Maximum Weighted Likelihood Estimation

by

Steven Xiaogang Wang

B.Sc., Beijing Polytechnic University, P.R. China, 1991.

M.S., University of California at Riverside, U.S.A., 1996.

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

Doctor of Philosophy

in

THE FACULTY OF GRADUATE STUDIES

Department of Statistics

We accept this thesis as conforming

to the required standard

THE UNIVERSITY OF BRITISH COLUMBIA

June 21, 2001

In presenting this thesis in partial fulfilment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the head of my department or by his or her representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Department of Statistics

The University of British Columbia
Vancouver, Canada

Date Aug 9th, 2001

DE-6 (2/88)

Maximum Weighted Likelihood Estimation

Steven X. Wang

Abstract

A maximum weighted likelihood method is proposed to combine all the relevant data from different sources to improve the quality of statistical inference especially when the sample sizes are moderate or small.

The linear *weighted likelihood estimator* (WLE), is studied in depth. The weak consistency, strong consistency and the asymptotic normality of the WLE are proved. The asymptotic properties of the WLE using adaptive weights are also established. A procedure for adaptively choosing the weights by using cross-validation is proposed in the thesis. The analytical forms of the "adaptive weights" are derived when the WLE is a linear combination of the MLE's. The weak consistency and asymptotic normality of the WLE with weights chosen by cross-validation criterion are established. The connection between WLE and theoretical information theory is discovered. The derivation of the *weighted likelihood* by using the maximum entropy principle is presented. The approximations of the distributions of the WLE by using saddlepoint approximation for small sample sizes are derived. The results of the application to the disease mapping are shown in the last chapter of this thesis.

Contents

List of Figures

List of Tables

Acknowledgments

I am most grateful to my supervisor, Dr. James V. Zidek, whose advice and support are a source of invaluable guidance, constant encouragement and great inspiration throughout my Ph.D. study at University of British Columbia. I am also very grateful to my co-supervisor, Dr. Constance van Eeden, who provides invaluable advice and rigorous training in technical writing.

I wish to thank my thesis committee members Dr. John Petkau and Dr. Paul Gustafson whose insights and helps are greatly appreciated. I also wish to thank Dr. Ruben Zamar for his suggestions.

Finally, I would like to thank my wife, Ying Luo, for her understanding and support. My other family members all have their shares in this thesis.

Chapter 1

Introduction

1.1 Introduction

Recently there has been increasing interest in combining information from diverse sources. Effective methods for combining information are needed in a variety of fields, including engineering, environmental sciences, geosciences and medicine. Cox (1981) gives an overview on some methods for the combination of data such as weighted means and pooling in the presence of over-dispersion. An excellent survey of more current techniques of combining information and some concrete examples can be found in the report *Combining Information* written by the Committee on Applied and Theoretical Statistics, U.S. National Research Council (1992).

This thesis will concentrate on the combination of information from separate sets of data. When two or more data sets derive from similar variables measured on different samples of subjects are available to answer a given question, a judgment must be made whether the samples and variables are sufficiently similar that the data sets may be directly merged or whether some other method of combining information that only partially merges them is more appropriate. For instance, data sets may be

time series data for ozone levels of each year. Or they may be hospital admissions for different geographical regions that are close to each other. The question is how to combine information from data sets collected under different conditions (and with differing degrees of precision and bias) to yield more reliable conclusions than those available from a single information source.

1.2 Local Likelihood and Related Methods

Local likelihood, introduced by Tibshirani and Hastie (1987), extends the idea of local fitting to likelihood-based regression models. Local regression may be viewed as a special case of the local likelihood procedure. Staniswalis (1989) defines her version of *local likelihood* in the context of non-parametric regression as follows:

$$W(\theta) = \sum_{i=1}^{n} W(\frac{x_0 - x_i}{b}) log f(y_i; \theta)$$

where x_i are fixed and b is a single unknown parameter. Recently, versions of local likelihood for estimation have been proposed and discussed. The general form of local likelihood was presented by Eguchi and Copas (1998). The basic idea is to infuse local adaptation into the likelihood by considering

$$L(t; x_1, x_2, ..., x_n) = \sum_{i=1}^{n} K(\frac{x_i - t}{h}) log f(x_i, \theta),$$

where $K = K(\frac{x_i - t}{h})$ is a kernel function with center t and bandwidth h. The local maximum likelihood estimate $\hat{\theta}_t$ of a parameter in a statistical model $f(x; \theta)$ is defined by maximizing the weighted version of the likelihood function which gives more weight to sample points near t. This does not give an unbiased estimating equation as it stands, and so the local likelihood approach introduces a correction factor to ensure consistent estimation. The resulting *local maximum likelihood estimator* (LMLE),

say $\hat{\theta}_{t,h}$, depends on the controllable variables t and h through the kernel function K, and, intuitively, it is natural to think that $\hat{\theta}_{t,h}$ gains more information about the data around t in the sample space. Detailed discussions of the local likelihood method can be found in Eguchi and Copas (1998). The LMLE might be related to the local M-estimator proposed by Hardle and Gasser (1984). The local M-estimator is defined as

$$\sum_{i=1}^{n} \psi(Y_i - \beta)\frac{1}{h}\int_{x_i-1}^{x_i} K(\frac{x-u}{h} = 0$$

where Y_i are observations obtained at point x_i.

The term *weighted likelihood* has been used in the statistics literature besides the local likelihood approach. Dickey and Lientz (1970) propose the use of what they called *weighted likelihood ratio* for tests of simple hypothesis. Dickey (1971) proposes the *weighted utility-likelihood* along the same line of argument. Assume a statistical model in which the observed data vector $\boldsymbol{D} \in E^n$ occurs with the probability mass or density function $\phi(D|\theta)$, depending continuously on an unknown parameter vector $\theta \in E^r$. Suppose that one suspects the unknown parameter θ of belonging to a given Borel set $H \subset E^r$. Let $\overline{H}$ denote a Borel measurable alternative such that $H \cap \overline{H} = \emptyset$ with $P(H) + P(\overline{H}) = 1$. The key feature of their proposal is to use *weighted likelihood ratio* defined as follows:

$$L = \frac{\Phi(D|H)}{\Phi(D|\overline{H})}$$

where $\Phi(D|H) = \int \phi(\boldsymbol{D}|\theta)dP(\theta|H)$ and $\Phi(D|\overline{H}) = \int \phi(\boldsymbol{D}|\theta)dP(\theta|\overline{H})$. The reason that they called it *weighted likelihood ratio* is because the quantity $\Phi(D|H)$ is the summary of the evidence in $\boldsymbol{D}$ for H. A modern name for their *weighted likelihood ratio* might be odds ratio.

Markatou, Basu and Lindsay (1997, 1998) propose a method based on the *weighted likelihood equation* in the context of robust estimation. Their approach can be de-

scribed as follows: Suppose that $\{X_1, X_2, ..., X_n\}$ is a random sample with distribution $f(x;\theta)$. The *weighted likelihood equation* is defined as

$$\sum_{i=1}^{n} w(x_i, \hat{F}) \frac{\partial}{\partial\theta} log f(x_i;\theta)$$

where $\hat{F}$ is the empirical cumulative distribution function. The weight function $w(X_i, \hat{F})$ is selected such that it has value close to 1 if there is no evidence of model violation at x from the empirical distribution function. The weight function will be very close to 0 or exactly 0 at X_i if the empirical cumulative distribution functions indicates lack of fit at or near X_i. Thus, the role of the weight function is to down-weight points in the sample that are inconsistent with the assumed model.

