ABSTRACT

Title of dissertation: SELECTED PROBLEMS IN MULTI-SAMPLE STATISTICAL INFERENCE

Carolina Franco, Doctor of Philosophy, 2012

Dissertation directed by: Professor Abram Kagan
Department of Mathematics

In Chapter 1, a natural semiparametric model for case control study data is discussed, and the asymptotic properties of two simple methods of estimation are explored. The probability element of the model can be factored into a known positive function $h(x, \boldsymbol{\theta})$ involving the finite dimensional structural parameter $\boldsymbol{\theta}$, an infinite dimensional nuisance parameter in the form of the probability element dP of a distribution, and a normalizing constant, i.e., $dP_{\boldsymbol{\theta}}(x) = C(\boldsymbol{\theta})h(x, \boldsymbol{\theta})dP(x)$. In the setup of interest, a sample of size n is available from a population with distribution $P_{\boldsymbol{\theta}}$. A second, independent sample of size m is available from a population with distribution P. The methods of estimation involve replacing P with its empirical version $\hat{P}_m$ based on the second sample, and constructing semiparametric analogs of the maximum likelihood estimator and the method of moments estimator. The simplicity of these semiparametric estimators permits analysis of their asymptotic distribution even when n and m grow at different rates, yielding very natural and interpretable asymptotic results. In the case where $n = o(m)$, the analog of the

Maximum Likelihood Estimator is asymptotically efficient.

Chapter 2 explores a related parametric asymptotic statistics problem. Suppose a sample $(Y_1, \ldots, Y_m)$ is available from a population with density $f_Y(y; \lambda)$, and an independent sample $(X_1, \ldots, X_n)$ is available from a population with density $f_X(x; \lambda, \psi)$. Here λ is regarded as a nuisance parameter and ψ is the structural parameter, where λ and ψ are scalars. One approach to estimation of ψ would be to compute the maximum likelihood estimator based on both samples, resulting in an estimator denoted as $\hat{\boldsymbol{\psi}}_{m,n}^{(1)}$. A second approach would be to first find the maximum likelihood estimator of λ from the sample $(Y_1, \ldots, Y_m)$, namely $\hat{\lambda}_m$, and to then treat $\hat{\lambda}_m$ as the true parameter. That is, one treats $(X_1, \ldots, X_n)$ as if the sample comes from $f_X(x; \hat{\lambda}_m, \psi)$, and then computes the maximum likelihood estimator of ψ, where we denote the resulting estimator as $\hat{\boldsymbol{\psi}}_{m,n}^{(2)}$. Chapter 2 compares the asymptotic behavior of $\hat{\boldsymbol{\psi}}_{m,n}^{(1)}$ and $\hat{\boldsymbol{\psi}}_{m,n}^{(2)}$ under different assumptions about the rate of growth of m relative to n.

Chapter 3 is about small area estimation, comparing an existing empirical Bayes method with a new empirical Bayes method for confidence interval construction for small area proportions based on data collected under stratified random sampling. Consider interval estimation of m small area proportions P_i $(i = 1, \cdots, m)$ where we assume a stratified random sampling design with equal number of observations n in each stratum, and where the domains of interest are the strata. In survey analysis, a commonly used 95% confidence interval for P_i is given by

$\hat{P}_i^{EB} \pm 1.96\sqrt{mse_i}$, where $\hat{P}_i^{EB}$ and mse_i are an empirical Bayes estimator of P_i and an associated second-order unbiased mean squared error estimator $(i = 1, \cdots, m)$. The underlying model is $p_i|P_i \sim \mathcal{N}(P_i, \psi_i)$, $P_i \sim \mathcal{N}(\mu, A)$, where p_i is the usual sample proportion for domain i $(i = 1, \ldots, m)$; ψ_i are known sampling variances; μ and A are unknown hyperparameters. The well-documented problems of the normal approximation to the binomial raise questions about the accuracy of these intervals when the domain sample sizes are small or when the true domain proportions are close to 0 or 1. We argue that a more reasonable model in this setting is to assume that the sampled stratum counts have binomial distributions and that the prior distribution of the true stratum proportions follows a beta distribution. We propose a new empirical Bayes confidence interval based on this model, and examine related simulation results.

SELECTED PROBLEMS IN MULTI-SAMPLE STATISTICAL INFERENCE

by

Carolina Franco

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
2012

Advisory Committee:
Professor Abram M. Kagan, Chair/Advisor
Professor Eric V. Slud
Professor Paul J. Smith
Professor Radu V. Balan
Professor Armand M. Makowski

Acknowledgments

First and foremost, I would like to thank Dr. Abram Kagan for having been my advisor. It has been a great honor to work under the supervision of such a brilliant mathematical statistician. In the many courses I took from him, Dr. Kagan taught me to appreciate the beauty of mathematical statistics proofs. I always admired Dr. Kagan's great intelligence and witty personality. In one my favorite courses as a graduate student, his Advanced Mathematical Statistics course, he brought to class extremely interesting problems which allowed me to reach a level in my mathematical statistics skills that I didn't think possible. Dr. Kagan is a great teacher, and I learned a lot from him when I was his Teaching Assistant for STAT 400. Moreover, he gave me many important lessons about becoming a statistical researcher–he taught me that research takes patience, perseverance and hard work, but that it can bring great rewards. He provided me with interesting problems and taught me how to formulate new ones. He met with me on countless occasions to discuss research and spent a lot of time helping me make my dissertation better. He encouraged me to become really focused on the research issues surrounding my dissertation and to push myself to become a better statistician, and as a result I grew not only as a researcher but as a person. He was kind and very supportive during the most difficult moments in my time as a PhD student. He always had my best interest at heart. I am very lucky to have had him as my advisor. For all this and more, I am forever grateful.

I would also like to thank Dr. Paul J. Smith and Dr. Eric Slud. Although

I never had the privilege of taking a class from Dr. Smith, he provided me with both statistical and career advice on numerous occasions. In fact, I met Dr. Smith even before starting the PhD program, when he gave me advice on how to get into the program. He continued to give me great advice over the years, served on my dissertation committee, and carefully read my dissertation and provided comments that significantly improved it.

Besides having been my professor, Dr. Slud gave me invaluable professional advice and feedback on my dissertation. He answered many questions, directed me to extremely relevant literature and provided many interesting suggestions and important corrections to my dissertation. He was also very helpful during my internship at the Census Bureau. Moreover, he provided great guidance throughout the PhD, and particularly when I began looking for a job.

It was through Dr. Slud and Dr. Kagan that I was put in contact with Dr. Partha Lahiri, under whose supervision I served as a Research Assistant in the Joint Program in Survey Methodology. Dr. Lahiri taught me a great deal about applied research (with a theoretical flavor), for which I am very grateful. He also provided many opportunities for me, such as the chance to participate in conferences, and helped me find my internship at the Census Bureau. He gave me much valuable advice on my job search when I was approaching graduation, and introduced me to many talented statisticians in the world of survey research. I really enjoyed doing research with Dr. Lahiri; it has been an honor. Chapter 3 was completed under his direction, supported by NIH Grant R01 CA129101 to P. Lahiri.

I also want to thank Dr. Benjamin Kedem for everything he taught me in the courses I took from him, for providing me with valuable career advice, for all his encouragement throughout the years, and for his kindness. I had the privilege of spending Passover Seder with Dr. Kedem, his family, and one of his students last year. His invitation to this event was a gesture which I much appreciated.

I am also much in debt to Dr. Peter Wolfe. He was the one who first encouraged me to apply to the PhD program at the University of Maryland's Department of Mathematics. Without him, I may have never thought possible to even enter the program without a bachelor's degree in mathematics/statistics. At the time, pursuing a PhD at the math department seemed like an unattainable dream, but he saw potential in me and made it possible. I also want to thank Dr. Jeffrey Cooper and Dr. David Levermore, who were very supportive and encouraging when I was in the process of applying.

Dr. Radu Balan and Dr. Armand Makowski also served in my dissertation committee, and I am grateful for their time and for all the interesting comments they made during the defense.

I cannot neglect to mention two other people who were crucial to my success in the PhD program, Dr. Ritaja Sur and Dr. Anastasia Voulgaraki. Anastasia and Ritaja shared an office with me for many years, but they were much more than officemates. They were great friends, always there during the good times and the times of crisis. Having entered the program earlier than me, they always had great advice and were always happy to help me through each of the challenges of the PhD. I would also like to thank some of the other great friends I made during the PhD program: Eduardo Zattara, Paula Casanovas, Joyce Hsiao, Neung Soo Ha, and Dr.

Yu-Ru Huang.

Lastly, I would like to thank my family: my sister Veronica, my twin brother Miguel, and my mother and father, Sara and Enrico. They were very understanding and supportive throughout my time at the PhD program, when often I was too busy to give them all the time and attention they deserved. In the most difficult times they helped in any way they could. Veronica made me promise to dedicate my first theorem to her, so Theorem 1 is dedicated to her.

Table of Contents

List of Tables vii

List of Figures viii

1 Asymptotic Properties of the Empirical Method of Moments and the Empirical Maximum Likelihood Estimator in Kernel Families 1

1.1 Introduction 1

1.2 Natural Exponential Families 15

1.2.1 Empirical Maximum Likelihood for Natural Exponential Families 18

1.3 Kernel Families 31

1.3.1 Empirical Maximum Likelihood for Kernel Families 34

1.3.2 Empirical Method of Moments Estimator for Kernel Families . 49

1.3.3 Relationship Between the Asymptotic Distribution of the EMLE and EMME 59

1.4 The Case of m-Sample Density Ratio Models 62

1.5 Future Research 71

1.6 Appendix: An Alternative Proof of Consistency of the EMLE. 73

2 A Related Two-Sample Parametric Problem 75

2.1 Introduction 75

2.2 Asymptotic Properties of Estimator from Method 1 77

2.3 Asymptotic Properties of Estimator from Method 2 87

2.4 Comparison of Method 1 and Method 2 92

3 Interval Estimation for Small Area Proportions with Small True Proportions from Stratified Random Sampling Survey Data 94

3.1 Introduction 94

3.2 Normal-Normal Empirical Bayes CI 100

3.3 Binomial-Beta Empirical Bayes CI 104

3.4 Simulation Results 107

3.4.1 Robust C for Fixed n and a Range of Small Prior Means and Variances 110

3.4.2 Numerous Small Domains 112

3.4.3 Unlucky n? 113

3.4.4 When the True Proportion is Not Close to the Extremes . . . 114

3.5 Discussion 114

3.6 Appendix: Tables 116

Bibliography 126

List of Tables

3.1 Two-sided coverages/average lengths for Binomial-Beta CI (top numbers in each cell) and Normal-Normal CI (bottom numbers in each cell) for one domain for $N = 1000$ replications, $m = 200$, $n = 20$, $c = .0001$, $\alpha = 0.05$. 117

3.2 Two-sided coverages/average lengths for Binomial-Beta CI (top numbers in each cell) and Normal-Normal CI (bottom numbers in each cell) for one domain for $N = 1000$ replications, $m = 200$, $n = 20$, $c = .0001$, $\alpha = 0.05$. 118

3.3 Two-sided coverages/average lengths for Binomial-Beta CI (top numbers in each cell) and Normal-Normal CI (bottom numbers in each cell) for one domain for $N = 1000$ replications, $m = 200$, $n = 5$, $c = .0001$, $\alpha = 0.05$. 119

3.4 Two-sided coverages/average lengths for Binomial-Beta CI (top numbers in each cell) and Normal-Normal CI (bottom numbers in each cell) for one domain for $N = 1000$ replications, $m = 200$, $n = 5$, $c = 0.001$, $\alpha = 0.05$. 120

3.5 Two-sided coverages/average lengths for Binomial-Beta CI (top numbers in each cell) and Normal-Normal CI (bottom numbers in each cell) for one domain for $N = 1000$ replications, $m = 200$, $n = 20$, $c = 0.001$, $\alpha = 0.05$. 121

3.6 Two-sided coverages/average lengths for Binomial-Beta CI (top numbers in each cell) and Normal-Normal CI (bottom numbers in each cell) for one domain for $N = 1000$ replications, $\mu = 0.01$, $\sigma^2 = .0001$, $c = 0.001$, $\alpha = 0.05$. 122

3.7 Two-sided coverages/average lengths for Binomial-Beta CI (top numbers in each cell) and Normal-Normal CI (bottom numbers in each cell) for one domain for $N = 1000$ replications, $\mu = 0.01$, $\sigma^2 = .0001$, $c = 0.001$, $\alpha = 0.05$. 123

3.8 Two-sided coverages/average lengths for Binomial-Beta CI (top numbers in each cell) and Normal-Normal CI (bottom numbers in each cell) for one domain for $N = 10000$ replications, $\mu = 0.005$, $\sigma^2 = .00001$, $c = .00001$, $\alpha = 0.05$. 124

3.9 Two-sided coverages/average lengths for Binomial-Beta CI (top numbers in each cell) and Normal-Normal CI (bottom numbers in each cell) for one domain for $N = 1000$ replications, $\mu = 0.4$, $\sigma^2 = 0.03$, $c = 0.035$, $\alpha = 0.05$. 125

List of Figures

3.1 Frequency of $\hat{\delta}$ out of range for various parameters: Simulation results for $N = 10,000$ replications. (a) $m = 100$, $n = 10$, $\sigma^2 = .0099$, (b) $\mu = .01$, $\sigma^2 = .0099$, $n = 10$, (c) $\mu = .01$, $\sigma^2 = .0099$, $m = 100$. In each case C of (3.21) is set equal to zero. 108

Chapter 1

Asymptotic Properties of the Empirical Method of Moments and the Empirical Maximum Likelihood Estimator in Kernel Families

1.1 Introduction

In this chapter we study the asymptotic properties of semiparametric estimators based on a sample from a distribution belonging to a *kernel family* with an unknown *generator*, and on an independent sample from the generator population. Besides being of theoretical interest, the situation arises in case-control studies.

Kernel families $\{P_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \Theta\}$ of distributions on $(\mathcal{X}, \mathcal{A})$ are defined as given by a probability element

$$dP_{\boldsymbol{\theta}}(x) = C(\boldsymbol{\theta})h(x; \boldsymbol{\theta})dP(x).$$

where P is a probability distribution on a measurable space $(\mathcal{X}, \mathcal{A})$ called the generator, $\boldsymbol{\theta}$ is the parameter of interest, and $h(x, \boldsymbol{\theta})$ is a positive function called the *kernel*. For a given kernel $h(x, \boldsymbol{\theta})$ a probability measure P with $\int h(x, \boldsymbol{\theta})dP < \infty$ generates a kernel family. The definition was inspired by the Natural Exponential Families (NEFs), given by the probability element:

$$dF(\mathbf{x}; \boldsymbol{\theta}) = e^{\boldsymbol{\theta}^T\mathbf{x} - \psi(\boldsymbol{\theta})}dF(\mathbf{x}), \boldsymbol{\theta} \in \Theta.$$

In the latter expression $\mathbf{x}$ is a vector of the same dimension as $\boldsymbol{\theta}$ and F can be any distribution function with a well-defined moment generating function.

The data are in the form of two independent samples, $(X_1, \ldots, X_n)$ of size n from a population P_θ from the kernel family, and $(Y_1, \ldots, Y_m)$ of size m from a population P from the generator.

In case-control studies, this setting can arise as follows: the control population can be modeled as having a generator distribution, and the case population as having a distribution from the corresponding kernel family with a particular kernel function.

In the case where θ is a scalar and the kernel family is a natural exponential family, $\theta > 0$ means that the observed characteristic in the case group is stochastically bigger than the observed characteristic in the control group since a natural exponential family is a family with a monotone likelihood ratio.

The idea behind semiparametric methods can be illustrated by the quote by Tukey: “It is better to be approximately right than exactly wrong." The advantage of the methods described in this chapter, and of other semiparametric methods, is that it is not necessary to specify the particular parametric distribution that is the basis of the model. When the kernel family of interest is the natural exponential family, for instance, the class includes many of the most commonly used family of distributions in practice–the Poisson family, the binomial family, the negative binomial family, the normal family with known variance, the gamma family with known shape parameter, and the multinomial family. Under the semiparametric approach we will discuss, it is not necessary to specify a distribution for the underlying model– the generator is assumed to be unknown. One must specify only the kernel function that defines the class of families.

More formally, the term semiparametric model is typically used to refer to a model containing a finite dimensional parameter of interest, in this case $\boldsymbol{\theta}$, and an infinite dimensional nuisance parameter, in this case the generator P.

The question arises as to what the information bound is on $\boldsymbol{\theta}$ in a semiparametric model. For estimators based a sample $(X_1, \ldots, X_n)$, Stein (1956) [68] proposed a method of obtaining such a lower bound. This is based on the observation that the Fisher information for estimating $\boldsymbol{\theta}$ in a semiparametric problem is no greater than the Fisher information for estimating $\boldsymbol{\theta}$ in any parametric submodel. Thus, by finding the "least favorable" submodel, a bound can be obtained on the asymptotic covariance of $\boldsymbol{\theta}$ in the semiparametric model. This approach has been further developed by Levit (1974) [43], Koshevnik and Levit (1976) [38], Lindsay (1980) [43], Pfanzagl (1982) [54], and Begun *et al.* (1983) [4], among others, and is discussed in more detail in Bickel *et al.* (1993) [6] and in Van der Vaart (2000) [70].

For the model we discuss here, an expression for the information bound for $\boldsymbol{\theta}$ when $m = cn(1 + o(1))$ is given in Gilbert (2000) [23].

An important theorem in asymptotic statistics, both in parametric and nonparametric models, is the Hájek-Le Cam Convolution Theorem. We will state a version in the parametric case when there is a sample $(X_1, \ldots, X_n)$:

Convolution Theorem. *Assume that the experiment $(P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta)$ is differentiable in quadratic mean at the point $\boldsymbol{\theta}$ with nonsingular Fisher information matrix $I_{\boldsymbol{\theta}}$. Let ψ be differentiable at $\boldsymbol{\theta}$. Let $\mathbf{T}_n$ be an at $\boldsymbol{\theta}$ regular estimator sequence in the experiments $(P^n_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta)$ with limit distribution $L_{\boldsymbol{\theta}}$. Then there exists a probability measure $M_{\boldsymbol{\theta}}$ such that*

$$L_{\boldsymbol{\theta}} = \mathcal{N}\left(\mathbf{0}, \psi'(\boldsymbol{\theta}) I_{\boldsymbol{\theta}}^{-1} \psi'(\boldsymbol{\theta})^T\right) * M_{\boldsymbol{\theta}}.$$

In particular, if $L_{\boldsymbol{\theta}}$ has covariance matrix $\Sigma_{\boldsymbol{\theta}}$, then the matrix $\Sigma_{\boldsymbol{\theta}} - \psi'(\boldsymbol{\theta}) I_{\boldsymbol{\theta}}^{-1} \psi'(\boldsymbol{\theta})^T$ is nonnegative definite.

Proof. See Van der Vaart (2000), p. 115. □

For a definition of differentiable in quadratic mean, see Van der Vaart (2000), p.93, and a definition of an at $\boldsymbol{\theta}$ regular estimator sequence is given in page 115.

An analogous result applies to semiparametric models based on one sample. For details, see Van der Vaart (2000), p. 366.

A specific example of how the kernel family model applies to case control studies can be found by specifying a logistic regression model for case control data, as follows: Suppose y is a binary response, and $\mathbf{x}$ a covariate. For instance, y could be the presence of a disease and $\mathbf{x}$ could be an environmental or genetic characteristic believed to be related to the incidence of the disease. The logistic regression model is frequently used to analyze relationships between diseases and environmental/genetic factors (Qin(1998) [56]). The logistic model is

$$P(Y=1|\mathbf{X}) = \frac{\exp\{\alpha^* + \mathbf{X}^T\boldsymbol{\beta}\}}{1+\exp\{\alpha^* + \mathbf{X}^T\boldsymbol{\beta}\}}.$$

The marginal density of $\mathbf{X}$, $f(\mathbf{x})$, is not specified.

Under the case-control study sampling scheme, the data consist of two independent samples of the observed values of $\mathbf{x}$ of sizes m and n from the subsets of the population with $Y=0$ and $Y=1$, respectively. The parameter $\boldsymbol{\beta}$ is of particular interest because it elucidates the relationship between Y and $\mathbf{X}$. When $\boldsymbol{\beta}=\mathbf{0}$ it implies that the covariate $\mathbf{X}$ does not influence Y.

Denote the sample from the control group ($Y=0$) as $(\mathbf{X}_1,\ldots,\mathbf{X}_m)$ with marginal density $f_0(\mathbf{x})$, and denote the sample from the case group ($Y=1$) as $(\mathbf{X}_{m+1},\ldots,\mathbf{X}_{n+m})$, with marginal density $f_1(\mathbf{x})$. Let π be the marginal probability of $Y=1$, that is,

$$\pi = P(Y=1) = \int P(Y=1|\mathbf{x})f(\mathbf{x})d\mathbf{x}$$

The probability π cannot be computed from this integral because f is not known, but it can be noted that π is a function of f and the parameters. By Bayes' rule we have

$$f_1(\mathbf{x}) = \frac{1}{\pi}\frac{\exp\{\alpha^* + \mathbf{x}^T\beta\}}{1+\exp\{\alpha^* + \mathbf{x}^T\beta\}}f(\mathbf{x})$$

and

$$f_0(\mathbf{x}) = \frac{1}{1-\pi} \frac{1}{1+\exp\{\alpha^* + \mathbf{x}^T \boldsymbol{\beta}\}} f(\mathbf{x})$$

Thus

$$\frac{f_1(\mathbf{x})}{f_0(\mathbf{x})} = \frac{1-\pi}{\pi} \exp\{\alpha^* + \mathbf{x}^T \boldsymbol{\beta}\} = \exp\{\alpha + \mathbf{x}^T \boldsymbol{\beta}\}.$$

Equivalently,

$$f_1(x) = \exp\{\alpha + x^T \beta\} f_0(\mathbf{x}).$$

In our kernel family setup, $\boldsymbol{\theta} = \boldsymbol{\beta}$, $C(\boldsymbol{\theta}) = \exp\{\alpha\}$, and $h(\mathbf{x}, \boldsymbol{\theta}) = \exp\{x^T \boldsymbol{\theta}\}$. Notice that α must be a function of $\boldsymbol{\beta}$ and f; this follows immediately since f_0 and f_1 must integrate to 1.

In the proposed methods of estimation, the normalizing constant $C(\boldsymbol{\theta})$ plays an important role in estimation, as will be seen in Section 2.

The logistic regression is one example of how a kernel family model might be appropriate in a case control study, but the analysis can be extended to other models. Suppose one assumes the model

$$P(Z = 1|\mathbf{x}) = g(\boldsymbol{\theta}, \mathbf{x})$$

where g is a known function. Then it is easy to see, again by Bayes' theorem, that

$$f_1(\mathbf{x}) = \frac{1}{\pi} g(\boldsymbol{\theta}, \mathbf{x}) f(\mathbf{x})$$

and

$$f_0(\mathbf{x}) = \frac{1}{1-\pi}(1 - g(\boldsymbol{\theta}, \mathbf{x}))f(\mathbf{x}).$$

Thus

$$\frac{f_1(\mathbf{x})}{f_0(\mathbf{x})} = \frac{1-\pi}{\pi}\frac{g(\boldsymbol{\theta}, \mathbf{x})}{1 - g(\boldsymbol{\theta}, \mathbf{x})}.$$

Again, π must be a function of $\boldsymbol{\theta}$ so we have the kernel family model

$$f_1(\mathbf{x}) = C(\boldsymbol{\theta})\frac{g(\boldsymbol{\theta}, \mathbf{x})}{1 - g(\boldsymbol{\theta}, \mathbf{x})}f_0(\mathbf{x})$$

Therefore, specifying $P(Y = 1|\mathbf{X})$ imposes a kernel function.

Returning to the general kernel family setup, let us discuss two new simple methods of estimation of the parameter $\boldsymbol{\theta}$. To find the Empirical Maximum Likelihood Estimator (EMLE) and the Empirical Method of Moments Estimator (EMME), P is replaced by its empirical version based on the sample $(Y_1, \ldots, Y_m)$. An alternative term for the EMLE could be the Maximum Likelihood Estimator via Empirical Generator (MLEEG), since the term Maximum Empirical Likelihood (MELE) is already available in the literature and thus confusion could ensue over the two very different estimators. For the purpose of this dissertation, we will use the shorter term, EMLE, for simplicity.

The empirical distribution function is defined as

$$\hat{F}_m(y) = \sum_{i=1}^{m} I(Y_i \leq y)$$

where $I(A)$ denotes the indicator of an event A.

The empirical distribution function is a classical nonparametric estimator of an unknown distribution function F based on a sample, in this case $(Y_1, \ldots, Y_m)$. Equal masses are placed on each of the observations $y_1, \ldots, y_m$. For any fixed y, $m\hat{F}_m(y)$ has a binomial distribution $Bi(F(y), m)$, so that $\hat{F}_m(y)$ is an unbiased estimator of $F(y)$, and, moreover $\hat{F}_m(y)$ is $\sqrt{m}$ consistent for $F(y)$. A stronger asymptotic result is the classical theorem by Glivenko and Cantelli that proves uniform convergence of $\hat{F}_m(y)$ to $F(y)$, that is,

$$\sup_y |\hat{F}_m(y) - F(y)| \overset{a.s.}{\to} 0.$$

The estimators that result from replacing the generator F by $\hat{F}_m$ have interesting asymptotic properties.

Although our definition of kernel families was inspired by the Natural Exponential Families, the same model has appeared in the literature in the context of selection bias models (Gilbert *et al.*, 1999) [22], (Gilbert, 2000) [23], and an equivalent model was studied in the literature by the label of density ratio models. Qin (1998), for instance, specifically uses the density ratio model for case control studies. In these papers, semiparametric likelihood ideas are used to derive estimators for $\boldsymbol{\theta}$.

Density ratio models have been used in several applications, particularly in the setting where two or more samples are available. A density ratio model when there are two samples available, one with probability element dQ and the second with probability element dP is of the form

$$\frac{dQ}{dP} = g(x, \boldsymbol{\theta})$$

where the function g is regarded as known. This is essentially the same as the kernel family model with $g(x, \boldsymbol{\theta}){=}C(\boldsymbol{\theta})h(x, \boldsymbol{\theta})$.

Multi-sample density ratio examples found in the literature typically handle estimation using empirical likelihood ideas pioneered by Owen (1988[48], 1990[49], 1991 [50]). These methods have been studied, applied, and extended by many including Qin and Zhang (1997) [60], Qin and Lawless (1994) [58], Qin (1998) [56], Qin (2000) [59], DiCiccio, Hall, and Romano (1989) [14], Cheng and Chu (2004) [12], Keziou and Aubin (2007) [35], Kedem, *et al.* (2009) [34], Qin *et al.* (2002), [59], Zhou *et al.* (2002) [72], among others, and appear to do well in many applications.

Empirical likelihood, and its many extensions by Qin and others, comprise a series of interesting nonparametric and semiparametric methods that yields estimators of parameters of interest, whether finite dimensional or infinite dimensional. These methods permit inference when aspects of the model are specified without specifying a fully parametric model. Empirical likelihood essentially involves writing the likelihood function in terms of the unknown probability elements $p_i = dF(x_i)$ of the observed data points with respect to a distribution F, subject to appropriate constraints. The main idea is that the distribution function F can be approximated by a discrete distribution concentrated on the observed data values.

Given a sample $(X_1, \ldots, X_n)$, absent any information on the distribution of the population, we can write the empirical likelihood function as

$$L(F) = \Pi_{i=1}^n p_i$$

with constraints

$$\sum_{i=1}^n p_i = 1, \quad p_i > 0.$$

In this case, it is easy to see that the empirical likelihood is maximized by the cumulative distribution function F_n. If we are interested in estimating $T(F)$, where T is a real functional of the distribution, we can estimate it by $T(F_n)$. For instance, the mean μ can be expressed as $\int x dF(x)$ so that the Non Parametric Maximum Likelihood Estimator (NPMLE) of μ is $\int x dF_n(x) = \bar{X}_n$.

Alternatively, if the true μ were known the constraint $\sum x_i p_i = \mu$ could be used to come up with an improved estimate of F.

Moreover, suppose that the observations are of the form (X, Y) and μ_x, the true value of the mean X, is known and one is interested in estimating the mean of Y, μ_y. This arises in survey sampling. The information about μ_x could be added as a constraint, and an additional constraint could be added of the form $\sum Y_i p_i = \mu_y$. One could maximize over p_i and then over μ_y to produce an estimator $\tilde{\mu}_y$ of μ_y, usually referred to as the Maximum Empirical Likelihood Estimator (MELE). The asymptotic variance of $\tilde{\mu}_y$ is at least as small as that of $\bar{Y}_n$, with equality holding if an only if Y is uncorrelated with X (Owen, 2001 [51]).

Qin (1998) [56] specifically extends these semiparametric likelihood ideas to two-sample problems under a density ratio model for case control studies. Namely, Qin's model is

$$f_1(x) = g(x, \boldsymbol{\theta}) f_0(x),$$

where $f_1(x)$ is the marginal density of the case group and $f_0(x)$ is the marginal density of the control group. The data are two independent samples $(X_1, \dots, X_{n_0})$ from the control group and $(X_{n_0+1}, \dots, X_n)$ from the case group, where $n_0 + n_1 = n$. The log likelihood is expressed as

$$l = \sum_{i=1}^{n} \log(p_i) + \sum_{i=n_0+1}^{n} \log g(x_i, \boldsymbol{\theta})$$

subject to the constraints

$$\sum_{i=1}^{n} p_i = 1, \quad \sum_{i=1}^{n} p_i(g(x_i, \boldsymbol{\theta}) - 1) = 0, \quad p_i > 0.$$

The first two constraints make f_0 and f_1 proper probability distributions. The optimization problem is solved through Lagrange multipliers, where the Lagrange multiplier λ can be estimated from the data by treating it as an ordinary parameter. The approach is first to obtain the optimal p_i while holding all other parameters constant. The estimates for p_i are then substituted in the likelihood, which becomes a function of the data, λ, and $\boldsymbol{\theta}$.

Qin (1998) studies the asymptotic properties of the estimators of $\boldsymbol{\theta}$ and λ, focusing only on the case where $n_0 = cn_1(1 + o(1))$, or, equivalently, where $n_0/n \to$

$\rho_0 > 0$ and $n_1/n \to \rho_1 > 0$. Although this is not proved in Qin (1998), it follows from Gilbert (2000) that Qin's estimator for $\boldsymbol{\theta}$ is efficient when $n_0 = cn_1(1 + o(1))$, because it is actually equivalent to that proposed in Gilbert (2000) under the setup of Qin (1998), where there are two samples. Gilbert proves efficiency of the estimator when $n_0 = cn_1(1 + o(1))$.

Gilbert (2000) has the following setup. The model comprises three components: a probability measure G, a set of nonnegative (measurable) stratum weight functions $w_1, \ldots, w_s$, and selection probabilities $\lambda_i, \quad i = 1, \ldots, s$, with $\sum_{i=1}^{s} \lambda_i = 1$. The data are assumed to be a sample $X_k = (I_k, Y_k), k = 1, \ldots, n$. The random variable $I \in \{1, \ldots, s\}$ denotes the stratum, selected with probability λ_i. Let $g = \frac{dG}{d\mu}$ for some measure μ dominating G. The density of $\mathbf{X}$ is given by

$$p(\mathbf{x}, \boldsymbol{\theta}, G) = p(i, y, \boldsymbol{\theta}, G) = \lambda_i \frac{w_i(y, \boldsymbol{\theta})}{W_i(\boldsymbol{\theta}, G)} g(y)$$

where $W_i(\boldsymbol{\theta}, G)$ is the ith normalizing function given by

$$W_i(\boldsymbol{\theta}, G) = \int_{\mathbf{Y}} w_i(u, \boldsymbol{\theta}) dG(\mu),$$

assumed to be positive and finite for all $\boldsymbol{\theta} \in \Theta$.

Conditional on $I = i$ the probability measure F_i of Y under this model satisfies

$$F_i(\mathbf{A}, \boldsymbol{\theta}, G) = \mathbf{W}_i^{-1}(\boldsymbol{\theta}, G) \int_{\mathbf{A}} w_i(u, \boldsymbol{\theta}) dG(u), \quad i = 1, \ldots, s.$$

If $s = 2$, and if one of the weight functions is constant, this model is equivalent to the kernel family model.

Returning to the general setup analyzed in Gilbert (2000), a previous paper, Gilbert *et al.* (1999) [22] provides an expression for the likelihood of the data and a procedure to maximize the likelihood over $(\boldsymbol{\theta}, G)$, assuming one of the weight functions is constant, under some assumptions. The maximization is over $\boldsymbol{\theta}$ and over all distributions concentrated at the observed sample points. Gilbert (2000) provides the large sample properties of this maximum likelihood estimator, and in particular, the author shows that the estimator for $(\boldsymbol{\theta}, G)$ is asymptotically efficient under several assumptions including the condition that $\lambda_{n_i} = n_i/n \to \lambda_i > 0, i = 1, \ldots, s$, where n_i is the ith sample size and $n = \sum_{i=1}^{s} n_i$. Thus, the estimator of $\boldsymbol{\theta}$ is asymptotically efficient when $m = cn(1 + o(1)), \quad c > 0$, using our notation. The method we propose is simpler and allows for analysis in cases where $n = o(m)$ and $m = o(n)$. The expression provided for the information bound in Gilbert (2000) is very complicated and extremely hard to compare to the expressions for the asymptotic covariances of the EMME and EMLE, although it is expected that the former two are not efficient when $m = cn(1 + o(1))$.

To our knowledge, no asymptotic theory is available for estimators based on semiparametric density ratio models/biased sampling models with weights depending on $\boldsymbol{\theta}$ when $n_0 = o(n_1)$ or $n_1 = o(n_0)$.

As stated by Qin, there are some specific examples of $h(x, \theta)$ that are of particular interest in applications. For instance, Qin (1998) [56] studies the multiplicative-intercept risk model:

$$g(x, \boldsymbol{\theta}) = e^{\alpha + \phi(x, \boldsymbol{\beta})}.$$

The quantity e^{α} is the normalizing constant, but because it is “unknown," α is treated as a separate parameter in Qin (1998). Under our approach, because $e^{\alpha} = C(\boldsymbol{\theta})$ where $\boldsymbol{\theta} = \boldsymbol{\beta}$, the multiplicative-intercept risk model is not a special case, but rather an equivalent expression of the kernel family. In the literature, specific forms of ϕ have been found to be useful in particular applications. Examples are Storer, Wacholder and Breslow (1983) [69], which focuses on some epidemiological studies, and Kay and Little (1987)[33], who analyzed a dataset on age of menarche in girls from Warsaw. The density ratio model has also been used to model meteorological data by Fokianos *et al.* (2001) [20] and on testicular germ cell data by Kedem *et al.* (2009), among many others.

Keziou and Aubin (2007) [35] build on Qin (2008) by constructing a test statistic for hypothesis testing. They study the asymptotic properties of both the estimate $\hat{\boldsymbol{\theta}}$ and the test statistics both under the null hypothesis $H_0 : \boldsymbol{\theta} = \mathbf{0}$, and the alternative hypotheses, once again assuming $m = cn(1 + o(1)), c > 0$. Cheng and Chu (2004) [12] also study two sample density ratio models, but they focus on density estimation using the kernel density estimates suggested by Jones (1991) [27].

In this chapter the asymptotic properties of the EMME and EMLE of $\boldsymbol{\theta}$ are explored. Section 1.2 explores the properties of the EMLE as it applies to the special case of natural exponential families. The estimating equations for the EMME in this case are identical to those of the EMLE, so that there is no need to consider this case separately. Section 1.3 generalizes the results related to the EMLE and EMME to kernel families, where the two estimators are not equivalent. The asymptotic distribution of the EMME and EMLE are compared, pointing out a special relationship between the two. Section 1.4 discusses the m-sample density ratio model, and Section 1.5 discusses future research.

1.2 Natural Exponential Families

The distribution of the random vector $\mathbf{X}$ is said to belong to a natural exponential family (NEF) with a generator F if the probability element of $\mathbf{X}$ has the form

$$dF(\mathbf{x};\boldsymbol{\theta}) = \exp\{\boldsymbol{\theta}^T\mathbf{x} - \psi(\boldsymbol{\theta})\}dF(\mathbf{x}), \boldsymbol{\theta} \in \Theta \tag{1.1}$$

where Θ is an open subset of R^p, $\psi : R^p \to R$, and $\mathbf{X} \in R^p$. Implicitly we assume that the generator has an exponential moment (i.e., that its moment generating function exists). The set of $\boldsymbol{\theta}$ for which this holds for a given generator is called the natural parameter space. We assume that $\boldsymbol{\theta}_0$, the true value of $\boldsymbol{\theta}$, is in the interior of the natural parameter space, and moreover that $\mathbf{0} \in \mathrm{Int}(\Theta)$ so that $F(\mathbf{x}, \mathbf{0}) = F(\mathbf{x})$.

The assumption that the moment generating function exists on a set with a nonempty interior containing zero also implies that the moment generating function of F is analytic and that all the moments of F exist. This allows for the computation of moments by differentiating the moment generating function and evaluating it at zero. For a discussion of this result in the multivariate case see Bickel and Doksum (2001) [5], p.105. A proof in the univariate case can be found in Billingsley (1995) [7] p. 278. Brown (1986) [8] also showed that in the interior of the natural parameter space, all the moments of $F_{\boldsymbol{\theta}}$ exist. Integration can be exchanged with differentiation in the following expression:

$$\frac{\partial}{\partial \boldsymbol{\theta}} \int e^{\boldsymbol{\theta}^T \mathbf{x}} dF(\mathbf{x}) \tag{1.2}$$

where the operator $\frac{\partial}{\partial \boldsymbol{\theta}}$ applied to a function $f(\boldsymbol{\theta}, x, y, \dots)$ denotes the vector of partial derivatives $\frac{\partial}{\partial \theta_i} f(\boldsymbol{\theta}, x, y, \dots)$, that is,

$$\frac{\partial}{\partial \boldsymbol{\theta}} f(\boldsymbol{\theta}, x, y, \dots) = \begin{bmatrix} \frac{\partial}{\partial \theta_1} f(\boldsymbol{\theta}, x, y, \dots) \\ \vdots \\ \frac{\partial}{\partial \theta_p} f(\boldsymbol{\theta}, x, y, \dots) \end{bmatrix}.$$

Expression (1.2) can be written as

$$\frac{\partial}{\partial \boldsymbol{\theta}} \int e^{\boldsymbol{\theta}^T \mathbf{x}} dF(\mathbf{x}) = \int \frac{\partial}{\partial \boldsymbol{\theta}} e^{\boldsymbol{\theta}^T \mathbf{x}} dF(\mathbf{x}) = \int \mathbf{x} e^{\boldsymbol{\theta}^T \mathbf{x}} dF(\mathbf{x}). \tag{1.3}$$

Alternatively, (1.2) can be expressed as

$$\frac{\partial}{\partial \boldsymbol{\theta}} (e^{\psi(\boldsymbol{\theta})}) = \frac{\partial}{\partial \boldsymbol{\theta}} \boldsymbol{\psi}(\boldsymbol{\theta}) e^{\psi(\boldsymbol{\theta})}. \tag{1.4}$$

Combining (1.3) and (1.4) we get the following relation:

$$E(\mathbf{X}) = \boldsymbol{\psi}'(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} \psi(\boldsymbol{\theta}). \tag{1.5}$$

If F is known, so is $\psi(\boldsymbol{\theta})$ and the MLE of $\boldsymbol{\theta}$ from a sample $(\mathbf{X}_1, \ldots, \mathbf{X}_n)$ is the solution of the system of equations

$$\bar{\mathbf{X}}_n = \boldsymbol{\psi}'(\boldsymbol{\theta}) \tag{1.6}$$

where both $\bar{\mathbf{X}}_n$ and $\boldsymbol{\psi}'(\boldsymbol{\theta})$ have dimension $p \times 1$.

This system of equations is identical to that obtained by the method of moments for natural exponential families, since the mean of the natural exponential family is $\boldsymbol{\psi}'(\boldsymbol{\theta})$.

