C.K. (2004), "Semiparametric density estimation under a two-sample density ratio model," *Bernoulli*, 10(4), 583-604.
- [12] Coale, A.J., and Kisker, E.E. (1990), "Defects in data on old-age mortality in the United States: New procedures for calculating mortality schedules and life tables at the highest ages," *Asian and Pacific Population Forum*, 4, 1-31.
- [13] Curtin, L.R. (1983), "Reliability considerations for state decennial life tables," *1983 Proceeding of the Social Statistics Section of the American Statistical Association*, 161-166.

- [14] Curtin, L.R., and Gonzalez F.J. (1986), "Smoothing procedures for life tables based on small numbers of deaths," *1986 Proceeding of the Social Statistics Section of the American Statistical Association*, 415-420.
- [15] Duan, N. (1983), "Smearing estimate: a non-parametric retransformation method," *Journal of the American Statistical Association*, 78, 605-610.
- [16] Efron, B., and Tibshirani, R.J. (1993), *An Introduction to the Bootstrap*, Chapman & Hall, New York.
- [17] Elo, I.T., and Preston, S.H (1997), "Racial and ethnic differences in mortality at older ages," in Martin LG, Soldo BJ (eds), *Racial and Ethnic Differences in the Health of Older Americans*, Washington DC: National Academy Press.
- [18] Fokianos, K., Kedem, B., Qin, J., and Short D.A. (2001), "A semiparametric approach to the one-way layout," *Technometrics* 2001, 43, 56-65.
- [19] Fokianos, K. (2004), "Merging information for semiparametric density estimation," *Journal of the Royal Statistical Society, Series B*, 66, 941-958.
- [20] Gage, T.B., and Mode, C.J. (1993), "Some laws of mortality: How well do they fit?" *Human Biology*, 65, 445-461.
- [21] Gilbert P.B., Lele S.R., Vardi Y. (1999), "Maximum likelihood estimation in semiparametric selection bias models with application to AIDS vaccine trials," *Biometrika*, 86, 27-43.
- [22] Gilbert P.B. (2000), "Large sample theory of maximum likelihood estimates in semiparametric biased sampling models," *Annals of Statistics* 2000, 28, 151-194.
- [23] Gilbert, P.B. (2004), "Goodness-of-fit tests for semiparametric biased sampling models," *Journal of Statistical Planning and Inference*, 118, 51-81.
- [24] Gompertz, B. (1825), "On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies," *Philosophical Transactions of the Royal Society of London*, 115, 513-585.
- [25] Hartmann, M. (1987), "Past and Recent Attempts to Model Mortality at All Ages," *Journal of Official Statistics*, Vol. 3, No 1, 19-36.
- [26] Hastie, T. and Tibshirani, R. (1986a), "Generalized additive models" With discussion, *Statist. Sci.*, 1, pp. 297-318.

- [27] Hastie, T. and Tibshirani, R. (1986b), “Generalized additive models, cubic splines and penalized likelihood,” *Technical Report, Div. Biostatistics, Univ. Toronto*.
- [28] Hastie, T. and Tibshirani, R. (1987), “Generalized additive models: Some applications,” *J. Amer. Statist. Assoc.*, 82, pp. 371-386.
- [29] Hastie, T. and Tibshirani, R. (1990), *Generalized Additive Models*, Chapman and Hall, London.
- [30] Heligman, L., and Pollard, J. H. (1980), “The age pattern of mortality,” *Journal of the Institute of Actuaries*, 107, Part I, 659-671.
- [31] Kedem, B., Chiu, L.S., and North, G. (1990), “Estimation of mean rain rate: Application to satellite observations,” *Journal of Geophysical Research*, 95, 1965-1972.
- [32] Kedem, B., and Fokianos, K. (2002), *Regression Models for Time Series Analysis*, Wiley, Hoboken.
- [33] Kedem, B., Pfeiffer, R., and Short, D.A. (1997), “Variability of space-time mean rain rate,” *Journal of Applied Meteorology*, 36, 443-451.
- [34] Kedem, B., and Wu, Y. (1997), “Minimum chi-square vs least squares in grouped data,” Technical Report TR. 97-37, Institute for Systems Research, University of Maryland, College Park.
- [35] Kedem B, Wen S. (2007), “Semi-parametric cluster detection,” *Journal of Statistical Theory and Practice* 2007, 1, 49-72.
- [36] Kedem B, Lu G, Wei R, Williams D. (2008), “Forecasting mortality rates via density ratio modeling,” *Canadian Jour. of Statistics* 2008, 36, 193-206.
- [37] Kedem, B., Kim, E., Voulgaraki, A., Graubard, B.I. (2009), “Two-dimensional semiparametric density ratio modeling of testicular germ cell data,” *Statistics in Medicine*, 28, 2147-2159.
- [38] Kestenbaum, B. (1992), “A description of the extreme aged population based on improved Medicare enrollment data,” *Demography*, 29, 565-580.
- [39] Kestenbaum, B. (1997), “Recent mortality of the oldest old, from Medicare data,” Paper presented at the 1997 meetings of the *Population Association of America*, March 27-29.