Hunsberger (1994) also uses the term "weighted likelihood" to arrive at kernel estimators for the parametric and non-parametric components of semi-parametric regression models. Consider the model with $X_i|(Y_i, T_i = t_i)$ having the distribution $f(X_i;\lambda_i)$ where $\lambda_i = y_i\beta_0 + g(t_i)$. Furthermore let f, the conditional density of $X|(Y,T)$, be arbitrary but known. Then $x\beta_0$ is the parametric portion, β_0 being the unknown parameter to be estimated that relates the covariate y to the response. Here g is the non-parametric portion of the model, the only assumption on g being that it is a smooth function of t. Assume $y_i = r(t_i) + \eta_i$ where r is a smooth funtion and the η_i are independent random error terms with $E(\eta_i) = 0$ and $E\eta_i^2 = \sigma^2$. Now λ_i can be rewritten using the model for the y's to obtain $\lambda_i = \eta_i\beta_0 + h(t_i)$, where $h(t_i) = r(t_i)\beta_0 + g(t_i)$ is the portion that depends on t. The main purpose is to estimate β_0 and $\theta_i = h(t_i), i = 1, 2, ..., n$ in the semi-parametric model by maximizing a *weighted likelihood* function

$$WL(\beta, \boldsymbol{\theta}) = \sum_i \sum_j w\left(\frac{t_i - t_j}{b}\right) log f(X_j;\beta,\theta_i)/n^2 b$$

with respect to β and $\boldsymbol{\theta}$, where $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_n)^t$. In the weighted likelihood function,

w is a kernel that assigns zero weights to the observations X_j that correspond to t_j outside a neighborhood of t_i.

Besides these "weighted likelihood" approaches, it should be noted that the term *weighted likelihood* has been used in other contexts as well. Newton and Raftery (1994) introduce what they called *weighted likelihood bootstrap* as a way to simulate approximately from a posterior distribution. The *weighted likelihood* function is defined as

$$\tilde{L}(\theta) := \prod_{i=1}^{n} f_i(x_i;\theta)^{w_{n,i}},$$

where the random weight vector $\boldsymbol{w} = (w_{n,1}, w_{n,2}, ..., w_{n,n})$ has some probability distribution determined by the statistician. The function $\tilde{L}$ is not a likelihood function in the usual sense. It is considered by Newton and Raftery to be good approximation to the posterior.

Rao (1991) introduces his definition of the *weighted maximum likelihood* to deal with irregularities of the observation times in the longitudinal studies of the growth rate. To be more specific, he defines the *weighted likelihood* as

$$L_n(\beta) = \prod_{i=1}^{k_n} f(x_{i,n}, t_{i,n}, \theta | \{x_{j,n}, t_{j,n} : 1 \leq j \leq i-1\})^{\lambda(t_{i,n}) - \lambda(t_{i-1,n})}$$

where $\lambda(.)$ is a known nondecreasing function on [a,b] and $t_{i,n}$ are observation times such that

$$a = t_{0,n} < t_{1,n} < ... < t_{k_n,n} = b.$$

The term *weighted likelihood* is also used by Warm (1987) in the context of item response theory. His proposal can be described as maximizing $w(\theta)L(\boldsymbol{x}|\theta)$ instead of the traditional likelihood function $L(\boldsymbol{x}|\theta)$. The *weight function*, $w(\theta)$ is carefully chosen such that the new estimator is unbiased to the order of n^{-1}.

1.3 Relevance Weighted Likelihood Method

Hu and Zidek (1997) give a very general method for using all relevant sample information in statistical inference. They base their theory on what they call *relevance-weighted likelihood* (REWL). Let $\mathbf{x} = (x_1, x_2, ..., x_n)$ denote a realization of the random vector $\mathbf{X} = (X_1, X_2, ..., X_n)$. Let f_i be the unknown probability density function of the X_i which are assumed to be independent. Inferential interest is on a probability density function f. At least in some qualitative sense, the f_i are thought to be "like" f. Consequently the x_i's are thought to be of value in our inferential analysis though the X_i are independently drawn from a population different from our study population. The *relevance-weighted likelihood* is defined as

$$\text{REWL}(\theta) = \prod_{i=1}^{n} f(x_i; \theta)^{\lambda_i}$$

where $\lambda_i, i = 1, 2, ..., n$, are weight functions.

REWL plays the same role in the "relevant-sample" case as the likelihood in the conventional ("exact-information") case. It can be seen that this method generalizes local likelihood. The asymptotic properties of the *Relevance Maximum Weighted Likelihood Estimator* can be found in Hu (1997). Hu, Rosenberger and Zidek (2000) extend the results for independent sequences in Hu (1997) to dependent sequences.

1.4 Weighted Likelihood Method

In this thesis, we are interested in a context which is different from that of all the methods described above. Suppose that we are interested in a single population. A parameter or parameter vector, θ_1, is of inferential interest. Information from other related populations, Population 2, Population 3 and so on, are available together with the direct information from Population 1. Let m denote the total number of popula-

tions whose distributions are thought to "resemble" Population 1. Let $n_1, n_2, ..., n_m$ denote the number of observations obtained from each population mentioned above. Let $\mathbf{X_1}, \mathbf{X_2}, ..., \mathbf{X_m}$ be random variables or vectors with marginal probability density functions $f_1(.;\theta_1), f_2(.;\theta_2), ..., f_m(.;\theta_m)$, where $\mathbf{X_i} = (X_{i1}, X_{i2}, ..., X_{in_i})^t$. The joint distribution of $(\mathbf{X_1}, \mathbf{X_2}, ..., \mathbf{X_m})$ is not assumed. We are interested in the probability density function $f_1(.;\theta_1) : \theta_1 \in \Theta$ of a study variable or vector of variables $\mathbf{X}, \theta_1$ being an unknown parameter or vector of parameters. At least in some qualitative sense, the $f_2(.;\theta_2), ..., f_m(.;\theta_m)$ are thought to be "similar to" $f_1(.;\theta_1)$.

For fixed $\mathbf{X} = \mathbf{x}$, the *weighted likelihood* (WL) is defined as:

$$\mathrm{WL}(\theta_1) = \prod_{i=1}^{m} f_1(\mathbf{x}_i;\theta_1)^{\lambda_i},$$

where $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, ..., \lambda_m)^t$ is the "weight vector" which must be specified by the analyst.

We say that $\tilde{\theta}_1$ is a maximum weighted likelihood estimator (WLE) for θ_1 if

$$\tilde{\theta}_1 = \arg\sup_{\theta_1 \in \Theta} \mathrm{WL}(\theta_1).$$

The uniqueness of the maximizer is not assumed.

In this thesis, we assume that $X_{i1}, X_{i2}, ..., X_{in_i}$, i=1,2,...,m, are independent and identically distributed random variables. The WL then becomes

$$\mathrm{WL}(\theta_1) = \prod_{i=1}^{m} \prod_{j=1}^{n_i} f_1(x_{ij};\theta_1)^{\lambda_i}.$$

Hu (1997) proposes a paradigm which abstracts that of non-parametric regression and function estimation. There information about θ_1 builds up because the number of populations grows with increasingly many in close proximity to that of θ_1. This is the paradigm commonly invoked in the context of non-parametric regression but it is not always the most natural one. In contrast we postulate a fixed number of

populations with an increasingly large number of samples from each population. Our paradigm may be more natural in situations like that in which James-Stein estimator obtained, where a specific set of populations is under study.

1.5 A Simple Example

The advantages of using WLE might be illustrated by the following example.