The asymptotic behavior of the MLE $\hat{\boldsymbol{\theta}}_n$ (and equivalently of the method of moments estimator $\tilde{\boldsymbol{\theta}}_n$) of $\boldsymbol{\theta}$ based on one sample $(\mathbf{X}_1, \ldots, \mathbf{X}_n)$ from a natural exponential family with a known generator is given by

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \overset{d}{\to} N_p(\mathbf{0}, \{\boldsymbol{\psi}''(\boldsymbol{\theta})\}^{-1}) \tag{1.7}$$

where $\boldsymbol{\psi}''(\boldsymbol{\theta})$ is the matrix of second partial derivatives of $\psi(\boldsymbol{\theta})$, i.e.,

$$\boldsymbol{\psi}''(\boldsymbol{\theta}) = \begin{bmatrix} \frac{\partial^2}{\partial \theta_1^2} \psi(\boldsymbol{\theta}) & \ldots & \frac{\partial^2}{\partial \theta_1 \partial \theta_p} \psi(\boldsymbol{\theta}) \\ \vdots & \vdots & \vdots \\ \frac{\partial^2}{\partial \theta_p \partial \theta_1} \psi(\boldsymbol{\theta}) & \ldots & \frac{\partial^2}{\partial \theta_p^2} \psi(\boldsymbol{\theta}) \end{bmatrix}$$

The asymptotic result (1.7) is governed by the standard theory of estimating equations.

Implicitly it is assumed that $\boldsymbol{\psi}''(\boldsymbol{\theta})$ is positive definite, so that its inverse is well-defined. This holds for any F that is nondegenerate. We will make the same assumption in the semiparametric analysis in the rest of this section. Note that $\hat{\boldsymbol{\theta}}_n$ is asymptotically efficient.

For the natural exponential family,

$$\boldsymbol{\psi}''(\boldsymbol{\theta}) = \mathbf{I}_{\mathbf{X}}(\boldsymbol{\theta}) = Cov_{\boldsymbol{\theta}}(\mathbf{X}). \tag{1.8}$$

The first relation follows immediately by differentiating the log likelihood and the second one can be shown using the cumulant generating function.

1.2.1 Empirical Maximum Likelihood for Natural Exponential Families

Suppose a sample $(\mathbf{X}_1, \ldots, \mathbf{X}_n)$ is available from a multivariate natural exponential family of dimension p. What happens if F is unknown but an independent sample $(\mathbf{Y}_1, \ldots, \mathbf{Y}_m)$ is available from its distribution?

The sample $(\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_m)$ allows us to construct the empirical distribution function $\hat{F}_m(\mathbf{y})$.

The empirical natural exponential family distribution can be constructed by replacing F by $\hat{F}_m(\mathbf{y})$ from the sample $(\mathbf{Y}_1, \ldots, \mathbf{Y}_m)$. Its probability element is

$$d\hat{F}_m(\mathbf{x}; \boldsymbol{\theta}) = \exp\{\boldsymbol{\theta}^T \mathbf{x} - \hat{\psi}_m(\boldsymbol{\theta})\} d\hat{F}_m(\mathbf{x}). \tag{1.9}$$

where

$$\exp\{\hat{\psi}_m(\boldsymbol{\theta})\} = \frac{1}{m} \sum_{i=1}^{m} \exp\{\boldsymbol{\theta}^T \mathbf{Y}_i\}. \tag{1.10}$$

This normalizing constant makes this a proper distribution, that is, it ensures that $\int \exp\{\boldsymbol{\theta}^T \mathbf{x} - \hat{\psi}_m(\boldsymbol{\theta})\} d\hat{F}_m(\mathbf{x}) = 1$.

The Strong Law of Large Numbers implies that $\hat{\psi}_m(\boldsymbol{\theta})$ is close to $\psi(\boldsymbol{\theta})$ for large m; that is,

$$\hat{\psi}_m(\boldsymbol{\theta}) \overset{a.s.}{\rightarrow} \psi(\boldsymbol{\theta}) \tag{1.11}$$

as $m \rightarrow \infty$.

This follows from applying the SLLN to $e^{\hat{\psi}_m(\boldsymbol{\theta})}$, so that

$$e^{\hat{\psi}_m(\boldsymbol{\theta})} \overset{a.s.}{\rightarrow} e^{\psi(\boldsymbol{\theta})} \tag{1.12}$$

as $m \rightarrow \infty$ since

$$E(e^{\boldsymbol{\theta}^T \mathbf{Y}_i}) = e^{\psi(\boldsymbol{\theta})}. \tag{1.13}$$

Notice that $\hat{F}_m(\mathbf{x}; \boldsymbol{\theta})$ is concentrated at the same points as $\hat{F}_m(\mathbf{x})$. The distribution $\hat{F}_m(\mathbf{x}, \boldsymbol{\theta})$ puts the masses

$$\frac{e^{\mathbf{Y}_1^T \boldsymbol{\theta}}}{\sum_{i=1}^{m} e^{\mathbf{Y}_1^T \boldsymbol{\theta}}}, \ldots, \frac{e^{\mathbf{Y}_m^T \boldsymbol{\theta}}}{\sum_{i=1}^{m} e^{\mathbf{Y}_i^T \boldsymbol{\theta}}}$$

at points $\mathbf{Y}_1, \ldots, \mathbf{Y}_m$, respectively.

For a continuous $F(\mathbf{x})$, with P_θ probability one none of the $\mathbf{X}$'s and $\mathbf{Y}$'s will match. Although this may seem counterintuitive, we recall that $\sup_{\mathbf{y}} |\hat{F}_m(\mathbf{y}) - F(\mathbf{y})| \overset{a.s.}{\to} 0$ as $m \to \infty$ and $e^{\hat{\psi}_m(\boldsymbol{\theta})} \overset{a.s.}{\to} e^{\psi(\boldsymbol{\theta})}$, so that the empirical NEF family probability element should be a good approximation to the true NEF family probability element provided m is large. The empirical NEF allows us to construct estimating equations analogous to those of the maximum likelihood estimator and the method of moments without the need to fully specify the parametric family.

To provide an expression for the empirical maximum likelihood system of equations, we compute $\frac{\partial}{\partial \boldsymbol{\theta}} \hat{\psi}_m(\boldsymbol{\theta})$, the derivative vector of $\hat{\psi}_m(\boldsymbol{\theta})$. Using (1.10), this expression has components

$$\frac{\partial}{\partial \theta_k} \hat{\psi}_m(\boldsymbol{\theta}) = \frac{\sum_{i=1}^m Y_{ik} e^{\boldsymbol{\theta}^T \mathbf{Y}_i}}{\sum_{i=1}^m e^{\boldsymbol{\theta}^T \mathbf{Y}_i}}, \quad k = 1, \ldots, p,$$

where Y_{ik} denotes the kth component of observation $\mathbf{Y}_i$. Thus the empirical ML system of equations is

$$\bar{X}_{n,k} = \frac{\sum_{i=1}^m Y_{ik} e^{\boldsymbol{\theta}^T \mathbf{Y}_i}}{\sum_{i=1}^m e^{\boldsymbol{\theta}^T \mathbf{Y}_i}}, \quad k = 1, \ldots, p$$

where $\bar{X}_{n,k}$ is the mean of the kth components of $\mathbf{X}_1, \ldots, \mathbf{X}_n$, or equivalently,

$$\bar{\mathbf{X}}_n = \frac{\sum_{i=1}^m \mathbf{Y}_i e^{\boldsymbol{\theta}^T \mathbf{Y}_i}}{\sum_{i=1}^m e^{\boldsymbol{\theta}^T \mathbf{Y}_i}}. \tag{1.14}$$

The following theorem explores the consistency and existence of the EMLE for an NEF. This is simply a special case of Theorem 3, but we include it here to illustrate the details of the argument for the special case of the NEFs.

Theorem 1. *With probability tending to one as $m, n \to \infty$ there exists a statistic $\hat{\boldsymbol{\theta}}_{m,n}$ that is a solution to the empirical maximum likelihood systems of equations and is a consistent estimator of $\boldsymbol{\theta}$.*

Proof. Let $\boldsymbol{\theta_0}$ denote the true parameter vector for $\boldsymbol{\theta}$, i.e., $\mathbf{X}_1, \ldots, \mathbf{X}_n$ is a sample from $F(\mathbf{x}; \boldsymbol{\theta}_0)$ and $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ is a sample from $F(\mathbf{x}; \mathbf{0})$

We will use Lemma 3, which will be stated and proved in section 1.3.

Consider the mapping

$$\mathbf{G}_{m,n}(\boldsymbol{\theta}, \bar{\mathbf{X}}_n, \mathbf{Y}_1, \ldots, \mathbf{Y}_m) = \bar{\mathbf{X}}_n - \frac{\sum_{i=1}^m e^{\boldsymbol{\theta}^T \mathbf{Y}_i} \mathbf{Y}_i / m}{\sum_{i=1}^m e^{\boldsymbol{\theta}^T \mathbf{Y}_i} / m}. \tag{1.15}$$

We will prove that conditions (i),(ii), and (iii) of Lemma 3 hold. To show (i), note that

$$\mathbf{G}_{m,n}(\boldsymbol{\theta}_0) = \bar{\mathbf{X}}_n - \frac{\sum_{i=1}^m e^{\boldsymbol{\theta}_0^T \mathbf{Y}_i} \mathbf{Y}_i / m}{\sum_{i=1}^m e^{\boldsymbol{\theta}_0^T \mathbf{Y}_i} / m} \overset{a.s.}{\to} \psi'(\boldsymbol{\theta}_0) - \frac{\psi'(\boldsymbol{\theta}_0) e^{\psi(\boldsymbol{\theta}_0)}}{e^{\psi(\boldsymbol{\theta}_0)}} = 0, \tag{1.16}$$

so that

$$\mathbf{G}_{m,n}(\boldsymbol{\theta}_0, \bar{\mathbf{X}}_n, \mathbf{Y}_1, \ldots, \mathbf{Y}_m) \overset{a.s}{\to} \mathbf{0} \tag{1.17}$$

We will now show that assumption (ii) of Lemma 3 holds; that is,

$$\frac{\partial}{\partial \boldsymbol{\theta}} \mathbf{G}_{m,n}(\boldsymbol{\theta}_0, \bar{\mathbf{X}}_n, \mathbf{Y}_1, \ldots, \mathbf{Y}_m) \overset{a.s.}{\to} -\boldsymbol{\psi}''(\boldsymbol{\theta}_0). \tag{1.18}$$

The $p \times p$ matrix of partial derivatives of $\mathbf{G}_{m,n}(\boldsymbol{\theta}, \bar{\mathbf{X}}_n, \mathbf{Y}_1, \ldots, \mathbf{Y}_m)$ with respect to the vector $\boldsymbol{\theta}$, $\partial / \partial \boldsymbol{\theta} \mathbf{G}_{m,n}(\boldsymbol{\theta})$ is given by entries

$$\frac{\partial}{\partial \theta_j} G_{m,n,k} = -\frac{\sum_{i=1}^m Y_{ij} Y_{ik} e^{\boldsymbol{\theta}^T \mathbf{Y}_i}}{\sum_{i=1}^m e^{\boldsymbol{\theta}^T \mathbf{Y}_i}} + \frac{\sum_{i=1}^m Y_{ij} e^{\boldsymbol{\theta}^T \mathbf{Y}_i} \sum_{i=1}^m Y_{ik} e^{\boldsymbol{\theta}^T \mathbf{Y}_i}}{(\sum_{i=1}^m e^{\boldsymbol{\theta}^T \mathbf{Y}_i})^2} \tag{1.19}$$

$$j, k = 1, \ldots, p.$$

From this expression it follows that as $m, n \to \infty$

$$\frac{\partial}{\partial \boldsymbol{\theta}} \mathbf{G}_{m,n}(\boldsymbol{\theta}_0, \bar{\mathbf{X}}_n, \mathbf{Y}_1, \ldots, \mathbf{Y}_m) \overset{a.s.}{\to} -\boldsymbol{\psi}''(\boldsymbol{\theta}_0). \tag{1.20}$$

The limiting matrix in (1.20) is negative definite. Relation (1.20) follows by applying the SLLN to each term in the expression (1.19) for the component functions. Note that using (1.8) it follows that

$$E(e^{\boldsymbol{\theta}^T \mathbf{Y}_i} Y_{ik} Y_{ij}) = \int e^{\boldsymbol{\theta}^T \mathbf{y}_i} y_{ik} y_{ij} dF(\mathbf{y}_i) = [\{\psi''(\boldsymbol{\theta})\}_{jk} + \frac{\partial}{\partial \theta_k} \psi(\boldsymbol{\theta}) \frac{\partial}{\partial \theta_j} \psi(\boldsymbol{\theta})] e^{\psi(\boldsymbol{\theta})}.$$

To show that assumption (iii) in Lemma 3 holds, we note that $\frac{\partial^2}{\partial \boldsymbol{\theta}^2} \mathbf{G}_{m,n}(\boldsymbol{\theta})$ will involve terms

$\sum_{i=1}^m e^{\boldsymbol{\theta}^T Y_i}$, $\sum_{i=1}^m e^{\boldsymbol{\theta}^T Y_i} Y_{ij}$,$\sum_{i=1}^m e^{\boldsymbol{\theta}^T Y_i} Y_{ij} Y_{ik}$, and $\sum_{i=1}^m e^{\boldsymbol{\theta}^T Y_i} Y_{ij} Y_{ik} Y_{il}$.

The summands are continuous and are thus bounded by integrable functions independent of $\boldsymbol{\theta}$ on a closed ball about $\boldsymbol{\theta}_0$, i.e. by $e^{\eta_1^T Y_i}$, $e^{\eta_2^T Y_i} Y_{ij}$, $e^{\eta_3^T Y_i} Y_{ij} Y_{ik}$ and $e^{\eta_4^T Y_i} Y_{ij} Y_{ik} Y_{il}$ where the η_i $i = 1, \ldots, 4$ refer to the maximizers of each function on the said closed ball. This implies assumption (iii) of Lemma 3 holds by the LLN.

□

We should note that the results stated here are asymptotic, and that the existence and uniqueness of $\hat{\boldsymbol{\theta}}_{m,n}$ for any fixed m and n and any given realization of $\mathbf{X}_1, \ldots, \mathbf{X}_n$ and $\mathbf{Y}_1, \ldots, \mathbf{Y}_m$ is not guaranteed. However, for a continuous distribution existence implies uniqueness. This follows from the fact that $\frac{\partial^2}{\partial \boldsymbol{\theta}^2} \hat{\psi}(\boldsymbol{\theta})$ is the covariance matrix of $\hat{F}_m(x, \boldsymbol{\theta})$. For a discrete distribution, it is possible some observations to be tied, and the covariance matrix of the empirical kernel family distribution can be degenerate. For a continuous distribution, this occurs with probability 0 so that uniqueness follows by the inverse function theorem from the fact that $\frac{\partial^2}{\partial \boldsymbol{\theta}^2}\hat{\psi}(\boldsymbol{\theta})$ is positive definite for all $\boldsymbol{\theta}$.

The following lemma will help establish the asymptotic distribution of the properly normalized EMLE.

Lemma 1. *Let*

$$\mathbf{G}_{m,n}(\boldsymbol{\theta}) = \bar{\mathbf{X}}_n \frac{\sum_{i=1}^m e^{\boldsymbol{\theta}^T \mathbf{Y}_i}}{m} - \frac{\sum_{i=1}^m e^{\boldsymbol{\theta}^T \mathbf{Y}_i} \mathbf{Y}_i}{m}.$$

(i) If $m = cn(1 + o(1)), \quad c > 0,$

$$\sqrt{n}\mathbf{G}_{m,n}(\boldsymbol{\theta}) \xrightarrow{d} N_p(\mathbf{0}, [\boldsymbol{\psi}''(\boldsymbol{\theta})]e^{2\psi(\boldsymbol{\theta})} + c^{-1}\mathbf{A})$$

(ii) If $m = o(n)$,

$$\sqrt{m}\mathbf{G}_{m,n}(\boldsymbol{\theta}) \xrightarrow{d} N_p(\mathbf{0}, \mathbf{A})$$

(iii) If $n = o(m)$,

$$\sqrt{n}\mathbf{G}_{m,n}(\boldsymbol{\theta}) \xrightarrow{d} N_p(\mathbf{0}, e^{2\psi(\boldsymbol{\theta})}[\boldsymbol{\psi}''(\boldsymbol{\theta})])$$

where

$$\mathbf{A} = \mathbf{A}(\boldsymbol{\theta}) = e^{\psi(2\boldsymbol{\theta})}([\boldsymbol{\psi}''(2\boldsymbol{\theta})] + [\boldsymbol{\psi}'(2\boldsymbol{\theta})][\boldsymbol{\psi}'(2\boldsymbol{\theta})]^T - [\boldsymbol{\psi}'(2\boldsymbol{\theta})][\boldsymbol{\psi}'(\boldsymbol{\theta})]^T$$

$$-[\boldsymbol{\psi}'(\boldsymbol{\theta})][\boldsymbol{\psi}'(2\boldsymbol{\theta})]^T + [\boldsymbol{\psi}'(\boldsymbol{\theta})][\boldsymbol{\psi}'(\boldsymbol{\theta})]^T).$$

Proof.

$$\begin{aligned}\mathbf{G}_{m,n}(\boldsymbol{\theta}) &= (\bar{\mathbf{X}}_n \frac{\sum_{i=1}^m \exp\{\boldsymbol{\theta}^T \mathbf{Y}_i\}}{m} - \frac{\sum_{i=1}^m \exp\{\boldsymbol{\theta}^T \mathbf{Y}_i\}\mathbf{Y}_i}{m}) \\ &= (\bar{\mathbf{X}}_n - \boldsymbol{\psi}'(\boldsymbol{\theta}))\frac{\sum_{i=1}^m \exp\{\boldsymbol{\theta}^T \mathbf{Y}_i\}}{m} - \frac{\sum_{i=1}^m \exp\{\boldsymbol{\theta}^T \mathbf{Y}_i\}(\mathbf{Y}_i - \boldsymbol{\psi}'(\boldsymbol{\theta}))}{m}.\end{aligned} \tag{1.21}$$

To prove (i) multiply (1.21) by $\sqrt{n}$:

$$\sqrt{n}\mathbf{G}_{m,n}(\boldsymbol{\theta}) = \sqrt{n}(\bar{\mathbf{X}}_n - \boldsymbol{\psi}'(\boldsymbol{\theta}))\frac{\sum_{i=1}^m \exp\{\boldsymbol{\theta}^T \mathbf{Y}_i\}}{m} - \sqrt{n}\frac{\sum_{i=1}^m \exp\{\boldsymbol{\theta}^T \mathbf{Y}_i\}(\mathbf{Y}_i - \boldsymbol{\psi}'(\boldsymbol{\theta}))}{m}. \tag{1.22}$$

By the Central Limit Theorem (CLT), provided $n \to \infty$,

$$\sqrt{n}(\bar{\mathbf{X}}_n - \boldsymbol{\psi}'(\boldsymbol{\theta})) \xrightarrow{d} \mathbf{N}_p(\mathbf{0}, \boldsymbol{\psi}''(\boldsymbol{\theta})). \tag{1.23}$$

By (1.12) if $m \to \infty$

$$\frac{1}{m}\sum_{i=1}^m \exp\{\boldsymbol{\theta}^T \mathbf{Y}_i\} \xrightarrow{a.s.} \exp\{\psi(\boldsymbol{\theta})\}.$$

Thus

$$\sqrt{n}(\bar{\mathbf{X}}_n - \boldsymbol{\psi}'(\boldsymbol{\theta}))\frac{\sum_{i=1}^m \exp\{\boldsymbol{\theta}^T \mathbf{Y}_i\}}{m} \xrightarrow{d} \mathbf{N}_p(0, \boldsymbol{\psi}''(\boldsymbol{\theta})e^{2\psi(\boldsymbol{\theta})}). \tag{1.24}$$

Because $m/n \to c$, we replace n with m/c in the second term of the right hand side of (1.22). The CLT applies to

$$\frac{\sum_{i=1}^m \exp\{\boldsymbol{\theta}^T \mathbf{Y}_i\}(\mathbf{Y}_i - \boldsymbol{\psi}'(\boldsymbol{\theta}))}{\sqrt{cm}}.$$

Note that $E(\exp\{\boldsymbol{\theta}^T \mathbf{Y}\}(\mathbf{Y} - \boldsymbol{\psi}'(\boldsymbol{\theta}))) = \mathbf{0}$.

The asymptotic covariance matrix is

$$E[((\mathbf{Y}_i - \boldsymbol{\psi}'(\boldsymbol{\theta}))\exp\{\boldsymbol{\theta}^T \mathbf{Y}_i\})((\mathbf{Y}_i - \boldsymbol{\psi}'(\boldsymbol{\theta}))\exp\{\boldsymbol{\theta}^T \mathbf{Y}_i\})^T]$$

$$= E[(\mathbf{Y}_i\mathbf{Y}_i^T - \mathbf{Y}_i\boldsymbol{\psi}'(\boldsymbol{\theta})^T - \boldsymbol{\psi}'(\boldsymbol{\theta})\mathbf{Y}_i^T + \psi'(\boldsymbol{\theta})\boldsymbol{\psi}'(\boldsymbol{\theta})^T)\exp\{2\boldsymbol{\theta}^T \mathbf{Y}_i\}]$$

Using (1.8) the following relations hold:

$$E(\mathbf{Y}_i\mathbf{Y}_i^T \exp\{2\boldsymbol{\theta}^T\mathbf{Y}_i\}) = e^{\psi(2\boldsymbol{\theta})}\int \mathbf{y}_i\mathbf{y}_i^T \exp\{2\boldsymbol{\theta}^T\mathbf{y}_i\}e^{-\psi(2\boldsymbol{\theta})}dF(\mathbf{y}_i)$$

$$= e^{\psi(2\boldsymbol{\theta})}(\boldsymbol{\psi}''(2\boldsymbol{\theta}) + \boldsymbol{\psi}'(2\boldsymbol{\theta})\boldsymbol{\psi}'(2\boldsymbol{\theta})^T).$$

By (1.13) it follows that

$$E(\boldsymbol{\psi}'(\boldsymbol{\theta})\boldsymbol{\psi}'(\boldsymbol{\theta})^T \exp\{2\boldsymbol{\theta}^T\mathbf{Y}_i\}) = \boldsymbol{\psi}'(\boldsymbol{\theta})\boldsymbol{\psi}'(\boldsymbol{\theta})^T E(\exp\{2\boldsymbol{\theta}^T\mathbf{Y}_i\}) = \boldsymbol{\psi}'(\boldsymbol{\theta})\boldsymbol{\psi}'(\boldsymbol{\theta})^T e^{\psi(2\boldsymbol{\theta})}.$$

Likewise,

$$E(\mathbf{Y}_i\boldsymbol{\psi}'(\boldsymbol{\theta})^T \exp\{2\boldsymbol{\theta}^T\mathbf{Y}_i\}) = \boldsymbol{\psi}'(2\boldsymbol{\theta})\boldsymbol{\psi}'(\boldsymbol{\theta})^T e^{\psi(2\boldsymbol{\theta})}$$

and

$$E(\boldsymbol{\psi}'(\boldsymbol{\theta})\mathbf{Y}_i^T \exp\{2\boldsymbol{\theta}^T\mathbf{Y}_i\}) = \boldsymbol{\psi}'(\boldsymbol{\theta})\boldsymbol{\psi}'(2\boldsymbol{\theta})^T e^{\psi(2\boldsymbol{\theta})}.$$

If follows that

$$\frac{\sum_{i=1}^m e^{\boldsymbol{\theta}^T\mathbf{Y}_i}(\mathbf{Y}_i - \boldsymbol{\psi}'(\boldsymbol{\theta}))}{\sqrt{m}} \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{A}). \tag{1.25}$$

Combining (1.24) and (1.25), (i) follows.

To prove (ii), we write:

$$\sqrt{m}\mathbf{G}_{m,n}(\boldsymbol{\theta}) = \sqrt{\frac{m}{n}}\sqrt{n}(\bar{\mathbf{X}}_n - \boldsymbol{\psi}'(\boldsymbol{\theta}))\frac{\sum_{i=1}^m e^{\boldsymbol{\theta}^T\mathbf{Y}_i}}{m} - \frac{1}{\sqrt{m}}\sum_{i=1}^m e^{\boldsymbol{\theta}^T\mathbf{Y}_i}(\mathbf{Y}_i - \boldsymbol{\psi}'(\boldsymbol{\theta})) \tag{1.26}$$

The first term in (1.26) converges in probability to zero so that the asymptotic distribution is determined entirely by the second term. Thus (ii) follows from (1.25).

To prove (iii), we multiply (1.21) by $\sqrt{n}$ to obtain

$$\sqrt{n}\mathbf{G}(\boldsymbol{\theta}) = \sqrt{n}(\bar{\mathbf{X}}_n - \boldsymbol{\psi}'(\boldsymbol{\theta}))\frac{\sum_{i=1}^m e^{\boldsymbol{\theta}^T\mathbf{Y}_i}}{m} - \sqrt{\frac{n}{m}}\frac{1}{\sqrt{m}}\sum_{i=1}^m e^{\boldsymbol{\theta}^T\mathbf{Y}_i}(\mathbf{Y}_i - \boldsymbol{\psi}'(\boldsymbol{\theta})) \tag{1.27}$$

Note that since $n = o(m)$ the second term converges in probability to zero, so that the asymptotic distribution is determined by the first term. Thus (iii) follows from (1.25). □

We will now explore the asymptotic distribution of the properly normalized EMLE, as well as the impact of the assumption on the rate of growth of m relative to n.

Lemma 2. *Let $(T_{1,n}, \dots, T_{s,n})$ be a sequence of random vectors converging in distribution to $(T_1, \dots, T_s)$ and suppose that for each fixed j and k $A_{j,k,n}$ is a sequence of random variables converging in probability to constants $A_{j,k}$ for which the matrix $A = \{A\}_{j,k}$ is nonsingular. Let $B = \{B\}_{j,k} = A^{-1}$. Then, if the distribution of $(T_1, \dots, T_s)$ has a density with respect to Lebesgue measure over E_s, where E_s is s-dimensional Euclidean space, the solutions $(Y_{1,n}, \dots, Y_{s,n})$ of*

$$\sum_{k=1}^{s} A_{j,k,n} Y_{k,n} = T_{j,n}, \quad j = 1, \dots, s$$

converge in probability to the solutions of

$$\sum_{k=1}^{s} A_{j,k} Y_k = T_j, \quad j = 1, \dots, s$$

given by

$$Y_j = \sum_{k=1}^{s} B_{j,k} T_k.$$

Proof. See Lehmann and Casella (1998), Lemma 5.2. □

Theorem 2. *Let $m, n \to \infty$.*

(i) If $m = cn(1 + o(1))$, and $c > 0$,

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_{m,n} - \boldsymbol{\theta}) \xrightarrow{d} N_p(\mathbf{0}, \boldsymbol{\psi}''(\boldsymbol{\theta})^{-1} + \frac{1}{c}\mathbf{A}).$$

(ii) If $m = o(n)$,

$$\sqrt{m}(\hat{\boldsymbol{\theta}}_{m,n} - \boldsymbol{\theta}) \xrightarrow{d} N_p\left(\mathbf{0}, \mathbf{A}\right).$$

(iii) If $n = o(m)$,

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_{m,n} - \boldsymbol{\theta}) \xrightarrow{d} N_p\left(\mathbf{0}, \boldsymbol{\psi}''(\boldsymbol{\theta})^{-1}\right).$$

where

$$\mathbf{A} = \mathbf{A}(\boldsymbol{\theta}) = e^{\psi(2\boldsymbol{\theta})-2\psi(\boldsymbol{\theta})}\boldsymbol{\psi}''(\boldsymbol{\theta})^{-1}$$

$$\times[\boldsymbol{\psi}''(2\boldsymbol{\theta}) + \boldsymbol{\psi}'(2\boldsymbol{\theta})\boldsymbol{\psi}'(2\boldsymbol{\theta})^T - \boldsymbol{\psi}'(2\boldsymbol{\theta})\boldsymbol{\psi}'(\boldsymbol{\theta})^T - \boldsymbol{\psi}'(\boldsymbol{\theta})\boldsymbol{\psi}'(2\boldsymbol{\theta})^T + \boldsymbol{\psi}'(\boldsymbol{\theta})\boldsymbol{\psi}'(\boldsymbol{\theta})^T]\boldsymbol{\psi}''(\boldsymbol{\theta})^{-1}$$

Proof. Let $\boldsymbol{\theta}_{\mathbf{0}}$ denote the true parameter vector for $\boldsymbol{\theta}$. That is, $\mathbf{X}_1, \ldots, \mathbf{X}_n$ is a sample from $F(\mathbf{x}; \boldsymbol{\theta}_0)$ and $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ is a sample from $F(\mathbf{x}; \mathbf{0})$. For the mapping $\mathbf{G}_{m,n}(\boldsymbol{\theta})$ as defined in the previous lemma, expand each component function of $G_{m,n,j}(\boldsymbol{\theta})$ about $\boldsymbol{\theta}_0$ and plug in $\hat{\boldsymbol{\theta}}_{m,n}$ to obtain equations

$$G_{m,n,j}(\hat{\boldsymbol{\theta}}_{m,n}) = 0 = G_{m,n,j}(\boldsymbol{\theta}_0) + \sum_{k=1}^{p}(\hat{\theta}_{m,n,k} - \theta_{0,k})\frac{\partial}{\partial\theta_k}G_{m,n,j}(\boldsymbol{\theta}_0)$$
$$+\frac{1}{2}\sum_{k=1}^{p}\sum_{l=1}^{p}(\hat{\theta}_{m,n,k} - \theta_{0,k})(\hat{\theta}_{m,n,l} - \theta_{0,l})\frac{\partial^2}{\partial\theta_k\partial\theta_l}G_{m,n,j}(\boldsymbol{\theta}_*)$$
$$j = 1, \ldots, p,$$

where $\boldsymbol{\theta}^*$ is a point in the line segment connecting $\boldsymbol{\theta}_0$ and $\hat{\boldsymbol{\theta}}_{m,n}$. We then rewrite the equality as

$$\sum_{k=1}^{p}(\hat{\theta}_{m,n,k} - \theta_{0,k})\left[\frac{\partial}{\partial\theta_k}G_{m,n,j}(\boldsymbol{\theta}_0) + \frac{1}{2}\sum_{l=1}^{p}(\hat{\theta}_{m,n,l} - \theta_{0,l})\frac{\partial^2}{\partial\theta_k\partial\theta_l}G_{m,n,j}(\boldsymbol{\theta}_*)\right]$$
$$= -G_{m,n,j}(\boldsymbol{\theta}_0). \tag{1.28}$$

For part (i), multiply the equation above by $\sqrt{n}$ and obtain

$$\sum_{k=1}^{p} \sqrt{n}(\hat{\theta}_{m,n,k} - \theta_{0,k}) \left[\frac{\partial}{\partial \theta_k} G_{m,n,j}(\boldsymbol{\theta}_0) + \frac{1}{2} \sum_{l=1}^{p} (\hat{\theta}_{m,n,l} - \theta_{0,l}) \frac{\partial^2}{\partial \theta_k \partial \theta_l} G_{m,n,j}(\boldsymbol{\theta}_*) \right]$$

$$= -\sqrt{n} G_{m,n,j}(\boldsymbol{\theta}_0).$$

Let $A_{j,k,m,n} = \frac{\partial}{\partial \theta_k} G_{m,n,j}(\boldsymbol{\theta}_0) + \frac{1}{2} \sum_{l=1}^{p} (\hat{\theta}_{m,n,l} - \theta_{0,l}) \frac{\partial^2}{\partial \theta_k \partial \theta_l} G_{m,n,j}(\boldsymbol{\theta}_*)$ $Y_{m,n,k} = \sqrt{n}(\hat{\theta}_{m,n,k} - \theta_{0,k})$ and $T_{m,n,j} = -\sqrt{n} G_{m,n,j}(\boldsymbol{\theta}_0)$.

The asymptotic distribution $\mathbf{T}_n$ under each of the three asymptotic settings was obtained in Lemma 1. We now need to establish the behavior of $A_{j,k,n}$. The matrix $\frac{\partial}{\partial \boldsymbol{\theta}} \mathbf{G}_{m,n}(\boldsymbol{\theta})$ is given by entries

$$\frac{\partial}{\partial \theta_j} G_{m,n,k} = \bar{X}_{n,k} \frac{\sum_{i=1}^{m} e^{\boldsymbol{\theta}^T \mathbf{Y}_i} Y_{ij}}{m} - \frac{\sum_{i=1}^{m} e^{\boldsymbol{\theta}^T \mathbf{Y}_i} Y_{ik} Y_{ij}}{m}, \quad j,k = 1, \ldots, p. \tag{1.29}$$

From this expression it follows that as $m, n \to \infty$

$$\frac{\partial}{\partial \boldsymbol{\theta}} \mathbf{G}_{m,n}(\boldsymbol{\theta}_0, \bar{\mathbf{X}}_n, \mathbf{Y}_1, \ldots, \mathbf{Y}_m) \overset{a.s.}{\to} -\boldsymbol{\psi}''(\boldsymbol{\theta}_0) e^{\psi(\boldsymbol{\theta}_0)}. \tag{1.30}$$

Thus, the first term in $A_{j,k,n}$ converges almost surely to $-e^{\psi(\boldsymbol{\theta})}\{\boldsymbol{\psi}''(\boldsymbol{\theta})\}_{kj}$ provided $m, n \to \infty$. As for the second term, we have established consistency of $\hat{\boldsymbol{\theta}}_{m,n}$, so all we need to show is that $\frac{\partial^2}{\partial \theta_k \partial \theta_l} G_{m,n,j}(\boldsymbol{\theta}_*)$ is bounded in probability for each $j, k = 1, \ldots, p$.

$$\left| \frac{\partial^2}{\partial \theta_k \partial \theta_l} G_{m,n,j}(\boldsymbol{\theta}) \right| = |\bar{X}_{n,k} \frac{\sum_{i=1}^{m} e^{\boldsymbol{\theta}^T \mathbf{Y}_i} Y_{ij} Y_{il}}{m} - \frac{\sum_{i=1}^{m} e^{\boldsymbol{\theta}^T \mathbf{Y}_i} Y_{ik} Y_{ij} Y_{il}}{m}|$$

Note that on a closed ball $\mathbf{B}_{\boldsymbol{\theta}_0}$ around $\boldsymbol{\theta}_0$ the functions of $\boldsymbol{\theta}$, $e^{\boldsymbol{\theta}^T \mathbf{Y}_i} Y_{ij} Y_{il}$ and $e^{\boldsymbol{\theta}^T \mathbf{Y}_i} Y_{ik} Y_{ij} Y_{il}$, being continuous in $\boldsymbol{\theta}$, must have a maximizer and are thus dominated by the functions $e^{\boldsymbol{\eta}_1^T \mathbf{Y_i}} Y_{ij} Y_{il}$ and $e^{\boldsymbol{\eta}_2^T \mathbf{Y_i}} Y_{ik} Y_{ij} Y_{il}$ where $\boldsymbol{\eta}_1$ and $\boldsymbol{\eta}_2$ are the respective maximizers. The $P_{\boldsymbol{\theta}_0}$ probability of the event $\{\hat{\boldsymbol{\theta}}_{m,n} \in \mathbf{B}_{\theta_0}\}$ tends to one. On this event

$$\begin{aligned}
&\left|\bar{X}_{n,k}\frac{\sum_{i=1}^{m} e^{\boldsymbol{\theta}*^T\mathbf{Y}_i}Y_{ij}Y_{il}}{m} - \frac{\sum_{i=1}^{m} e^{\boldsymbol{\theta}*^T\mathbf{Y}_i}Y_{ik}Y_{ij}Y_{il}}{m}\right| \\
\leq &\left|\bar{X}_{n,k}\frac{\sum_{i=1}^{m} e^{\boldsymbol{\theta}*^T\mathbf{Y}_i}Y_{ij}Y_{il}}{m}\right| + \left|\frac{\sum_{i=1}^{m} e^{\boldsymbol{\theta}*^T\mathbf{Y}_i}Y_{ik}Y_{ij}Y_{il}}{m}\right| \\
\leq &\left|\bar{X}_{n,k}\frac{\sum_{i=1}^{m} e^{\boldsymbol{\eta}_1^T\mathbf{Y}_i}Y_{ij}Y_{il}}{m}\right| + \left|\frac{\sum_{i=1}^{m} e^{\boldsymbol{\eta}_2^T\mathbf{Y}_i}Y_{ik}Y_{ij}Y_{il}}{m}\right|
\end{aligned}$$

The SLLN guarantees the convergence of these sums provided that the expected values of the summands are finite. Thus this term is bounded in probability provided that $|E(e^{\boldsymbol{\theta}^T\mathbf{Y}_i}Y_{ij}Y_{il})| < \infty$ and $|E(e^{\boldsymbol{\theta}^T\mathbf{Y}_i}Y_{ik}Y_{ij}Y_{il})| < \infty$ for all $\boldsymbol{\theta}$ in a neighborhood of $\boldsymbol{\theta}_0$. We have already established the finiteness of these quantities on the interior of the parameter space. Applying Lemma 2, the result follows.

To prove part (ii), we multiply (1.28) by $\sqrt{m}$ to obtain

$$\begin{aligned}
&\sum_{k=1}^{p}\sqrt{m}(\hat{\theta}_{m,n,k}-\theta_{0,k})[\frac{\partial}{\partial\theta_k}G_{m,n,j}(\boldsymbol{\theta}_0)+\frac{1}{2}\sum_{l=1}^{p}(\hat{\theta}_{m,n,l}-\theta_{0,l})\frac{\partial^2}{\partial\theta_k\partial\theta_l}G_{m,n,j}(\boldsymbol{\theta}_*)] \\
&= -\sqrt{m}G_{m,n,j}(\boldsymbol{\theta}_0)
\end{aligned}$$

The rest of the arguments are identical to those of part (i) using part (ii) of Lemma 1. Similarly, part (iii) uses the the same arguments as part (i), where now we use part (iii) of Lemma 1. □

Theorem 2 has interesting implications. When m is very large relative to n, the asymptotic distribution of $\hat{\boldsymbol{\theta}}_{m,n}$ is the same as that of the MLE based on the sample $(\mathbf{X}_1, \ldots, \mathbf{X}_n)$ when the generator is known. When m and n are of the same order, the asymptotic covariance matrix is equal to $\boldsymbol{\psi}''(\boldsymbol{\theta})$ plus an additional term which depends on c. Note that the additional term must be positive semidefinite because it was obtained as $E(\mathbf{z}\mathbf{z}^T)$ for a nonzero random vector $\mathbf{z}$. This positive semidefinite term is the price of having to estimate the generator. When m grows at a slower rate than n, the asymptotic distribution is determined by m; that is, the normalizing factor is $\sqrt{m}$.

1.3 Kernel Families

The analysis performed in Section 1.2 for the NEFs can be applied to a broader class of families. Suppose P is a probability distribution on a measurable space $(\mathcal{X}, \mathcal{A})$ and $\boldsymbol{\theta} \in \Theta$, where $\Theta \in R^p$, and contains an open subset with the true parameter in the interior. Consider a positive function h such that

$$\int h(x; \boldsymbol{\theta}) dP(x) < \infty, \boldsymbol{\theta} \in \Theta. \tag{1.31}$$

The distribution P is generates a family parameterized by $\boldsymbol{\theta}$ with

$$dP_{\boldsymbol{\theta}}(x) = C(\boldsymbol{\theta}) h(x; \boldsymbol{\theta}) dP(x) \tag{1.32}$$

where the normalizing constant, $C(\boldsymbol{\theta})$ is given by

$$C(\boldsymbol{\theta}) = \left(\int h(x; \boldsymbol{\theta}) dP(x) \right)^{-1}. \tag{1.33}$$

We'll call the function h the kernel, and the distribution P the generator of (1.32). We assume that the model is identifiable. Conditions for identifiability are given in Gilbert *et al.* (1999).