- [40] Keziou, A. and Leoni-Aubin, S. (2005), "Test of homogeneity in semiparametric two-sample density ratio models," *Comptes Rendus de l'Academie des Sciences, Paris, Ser. I*, 340, 905-910.
- [41] Lambert, D. (1992), "Zero-inflated Poisson regression, with an application to defects in manufacturing," *Technometrics*, 34, 1-14.
- [42] Lu, G. (2007), *Asymptotic Theory for Multiple-sample Semiparametric Density Ratio Models and its Application to Mortality Forecasting*, Ph.D. Dissertation, University of Maryland, College Park, 2007.
- [43] Makeham, W. M. (1860), "Out the law of mortality," *Journal of the Institute of Actuaries*, 13, 325-358.
- [44] Manning, W. G. (1997), "The logged dependent variable, heteroscedasticity, and the transformation problem," *Journal of Health Economics*, 17, 283-295.
- [45] McCullagh, P., and Nelder, J.A. (1989), *Generalized Linear Models*, 2nd Ed, Chapman and Hall, London.
- [46] McGlynn K. A., Devesa S. S., Sigurdson A. J., Brown L. M., Tsao L., Tarone R. E. (2003), "Trends in the incidence of testicular germ cell tumors in the United States," *Cancer*, 97, 63-70.
- [47] McGlynn K. A., Sakoda L. C., Rubertone M. V., Sesterhenn I. A., Lyu C., Graubard B. I., Erickson R. L. (2007), "Body size, dairy consumption, puberty, and risk of testicular germ cell tumors," *American Journal of Epidemiology*, 165, 355-363.
- [48] Mode, C. J., and Jacobson, M. E. (1984), "A parametric algorithm for computing model period and cohort human survival functions," *International Journal of Bio-medical Computing*, 15, 341-356.
- [49] Mullahy, J. (1998), "Much ado about two: reconsidering retransformation and the two-part model in health econometrics," *Journal of Health Economics*, 17, 247-281.
- [50] Mullahy, J. (1986), "Specification and testing of some modified count data models," *Journal of Econometrics*, 33, 341-365.
- [51] Nadaraya, E.A. (1964), "On estimating regression," *Theory of Probability and its Applications*, 9, 141-142.

- [52] Owen, A.B. (2001), *Empirical Likelihood*, Chapman & Hall, New York.
- [53] Panel on Nonstandard Mixtures of Distributions (1989). Statistical models and analysis in auditing. *Statistical Science*, 4, 2-33.
- [54] Parzen, E. (1962). "On estimation of a probability density function and mode," *Annals of Mathematical Statistics*, 33, 1065-1076.
- [55] Patil, G. P. and Rao, C. R. (1977), "The weighted distributions: A survey of their applications," *Applications of Statistics*, Ed. P.R. Krishnaiah, 383-405. Amsterdam: North-Holland.
- [56] Phue J.N., Kedem B., Jaluria P., Shiloach J. (2006), "Evaluating microarrays using a semiparametric approach: Application to the central carbon metabolism of *Escherichia coli* BL21 and JM109," *GENOMICS*, 89, 300-305.
- [57] Prentice, R.L. and Pyke, R. (1979), "Logistic disease incidence models and case-control studies," *Biometrika*, 66, 403-411.
- [58] Preston, S.H., Elo, I.T., Rosenwaike I., and Hill M. (1996), "African-American mortality at older ages: Results of a matching study," *Demography*, 33, 193-209.
- [59] Qin, J., and Lawless, J.F. (1994), "Empirical likelihood and general estimating functions," *Annals of Statistics*, 22, 300-325.
- [60] Qin J, Zhang B. (1997), "A goodness of fit test for logistic regression models based on case-control data," *Biometrika* 1997, 84, 609-618.
- [61] Qin J. (1998), "Inferences for case-control and semiparametric two-sample density ratio models," *Biometrika* 1998, 85, 619-630.
- [62] Qin, J., and Zhang, B. (2005), "Density estimation under a two-sample semi-parametric model," *Nonparametric Statistics*, 17, 665-683.
- [63] Rao, C.R. (1965), "On discrete distributions arising out of methods of ascertainment," *Classical and Contagious Discrete Distributions*, G.P. Patil, ed., Pergamon Press and Statistical Publishing Society, Calcutta, pp. 320-332.
- [64] Rao, C.R. (1973), *Linear Statistical Inference and its Applications*, Wiley, New York.
- [65] Rencher, A.C. (2000), *Linear Models in Statistics*, Wiley, New York.