A coin is tossed twice. Let $\theta_1 = P(H)$ for this coin. Let $\hat{\theta}_1$ denote the MLE of θ_1. It then follows that $\hat{\theta}_1 = S_1/2$ where $S_1 = X_1 + X_2$ and the X's are independent Bernoulli random variables. If unknown to the statistician, $\theta = 1/2$, then

$$\begin{aligned} P\left(\hat{\theta}_1 = 0\right) &= P\left(X_1 = X_2 = 0\right) = 1/4; \\ P\left(\hat{\theta}_1 = 1/2\right) &= P\left(\{X_1 = 0; X_2 = 1\} \cap \{X_1 = 1; X_2 = 0\}\right) = 1/2; \\ P\left(\hat{\theta}_1 = 1\right) &= P\left(X_1 = X_2 = 1\right) = 1/4. \end{aligned}$$

The probability for the MLE to conclude that either the fair coin has no head or no tail is 50%. It is clear that the probability of making a nonsensical decision about the fair coin in this case is extremely high due to the small sample size.

Suppose that another coin, not necessarily a fair one, is tossed twice as well. The question here is whether we can use the result from the second experiment to derive a better estimate for θ_1. The answer is affirmative if we combine the information obtained from the two experiments.

Suppose for definiteness $\theta_2 = P(H) = 0.6$ for the second coin. Let $\hat{\theta}_2$ denote the MLE of θ_2. Thus, $\hat{\theta}_2 = S_2/2 = (Y_1 + Y_2)/2$ where Y_1 and Y_2 are independent Bernoulli

random variables. It then follows that

$$\begin{aligned} P\left(\hat{\theta}_2 = 0\right) &= P(Y_1 = Y_2 = 0) = 0.16; \\ P\left(\hat{\theta}_2 = 1/2\right) &= P\left(\{Y_1 = 0; Y_2 = 1\} \cap \{Y_1 = 1; Y_2 = 0\}\right) = 0.48; \\ P\left(\hat{\theta}_2 = 1\right) &= P\left(Y_1 = Y_2 = 1\right) = 0.36. \end{aligned}$$

Consider a new estimate which is a linear combination of $\hat{\theta}_1$ and $\hat{\theta}_2$:

$$\tilde{\theta}_1 = \lambda_1 \hat{\theta}_1 + \lambda_2 \hat{\theta}_2,$$

where λ_1 and λ_2 are relevant weights.

The optimum weights will be discussed in later chapters. Pretend that we do not know how to choose the best weights. We might just set each of the weights to be 1/2. It follows that

$$\begin{aligned} P\left(\tilde{\theta}_1 = 0\right) &= P\left(S_1 = 0; S_2 = 0\right) = 0.04; \\ P\left(\tilde{\theta}_1 = 1/4\right) &= P\left(\{S_1 = 1; S_2 = 0\} \cap \{S_1 = 0; S_2 = 1\}\right) = 0.20; \\ P\left(\tilde{\theta}_1 = 1/2\right) &= P\left(\{S_1 = 1; S_2 = 1\} \cap \{S_1 = 2; S_2 = 0\} \cap \{S_1 = 0; S_2 = 1\}\right) = 0.37; \\ P\left(\tilde{\theta}_1 = 3/4\right) &= P\left(S_1 = 1; S_2 = 2\} \cap \{S_1 = 2; S_2 = 1\}\right) = 0.30; \\ P\left(\tilde{\theta}_1 = 1\right) &= P\left(S_1 = 2; S_2 = 2\right) = 0.09. \end{aligned}$$

Note that the probability of making a nonsensical decision has been greatly reduced to 0.13 (0.04+0.09) through the introduction of the information obtained from the second coin. Furthermore, it can be verified that

$$MSE(\hat{\theta}_1) = 1/8; \qquad MSE(\tilde{\theta}_1) = 1.02 \times 1/16 \approx 1/16.$$

Thus,

$$MSE(\tilde{\theta}_1)/MSE(\hat{\theta}_1) \approx 0.5.$$

We see that the MSE of the WLE is only about 50% of that of the MLE.

Due to the small sample size of the first experiment, for arbitrary θ_1, the probability of making a nonsensical decision is $\theta_1^2 + (1-\theta_1)^2$ with a minimum value of 50%. By incorporating the relevant information in a very simple way, that probability is greatly reduced. In particular, if the second coin is indeed a fair coin and θ_1 is arbitrary, the probability of making a nonsensical decision is then reduced to $\frac{1}{4}[\theta_1^2 + (1-\theta_1)^2]$ which is less or equal to 12.5%.

We would like to consider the reduction of MSE by using the simple average of the two MLE's in this case. Let $\tilde{\theta}_1 = \frac{1}{2}\hat{\theta}_1 + \frac{1}{2}\hat{\theta}_2$. For arbitrary θ_1 and θ_2 with sample size of 2, we have

$$\begin{aligned} MSE(\hat{\theta}_1) &= Var(\hat{\theta}_1) = \frac{1}{2}\theta_1(1-\theta_1) \\ MSE(\tilde{\theta}_1) &= Var(\tilde{\theta}_1) + Bias(\tilde{\theta}_1)^2 \\ &= = \frac{1}{8}\theta_1(1-\theta_1) + \frac{1}{8}\theta_2(1-\theta_2) + \frac{1}{4}(\theta_1-\theta_2)^2. \end{aligned}$$

It follows that, for $\theta_1 \neq 0$ or 1,

$$\begin{aligned} \frac{MSE(\tilde{\theta}_1)}{MSE(\hat{\theta}_1)} &= \frac{\frac{1}{8}\theta_1(1-\theta_1) + \frac{1}{8}\theta_2(1-\theta_2) + \frac{1}{4}(\theta_1-\theta_2)^2}{\frac{1}{2}\theta_1(1-\theta_1)} \\ &= \frac{1}{4} + \frac{\frac{1}{8}\theta_2(1-\theta_2) + \frac{1}{4}(\theta_1-\theta_2)^2}{\frac{1}{2}\theta_1(1-\theta_1)} \end{aligned}$$

Assume that $\theta_1, \theta_2 \in [0.35, 0.65]$, it then follows that

$$\begin{aligned} 0.35 * 0.65 &\leq (1-\theta_1)\theta_1; \\ (\theta_1-\theta_2)^2 &\leq 0.09; \\ \theta_2(1-\theta_2) &\leq 0.25. \end{aligned}$$

We then have

$$\frac{MSE(\tilde{\theta}_1)}{MSE(\hat{\theta}_1)} \leq 0.72.$$

Therefore, for a wide range of values of θ_1 and θ_2, the simple average of $\hat{\theta}_1$ and $\hat{\theta}_2$ will produce at least 28% reduction in the MSE compared with the traditional MLE, $\hat{\theta}_1$. The maximum reduction is achieved if these two coins happen to be of the same type, *i.e.* $\theta_1 = \theta_2$. We remark that the upper bound on the bias in this case is 0.15. Notice that the weights are chosen to be 0.5. However they are not the optimum weights which minimize the MSE of a weighted average of $\hat{\theta}_1$ and $\hat{\theta}_2$. The optimum weights will be studied in later chapters.

1.6 The Scope of the Thesis

In Chapter 2 we will show that certain linear combinations of the MLE's derived from two possibly different normal populations achieves smaller MSE than that of the traditional one sample MLE for finite sample sizes. A criterion for assessing the relevance of two related samples is proposed to control the magnitude of possible bias introduced by the combination of data. Results for two normal populations are shown in this chapter.

The *weighted likelihood function* which uses all the relevant information is formally proposed in Chapter 3. Our weighted likelihood generalizes the REWL of Hu (1994). Results for exponential families are presented in this chapter. The advantages of using a linear WLE on restricted parameter spaces are demonstrated. A set of optimum weights for the linear WLE is proposed and the non-stochastic limits of the proposed optimum weights when the sample size of the first population goes to infinity are found.