The score of a random element X with probability element (1.32) $\mathbf{J}_X$ is given by

$$\mathbf{J}(x, \boldsymbol{\theta}) = \frac{\mathbf{C}'(\boldsymbol{\theta})}{C(\boldsymbol{\theta})} + \frac{\frac{\partial}{\partial \boldsymbol{\theta}} h(x, \boldsymbol{\theta})}{h(x, \boldsymbol{\theta})}. \tag{1.34}$$

Here, $\frac{\partial}{\partial \boldsymbol{\theta}} h(x, \boldsymbol{\theta})$ is the vector of partial derivatives of h with respect to the components of $\boldsymbol{\theta}$ and $C'(\boldsymbol{\theta})$ is defined analogously.

We assume that differentiation and integration can be interchanged in the following expression:

$$\int \frac{\partial}{\partial \boldsymbol{\theta}} h(x, \boldsymbol{\theta}) dP(x). \tag{1.35}$$

This implies $E(\mathbf{J}_X) = \mathbf{0}$ for a random element X. Note that the expression (1.35) is equal to

$$C(\boldsymbol{\theta})^{-1} \int \frac{\frac{\partial}{\partial \boldsymbol{\theta}} h(x, \boldsymbol{\theta})}{h(x, \boldsymbol{\theta})} h(x, \boldsymbol{\theta}) C(\boldsymbol{\theta}) dP(x) = C(\boldsymbol{\theta})^{-1} E_{\boldsymbol{\theta}}(\frac{\frac{\partial}{\partial \boldsymbol{\theta}} h(X, \boldsymbol{\theta})}{h(X, \boldsymbol{\theta})})$$

where here X denotes a random element from the kernel family population $P_{\boldsymbol{\theta}}$ and where we write the subscript $\boldsymbol{\theta}$ to emphasize that the expectation is taken with respect to the probability distribution $P_{\boldsymbol{\theta}}$. On the other hand, interchanging integration and differentiation in (1.35), gives

$$\int \frac{\partial}{\partial \boldsymbol{\theta}} h(x, \boldsymbol{\theta}) dP(x) = \frac{\partial}{\partial \boldsymbol{\theta}} \int h(x, \boldsymbol{\theta}) dP(x) = \frac{\partial}{\partial \boldsymbol{\theta}} \{C(\boldsymbol{\theta})^{-1}\} = -\frac{\mathbf{C}'(\boldsymbol{\theta})}{C(\boldsymbol{\theta})^2}.$$

Combining these relations gives

$$E_{\boldsymbol{\theta}} \left(\frac{\frac{\partial}{\partial \boldsymbol{\theta}} h(X, \boldsymbol{\theta})}{h(X, \boldsymbol{\theta})} \right) = -\frac{\mathbf{C}'(\boldsymbol{\theta})}{C(\boldsymbol{\theta})} \tag{1.36}$$

which implies that $E_{\boldsymbol{\theta}}(\mathbf{J}_{\mathbf{X}}) = \mathbf{0}$.

The matrix of Fisher information, defined as $\mathbf{I}_X(\boldsymbol{\theta}) = E(\mathbf{J}_X \mathbf{J}_X^T)$ is

$$\begin{aligned}
&E_{\boldsymbol{\theta}} \left((\frac{\mathbf{C}'(\boldsymbol{\theta})}{C(\boldsymbol{\theta})} + \frac{\frac{\partial}{\partial \boldsymbol{\theta}} h(X, \boldsymbol{\theta})}{h(X, \boldsymbol{\theta})})(\frac{\mathbf{C}'(\boldsymbol{\theta})}{C(\boldsymbol{\theta})} + \frac{\frac{\partial}{\partial \boldsymbol{\theta}} h(X, \boldsymbol{\theta})}{h(X, \boldsymbol{\theta})})^T \right) \\
&= \frac{\mathbf{C}'(\boldsymbol{\theta})}{C(\boldsymbol{\theta})} \frac{\mathbf{C}'(\boldsymbol{\theta})}{C(\boldsymbol{\theta})}^T + \frac{\mathbf{C}'(\boldsymbol{\theta})}{C(\boldsymbol{\theta})} E_{\boldsymbol{\theta}} \left(\frac{\frac{\partial}{\partial \boldsymbol{\theta}} h(X, \boldsymbol{\theta})}{h(X, \boldsymbol{\theta})} \right)^T + E_{\boldsymbol{\theta}} \left(\frac{\frac{\partial}{\partial \boldsymbol{\theta}} h(X, \boldsymbol{\theta})}{h(X, \boldsymbol{\theta})} \right) \frac{\mathbf{C}'(\boldsymbol{\theta})}{C(\boldsymbol{\theta})}^T \\
&+ E_{\boldsymbol{\theta}} \left[\left(\frac{\frac{\partial}{\partial \boldsymbol{\theta}} h(X, \boldsymbol{\theta})}{h(X, \boldsymbol{\theta})} \right) \left(\frac{\frac{\partial}{\partial \boldsymbol{\theta}} h(X, \boldsymbol{\theta})}{h(X, \boldsymbol{\theta})} \right)^T \right].
\end{aligned}$$

Using (1.36) we obtain:

$$\mathbf{I}_X(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}(\frac{\frac{\partial}{\partial \boldsymbol{\theta}} h(X,\boldsymbol{\theta}) \frac{\partial}{\partial \boldsymbol{\theta}} h(X,\boldsymbol{\theta})^T}{h(X,\boldsymbol{\theta})^2}) - \frac{\mathbf{C}'(\boldsymbol{\theta})\mathbf{C}'(\boldsymbol{\theta})^T}{C(\boldsymbol{\theta})^2}. \tag{1.37}$$

In this chapter, we assume $\mathbf{I}_X(\boldsymbol{\theta})$ is positive definite, $\mathbf{I}_X(\boldsymbol{\theta}) > 0$.

When the generator is known, since the kernel is always assumed to be known, (1.32) defines a parametric family. The consistency and asymptotic distributions of the maximum likelihood estimator $\hat{\boldsymbol{\theta}}_n$ and the method of moments estimator $\tilde{\boldsymbol{\theta}}_n$ are well known under regularity conditions when we have a sample $(\mathbf{X}_1, \ldots, \mathbf{X}_n)$ from a parametric family where the dimension of $\boldsymbol{\theta}$ is fixed. When F is known we have that under regularity conditions

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \xrightarrow{d} N_p(\mathbf{0}, \mathbf{I}_X(\boldsymbol{\theta})^{-1}) \tag{1.38}$$

and

$$\sqrt{n}(\tilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \xrightarrow{d} N_p\left[\mathbf{0}, \left(\frac{\partial}{\partial \boldsymbol{\theta}} \boldsymbol{\mu}_{\mathbf{X}}(\boldsymbol{\theta})^{-1}\right) \boldsymbol{\Sigma}_{\mathbf{X}}(\boldsymbol{\theta}) \left(\frac{\partial}{\partial \boldsymbol{\theta}} \boldsymbol{\mu}_{\mathbf{X}}(\boldsymbol{\theta})^{-1}\right)^T\right] \tag{1.39}$$

where $\boldsymbol{\mu}_{\mathbf{X}}(\boldsymbol{\theta})$ and $\boldsymbol{\Sigma}_{\mathbf{X}}(\boldsymbol{\theta})$ represent the mean vector and covariance matrix of observations $\mathbf{X}_i$.

The asymptotic relations (1.38) and (1.39) follow from the asymptotic theory of estimating equations (see, for instance, Van der Vaart (2000) [70]). In (1.39) it is implicitly assumed that the dimension of the $\mathbf{X}_i$ is the same as the dimension of $\boldsymbol{\theta}$, and that the square matrix $\frac{\partial}{\partial \boldsymbol{\theta}} \boldsymbol{\mu}_{\mathbf{X}}(\boldsymbol{\theta})$ is nonsingular so that its inverse is well-defined. These assumptions will be made throughout the rest of Chapter 1. Note that $\hat{\boldsymbol{\theta}}_n$ is asymptotically efficient.

We will provide an expression for $\frac{\partial}{\partial\boldsymbol{\theta}}\boldsymbol{\mu}_{\mathbf{X}}(\boldsymbol{\theta})$ which will be used subsequently. We assume the exchange of differentiation and integration is valid in the expression $\frac{\partial}{\partial\boldsymbol{\theta}}\int \mathbf{x}h(\mathbf{x},\boldsymbol{\theta})dP(\mathbf{x})$. This yields an expression for the components of matrix $\frac{\partial}{\partial\boldsymbol{\theta}}\mu_{\mathbf{X}}(\boldsymbol{\theta}_0)$:

$$\begin{aligned}\frac{\partial}{\partial\theta_k}\{\boldsymbol{\mu}_{\mathbf{X}}(\boldsymbol{\theta})\}_j &= \int x_j\frac{\partial}{\partial\theta_k}h(\mathbf{x},\boldsymbol{\theta})C(\boldsymbol{\theta})dP(\mathbf{x}) + \int x_j h(\mathbf{x},\boldsymbol{\theta}_0)\frac{\partial}{\partial\theta_k}C(\boldsymbol{\theta})dP(\mathbf{x}) \\ &= E(\frac{X_j\frac{\partial}{\partial\theta_k}h(\mathbf{X},\boldsymbol{\theta})}{h(\mathbf{X},\boldsymbol{\theta})}) + \{\boldsymbol{\mu}_{\mathbf{X}}(\boldsymbol{\theta})\}_j\frac{\frac{\partial}{\partial\theta_{\mathbf{k}}}C(\boldsymbol{\theta})}{C(\boldsymbol{\theta})}.\end{aligned}$$

Thus,

$$\frac{\partial}{\partial\theta_k}\{\boldsymbol{\mu}_{\mathbf{X}}(\boldsymbol{\theta})\}_j = E_{\boldsymbol{\theta}}(\frac{X_j\frac{\partial}{\partial\theta_k}h(\mathbf{X},\boldsymbol{\theta})}{h(\mathbf{X},\boldsymbol{\theta})}) + \{\boldsymbol{\mu}_{\mathbf{X}}(\boldsymbol{\theta})\}_j\frac{\frac{\partial}{\partial\theta_{\mathbf{k}}}C(\boldsymbol{\theta})}{C(\boldsymbol{\theta})}. \tag{1.40}$$

1.3.1 Empirical Maximum Likelihood for Kernel Families

Suppose the data are independent samples $(Y_1,\ldots,Y_m)$ from a population P and $(X_1,\ldots,X_n)$ from a population $P_{\boldsymbol{\theta}}$ from the corresponding kernel family.

The empirical kernel family can be constructed as given by the probability element

$$d\hat{P}_m(x;\boldsymbol{\theta}) = \hat{C}_m(\boldsymbol{\theta})h(x,\boldsymbol{\theta})d\hat{P}_m(x) \tag{1.41}$$

where $\hat{P}_m(x)$ is the empirical distribution of $(Y_1,\ldots,Y_m)$, and the normalizing constant is

$$\hat{C}_m(\boldsymbol{\theta}) = \left(\frac{1}{m}\sum_{i=1}^{m}h(y_i,\boldsymbol{\theta})\right)^{-1} \tag{1.42}$$

It follows that as $m \to \infty$,

$$\hat{C}_m(\boldsymbol{\theta}) \overset{a.s.}{\to} C(\boldsymbol{\theta}).$$

The empirical maximum likelihood estimator (EMLE) is defined as

$$\hat{\boldsymbol{\theta}}_{m,n} = \arg\max_{\boldsymbol{\theta}} \left[(\hat{C}_m(\boldsymbol{\theta}))^n \prod_{i=1}^{n} h(x_i, \boldsymbol{\theta}) \right]. \tag{1.43}$$

Taking the log of the quantity in brackets in (1.43) gives

$$l(\boldsymbol{\theta}) = n \log \hat{C}_m(\boldsymbol{\theta}) + \sum_{i=1}^{n} \log h(x_i, \boldsymbol{\theta})$$

and taking the derivative with respect to the vector parameter $\boldsymbol{\theta}$ gives the vector equation

$$\frac{\partial}{\partial \boldsymbol{\theta}} l(\boldsymbol{\theta}) = -n \frac{\sum_{i=1}^{m} \frac{\partial}{\partial \boldsymbol{\theta}} h(y_i, \boldsymbol{\theta})}{\sum_{i=1}^{m} h(y_i, \boldsymbol{\theta})} + \sum_{i=1}^{n} \frac{\frac{\partial}{\partial \boldsymbol{\theta}} h(x_i, \boldsymbol{\theta})}{h(x_i, \boldsymbol{\theta})}. \tag{1.44}$$

The empirical maximum likelihood system of p equations is found by setting this expression equal to $\mathbf{0}$.

Note that an equivalent expression is

$$\frac{1}{m} \sum_{i=1}^{m} \frac{\partial}{\partial \boldsymbol{\theta}} h(y_i, \boldsymbol{\theta}) - \frac{1}{nm} \sum_{i=1}^{m} h(y_i, \boldsymbol{\theta}) \sum_{i=1}^{n} \frac{\frac{\partial}{\partial \boldsymbol{\theta}} h(x_i, \boldsymbol{\theta})}{h(x_i, \boldsymbol{\theta})} = 0. \tag{1.45}$$

It is easy to see that (1.45) is an unbiased estimating function.

The following Lemma will be used to establish consistency of the EMLE.

Lemma 3. *Let $\boldsymbol{\theta}_0$ denote the true parameter.*

Let $g_{m,n}(x_1, \ldots, x_n, y_1, \ldots, y_m, \boldsymbol{\theta})$ be a function that is thrice differentiable with respect to $\boldsymbol{\theta}$ in a neighborhood of $\boldsymbol{\theta}_0$ for all m, n.

Let $\mathbf{G}_{m,n}(x_1, \ldots, x_n, y_1, \ldots, y_m, \boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} g_{m,n}(y_1, \ldots, x_n, y_1, \ldots, y_m, \boldsymbol{\theta})$.

Suppose that as $m, n \to \infty$

(i) $\mathbf{G}_{m,n}(X_1, \ldots, X_n, Y_1, \ldots, Y_m, \boldsymbol{\theta}_0) \xrightarrow{p} \mathbf{0}$.

(ii) $\frac{\partial}{\partial \boldsymbol{\theta}} \mathbf{G}_{m,n}(X_1, \ldots, X_n, Y_1, \ldots, Y_m, \boldsymbol{\theta}_0) \xrightarrow{p} \mathbf{A}$ where $\mathbf{A}$ is negative definite.

(iii) There exist functions $M_{m,n,j,k,l}(x_1, \ldots, x_n, y_1, \ldots, y_m)$ such that

$$\left| \frac{\partial^2}{\partial \theta_k \partial \theta_l} G_{m,n,j}(x_1, \ldots, x_n, y_1, \ldots, y_m, \boldsymbol{\theta}) \right| \leq M_{m,n,j,k,l}(x_1, \ldots, x_n, y_1, \ldots, y_m)$$

for all m, n, j, k, l and for all $\boldsymbol{\theta}$ in a neighborhood of $\boldsymbol{\theta}_0$ where as $m, n \to \infty$, $M_{m,n,j,k,l}(X_1, \ldots, X_n, Y_1, \ldots, Y_m) \xrightarrow{p} m_{j,k,l}$.

Then with probability tending to 1 as $m, n \to \infty$ there exists a solution of $\hat{\boldsymbol{\theta}}_{m,n} = \hat{\boldsymbol{\theta}}_{m,n}(X_1, \ldots, X_n, Y_1, \ldots, Y_m)$ of $G_{m,n}(\boldsymbol{\theta}) = \mathbf{0}$ such that $\hat{\boldsymbol{\theta}}_{m,n}$ is a consistent estimator of $\boldsymbol{\theta}_0$.

Proof. Let $\mathcal{Q}_a$ be a sphere of radius a. We will show that for any a sufficiently small, the probability tends to 1 that

$$g_{m,n}(\boldsymbol{\theta}) < g_{m,n}(\boldsymbol{\theta}_0)$$

at all points $\boldsymbol{\theta}$ on the surface of $\mathcal{Q}_a$, so that $g_{m,n}(\boldsymbol{\theta})$ has a local maximum in the interior of $\mathcal{Q}_a$. At a local maximum $G_{m,n}(\boldsymbol{\theta}) = 0$, so that with probability tending to one there is a solution $\hat{\boldsymbol{\theta}}_{m,n}$ which is consistent.

To show that with probability tending to one $g_{m,n}(\boldsymbol{\theta}) < g_{m,n}(\boldsymbol{\theta}_0)$ at all points $\boldsymbol{\theta}$ on the surface of $\mathcal{Q}_a$ for any a sufficiently small, we expand $g_{m,n}(\boldsymbol{\theta})$ at $\boldsymbol{\theta}_0$:

$$\begin{aligned} g_{m,n}(\boldsymbol{\theta}) - g_{m,n}(\boldsymbol{\theta}_0) &= \sum_j G_j(\boldsymbol{\theta}_0)(\theta_j - \theta_{0,j}) + \frac{1}{2}\sum_{j,k} \frac{\partial}{\partial \theta_k} G_j(\boldsymbol{\theta}_0)(\theta_j - \theta_{0,j})(\theta_k - \theta_{0,k}) \\ &+ \frac{1}{6}\sum_{j,k,l} (\theta_j - \theta_{0,j})(\theta_k - \theta_{0,k})(\theta_l - \theta_{0,l})\gamma_{j,k,l,m,n} M_{m,n,j,k,l} \\ &= S_1 + S_2 + S_3 \end{aligned}$$

where by assumption (iii), $|\gamma_{j,k,l,m,n}| \leq 1$ with probability tending to one.

Note that the dependence of $\gamma_{j,k,l,m,n} = \gamma_{j,k,l,m,n}(X_1, \ldots, X_m, Y_1, \ldots, Y_m)$ and $M_{j,k,l,m,n} = M_{j,k,l,m,n}(X_1, \ldots, X_m, Y_1, \ldots, Y_m)$ on the data are suppressed in the last equation for simplicity of exposition.

We will show that $\max_{\boldsymbol{\theta}}(S_1 + S_2 + S_3) < 0$ for all $\boldsymbol{\theta}$ in the surface of $\mathcal{Q}_a$ for all a sufficiently small with probability tending to 1.

Let's consider S_1. By assumption (i) it follows that with probability tending to one for any given a, $|G_{m,n,j}(\boldsymbol{\theta}_0)| < a^2$ and hence $|S_1| < pa^3$.

For S_3 we have that by assumption (iii) with probability tending to one $|S_3| < ba^3$ where $b = \frac{p^3}{3}\sum_{j,k,l} m_{j,k,l}$ for all points on $\mathcal{Q}_a$.

We will express $2S_2$ as follows:

$$2S_2 = \sum_{j,k} [A_{j,k}](\theta_j - \theta_{0,j})(\theta_k - \theta_{0,k}) + \sum_{j,k} \left[\frac{\partial}{\partial \theta_k} G_j(\boldsymbol{\theta}_0) - A_{j,k}\right](\theta_j - \theta_{0,j})(\theta_k - \theta_{0,k}).$$

For the second term, it follows from assumption (ii) that its absolute value is less than p^2a^3 with probability tending to one. The first term is a negative definite quadratic form, which, by an orthogonal transformation, can be reduced to $\sum_p \lambda_i \xi_i^2$ with $\sum_{i=1}^p \xi_i^2 = a^2$. Let $\lambda_{(1)}$ denote the smallest λ_i, and note that it must be negative. Then $\sum_{i=1}^p \lambda_i \xi_i^2 \leq \lambda_{(1)} \sum_{i=1}^p \xi_i^2 = \lambda_{(1)} a^2$. From this it follows that there exists constants $c > 0$ and $a_0 > 0$ such that for $a < a_o$

$$S_2 < -ca^2.$$

Combining these inequalities we obtain that with probability tending to one $\max(S_1 + S_2 + S_3) < -ca^2 + (b+s)a^3$, and the right hand side of the inequality is negative for any a that satisfies $a < \frac{c}{b+s}$. □

Remark: Although condition (iii) may seem artificial, it is satisfied by assuming very simple and natural conditions on $h(x, \boldsymbol{\theta})$. The conditions to satisfy assumption (iii) of Lemma 3 are satisfied, for instance, for the NEF's, as was illustrated in Section 1.2.

Theorem 3. *Under regularity conditions, with probability tending to one as $m, n \to \infty$ there exists a statistic $\hat{\boldsymbol{\theta}}_{m,n}$ which is a solution to the empirical maximum likelihood systems of equations and is a consistent estimator of $\boldsymbol{\theta}$.*

Proof. Let $\boldsymbol{\theta}_0$ be the true value of $\boldsymbol{\theta}$. We will use Lemma 3.

Consider the mapping given by

$$\mathbf{G}_{m,n} = \frac{1}{n} \sum_{i=1}^{n} \frac{\frac{\partial}{\partial \boldsymbol{\theta}} h(x_i, \boldsymbol{\theta})}{h(x_i, \boldsymbol{\theta})} - \frac{\sum_{i=1}^{m} \frac{\partial}{\partial \boldsymbol{\theta}} h(y_i, \boldsymbol{\theta})/m}{\sum_{i=1}^{m} h(y_i, \boldsymbol{\theta})/m} \tag{1.46}$$

We will show that:

$$\mathbf{G}_{m,n}(\boldsymbol{\theta}_0, Y_1, \ldots, Y_m, X_1, \ldots, X_n) \xrightarrow{p} \mathbf{0} \tag{1.47}$$

and

$$\frac{\partial}{\partial\boldsymbol{\theta}}\mathbf{G}_{m,n}(\boldsymbol{\theta}_0, Y_1, \ldots, Y_m, X_1, \ldots, X_n) \overset{p}{\to} -\mathbf{I}_X(\boldsymbol{\theta}_0). \tag{1.48}$$

We will also explore the conditions under which condition (iii) of Lemma 3 hold.

Relation (1.47) follows from the SLLN. That is,

$$\mathbf{G}_{m,n}(\boldsymbol{\theta}_0, Y_1, \ldots, Y_m, X_1, \ldots, X_n) \overset{p}{\to}$$
$$\left[\int \frac{\frac{\partial}{\partial\boldsymbol{\theta}}h(x,\boldsymbol{\theta}_0)}{h(x,\boldsymbol{\theta}_0)} h(x,\boldsymbol{\theta}_0)C(\boldsymbol{\theta}_0)dP(x)\right] - \int \frac{\partial}{\partial\boldsymbol{\theta}}h(x,\boldsymbol{\theta}_0)dP(x)C(\boldsymbol{\theta}_0) = \mathbf{0}$$

To prove (1.48), note that the jth component function of $\mathbf{G}_{m,n}$ is given by:

$$G_{m,n,j}(\boldsymbol{\theta}_0) = \frac{1}{n}\sum_{i=1}^{n} \frac{\frac{\partial}{\partial\theta_j}h(x_i,\boldsymbol{\theta})}{h(x_i,\boldsymbol{\theta})} - \frac{\sum_{i=1}^{m}\frac{\partial}{\partial\theta_j}h(y_i,\boldsymbol{\theta})}{\sum_{i=1}^{m}h(y_i,\boldsymbol{\theta})}.$$

Differentiating with respect to θ_k and evaluating at $\boldsymbol{\theta}_0$ gives:

$$\begin{aligned}
&\frac{\partial}{\partial\theta_k}G_{m,n,j}(\boldsymbol{\theta}_0) = \\
&\frac{1}{n}\sum_{i=1}^{n}\Big(\frac{\frac{\partial^2}{\partial\theta_j\partial\theta_k}h(X_i,\boldsymbol{\theta}_0)h(X_i,\boldsymbol{\theta}_0) - \frac{\partial}{\partial\theta_j}h(X_i,\boldsymbol{\theta}_0)\frac{\partial}{\partial\theta_k}h(X_i,\boldsymbol{\theta}_0)}{h(X_i,\boldsymbol{\theta}_0)^2}\Big) \\
&- \frac{\sum_{i=1}^{m}\frac{\partial^2}{\partial\theta_j\partial\theta_k}h(Y_i,\boldsymbol{\theta}_0)\sum_{i=1}^{m}h(Y_i,\boldsymbol{\theta}_0) - \sum_{i=1}^{m}\frac{\partial}{\partial\theta_j}h(Y_i,\boldsymbol{\theta}_0)\sum_{i=1}^{m}\frac{\partial}{\partial\theta_k}h(Y_i,\boldsymbol{\theta}_0)}{[\sum_{i=1}^{m}h(\mathbf{Y}_i,\boldsymbol{\theta}_0)]^2} \\
&\overset{a.s.}{\to} E_{\boldsymbol{\theta}_0}\Big(\frac{\frac{\partial^2}{\partial\theta_j\partial\theta_k}h(X,\boldsymbol{\theta}_0)}{h(x,\boldsymbol{\theta}_0)}\Big) - E_{\boldsymbol{\theta}_0}\Big(\frac{\frac{\partial}{\partial\theta_j}h(X,\boldsymbol{\theta}_0)\frac{\partial}{\partial\theta_k}h(X,\boldsymbol{\theta}_0)}{h(X,\boldsymbol{\theta}_0)^2}\Big) \\
&- E\left[\frac{\partial^2}{\partial\theta_j\partial\theta_k}h(Y,\boldsymbol{\theta}_0)\right]C(\boldsymbol{\theta}_0) + \left[E\left(\frac{\partial}{\partial\theta_j}h(Y_i,\boldsymbol{\theta}_0)\right)E\left(\frac{\partial}{\partial\theta_k}h(Y_i,\boldsymbol{\theta}_0)\right)\right]C(\boldsymbol{\theta}_0)^2.
\end{aligned}$$

The first and the third terms in the limit cancel since

$$\begin{aligned}
E_{\boldsymbol{\theta}_0}\Big(\frac{\frac{\partial^2}{\partial\theta_j\partial\theta_k}h(X,\boldsymbol{\theta}_0)}{h(X,\boldsymbol{\theta}_0)}\Big) &= \int \frac{\frac{\partial^2}{\partial\theta_j\partial\theta_k}h(x,\boldsymbol{\theta}_0)}{h(x,\boldsymbol{\theta}_0)}h(x,\boldsymbol{\theta}_0)C(\boldsymbol{\theta}_0)dP(x) \\
&= E\left[\frac{\partial^2}{\partial\theta_j\partial\theta_k}h(Y,\boldsymbol{\theta}_0)\right]C(\boldsymbol{\theta}_0).
\end{aligned}$$

The last term in the limit can be simplified. Note that

$$E\left[\frac{\partial}{\partial\theta_k}h(Y,\boldsymbol{\theta}_0)\right] = \int \frac{\partial}{\partial\theta_k}h(y,\boldsymbol{\theta}_0)dP(y)$$
$$= \frac{\partial}{\partial\theta_k}\int h(y,\boldsymbol{\theta}_0)dP(y) = \frac{\partial}{\partial\theta_k}\{C(\boldsymbol{\theta}_0)^{-1}\} = -\frac{\frac{\partial}{\partial\theta_k}C(\boldsymbol{\theta}_0)}{C(\boldsymbol{\theta}_0)^2}$$

Thus

$$\frac{\partial}{\partial\theta_k}G_{m,n,j} \xrightarrow{p} -E_{\boldsymbol{\theta}_0}(\frac{\frac{\partial}{\partial\theta_j}h(X,\boldsymbol{\theta}_0)\frac{\partial}{\partial\theta_k}h(X,\boldsymbol{\theta}_0)}{h(X,\boldsymbol{\theta}_0)^2}) + C(\boldsymbol{\theta}_0)^{-2}\frac{\partial}{\partial\theta_j}C(\boldsymbol{\theta}_0)\frac{\partial}{\partial\theta_k}C(\boldsymbol{\theta}_0)$$
$$= -\{\mathbf{I}_X(\boldsymbol{\theta}_0)\}_{j,k}$$

where the last equality follows from (1.37), so that

$$\frac{\partial}{\partial\boldsymbol{\theta}}\mathbf{G}_{m,n}(\boldsymbol{\theta}_0, X_1, \ldots, X_n, Y_1, \ldots, Y_m) \xrightarrow{p} -\mathbf{I}_X(\boldsymbol{\theta}_0).$$

We assume that all expected values involved in the use of the LLN in the expressions above are finite.

The matrix $\mathbf{I}_X(\boldsymbol{\theta}_0)$ is positive definite by assumption. Thus, condition (ii) in Lemma 3 is satisfied.

Condition (iii) follows by imposing some dominability conditions, which will be summarized shortly. □

Remark: For an alternative proof of consistency of the EMLE, see Appendix at the end of this chapter.

We will now list the regularity conditions that were involved in the preceding Theorem. In some of the theorems that follow, the regularity conditions are of a similar flavor, so we will not list them in order to avoid too many extensive list of conditions.

Regularity Conditions of Theorem 3.

- The matrix $\mathbf{I}_X(\boldsymbol{\theta}_0)$ is positive definite.

- The following expected values are finite:

$$E\left(\frac{\partial}{\partial\boldsymbol{\theta}}h(\mathbf{Y},\boldsymbol{\theta}_0)\right), \quad E\left(\frac{\partial^2}{\partial\boldsymbol{\theta}^2}h(\mathbf{Y},\boldsymbol{\theta}_0)\right), \quad E\left(\frac{[\frac{\partial}{\partial\boldsymbol{\theta}}h(\mathbf{Y},\boldsymbol{\theta}_0)][\frac{\partial}{\partial\boldsymbol{\theta}}h(\mathbf{Y},\boldsymbol{\theta}_0)]^T}{h(\mathbf{Y},\boldsymbol{\theta}_0)}\right).$$

  These conditions allow for the use of the LLN where needed.

- There exists functions $M_{j,k,l}(x)$, $N_{j,k,l}(x)$, and $L_{j,k,l}(x)$ such that for all $\boldsymbol{\theta}$ in a neighborhood of $\boldsymbol{\theta}_0$:

$$\begin{gathered}|\frac{\frac{\partial^3}{\partial\theta_j\partial\theta_k\partial\theta_l}h(\mathbf{x},\boldsymbol{\theta})}{h(\mathbf{x},\boldsymbol{\theta})}| < M_{j,k,l}(x), \quad |\frac{\frac{\partial^2}{\partial\theta_j\partial\theta_k}h(\mathbf{x},\boldsymbol{\theta})\frac{\partial}{\partial\theta_l}h(\mathbf{x},\boldsymbol{\theta})}{h(\mathbf{x},\boldsymbol{\theta})^2}| < N_{j,k,l}(x), \\ |\frac{\frac{\partial}{\partial\theta_j}h(\mathbf{x},\boldsymbol{\theta})\frac{\partial}{\partial\theta_k}h(\mathbf{x},\boldsymbol{\theta})\frac{\partial}{\partial\theta_l}h(\mathbf{x},\boldsymbol{\theta})}{h(\mathbf{x},\boldsymbol{\theta})^3}| < L_{j,k,l}(x),\end{gathered}$$

  with $E_{\boldsymbol{\theta}_0}(M_{j,k,l}(X)) < \infty$, $E_{\boldsymbol{\theta}_0}(N_{j,k,l}(X)) < \infty$, and $E_{\boldsymbol{\theta}_0}(L_{j,k,l}(X)) < \infty$,

$$j,k,l = 1,\ldots,p.$$

- There exists functions $R_j(y)$, $Q_{j,k}(y)$, $S_{j,k,l}(y)$ and $H_j(y)$ such that for all $\boldsymbol{\theta}$ in a neighborhood of $\boldsymbol{\theta}_0$

$$\begin{gathered}|\frac{\partial}{\partial\theta_j}h(\mathbf{y},\boldsymbol{\theta})| < R_j(y), \quad |\frac{\partial^2}{\partial\theta_j\partial\theta_k}h(\mathbf{y},\boldsymbol{\theta})| < Q_{j,k}(y), \\ |\frac{\partial^3}{\partial\theta_j\partial\theta_k\partial\theta_l}h(\mathbf{y},\boldsymbol{\theta})| < S_{j,k,l}(y), \quad |h(\mathbf{x},\boldsymbol{\theta})\frac{\partial}{\partial\theta_j}h(\mathbf{y},\boldsymbol{\theta})| < H_j(y),\end{gathered}$$

  with $E(R_j(Y)) < \infty$, $E(Q_{j,k}(Y)) < \infty$, $E(S_{j,k,l}(Y)) < \infty$ and $E(H_j(Y)) < \infty$,

$$j,k,l = 1,\ldots,p.$$

The last two bullets ensure that condition (iii) of Lemma 3 is met. Obviously, it is implicitly assumed that h is thrice differentiable.

We will now explore the asymptotic distribution of the properly normalized EMLE. We will first determine the distribution of $\mathbf{G}_{m,n}(\boldsymbol{\theta}, Y_1, \ldots, Y_m, X_1, \ldots, X_n)$.

Lemma 4. *Consider the mapping*

$$\mathbf{G}_{m,n}(\boldsymbol{\theta}, Y_1, \ldots, Y_m, X_1, \ldots, X_n) = \frac{\sum_{i=1}^{n} \frac{\frac{\partial}{\partial \boldsymbol{\theta}} h(X_i,\boldsymbol{\theta})}{h(X_i,\boldsymbol{\theta})}}{n} \frac{\sum_{i=1}^{m} h(Y_i,\boldsymbol{\theta})}{m} - \frac{\sum_{i=1}^{m} \frac{\partial}{\partial \boldsymbol{\theta}} h(Y_i,\boldsymbol{\theta})}{m}$$

Suppose $m, n \to \infty$. Under regularity conditions,

(i) If $m = cn(1 + o(1)), \quad c > 0,$

$$\begin{aligned} &\sqrt{n}\mathbf{G}_{m,n}(\boldsymbol{\theta}, Y_1, \ldots, Y_m, X_1, \ldots, X_n) \\ &\xrightarrow{d} N_p\left[\mathbf{0}, \mathbf{I}_X(\boldsymbol{\theta})C(\boldsymbol{\theta})^{-2} + \frac{1}{c}\int h(x,\boldsymbol{\theta})^2 \mathbf{J}(x;\boldsymbol{\theta})\mathbf{J}(x;\boldsymbol{\theta})^T dP(x)\right]. \end{aligned}$$

(ii) If $m = o(n)$, ,

$$\sqrt{m}\mathbf{G}_{m,n}(\boldsymbol{\theta}, Y_1, \ldots, Y_m, X_1, \ldots, X_n) \xrightarrow{d} N_p\left[\mathbf{0}, \int h(x,\boldsymbol{\theta})^2 \mathbf{J}(x;\boldsymbol{\theta})\mathbf{J}(x;\boldsymbol{\theta})^T dP(x)\right].$$

(iii) If $n = o(m)$

$$\sqrt{n}\mathbf{G}_{m,n}(\boldsymbol{\theta}, Y_1, \ldots, Y_m, X_1, \ldots, X_n) \xrightarrow{d} N_p(\mathbf{0}, C(\boldsymbol{\theta})^{-2}\mathbf{I}_X(\boldsymbol{\theta})).$$

Proof.

$$\begin{aligned} &\mathbf{G}_{m,n}(\boldsymbol{\theta}, Y_1, \ldots, Y_m, X_1, \ldots, X_n) \\ &= \frac{\sum_{i=1}^{n} \frac{\frac{\partial}{\partial \boldsymbol{\theta}} h(X_i,\boldsymbol{\theta})}{h(X_i,\boldsymbol{\theta})}}{n} \frac{\sum_{i=1}^{m} h(Y_i,\boldsymbol{\theta})}{m} - \frac{\sum_{i=1}^{m} \frac{\partial}{\partial \boldsymbol{\theta}} h(Y_i,\boldsymbol{\theta})}{m} \qquad (1.49) \\ &= (\sum_{i=1}^{n} \frac{\frac{\frac{\partial}{\partial \boldsymbol{\theta}} h(X_i,\boldsymbol{\theta})}{h(X_i,\boldsymbol{\theta})}}{n} + \frac{\mathbf{C}'(\boldsymbol{\theta})}{C(\boldsymbol{\theta})}) \frac{\sum_{i=1}^{m} h(Y_i,\boldsymbol{\theta})}{m} - \sum_{i=1}^{m} (\frac{\frac{\partial}{\partial \boldsymbol{\theta}} h(Y_i,\boldsymbol{\theta}) + h(Y_i,\boldsymbol{\theta})\frac{\mathbf{C}'(\boldsymbol{\theta})}{C(\boldsymbol{\theta})}}{m}) \end{aligned}$$

For part (i), we multiply equation (1.49) by $\sqrt{n}$ and substitute $n = m/c$ in the second term to obtain

$$\begin{aligned}\sqrt{n}\mathbf{G}_{m,n} = \sqrt{n}\left[\frac{1}{n}\sum_{i=1}^{n}\frac{\frac{\partial}{\partial\boldsymbol{\theta}}h(X_i,\boldsymbol{\theta})}{h(X_i,\boldsymbol{\theta})} + \frac{\mathbf{C}'(\boldsymbol{\theta})}{C(\boldsymbol{\theta})}\right]\frac{\sum_{i=1}^{m}h(Y_i,\boldsymbol{\theta})}{m} \\ - \frac{1}{\sqrt{m}\sqrt{c}}\sum_{i=1}^{m}(\frac{\partial}{\partial\boldsymbol{\theta}}h(Y_i,\boldsymbol{\theta}) + h(Y_i,\boldsymbol{\theta})\frac{\mathbf{C}'(\boldsymbol{\theta})}{C(\boldsymbol{\theta})}).\end{aligned} \tag{1.50}$$

By (1.34)

$$\sqrt{n}\left[\sum_{i=1}^{n}\frac{\frac{\frac{\partial}{\partial\boldsymbol{\theta}}h(X_i,\boldsymbol{\theta})}{h(X_i,\boldsymbol{\theta})}}{n} + \frac{\mathbf{C}'(\boldsymbol{\theta})}{C(\boldsymbol{\theta})}\right] \xrightarrow{d} N_p(\mathbf{0}, \mathbf{I}_X(\boldsymbol{\theta})). \tag{1.51}$$

By the SLLN

$$\frac{\sum_{i=1}^{m}h(Y_i,\boldsymbol{\theta})}{m} \xrightarrow{a.s.} C(\boldsymbol{\theta})^{-1}. \tag{1.52}$$

The last term of (1.50), namely

$$-\frac{1}{\sqrt{m}\sqrt{c}}\sum_{i=1}^{m}\left[\frac{\partial}{\partial\boldsymbol{\theta}}h(Y_i,\boldsymbol{\theta}) + h(Y_i,\boldsymbol{\theta})\frac{\mathbf{C}'(\boldsymbol{\theta})}{C(\boldsymbol{\theta})}\right], \tag{1.53}$$

is asymptotically normal. Note by (1.33) and by the interchange of integration and differentiation in (1.35),

$$E(\frac{\partial}{\partial\boldsymbol{\theta}}h(Y,\boldsymbol{\theta})) + \frac{1}{C(\boldsymbol{\theta})}E(h(Y,\boldsymbol{\theta}))\mathbf{C}'(\boldsymbol{\theta}_0) = -\frac{\mathbf{C}'(\boldsymbol{\theta})}{C(\boldsymbol{\theta})^2} + \frac{\mathbf{C}'(\boldsymbol{\theta})}{C(\boldsymbol{\theta})^2} = \mathbf{0}.$$

The asymptotic covariance matrix of (1.53) is given by the constant c times

$$
\begin{aligned}
& E\left\{\left[\frac{\partial}{\partial\boldsymbol{\theta}}h(Y,\boldsymbol{\theta})+\frac{h(Y,\boldsymbol{\theta})}{C(\boldsymbol{\theta})}\mathbf{C}'(\boldsymbol{\theta})\right]\left[\frac{\partial}{\partial\boldsymbol{\theta}}h(Y,\boldsymbol{\theta})+\frac{h(Y,\boldsymbol{\theta})}{C(\boldsymbol{\theta})}\mathbf{C}'(\boldsymbol{\theta})\right]^T\right\} \\
& =\int\left\{\frac{\partial}{\partial\boldsymbol{\theta}}h(y,\boldsymbol{\theta})\left[\frac{\partial}{\partial\boldsymbol{\theta}}h(y,\boldsymbol{\theta})\right]^T+\frac{h(y,\boldsymbol{\theta})}{C(\boldsymbol{\theta})}\mathbf{C}'(\boldsymbol{\theta})\left[\frac{\partial}{\partial\boldsymbol{\theta}}h(y,\boldsymbol{\theta})\right]^T\right\}dP(y) \\
& +\int\left\{\frac{h(y,\boldsymbol{\theta})}{C(\boldsymbol{\theta})}\left[\frac{\partial}{\partial\boldsymbol{\theta}}h(y,\boldsymbol{\theta})\right]\mathbf{C}'(\boldsymbol{\theta})^T+\left[\frac{h(y,\boldsymbol{\theta})}{C(\boldsymbol{\theta})}\right]^2\mathbf{C}'(\boldsymbol{\theta})\mathbf{C}'(\boldsymbol{\theta})^T\right\}dP(y) \\
& =\int h(y,\boldsymbol{\theta})^2\left\{\frac{1}{h(y,\boldsymbol{\theta})^2}\left[\frac{\partial}{\partial\boldsymbol{\theta}}h(y,\boldsymbol{\theta})\right]\frac{\partial}{\partial\boldsymbol{\theta}}h(y,\boldsymbol{\theta})^T+\frac{\mathbf{C}'(\boldsymbol{\theta})\frac{\partial}{\partial\boldsymbol{\theta}}h(y,\boldsymbol{\theta})^T}{h(y,\boldsymbol{\theta})C(\boldsymbol{\theta})}\right\}dP(y) \\
& +\int h(y,\boldsymbol{\theta})^2\left\{\frac{1}{C(\boldsymbol{\theta})h(y,\boldsymbol{\theta})}\left[\frac{\partial}{\partial\boldsymbol{\theta}}h(y,\boldsymbol{\theta})\right]\mathbf{C}'(\boldsymbol{\theta})^T+(\frac{1}{C(\boldsymbol{\theta})})^2\mathbf{C}'(\boldsymbol{\theta})\mathbf{C}'(\boldsymbol{\theta})^T\right\}dP(y) \\
& =\int h(y,\boldsymbol{\theta})^2(\frac{\mathbf{C}'(\boldsymbol{\theta})}{C(\boldsymbol{\theta})}+\frac{\frac{\partial}{\partial\boldsymbol{\theta}}h(y,\boldsymbol{\theta})}{h(y,\boldsymbol{\theta})})(\frac{\mathbf{C}'(\boldsymbol{\theta})}{C(\boldsymbol{\theta})}+\frac{\frac{\partial}{\partial\boldsymbol{\theta}}h(y,\boldsymbol{\theta})}{h(y,\boldsymbol{\theta})})^T dP(y) \\
& =\int h(y,\boldsymbol{\theta})^2\mathbf{J}(y;\boldsymbol{\theta})\mathbf{J}(y;\boldsymbol{\theta})^T dP(y).
\end{aligned}
$$

The asymptotic distribution of (1.53) is then given by:

$$
\frac{1}{\sqrt{m}\sqrt{c}}\sum_{i=1}^{m}(\frac{\partial}{\partial\boldsymbol{\theta}}h(Y_i,\boldsymbol{\theta})+h(Y_i,\boldsymbol{\theta})\frac{\mathbf{C}'(\boldsymbol{\theta})}{C(\boldsymbol{\theta})})\xrightarrow{d}\mathcal{N}_P(\mathbf{0},\frac{1}{c}\int h(y,\boldsymbol{\theta})^2\mathbf{J}(x;\boldsymbol{\theta})\mathbf{J}(x;\boldsymbol{\theta})^T dP(x)). \tag{1.54}
$$

Combining (1.51), (1.52), and (1.54) gives (i).