- [66] Rosenwaike, I., and Hill, M.E (1996), “The accuracy of age reporting among elderly African Americans: Evidence of a birth registration effect,” *Research on Aging*, 18, 310-324.
- [67] Shao, J. (2003), *Mathematical Statistics*, 2nd Ed, Springer, New York.
- [68] Siler, W. (1979), “A competing-risk model for animal mortality,” *Ecology*, 60, 750-757.
- [69] Silverman, B. W. (1986), *ensity Estimation for Statistics and Data Analysis*, Chapman & Hall/CRC Monographs on Statistics & Applied Probability, New York.
- [70] Tan, Z. (2009), “A note on profile likelihood for exponential tilt mixture models,” *Biometrika*, 96, 229-236.
- [71] Thiele, T. N. (1872), “On a mathematical formula to express the rate of mortality throughout life,” *Journal of the Institute of Actuaries*, XVI, 313-329.
- [72] Vardi, Y. (1982), “Nonparametric estimation in the presence of length bias,” *Annals of Statistics*, 10, 616-20.
- [73] Vardi, Y. (1985), “Empirical distribution in selection bias models,” *Annals of Statistics*, 13, 178-203.
- [74] Watson, G.S. (1964), “Smooth regression analysis,” *Sankhya A*, 26, 359-372.
- [75] Wei, R., Curtin, L. R., and Anderson R. (2003), “Model US mortality data for building life tables and further studies,” *2003 Joint Statistical Meetings of the American Statistical Association, Biometrics Section [CD-ROM]*, Alexandria, VA., American Statistical Association, 4458-4464.
- [76] Wei, R., Curtin, L. R., Anderson R., and Arias E. (2006), “Smoothing state life tables based on small numbers of death,” *2006 Joint Statistical Meetings of the American Statistical Association, Biometrics Section [CD-ROM]*, Alexandria, VA., American Statistical Association, 2650-2656.
- [77] Wen, S., and Kedem, B. (2009), “A semiparametric cluster detection method - a comprehensive power comparison with Kulldorff’s method,” *International Journal of Health Geographics*, 8:73 (31 December 2009), online journal without page numbers.

- [78] Wood, S. N. (2006), *Generalized Additive Models: An introduction with R*, Chapman and Hall, London.
- [79] Zhang, B. (1999), “A chi-squared goodness-of-fit test for logistic regression models based on case-control data,” *Biometrika*, 86, 531-539.
- [80] Zhang B. (2000), “A goodness of fit test for multiplicative-intercept risk models based on case-control data,” *Statistica Sinica*, 10, 839-866.
- [81] Zhang B. (2001), “An information matrix for logistic regression models based on case-control data,” *Biometrika*, 88, 921-932.
- [82] Zhang, B. (2002), “Assessing goodness-of-fit of generalized logit models based on case-control data,” *Journal of Multivariate Analysis*, 82, 17-38.
- [83] Zeileis, A., Kleiber, C., and Jackman S. (2007), “Regression Models for Count Data in R,” *Research Report Series / Department of Statistics and Mathematics.*, 53, Wien, Wirtshaftsuniversität., 2007.