Chapter 4 is concerned with the asymptotic properties of the WLE . The weak

consistency, strong consistency and asymptotic normality of the WLE are proved when the parameter space is a subset of $R^p, p \geq 1$. The asymptotic results proved here differ from those of Hu (1997) because a different asymptotic paradigm is used. Hu's paradigm abstracts that of non-parametric regression and function estimation. There information about θ_1 builds up because the number of populations grows with increasingly many in close proximity to that of θ_1. This is the paradigm commonly invoked in the context of non-parametric regression but it is not always the most natural one. In contrast we postulate a fixed number of populations with an increasingly large number samples from each. Asymptotically, the procedure can rely on just the data from the population of interest alone. The asymptotic properties of the WLE using adaptive weights, *i.e.* weights determined from the sample, are also established in this chapter. These results offer guidance on the difficult problem of specifying λ.

In Chapter 5 we address the of choosing the adaptive weights by using the *cross-validation* criterion. Stone (1974) introduces and studies in detail the cross-validatory choice and assessment of statistical predictions. The *K-group* estimators in Stone (1974) and Geisser (1975) are closely related to the linear WLE . Breiman and Friedman (1997) also demonstrate the benefit of using cross-validation to obtain the linear combination of predictions that achieve better estimation in the context of multivariate regression. Although there are many ways of dividing the entire sample into subsamples such as a random selection of the validation sample, we use the simplest *leave-one-out* approach in this chapter since the analytic forms of the optimum weights are tractable for the linear WLE . The weak consistency and asymptotic normality of the WLE based on cross-validated weights are established in this chapter.

We develop a theoretical foundation for the WLE in Chapter 6. Akaike (1985) reviewed the historical development of entropy and discussed the importance of the maximum entropy principle. Hu and Zidek (1997) discovered the connection between

relevance weighted likelihood and *maximum entropy principle* for the discrete case. We shall show that the *weighted likelihood function* can be derived from the *maximum entropy principle* for the continuous case.

In the context of *weighted likelihood estimation*, the *i.i.d.* assumption is no longer valid. Observations from different samples follow different distributions. The saddlepoint approximation technique in Daniels (1954) is then generalized for the non *i.i.d.* case to derive very accurate approximate distributions of the WLE for exponential families in Chapter 7. The saddlepoint approximation for estimating equations proposed in Daniels (1983) is also generalized to derive the approximate density of the WLE derived from estimating equations.

The last chapter of this thesis applies the WLE approach to disease mapping data. Weekly hospital admission data are analyzed. The data from a particular site and neighboring sites are combined to yield a more reliable estimate to the average weekly hospital admissions compared with the traditional MLE.

Chapter 2

Motivating Example: Normal Populations

Combining information from disparate sources is a fundamental activity in both scientific research and policy decision making. The process of learning, for example, is one of combining information: we are constantly called upon to update our beliefs in the light of new evidence, which may come in various forms. In some cases, the nature of the similarity among different populations is revealed through some geometrical structure in the parameter space, *e.g.* the means of several populations are all points on a circular helix. From the value of relevant variables it might then be possible to obtain a great deal of information about the parameter of primary inferential interest. Therefore, we should be able to construct a better estimate of the parameter of primary interest by combining data derived from different sources. How might one combine the data in a sensible way? To illustrate some of the fundamental characteristics of our research, it is useful to consider a simple example and examine the inferences that can be made.

2.1 A Motivating Example

In this subsection we consider the following simple example:

$$X_i = \alpha\, t_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma_1^2), \quad i = 1, ..., n \tag{2.1}$$

$$Y_i = \beta\, t_i + \epsilon_i^{'}; \quad \epsilon_i^{'} \sim N(0, \sigma_2^2), \quad i = 1, ..., n, \tag{2.2}$$

where the $\{t_i\}_{i=1}^n$ are fixed. The $\{\epsilon_i\}$'s are i.i.d.. So are the $\{\epsilon_i^{'}\}$'s. While $Cov(\epsilon_i, \epsilon_j^{'}) = 0$ if $i \neq j$; $Cov(\epsilon_i, \epsilon_i^{'}) = \rho\sigma_1\sigma_2$ for all i, and for the purpose of our demonstration we assume ρ, σ_1 and σ_2 are known although that would rarely be the case in practice.

Note that a bivariate normal distribution is not assumed in the above model. In fact, only the marginal distributions are specified; no joint distribution is assumed although we do assume the correlation structure in this case.

The parameters, α and β, are of primary interest. They are thought or expected to be not too different due to the "similarity" of the two experiments. The error terms are assumed to be *i.i.d.* within each sample. The joint distributions of the $(X_i, Y_i), i = 1, ..., n$ are not specified. Only the correlations between samples are assumed to be known. The objective is to get reasonably good estimates for the parameters without assuming a functional form of the joint distribution of (X_i, Y_i) which may be unknown to the investigator.

Assuming marginal normality, the marginal likelihoods for α and β are

$$L_1(x_1, x_2, ..., x_n; \alpha) \propto \prod_{i=1}^{n} exp(-\frac{(x_i - \alpha\, t_i)^2}{2\,\sigma_1^2}),$$

$$L_2(y_1, y_2, ..., y_n; \beta) \propto \prod_{i=1}^{n} exp(-\frac{(y_i - \beta\, t_i)^2}{2\,\sigma_2^2}).$$

Therefore, ignoring the constants,

$$\ln\ L_1(\alpha) = -\sum_{i=1}^{n} \frac{(x_i - \alpha\, t_i)^2}{2\,\sigma_1^2},$$

$$\ln\ L_2(\beta) = -\sum_{i=1}^{n} \frac{(y_i - \beta\, t_i)^2}{2\,\sigma_2^2}.$$

The MLE's based on the X_i and Y_i respectively for α and β are

$$\hat{\alpha} = \frac{\sum_{i=1}^{n} t_i\, x_i}{\sum_{i=1}^{n} t_i^2},$$

$$\hat{\beta} = \frac{\sum_{i=1}^{n} t_i\, y_i}{\sum_{i=1}^{n} t_i^2}.$$

2.2 Weighted Likelihood Estimation

If we know that $|\alpha - \beta| \leq C$, where C is a known constant according to past studies or expert opinions, this information ("direct information" or "prior information") might be used to yield a better estimate of the parameter. An extremely important aspect of the problem of combining information that we have just described is that we can incorporate the relevant information into the likelihood function by assigning possibly different weights to different samples. We next show how it is done.

The weighted likelihood (WL) for inference about α is defined as:

$$\mathrm{WL}(\alpha) = L_1(x_1, x_2, ..., x_n; \alpha)^{\lambda_1} L_1(y_1, y_2, ..., y_n; \alpha)^{\lambda_2}, \tag{2.3}$$

where λ_1 and λ_2 are weights selected according to the relevance of the likelihood to which they attached. The non-negativeness of the weights is not assume although the optimum weights are actually non-negative.

Note that $L_1((y_1, y_2, ..., y_n; \alpha)$ instead of $L_2((y_1, y_2, ..., y_n; \beta)$ is used to define $\mathrm{WL}(\alpha)$ since α is of our primary interest at this stage and the marginals distributions of the Y's are thought to resemble the marginal distributions of the X's. Note that the WL depends on the distributions of the X's. But it does not depend

on the distribution of Y's. Since X and Y are not independent, the weights of the likelihood functions are designed to reflect the dependence that is not expressed in the marginals. Notice that the joint distribution of the X's and Y's does not appear in the WL and no assumptions are made about it.