To prove part (ii) express (1.49) as

$$
\begin{aligned}
& \sqrt{m}\mathbf{G}_{m,n}(\boldsymbol{\theta},Y_1,\ldots,Y_m,X_1,\ldots,X_n) \\
& =\frac{\sqrt{m}}{\sqrt{n}}\sqrt{n}\left[\frac{1}{n}\sum_{i=1}^{n}\frac{\frac{\partial}{\partial\boldsymbol{\theta}}h(X_i,\boldsymbol{\theta})}{h(X_i,\boldsymbol{\theta})}+\frac{\mathbf{C}'(\boldsymbol{\theta})}{C(\boldsymbol{\theta})}\right]\frac{\sum_{i=1}^{m}h(Y_i,\boldsymbol{\theta})}{m} \\
& -\frac{1}{\sqrt{m}}\sum_{i=1}^{m}\left[\frac{\partial}{\partial\boldsymbol{\theta}}h(Y_i,\boldsymbol{\theta})+h(Y_i,\boldsymbol{\theta})\frac{\mathbf{C}'(\boldsymbol{\theta})}{C(\boldsymbol{\theta})}\right]
\end{aligned}
$$

and notice that the first term converges in probability to zero so that the asymptotic distribution is determined by the second term. Its asymptotic distribution was already derived, yielding (ii).

For part (iii) we write:

$$\sqrt{n}\mathbf{G}_{m,n}(\boldsymbol{\theta}, Y_1, \ldots, Y_m, X_1, \ldots, X_n) = \sqrt{n}(\sum_{i=1}^{n} \frac{\frac{\frac{\partial}{\partial \boldsymbol{\theta}} h(X_i, \boldsymbol{\theta})}{h(X_i, \boldsymbol{\theta})}}{n} + \frac{\mathbf{C}'(\boldsymbol{\theta})}{C(\boldsymbol{\theta})}) \frac{\sum_{i=1}^{m} h(Y_i, \boldsymbol{\theta})}{m}$$

$$-\frac{\sqrt{n}}{\sqrt{m}} \frac{1}{\sqrt{m}} \sum_{i=1}^{m} (\frac{\partial}{\partial \boldsymbol{\theta}} h(Y_i, \boldsymbol{\theta}) + h(Y_i, \boldsymbol{\theta}) \frac{\mathbf{C}'(\boldsymbol{\theta})}{C(\boldsymbol{\theta})})$$

and notice the second term converges in probability to zero, so that the asymptotic distribution of $\sqrt{m}\mathbf{G}_{m,n}$ is determined by the first term when $n = o(m)$. □

We are now ready to state the asymptotic distribution of the normalized EMLE for kernel families.

Theorem 4. *Suppose $m, n \to \infty$.*

(i) If $m = cn(1 + o(1)), \quad c > 0,$

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_{m,n} - \boldsymbol{\theta})$$
$$\xrightarrow{d} N_p \left[\mathbf{0}, \mathbf{I}_X(\boldsymbol{\theta})^{-1} + C(\boldsymbol{\theta})^2 \frac{1}{c} \mathbf{I}_X(\boldsymbol{\theta})^{-1} \int h(x, \boldsymbol{\theta})^2 \mathbf{J}(x; \boldsymbol{\theta}) \mathbf{J}(x; \boldsymbol{\theta})^T dP(x) \mathbf{I}_X(\boldsymbol{\theta})^{-1} \right].$$

(ii) If $m = o(n)$,

$$\sqrt{m}(\hat{\boldsymbol{\theta}}_{m,n} - \boldsymbol{\theta}) \xrightarrow{d} N_p(\mathbf{0}, C(\boldsymbol{\theta})^2 \mathbf{I}_X(\boldsymbol{\theta})^{-1} \int h(x, \boldsymbol{\theta})^2 \mathbf{J}(x; \boldsymbol{\theta}) \mathbf{J}(x; \boldsymbol{\theta})^T dP(x) \mathbf{I}_X(\boldsymbol{\theta})^{-1}).$$

(iii) If $n = o(m)$,

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_{m,n} - \boldsymbol{\theta}) \xrightarrow{d} N_p(\mathbf{0}, \mathbf{I}_X(\boldsymbol{\theta})^{-1}).$$

Proof. Let $\boldsymbol{\theta}_0$ be the true parameter. Using the mapping $\mathbf{G}_{m,n}(\boldsymbol{\theta})$ as defined in the previous lemma, we expand each component function $G_{m,n,j}(\boldsymbol{\theta})$ about $\boldsymbol{\theta}_0$ and plug in $\hat{\boldsymbol{\theta}}_{m,n}$ to obtain equations

$$\begin{aligned} G_{m,n,j}(\hat{\boldsymbol{\theta}}) = 0 & \\ &= G_{m,n,j}(\boldsymbol{\theta}_0) + \sum_{k=1}^{p}(\hat{\theta}_{m,n,k} - \theta_{0,k})\frac{\partial}{\partial\theta_k}G_{m,n,j}(\boldsymbol{\theta}_0) \\ &+ \frac{1}{2}\sum_{k=1}^{p}\sum_{l=1}^{p}(\hat{\theta}_{m,n,k} - \theta_{0,k})(\hat{\theta}_{m,n,l} - \theta_{0,l})\frac{\partial^2}{\partial\theta_k\partial\theta_l}G_j(\boldsymbol{\theta}_*), \\ & j = 1, \ldots, p, \end{aligned} \tag{1.55}$$

where $\boldsymbol{\theta}_*$ is a point in the line segment connecting $\hat{\boldsymbol{\theta}}_{m,n}$ and $\boldsymbol{\theta}_0$.

These equations can be expressed as

$$\begin{aligned} &\sum_{k=1}^{p}(\hat{\theta}_{m,n,k} - \theta_{0,k})[\frac{\partial}{\partial\theta_k}G_{m,n,j}(\boldsymbol{\theta}_0) + \frac{1}{2}\sum_{l=1}^{p}(\hat{\theta}_{m,n,l} - \theta_{0,l})\frac{\partial^2}{\partial\theta_k\partial\theta_{m,n,l}}G_{m,n,j}(\boldsymbol{\theta}_*)] \\ &= -G_{m,n,j}(\boldsymbol{\theta}_0) \end{aligned} \tag{1.56}$$

for $j = 1, \ldots, p$.

Again, we use Lemma 2. We must consider the behavior of

$$A_{j,k,n,m} = \frac{\partial}{\partial\theta_k}G_{m,n,j}(\boldsymbol{\theta}_0) + \frac{1}{2}\sum_{l=1}^{p}(\hat{\theta}_{m,n,l} - \theta_{0,l})\frac{\partial^2}{\partial\theta_k\partial\theta_l}G_{m,n,j}(\boldsymbol{\theta}_*).$$

Recall that the jth component function $\mathbf{G}_{m,n}$ is given by:

$$G_{m,n,j}(\boldsymbol{\theta}_0) = \frac{1}{n}\sum_{i=1}^{n}\frac{\frac{\partial}{\partial\theta_j}h(X_i, \boldsymbol{\theta}_0)}{h(X_i, \boldsymbol{\theta}_0)}\frac{\sum_{i=1}^{m}h(Y_i, \boldsymbol{\theta}_0)}{m} - \frac{\sum_{i=1}^{m}\frac{\partial}{\partial\theta_j}h(Y_i, \boldsymbol{\theta}_0)}{m}.$$

Differentiating with respect to θ_k and evaluating at $\boldsymbol{\theta}_0$ gives:

$$
\begin{aligned}
&\frac{\partial}{\partial\theta_k} G_{m,n,j}(\boldsymbol{\theta}_0) = \\
&\left[\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\frac{\partial}{\partial\theta_j}kh(X_i,\boldsymbol{\theta}_0)h(X_i,\boldsymbol{\theta}_0) - \frac{\partial}{\partial\theta_j}h(X_i,\boldsymbol{\theta}_0)\frac{\partial}{\partial\theta_k}h(X_i,\boldsymbol{\theta}_0)}{h(X_i,\boldsymbol{\theta}_0)^2}\right)\right] \frac{\sum_{i=1}^{m} h(Y_i,\boldsymbol{\theta}_0)}{m} \\
&+ \left(\frac{\sum_{i=1}^{m}\frac{\partial}{\partial\theta_k}h(Y_i,\boldsymbol{\theta}_0)}{m}\right)\frac{\sum_{i=1}^{n}\frac{\frac{\partial}{\partial\theta_j}h(X_i,\boldsymbol{\theta}_0)}{h(X_i,\boldsymbol{\theta}_0)}}{n} - \frac{\sum_{i=1}^{m}\frac{\partial}{\partial\theta_j}kh(Y_i,\boldsymbol{\theta}_0)}{m} \\
&\overset{a.s.}{\to} \left[E_{\boldsymbol{\theta}_0}\left(\frac{\frac{\partial}{\partial\theta_j}kh(X,\boldsymbol{\theta}_0)}{h(x,\boldsymbol{\theta}_0)}\right) - E_{\boldsymbol{\theta}_0}\left(\frac{\frac{\partial}{\partial\theta_j}h(X,\boldsymbol{\theta}_0)\frac{\partial}{\partial\theta_k}h(X,\boldsymbol{\theta}_0)}{h(X,\boldsymbol{\theta}_0)^2}\right)\right] C(\boldsymbol{\theta}_0)^{-1} \\
&+ E\left(\frac{\partial}{\partial\theta_k}h(Y,\boldsymbol{\theta}_0)\right)E_{\boldsymbol{\theta}_0}\left(\frac{\frac{\partial}{\partial\theta_j}h(X,\boldsymbol{\theta}_0)}{h(X,\boldsymbol{\theta}_0)}\right) - E\left(\frac{\partial^2}{\partial\theta_j\partial\theta_k}h(Y,\boldsymbol{\theta}_0)\right).
\end{aligned}
$$

The first and the last terms in the limit cancel since

$$
\begin{aligned}
E_{\boldsymbol{\theta}_0}\left(\frac{\frac{\partial^2}{\partial\theta_j\partial\theta_k}h(X,\boldsymbol{\theta}_0)}{h(X,\boldsymbol{\theta}_0)}\right) &= \int \frac{\frac{\partial^2}{\partial\theta_j\partial\theta_k}h(x,\boldsymbol{\theta}_0)}{h(x,\boldsymbol{\theta}_0)} h(x,\boldsymbol{\theta}_0)C(\boldsymbol{\theta}_0)dP(x) \\
&= E\left(\frac{\partial^2}{\partial\theta_j\partial\theta_k}h(Y,\boldsymbol{\theta}_0)\right)C(\boldsymbol{\theta}_0).
\end{aligned}
$$

The third term in the limit can be simplified. Note that

$$
\begin{aligned}
E\left(\frac{\partial}{\partial\theta_k}h(Y,\boldsymbol{\theta}_0)\right) &= \int \frac{\partial}{\partial\theta_k}h(y,\boldsymbol{\theta}_0)dP(y) \\
&= C(\boldsymbol{\theta}_0)^{-1}\int \frac{\frac{\partial}{\partial\theta_k}h(y,\boldsymbol{\theta}_0)}{h(y,\boldsymbol{\theta}_0)} h(y,\boldsymbol{\theta}_0)C(\boldsymbol{\theta}_0)dP(y) \\
&= C(\boldsymbol{\theta}_0)^{-1}E_{\boldsymbol{\theta}_0}\left(\frac{\frac{\partial}{\partial\theta_k}h(X,\boldsymbol{\theta}_0)}{h(X,\boldsymbol{\theta}_0)}\right).
\end{aligned}
$$

Moreover, by (1.36)

$$
E_{\boldsymbol{\theta}_0}\left(\frac{\frac{\partial}{\partial\theta_j}h(X,\boldsymbol{\theta}_0)}{h(X,\boldsymbol{\theta}_0)}\right) = -\frac{\frac{\partial}{\partial\theta_j}C(\boldsymbol{\theta}_0)}{C(\boldsymbol{\theta}_0)}
$$

so that the third term in the limit can be expressed as

$$
C(\boldsymbol{\theta}_0)^{-1}\frac{\frac{\partial}{\partial\theta_j}C(\boldsymbol{\theta}_0)}{C(\boldsymbol{\theta}_0)}\frac{\frac{\partial}{\partial\theta_k}C(\boldsymbol{\theta}_0)}{C(\boldsymbol{\theta}_0)}.
$$

Thus

$$\begin{aligned}
&\frac{\partial}{\partial \theta_k} G_{m,n,j}(\boldsymbol{\theta}_0) \\
&\overset{a.s.}{\rightarrow} -E_{\boldsymbol{\theta}_0}\Big(\frac{\frac{\partial}{\partial \theta_j} h(X,\boldsymbol{\theta}_0)\frac{\partial}{\partial \theta_k} h(X,\boldsymbol{\theta}_0)}{h(X,\boldsymbol{\theta}_0)^2}\Big)C(\boldsymbol{\theta}_0)^{-1} + C(\boldsymbol{\theta}_0)^{-3}\frac{\partial}{\partial \theta_j}C(\boldsymbol{\theta}_0)\frac{\partial}{\partial \theta_k}C(\boldsymbol{\theta}_0) \\
&= -C(\boldsymbol{\theta}_0)^{-1}\{\mathbf{I}_X(\boldsymbol{\theta}_0)\}_{jk}
\end{aligned}$$

where the last equality follows from (1.37), so that

$$\frac{\partial}{\partial \boldsymbol{\theta}}\mathbf{G}_{m,n}(\boldsymbol{\theta}_0, X_1, \ldots, X_n, Y_1, \ldots, Y_m) \overset{a.s.}{\rightarrow} -C(\boldsymbol{\theta}_0)^{-1}\mathbf{I}_X(\boldsymbol{\theta}_0).$$

We now show that the second term in $A_{j,k,n,m}$ converges in probability to 0. Because we have already shown the consistency of $\hat{\boldsymbol{\theta}}_{m,n}$, we just need to show that $\frac{\partial^2}{\partial \theta_k \partial \theta_l} G_{m,n,j}(\boldsymbol{\theta}_*)$ is bounded in probability. As previously discussed, this follows by imposing some conditions on h.

The rest of the argument is the same as that of Theorem 2, using Lemma 2 and Lemma 4. □

1.3.2 Empirical Method of Moments Estimator for Kernel Families

Again, we assume that the sample $(\mathbf{Y}_1, \ldots, \mathbf{Y}_m)$ is from the population P and that the sample $(\mathbf{X}_1, \ldots, \mathbf{X}_n)$ is from the kernel family population $P_{\boldsymbol{\theta}}$, where $\mathbf{X}_i$, $\mathbf{Y}_i$, $\in R^p$, $\boldsymbol{\theta} \in \Theta \subset \mathcal{R}^p$, and where the two samples are independent, and the true parameter value is in the interior of the parameter space. As in the method of moments for parametric models, we assume that the observations are random vectors of the same dimension as that of $\boldsymbol{\theta}$, that $\boldsymbol{\mu}_{\mathbf{X}}(\boldsymbol{\theta})$ exists, and that the square matrix $\frac{\partial}{\partial \boldsymbol{\theta}} \boldsymbol{\mu}_{\mathbf{X}}(\boldsymbol{\theta})$ is nonsingular so that its inverse is well-defined.

As in Section (1.3.1), the probability element associated with the sample $(\mathbf{X}_1, \ldots, \mathbf{X}_n)$ is

$$dP_{\boldsymbol{\theta}}(\mathbf{x}) = C(\boldsymbol{\theta}) h(\mathbf{x}; \boldsymbol{\theta}) dP(\mathbf{x}) \tag{1.57}$$

and the probability element associated with the sample $(\mathbf{Y}_1, \ldots, \mathbf{Y}_m)$ is $dP(\mathbf{x})$. We construct the empirical kernel family which has the probability element:

$$d\hat{P}(\mathbf{x}; \boldsymbol{\theta}) = \hat{C}_m(\boldsymbol{\theta}) h(\mathbf{x}, \boldsymbol{\theta}) d\hat{P}_m(\mathbf{x}) \tag{1.58}$$

where $\hat{P}_m(\mathbf{x})$ is the empirical distribution based on the sample from P, and the normalizing constant is

$$\hat{C}_m(\boldsymbol{\theta}) = \left(\frac{1}{m} \sum_{i=1}^{m} h(\mathbf{y}_i, \boldsymbol{\theta}) \right)^{-1}. \tag{1.59}$$

Since the empirical kernel family distribution puts the masses $h(\mathbf{y}_i, \boldsymbol{\theta}) / \sum h(\mathbf{y}_i, \boldsymbol{\theta})$

at the points $\mathbf{y}_i$ for $i = 1, \ldots, m$,

$$\int \mathbf{x} d\hat{P}_m(\mathbf{x}; \boldsymbol{\theta}) = \frac{\sum_{i=1}^m \mathbf{y}_i h(\mathbf{y}_i, \boldsymbol{\theta})}{\sum_{i=1}^m h(\mathbf{y}_i, \boldsymbol{\theta})}.$$

Thus the semiparametric analog of the method of moments system of estimating equations is

$$\bar{\mathbf{X}}_n = \frac{\sum_{i=1}^m \mathbf{Y}_i h(\mathbf{Y}_i, \boldsymbol{\theta})}{\sum_{i=1}^m h(\mathbf{Y}_i, \boldsymbol{\theta})}. \tag{1.60}$$

Note that the expected value of the left hand side is equal to $\boldsymbol{\mu}_{\mathbf{X}}(\boldsymbol{\theta})$, so that this system is unbiased.

We will call the solution of this system of equations the Empirical Method of Moments Estimator (EMME), denoted by $\tilde{\boldsymbol{\theta}}_{m,n}$.

We will now give a simple proof that shows that, under regularity conditions, the EMME is consistent when θ is a scalar. In this case, the EMME equation can be expressed as

$$G_{m,n}(X_1, \ldots, X_n, Y_1, \ldots, Y_m, \theta) = \bar{X}_n \frac{\sum_{i=1}^m h(Y_i, \theta)}{m} - \frac{\sum_{i=1}^m Y_i h(Y_i, \theta)}{m}. \tag{1.61}$$

Theorem 5. *Suppose θ is a scalar. Let θ_0 be the true parameter value and assume $E(Yh(Y, \theta)) < \infty$ and $\mu'(\theta) \neq 0$ in a neighborhood of θ_0. Suppose h is continuous. As $m, n \to \infty$, with probability tending to one there exists a zero of (1.61), denoted as $\tilde{\theta}_{m,n}$, which is a consistent estimator of θ_0.*

Proof. Notice that

$$G_{m,n}(\theta_0+\epsilon) \xrightarrow{d} \mu(\theta_0)C(\theta+\epsilon)^{-1} - \mu(\theta_0+\epsilon)C(\theta_0+\epsilon)^{-1} = C(\theta_0+\epsilon)^{-1}(\mu(\theta_0)-\mu(\theta_0+\epsilon)) \tag{1.62}$$

$$G_{m,n}(\theta_0-\epsilon) \xrightarrow{d} \mu(\theta_0)C(\theta_0-\epsilon)^{-1} - \mu(\theta_0-\epsilon)C(\theta-\epsilon)^{-1} = C(\theta_0-\epsilon)^{-1}(\mu(\theta_0)-\mu(\theta_0+\epsilon)) \tag{1.63}$$

Since $\mu'(\theta) \neq 0$ on an open interval containing θ_0, it follows that the limits in (1.62) and (1.63) are of opposite signs since $C(\theta)$ is strictly positive for all θ. By continuity of $G_{m,n}(\theta)$ it follows that with probability tending to one there is a zero $\tilde{\theta}_{m,n}$ of $G_{m,n}(\theta)$ which is consistent. □

We will now prove consistency when $\boldsymbol{\theta}$ is a vector for a special case: namely, when

$$h(\mathbf{x}, \boldsymbol{\theta}) = \sum_{k=1}^{s}(W_k(\boldsymbol{\theta})\psi_k(\mathbf{x})). \tag{1.64}$$

In this case the EMME equation is

$$\bar{\mathbf{X}}_n = \frac{\sum_{i=1}^{m} \mathbf{Y}_i \sum_{k=1}^{s}(W_k(\boldsymbol{\theta})\psi_k(\mathbf{Y}_i))}{\sum_{i=1}^{m}\sum_{k=1}^{s}(W_k(\boldsymbol{\theta})\psi_k(\mathbf{Y}_i))}.$$

Theorem 6 follows from Theorem 7, but we include it to illustrate how the implicit function theorem can be applied in this special case.

Theorem 6. *Suppose h is of the form (1.64). Then under regularity conditions, with probability tending to one as $m, n \to \infty$ there exists a statistic $\tilde{\boldsymbol{\theta}}_{m,n}$ which is a solution to the empirical method of moments system of equations and is a consistent estimator of $\boldsymbol{\theta}$.*

Proof. Let $\boldsymbol{\theta}_0$ be the true parameter value. Consider the function

$$\begin{aligned}
&\mathbf{G}_{m,n}(\bar{\mathbf{X}}_n, \mathbf{Y}_1, \ldots, \mathbf{Y}_n, \boldsymbol{\theta}) \\
&= \frac{\bar{\mathbf{X}}_n \sum_{i=1}^m \sum_{k=1}^s (W_k(\boldsymbol{\theta})\psi_k(\mathbf{Y}_i))}{m} - \frac{\sum_{i=1}^m \mathbf{Y}_i \sum_{k=1}^s (W_k(\boldsymbol{\theta})\psi_k(\mathbf{Y}_i))}{m}. \qquad (1.65) \\
&= \bar{\mathbf{X}}_n \frac{1}{m} \sum_{k=1}^s W_k(\boldsymbol{\theta}) \sum_{i=1}^m \psi_k(\mathbf{Y}_i) - \frac{1}{m} \sum_{k=1}^s W_k(\boldsymbol{\theta}) \sum_{i=1}^m \eta_k(\mathbf{Y}_i)
\end{aligned}$$

where $\eta_k(\mathbf{y}) = \mathbf{y}\psi_k(\mathbf{y})$.

Consider the functional

$$G(\bar{\mathbf{X}}_n, \bar{\psi}_{m,1}, \ldots, \bar{\psi}_{m,s}, \bar{\eta}_{m,1}, \ldots, \bar{\eta}_{m,s}, \boldsymbol{\theta}) = \bar{\mathbf{X}}_n \sum_{k=1}^s W_k(\boldsymbol{\theta})\bar{\psi}_{m,k} - \sum_{k=1}^s W_k(\boldsymbol{\theta})\bar{\eta}_{m,k} \tag{1.66}$$

where

$$\bar{\psi}_{m,k} = \frac{1}{m} \sum_{i=1}^m \psi_k(\mathbf{Y}_i) \tag{1.67}$$

$$\bar{\eta}_{m,k} = \frac{1}{m} \sum_{i=1}^m \eta_k \mathbf{Y}_i) = \frac{1}{m} \sum_{i=1}^m \mathbf{Y_i}\psi_k(\mathbf{Y}_i) \tag{1.68}$$

Clearly (1.65) and (1.66) are equivalent.

Consider the point $\bar{\mathbf{X}}_n = \mu_{\mathbf{X}}(\boldsymbol{\theta}_0)$, $\bar{\psi}_{m,k} = E_{\boldsymbol{\theta}_0}(\psi_{m,k})$, $k = 1, \ldots, s$, $\bar{\eta}_{m,k} = E_{\boldsymbol{\theta}_0}(\eta_{m,k})$, $k = 1, \ldots, s$, $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. Note that

$$\mathbf{G}(\mu_{\mathbf{X}}(\boldsymbol{\theta}_0), E_{\boldsymbol{\theta}_0}(\psi_{m,1}), \ldots, E_{\boldsymbol{\theta}_0}(\psi_{m,s}), E_{\boldsymbol{\theta}_0}(\eta_{m,1}), \ldots, E_{\boldsymbol{\theta}_0}(\eta_{m,s}), \boldsymbol{\theta}_0) = \mathbf{0} \tag{1.69}$$

since $E_{\boldsymbol{\theta}_0}(G(\bar{\mathbf{X}}_n, \bar{\psi}_{m,1}, \ldots, \bar{\psi}_{m,s}, \bar{\eta}_{m,1}, \ldots, \bar{\eta}_{m,s}, \boldsymbol{\theta}_0)) = \mathbf{0}$. Moreover, it can be shown using (1.40) that

$$\frac{\partial}{\partial \boldsymbol{\theta}} \mathbf{G}(\mu_{\mathbf{X}}(\boldsymbol{\theta}_0), E_{\boldsymbol{\theta}_0}(\psi_{m,1}), \ldots, E_{\boldsymbol{\theta}_0}(\psi_{m,s}), E_{\boldsymbol{\theta}_0}(\eta_{m,1}), \ldots, E_{\boldsymbol{\theta}_0}(\eta_{m,s}), \boldsymbol{\theta}_0) = -\frac{\partial}{\partial \boldsymbol{\theta}} \mu_{\mathbf{X}}(\boldsymbol{\theta}_0). \tag{1.70}$$

Also, h is continuously differentiable and $\frac{\partial}{\partial \boldsymbol{\theta}}\mu_{\mathbf{X}}(\boldsymbol{\theta}_0)$ is nonsingular by assumption. Thus we can apply the implicit function theorem at the aforementioned point. There exists a statistic $\tilde{\boldsymbol{\theta}}_{m,n}$ that is a zero of (1.65) whenever

$$||\bar{\mathbf{X}}_n - \mu_{\mathbf{X}}(\boldsymbol{\theta}_0)|| < r, ||\bar{\eta}_{m,k} - E_{\boldsymbol{\theta}_0}(\eta_k)|| < r, ||\bar{\psi}_{m,k} - E_{\boldsymbol{\theta}_0}(\psi_k)|| < r, ||\boldsymbol{\theta} - \boldsymbol{\theta}_0|| < r$$

Since, under regularity conditions, $\bar{\mathbf{X}}_n \xrightarrow{p} \mu_{\mathbf{X}}(\boldsymbol{\theta}_0)$, $\bar{\eta}_{m,k} \xrightarrow{p} E_{\boldsymbol{\theta}_0}(\eta_k)$, and $\bar{\psi}_{m,n} \xrightarrow{p} E_{\boldsymbol{\theta}_0}(\psi_k)$, this implies that the $P_{\boldsymbol{\theta}_0}$ probability tends to one that a statistic $\tilde{\boldsymbol{\theta}}_{m,n}$ which is a solution to the EMME system of equation exists and satisfies

$$||\tilde{\boldsymbol{\theta}}_{m,n} - \boldsymbol{\theta}_0|| < \epsilon.$$

□

Remark: For a general h, the result should also follow from the implicit function theorem, but there are many technical difficulties to prove it. However, we will now prove the result through a different approach.

Theorem 7. *Let $\boldsymbol{\theta}_0$ be the true parameter value. Suppose that*

(i) $h(\mathbf{x}, \boldsymbol{\theta})$ is continuously differentiable in a neighborhood of $\boldsymbol{\theta}_0$.

(ii) $E_{\boldsymbol{\theta}_0}(\mathbf{Y}h(\mathbf{Y}, \boldsymbol{\theta}_0)) < \infty$.

(iii) $E_{\boldsymbol{\theta}_0}\left[\frac{\partial}{\partial \boldsymbol{\theta}} h(\mathbf{Y}, \boldsymbol{\theta}_0)\right] < \infty$, $E_{\boldsymbol{\theta}_0}\left[\mathbf{Y}\frac{\partial}{\partial \boldsymbol{\theta}} h(\mathbf{Y}, \boldsymbol{\theta}_0)\right] < \infty$.

(iv) $\boldsymbol{\mu}'(\boldsymbol{\theta}_0)$ is nonsingular.

Then with $P_{\boldsymbol{\theta}_0}$ probability tending to one there exists a solution $\tilde{\boldsymbol{\theta}}_{m,n}$ to the EMME system of equations that is consistent for $\boldsymbol{\theta}_0$ as $m, n \to \infty$

Proof. Recall that the EMLE system of equations is

$$\bar{\mathbf{X}}_n = \frac{\sum_{i=1}^m \mathbf{Y}_i h(\mathbf{Y}_i, \boldsymbol{\theta})}{\sum_{i=1}^m h(\mathbf{Y}_i, \boldsymbol{\theta})}. \tag{1.71}$$

Let $\mathbf{G}_m(\mathbf{Y}_1,\ldots,\mathbf{Y}_m,\boldsymbol{\theta}) = \sum_{i=1}^m \mathbf{Y}_i h(\mathbf{Y}_i,\boldsymbol{\theta})/\sum_{i=1}^m h(\mathbf{Y}_i,\boldsymbol{\theta})$

so that

$$\frac{\partial}{\partial\boldsymbol{\theta}}\mathbf{G}_m(\mathbf{Y}_1,\ldots,\mathbf{Y}_m,\boldsymbol{\theta}) = \frac{\sum_{i=1}^m \mathbf{Y}_i \frac{\partial}{\partial\boldsymbol{\theta}} h(\mathbf{Y}_i,\boldsymbol{\theta})}{\sum_{i=1}^m h(\mathbf{Y}_i,\boldsymbol{\theta})} - \frac{\sum_{i=1}^m \mathbf{Y}_i h(\mathbf{Y}_i,\boldsymbol{\theta}) \sum_{i=1}^m \frac{\partial}{\partial\boldsymbol{\theta}} h(\mathbf{Y}_i,\boldsymbol{\theta})}{(\sum_{i=1}^m h(\mathbf{Y}_i,\boldsymbol{\theta}))^2}.$$

Note that the for any fixed $\mathbf{X}_1,\ldots,\mathbf{X}_n,\mathbf{Y}_1,\ldots,\mathbf{Y}_m$, the statistic $\tilde{\boldsymbol{\theta}}_{m,n}$, if it exists, is given by:

$$\tilde{\boldsymbol{\theta}}_{m,n} = \mathbf{G}_m^{-1}(\mathbf{Y}_1,\ldots,\mathbf{Y}_m,\bar{\mathbf{X}}_n).$$

The LLN implies, by assumption (ii), that

$$\bar{\mathbf{X}}_n \xrightarrow{p} \boldsymbol{\mu}(\boldsymbol{\theta}_0) \tag{1.72}$$

as $n \to \infty$ and assumptions (ii) and (1.31) imply

$$\mathbf{G}_m(\mathbf{Y}_1,\ldots,\mathbf{Y}_m,\boldsymbol{\theta}_0) \xrightarrow{p} \boldsymbol{\mu}(\boldsymbol{\theta}_0) \tag{1.73}$$

as $m \to \infty$.

Moreover, (1.40) and assumptions (ii)-(iii) imply, after some computation, that

$$\frac{\partial}{\partial\boldsymbol{\theta}}\mathbf{G}_m(\mathbf{Y}_1,\ldots,\mathbf{Y}_m,\boldsymbol{\theta}_0) \xrightarrow{p} \boldsymbol{\mu}'(\boldsymbol{\theta}_0) \tag{1.74}$$

as $m \to \infty$.

Since $\boldsymbol{\mu}'(\boldsymbol{\theta}_0)$ is nonsingular, with probability tending to one, $\frac{\partial}{\partial\boldsymbol{\theta}}\mathbf{G}_m(\mathbf{Y}_1,\ldots,\mathbf{Y}_m,\boldsymbol{\theta}_0)$ is nonsingular. Moreover, with probability tending to one, for any $\delta > 0$

$$||\mathbf{X}_n - \mathbf{G}_m(\mathbf{Y}_1,\ldots,\mathbf{Y}_m,\boldsymbol{\theta}_0)|| < \delta.$$

Note that $\mathbf{G}_m(\mathbf{Y}_1,\ldots,\mathbf{Y}_m,\boldsymbol{\theta})$ is continuously differentiable in a neighborhood of $\boldsymbol{\theta}_0$ for all m and n by assumption (i). Choose δ such that, by virtue of the inverse function theorem, $\mathbf{G}_m(\mathbf{Y}_1, \dots, \mathbf{Y}_m, \boldsymbol{\theta})$ is one to one and onto on a neighborhood of $\boldsymbol{\theta}_0$ and $\mathbf{G}_m^{-1}(\mathbf{Y}_1, \dots, \mathbf{Y}_m, \mathbf{a})$ exists and is continuous in its last argument in a neighborhood of $\mathbf{G}_m(\boldsymbol{\theta}_0)$ including $\bar{\mathbf{X}}_n$. Moreover let δ be such that for any $\epsilon > 0$, by continuity of $\mathbf{G}_m^{-1}(\mathbf{Y}_1, \dots, \mathbf{Y}_m, \mathbf{a})$,

$$||\mathbf{G}_m^{-1}(\mathbf{Y}_1, \dots, \mathbf{Y}_m, \bar{\mathbf{X}}_n) - \mathbf{G}_m^{-1}(\mathbf{Y}_1, \dots, \mathbf{Y}_m, \mathbf{G}_m(\mathbf{Y}_1, \dots, \mathbf{Y}_m, \boldsymbol{\theta}_0))|| < \epsilon.$$

i.e.,

$$||\tilde{\boldsymbol{\theta}}_{m,n} - \boldsymbol{\theta}_0|| < \epsilon.$$

□

We will now explore the asymptotic distribution of $\tilde{\boldsymbol{\theta}}_{m,n}$.

Lemma 5. *Let*

$$\mathbf{G}_{m,n}(\bar{\mathbf{X}}_n, \mathbf{Y}_1, \dots, \mathbf{Y}_n, \boldsymbol{\theta}) = \bar{\mathbf{X}}_n \frac{\sum_{i=1}^m h(\mathbf{Y}_i, \boldsymbol{\theta})}{m} - \frac{\sum_{i=1}^m \mathbf{Y}_i h(\mathbf{Y}_i, \boldsymbol{\theta})}{m}.$$

Suppose $m, n \to \infty$. Under regularity conditions,

(i) If $m = cn(1 + o(1)), \quad c > 0$,

$$\begin{aligned}&\sqrt{n}\mathbf{G}_{m,n}(\bar{\mathbf{X}}_n, \mathbf{Y}_1, \dots, \mathbf{Y}_n, \boldsymbol{\theta}) \\ &\xrightarrow{d} N_p\left[\mathbf{0}, \frac{\boldsymbol{\Sigma}_{\mathbf{X}}(\boldsymbol{\theta})}{C(\boldsymbol{\theta})^2} + \frac{1}{c}\int h(\mathbf{x}, \boldsymbol{\theta})^2(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{X}}(\boldsymbol{\theta}))(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{X}}(\boldsymbol{\theta}))^T dP(\mathbf{x})\right].\end{aligned}$$

(ii) If $m = o(n)$,

$$\sqrt{m}\mathbf{G}_{m,n}(\bar{\mathbf{X}}_n, \mathbf{Y}_1, \dots, \mathbf{Y}_n, \boldsymbol{\theta}) \xrightarrow{d} N_p\left[\mathbf{0}, \int h(\mathbf{x}, \boldsymbol{\theta})^2(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{X}}(\boldsymbol{\theta}))(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{X}}(\boldsymbol{\theta}))^T dP(\mathbf{x}))\right].$$

(iii) If $n = o(m)$,

$$\sqrt{n}\mathbf{G}_{m,n}(\bar{\mathbf{X}}_n, \mathbf{Y}_1, \dots, \mathbf{Y}_n, \boldsymbol{\theta}) \xrightarrow{d} N_p\left[\mathbf{0}, \frac{\boldsymbol{\Sigma}_{\mathbf{X}}(\boldsymbol{\theta})}{C(\boldsymbol{\theta})^2}\right].$$

Proof.

$$\begin{aligned}
&\mathbf{G}_{m,n}(\bar{\mathbf{X}}_n, \mathbf{Y}_1, \ldots, \mathbf{Y}_n, \boldsymbol{\theta}) \\
&= \bar{\mathbf{X}}_n \frac{\sum_{i=1}^m h(\mathbf{Y}_i, \boldsymbol{\theta})}{m} - \frac{\sum_{i=1}^m \mathbf{Y}_i h(\mathbf{Y}_i, \boldsymbol{\theta})}{m} \qquad (1.75) \\
&= (\bar{\mathbf{X}}_n - \boldsymbol{\mu}_{\mathbf{X}}(\boldsymbol{\theta})) \frac{\sum_{i=1}^m h(\mathbf{Y}_i, \boldsymbol{\theta})}{m} - \sum_{i=1}^m \frac{(\mathbf{Y}_i h(\mathbf{Y}_i, \boldsymbol{\theta}) - \boldsymbol{\mu}_{\mathbf{X}}(\boldsymbol{\theta}) h(\mathbf{Y}_i, \boldsymbol{\theta}))}{m}
\end{aligned}$$

For part (i) we multiply both sides of (1.75) by $\sqrt{n}$ and replace n with m/c in the second term on the right hand side to obtain

$$\begin{aligned}
&\sqrt{n}\mathbf{G}_{m,n}(\bar{\mathbf{X}}_n, \mathbf{Y}_1, \ldots, \mathbf{Y}_n, \boldsymbol{\theta}) \\
&= \sqrt{n}(\bar{\mathbf{X}}_n - \boldsymbol{\mu}_{\mathbf{X}}(\boldsymbol{\theta})) \frac{\sum_{i=1}^m h(\mathbf{Y}_i, \boldsymbol{\theta})}{m} - \frac{1}{\sqrt{c}} \sum_{i=1}^m \frac{(\mathbf{Y}_i h(\mathbf{Y}_i, \boldsymbol{\theta}) - \boldsymbol{\mu}_{\mathbf{X}}(\boldsymbol{\theta}) h(\mathbf{Y}_i, \boldsymbol{\theta}))}{\sqrt{m}}.
\end{aligned} \qquad (1.76)$$

Clearly,

$$\sqrt{n}(\bar{\mathbf{X}}_n - \boldsymbol{\mu}_{\mathbf{X}}(\boldsymbol{\theta})) \frac{\sum_{i=1}^m h(\mathbf{Y}_i, \boldsymbol{\theta})}{m} \xrightarrow{d} \mathcal{N}_p \left[\mathbf{0}, \frac{1}{C(\boldsymbol{\theta})^2} \boldsymbol{\Sigma}_{\mathbf{X}}(\boldsymbol{\theta}) \right] \qquad (1.77)$$

by the CLT and LLN and Slutsky's Theorem.