The maximum weighted likelihood estimator (WLE) is obtained by maximizing the weighted likelihood function for given weights λ_1 and λ_2.

From (2.3) we get

$$\ln \mathrm{WL}(\alpha) = \lambda_1 \ln L_1(x_1, x_2, ..., x_n; a) + \lambda_2 \ln L_1(y_1, y_2, ..., y_n; a),$$

$$\frac{\partial \ln \mathrm{WL}(\alpha)}{\partial \alpha} = \frac{\lambda_1}{2\sigma_1^2} \sum_{i=1}^{n} t_i \, (x_i - \alpha \, t_i) + \frac{\lambda_2}{2\sigma_1^2} \sum_{i=1}^{n} t_i \, (y_i - \alpha \, t_i).$$

So the WLE for α is

$$\tilde{\alpha} = \frac{\lambda_1}{\lambda_1 + \lambda_2} \hat{\alpha} + \frac{\lambda_2}{\lambda_1 + \lambda_2} \hat{\beta}.$$

Without loss of generality, we can write

$$\tilde{\alpha} = \lambda_1 \, \hat{\alpha} + \lambda_2 \, \hat{\beta},$$

where $\lambda_1 + \lambda_2 = 1$.

The WLE is a linear combination of the MLE for α and β, $\hat{\alpha}$ and $\hat{\beta}$ under the over-simplified model. The weights λ_1 and λ_2 reflect the importance of $\hat{\alpha}$ and $\hat{\beta}$. Intuitively, the inequality $\lambda_2 \leq \lambda_1$ should be satisfied because direct sample information from $\{X_i\}_{i=1}^n$ should be more reliable than relevant sample information from $\{Y_i\}_{i=1}^n$. Obviously, the WLE is the MLE obtained from the marginal distribution if the weight for the second marginal likelihood function is set to be zero. This may happen when evidence suggests that the seemingly relevant sample information is actually totally irrelevant. In that case, we would not want to include that information and thereby accrue too much bias into our estimator.

We call the estimator derived from the weighted likelihood function the WLE in line with the terminology found in the literature although the estimator described here differs from others proposed in those published papers. In particular, we work with the problem in which the number of samples are fixed in advance while the authors of the REWLE were interested in the problem where the number of populations goes to infinity. The distinction here should be clear.

2.3 A Criterion for Assessing Relevance

The weighted likelihood estimator is a linear combination of the MLE's derived from the likelihood function under the condition of marginal normality. We would like to find a good WLE under our model since there is no guarantee that any WLE will outperform the MLE in the sense of achieving a smaller MSE. Obviously, the WLE will be determined by the weights assigned to the likelihood function and the MLE's obtained from the marginals. The value of any estimator will depend on our choice of a loss function. The most commonly used criterion, the mean squared error (MSE), is selected as the measure of the performance of the WLE. The next proposition will give a lower bound for the ratio $\frac{\lambda_1}{\lambda_2}$ such that the WLE will outperform the MLE obtained from the marginal distribution. For simplicity, we make the assumption of equal variances, *i.e.* $\sigma_1^2 = \sigma_2^2$.

Proposition 2.1 *Let $\tilde{\alpha}$ be the WLE and $\hat{\alpha}$, the MLE of α. If $|\alpha - \beta| \leq C$, then*

$$\begin{aligned} E(\tilde{\alpha} - \alpha) &= \lambda_2 \, (\beta - \alpha) \\ |E(\tilde{\alpha} - \alpha)| &\leq \lambda_2 \, C \\ Var(\tilde{\alpha}) &\leq Var(\hat{\alpha}). \end{aligned}$$

If $\rho < 1$ *and* $\lambda_2 > 0$, *then*

$$\max_{|\alpha-\beta|\leq C} MSE(\tilde{\alpha}) < MSE(\hat{\alpha}) \ \ \textit{iff} \ \ \frac{\lambda_1}{\lambda_2} > \frac{C^2\, S_t^2}{2(1-\rho)\sigma^2}$$

where $S_t^2 = \sum_{i=1}^{n} t_i^2$.

If $\rho = 1$, *then* $\max_{|a-b|\leq C} MSE(\tilde{\alpha}) \geq MSE(\hat{\alpha})$ *with equality iff* $\lambda_1 = 1$ *and* $\lambda_2 = 0$.

Proof: It can be verified that

$$\begin{aligned} E(\hat{\alpha}) &= a, \\ E(\hat{\beta}) &= b, \\ E(\tilde{\alpha}) &= \lambda_1\, \alpha + \lambda_2\, \beta. \end{aligned}$$

It follows that

$$E(\tilde{\alpha} - \alpha) = \lambda_2\, (\beta - \alpha).$$

Thus $\tilde{\alpha}$ is not an unbiased estimator of α unless $\alpha = \beta$. However the absolute bias is bounded by $\lambda_2 C$ if $|\beta - \alpha| \leq C$.

It can also be checked that

$$\begin{aligned} Var(\hat{\alpha}) &= \frac{\sigma^2}{S_t^2}, \\ Var(\hat{\beta}) &= \frac{\sigma^2}{S_t^2}, \\ Cov(\hat{\alpha}, \hat{\beta}) &= \frac{Cov(\epsilon_1, \epsilon_1')}{S_t^2} = \frac{\rho\sigma^2}{S_t^2}, \end{aligned}$$

where $S_t^2 = \sum_{i=1}^{n} t_i^2$.

Furthermore, we have

$$\begin{aligned} Var(\tilde{\alpha}) &= \lambda_1^2\, Var(\hat{\alpha}) + \lambda_2^2\, Var(\hat{\beta}) + 2\lambda_1\, \lambda_2\, Cov(\hat{\alpha}, \hat{\beta}) \\ &= (\lambda_1^2 + \lambda_2^2 + 2\, \lambda_1\, \lambda_2\, \rho)\frac{\sigma^2}{S_t^2}. \end{aligned}$$

It can be seen that the MSE of WLE is a function of α and β. So is the bias function for $\tilde{\alpha}$. We would like to choose weights that are independent of α and β which are unknown.

Let us consider the MSE's of $\hat{\alpha}$ and $\tilde{\alpha}$ under the assumption $|\alpha - \beta| \leq C$,

$$\begin{aligned} E(\hat{\alpha} - \alpha)^2 &= Var(\hat{\alpha}) = \frac{\sigma^2}{S_t^2}, \\ E(\tilde{\alpha} - \alpha)^2 &= Var(\tilde{\alpha}) + (Bias)^2 \\ &\leq (\lambda_1^2 + \lambda_2^2 + 2\,\lambda_1\,\lambda_2\,\rho)\,\frac{\sigma^2}{S_t^2} + \lambda_2^2\,C^2. \end{aligned}$$

(i) If $\rho = 1$, we have $Var(\tilde{\alpha}) = Var(\hat{\alpha})$, as well as $max_{|\beta-\alpha|\leq C} MSE(\tilde{\alpha}) \geq MSE(\hat{\alpha}) = Var(\hat{\alpha})$. Equality is achieved only when λ_1 is set to 1 and λ_2 to 0.

(ii) If $\rho < 1$, it follows that

$$Var(\tilde{a}) < Var(\hat{a}).$$

Furthermore,

$$\max_{|\beta-\alpha|\leq C} E(\tilde{\alpha} - \alpha)^2 < E(\hat{\alpha} - \alpha)^2 \Leftrightarrow 2\,\lambda_1\,\lambda_2\,\rho\,\frac{\sigma^2}{S_t^2} + \lambda_2^2\,C^2 < 2\,\lambda_1\,\lambda_2\,\frac{\sigma^2}{S_t^2}.$$

We then have

$$\max_{|\beta-\alpha|\leq C} E(\tilde{\alpha} - \alpha)^2 < E(\hat{\alpha} - \alpha)^2 \;\Leftrightarrow\; \frac{\lambda_1}{\lambda_2} > \frac{C^2\,S_t^2}{2(1-\rho)\sigma^2}.$$

In conclusion, we have $max_{|\beta-\alpha|\leq C} MSE(\tilde{\alpha}) \leq MSE(\hat{\alpha})$ iff $\frac{\lambda_1}{\lambda_2} \geq \frac{C^2\,S_t^2}{2(1-\rho)\sigma^2}$. $\diamond$

The optimum weights which achieve the minimum of the MSE in this case will be positive as will be shown in the next section. Thus, we will not consider the case when $\lambda_2 < 0$.

We remark that the reduction in the MSE is independent of the assumption that σ_1^2 equals σ_2^2. In general, it can be verified that , if $\rho < \sigma_1/\sigma_2$, then

$$\max_{|\beta-\alpha|\leq C} MSE(\tilde{\alpha}) < MSE(\hat{\alpha}) \;\text{ iff }\; \frac{\lambda_1}{\lambda_2} > \frac{C^2 S_t^2 + (\sigma_2^2 - \sigma_1^2)}{2(\sigma_1^2 - \rho\sigma_1\sigma_2)}. \tag{2.4}$$

If $\sigma_1^2 = \sigma_2^2 = \sigma^2$, then the equation (2.4) is reduced to the inequality in the previous Proposition, *i.e.*,

$$\max_{|\alpha-\beta|\leq C} MSE(\tilde{\alpha}) < MSE(\hat{\alpha}) \text{ iff } \frac{\lambda_1}{\lambda_2} > \frac{C^2\, S_t^2}{2(1-\rho)\sigma^2}$$

From equation (2.4), observe that we will have $\lambda_1 = 1$ and $\lambda_2 = 0$ for a number of special cases:

1) $C \to \infty$ and $\rho < 1$. Here we have little information about the upper bound for the distance between the two parameters, or we do not have any prior information at all.

2) $S_t^2 \to \infty$ and $\rho < 1$. We already have enough data to make inference and the additional relevant data has little value in terms of estimating the parameter of primary interest.

3) $\sigma_1^2 \to 0$ and $\rho < 1$. This means that the precision of the data from the first sample is good enough already.

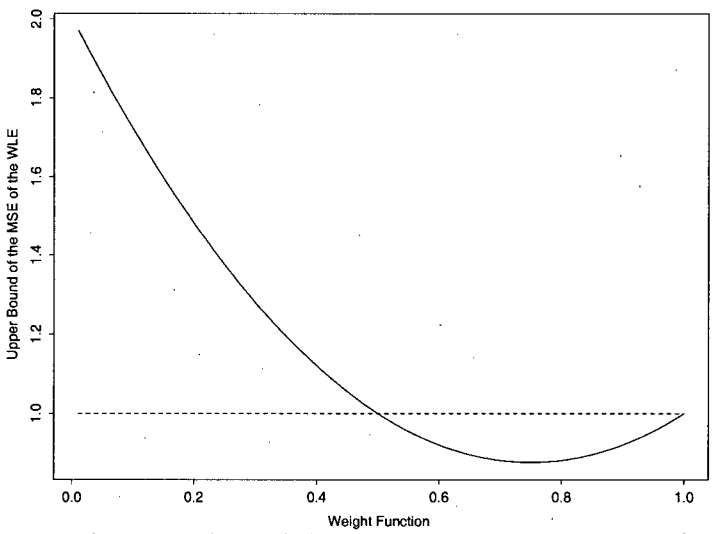

Figure 2.1: *A special case: the solid line is the* $max_{|\beta-\alpha|\leq C} MSE(WLE)$, *and the broken line represents the* $MSE(MLE)$. *The X-axis represents the value of* λ_1. *The parameters have been set to be the following:* $\frac{S_t^2}{\sigma^2} = 1, \rho = 0.5$ *and* $C = 1$.

A special case is discussed here to gain some insight into the situation. The upper bound of mean square error for WLE as a function of λ_1 is shown in the Figure 1. It can be seen that the upper bound of the MSE for the WLE will be less than that of the MLE if λ_1 lies in a certain range. From the Proposition 2.1, the cut-off point should be 0.5 which is exactly the case shown in the graph. In fact, the minimum of the $max_{|\beta-\alpha|\leq C}$ MSE can be obtained by using the Proposition 2.2 which will be established in the next section.

2.4 The Optimum WLE

In the previous section we found the threshold for λ_1 and λ_2 in order to make WLE outperform the MLE in the sense of obtaining a MSE and variance which are smaller than that of MLE uniformly on the set $\{(\alpha, \beta) : |\beta - \alpha| \leq C\}$. Because there are numerous cases under which λ_1 and λ_2 satisfy the threshold, we might want to find the optimum WLE in the sense of minimizing the maximum of MSE and variance.

Proposition 2.2 *The optimum WLE under mean squared error is*

$$\tilde{\alpha} = \frac{1+K}{2+K}\,\hat{\alpha} + \frac{1}{2+K}\,\hat{\beta},$$

$$\tilde{\beta} = \frac{1+K}{2+K}\,\hat{\beta} + \frac{1}{2+K}\,\hat{\alpha},$$

where $K = \frac{C^2\, S_t^2}{(1-\rho)\sigma^2}$.

Proof: $max_{|\beta-\alpha|\leq C} MSE(\tilde{\alpha})$ is a function of λ_1 and λ_2 for fixed σ, S_t^2 and ρ. Let $G(\lambda_1, \lambda_2) = max_{|\beta-\alpha|\leq C} MSE(\tilde{\alpha})$. The saddle point (stationary point) of this function can be obtained as the solution of the following equations,

$$\begin{cases} \frac{\partial G(\lambda_1,\lambda_2)}{\partial \lambda_1} &= 0, \\ \frac{\partial G(\lambda_1,\lambda_2)}{\partial \lambda_2} &= 0, \\ \lambda_1 + \lambda_2 &= 1. \end{cases}$$

In fact, the above equation has a unique solution. It can be verified that the minimum is achieved at that point by checking the second derivative. The solution to these equations are $\lambda_1^* = \frac{1+K}{2+K}$ and $\lambda_2^* = \frac{1}{2+K}$, where $K = \frac{C^2 S_t^2}{(1-\rho)\sigma^2}$. Note that the weights λ_1^* and λ_2^* satisfy the condition of $\frac{\lambda_1}{\lambda_2} > \frac{K}{2}$ because $\frac{\lambda_1^*}{\lambda_2^*} = K + 1 > \frac{K}{2}$. This implies that with the optimum weights given above the WLE will achieve a smaller MSE than the MLE according to Proposition 2.1

Thus, the optimum weights for estimating α are

$$\begin{cases} \lambda_1^* &= \frac{1+K}{2+K}, \\ \lambda_2^* &= \frac{1}{2+K}. \end{cases}$$

The optimum WLE's for α and β are

$$\begin{cases} \tilde{\alpha} &= \frac{1+K}{2+K}\,\hat{\alpha} + \frac{1}{2+K}\,\hat{\beta}, \\ \tilde{\beta} &= \frac{1+K}{2+K}\,\hat{\beta} + \frac{1}{2+K}\,\hat{\alpha}, \end{cases}$$

where $K = \frac{C^2 S_t^2}{(1-\rho)\sigma^2}$. Note that the optimum WLE for β is obtained by the argument of symmetry. $\diamond$

Notice that $\lambda_2^* < \frac{1}{2}$ for all $K > 0$. This implies that, for the purpose of estimating α, $\hat{\alpha}$ should never get a weight which is less than that of $\hat{b}$ if we want to obtain the optimum WLE. This is consistent with our intuition since relevant sample information (the $\{Y_i\}$'s) should never get larger weight than direct sample information (the $\{X_i\}$'s) when the sample sizes and variances are all equal. It should be noted that the WLE under a marginal normality assumption is a linear combination of the MLE's. Furthermore, the optimum WLE is the best linear combination of MLE's in the sense of achieving the smallest upper bound for MSE compared to other linear combination of MLE's. The optimization procedure is used to obtain the *minimax* solution in the sense that basically we are minimizing the upper bound of the MSE function over the set $\{(\alpha, \beta) : |\alpha - \beta| \leq C\}$.