It's easy to check that the second term in (1.76) has mean $\mathbf{0}$ since

$$E(\mathbf{Y} h(\mathbf{Y}, \boldsymbol{\theta})) = C(\boldsymbol{\theta})^{-1} E_{\boldsymbol{\theta}}(\mathbf{X}) = E h(\mathbf{Y}, \boldsymbol{\theta}) \boldsymbol{\mu}_{\mathbf{X}}(\boldsymbol{\theta}).$$

The CLT implies that this term has an asymptotically normal distribution with mean $\mathbf{0}$ and with an asymptotic covariance matrix given by:

$$\begin{aligned}
&E[\mathbf{Y} h(\mathbf{Y}, \boldsymbol{\theta}) - \boldsymbol{\mu}_{\mathbf{X}}(\boldsymbol{\theta}) h(\mathbf{Y}, \boldsymbol{\theta})][\mathbf{Y} h(\mathbf{Y}, \boldsymbol{\theta}) - \boldsymbol{\mu}_{\mathbf{X}}(\boldsymbol{\theta}) h(\mathbf{Y}, \boldsymbol{\theta})]^T \\
&= E[h(\mathbf{Y}, \boldsymbol{\theta}_0)^2 (\mathbf{Y} - \boldsymbol{\mu}_{\mathbf{X}}(\boldsymbol{\theta}))(\mathbf{Y} - \boldsymbol{\mu}_{\mathbf{X}}(\boldsymbol{\theta}))^T] \\
&= \int h(\mathbf{x}, \boldsymbol{\theta})^2 (\mathbf{x} - \boldsymbol{\mu}_{\mathbf{X}}(\boldsymbol{\theta}))(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{X}}(\boldsymbol{\theta}))^T dP(\mathbf{x})
\end{aligned}$$

Thus,

$$\begin{aligned}&\sum_{i=1}^{m}\frac{(\mathbf{Y}_i h(\mathbf{Y}_i,\boldsymbol{\theta})-\boldsymbol{\mu}_{\mathbf{X}}(\boldsymbol{\theta})h(\mathbf{Y}_i,\boldsymbol{\theta}))}{\sqrt{m}}\\ &\xrightarrow{d}\mathcal{N}_p(\mathbf{0},\int h(\mathbf{x},\boldsymbol{\theta})^2(\mathbf{x}-\boldsymbol{\mu}_{\mathbf{X}}(\boldsymbol{\theta}))(\mathbf{x}-\boldsymbol{\mu}_{\mathbf{X}}(\boldsymbol{\theta}))^T dP(\mathbf{x})).\end{aligned} \tag{1.78}$$

Part (i) follows from (1.77) and (1.78).

For part (ii) we write

$$\begin{aligned}&\sqrt{m}\mathbf{G}_{m,n}(\bar{\mathbf{X}}_n,\mathbf{Y}_1,\ldots,\mathbf{Y}_n,\boldsymbol{\theta})\\ &=\frac{\sqrt{m}}{\sqrt{n}}\sqrt{n}(\bar{\mathbf{X}}_n-\boldsymbol{\mu}_{\mathbf{X}}(\boldsymbol{\theta}))\frac{\sum_{i=1}^{m}h(\mathbf{Y}_i,\boldsymbol{\theta})}{m}-\sum_{i=1}^{m}\frac{(\mathbf{Y}_i h(\mathbf{Y}_i,\boldsymbol{\theta})-\boldsymbol{\mu}_{\mathbf{X}}(\boldsymbol{\theta})h(\mathbf{Y}_i,\boldsymbol{\theta}))}{\sqrt{m}}\end{aligned}$$

and notice that the first term in the right hand side converges in probability to $\mathbf{0}$ so that the result follows by (1.78).

For part (iii) we write

$$\begin{aligned}&\sqrt{n}\mathbf{G}_{m,n}(\bar{\mathbf{X}}_n,\mathbf{Y}_1,\ldots,\mathbf{Y}_n,\boldsymbol{\theta})\\ &=\sqrt{n}(\bar{\mathbf{X}}_n-\boldsymbol{\mu}_{\mathbf{X}}(\boldsymbol{\theta}))\frac{\sum_{i=1}^{m}h(\mathbf{Y}_i,\boldsymbol{\theta})}{m}-\frac{\sqrt{n}}{\sqrt{m}}\sum_{i=1}^{m}\frac{(\mathbf{Y}_i h(\mathbf{Y}_i,\boldsymbol{\theta})-\boldsymbol{\mu}_{\mathbf{X}}(\boldsymbol{\theta})h(\mathbf{Y}_i,\boldsymbol{\theta}))}{\sqrt{m}}\end{aligned}$$

and notice that the second term in the right hand side converges in probability to $\mathbf{0}$, so that the result follows by (1.77). □

Theorem 8. *Let $m,n\to\infty$. Under regularity conditions,*

(i) If $m=cn(1+o(1)),\quad c>0,$

$$\sqrt{n}(\tilde{\boldsymbol{\theta}}_{m,n}-\boldsymbol{\theta})\xrightarrow{d}N_p\left[\mathbf{0},\frac{\partial}{\partial\boldsymbol{\theta}}\boldsymbol{\mu}_{\mathbf{X}}(\boldsymbol{\theta})^{-1}\boldsymbol{\Sigma}_{\mathbf{X}}(\boldsymbol{\theta})(\frac{\partial}{\partial\boldsymbol{\theta}}\boldsymbol{\mu}_{\mathbf{X}}(\boldsymbol{\theta})^{-1})^T+c^{-1}\mathbf{A}\right]$$

(ii) If $m=o(n)$,

$$\sqrt{m}(\tilde{\boldsymbol{\theta}}_{m,n}-\boldsymbol{\theta})\xrightarrow{d}N_p(\mathbf{0},\mathbf{A})$$

(iii) If $n = o(m)$,

$$\sqrt{n}(\tilde{\boldsymbol{\theta}}_{m,n} - \boldsymbol{\theta}) \xrightarrow{d} N_p(\mathbf{0}, \frac{\partial}{\partial \boldsymbol{\theta}} \boldsymbol{\mu}_{\mathbf{X}}(\boldsymbol{\theta})^{-1} (\frac{\partial}{\partial \boldsymbol{\theta}} \boldsymbol{\mu}_{\mathbf{X}}(\boldsymbol{\theta})^{-1})^T)$$

where

$$\mathbf{A} = C(\boldsymbol{\theta})^2 \left((\frac{\partial}{\partial \boldsymbol{\theta}} \boldsymbol{\mu}_{\mathbf{X}}(\boldsymbol{\theta}))^{-1} \right) \left[\int h(\mathbf{x}, \boldsymbol{\theta})^2 (\mathbf{x} - \boldsymbol{\mu}_{\mathbf{X}}(\boldsymbol{\theta}))(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{X}}(\boldsymbol{\theta}))^T dP(\mathbf{x}) \right] \left((\frac{\partial}{\partial \boldsymbol{\theta}} \boldsymbol{\mu}_{\mathbf{X}}(\boldsymbol{\theta}))^{-1} \right)^T$$

Proof. Let $\boldsymbol{\theta}_0$ be the true parameter value. Again, we expand $\mathbf{G}_{m,n}(\boldsymbol{\theta})$ about $\boldsymbol{\theta}_0$ and plug in $\tilde{\boldsymbol{\theta}}_{m,n}$ to obtain equations

$$G_{m,n,j}(\tilde{\boldsymbol{\theta}}_{m,n}) = 0 = G_{m,n,j}(\boldsymbol{\theta}_0) + \sum_{k=1}^{p} (\tilde{\theta}_k - \theta_{0,k}) \frac{\partial}{\partial \theta_k} G_{m,n,j}(\boldsymbol{\theta})$$
$$+ \frac{1}{2} \sum_{k=1}^{p} \sum_{l=1}^{p} (\tilde{\theta}_k - \theta_{0,k})(\tilde{\theta}_l - \theta_{0,l}) \frac{\partial^2}{\partial \theta_k \partial \theta_l} G_{m,n,j}(\boldsymbol{\theta}_*),$$
$$j = 1, \ldots, p.$$

This is equivalent to

$$\sum_{k=1}^{p} (\tilde{\theta}_k - \theta_k) [\frac{\partial}{\partial \theta_k} G_{m,n,j}(\boldsymbol{\theta}_0) + \frac{1}{2} \sum_{l=1}^{p} (\hat{\theta}_l - \theta_{0,l}) \frac{\partial^2}{\partial \theta_k \partial \theta_l} G_{m,n,j}(\boldsymbol{\theta}_*)] = -G_{m,n,j}(\boldsymbol{\theta}_0)$$

Again, we use Lemma 2. We must consider the behavior of

$$A_{j,k,m,n} = \frac{\partial}{\partial \theta_k} G_{m,n,j}(\boldsymbol{\theta}_0) + \frac{1}{2} \sum_{l=1}^{p} (\tilde{\theta}_l - \theta_{0,l}) \frac{\partial^2}{\partial \theta_k \partial \theta_l} G_{m,n,j}(\boldsymbol{\theta}_*)$$

As previously shown,

$$\frac{\partial}{\partial \boldsymbol{\theta}} \mathbf{G}_{m,n}(\boldsymbol{\theta}_0) \xrightarrow{p} -C(\boldsymbol{\theta}_0)^{-1} \frac{\partial}{\partial \boldsymbol{\theta}} \mu_{\mathbf{X}}(\boldsymbol{\theta}_0)$$

We now show that the rest converges in probability to $\mathbf{0}$.

The function $\frac{\partial^2}{\partial\theta_k\partial\theta_l}G_{m,n,j}(\boldsymbol{\theta})$ can be expressed as

$$\begin{aligned}&\frac{\partial^2}{\partial\theta_k\partial\theta_l}\mathbf{G}_{m,n}(\bar{\mathbf{X}}_n, \mathbf{Y}_1, \ldots, \mathbf{Y}_n, \boldsymbol{\theta})_j \\ &= \frac{\bar{X}_{n,j}\sum_{i=1}^m \frac{\partial^2}{\partial\theta_k\partial\theta_l}h(\mathbf{Y}_i, \boldsymbol{\theta}) - \sum_{i=1}^m Y_{ij}\frac{\partial^2}{\partial\theta_k\partial\theta_l}h(\mathbf{Y}_i, \boldsymbol{\theta})}{m}.\end{aligned} \tag{1.79}$$

Note that under the assumptions that $|\frac{\partial^2}{\partial\theta_k\partial\theta_l}h(\mathbf{y}, \boldsymbol{\theta})| < M_{k,l}(\mathbf{y})$ and $|y_j\frac{\partial^2}{\partial\theta_k\partial\theta_l}h(\mathbf{y}, \boldsymbol{\theta})| < N_{j,k,l}(\mathbf{y})$ for all $\boldsymbol{\theta}$ in a neighborhood of $\boldsymbol{\theta}_0$, with $EM_{k,l}(\mathbf{Y}) < \infty$ and $E(N_{j,k,l}(\mathbf{Y})) < \infty$, for $j, k, l = 1, \ldots, p$, (1.79) is bounded in probability by the Law of Large Numbers. By consistency of $\tilde{\boldsymbol{\theta}}_{m,n}$ the second term of $A_{j,k,n}$ converges in probability to 0. □

1.3.3 Relationship Between the Asymptotic Distribution of the EMLE and EMME

There is an interesting relationship between the asymptotic covariance of the EMME and the EMLE. Let's consider the case where $m = cn(1 + o(1))$, $c > 0$. By Theorem 4, we have that under regularity conditions the EMLE has the following asymptotic distribution:

$$\begin{aligned}&\sqrt{n}(\hat{\boldsymbol{\theta}}_{m,n} - \boldsymbol{\theta}) \\ &\xrightarrow{d} N_p\left[\mathbf{0}, \mathbf{I}_X(\boldsymbol{\theta})^{-1} + C(\boldsymbol{\theta})^2\frac{1}{c}\mathbf{I}_X(\boldsymbol{\theta})^{-1}\left\{\int h(x, \boldsymbol{\theta})^2\mathbf{J}(x; \boldsymbol{\theta})\mathbf{J}(x; \boldsymbol{\theta})^T dP(x)\right\}\mathbf{I}_X(\boldsymbol{\theta})^{-1}\right].\end{aligned} \tag{1.80}$$

By Theorem 8, under regularity conditions the EMME satisfies the following relation:

$$\sqrt{n}(\tilde{\boldsymbol{\theta}}_{m,n} - \boldsymbol{\theta}) \xrightarrow{d} N_p(\mathbf{0}, (\frac{\partial}{\partial\boldsymbol{\theta}}\boldsymbol{\mu}_{\mathbf{X}}(\boldsymbol{\theta}))^{-1}(\boldsymbol{\Sigma}_{\mathbf{X}}(\boldsymbol{\theta})(\frac{\partial}{\partial\boldsymbol{\theta}}\boldsymbol{\mu}_{\mathbf{X}}(\boldsymbol{\theta}))^{-1})^T + \frac{1}{c}\mathbf{A})$$

where

$$\mathbf{A} = C(\boldsymbol{\theta})^2(\frac{\partial}{\partial \boldsymbol{\theta}}\boldsymbol{\mu}_{\mathbf{X}}(\boldsymbol{\theta}))^{-1}\int h(\mathbf{x},\boldsymbol{\theta})^2(\mathbf{x}-\boldsymbol{\mu}_{\mathbf{X}}(\boldsymbol{\theta}))(\mathbf{x}-\boldsymbol{\mu}_{\mathbf{X}}(\boldsymbol{\theta}))^T dP(\mathbf{x})((\frac{\partial}{\partial \boldsymbol{\theta}}\boldsymbol{\mu}_{\mathbf{X}}(\boldsymbol{\theta}))^{-1})^T. \tag{1.81}$$

Let $\hat{J}(\mathbf{X},\boldsymbol{\theta})$ be the linear score and $\hat{\mathbf{I}}_{\mathbf{X}}(\boldsymbol{\theta})$ be the linear version of the Fisher information matrix. For a random element X, $\hat{J}_X$ is the projection of the score J_X onto the linear space spanned by X. The linear version of the Fisher information is defined analogously to the Fisher information with the score J_X replaced by $\hat{J}_X$.

The asymptotic distribution of the EMME can be expressed as

$$\begin{aligned}&\sqrt{n}(\tilde{\boldsymbol{\theta}}_{m,n} - \boldsymbol{\theta})\\ &\xrightarrow{d} N_p\left[\mathbf{0}, \hat{\mathbf{I}}_{\mathbf{X}}(\boldsymbol{\theta})^{-1} + C(\boldsymbol{\theta})^2\frac{1}{c}\hat{\mathbf{I}}_{\mathbf{X}}(\theta)^{-1}\left\{\int h(\mathbf{x},\boldsymbol{\theta})^2\hat{\mathbf{J}}(\mathbf{x};\boldsymbol{\theta})\hat{\mathbf{J}}(\mathbf{x};\boldsymbol{\theta})^T dP(\mathbf{x})\right\}\hat{\mathbf{I}}_{\mathbf{X}}(\boldsymbol{\theta})^{-1}\right].\end{aligned} \tag{1.82}$$

The asymptotic distribution of the EMME (1.82) is the same as that of the EMLE (1.80) with $\mathbf{I}_{\mathbf{X}}$ replaced with $\hat{\mathbf{I}}_{\mathbf{X}}$ and $\mathbf{J}$ replaced with $\hat{\mathbf{J}}$.

The second terms in the asymptotic covariance matrices of (1.80) and (1.81) are positive semidefinite, and represent the price paid for having to estimate the generator. It is unclear from the covariance expressions whether the EMME or EMLE is superior.

In both (1.80) and (1.82), the second term vanishes as $c \to \infty$, which corresponds $n = o(m)$. In this case, the asymptotic distributions are equivalent to those of the estimators that would have been obtained by applying maximum likelihood and the method of moments on the sample $(X_1, \ldots, X_n)$ if the generator were known. When $n = o(m)$, the EMLE is efficient.

When $m = o(n)$, both the EMLE and EMME have a rate of convergence of $1/\sqrt{m}$ instead of $1/\sqrt{n}$.

1.4 The Case of m-Sample Density Ratio Models

Several papers deal with semiparametric estimation for density ratio models when there are several samples, where the model is of the form

$$f_i = g(x, \theta_i) f_m, \quad i = 1, \ldots, m-1 \tag{1.83}$$

where the f_i are unknown densities with available associated samples and the θ_i are vectors of parameters of finite dimension (See, for instance, Fokianos 2004 [19], Kedem *et al.* (2009) [34], among others). The distribution f_m is usually called the reference distribution. Fokianos (2004) extends the asymptotic normality results of Qin (1998) to the model (1.83). Denoting the size of sample i as n_i, for $i = 1, \ldots, m$, and $n = \sum_{i=1}^{m} n_i$, he proves that if $n_i/n \to \rho_i$, under some conditions, the estimators of $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_{m-1})^T$ and $\boldsymbol{\lambda}$=$(\lambda_1, \ldots, \lambda_{m-1})^T$ are asymptotically normal with a normalizing factor of $\sqrt{n}$. As in Qin's case, for the asymptotic variance to be well defined it is necessary for $\rho_i > 0$.

As pointed out by Fokianos, the sample corresponding to f_m is arbitrarily chosen. A property that characterizes the multinomial logit models, usually described as independence of irrelevant alternatives, is that the choice of f_m does not affect inferential results since the difference of the slopes remains constant. For other models, the choice of reference measure affects inference, an undesirable property. In the simple calculations that follow, we will show that, using empirical likelihood, it is not necessary to choose a sample which corresponds to the reference measure.

In fact if we have m samples, we can formulate the model as

$$f_i = g_i(x, \theta_i) f_0, \quad i = 1, \ldots, m \tag{1.84}$$

where there is no need to have a sample from f_0. This removes the subjective choice of which sample to ascribe as the reference measure. We will illustrate this in the case where $m = 2$, for simplicity, and show that the resulting method also results in estimators that are asymptotically normal in the case where

$$n_1/n \to \rho_1, n_2/n \to \rho_2, n_1 + n_2 = n. \tag{1.85}$$

Suppose $X_1, \ldots, X_{n_1} \sim f_1(x, \theta_1) = g_1(x, \theta_1) f(x)$ and $X_{n_1+1}, \ldots, X_n \sim f_2(x, \theta_2) = g_2(x, \theta_2) f(x)$.

For simplicity we assume all parameters are scalars. Following Qin's approach we can let $p_i = dF(x_i)$ and express the empirical log likelihood as

$$l(\mathbf{X}, \boldsymbol{\theta}) = \sum_{i=1}^{n} \log p_i + \sum_{i=1}^{n_1} \log g_1(x_i, \theta_1) + \sum_{i=n_1+1}^{n} \log g_2(x_i, \theta_2) \tag{1.86}$$

Subject to the constraints

$$\sum_{i=1}^{n} p_i = 1, \quad \sum_{i=1}^{n} (g_1(x_i, \theta_1) - 1) p_i = 0, \quad \sum_{i=1}^{n} (g_2(x_i, \theta_2) - 1) p_i = 0, \quad 0 \le p_i \le 1. \tag{1.87}$$

By Lagrange multipliers, it is straightforward to show that for given θ_1, θ_2

$$p_i = \frac{1}{n} \frac{1}{1 + \lambda_1 (g_1(x_i, \theta_1) - 1) + \lambda_2 (g_2(x_i, \theta_2) - 1)} \tag{1.88}$$

where λ_1, λ_2 are given by the equations

$$-\sum_{i=1}^{n} \frac{(g_1(x_i, \theta_1) - 1)}{1 + \lambda_1 (g_1(x_i, \theta_1) - 1) + \lambda_2 (g_2(x_i, \theta_2) - 1)} = 0, \tag{1.89}$$

$$-\sum_{i=1}^{n} \frac{(g_2(x_i,\theta_2)-1)}{1+\lambda_1(g_1(x_i,\theta_1)-1)+\lambda_2(g_2(x_i,\theta_2)-1)} = 0. \tag{1.90}$$

Thus the log likelihood can be expressed as a function of $\boldsymbol{\theta} = (\theta_1, \theta_2)$ and $\boldsymbol{\lambda} = (\lambda_1, \lambda_2)$:

$$\begin{aligned} l(\boldsymbol{\theta}, \boldsymbol{\lambda}) = -\sum_{i=1}^{n} \log[1 + \lambda_1(g_1(x_i,\theta_1)-1) + \lambda_2(g_2(x_i,\theta_2)-1) \\ + \sum_{i=1}^{n_1} \log g_1(x_i,\theta_1) + \sum_{i=n_1+1}^{n} \log g_2(x_i,\theta_2). \end{aligned} \tag{1.91}$$

Let $\boldsymbol{\psi}_0^T = (\boldsymbol{\rho}^T, \boldsymbol{\theta}_0)$. Under regularity conditions the vector $\hat{\boldsymbol{\psi}}^T = (\hat{\boldsymbol{\lambda}}^T, \hat{\boldsymbol{\theta}}^T)$ of maximizers of $l(\boldsymbol{\theta}, \boldsymbol{\lambda})$ is asymptotically normal with a normalizing constant $\sqrt{n}$, where $\hat{\lambda}_0 \xrightarrow{p} \rho_1$, $\hat{\lambda}_1 \xrightarrow{p} \rho_2$, and $\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_0$. The proof follows a standard Taylor expansion argument.

Theorem 9. *Under regularity conditions*

$$\sqrt{n}(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}) \xrightarrow{d} \mathcal{N}_4(\mathbf{0}, \mathbf{V})$$

where $\mathbf{V} = \mathbf{S}^{-1}\mathbf{W}\mathbf{S}^{-1}$, *and* $\mathbf{S}$ *and* $\mathbf{W}$ *will be stated subsequently.*

Proof. The system of estimating equations, comprising the partial derivatives of l is given by:

$$\mathbf{G}_{n_1,n_2} = \begin{bmatrix} -\frac{1}{n}\sum_{i=1}^{n} \frac{(g_1(x_i,\theta_1)-1)}{1+\lambda_1(g_1(x_i,\theta_1)-1)+\lambda_2(g_2(x_i,\theta_2)-1)} \\ -\frac{1}{n}\sum_{i=1}^{n} \frac{(g_2(x_i,\theta_2)-1)}{1+\lambda_1(g_1(x_i,\theta_1)-1)+\lambda_2(g_2(x_i,\theta_2)-1)} \\ -\frac{1}{n}\sum_{i=1}^{n} \lambda_1 \frac{\frac{\partial}{\partial\theta_1} g_1(x_i,\theta_1)}{1+\lambda_1(g_1(x_i,\theta_1)-1)+\lambda_2(g_2(x_i,\theta_2)-1)} + \frac{1}{n}\sum_{i=1}^{n_1} \frac{\partial}{\partial\theta_1} \log g_1(x_i,\theta_1) \\ -\frac{1}{n}\sum_{i=1}^{n} \lambda_2 \frac{\frac{\partial}{\partial\theta_2} g_2(x_i,\theta_2)}{1+\lambda_2(g_1(x_i,\theta_1)-1)+\lambda_2(g_2(x_i,\theta_2)-1)} + \frac{1}{n}\sum_{i=n_1+1}^{n} \frac{\partial}{\partial\theta_2} \log g_2(x_i,\theta_2) \end{bmatrix} \tag{1.92}$$

$$\mathbf{G}_{n_1,n_2}|_{\boldsymbol{\psi}=\boldsymbol{\psi}_0} = \begin{bmatrix} -\frac{1}{n}\sum_{i=1}^{n}\frac{(g_1(x_i,\theta_{0,1})-1)}{\rho_1(g_1(x_i,\theta_{0,1}))+\rho_2(g_2(x_i,\theta_{0,2}))} \\ -\frac{1}{n}\sum_{i=1}^{n}\frac{(g_2(x_i,\theta_{0,2})-1)}{\rho_1(g_1(x_i,\theta_{0,1}))+\rho_2(g_2(x_i,\theta_{0,2}))} \\ -\frac{1}{n}\sum_{i=1}^{n}\rho_1\frac{\frac{\partial}{\partial\theta_1}g_1(x_i,\theta_{0,1})}{\rho_1(g_1(x_i,\theta_{0,1}))+\rho_2(g_2(x_i,\theta_{0,2}))} + \frac{1}{n}\sum_{i=1}^{n_1}\frac{\partial}{\partial\theta_1}\log g_1(x_i,\theta_{0,1}) \\ -\frac{1}{n}\sum_{i=1}^{n}\rho_2\frac{\frac{\partial}{\partial\theta_2}g_2(x_i,\theta_{0,2})}{\rho_1(g_1(x_i,\theta_{0,1}))+\rho_2(g_2(x_i,\theta_{0,2}))} + \frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\theta_2}\log g_2(x_i,\theta_{0,2}) \end{bmatrix}. \tag{1.93}$$

Note that for any function $V(x)$, we have that

$$\frac{1}{n}\sum_{i=1}^{n}V(x_i) = \frac{n_1}{n}\frac{1}{n_1}\sum_{i=1}^{n_1}V(x_i) + \frac{n_2}{n}\frac{1}{n_2}\sum_{i=1}^{n_2}V(x_i) \xrightarrow{p} E_F[(\rho_1 g_1 + \rho_2 g_2)V]. \tag{1.94}$$

From this, and from the properties of the Fisher score, it follows, under regularity conditions

$$\mathbf{G}_{n_1,n_2}|_{\boldsymbol{\psi}=\boldsymbol{\psi}_0} \xrightarrow{p} \mathbf{0} \tag{1.95}$$

Consistency follows from the Lemma 3 under regularity conditions.

The proof of asymptotic normality follows a standard Taylor expansion argument, using Lemma 2. Through a Taylor expansion of component function $\mathbf{G}_{n_1,n_2,j}$ we have:

$$\begin{aligned}&\sqrt{n}G_{m,n,j}(\boldsymbol{\psi}) \\ &= -\left\{\sum_{k=1}^{4}\sqrt{n}(\hat{\psi}_{mk}-\psi_k)[\frac{\partial}{\partial\psi_k}G_{m,n,j}(\boldsymbol{\psi}) + \frac{1}{2}\sum_{l=1}^{4}(\hat{\psi}_{ml}-\psi_l)\frac{\partial^2}{\partial\psi_k\partial\psi_l}G_{m,n,j}(\boldsymbol{\psi}_*)]\right\}\end{aligned} \tag{1.96}$$

The CLT applies to $\sqrt{n}\mathbf{G}_{n_1,n_2}|_{\boldsymbol{\psi}=\boldsymbol{\psi}_0}$ by splitting each vector entry into sums of $i.i.d.$ random variables:

$$\frac{1}{\sqrt{n}}\frac{\partial}{\partial\boldsymbol{\psi}}l|_{\boldsymbol{\psi_0}} =$$

$$\begin{bmatrix} -\frac{\sqrt{n_1}}{\sqrt{n}}\frac{1}{\sqrt{n_1}}\sum_{i=1}^{n_1}\frac{(g_1-1)}{\rho_1 g_1+\rho_2 g_2} - \frac{\sqrt{n_2}}{\sqrt{n}}\frac{1}{\sqrt{n_2}}\sum_{i=n_1+1}^{n}\frac{(g_1-1)}{\rho_1 g_1+\rho_2 g_2} \\ -\frac{\sqrt{n_1}}{\sqrt{n}}\frac{1}{\sqrt{n_1}}\sum_{i=1}^{n_1}\frac{(g_2-1)}{\rho_1 g_1+\rho_2 g_2} - \frac{\sqrt{n_2}}{\sqrt{n}}\frac{1}{\sqrt{n_2}}\sum_{i=n_1+1}^{n}\frac{(g_2-1)}{\rho_1 g_1+\rho_2 g_2} \\ -\frac{\sqrt{n_1}}{\sqrt{n}}\frac{1}{\sqrt{n_1}}\sum_{i=1}^{n_1}\rho_1\frac{\frac{\partial}{\partial\theta_1}g_1}{\rho_1 g_1+\rho_2 g_2} - \frac{\sqrt{n_2}}{\sqrt{n}}\frac{1}{\sqrt{n_2}}\sum_{i=n_1+1}^{n}\rho_1\frac{\frac{\partial}{\partial\theta_1}g_1}{\rho_1 g_1+\rho_2 g_2} + \frac{\sqrt{n_1}}{\sqrt{n}}\frac{1}{\sqrt{n_1}}\sum_{i=1}^{n_1}\frac{\partial}{\partial\theta_1}\log g_1 \\ -\frac{\sqrt{n_1}}{\sqrt{n}}\frac{1}{\sqrt{n_1}}\sum_{i=1}^{n_1}\rho_2\frac{\frac{\partial}{\partial\theta_2}g_2}{\rho_1 g_1+\rho_2 g_2} - \frac{\sqrt{n_2}}{\sqrt{n}}\frac{1}{\sqrt{n_2}}\sum_{i=n_1+1}^{n}\rho_2\frac{\frac{\partial}{\partial\theta_2}g_2}{\rho_1 g_1+\rho_2 g_2} + \frac{\sqrt{n_2}}{\sqrt{n}}\frac{1}{\sqrt{n_2}}\sum_{i=n_1+1}^{n}\frac{\partial}{\partial\theta_2}\log g_2 \\ \\ \end{bmatrix}. \tag{1.97}$$

The vector is asymptotically normal, with a 4×4 asymptotic covariance matrix $\boldsymbol{\Sigma} = \mathbf{W}(\boldsymbol{\theta}_0)$, which will be given below. Also, $\frac{\partial}{\partial\boldsymbol{\psi}}\mathbf{G}_{n_1,n_2}|_{\boldsymbol{\psi}=\boldsymbol{\psi}} \xrightarrow{p} \mathbf{S}(\boldsymbol{\theta}_0)$, where $S(\boldsymbol{\theta})$ will be given below. Thus, the result follows from Lemma 2 in Chapter 1 under the regularity conditions that imply that all the second partial derivatives of $\mathbf{G}_{n_1,n_2}(\boldsymbol{\psi})$, $\frac{\partial^2}{\partial\psi^2}\mathbf{G}_{n_1,n_2}(\boldsymbol{\psi})$ are bounded in probability in a neighborhood of $\boldsymbol{\psi}_0$. It is easy to see that regularity conditions can be imposed to ensure this.

In what follows, we will provide expressions for $\mathbf{S}(\boldsymbol{\theta}_0)$ and $\mathbf{W}(\boldsymbol{\theta}_0)$. This will show that, under regularity conditions, $\mathbf{S}(\boldsymbol{\theta}_0)$ is invertible, and that the asymptotic covariance $\Sigma = \mathbf{W}(\boldsymbol{\theta}_0)$ of the random vector given in (1.97) is well defined, so that the result follows. We will drop the subscript 0 for $\boldsymbol{\theta}_0$ for simplicity of exposition.

The calculation of $\mathbf{W}(\boldsymbol{\theta}_0)$, is straight forward. Its entries are given by:

$$W_{1,1} = \rho_1 var_{F1}(\frac{(g_1(x,\theta_1)-1)}{\rho_1(g_1(x_i,\theta_1))+\rho_2(g_2(x,\theta_2))}) + \rho_2 var_{F2}(\frac{(g_1(x,\theta_1)-1)}{\rho_1(g_1(x,\theta_1))+\rho_2(g_2(x,\theta_2))})$$

$$W_{2,2} = \rho_1 var_{F1}(\frac{(g_2(x,\theta_1)-1)}{\rho_1(g_1(x,\theta_1)) + \rho_2(g_2(x,\theta_2))}) + \rho_2 var_{F2}(\frac{(g_2(x_i,\theta_1)-1)}{\rho_1(g_1(x_i,\theta_1)) + \rho_2(g_2(x_i))})$$

$$W_{3,3} = \rho_1^3 var_{F1}(\frac{\frac{\partial}{\partial\theta_1} g_1(x,\theta_2)}{\rho_1(g_1(x_i,\theta_1)) + \rho_2(g_2(x,\theta_2))}) + \rho_2\rho_1^2 var_{F_2}(\frac{\frac{\partial}{\partial\theta_1} g_1(x,\theta_2)}{\rho_1(g_1(x,\theta_1)) + \rho_2(g_2(x_i))})$$

$$+\rho_1 I^{F_1}_{\theta_1\theta_1}$$

$$W_{4,4} = \rho_1\rho_2^2 var_{F1}(\frac{\frac{\partial}{\partial\theta_2} g_2(x,\theta_2)}{\rho_1(g_1(x,\theta_1)) + \rho_2(g_2(x_i))}) + \rho_2^3 var_{F_2}(\frac{\frac{\partial}{\partial\theta_2} g_2(x,\theta_2)}{\rho_1(g_1(x_i,\theta_1)) + \rho_2(g_2(x_i))})$$

$$+\rho_2 I^{F_2}_{\theta_2\theta_2}$$

$$W_{1,2} = W_{2,1} = \rho_1 cov_{F_1}(\frac{(g_1(x,\theta_1)-1)}{\rho_1(g_1(x,\theta_1)) + \rho_2(g_2(x,\theta_2))}, \frac{(g_2(x,\theta_2)-1)}{\rho_1(g_1(x,\theta_1)) + \rho_2(g_2(x,\theta_2))}) +$$
$$\rho_2 cov_{F_2}(\frac{(g_1(x,\theta_1)-1)}{\rho_1(g_1(x,\theta_1)) + \rho_2(g_2(x,\theta_2))}, \frac{(g_2(x,\theta_2)-1)}{\rho_1(g_1(x,\theta_1)) + \rho_2(g_2(x,\theta_2))})$$

$$W_{1,3} = W_{3,1} =$$

$$\rho_1 cov_{F_1}(\frac{(g_1(x,\theta_1)-1)}{\rho_1(g_1(x,\theta_1)) + \rho_2(g_2(x,\theta_2))}, \rho_1\frac{\frac{\partial}{\partial\theta_1} g_1(x,\theta_2)}{\rho_1(g_1(x_i,\theta_1)) + \rho_2(g_2(x,\theta_2))} - \frac{\partial}{\partial\theta_1}\log g_1(x_i,\theta_1))$$
$$+\rho_2 cov_{F_2}(\frac{(g_1(x,\theta_1)-1)}{\rho_1(g_1(x,\theta_1)) + \rho_2(g_2(x,\theta_2))}, \rho_1\frac{\frac{\partial}{\partial\theta_1} g_1(x_i,\theta_1)}{\rho_1(g_1(x_i,\theta_1)) + \rho_2(g_2(x_i))})$$

$$W_{2,3} = W_{3,2} =$$

$$\rho_1 cov_{F_1}(\frac{(g_2(x,\theta_2)-1)}{\rho_1(g_1(x,\theta_1)) + \rho_2(g_2(x,\theta_2))}, \rho_1\frac{\frac{\partial}{\partial\theta_1} g_1(x,\theta_1)}{\rho_1(g_1(x_i,\theta_1)) + \rho_2(g_2(x,\theta_2))} - \frac{\partial}{\partial\theta_1}\log g_1(x_i,\theta_1))$$
$$+\rho_2 cov_{F_2}(\frac{(g_2(x,\theta_2)-1)}{\rho_1(g_1(x,\theta_1)) + \rho_2(g_2(x,\theta_2))}, \rho_1\frac{\frac{\partial}{\partial\theta_1} g_1(x,\theta_1)}{\rho_1(g_1(x_i,\theta_1)) + \rho_2(g_2(x,\theta_2))})$$

$$W_{2,4} = W_{4,2} = \rho_1 cov_{F_1}(\frac{(g_2(x,\theta_2)-1)}{\rho_1(g_1(x,\theta_1)) + \rho_2(g_2(x,\theta_2))}, \rho_2\frac{\frac{\partial}{\partial\theta_2} g_2(x,\theta_2)}{\rho_1(g_1(x_i,\theta_1)) + \rho_2(g_2(x,\theta_2))})$$
$$+\rho_2 cov_{F_2}(\frac{(g_2(x,\theta_2)-1)}{\rho_1(g_1(x,\theta_1)) + \rho_2(g_2(x,\theta_2))}, \rho_2\frac{\frac{\partial}{\partial\theta_2} g_2(x,\theta_2)}{\rho_1(g_1(x_i,\theta_1)) + \rho_2(g_2(x_i))} - \frac{\partial}{\partial\theta_2}\log g_2(x,\theta_2))$$

$$W_{1,4} = W_{4,1} = \rho_1 cov_{F_1}(\frac{(g_1(x,\theta_1)-1)}{\rho_1(g_1(x,\theta_1)) + \rho_2(g_2(x,\theta_2))}, \rho_2\frac{\frac{\partial}{\partial\theta_2} g_2(x,\theta_2)}{\rho_1(g_1(x_i,\theta_1)) + \rho_2(g_2(x,\theta_2))})$$
$$+\rho_2 cov_{F_2}(\frac{(g_1(x,\theta_1)-1)}{\rho_1(g_1(x,\theta_1)) + \rho_2(g_2(x,\theta_2))}, \rho_2\frac{\frac{\partial}{\partial\theta_2} g_2(x,\theta_2)}{\rho_1(g_1(x_i,\theta_1)) - \rho_2(g_2(x_i))} - \frac{\partial}{\partial\theta_2}\log g_2(x,\theta_2))$$

$$W_{3,4} = W_{4,3}$$

$$= \rho_1 cov_{F_1}(\rho_1 \frac{\frac{\partial}{\partial \theta_1} g_1(x,\theta_1)}{\rho_1(g_1(x,\theta_1)) - \rho_2(g_2(x,\theta_2))} - \frac{\partial}{\partial \theta_1} \log g_1(x,\theta_1), \rho_2 \frac{\frac{\partial}{\partial \theta_2} g_2(x,\theta_2)}{\rho_1(g_1(x_i,\theta_1)) + \rho_2(g_2(x,\theta_2))})$$

$$+\rho_2 cov_{F_2}(\rho_1 \frac{\frac{\partial}{\partial \theta_1} g_1(x,\theta_1)}{\rho_1(g_1(x,\theta_1)) + \rho_2(g_2(x,\theta_2))}, \rho_2 \frac{\frac{\partial}{\partial \theta_2} g_2(x,\theta_2)}{\rho_1(g_1(x_i,\theta_1)) + \rho_2(g_2(x,\theta_2))} - \frac{\partial}{\partial \theta_2} \log g_2(x,\theta_2))$$

We now provide expressions for $\mathbf{S}(\boldsymbol{\theta})$. Each term in this 4×4 matrix can be easily found by using the relation

$$\frac{1}{n} \sum_{k=1}^{n} V(\theta, x_k) \xrightarrow{p} E_F[(\rho_1 g_1 + \rho_2 g_2)V],$$

namely,

$$\frac{\partial^2 l}{\partial \theta_i^2} = -\lambda_i \sum_{k=1}^{n} \frac{\frac{\partial^2}{\partial \theta_i^2} g_i(x_k,\theta_i)}{1 + \lambda_1(g_1(x_k,\theta_1) - 1) + \lambda_2(g_2(x_k,\theta_2) - 1)}$$

$$+ \lambda_i^2 \sum_{k=1}^{n} \frac{(\frac{\partial}{\partial \theta_i} g_i(x_k,\theta_i))^2}{(1 + \lambda_1(g_1(x_k,\theta_1) - 1) + \lambda_2(g_2(x_k,\theta_2) - 1))^2}$$

$$+ \sum_{k \in \mathcal{S}_i} \frac{\partial^2}{\partial \theta_i^2} \log g_i(x_k,\theta_i)$$

where $\mathcal{S}_i$ is the set of indexes in sample i.

$$\frac{1}{n} \frac{\partial^2 l}{\partial \theta_i^2}|_{\boldsymbol{\psi}=\boldsymbol{\psi}_0}$$

$$= -\frac{1}{n} \rho_i \sum_{k=1}^{n} \frac{\frac{\partial^2}{\partial \theta_i^2} g_i(x_k,\theta_i)}{\rho_1 g_1(x_k,\theta_1) + \rho_2(g_2(x_k,\theta_2))}$$

$$+\frac{1}{n} \rho_i^2 \sum_{k=1}^{n} \frac{(\frac{\partial}{\partial \theta_i} g_i(x_k,\theta_i))^2}{(\rho_1 g_1(x_k,\theta_1) + \rho_2(g_2(x_k,\theta_2)))^2}$$

$$+\frac{1}{n} \sum_{k \in \mathcal{S}_i} \frac{\partial^2}{\partial \theta_i^2} \log g_i(x_k,\theta_i)$$

$$\xrightarrow{p} -\rho_i E_F(\frac{\partial^2}{\partial \theta_i^2} g_i(x,\theta_i))$$

$$+\rho_i^2 E_F \frac{(\frac{\partial}{\partial \theta_i} g_i(x,\theta_i))^2}{(\rho_1 g_1(x,\theta_1) + \rho_2(g_2(x,\theta_2)))} - \rho_i I_{\theta_i \theta_i}^{F_i},$$

$$i = 1, 2.$$

$$\frac{1}{n}\frac{\partial^2 l}{\partial\theta_i\partial\theta_j}|_{\boldsymbol{\psi}=\boldsymbol{\psi}_0} =$$

$$\frac{1}{n}\rho_1\rho_2\sum_{k=1}^{n}\frac{\frac{\partial}{\partial\theta_i}g_i(x,\theta_i)\frac{\partial}{\partial\theta_j}g_i(x,\theta_i)}{(\rho_1 g_1(x_k,\theta_1)+\rho_2(g_2(x_k,\theta_2)))^2}$$

$$\xrightarrow{p}\rho_1\rho_2 E_F(\frac{\frac{\partial}{\partial\theta_i}g_i(x,\theta_i)\frac{\partial}{\partial\theta_j}g_i(x,\theta_i)}{(\rho_1 g_1(x_k,\theta_1)+\rho_2(g_2(x_k,\theta_2)))}), i \neq j$$

$$\frac{1}{n}\frac{\partial^2 l}{\partial\lambda_i\lambda_j}|_{\boldsymbol{\psi}=\boldsymbol{\psi}_0} =$$

$$\frac{1}{n}\sum_{k=1}^{n}\frac{(g_i(x_k,\theta_i)-1)(g_j(x_k,\theta_j)-1)}{(\rho_1(g_1(x_k,\theta_1))+\rho_2(g_2(x_k)))^2}$$

$$\xrightarrow{d} E_F(\frac{(g_i(x,\theta_i)-1)(g_j(x,\theta_j)-1)}{(\rho_1(g_1(x,\theta_1))+\rho_2(g_2(x)))}), i, j = 1, 2$$

Likewise,

$$\frac{1}{n}\frac{\partial^2 l}{\partial\theta_i\lambda_j}|_{\boldsymbol{\psi}=\boldsymbol{\psi}_0} =$$

$$= -\frac{1}{n}\sum_{k=1}^{n}\frac{\frac{\partial}{\partial\theta_i}g_j(x_k,\theta_i)}{(\rho_1(g_1(x,\theta_1))+\rho_2(g_2(x)))} + \frac{1}{n}\rho_i\sum_{k=1}^{n}\frac{(g_j(x_k,\theta_j)-1)\frac{\partial}{\partial\theta_i}g_i(x_k,\theta_i)}{(\rho_1(g_1(x,\theta_1))+\rho_2(g_2(x)))^2}$$

$$\xrightarrow{p}\rho_i E_F(\frac{(g_j(x_k,\theta_j)-1)\frac{\partial}{\partial\theta_i}g_i(x_k,\theta_i)}{(\rho_1(g_1(x,\theta_1))+\rho_2(g_2(x)))}), i, j = 1, 2.$$

It follows that under regularity conditions,

$$\sqrt{n}(\tilde{\boldsymbol{\psi}}-\boldsymbol{\psi}) \xrightarrow{d} \mathcal{N}(0, V) \qquad (1.98)$$

where $\mathbf{V} = \mathbf{S}(\boldsymbol{\theta}_0)^{-1}\mathbf{W}(\boldsymbol{\theta}_0)\mathbf{S}(\boldsymbol{\theta}_0)^{-1}$, and $\mathbf{W}(\boldsymbol{\theta}_0)$ and $\mathbf{S}(\boldsymbol{\theta}_0)$ are as specified in the preceding pages. □

A more general setup would be the model

$$f_i(x) = g_i(\boldsymbol{\theta}, x) f_0(x), i = 1, ..., m, \tag{1.99}$$

where $\boldsymbol{\theta}$ is a vector of dimension p and samples of size n_i are available from distributions $f_i(x)$,$i = 1, \ldots, m$, with $\sum_{i=1}^{m} n_i = n$ and $n_i/n \to \rho_i > 0$. No sample is available from f_0.