2.5 Results for Bivariate Normal Populations

This subsection contains some results for bivariate normal distributions. The following results should be considered as immediate corollaries of Propositions 2.1 and 2.2 established in the previous section.

Corollary 2.1 *Let*

$$\begin{pmatrix} X_i \\ Y_i \end{pmatrix} \sim N\left(\begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix} \begin{bmatrix} \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 \end{bmatrix} \right), i = 1, .., n.$$

If $|\mu_X - \mu_Y| \leq C$ *and* $\rho < 1$, *then the optimum WLE's for estimating the marginal means are*

$$\begin{cases} \tilde{\mu}_X &= \frac{1+M}{2+M}\,\bar{X} + \frac{1}{2+M}\,\bar{Y}, \\ \tilde{\mu}_Y &= \frac{1+M}{2+M}\,\bar{Y} + \frac{1}{2+M}\,\bar{X}, \end{cases}$$

where $M = \frac{C^2\, n}{2\,(1-\rho)\,\sigma^2}$.

Furthermore,

$$max_{|\mu_X - \mu_Y| \leq C}\, E(\tilde{\mu}_X - \mu_X)^2 < E(\bar{X} - \mu_X)^2;$$

$$max_{|\mu_X - \mu_Y| \leq C}\, E(\tilde{\mu}_Y - \mu_Y)^2 < E(\bar{Y} - \mu_Y)^2;$$

$$Var(\tilde{\mu}_X) < Var(\bar{X}) = \frac{\sigma^2}{n};$$

$$Var(\tilde{\mu}_Y) < Var(\bar{Y}) = \frac{\sigma^2}{n};$$

$$Cov(\tilde{\mu}_X, \tilde{\mu}_Y) > Cov(\bar{X}, \bar{Y}).$$

Proof: Let $t_i = 1$ in Proposition 2.1. Then $S_t^2 = \sum_{i=1}^{n} 1 = n$. By letting $\hat{a} = \bar{X}$ and $\hat{b} = \bar{Y}$, we can apply Proposition 2.2 to get the optimum WLE . Therefore the optimum WLE will have smaller MSE and variance than the MLE, $\bar{X}$, obtained from the marginal normal distribution.

Observe that $\tilde{\mu}_X + \tilde{\mu}_Y = \bar{X} + \bar{Y}$. If we take Var on both sides, we then have the following

$$Var(\tilde{\mu}_X) + Var(\tilde{\mu}_Y) + 2\,Cov(\tilde{\mu}_X, \tilde{\mu}_Y) = Var(\bar{X}) + Var(\bar{Y}) + 2\,Cov(\bar{X}, \bar{Y}).$$

It follows that

$$Cov(\tilde{\mu}_X, \tilde{\mu}_Y) > Cov(\bar{X}, \bar{Y}). \quad \diamond$$

From Corollary 2.1, we draw the conclusion that the optimum WLE out-performs the MLE when $|\mu_X - \mu_Y| \leq C$ is true.

Corollary 2.2 *Under the conditions of Corollary 2.1,*

$$\tilde{\mu}_X - \bar{X} \xrightarrow{P} 0 \quad as \quad n \to \infty.$$

Proof: Let $M = \frac{C^2\, n}{2\,(1-\rho)\,\sigma^2}$. Thus, $M \to \infty$ as $n \to \infty$. Therefore, $\frac{1}{2+M} \to 0$ as $n \to \infty$. By Markov's inequality, we have, for all $\epsilon > 0$,

$$P\left(|\tilde{\mu}_X - \bar{X}| > \epsilon\right) = P\left(\frac{1}{2+M}|\bar{X} - \bar{Y}| > \epsilon\right) \leq \frac{\frac{1}{2+M} E\left(|\bar{X} - \bar{Y}|\right)}{\epsilon} \longrightarrow 0 \quad n \to \infty. \diamond$$

Corollary 2.3 *The optimum WLE is strongly consistent under the conditions of Corollary 2.1.*

Proof: Consider

$$\begin{aligned} P\left(|\tilde{\mu}_X - \mu_X| > \epsilon\right) &= P\left(|(\tilde{\mu}_X - \bar{X}) - (\mu_X - \bar{X})| > \epsilon\right) \\ &\leq P\left(|\frac{\bar{X} - \bar{Y}}{2+M}| > \frac{\epsilon}{2}\right) + P\left(|\mu_X - \bar{X}| > \frac{\epsilon}{2}\right), \end{aligned}$$

where $M \doteq \frac{C^2\, n}{2\,(1-\rho)\,\sigma^2}$ as before. But

$$P\left(|\tilde{\mu}_X - \bar{X}| > \frac{\epsilon}{2}\right) = P\left(|\frac{\bar{X} - \bar{Y}}{2+M}| > \frac{\epsilon}{2}\right) \leq \left(\frac{1}{2+M}\right)^2 E\left(\bar{X} - \bar{Y}\right)^2 \frac{4}{\epsilon^2} < \frac{M_1^2}{n^2},$$

where M_1 is a constant. Furthermore

$$P\left(|\bar{X} - \mu_X| > \frac{\epsilon}{2}\right) \leq \frac{16\, E\{\sum_{i=1}^{n}(X_i - \mu_X)\}^4}{(n\,\epsilon)^4} \leq \frac{M_2'\, n^2}{(n\,\epsilon)^4} = \frac{M_2}{n^2},$$

where M_2 is a constant. Thus,

$$\sum_{n=1}^{\infty} P\left(|\tilde{\mu}_X - \mu_X| > \epsilon\right) < \infty.$$

By the Borel-Cantelli Lemma (Chung 1968), we conclude that $\tilde{\mu}_X \xrightarrow{a.s} \mu_X$. $\diamond$

It follows from Corollary 2.1 that the optimum WLE is preferable to the MLE. However the relevant information will play a decreasingly important role as the direct sample size increases.

Motivated by our previous results, it seems that the linear combination of MLE's is a convenient way to combine information if the marginal distributions are assumed to be normal. A more general result is given as follows when the normality condition is relaxed.