Sufficient conditions for identifiability of this model can be found in Gilbert *et al.* (1999)[22].

Defining $\boldsymbol{g}(\boldsymbol{\theta}, x)$ as the mapping with component functions g_i, and defining the vector $\boldsymbol{\lambda}^T = (\lambda_2, \ldots, \lambda_m)$ it is easy to see that we can obtain the following expression for the probability element $p_i = dF_0(x_i)$:

$$p_i = \frac{1}{n} \frac{1}{1 + \boldsymbol{\lambda}^T (\mathbf{g}(x_i, \boldsymbol{\theta}) - \mathbf{1})} \tag{1.100}$$

and the log likelihood

$$l(\boldsymbol{\lambda}, \boldsymbol{\theta}) = -\sum_{i=1}^{n} \log[1 + \boldsymbol{\lambda}^T (\mathbf{g}(x_i, \boldsymbol{\theta}) - \mathbf{1})] + \sum_{k=1}^{m} \sum_{i \in \mathcal{S}_k} \log g_i(x_i, \boldsymbol{\theta}). \tag{1.101}$$

Here $\mathcal{S}_k$ represents the set of indexes corresponding to the sample k for $k = 1, \ldots, m$. This model has the advantage that there is no need to have a reference sample and is also more general than the model expressed in Fokianos (2004) in that it allows the known weight functions $g_i(x_i, \boldsymbol{\theta})$ to vary, whereas Fokianos considers weights of the form $g(x_i, \theta_i)$. This expression for the likelihood can the be differentiated with respect to $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}$ to obtain estimators for the two parameters. The extension of the asymptotic results to this more general case is straightforward, but will not be pursued here.

1.5 Future Research

A question of interest is to ascertain the asymptotic properties of the estimator defined in Qin (1998) when $m = o(n)$ and $n = o(m)$. Also, analytical comparison of the asymptotic covariance matrix of the EMLE, EMME with the former estimator, although difficult, would be of interest, and moreover, a comparison with the information bounds provided in Gilbert (2000) [22] would be revealing in the case where $m = cn(1 + o(1))$, but this is also a very difficult task due to the complexity of the information bounds provided in Gilbert (2000). It is expected that the EMLE and the EMME will not meet those information bounds. Simulation studies comparing the performance of the aforementioned estimators would shed some light on how the methods perform in finite samples.

The extension of the EMLE to the s-sample case should be straightforward, and its asymptotic properties should be analogous to those in the two-sample case. Specifically, suppose s samples are available where sample i has probability element $dF_i(x, \boldsymbol{\theta}) = C(\boldsymbol{\theta})h_i(x, \boldsymbol{\theta})dF(x)$, $i = 1, \ldots, s$, and a sample is available from the generator, where all samples are independent. One may use the same approach as in the two-sample case to compute the EMLE.

A question of theoretical interest is to study the EMLE and EMME when the sample from the generator is $(Y_1, \ldots, Y_m)$ but the researcher instead collects a sample from $(Y_1^*, \ldots, Y_m^*)$ which is treated as the true generator sample, where the degree of misspecification is small. That is, suppose the true distribution is not F but rather $F^* = F(x, \eta_m)$, where $\eta_m \to 0$ as $m \to \infty$. One may study the asymptotic properties of the resulting estimators under various assumptions of the rate of convergence of $\eta_m \to 0$.

1.6 Appendix: An Alternative Proof of Consistency of the EMLE.

Theorem 10. *Let $\boldsymbol{\theta}_0$ be the true parameter value. Suppose that*

(i) $h(x, \boldsymbol{\theta})$ is continuously differentiable in a neighborhood of $\boldsymbol{\theta}_0$.

(ii) $\mathbf{I}_X(\boldsymbol{\theta}_0)$ is nonsingular.

(iv) $E(\frac{\partial}{\partial \boldsymbol{\theta}} h(Y, \boldsymbol{\theta}_0)) < \infty$

(v) The expected values are finite of each of the i.i.d averages comprising

$\frac{\partial}{\partial \boldsymbol{\theta}} \mathbf{G}_{m,n}(\boldsymbol{\theta}_0, Y_1, \ldots, Y_m, X_1, \ldots, X_n)$.

Then as $m, n \to \infty$, with $P_{\boldsymbol{\theta}_0}$ probability tending to one there exists a solution to the EMLE system of estimating equations, $\hat{\boldsymbol{\theta}}_{m,n}$, which is a consistent estimator of $\boldsymbol{\theta}_0$.

Proof. Consider the mapping

$$\mathbf{G}_{m,n} = \sum_{i=1}^{n} \frac{\frac{\partial}{\partial \boldsymbol{\theta}} h(X_i, \boldsymbol{\theta})}{h(X_i, \boldsymbol{\theta})} / n - \frac{\sum_{i=1}^{m} \frac{\partial}{\partial \boldsymbol{\theta}} h(Y_i, \boldsymbol{\theta})/m}{\sum_{i=1}^{m} h(Y_i, \boldsymbol{\theta})/m} \tag{1.102}$$

Notice that, for any fixed $X_1, \ldots, X_n, Y_1, \ldots, Y_m$, if $\hat{\boldsymbol{\theta}}_{m,n}$ exists, it is given by

$$\hat{\boldsymbol{\theta}}_{m,n} = \mathbf{G}_{m,n}^{-1}(\mathbf{0}, Y_1, \ldots, Y_m, X_1, \ldots, X_n).$$

Note that as $m, n \to \infty$,

$$\mathbf{G}_{m,n}(\boldsymbol{\theta}_0, Y_1, \ldots, Y_m, X_1, \ldots, X_n) \xrightarrow{p} \mathbf{0} \tag{1.103}$$

$$\frac{\partial}{\partial \boldsymbol{\theta}} \mathbf{G}_{m,n}(\boldsymbol{\theta}_0, Y_1, \ldots, Y_m, X_1, \ldots, X_n) \xrightarrow{p} -\mathbf{I}_X(\boldsymbol{\theta}_0). \tag{1.104}$$

Since $\mathbf{I}_X(\boldsymbol{\theta}_0)$ is nonsingular, with $P_{\boldsymbol{\theta}_0}$ probability tending to one, $\frac{\partial}{\partial \boldsymbol{\theta}} \mathbf{G}_{m,n}(\boldsymbol{\theta}_0, Y_1, \ldots, Y_m, X_1, \ldots, X_n)$ is nonsingular. Moreover, with $P_{\boldsymbol{\theta}_0}$ probability tending to one, for any $\delta > 0$

$$||\mathbf{G}_{m,n}(\boldsymbol{\theta}_0, Y_1, \ldots, Y_m, X_1, \ldots, X_n)|| < \delta.$$

Note that $\mathbf{G}_{m,n}(\boldsymbol{\theta}, Y_1, \ldots, Y_m, X_1, \ldots, X_n)$ is continuously differentiable in a neighborhood of $\boldsymbol{\theta}_0$ for all m, n by assumption (i). Choose δ such that by virtue of the inverse function theorem, $\mathbf{G}_{m,n}(\boldsymbol{\theta}, Y_1, \ldots, Y_m, X_1, \ldots, X_n)$ is one to one and onto in a neighborhood of $\boldsymbol{\theta}_0$, $\mathbf{G}^{-1}_{m,n}(\mathbf{a}, Y_1, \ldots, Y_m, X_1, \ldots, X_n)$ exists and is continuous for $\mathbf{a}$ in a neighborhood of $\mathbf{G}_{m,n}(\boldsymbol{\theta}_0, Y_1, \ldots, Y_m, X_1, \ldots, X_n)$ including $\mathbf{0}$, and such that for $\epsilon > 0$, by continuity of $\mathbf{G}^{-1}_{m,n}(\mathbf{a}, Y_1, \ldots, Y_m, X_1, \ldots, X_n)$,

$$||\mathbf{G}^{-1}_{m,n}(\mathbf{0}, Y_1, \ldots, Y_m, X_1, \ldots, X_n) - \mathbf{G}^{-1}_{m,n}(\mathbf{G}_{m,n}(\boldsymbol{\theta}_0), Y_1, \ldots, Y_m, X_1, \ldots, X_n)|| < \epsilon.$$

Then

$$||\hat{\boldsymbol{\theta}}_{m,n} - \boldsymbol{\theta}_0|| < \epsilon.$$

□

Chapter 2

A Related Two-Sample Parametric Problem

2.1 Introduction

A shortcoming of the EMLE and EMME is that the infinite dimensional nuisance parameter, the generator, is estimated from the sample $(Y_1, \ldots, Y_m)$ alone, whereas information about the generator is provided by both samples. The method described in Qin (1998) [56], on the other hand, estimates the generator using the data from both samples. Although direct comparison of the resulting asymptotic covariances is challenging in the semiparametric case, the issue raises an interesting analogous parametric problem.

Suppose a sample $(Y_1, \ldots, Y_m)$ is available from a population with density $f_Y(y; \lambda)$, and an independent sample $(X_1, \ldots, X_n)$ is available from a population with density $f_X(x, \lambda; \psi)$. Here λ is regarded as a nuisance parameter and ψ is the structural parameter. For simplicity we deal with the case where λ and ψ are scalars.

One approach to estimation of ψ would be to compute the maximum likelihood estimator based on both samples (Method 1). A second approach would be to first find the maximum likelihood estimator of λ from the sample $(Y_1, \ldots, Y_m)$, namely $\hat{\lambda}_m$, and then to treat $\hat{\lambda}_m$ as the true parameter. That is, one treats $(X_1, \ldots, X_n)$ as if the sample comes from $f_X(x; \hat{\lambda}_m, \psi)$, and then computes the maximum likelihood of ψ (Method 2). Method 2 resembles the EMLE approach in the semiparametric case, where the generator (the infinite dimensional nuisance parameter) is replaced by an estimator that is treated as if it were the true distribution.

Intuitively, Method 2 should result in inferior estimation for ψ (as well as λ) due to the fact that it does not use the information provided by the second sample when estimating λ. On the other hand, it has the advantage of reducing the dimension of the system of equations (in this case, from 2 dimensions to 1 dimension). It is interesting to explore the asymptotic distributions of both methods. The results are rather natural. It should be noted that in the case where $m = cn(1 + o(1))$, Method 1 is efficient, so that Method 2 cannot outperform Method 1 in estimating ψ in terms of first-order asymptotics (see for instance, Lehmann and Casella (1998) [40], Theorem 7.1).

For convenience, let

$$J_\lambda^{y_i} = \frac{\partial}{\partial \lambda} \log f_y(y_i, \lambda)$$

$J_\lambda^{y_i}$ is the Fisher score for λ from observation y_i. Let $I_\lambda^y = E(J_\lambda^y)^2$ be the Fisher information on λ from one observation y_i from the first sample. Likewise

$$J_\psi^{x_i} = \frac{\partial}{\partial \psi} \log f_x(x_i, \lambda, \psi)$$

$$J_\lambda^{x_i} = \frac{\partial}{\partial \lambda} \log f_x(x_i, \lambda, \psi)$$

$$J_{\psi\psi}^{x_i} = \frac{\partial^2}{\partial \psi^2} \log f_x(x_i, \lambda, \psi)$$

$$J_{\lambda\lambda}^{x_i} = \frac{\partial^2}{\partial \lambda^2} \log f_x(x_i, \lambda, \psi)$$

$$J^{x_i}_{\psi\lambda} = J^{x_i}_{\lambda\psi} = \frac{\partial^2}{\partial\psi\partial\lambda}\log f_x(x_i,\lambda,\psi)$$
$$J^{x_i}_{\psi\lambda\lambda} = \frac{\partial^3}{\partial\psi\partial\lambda^2}\log f_x(x_i,\lambda,\psi)$$
$$J^{x_i}_{\lambda\lambda\lambda} = \frac{\partial^3}{\partial\lambda\partial\lambda\partial\lambda}\log f_x(x_i,\lambda,\psi)$$

Other derivatives of the scores based on the samples from f_x and f_y are denoted analogously. Then $I^x_{\psi\psi} = E(J^{x_i}_{\psi})^2 = -E(J^{x_i}_{\psi\psi})$ and $I^x_{\lambda\psi}$ $I^x_{\lambda\lambda}$, and $I^x_{\psi\lambda}$ are defined analogously in obvious ways.

2.2 Asymptotic Properties of Estimator from Method 1

Let the estimators resulting from method 1 be denoted as $\hat{\lambda}^{(1)}_{m,n}$ and $\hat{\psi}^{(1)}_{m,n}$, and let $\boldsymbol{\theta} = (\psi,\lambda)$ and $\hat{\boldsymbol{\theta}}^{(1)}_{m,n} = (\hat{\psi}^{(1)}_{m,n}, \hat{\lambda}^{(1)}_{m,n})$ The likelihood is given by

$$L(\mathbf{X},\mathbf{Y},\psi,\lambda) = \Pi^m_{i=1} f_y(y_i,\lambda)\Pi^n_{i=1} f_x(x_i,\lambda,\psi).$$

The log likelihood is given by

$$l(\psi,\lambda) = \sum_{i=1}^{m}\log f_y(y_i,\lambda) + \sum_{i=1}^{n}\log f_x(x_i,\lambda,\psi). \tag{2.1}$$

Thus, the estimating equations are given by

$$\frac{\partial l}{\partial\psi} = \sum_{i=1}^{n}\frac{\partial}{\partial\psi}\log f_x(x_i,\lambda,\psi) = \sum_{i=1}^{n} J^{x_i}_{\psi} = 0, \tag{2.2}$$

$$\frac{\partial l}{\partial\lambda} = \sum_{i=1}^{m}\frac{\partial}{\partial\lambda}\log f_y(y_i,\lambda) + \sum_{i=1}^{n}\frac{\partial}{\partial\lambda}\log f_x(x_i,\lambda,\psi) = \sum_{i=1}^{m} J^{y_i}_{\lambda} + \sum_{i=1}^{n} J^{x_i}_{\lambda} = 0. \tag{2.3}$$

We also compute the Jacobian of l, which will be needed both to prove consistency and to find the asymptotic covariance of $\hat{\psi}^{(1)}_{m,n}$. Its entries are:

$$\frac{\partial^2 l}{\partial\psi^2} = \sum_{i=1}^{n}\frac{\partial^2}{\partial\psi^2}\log f_x(x_i,\lambda), \tag{2.4}$$

$$\frac{\partial^2 l}{\partial \lambda^2} = \sum_{i=1}^{m} \frac{\partial^2}{\partial \lambda^2} \log f_y(y_i, \lambda) + \sum_{i=1}^{n} \frac{\partial^2}{\partial \lambda^2} \log f_x(x_i, \lambda, \psi), \tag{2.5}$$

$$\frac{\partial^2 l}{\partial \lambda \partial \psi} = \sum_{i=1}^{n} \frac{\partial^2}{\partial \lambda \partial \psi} \log f_x(x_i, \lambda, \psi). \tag{2.6}$$

Thus the Jacobian of l is

$$\begin{bmatrix} \sum_{i=1}^{n} J^{x_i}_{\psi\psi} & \sum_{i=1}^{n} J^{x_i}_{\psi\lambda} \\ \sum_{i=1}^{n} J^{x_i}_{\psi\lambda} & \sum_{i=1}^{n} J^{y_i}_{\lambda\lambda} + \sum_{i=1}^{m} J^{x_i}_{\lambda\lambda} \end{bmatrix}. \tag{2.7}$$

Consider the following mapping and corresponding partial derivative matrix:

$$G_{m,n} = \begin{bmatrix} \frac{1}{n} \sum_{i=1}^{n} J^{x_i}_{\psi} \\ \frac{m}{n} \frac{1}{m} \sum_{i=1}^{m} J^{y_i}_{\lambda} + \frac{1}{n} \sum_{i=1}^{n} J^{x_i}_{\lambda} \end{bmatrix}, \tag{2.8}$$

$$\frac{\partial}{\partial \boldsymbol{\theta}} G_{m,n} = \begin{bmatrix} \frac{1}{n} \sum_{i=1}^{n} J^{x_i}_{\psi\psi} & \frac{1}{n} \sum_{i=1}^{n} J^{x_i}_{\psi\lambda} \\ \frac{1}{n} \sum_{i=1}^{n} J^{x_i}_{\psi\lambda} & \frac{m}{n} \frac{1}{m} \sum_{i=1}^{m} J^{y_i}_{\lambda\lambda} + \frac{1}{n} \sum_{i=1}^{m} J^{x_i}_{\lambda\lambda} \end{bmatrix}. \tag{2.9}$$

It is clear that if $m = cn(1 + o(1))$, $G_{m,n}(\boldsymbol{\theta}_0) \xrightarrow{p} \mathbf{0}$ and that

$$\frac{\partial}{\partial \boldsymbol{\theta}} G_{m,n}(\boldsymbol{\theta}_0) \xrightarrow{p} \begin{bmatrix} -I^x_{\psi\psi} & -I^x_{\psi\lambda} \\ -I^x_{\psi\lambda} & -cI^y_{\lambda} - I^x_{\lambda\lambda} \end{bmatrix}. \tag{2.10}$$

Under some regularity conditions, it follows from Lemma 3 that $\hat{\boldsymbol{\theta}}^{(1)}_{m,n}$ is a consistent estimator of $\boldsymbol{\theta}$. Note that the argument is valid when $m = cn(1+o(1)), c > 0$, and when $m = o(n)$ so that $m/n \to 0$ which implies $c = 0$. To prove that consistency holds for when $n = o(m)$, the same argument does not hold because $m/n \to \infty$. In such a case, we need to consider a slightly different system of equations whose solution is also $\hat{\boldsymbol{\theta}}^{(1)}_{m,n}$, namely

$$G^*_{m,n} = \begin{bmatrix} \frac{1}{n}\sum_{i=1}^{n} J^{x_i}_{\psi} \\ \frac{1}{m}\sum_{i=1}^{m} J^{y_i}_{\lambda} + \frac{n}{m}\frac{1}{n}\sum_{i=1}^{n} J^{x_i}_{\lambda} \end{bmatrix}. \tag{2.11}$$

Note that $G^*_{m,n}(\boldsymbol{\theta}_0) \xrightarrow{p} \mathbf{0}$. Moreover, its derivative matrix is given by

$$\frac{\partial}{\partial\boldsymbol{\theta}} G^*_{m,n} = \begin{bmatrix} \frac{1}{n}\sum_{i=1}^{n} J^{x_i}_{\psi\psi} & \frac{1}{n}\sum_{i=1}^{n} J^{x_i}_{\psi\lambda} \\ \frac{n}{m}\frac{1}{n}\sum_{i=1}^{n} J^{x_i}_{\psi\lambda} & \frac{1}{m}\sum_{i=1}^{m} J^{y_i}_{\lambda\lambda} + \frac{n}{m}\frac{1}{n}\sum_{i=1}^{n} J^{x_i}_{\lambda\lambda} \end{bmatrix}, \tag{2.12}$$

so that as $n, m \to \infty$,

$$\frac{\partial}{\partial\boldsymbol{\theta}} G^*_{m,n}(\boldsymbol{\theta}_0) \xrightarrow{p} \begin{bmatrix} -I^x_{\psi\psi} & -I^x_{\psi\lambda} \\ 0 & -I^y_{\lambda} \end{bmatrix}. \tag{2.13}$$

This matrix is negative definite provided $I^x_{\psi\psi} I^y_{\lambda} > 0$. By Lemma 3, under some conditions, this proves consistency of $\hat{\boldsymbol{\theta}}^{(1)}_{m,n}$ when $n = o(m)$. Thus it follows that $\hat{\boldsymbol{\theta}}^{(1)}_{m,n}$ is consistent provided $m, n \to \infty$ regardless of the rate of growth of m relative to n

We will now explore the asymptotic distribution of $\hat{\boldsymbol{\theta}}_{m,n}^{(1)}$ under each of the three settings.

Theorem 11. *Let $m, n \to \infty$. Suppose $m = cn(1 + o(1)), c > 0$. Under some regularity conditions*

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_{m,n}^{(1)} - \boldsymbol{\theta}) \xrightarrow{d} N(\mathbf{0}, \mathbf{I}(c, \boldsymbol{\theta})^{-1})$$

where $\mathbf{I}(c, \boldsymbol{\theta})$ is given by

$$\begin{bmatrix} I_{\psi\psi}^x & I_{\psi\lambda}^x \\ I_{\psi\lambda}^x & cI_{\lambda}^y + I_{\lambda\lambda}^x \end{bmatrix}.$$

Proof. Let $\boldsymbol{\theta}_0$ be the true value. Consider the Taylor expansion of each component function $G_{m,n,j}$ at $\boldsymbol{\theta}_0$:

$$\begin{aligned} 0 &= G_{m,n,j}(\hat{\boldsymbol{\theta}}_{m,n}) \\ &= G_{m,n,j}(\boldsymbol{\theta}_0) + \sum_{k=1}^{2} (\hat{\theta}_{m,n,k}^{(1)} - \theta_{0,k}) \frac{\partial}{\partial \theta_k} G_{m,n,j}(\boldsymbol{\theta}_0) \\ &+ \frac{1}{2} \sum_{k=1}^{2} \sum_{l=1}^{2} (\hat{\theta}_{m,n,k}^{(1)} - \theta_{0,k})(\hat{\theta}_{m,n,l}^{(1)} - \theta_{0,l}) \frac{\partial^2}{\partial \theta_k \theta_l} G_{m,n,j}(\boldsymbol{\theta}^*) \end{aligned}$$

where $\boldsymbol{\theta}^*$ is a point on the line segment connecting $\boldsymbol{\theta}_0$ and $\hat{\boldsymbol{\theta}}_{m,n}^{(1)}$. Rearranging, we obtain

$$\begin{aligned} &\sqrt{n} G_{m,n,j}(\boldsymbol{\theta}_0) \\ &= -\sum_{k=1}^{2} \sqrt{n}(\hat{\theta}_{m,n,k}^{(1)} - \theta_{0,k})[\frac{\partial}{\partial \theta_k} G_{m,n,j}(\boldsymbol{\theta}_0) + \frac{1}{2} \sum_{l=1}^{2} (\hat{\theta}_{m,n,l}^{(1)} - \theta_{0,l}) \frac{\partial^2}{\partial \theta_k \theta_l} G_{m,n,j}(\boldsymbol{\theta}^*)]. \end{aligned} \tag{2.14}$$

Again we use Lemma 2. We already found the asymptotic distribution of $\frac{\partial}{\partial \boldsymbol{\theta}} G_{m,n}(\boldsymbol{\theta}_0)$. We must now find the distribution of $\sqrt{n} G_{m,n}(\boldsymbol{\theta}_0)$, which can be expressed as:

$$\sqrt{n}G_{m,n} = \begin{bmatrix} \frac{1}{\sqrt{n}}\sum_{i=1}^{n} J_{\psi}^{x_i} \\ \frac{\sqrt{m}}{\sqrt{n}}\frac{1}{\sqrt{m}}\sum_{i=1}^{m} J_{\lambda}^{y_i} + \frac{1}{\sqrt{n}}\sum_{i=1}^{n} J_{\lambda}^{x_i} \end{bmatrix}. \tag{2.15}$$

Clearly the following asymptotic result follows:

$$\sqrt{n}G_{m,n}(\boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}_2(\mathbf{0}, \mathbf{I}(c, \boldsymbol{\theta}_0)). \tag{2.16}$$

Referring back to Lemma 2, we must consider

$$A_{j,k,m,n} = \frac{\partial}{\partial\theta_k}G_{m,n,j}(\boldsymbol{\theta}_0) + \frac{1}{2}\sum_{l=1}^{2}(\hat{\theta}_l^{(1)} - \theta_l)(\frac{\partial^2}{\partial\theta_k\theta_l})G_{m,n,j}(\boldsymbol{\theta}^*). \tag{2.17}$$

We have already found that first term converges in probability to $\{\mathbf{I}(c, \boldsymbol{\theta}_0)\}_{\{j,k\}}$.

The second term converges in probability to 0 if $\frac{\partial^2}{\partial\theta_k\theta_l}G_{m,n,j}(\boldsymbol{\theta}^*)$ are bounded in probability for all j, k, l, which holds under the assumption that all the second partial derivatives with respect to $\boldsymbol{\theta}$ of the functions J_{ψ}^{x}, J_{λ}^{x}, and J_{λ}^{y} are dominated by integrable functions not depending on $\boldsymbol{\theta}$ in a neighborhood of $\boldsymbol{\theta_0}$.

□

Corollary 12. *Suppose $m, n \to \infty$ such that $m = cn(1 + o(1))$. Then under regularity conditions*

$$\sqrt{n}(\hat{\psi}_{m,n}^{(1)} - \psi) \xrightarrow{d} \mathcal{N}\left[0, \frac{cI_{\lambda}^{y} + I_{\lambda\lambda}^{x}}{I_{\psi\psi}^{x}(cI_{\lambda}^{y} + I_{\lambda\lambda}^{x}) - (I_{\psi\lambda}^{x})^2}\right]$$

Proof. Take the appropriate entry of the inverse of the asymptotic variance in Theorem 11. □

Corollary 13. *Suppose $m, n \to \infty$ such that $m = o(n)$. Then*

$$\sqrt{n}(\hat{\psi}_{m,n}^{(1)} - \psi) \xrightarrow{d} \mathcal{N}(0, (\hat{I}_{\psi|\lambda})^{-1})$$

where $\hat{I}^x_{\psi|\lambda}$ is the efficient information of ψ in the presence of λ given by the expression

$$\hat{I}^x_{\psi|\lambda} = I^x_{\psi\psi} - I^x_{\psi\lambda}(I^x_{\lambda\lambda})^{-1} I^x_{\lambda\psi}.$$

Proof. Let $c \to 0$. Notice that this is equivalent to $m/n \to 0$. □

There is an interesting interpretation to Corollary 13: When the sample from $f_y(y, \lambda)$ is very small relative to the sample from $f_x(x, \theta, \lambda)$, the asymptotic behavior of the resulting estimator $\hat{\psi}_{m,n}^{(1)}$ is the same as that of the MLE from the sample from $f_x(x, \theta, \lambda)$ alone, where λ is an unknown nuisance parameter. That is, the asymptotic distribution of $\hat{\psi}_{m,n}^{(1)}$ is the same as if the sample from $f_y(y, \lambda)$ were not available and maximum likelihood were used in the sample from $f_x(x, \lambda, \psi)$.

Method 1 is asymptotically efficiency under regularity conditions when $m = cn(1+o(1))$, $c > 0$, as noted in the introduction. Thus $\hat{\boldsymbol{\psi}}_{m,n}$ must have an asymptotic covariance at least as small as of that of the MLE $\hat{\boldsymbol{\psi}}_n$ from the sample $(X_1, \ldots, X_n)$ alone. The following corollary provides a direct proof based on the derived asymptotic variances.

Corollary 14. *Denote the asymptotic variance of $\hat{\psi}_{m,n}^{(1)}$ when $m = cn(1 + o(1))$, $c > 0$, as σ_1^2. Assume $I^y_\lambda > 0$, $\mathbf{I}_x$ is nonsingular, and $I^x_{\psi\lambda} \neq 0$.*

Then $\sigma_1^2 < \hat{I}^x_{\psi|\lambda}$.

Proof. If

$$\sigma_1^2 = \frac{cI^y_\lambda + I^x_{\lambda\lambda}}{I^x_{\psi\psi}(cI^y_\lambda + I^x_{\lambda\lambda}) - (I^x_{\psi\lambda})^2} < \frac{1}{I^x_{\psi\psi} - I^x_{\psi\lambda}(I^x_{\lambda\lambda})^{-1}I^x_{\lambda\psi}},$$

then equivalently,

$$\begin{aligned}
\frac{I^x_{\psi\psi}(cI^y_\lambda + I^x_{\lambda\lambda}) - (I^x_{\psi\lambda})^2}{cI^y_\lambda + I^x_{\lambda\lambda}} &> I^x_{\psi\psi} - I^x_{\psi\lambda}(I^x_{\lambda\lambda})^{-1}I^x_{\lambda\psi}, \\
\frac{(I^x_{\psi\lambda})^2}{cI^y_\lambda + I^x_{\lambda\lambda}} &< I^x_{\psi\lambda}(I^x_{\lambda\lambda})^{-1}I^x_{\lambda\psi}, \\
cI^y_\lambda + I^x_{\lambda\lambda} &> I^x_{\lambda\lambda}, \\
cI^y_\lambda &> 0.
\end{aligned}$$

Since $c > 0$ and $I^y_\lambda > 0$ the result follows. □

Note that if $I^x_{\lambda\psi} = 0$, it is easy to see that $\sigma_1^2 = \hat{I}^x_{\psi|\lambda}$.

Theorem 15. *Let $n, m \to$ such that $n = o(m)$. Under regularity conditions,*

$$\begin{bmatrix} \sqrt{n}(\hat{\psi}^{(1)}_{m,n} - \psi) \\ \sqrt{m}(\hat{\lambda}^{(1)}_{m,n} - \lambda) \end{bmatrix} \xrightarrow{d} \mathcal{N}\left\{\mathbf{0}, \begin{bmatrix} \frac{1}{I^x_{\psi\psi}} & 0 \\ 0 & \frac{1}{I^y_\lambda} \end{bmatrix}\right\}$$

Proof. Let $\boldsymbol{\theta}_0$ be the true parameter. Let

$$G_{m,n,1} = \frac{1}{n}\sum_{i=1}^{n} J^{x_i}_\psi \tag{2.18}$$

$$G_{m,n,2} = \frac{1}{m}\sum_{i=1}^{n} J^{x_i}_\lambda + \frac{1}{m}\sum_{i=1}^{m} J^{y_i}_\lambda \tag{2.19}$$

and $G_{m,n} = (G_{m,n,1}, G_{m,n,2})^T$.

A Taylor expansion of each component function about the true $\boldsymbol{\theta}_0$ evaluated at $\hat{\boldsymbol{\theta}}^{(1)}_{m,n}$

gives

$$0 = \sqrt{n}G_{m,n,1}(\boldsymbol{\theta}_0) \tag{2.20}$$

$$+\sqrt{n}(\hat{\psi}^{(1)}_{m,n}-\psi_0)\{\frac{\partial}{\partial\psi}G_{m,n,1}(\boldsymbol{\theta}_0)+\frac{1}{2}\frac{\partial^2}{\partial\psi^2}G_{m,n,1}(\boldsymbol{\xi})(\hat{\psi}^{(1)}_{m,n}-\psi_0)+\frac{1}{2}\frac{\partial^2}{\partial\psi\partial\lambda}G_{m,n,1}(\boldsymbol{\xi})(\hat{\lambda}^{(1)}_{m,n}-\lambda_0)\} \tag{2.21}$$

$$+\sqrt{m}(\hat{\lambda}^{(1)}_{m,n}-\lambda_0)\{\frac{\sqrt{n}}{\sqrt{m}}(\frac{\partial}{\partial\lambda}G_{m,n,1}(\boldsymbol{\theta}_0)+\frac{1}{2}\frac{\partial^2}{\partial\lambda^2}G_{m,n,1}(\boldsymbol{\xi})(\hat{\lambda}^{(1)}_{m,n}-\lambda_0)+\frac{1}{2}\frac{\partial^2}{\partial\lambda\partial\psi}G_{m,n,1}(\boldsymbol{\xi})(\hat{\psi}^{(1)}_{m,n}-\psi_0))\} \tag{2.22}$$

$$0 = \sqrt{m}G_{m,n,2}(\boldsymbol{\theta}_0) \tag{2.23}$$

$$+\sqrt{n}(\hat{\psi}^{(1)}-\psi_0)\{\frac{\sqrt{m}}{\sqrt{n}}(\frac{\partial}{\partial\psi}G_{m,n,2}(\boldsymbol{\theta}_0)+\frac{1}{2}\frac{\partial^2}{\partial\psi^2}G_{m,n,2}(\boldsymbol{\eta})(\hat{\psi}^{(1)}_{m,n}-\psi_0)+\frac{1}{2}\frac{\partial^2}{\partial\psi\partial\lambda}G_{m,n,2}(\boldsymbol{\eta})(\hat{\lambda}^{(1)}_{m}-\lambda_0))\} \tag{2.24}$$

$$+\sqrt{m}(\hat{\lambda}^{(1)}_{m,n}-\lambda_0)\{(\frac{\partial}{\partial\lambda}G_{m,n,2}(\boldsymbol{\theta}_0)+\frac{1}{2}\frac{\partial^2}{\partial\lambda^2}G_{m,n,2}(\boldsymbol{\eta})(\hat{\lambda}^{(1)}_{m,n}-\lambda_0)+\frac{1}{2}\frac{\partial^2}{\partial\lambda\partial\psi}G_{m,n,2}(\boldsymbol{\eta})(\hat{\psi}^{(1)}_{m,n}-\psi_0))\} \tag{2.25}$$

where $\boldsymbol{\eta}$ and $\boldsymbol{\xi}$ are points on the line segment between $\boldsymbol{\theta}_0$ and $\hat{\boldsymbol{\theta}}_{m,n}$. Denote the term in brackets in (2.21) as $A_{m,n}$, the term in brackets in (2.22) as $B_{m,n}$, the term in brackets in (2.24) as $C_{m,n}$, and the term in brackets (2.25) as $D_{m,n}$. Note that we can express the system of equation given by (2.20)-(2.25) as follows:

$$\begin{bmatrix} A_{m,n} & B_{m,n} \\ C_{m,n} & D_{m,n} \end{bmatrix} \begin{bmatrix} \sqrt{n}(\hat{\psi}^{(1)}_{m,n}-\psi_0) \\ \sqrt{m}(\hat{\lambda}^{(1)}_{m,n}-\lambda_0) \end{bmatrix} = \begin{bmatrix} -\sqrt{n}G_{m,n,1} \\ -\sqrt{m}G_{m,n,2} \end{bmatrix}. \tag{2.26}$$

□

We will show that under regularity conditions, $A_{m,n} \xrightarrow{p} -I^{x}_{\psi\psi}$, $B_{m,n} \xrightarrow{p} 0$,

$C_{m,n} \xrightarrow{p} 0$, $D_{m,n} \xrightarrow{p} -I^y_\lambda$, and that

$$\begin{bmatrix} \sqrt{n}G_{m,n,1} \\ \\ \sqrt{m}G_{m,n,2} \end{bmatrix} \xrightarrow{d} \mathcal{N}\left(0, \begin{bmatrix} cI^x_{\psi\psi} & 0 \\ \\ 0 & I^y_\lambda \end{bmatrix}\right).$$

The result then follows from Lemma 2, Chapter 1.

We note that

$$\begin{aligned} A_{m,n} &= \frac{\partial}{\partial\psi}G_{m,n,1}(\boldsymbol{\theta}_0) + \frac{1}{2}\frac{\partial^2}{\partial\psi^2}G_{m,n,1}(\boldsymbol{\xi})(\hat{\psi}^{(1)}_{m,n} - \psi_0) + \frac{1}{2}\frac{\partial^2}{\partial\psi\partial\lambda}G_{m,n,1}(\boldsymbol{\xi})(\hat{\lambda}^{(1)}_{m,n} - \lambda_0) \\ &= \frac{1}{n}\sum_{i=1}^n J^{x_i}_{\psi\psi}(\boldsymbol{\theta}_0) + \frac{1}{2}\frac{1}{n}\sum_{i=1}^n J^{x_i}_{\psi\psi\psi}(\boldsymbol{\xi})(\hat{\psi}^{(1)}_{m,n} - \psi_0) + \frac{1}{2}\frac{1}{n}\sum_{i=1}^n J^{x_i}_{\psi\psi\lambda}(\boldsymbol{\xi})(\hat{\lambda}^{(1)}_{m,n} - \lambda_0) \end{aligned} \tag{2.27}$$

so that

$$A_{m,n} \xrightarrow{p} -I^x_{\psi\psi} \tag{2.28}$$

provided that $|J^x_{\psi\psi\psi}| < Q(x)$ and $|J^x_{\psi\psi\lambda}| < R(x)$ in a neighborhood of $\boldsymbol{\theta}_0$, with $E_{\boldsymbol{\theta}_0}(Q(X)) < \infty$ and $E_{\boldsymbol{\theta}_0}(R(X)) < \infty$. Also

$$\begin{aligned} B_{m,n} &= \frac{\sqrt{n}}{\sqrt{m}}(\frac{\partial}{\partial\lambda}G_{m,n,1}(\boldsymbol{\theta}_0) + \frac{1}{2}\frac{\partial^2}{\partial\lambda^2}G_{m,n,1}(\boldsymbol{\xi})(\hat{\lambda}^{(1)}_{m,n} - \lambda_0) + \frac{1}{2}\frac{\partial^2}{\partial\lambda\partial\psi}G_{m,n,1}(\boldsymbol{\xi})(\hat{\psi}^{(1)}_{m,n} - \psi_0)) \\ &= \frac{\sqrt{n}}{\sqrt{m}}(\frac{1}{n}\sum_{i=1}^n J^{x_i}_{\lambda\psi}(\boldsymbol{\theta}_0) + \frac{1}{2n}\sum_{i=1}^n J^{x_i}_{\psi\lambda\lambda}(\boldsymbol{\xi})(\hat{\lambda}^{(1)}_{m,n} - \lambda_0) + \frac{1}{2n}\sum_{i=1}^n J^{x_i}_{\psi\psi\lambda}(\boldsymbol{\xi})(\hat{\psi}^{(1)}_{m,n} - \psi_0)) \end{aligned} \tag{2.29}$$

so that

$$B_{m,n} \xrightarrow{p} 0. \tag{2.30}$$

provided $|J^x_{\psi\lambda\lambda}| < M(x)$ and $|J^x_{\psi\lambda\lambda}| < N(x)$ in a neighborhood of $\boldsymbol{\theta}_0$, with $E_{\boldsymbol{\theta}_0}(M(X)) < \infty$ and $E_{\boldsymbol{\theta}_0}(N(X)) < \infty$.

$$
\begin{aligned}
&C_{m,n} \\
&= \frac{\sqrt{m}}{\sqrt{n}}\left(\frac{\partial}{\partial\psi}G_{m,n,2}(\boldsymbol{\theta}_0) + \frac{1}{2}\frac{\partial^2}{\partial\psi^2}G_{m,n,2}(\boldsymbol{\eta})(\hat{\psi}^{(1)}_{m,n} - \psi_0) + \frac{1}{2}\frac{\partial^2}{\partial\psi\partial\lambda}G_{m,n,2}(\boldsymbol{\eta})(\hat{\lambda}^{(1)}_{m,n} - \lambda_0)\right) \\
&= \frac{1}{\sqrt{m}\sqrt{n}}\sum_{i=1}^{n} J^{x_i}_{\psi\lambda} + \frac{1}{2\sqrt{m}\sqrt{n}}\sum_{i=1}^{n} J^{x_i}_{\psi\psi\lambda}(\boldsymbol{\eta})(\hat{\psi}^{(1)}_{m,n} - \psi_0) + \frac{1}{2\sqrt{m}\sqrt{n}}\sum_{i=1}^{n} J^{x_i}_{\psi\lambda\lambda}(\boldsymbol{\eta})(\hat{\lambda}^{(1)}_{m,n} - \lambda_0)\}^{\cdot}
\end{aligned}
\tag{2.31}
$$

The first term clearly converges in probability to zero. Under the assumption that $|J^{x}_{\psi\psi\lambda}| < H(x)$ and $|J^{x}_{\psi\lambda\lambda}| < W(x)$ in a neighborhood of $\boldsymbol{\theta}_0$ with $E_{\boldsymbol{\theta}_0}(H(X)) < \infty$ and $E_{\boldsymbol{\theta}_0}(W(X)) < \infty$, the second and third terms also converge in probability to 0 due to consistency of $\hat{\boldsymbol{\theta}}^{(1)}_{m,n}$. Thus

$$
C_{m,n} \xrightarrow{p} 0. \tag{2.32}
$$

$$
\begin{aligned}
D_{m,n} &= \left(\frac{\partial}{\partial\lambda}G_{m,n,2}(\boldsymbol{\theta}_0) + \frac{1}{2}\frac{\partial^2}{\partial\lambda^2}G_{m,n,2}(\boldsymbol{\eta})(\hat{\lambda}^{(1)}_{m,n} - \lambda_0) + \frac{1}{2}\frac{\partial^2}{\partial\lambda\partial\psi}G_{m,n,2}(\boldsymbol{\eta})(\hat{\psi}^{(1)}_{m,n} - \psi_0)\right) \\
&= \frac{1}{m}\sum_{i=1}^{n} J^{x_i}_{\lambda\lambda}(\boldsymbol{\theta}_0) + \frac{1}{m}\sum_{i=1}^{m} J^{y_i}_{\lambda\lambda}(\boldsymbol{\theta}_0) \\
&+ \frac{1}{2}\left(\frac{1}{m}\sum_{i=1}^{n} J^{x_i}_{\lambda\lambda\lambda}(\boldsymbol{\eta}) + \frac{1}{m}\sum_{i=1}^{m} J^{y_i}_{\lambda\lambda\lambda}(\boldsymbol{\eta})(\hat{\lambda}^{(1)}_{m,n} - \lambda_0) + \frac{1}{2}\frac{1}{m}\sum_{i=1}^{n} J^{x_i}_{\psi\lambda\lambda}(\boldsymbol{\eta})(\hat{\psi}^{(1)}_{m,n} - \psi_0)\right)
\end{aligned}
\tag{2.33}
$$

so that

$$
D_{m,n} \xrightarrow{p} -I^{y}_{\lambda} \tag{2.34}
$$

under the conditions that $|J^{x}_{\lambda\lambda\lambda}| < P(x)$, $|J^{x}_{\lambda\lambda\lambda}| < Q(x)$, $|J^{x}_{\psi\lambda\lambda}| < R(x)$, $|J^{y}_{\lambda\lambda\lambda}| < S(Y)$ and $|J^{x}_{\psi\lambda\lambda}| < T(x)$ in neighborhoods of $\boldsymbol{\theta}_0$, where

$$
E_{\boldsymbol{\theta}_0}(P(X)), E_{\boldsymbol{\theta}_0}(Q(X)), E_{\boldsymbol{\theta}_0}(S(Y)), E_{\boldsymbol{\theta}_0}(R(X)), E_{\boldsymbol{\theta}_0}(S(Y)), E_{\boldsymbol{\theta}_0}(T(X))
$$

are finite, due to the LLN and the consistency of $\hat{\lambda}^{(1)}_{m,n}$ and $\hat{\psi}^{(1)}_{m,n}$.

Notice also that

$$\begin{bmatrix} \sqrt{n}G_{m,n,1} \\ \\ \sqrt{m}G_{m,n,2} \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{n}}\sum_{i=1}^{n} J_{\psi}^{x_i} \\ \\ \frac{1}{\sqrt{m}}\sum_{i=1}^{n} J_{\lambda}^{x_i} + \frac{1}{\sqrt{m}}\sum_{i=1}^{m} J_{\lambda}^{y_i} \end{bmatrix} \xrightarrow{d} \mathcal{N}\left(0, \begin{bmatrix} I_{\psi\psi}^{x} & 0 \\ \\ 0 & I_{\lambda}^{y} \end{bmatrix}\right) \tag{2.35}$$

Using Lemma 2 from Chapter 1, the result follows under some additional regularity conditions.

2.3 Asymptotic Properties of Estimator from Method 2

We now turn our attention to Method 2, which is perhaps a more naive method of estimation. We denote the resulting estimator as $\hat{\boldsymbol{\theta}}_{m,n}^{(2)} = (\hat{\psi}_{m,n}^{(2)}, \hat{\lambda}_{m}^{(2)})$.

Under the same data setup as before, we compute the MLE of λ, namely $\hat{\lambda}_{m}^{(2)}$, from the sample $(Y_1, \ldots, Y_m)$ alone. In the sample $(X_1, \ldots, X_n)$ we treat λ as if it were known and equal to $\hat{\lambda}_{m}^{(2)}$, and then compute the maximum estimator of ψ. This reduces the dimension of the system of equations to estimate ψ, which when λ and ψ are scalars becomes a one dimensional equation:

$$G_{m,n}^{(2)}(\mathbf{X}, \hat{\lambda}_{m}^{(2)}, \psi) = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial\psi} \log f_x(x_i, \psi, \hat{\lambda}_{m}^{(2)}) = \frac{1}{n}\sum_{i=1}^{n} J_{\psi}^{x_i}(\psi, \hat{\lambda}_{m}^{(2)}) = 0 \tag{2.36}$$

Under regularity conditions, the asymptotic distribution of $\hat{\lambda}_{m}^{(2)}$ is given by

$$\sqrt{m}(\hat{\lambda}_{m}^{(2)} - \lambda) \xrightarrow{d} \mathcal{N}\left(0, \frac{1}{I_{\lambda}^{y}}\right). \tag{2.37}$$

To prove consistency of $\hat{\psi}_{m,n}^{(2)}$, we observe that if we let $\boldsymbol{\theta}_0$ be the true parameter,

$$\frac{1}{n}\sum_{i=1}^{n} J_{\psi}^{x_i}(\psi_0, \hat{\lambda}_{m}^{(2)}) = \frac{1}{n}\sum_{i=1}^{n} J_{\psi}^{x_i}(\psi_0, \lambda_0) + \frac{1}{n}\sum_{i=1}^{n} J_{\psi\lambda}^{x_i}(\psi_0, \eta)(\hat{\lambda}_{m}^{(2)} - \lambda_0) \tag{2.38}$$

for η strictly between λ_0 and $\hat{\lambda}_m^{(2)}$. The first term converges in probability to its mean, 0, when evaluated at the true parameter value $\boldsymbol{\theta}_0$. The second term also converges in probability to 0, assuming $\hat{\lambda}_m^{(2)}$ is consistent and assuming that $|J^x_{\psi\lambda}(\psi_0, \lambda)| < M(x)$ with $E_{\boldsymbol{\theta}_0}(M(X)) < \infty$ in a neighborhood of the true λ_0.

To check condition (ii) of Lemma 3, we note that

$$\frac{1}{n}\sum_{i=1}^{n} J^{x_i}_{\psi\psi}(\psi_0, \hat{\lambda}_m^{(2)}) = \frac{1}{n}\sum_{i=1}^{n} J^{x_i}_{\psi\psi}(\psi_0, \lambda_0) + \frac{1}{n}\sum_{i=1}^{n} J^{x_i}_{\psi\psi\lambda}(\psi_0, \eta_2)(\hat{\lambda}_m^{(2)} - \lambda). \qquad (2.39)$$

It follows that if $|J^x_{\psi\psi\lambda}(\psi_0, \lambda)| < N(x)$ in a neighborhood of λ_0, with $E_{\boldsymbol{\theta}_0}(N(X)) < \infty$, this expression converges in probability to $-I^x_{\psi\psi}$.

Moreover, if $|J^x_{\psi\psi\psi\lambda}(\psi_0, \lambda)| < P(x)$ on neighborhood of λ_0, with $E_{\boldsymbol{\theta}_0}(P(X)) < \infty$, condition (iii) of Lemma 3 is satisfied and it follows that with probability tending to one there exists a consistent solution $\hat{\psi}^{(2)}_{m,n}$ of the equation $G_{m,n}(\psi) = 0$.

Note that for consistency of $\hat{\boldsymbol{\theta}}^{(2)}_{m,n}$ to hold, we need $n, m \to \infty$ but we do not need to make any assumption about the rate of growth of m relative to n. The same is true for $\hat{\boldsymbol{\theta}}^{(1)}_{m,n}$

Theorem 16. *Let $m, n \to \infty$ such that $m = cn(1 + o(1))$. Under regularity conditions,*

$$\sqrt{n}(\hat{\psi}^{(2)}_{m,n} - \psi) \xrightarrow{d} \mathcal{N}\left[0, (I^x_{\psi\psi})^{-1} + \frac{1}{c}\frac{(I^x_{\psi\lambda})^2}{(I^x_{\psi\psi})^2 I^y_{\lambda}}\right]$$

.

Proof. Let $\boldsymbol{\theta}_0$ be the true value of $\boldsymbol{\theta}$. Then

$$\begin{aligned}0 = G^{(2)}_{m,n}(\hat{\psi}^{(2)}_{m,n}) &= G^{(2)}_{m,n}(\psi_0) + \frac{\partial}{\partial\psi}G^{(2)}_{m,n}(\psi_0)(\hat{\psi}^{(2)}_{m,n} - \psi_0) + \frac{1}{2}\frac{\partial^2}{\partial\psi^2}G^{(2)}_{m,n}(\eta)(\hat{\psi}^{(2)}_{m,n} - \psi_0)^2 \\ &= G^{(2)}_{m,n}(\psi_0) + (\hat{\psi}^{(2)}_{m,n} - \psi_0)(\frac{\partial}{\partial\psi}G^{(2)}_{m,n}(\psi_0) + \frac{1}{2}\frac{\partial^2}{\partial\psi^2}G^{(2)}_{m,n}(\eta)(\hat{\psi}^{(2)}_{m,n} - \psi_0))\end{aligned}$$

so that

$$-\sqrt{n}(\hat{\psi}^{(2)}_{m,n} - \psi_0) = \frac{\sqrt{n}G^{(2)}_{m,n}(\psi_0)}{\frac{\partial}{\partial\psi}G^{(2)}_{m,n}(\psi_0) + \frac{1}{2}\frac{\partial^2}{\partial\psi^2}G^{(2)}_{m,n}(\eta)(\hat{\psi}^{(2)}_{m,n} - \psi_0)}, \tag{2.40}$$

where

$$\begin{aligned}G^{(2)}_{m,n}(\psi) &= \frac{1}{n}\sum_{i=1}^{n} J^{x_i}_{\psi}(\psi_0, \hat{\lambda}^{(2)}_m) \\ &= \frac{1}{n}\sum_{i=1}^{n} J^{x_i}_{\psi}(\psi_0, \lambda_0) + \frac{1}{n}\sum_{i=1}^{n} J^{x_i}_{\psi\lambda}(\psi_0, \lambda_0)(\hat{\lambda}^{(2)}_m - \lambda) + \frac{1}{n}\sum_{i=1}^{n} J^{x_i}_{\psi\lambda\lambda}(\psi_0, \xi)(\hat{\lambda}^{(2)}_m - \lambda_0)^2.\end{aligned} \tag{2.41}$$

Notice that

$$\begin{aligned}\sqrt{n}G^{(2)}_{m,n}(\psi_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} J^{x_i}_{\psi}(\psi_0, \lambda_0) &+ \sqrt{\frac{n}{m}}\frac{1}{n}\sum_{i=1}^{n} J^{x_i}_{\psi\lambda}(\psi_0, \lambda_0)\sqrt{m}(\hat{\lambda}^{(2)}_m - \lambda_0) \\ &+ \frac{1}{\sqrt{m}}\sqrt{\frac{n}{m}}\frac{1}{n}\sum_{i=1}^{n} J^{x_i}_{\psi\lambda\lambda}(\psi_0, \eta)m(\hat{\lambda}^{(2)}_m - \lambda)^2.\end{aligned} \tag{2.42}$$

Note that the first term and the second term converge in distribution to normal distributions with 0 mean and with variances $I^x_{\psi\psi}$ and $c^{-1}(I^x_{\psi\lambda})^2/I^y_\lambda$ respectively, and the last term converges in probability to 0 under regularity conditions. Thus,

$$\sqrt{n}G^{(2)}_{m,n}(\psi) \xrightarrow{d} \mathcal{N}\left[0, I^x_{\psi\psi} + \frac{1}{c}\frac{(I^x_{\psi\lambda})^2}{I^y_\lambda}\right] \tag{2.43}$$

We have already shown that $\partial/\partial\psi G^{(2)}_{m,n} \xrightarrow{p} -I^x_{\psi\psi}$ and clearly the second term in the denominator converges in probability to 0 under some regularity conditions.

This follows since

$$\frac{1}{2}\frac{1}{n}\sum_{i=1}^{n} J^{x_i}_{\psi\psi\psi}(\eta, \hat{\lambda}^{(2)}_m)(\hat{\psi}^{(2)}_{m,n} - \psi_0)$$

$$= \frac{1}{2}\frac{1}{n}\sum_{i=1}^{n} J^{x_i}_{\psi\psi\psi}(\eta, \lambda_0)(\hat{\psi}^{(2)}_{m,n} - \psi_0) + \frac{1}{2}\frac{1}{n}\sum_{i=1}^{n} J^{x_i}_{\psi\psi\psi\lambda}(\eta, \xi)(\hat{\lambda}^{(2)}_m - \lambda_0)(\hat{\psi}^{(2)}_{m,n} - \psi_0). \quad (2.44)$$

Thus assuming $|J^{x}_{\psi\psi\psi}(x, \psi, \lambda_0)| < Q(x)$ and $|J_{\psi\psi\psi\lambda}(x, \psi, \lambda)| < R(x)$ on neighborhoods of the true ψ_0 and $\boldsymbol{\theta}_0$, respectively, where $E_{\boldsymbol{\theta}_0}(Q(X)) < \infty$ and $E_{\boldsymbol{\theta}_0}(P(X)) < \infty$, (2.44) converges in probability to 0 due to the consistency of $\hat{\boldsymbol{\theta}}^{(2)}_{m,n}$. □

Corollary 17. *Let $m, n \to \infty$ such that $n = o(m)$. Under regularity conditions,*

$$\sqrt{n}(\hat{\psi}^{(2)}_{m,n} - \psi) \xrightarrow{d} \mathcal{N}(0, \frac{1}{I^{x}_{\psi\psi}}).$$

Proof. This follows by letting $c \to \infty$, and is also easy to see from the proof of Theorem 14 by letting $m/n \to 0$ □

Remark: Theorem 15 and Corollary 17 imply that the asymptotic distributions of $\hat{\psi}^{(1)}_{m,n}$ and $\hat{\psi}^{(2)}_{m,n}$ are the same when $n = o(m)$. When the sample $(Y_1, \ldots, Y_m)$ is very large, both estimators have asymptotic distributions equivalent to those that would have been obtained through maximum likelihood from the sample $(X_1, \ldots, X_n)$ if λ were known.

Theorem 18. *Let $m, n \to \infty$ such that $m = o(n)$. Under regularity conditions,*

$$\sqrt{m}(\hat{\psi}^{(2)}_{m,n} - \psi) \xrightarrow{d} \mathcal{N}\left[0, \frac{(I^{x}_{\psi\lambda})^2}{(I^{x}_{\psi\psi})^2 I^{y}_{\lambda}}\right]$$

.

Proof. By a Taylor expansion:

$$\begin{aligned}
0 &= \sum_{i=1}^{n} J_{\psi}^{x_i}(\hat{\psi}_{m,n}^{(2)}, \hat{\lambda}_m^{(2)}) \\
&= \sum_{i=1}^{n} J_{\psi}^{x_i}(\psi_0, \hat{\lambda}_m^{(2)}) + \sum_{i=1}^{n} J_{\psi\psi}^{x_i}(\psi_0, \hat{\lambda}_m^{(2)})(\hat{\psi}_{m,n}^{(2)} - \psi_0) \\
&+ \frac{1}{2}\sum_{i=1}^{n} J_{\psi\psi\psi}^{x_i}(\xi, \hat{\lambda}_m^{(2)})(\hat{\psi}_{m,n}^{(2)} - \psi_0)^2
\end{aligned}$$

Thus we have

$$\sqrt{m}(\hat{\psi}_{m,n}^{(2)} - \psi_0) = \frac{-\sqrt{m}\frac{1}{n}\sum_{i=1}^{n} J_{\psi}^{x_i}(\psi_0, \hat{\lambda}_m^{(2)})}{\frac{1}{n}\sum_{i=1}^{n} J_{\psi\psi}^{x_i}(\psi_0, \hat{\lambda}_m^{(2)}) + \frac{1}{2}\frac{1}{n}\sum_{i=1}^{n} J_{\psi\psi\psi}^{x_i}(\xi, \hat{\lambda}_m^{(2)})(\hat{\psi}_{m,n}^{(2)} - \psi_0)}. \quad (2.45)$$

Let's examine the numerator first. To do so, we perform another Taylor expansion:

$$\begin{aligned}
&\sqrt{m}\frac{1}{n}\sum_{i=1}^{n} J_{\psi}^{x_i}(\psi_0, \hat{\lambda}_m^{(2)}) \\
&= \frac{\sqrt{m}}{\sqrt{n}}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} J_{\psi}^{x_i}(, \psi_0, \lambda_0) + \frac{1}{n}\sum_{i=1}^{n} J_{\psi\lambda}^{x_i}(\psi_0, \lambda_0)\sqrt{m}(\hat{\lambda}_m^{(2)} - \lambda_0) \qquad (2.46) \\
&+ \frac{1}{2}\frac{1}{\sqrt{m}}\frac{1}{n}\sum_{i=1}^{n} J_{\lambda\lambda\psi}^{x_i}(\psi_0, \eta) m(\hat{\lambda}_m^{(2)} - \lambda_0)^2.
\end{aligned}$$

Note that the first sum converges in probability to 0, the second term converges in distribution to $\mathcal{N}(0, (I_{\psi\lambda}^{x})^2/I_{\lambda}^{y})$, and the third term converges in probability to 0 under some assumptions analogous to those given for other Taylor expansions.

□

Theorem 16, Corollary 17, and Theorem 18 show asymptotic results analogous to those of the EMLE and EMME. For instance, when $n = o(m)$ the asymptotic distribution of $\hat{\psi}_{m,n}^{(2)}$ is the asymptotic distribution of the MLE based on $(X_1, \ldots, X_n)$ if λ were fully known, and is thus efficient. When $m = cn(1 + o(1))$ the asymptotic variance of $\hat{\psi}^{(2)}_{m,n}$ has an additional nonnegative term. When $m = o(n)$ the rate of convergence is $1/\sqrt{m}$ instead of $1/\sqrt{n}$.

2.4 Comparison of Method 1 and Method 2

Comparing Method 1 and Method 2 when $m = o(n)$ show that, in large samples, $\hat{\psi}^{(1)}$ is a better estimator than $\hat{\psi}^{(2)}$ because its rate of convergence is faster. On the other hand, when $n = o(m)$ our results show that the two estimators have the same asymptotic distribution. In this case, $\hat{\lambda}^{(2)}_m$ is a very accurate estimator of λ.

In the case where $m = cn(1 + o(1)), \quad c > 0$, the asymptotic variance of $\sqrt{n}(\hat{\psi}^{(2)}_{m,n} - \psi_0)$, which we will denote as σ_2^2, should not be greater than that of $\sqrt{n}(\hat{\psi}^{(1)}_{m,n} - \psi_0)$, denoted as σ_1^2, since the former is asymptotically efficient under regularity conditions, as previously noted. In the following theorem, we compare the covariances of the two estimators based on direct computation.

Theorem 19. *For $m = cn(1 + o(1)), c > 0$, suppose $\mathbf{I}^x$ is nonsingular, $I^x_\lambda > 0$, and $I^x_{\psi\lambda} \neq 0$. Then $\sigma_2^2 > \sigma_1^2$.*

Proof. Suppose

$$\sigma_1^2 = \frac{cI^y_\lambda + I^x_{\lambda\lambda}}{I^x_{\psi\psi}(cI^y_\lambda + I^x_{\lambda\lambda}) - (I^x_{\psi\lambda})^2} < (I^x_{\psi\psi})^{-1} + \frac{1}{c}\frac{(I^x_{\psi\lambda})^2}{(I^x_{\psi\psi})^2 I^y_\lambda} = \sigma_2^2.$$

Then, equivalently,

$$\begin{aligned}
\frac{cI^y_\lambda + I^x_{\lambda\lambda}}{(cI^y_\lambda + I^x_{\lambda\lambda}) - (I^x_{\psi\lambda})^2/I^x_{\psi\psi}} &< 1 + \frac{1}{c}\frac{(I^x_{\psi\lambda})^2}{(I^x_{\psi\psi})I^y_\lambda},\\
\frac{cI^y_\lambda + I^x_{\lambda\lambda} - (I^x_{\psi\lambda})^2/I^x_{\psi\psi} + (I^x_{\psi\lambda})^2/I^x_{\psi\psi}}{cI^y_\lambda + I^x_{\lambda\lambda} - (I^x_{\psi\lambda})^2/I^x_{\psi\psi}} &< 1 + \frac{1}{c}\frac{(I^x_{\psi\lambda})^2}{(I^x_{\psi\psi})I^y_\lambda},\\
1 + \frac{(I^x_{\psi\lambda})^2/I^x_{\psi\psi}}{cI^y_\lambda + I^x_{\lambda\lambda} - (I^x_{\psi\lambda})^2/I^x_{\psi\psi}} &< 1 + \frac{1}{c}\frac{(I^x_{\psi\lambda})^2}{(I^x_{\psi\psi})I^y_\lambda},\\
\frac{(I^x_{\psi\lambda})^2/I^x_{\psi\psi}}{cI^y_\lambda + I^x_{\lambda\lambda} - (I^x_{\psi\lambda})^2/I^x_{\psi\psi}} &< \frac{1}{c}\frac{(I^x_{\psi\lambda})^2}{(I^x_{\psi\psi})I^y_\lambda},\\
\frac{cI^y_\lambda}{cI^y_\lambda + I^x_{\lambda\lambda} - (I^x_{\psi\lambda})^2/I^x_{\psi\psi}} &< 1,\\
\frac{cI^y_\lambda}{cI^y_\lambda + \hat{I}^x_{\lambda|\psi}} &< 1,
\end{aligned}$$

and finally

$$\hat{I}^x_{\lambda|\psi} > 0.$$

The latter is true because $\mathbf{I}^x$ is nonsingular, so its inverse must be nonsingular. □

Remark: These results show that when there are two samples providing information about a nuisance parameter, for large samples, estimation improves when using both samples for its estimation. However, note that if $I^x_{\psi\lambda} = 0$, $\sigma_2^2 = \sigma_1^2$ when $m = cn(1 + o(1))$, $c > 0$, and that the two estimators $\hat{\psi}^{(1)}_{m,n}$ and $\hat{\psi}^{(2)}_{m,n}$ are asymptotically equivalent when $n = o(m)$.

Chapter 3

Interval Estimation for Small Area Proportions with Small True Proportions from Stratified Random Sampling Survey Data

3.1 Introduction

In this chapter we study two small area estimation models to construct confidence intervals for domain proportions from data collected from stratified random sampling surveys.

A commonly used 95% confidence interval for the true proportion P_i of a domain i in a survey is $\hat{P}_i^{EB} \pm 1.96\sqrt{mse_i}$, where $\hat{P}_i^{EB}$ and mse_i are an empirical Bayes estimator of P_i and an associated second-order unbiased mean squared error estimator $(i = 1, \dots, m)$ (Rao, 2003) [61]. There are different choices of $\hat{P}_i^{EB}$ and mse_i in the literature–the specific estimators we will study here are described in Section 3.2. The underlying model for this popular method uses normal approximations: $p_i|P_i \sim N(P_i, \psi_i)$, $P_i \sim N(\mu, A)$, where p_i is the sample proportion for domain i based on a sample of size n_i; $i = 1, \dots, m$; ψ_i are known smoothed sampling variances; μ and A are unknown hyperparameters. This interval, which we will call the Normal-Normal interval, relies on the accuracy of the aforementioned normal approximations.

As is well-known to statisticians, the normal approximation to the binomial can be problematic when the probability of success is in the extremes and the number of Bernoulli trials is small. Students in elementary statistics classes learn that the normal approximation to a binomial random variable Y with parameters r and p, where r is the number of replications and p is the probability of success, is only reasonable when rp and rq are "large". Brown *et al.* (2001 [9],2002[10]) showed that problems may ensue even in cases where rp and rq are quite large. In particular, they showed that the actual coverage of the Wald interval, given by $Y/r \pm z_{\alpha/2}\sqrt{(Y/r)(1-Y/r)/n}$, may fall well below the nominal coverage in several examples. These include cases where p is not close to 0 or 1 and where rp and rq are greater than 10, a rule of thumb sometimes given in introductory statistics books. In fact, the actual coverage tends to oscillate both as r increases with p being fixed and as p varies with r fixed, making the actual coverage for a particular problem difficult to predict. This phenomenon is due to the discreteness and skewness of the binomial distribution. The erratic coverage properties of the normal approximation to the binomial raises questions about the performance of the Normal-Normal CI when the underlying true proportions are small.

Brown *et al.* (2000) discuss several other methods of constructing confidence intervals for proportions, many of which perform better than the Wald interval. One of note is also based on a normal approximation–the Wilson Interval. Like the Wald interval, the Wilson interval is derived from an asymptotic pivot. For the Wald interval the pivot is $\frac{p-\hat{p}}{\sqrt{\hat{p}(1-\hat{p})/r}}$ and for the Wilson interval the pivot is $\frac{p-\hat{p}}{\sqrt{p(1-p)/r}}$. The Wilson interval also shows oscillation as p and r vary, but it is less severe than that of the Wald interval. It is one of the two intervals recommended by Brown *et al.* for small sample sizes. Although both intervals are based on a normal approximation to the binomial, it should be noted that the Wald Interval bears a greater resemblance to Normal-Normal CI, which is of the form $\hat{P}_i^{EB} \pm 1.96\sqrt{mse_i}$.

A related paper is Liu and Kott (2009) [42]. Kott and Liu study one-sided coverage intervals for a proportion based on data collected from simple random sampling and stratified random sampling. In particular, they compare several methods for confidence interval construction through simulations. They note that the intervals that are effective for two-sided intervals differ than those that are effective for one-sided intervals. The intervals that are explored in the stratified random sampling case are not based on small area estimation models–the focus is to provide an interval for the overall proportion based on data collected form stratified random sampling surveys. For the simple random sampling case, they provide a comparison of several intervals for $n = 30$ over $0 \leq p \leq 1$, providing simulation results for lower bound estimates. They point out that for $n = 20, 60$, and 120, the trends are similar than those of the case $n = 30$. They are interested in finding methods that perform well over the range of p in the sense of having a coverage that is closest to the nominal, and do not focus on finding the method that is best when p is in the extremes. Using this criteria, among their conclusions is that the Wald interval is systematically biased, as are the Angular method and the Logit method, which are based on normality approximations with variance stabilizing transformations, although the Logit method's systematic bias is of lesser degree and of the opposite direction than that of the Wald Interval. Their preferred intervals for the SRS case are the Cai and the Kott-Liu intervals, which are modifications to the Wald interval.

We study the coverage properties of the Normal-Normal CI in the setting where there are are a large number of domains and we are interested in producing confidence intervals for each of the true domain proportions. We focus primarily in the case of small areas (i.e., small domain sample sizes) where the true proportion for each domain is small. Analogous results should hold for small areas where the true proportions are close to one.

We assume a stratified random sampling design, where a simple random sample is taken from each domain of interest. Typically, stratified random sampling designs are used in situations where the strata differ significantly from each other with respect to the variable of interest. Because we assume all domain proportions are small, our strata are fairly similar to each other but are sampled separately to ensure that data is collected from all domains of interest. Thus, the domains are the strata. Such a situation may arise, for instance, when studying the proportion of a rare disease in each of several domains.

For simplicity we assume equal stratum sample sizes (i.e., $n_i = n$). Small area models benefit greatly from the presence of relevant covariates but here we will assume that covariates are not available but that we still wish to “borrow strength" from different domains. For examples in the literature of such an approach, see Efron and Morris (1975) [16], and Carter and Rolph (1974) [11].

In a stratified random sampling setting with small stratum sample sizes, it is reasonable to assume the sampled domain counts Y_i, given P_i, follow a binomial distribution. To use normal approximations to the binomial in this case is not warranted due to the small sample sizes.

To apply an empirical Bayes approach, we must specify the distribution of the true proportions P_i. The beta distribution is a reasonable choice since its support is $(0, 1)$ and since its shape varies greatly with different choices of its parameters, allowing some flexibility in the model.

Thus a reasonable model in this setting is:

$$Y_i|P_i \sim Bin(n, P_i), \tag{3.1}$$

and

$$P_i \sim Beta(a, b) \tag{3.2}$$

Assumption (3.1) is appropriate due to the assumption that small SRS samples are taken from each domain.

It will be shown subsequently that assuming a beta prior implies the P_i's are not likely to differ by much since a small prior mean μ implies a small prior variance σ^2 for the beta prior. This feature of the model is consistent with the scenario of interest, where all domain proportions are small.

In Section 3.2 we give more details on the Normal-Normal CI. In Section 3.3 we derive an empirical Bayes confidence interval based on the assumed model given by (3.1) and (3.2), which we call the Binomial-Beta confidence interval. In Section 3.4 we compare through simulation studies the Normal-Normal CI to the Binomial-Beta CI, under the assumption that the true model is given by (3.1) and (3.2). We find that in our simulations the Normal-Normal CI does suffer from undercoverage and performs worse than the Binomial-Beta CI in terms of coverage in cases of interest. There is also some evidence of the oscillatory behavior of the Normal-Normal CI coverages both as the prior mean μ changes with n and m held fixed, as m increases with all other parameters held fixed, and as n increases with everything else held fixed.

Moreover, although our interest is primarily in small areas, we provide an example where there is some undercoverage even for very large n.

The Normal-Normal CI does appear to perform well in terms of coverage when the prior mean is not close to 0 or 1 and when the domains are large, according to our simulation results.

For the Binomial-Beta CI the proposed estimators for the hyperparameters (a,b) depend on the estimator for a related hyperparameter, δ. The proposed method of estimation appears to yield an estimator $\hat{\delta}$ that is always in the admissible range– specifically, δ has the property that $0 < \delta < 1$, and based on extensive simulations $\hat{\delta}$ appears to share the same property under the assumed model. The equation used to estimate δ depends on a constant, C, which is a small positive number. The best choice of C depends on the true value of the parameters and on n, the domain sample size. This is a drawback of the model, but in our simulations, we provide evidence that it is possible to find a C that yields the desired coverages for a range of small prior means and variances for a given n, so that the method may be of practical use in the situations of interest.

3.2 Normal-Normal Empirical Bayes CI

This widely used CI rests on the following assumptions:

$$p_i|P_i \sim N(P_i, \psi_i), \tag{3.3}$$

$$P_i \sim N(\mu, A). \tag{3.4}$$

The model is a special case of the famous Fay-Herriot Model (Fay and Herriot, 1979) [17] and is closely related to the Efron-Morris model (Efron and Morris, 1975) [16], and to the models described in Carter and Rolph (1974) [11]. In particular, the Normal-Normal model is the balanced domain sample size version of a model which is chosen to analyze a fire alarm dataset in the former paper. Three other related models are also discussed and considered for analysis of the data, including a two level model using an arcsine variance stabilizing transformation.

Level 1 (3.3) is usually referred to as the sampling model and level 2 (3.4) is usually referred to the linking model (Jiang and Lahiri, 2006) [25]. The sampling variabilities ψ_i are assumed to be known, although they typically need to be estimated. This is a weakness of the Fay-Herriot model–it does not incorporate the uncertainty due to estimation of ψ_i.

Returning to the empirical Bayes setup defined by (3.3) and (3.4), we note that because the Normal distribution is its own conjugate prior, the posterior distribution of $P_i|p_i$ is normal with mean

$$\gamma_i p_i + (1-\gamma_i)\mu \tag{3.5}$$

where

$$\gamma_i = \frac{A}{A+\psi_i}. \tag{3.6}$$

The parameter γ_i is called the shrinkage factor. Note that γ_i determines weights to the area-specific estimator and the prior mean μ. An estimator for P_i is given by

$$\hat{\gamma}_i p_i + (1-\hat{\gamma}_i)\bar{p} \tag{3.7}$$

where

$$\bar{p} = \frac{\sum_{i=1}^{m} p_i}{m}, \tag{3.8}$$

and where an estimator for γ_i will be given subsequently.

A two-sided $100(1-\alpha)\%$ empirical Bayes CI is given by:

$$(\hat{P}_i^{EB} \pm z_{\alpha/2}\sqrt{mse_i}) \tag{3.9}$$

where $z_{\alpha/2}$ represents the appropriate quantile of the standard normal distribution. The following estimators are used for A, γ_i, and the mse_i:

$$\hat{\psi}_i = \hat{\psi} = (\bar{p})(1-\bar{p})/n. \tag{3.10}$$

$$\hat{A} = \max\left\{0, (m-1)^{-1}\sum_{i=1}^{m}(p_i-\bar{p})^2 - \hat{\psi}\right\}. \tag{3.11}$$

$$\hat{\gamma}_i = \hat{\gamma} = \frac{\hat{A}}{\hat{\psi}+\hat{A}}. \tag{3.12}$$

$$mse_i^{EB} = \sqrt{g_1(\hat{A}) + g_2(\hat{A}) + 2g_3(\hat{A})}, \tag{3.13}$$

where

$$g_1(\hat{A}) = \hat{\gamma}\hat{\psi}, \tag{3.14}$$

$$g_2(\hat{A}) = \left(\frac{\hat{\psi}}{\hat{\psi}+\hat{A}}\right)^2 \left(\sum_{j=1}^{m}\frac{1}{\hat{\psi}+\hat{A}}\right)^{-1} = \frac{\hat{\psi}^2}{m(\hat{\psi}+\hat{A})}, \tag{3.15}$$

$$g_3(\hat{A}) = \left[\frac{(1-\hat{\gamma})^2}{\hat{\psi}+\hat{A}}\right]\widehat{Var(A)} = \frac{2(1-\hat{\gamma})^2(\hat{\psi}+\hat{A})}{m}. \tag{3.16}$$

Discussion of formulas (3.11-3.16) can be found in Rao (2001) [62] or Jiang and Lahiri (2006) [25]. The sum of (3.14) and (3.15) alone, as an estimator of mse_i, would be a naive estimator because it would not account for the uncertainty due to the estimation of A (Rao, 2003) [61]. If one were to use the sum of these two terms alone the bias would be of order $O(1/m)$. The estimator (3.13) was proposed by

Prasad and Rao (1990) [53] and has bias of order $o(1/m)$.

The estimator for A given by (3.11) truncates the unbiased estimator $(m - 1)^{-1}\sum_{i=1}^{m}(p_i - \bar{p})^2 - \hat{\psi}$ whenever it is negative.

Although the sampling variances ψ_i are not equal, they are all estimated by the same quantity as illustrated in (3.10). This is because this estimator is more stable than estimators based on the data from only a single domain. The former are unreliable due to the small domain sample sizes. This approach is similar to that discussed in Carter and Rolph (1974) [11], although there the domain sample sizes are not equal which results in differing estimates for ψ_i. If covariates were available, a better approach for estimating the ψ_i would be possible. For instance, one may estimate the ψ_i using the generalized variance function (see Wolter 1985 [71], Chapter 5).

Variants of the Fay-Herriot model are frequently used in surveys, so that the model defined by (3.4) and (3.5) can be viewed as a benchmark.

Our focus is on two-sided intervals. When dealing with one binomial proportion, the source of the undercoverage problem is twofold: it is due to the discreteness and to the skewness of the binomial distribution. The former plays a dominant role in two-sided intervals. In one-sided intervals the error due to skewness can be dominant, so that the best intervals in the one-sided case may differ from those in the two-sided case (Brown *et al.*, 2000 [9], 2001 [10]). This phenomenon should carry over to normality-based small area confidence intervals, but investigation of this issue is beyond the scope of this chapter.

3.3 Binomial-Beta Empirical Bayes CI

The Binomial-Beta confidence interval is built under the model given by (3.1) and (3.2); that is $Y_i|P_i \sim \text{Bin}(n, P_i)$, and $P_i \sim \text{Beta}(a, b)$. This is regarded as the "true" model. As previously discussed, the underlying model assumptions are more reasonable than the normal distribution assumptions, particularly in the cases of interest where the normal approximation to the binomial is inappropriate.

Since the beta distribution is a conjugate prior for the binomial, the posterior distribution $P_i|Y_i$ follows a beta distribution. If a and b were known, a credible interval for P_i would be

$$L_i = B(\alpha/2, y_i + a, n - y_i + b) \tag{3.17}$$

$$U_i = B(1 - \alpha/2, y_i + a, n - y_i + b) \tag{3.18}$$

Under our proposed method of estimation, to estimate the hyperparameters a and b we first estimate δ, given by

$$\delta = \frac{1}{a + b + 1}. \tag{3.19}$$

The hyperparameter δ specifies a relationship between the prior mean μ and the prior variance σ^2:

$$\sigma^2 = \mu(1-\mu)\delta. \tag{3.20}$$

The hyperparameters μ and σ^2 can be expressed in terms of a and b, where

$$\mu = \frac{a}{a+b}, \quad \sigma^2 = \frac{a-1}{a+b-2}.$$

It is easy to see that δ has the property that $0 < \delta < 1$. Also, note that δ is directly proportional to the prior variance, so the larger the δ the less confidence it reflects on the prior distribution, i.e., the less informative the prior.

Under our model, the prior variance must be smaller than the prior mean. This relationship holds because $0 < \delta < 1$. Since $\sigma^2 = \mu(1-\mu)\delta$ it follows that $\sigma^2 < \mu(1-\mu) < \mu$ because $0 < \mu < 1$. A small prior mean implies that the variability between the proportions for the domains will be small. However, in the situations that we are interested in, all domains have small true proportions, so a small variance across domains is reasonable.

We estimate δ through the following equation:

$$\left[1 - \frac{MSW}{\bar{p}(1-\bar{p})} - \delta\right] + \frac{C}{\delta} = 0 \tag{3.21}$$

where

$$MSW = \frac{n}{nm-m}\sum_{i=1}^{m} p_i(1-p_i).$$

The above equation can be solved in closed form, and it has two solutions, one which is negative and the other one which is:

$$\hat{\delta} = \frac{K + \sqrt{K^2 + 4C}}{2} \tag{3.22}$$

where $K = 1 - MSW/((\bar{p})(1 - \bar{p}))$.

When n is fixed and $m \to \infty$, as is typically assumed in small area estimation asymptotic problems due to the fact that typically the domain sample size n is much smaller than the number of domains m, $MSW \xrightarrow{p} \mu(1 - \mu)(1 - \delta)$ and thus $K \xrightarrow{p} \delta$. Note that $\hat{\delta}$ is consistent provided $C = o(1)$.

Based on extensive empirical evidence gathered through a large number of simulations, for small $C > 0$, $\hat{\delta}$ appears to have the property that it is always in the desired range (i.e., $0 \leq \hat{\delta} \leq 1$). A similar technique was used by Gabler *et al.* (2011) [21] in the estimation of intra-cluster correlation for the balanced one-way random effects model.

According to our simulation results, the values of C that give good coverages depend on the true parameters and on n. However, our simulation studies also suggest that in cases of interest it is possible to find a C that works well for a range of prior means and variances.

We could also have estimated δ more simply by:

$$1 - MSW/(\bar{p}(1 - \bar{p}))$$

This corresponds to $C = 0$. The problem with this method of estimation is that it frequently yields values for δ that are outside the admissible range, particularly when μ, n, and m are small. In fact, the frequency with which δ is out of the range increases when μ approaches 0 (or 1), as n decreases and as m decreases, as is illustrated by Figure 3.1, which is based on simulations under the assumed model.

One could arbitrarily set $\hat{\delta}$ to be a particular constant, such as .5, whenever an inadmissible value is obtained but this method results in undercoverage of the corresponding confidence intervals, according to our simulations.

Estimators for a and b are derived from the relations between a, b, μ and δ, as follows:

$$\hat{a} = \bar{p}\left(\frac{1}{\hat{\delta}} - 1\right) \tag{3.23}$$

and

$$\hat{b} = (1 - \bar{p})\left(\frac{1}{\hat{\delta}} - 1\right) \tag{3.24}$$

Care must be exercised to select a C that is appropriate for the cases of interest. A poor choice of C may result in coverages that are far below the nominal.

3.4 Simulation Results

For each replication, we generated data using the Binomial-Beta model, with $Y_i|P_i \sim \text{Bin}(n, P_i)$ and $P_i \sim \text{Beta}(a, b)$ for a variety of choices of m, n, μ, and σ^2, focusing primarily in small-area examples with small μ and σ^2. We computed cover-

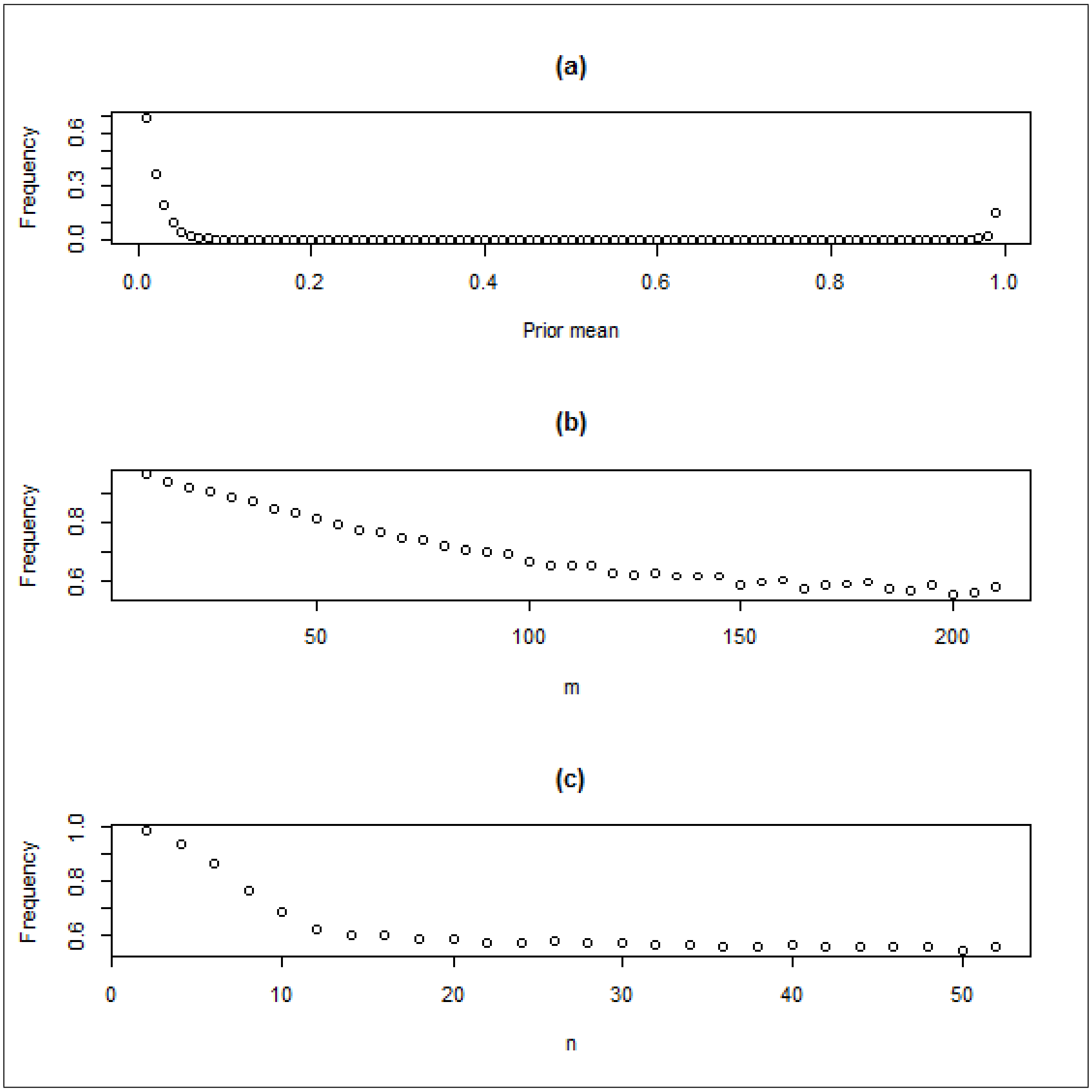

Figure 3.1: Frequency of $\hat{\delta}$ out of range for various parameters: Simulation results for $N = 10,000$ replications. (a) $m = 100$, $n = 10$, $\sigma^2 = .0099$, (b) $\mu = .01$, $\sigma^2 = .0099$, $n = 10$, (c) $\mu = .01$, $\sigma^2 = .0099$, $m = 100$. In each case C of (3.21) is set equal to zero.

ages (computed as the proportion of replications that capture the true domain proportion) and average lengths for the Binomial-Beta CI and for the Normal-Normal CI for one domain. The coverages and average lengths for all other domains should be similar under this model. Each simulation was performed for N replications, where N is typically 1000 or 10000 depending on the desired accuracy. All our CI's have a nominal 95% coverage and are two-sided.

3.4.1 Robust C for Fixed n and a Range of Small Prior Means and Variances

Table 3.1 displays the coverage frequency and average lengths for the Binomial-Beta CI and for the Normal-Normal CI for domain 1.

As was previously mentioned, the optimal value of C for estimating δ depends on the prior parameters, and in fact when C is inappropriately chosen the coverages of the Binomial-Beta CI can be quite poor, according to our simulations. However, Table 3.1 suggests it is possible to choose a C that provides appropriate coverage for the Binomial-Beta CI for a range of small prior means with small prior variances, and that the coverage of the Normal-Normal CI can be lacking in such situations.

In many cells in Table 3.1, the Normal-Normal CI falls below 90% coverage, although the nominal level is 95%. Smaller prior means were not considered for this size of m and n to eliminate the effect of having a high occurrence of zero counts in all domain samples, which can bias the results in favor of the Binomial-Beta CI in terms of coverage since the Binomial-Beta CI gives an interval of $[0, 1]$ in this case and the Normal-Normal CI gives a CI of $[0, 0]$.

This Table also suggests that the Normal-Normal coverage is oscillatory as one increases μ.

The Binomial-Beta CI performs fairly well within this range, although there is a trend for the coverage to decrease as μ increases. Larger values of μ than those in this Table are not the primary focus of this paper, but our simulations suggest that for larger values it will be necessary to increase C to maintain good coverages.

The prior variances in Table 3.1 are all small relative to the prior means. These were chosen such that for any given cell the inequality $\sigma^2 < \mu(1-\mu)$ holds. To check that the Binomial-Beta CI will still perform well when the prior variances are closer to the mean, we take as an example $\mu = .04$ and we take prior variances that are a bit larger than those in Table 3.1, starting with $\sigma^2 = \mu(1-\mu) - \epsilon$, where ϵ is a small number. The beta-binomial CI still performs reasonably well with the same choice of C, as shown in Table 3.2.

Table 3.3 illustrates that the appropriate value for C for a given range of hyperparameters depends on n. In this Table, we change n from 20 to 5, but all other parameters are the same as in Table 1, except that the first row of Table 1, with $\mu = .002$, was omitted to avoid situations with a high proportion of all zero counts. In this case $C = .0001$ yields poor results for the Binomial-Beta CI. However, Table 3.4 suggests that by changing the value of C to .001 one can obtain better results. Table 3.5 shows that $C = .001$ would have worked well for $n = 20$ as well.

Caution must be exercised when choosing the value of C as the interval can perform very poorly with a misspecified C.

3.4.2 Numerous Small Domains

Table 3.6 illustrates the impact on the coverages as m increases in small area problems, where n is small, and where $\mu = .01$. Very large values of m are included to study the behavior of the coverage as m increases and n is fixed. In this situation, the Normal-Normal CI can exhibit significant undercoverage even when m is large. The coverage of the Normal-Normal CI as m increases seems to be oscillatory when holding everything else fixed.

The Binomial-Beta CI performs well throughout Table 3.6 in terms of coverage, regardless of the value of n. $C = .001$ worked well in the previously discussed simulations with similar hyperparameters, and works well here as well.

This Table not only suggests oscillation for the Normal-Normal CI coverage as m increases and n is held fixed–there is also some evidence that as n increases with m held fixed the coverage oscillates as well. To investigate this further, Table 3.7 provides a wider range of n, and the results reinforce the oscillatory nature of the true coverage as n increases, although the oscillations are less pronounced than when there is a single binomial random variable, most likely because of the large number of domains and the fact that these intervals "borrow strength" from other domains when a estimating a domain proportion. Moreover, in Table 3.6 we can once again see that the appropriate value for C for the Binomial-Beta CI can depend on n as well as on the prior parameters. In this example, coverage for the Binomial-Beta CI seems to decrease as n increases, when everything else is fixed.

3.4.3 Unlucky n?

Brown *et al.* (2001) [9] show that the Wald-interval for building a confidence interval for the probability of success based on one binomial(r, p) observation can have poor coverage even when rp is fairly large. We investigate whether this phenomenon extends to our scenario. We select some of the "unlucky" pairings of (p, r) from Brown *et al.* and set our prior mean μ to equal their p and our domain sample size n to be equal to the corresponding value of r.

Table 3.8 shows the Normal-Normal CI can have undercoverage even when m and n are both large. The values $n = 592$ and $n = 954$ correspond to an example given Brown *et al* (2001) [9] to illustrate that the Wald interval can fail to yield the desired coverages even when rp and rq are large (a value in between these two was also included). The binomial probability of success in their example is $p = .005$, and we set our prior mean accordingly and our prior variance to be very small. Although the undercoverage is slight, it may be surprising due to the large n. Another interesting observation is that in this Table the Binomial-Beta CI's average lengths seem to be smaller than the Normal-Normal average lengths, despite the fact that the coverages of the latter are inferior. For this particular simulation we increased

N to 10,000 to increase the accuracy, since the undercoverage is of a lesser magnitude.

3.4.4 When the True Proportion is Not Close to the Extremes

The Normal-Normal CI does not show undercoverage in all cases, and seems to do well when the prior mean is not close to zero. Table 3.9 is one illustration of that.

3.5 Discussion

Through simulations, we have provided evidence that in the balanced case ($n_i = n$) and under our assumed model given by (3.1) and (3.2), the Normal-Normal CI for estimating the true proportion of a given domain can display significant undercoverage when the mean and variance of the prior distributions of the domains' true proportions are small. Moreover, in such cases the coverage appears to be oscillatory both as μ increases from 0, keeping everything else fixed, as n increases with all else held constant, and as m increases with all else held constant. We have shown that the undercoverage can be significant even when the number of sampled domains is very large. And we provided an example where there is undercoverage even when n and m are high. In situations where all domain true proportions are suspected to be small, the Normal-Normal CI should not be trusted.

We have provided a competing confidence interval that can display better coverage. The Binomial-Beta CI is appropriate when the statistician has some idea of the magnitude of the true proportions and their variability across domains, in particular when all domain proportions are known to be small as may occur when the proportion refers to a rare characteristic in the population, such as a rare disease. The interval depends on a constant C, but simulation studies suggest that it is possible to find a C that works well for a range of small prior means and variances, so that the statistician may be able to apply the Binomial-Beta CI to obtain better results than those yielded by the Normal-Normal CI provided they know the true proportions and their variability fall within certain ranges. The optimal C also depends on n and can be determined through simulations for a given problem. Misspecification of C can result in undercoverage. For a particular application where the Binomial-Beta model is deemed appropriate, the statistician can determine the value of C based on the expected range where the proportions fall by simulation.

The erratic behavior of the Normal-Normal CI may be due to its reliance on the normal approximation to the binomial. Practitioners must be aware that applying these normality-based empirical Bayes confidence intervals in small area problems where the true proportions and the domain sample sizes are small may result in undercoverage.

In this chapter, we tackled the simplest case of a much more complex problem– small area estimation for proportions in complex surveys where the true proportions are in the extremes and the domain sample sizes are small. The problem is even more challenging in the presence of additional complex survey features such as clustering, weights, varying domain sample sizes, and covariates. Constructing confidence intervals for proportions in complex surveys is very challenging both from a modeling and a mathematics perspective. Interesting areas for future research are to adapt the Binomial-Beta model to surveys with more complex designs, and to incorporate covariates into the model.

3.6 Appendix: Tables

Table 3.1: Two-sided coverages/average lengths for Binomial-Beta CI (top numbers in each cell) and Normal-Normal CI (bottom numbers in each cell) for one domain for $N = 1000$ replications, $m = 200$, $n = 20$, $c = .0001$, $\alpha = 0.05$.

μ \ σ^2	0.001	.00001	.00000001
0.002	0.998 / 0.014 0.939 / 0.008	0.997 / 0.013 0.926 / 0.009	0.993 / 0.013 0.927 / 0.0089
0.005	0.987 / 0.024 0.904 / 0.017	0.985 / 0.024 0.899 / 0.019	0.986 / 0.024 0.912 / 0.019
0.01	0.975 / 0.038 0.878 / 0.032	0.976 / 0.038 0.903 / 0.034	0.971 / 0.038 0.887 / 0.034
0.02	0.969 / 0.061 0.945 / 0.062	0.951 / 0.062 0.927 / 0.063	0.955 / 0.063 0.924 / 0.063
0.03	0.957 / 0.083 0.935 / 0.088	0.95 / 0.084 0.941 / 0.089	0.945 / 0.084 0.938 / 0.089
0.04	0.946 / 0.1 0.95 / 0.11	0.946 / 0.1 0.946 / 0.11	0.946 / 0.1 0.94 / 0.11
0.05	0.943 / 0.12 0.934 / 0.13	0.952 / 0.12 0.94 / 0.13	0.954 / 0.12 0.952 / 0.13

Table 3.2: Two-sided coverages/average lengths for Binomial-Beta CI (top numbers in each cell) and Normal-Normal CI (bottom numbers in each cell) for one domain for $N = 1000$ replications, $m = 200$, $n = 20$, $c = .0001$, $\alpha = 0.05$.

μ \ σ^2	0.0384	0.01	0.005
0.04	1 / 0.071 0.947 / 0.044	0.948 / 0.1 0.936 / 0.1	0.947 / 0.1 0.945 / 0.11

Table 3.3: Two-sided coverages/average lengths for Binomial-Beta CI (top numbers in each cell) and Normal-Normal CI (bottom numbers in each cell) for one domain for $N = 1000$ replications, $m = 200$, $n = 5$, $c = .0001$, $\alpha = 0.05$.

μ \ σ^2	0.001	.00001	.00000001
0.005	0.971 / 0.03 0.922 / 0.023	0.973 / 0.028 0.919 / 0.023	0.952 / 0.029 0.905 / 0.023
0.01	0.913 / 0.039 0.909 / 0.04	0.887 / 0.04 0.905 / 0.042	0.921 / 0.039 0.914 / 0.041
0.02	0.875 / 0.066 0.897 / 0.069	0.867 / 0.066 0.903 / 0.069	0.863 / 0.067 0.903 / 0.069
0.03	0.865 / 0.093 0.886 / 0.096	0.865 / 0.097 0.871 / 0.1	0.874 / 0.094 0.879 / 0.099
0.04	0.853 / 0.12 0.867 / 0.12	0.887 / 0.12 0.898 / 0.12	0.863 / 0.12 0.882 / 0.13
0.05	0.887 / 0.14 0.899 / 0.15	0.877 / 0.15 0.888 / 0.15	0.889 / 0.15 0.903 / 0.15

Table 3.4: Two-sided coverages/average lengths for Binomial-Beta CI (top numbers in each cell) and Normal-Normal CI (bottom numbers in each cell) for one domain for $N = 1000$ replications, $m = 200$, $n = 5$, $c = 0.001$, $\alpha = 0.05$.

μ \ σ^2	0.001	.00001	.00000001
0.005	0.996 / 0.039 0.943 / 0.023	0.988 / 0.038 0.939 / 0.024	0.991 / 0.041 0.93 / 0.025
0.01	0.993 / 0.058 0.918 / 0.04	0.991 / 0.058 0.903 / 0.041	0.983 / 0.057 0.913 / 0.041
0.02	0.987 / 0.089 0.892 / 0.069	0.977 / 0.09 0.91 / 0.072	0.983 / 0.09 0.896 / 0.072
0.03	0.969 / 0.12 0.895 / 0.098	0.972 / 0.12 0.884 / 0.098	0.964 / 0.12 0.886 / 0.099
0.04	0.966 / 0.14 0.883 / 0.13	0.968 / 0.14 0.894 / 0.13	0.968 / 0.14 0.888 / 0.12
0.05	0.963 / 0.17 0.896 / 0.15	0.963 / 0.17 0.901 / 0.15	0.961 / 0.17 0.905 / 0.15

Table 3.5: Two-sided coverages/average lengths for Binomial-Beta CI (top numbers in each cell) and Normal-Normal CI (bottom numbers in each cell) for one domain for $N = 1000$ replications, $m = 200$, $n = 20$, $c = 0.001$, $\alpha = 0.05$.

μ σ^2	0.001	.00001	.00000001
0.002	0.99 / 0.016 0.939 / 0.0082	0.985 / 0.017 0.919 / 0.0088	0.988 / 0.015 0.918 / 0.009
0.005	0.997 / 0.03 0.929 / 0.018	0.993 / 0.031 0.921 / 0.019	0.985 / 0.03 0.904 / 0.019
0.01	0.987 / 0.047 0.887 / 0.032	0.989 / 0.047 0.895 / 0.035	0.987 / 0.048 0.896 / 0.033
0.02	0.974 / 0.071 0.924 / 0.062	0.978 / 0.072 0.943 / 0.063	0.98 / 0.071 0.926 / 0.063
0.03	0.976 / 0.092 0.952 / 0.089	0.952 / 0.092 0.927 / 0.089	0.976 / 0.093 0.944 / 0.089
0.04	0.97 / 0.11 0.93 / 0.11	0.967 / 0.11 0.945 / 0.11	0.97 / 0.11 0.941 / 0.11
0.05	0.964 / 0.13 0.944 / 0.13	0.958 / 0.12 0.943 / 0.13	0.968 / 0.13 0.945 / 0.13

Table 3.6: Two-sided coverages/average lengths for Binomial-Beta CI (top numbers in each cell) and Normal-Normal CI (bottom numbers in each cell) for one domain for $N = 1000$ replications, $\mu = 0.01$, $\sigma^2 = .0001$, $c = 0.001$, $\alpha = 0.05$.

m n	3	5	10	20
100	0.976 / 0.1 0.901 / 0.055	0.979 / 0.061 0.915 / 0.048	0.993 / 0.053 0.931 / 0.043	0.979 / 0.048 0.918 / 0.036
500	0.992 / 0.059 0.882 / 0.037	0.994 / 0.059 0.857 / 0.036	0.991 / 0.055 0.85 / 0.033	0.994 / 0.047 0.923 / 0.034
1000	0.996 / 0.061 0.776 / 0.033	0.996 / 0.059 0.792 / 0.033	0.993 / 0.055 0.886 / 0.033	0.988 / 0.047 0.95 / 0.034
5000	0.998 / 0.062 0.755 / 0.029	0.997 / 0.06 0.875 / 0.033	0.995 / 0.054 0.939 / 0.035	0.996 / 0.048 0.951 / 0.034
8000	0.994 / 0.062 0.785 / 0.029	1 / 0.06 0.892 / 0.032	0.996 / 0.056 0.934 / 0.035	0.991 / 0.047 0.947 / 0.034
10000	0.994 / 0.062 0.793 / 0.029	0.998 / 0.06 0.909 / 0.033	0.997 / 0.054 0.954 / 0.035	0.986 / 0.046 0.93 / 0.035

Table 3.7: Two-sided coverages/average lengths for Binomial-Beta CI (top numbers in each cell) and Normal-Normal CI (bottom numbers in each cell) for one domain for $N = 1000$ replications, $\mu = 0.01$, $\sigma^2 = .0001$, $c = 0.001$, $\alpha = 0.05$.

n \ m	100	300	500
3	0.983 / 0.11 0.905 / 0.053	0.992 / 0.058 0.9 / 0.038	0.994 / 0.061 0.895 / 0.037
5	0.984 / 0.059 0.913 / 0.048	0.997 / 0.058 0.914 / 0.039	0.991 / 0.058 0.875 / 0.036
10	0.991 / 0.053 0.926 / 0.041	0.992 / 0.054 0.877 / 0.036	0.996 / 0.054 0.859 / 0.034
20	0.985 / 0.046 0.919 / 0.036	0.991 / 0.046 0.905 / 0.034	0.987 / 0.047 0.924 / 0.034
30	0.982 / 0.042 0.915 / 0.033	0.98 / 0.042 0.915 / 0.033	0.972 / 0.042 0.93 / 0.033
40	0.982 / 0.038 0.924 / 0.032	0.971 / 0.039 0.928 / 0.032	0.983 / 0.038 0.948 / 0.032
50	0.963 / 0.035 0.925 / 0.031	0.972 / 0.036 0.951 / 0.031	0.956 / 0.036 0.928 / 0.031
100	0.962 / 0.028 0.941 / 0.027	0.958 / 0.028 0.949 / 0.027	0.958 / 0.029 0.948 / 0.027
500	0.947 / 0.015 0.937 / 0.016	0.939 / 0.014 0.944 / 0.016	0.954 / 0.015 0.936 / 0.016

Table 3.8: Two-sided coverages/average lengths for Binomial-Beta CI (top numbers in each cell) and Normal-Normal CI (bottom numbers in each cell) for one domain for $N = 10000$ replications, $\mu = 0.005$, $\sigma^2 = .00001$, $c = .00001$, $\alpha = 0.05$.

m \ n	592	700	954
50	0.953 / 0.009 0.937 / 0.0098	0.951 / 0.0083 0.936 / 0.0092	0.952 / 0.0073 0.94 / 0.0081
100	0.953 / 0.009 0.936 / 0.0098	0.952 / 0.0084 0.936 / 0.0092	0.955 / 0.0073 0.938 / 0.0081
200	0.955 / 0.0089 0.939 / 0.0098	0.954 / 0.0084 0.941 / 0.0092	0.955 / 0.0073 0.94 / 0.0081

Table 3.9: Two-sided coverages/average lengths for Binomial-Beta CI (top numbers in each cell) and Normal-Normal CI (bottom numbers in each cell) for one domain for $N = 1000$ replications, $\mu = 0.4$, $\sigma^2 = 0.03$, $c = 0.035$, $\alpha = 0.05$.

n m	50	100	1000
3	0.963 / 0.65 0.944 / 0.71	0.961 / 0.65 0.932 / 0.69	0.967 / 0.66 0.939 / 0.67
5	0.955 / 0.56 0.967 / 0.65	0.969 / 0.56 0.976 / 0.64	0.952 / 0.56 0.962 / 0.63
10	0.965 / 0.44 0.978 / 0.52	0.952 / 0.44 0.966 / 0.52	0.961 / 0.44 0.975 / 0.51
20	0.951 / 0.33 0.972 / 0.4	0.952 / 0.33 0.974 / 0.4	0.95 / 0.33 0.974 / 0.39
30	0.953 / 0.28 0.972 / 0.33	0.952 / 0.27 0.973 / 0.33	0.951 / 0.28 0.969 / 0.33
40	0.964 / 0.24 0.981 / 0.29	0.947 / 0.24 0.968 / 0.29	0.953 / 0.24 0.965 / 0.29
50	0.933 / 0.21 0.967 / 0.26	0.951 / 0.21 0.968 / 0.26	0.943 / 0.22 0.97 / 0.26
100	0.947 / 0.16 0.966 / 0.19	0.945 / 0.15 0.972 / 0.19	0.945 / 0.15 0.973 / 0.19
1000	0.944 / 0.05 0.968 / 0.06	0.948 / 0.05 0.972 / 0.061	0.954 / 0.05 0.977 / 0.061

Bibliography

[1] Agresti, A. (2007) *An Introduction to Categorical Data Analysis, 2nd ed.*. New York: Wiley.

[2] Barndorff-Nielsen, O.E., and Cox, D. R. (1989). *Asymptotic Techniques for Use in Statistics.* New York: Chapman and Hall.

[3] Barvinek, E. Daler, I., Francu, J. (1991). Convergence of Sequences of Inverse Functions. *Archivum Mathematicum*, 27, 3-4, pp. 201-204.

[4] Begin, J. M., Hall, W. J., Huang, W. M. and Wellner, J. A. (1983). Information and asymptotic efficiency in parametric-nonparametric models. *Annals of Statistics.* 10, pp. 432-452.

[5] Bickel, P. J., and Docksum, K. L. (2001).*Mathematical Statistics: Basic Ideas and Selected Topics, Volume I.* New Jersey: Prentice Hall.

[6] Bickel, P., Klassen, C., Ya'acov, R, Wellner, J. (1993). *Efficient and Adaptive Estimation for Semiparametric Models.* New York: Springer.

[7] Billingsley, P. (1995). *Probability and Measure, 3rd ed.*. New York: Wiley.

[8] Brown, L.D., (1986). *Fundamentals of Statistical Exponential Families with Applications in Statistical Decision Theory.* California: IMS Lecture Notes Monograph Series.

[9] Brown, L. D., Cai, T. T, and DasGupta, A. (2001). Interval Estimation for a Binomial Proportion. *Statistical Science*,16, 2, 101-133.

[10] Brown, L. D., Cai, T. T, and DasGupta, A. (2002). Confidence Intervals for a Binomial Proportion and Asymptotic Expansions. *Annals of Statistics* 30, 1, 160-201.

[11] Carter, G. M. and Rolph, J. F. (1974). Empirical Bayes Methods Applied to Estimating Fire Alarm Probabilities. *Journal of the American Statistical Association*, 67, 348, pp. 880-885.

[12] Cheng, K. F., and Chu, C. K. (2004). Semiparametric Density Estimation under a Two-Sample Density Ratio Model. *Bernoulli*, 10, 4, pp. 583-604.

[13] DasGupta, Anirban. (2008). *Asymptotic Theory of Statistics and Probability.* New York: Springer.

[14] Diciccio, T. J., Hall, P. and Romano, J. P. (1989). Comparison of parametric and empirical likelihood functions. *Biometrika*, 76, pp. 465-476.

[15] Diciccio, T. J., Hall, P. and Romano, J. P. (1991). Empirical likelihood is Bartlett-correctable. *Annals of Statistics.* 19, pp. 1053-1061.

[16] Efron B. and Morris C. (1975). Data Analysis Using Stein's Estimator and its Generalizations. *Journal of the American Statistical Association*, 70, 350, pp. 311-319.

[17] Fay, R. and Herriot, R. (1979) Estimates of Income for Small Places: An application of James-Stein Procedures to Census Data. *Journal of the American Statistical Association.* 74, 366, 269-277.

[18] Fitzpatrick, P. M. (2006) *Advanced Calculus.* California: Thomson.

[19] Fokianos, K. (2004). Merging Information for Semi-Parametric Density Estimation. *Journal of the Royal Statistical Society* 66B, 4, pp.941-958.

[20] Fokianos, K., Kedem, B., Qin, J. Shore, D. A. (2001). A semiparametric approach to the one-way layout. *Technometrics*, 43(1), 56-65.

[21] Gabler, S. Ganningher, M., and Lahiri, P. (2011). A strictly Positive Estimator of Intra-Cluster Correlation for the One-Way Random Effects Model.Preprint.

[22] Gilbert, P. B., Lele, S. R., Vardi, Y. V. (1999) Maximum Likelihood Estimation in Semiparametric Bias Models with Application to AIDS Vaccine Trials. *Biometrika.* 86, 1, pp. 27-43.

[23] Gilbert, P. B., (2000) Large Sample Theory of Maximum Likelihood Estimates in Semiparametric Biased Sampling Models. *The Annals of Statistics.* 28,1, pp. 181-194.

[24] Ghosh, M, and Lahiri, P (1987). Robust Empirical Bayes Estimation of Means From Stratified Samples *Journal of the American Statistical Association*, 82, 400, pp. 1153-1162

[25] Jiang, J., and Lahiri, P. (2006). Mixed Model Prediction and Small Area Estimation. *Sociedad de Estadistica e Investigacion Operativa*,12, 1, pp. 1-96.

[26] Jiang, J., and Lahiri, P. (2001). Empirical Best Prediction for Small Area Inference with Binary Data. *Ann. Inst. Statist. Math.*,53, 2, pp. 217-243.

[27] Jones, M.C. (1991). Kernel Density Estimation For Length Biased Data. *Biometrika*, 78, pp. 511-519.

[28] Kagan, A. (2011). Semiparametric estimation for kernel families. *Preprint.*

[29] Ibrahimov, I. A., and Has'Minskii, R. Z. (1981). *Statistical Estimation, Asymptotic Theory.* New York: Springer.

[30] Kagan, A. and Rao, C. R. (2003) Some properties and applications of the efficient Fisher Score. *Journal of Statistical Planning and Inference.* 116. 342-352.

[31] Kagan, A. (1976). Fisher information contained in a finite-dimensional linear space, and a correctly posed version of the method of moments. Translated from *Problemi Peredachi Informatsii.* 12, 2, pp. 20-42.

[32] Kagan, A. (1985). An information property of the exponential families. *Theory of Probability Applications*, 30.

[33] Kay, R., and Little, S. (1987). Transformation of the explanatory variables in the logistic regression model for binary data. Biometrika, 74, pp. 495-501.

[34] Kedem, B. K., Kim, E., Voulgaraki, A., and Graubard, B. I. (2009). Two-dimensional semiparametric density ratio modeling of testicular germ cell data. *Statistics in Medicine*, 28, pp. 2147-2159.

[35] Keziou, A. and Leoni-Aubin, Samuela. (2008). On empirical likelihood for semiparametric two-sample density ratio models. *Journal of Statistical Planning and Inference*, 138, pp. 915-928.

[36] Koralov, L. B. and Sinai, Y. G. (2007). *Theory of Probability and Random Processes.* New York: Springer.

[37] Liu, Y. K, and Kott, P. S. Evaluating Alternative One-Sided Coverage Intervals for a Proportion. *Journal of Official Statistics.* 25, 4, pp. 569-588.

[38] Koshevnik, Y. A., and Levit, B. Y. (1976). On a nonparametric analogue of the information matrix. *Theory of Probability Applications*, 20, pp. 723-740.

[39] Korn, E., and Graubad, B. (1998). Confidence Intervals for Proportions with small Expected Number of Positive Counts Estimated from Survey Data. *Survey Methodology*, 24, pp. 1030-1039.

[40] Lehmann, E.L. (1999). *Elements of Large Sample Theory.* New York: Springer-Verlag.

[41] Lehmann, E. L. and Casella, G. (1993). *Theory of Point Estimation.* New York: Springer.

[42] Liu, Y. K., and Kott, P. S.(2009). Evaluating One-Sided Coverage Intervals for a Proportion. *Journal of Official Statistics*, 25, pp. 569-588.

[43] Levit, B. Y. (1974). On optimality of some statistical estimates. *Proceedings of the Prague Symposium on Asymptotic Statistics* (J. Hájek, ed.), 2, pp. 215-238. University of Karlove, Prague.

[44] Morris, C. N. (1982). Natural exponential families with quadratic variance functions. *Annals of Statistics.* 10, 1, pp. 65-80.

[45] Morris, C. N. (1983). Natural exponential families with quadratic variance functions: statistical theory. *Annals of Statistics*, 11, 2, pp. 515-329.

[46] Morris, C. N., and Lock, K. F. (2009). Unifying the named natural exponential families and their relatives. *American Statistician*, 63, 3, pp. 247-253.

[47] Mukhopadhyay, P. (2004). *An Introduction to Estimating Functions.* Harrow U.K.: Alpha Science

[48] Owen, A. B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, 18, pp. 90-120.

[49] Owen, A. B. (1990). Empirical likelihood confidence regions. *Annals of Statistics*, 19, pp. 90-120.

[50] Owen, A. B. (1991). Empirical likelihood for linear models. *Annals of Statistics*, 19, pp. 1725-1747.

[51] Owen, A. B. (2001). *Empirical Likelihood.* Boca Raton: Chapman and Hall/CRC.

[52] Pace, L., and Salvan, A. *Advanced Series on Statistical Science and Applied Probability, Volume 4: Principles of Statistical Inference from a Neo-Fisherian Perspective*. Singapore: World Scientific.

[53] Prasad, N. G. N., and Rao, J. N. K. (1990). The estimation of mean squared errors of small area estimators. *Journal of the American Statistical Association* 85, pp. 163-171.

[54] Pfanzagl, J. (1982). *Contributions to a General Asymptotic Statistical Theory.* New York: Springer.

[55] Prentice, R. L., and Pyke, R. (1979) Logistic Inference Models and Case-Control Studies. *Biometrika*, 66, 3, pp. 403-411.

[56] Qin, J. (1998). Inference for Case-Control and Semiparametric Two-Sample Density Ratio Models. *Biometrika*, 85, 3, pp. 619-630.

[57] Qin, J. (2000). Combining Parametric and Empirical Likelihoods. Biometrika, 87, 2, pp. 484-490.

[58] Qin J. and Lawless, J. (1994). Empirical Likelihood and General Estimating Equations. *Annals of Statistics*, 22, 1, pp. 300-325.

[59] Qin, J., Berwick, M., Ashbolt, and R., Dwyer, T., (2002). Quantifying the change of melanoma incidence by Breslow thickness. *Biometrics* 58,3, pp.665-670.

[60] Qin, J., Zhang, B. (1997). A goodness-of-fit test for logistic regression models based on case-control data. *Biometrika*, 84, 3, pp. 609-618.

[61] Rao, C. R. (1973). *Linear Statistical Inference and its Applications, 2nd ed.*. New York: Wiley.

[62] Rao, J. N. K. (2003). *Small Area Estimation.* New York: Wiley.

[63] Severini, T. A., and Wong, H. W. (1992). Profile likelihood and conditionally parametric models. *Annals of Statistics*, 20, 4, pp. 1768-1802.

[64] Shao, J. (2003). *Mathematical Statistics.* New York: Springer-Verlag.

[65] Shiryaev, A.N. (1996). *Probability, 2nd ed.* New York: Springer Verlag.

[66] Shorack, R. G., and Wellner, J.A. (1986). *Empirical Processes with Applications to Statistics*. New York: Wiley.

[67] Slud, E. V., and Vonta F. (2005). Efficient Semiparametric Estimators Via Modified Profile Likelihood. *Journal of Statistical Planning and Inference* 129. pp. 339-367.

[68] Stein (1956). Efficient nonparametric testing and estimation. *Proceedings of the Third Berkely Symposium on Mathematical Statistics and Probability*. University of Californiaa Press, Berkely.

[69] Storer, B.E., Wachholder, S. and Breslow, N. (1983). Maximum likelihood fitting of general risk models to stratified data. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 32, 2, pp 172-181.

[70] Van der Vaart, A. W. (2000). *Asymptotic Statistics*. Cambridge: Cambridge University Press.

[71] Wolter, K. (1985). *Introduction to Variance Estimation*. New York: Springer.

[72] Zhou, H., Weaver, M. A., Qin, J., Longnecker, M.P., Wang, M.C. (2002). A semiparametric empirical likelihood method for data from an outcome-dependent sampling scheme with continuous outcome. *Biometrics*, 58, 2, pp. 413-421.