Theorem 2.1 *Let $X_1, X_2, ..., X_n$ be i.i.d. random variables with $E(X_i) = \alpha$ and $Var(X_i) = \sigma_X^2 < \infty$, and let $Y_1, Y_2, ..., Y_n$ be i.i.d. random variables with $E(Y_i) = \beta$ and $Var(Y_i) < \infty$. Let $\hat{\alpha}$ and $\hat{\beta}$ denote some estimates of α and β respectively. Assume $E(\hat{\alpha}) = \alpha$, $E(\hat{\beta}) = \beta$, and $|\beta - \alpha| \leq C$. Let $\tilde{\alpha} = \lambda_1\ \hat{\alpha} + \lambda_2\ \hat{\beta}$, where $\lambda_1 = \frac{1+K}{2+K}$, $\lambda_2 = \frac{1}{2+K}$ and $K = \frac{C^2}{(1-\rho)\,Var(\hat{\alpha})}$. Also suppose $\rho = cor(\hat{\alpha}, \hat{\beta}) < 1$ while $Var(\hat{\alpha})$ is assumed known.*

If $Var(\hat{\beta}) \leq Var(\hat{\alpha})$, then $|E(\tilde{\alpha} - \alpha)| \leq \lambda_2\, C$ and

$$Var(\tilde{\alpha}) < Var(\hat{\alpha}),$$

$$max_{|\alpha-\beta|\leq C}\ E(\tilde{\alpha} - \alpha)^2 < E(\hat{\alpha} - \alpha)^2, while$$

$$\tilde{\alpha} - \alpha \xrightarrow{P} 0\ \ if\ \ Var(\hat{\alpha}) \to\ 0\ \ as\ \ \ n \to \infty.$$

Proof: We consider the variance of weighted estimate $\tilde{\alpha}$ of α:

$$\begin{aligned}
Var(\tilde{\alpha}) &= Var(\lambda_1\ \hat{\alpha} + \lambda_2\ \hat{\beta}) \\
&= \lambda_1^2\, Var(\hat{\alpha}) \ +\ \lambda_2^2\, Var(\hat{\beta}) \ +\ 2\ \lambda_1\ \lambda_2\, Cov(\hat{\alpha}, \hat{\beta}) \\
&\leq \lambda_1^2\, Var(\hat{\alpha}) \ +\ \lambda_2^2\, Var(\hat{\alpha}) \ +\ 2\ \lambda_1\ \lambda_2\, \rho\, \sqrt{Var(\hat{\alpha})}\, \sqrt{Var(\hat{\beta})} \\
&< \lambda_1^2\, Var(\hat{\alpha}) \ +\ \lambda_2^2\, Var(\hat{\alpha}) \ +\ 2\ \lambda_1\ \lambda_2\, Var(\hat{a}) \\
&= Var(\hat{\alpha}).
\end{aligned}$$

Thus, we have $Var(\tilde{\alpha}) < Var(\hat{\alpha})$. Furthermore,

$$\begin{aligned} MSE_{|\beta-\alpha|\leq C}(\tilde{\alpha}) &= Var(\tilde{\alpha}) + Bias^2(\tilde{\alpha}) \\ &= \lambda_1^2\, Var(\hat{\alpha}) \;+\; \lambda_2^2\, Var(\hat{\beta}) \;+\; 2\,\lambda_1\,\lambda_2\, Cov(\hat{\alpha},\hat{\alpha}) \;+\; \lambda_2^2\,(\beta-\alpha)^2 \\ &\leq (\lambda_1^2 \;+\; \lambda_2^2 \;+\; 2\,\lambda_1\,\lambda_2\,\rho) Var(\hat{a}) \;+\; \lambda_2^2\, C^2. \end{aligned}$$

Now we are in a position to apply Proposition 2.2 to get the weights for the best linear combination of $\hat{\alpha}$ and $\hat{\beta}$. Consequently we have the best linear combination, $\tilde{\alpha} = \lambda_1\,\hat{\alpha} + \lambda_2\,\hat{\beta}$, where the weights $\lambda_1 = \frac{1+K}{2+K}$, $\lambda_2 = \frac{1}{2+K}$ and $K = \frac{C^2}{(1-\rho)\,Var(\hat{a})}$. Also, we have $max_{|\beta-\alpha|\leq C}\; E(\tilde{\alpha}-\alpha)^2 < E(\hat{\alpha}-\alpha)^2$. The last statement of this theorem can be established in the same way as shown in Corollary 2.2. $\diamond$

The above theorem implies that any information contained in $\hat{\beta}$ can be used to improve the quality of inference provided it is as reliable as $\hat{a}$. Normality assumptions for the marginals are no longer necessary to obtain a better estimate of a under the conditions of Theorem 2.1. Moreover, we can correct for our "working assumption" of independent samples through likelihood weights.

Chapter 3

Maximum Weighted Likelihood Estimation

3.1 Weighted Likelihood Estimation

Suppose that we are interested in a single population. A parameter or parameter vector, θ_1, is of inferential interest. Information from other related populations, Population 2, Population 3 and so on, are available together with the direct information from Population 1. Let m denote the total number of populations whose distributions are thought to "resemble" Population 1. Let $n_1, n_2, ..., n_m$ denote the number of observations obtained from each population mentioned above. Let $\mathbf{X_1}, \mathbf{X_2}, ..., \mathbf{X_m}$ be random variables or vectors with marginal probability density functions $f_1(.;\theta_1), f_2(.;\theta_2), ..., f_m(.;\theta_m)$, where $\mathbf{X_i} = (X_{i1}, X_{i2}, ..., X_{in_i})^t$. We assume that $X_{i1}, X_{i2}, ..., X_{in_i}$ are $i.i.d$ random variables in this thesis. The joint distribution of $(X_{1j}, X_{2j}, ..., X_{mj})$ is not assumed. We are interested in the probability density function $f_1(.;\theta_1) : \theta_1 \in \Theta$ of a study variable or vector of variables X, θ_1 being an unknown parameter or vector of parameters. At least in some qualitative sense, the $f_2(.;\theta_2), ..., f_m(.;\theta_m)$ are thought to be "similar to" $f_1(.;\theta_1)$.

For fixed $\mathbf{X} = \mathbf{x}$, the weighted likelihood (WL) is defined as:

$$\mathrm{WL}(\theta_1) = \prod_{i=1}^{m} \prod_{j=1}^{n_i} f_1(x_{ij}; \theta_1)^{\lambda_i}, \tag{3.1}$$

where $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, ..., \lambda_m)^t$ is the "weight vector". This must be specified by the analyst. It follows that

$$log\ \mathrm{WL}(\theta_1) = \sum_{i=1}^{m} \sum_{j=1}^{n_i} \lambda_i\ log\ f_1(x_{ij}; \theta_1).$$

We say that $\tilde{\theta}_1$ is a maximum weighted likelihood estimator (WLE) for θ_1 if

$$\tilde{\theta}_1 = arg \sup_{\theta_1 \in \Theta} \mathrm{WL}(\theta_1).$$

In many cases the WLE may be obtained by solving the *estimating equation:*

$$(\partial/\partial\theta_1)\ log\ \mathrm{WL}(\theta_1) = 0.$$

Note that the uniqueness of the WLE is not assumed.

Throughout this thesis, θ_1 is the parameter of primary inferential interest. We will use $\tilde{\theta}_1$ and θ_1^0 to denote the WLE and the true value of θ_1 in the sequel.

3.2 Results for One-Parameter Exponential Families

Exponential family models play a central role in classical statistical theory for independent observations. Many models used in statistical practice are from exponential families and they are analytically tractable.

Assume that $X_{11}, X_{12}, ..., X_{1n_i}$ are independent random variables which follow the same distribution from the exponential family with one parameter, θ_1, that is,

$$f_1(x; \theta_1) = exp\left(A_1(\theta_1)S(x) + B_1(\theta_1) + C_1(x)\right).$$

The likelihood function for n_1 observations which are *i.i.d* from the above distribution family can be written as

$$L_1(X_{11}, X_{12}, ..., X_{1n_1}; \theta_1) = exp\left(A_1(\theta_1) \sum_{i=1}^{n_1} S(x_{1i}) + n_1 \, B_1(\theta_1) + \sum_{i=1}^{n_1} C_1(x_{1i}) \right).$$

It follows that

$$\ln L_1(X_{11}, X_{12}, ..., X_{1n_1}; \theta_1) = A_1(\theta_1) \sum_{j=1}^{n_1} S(x_{1j}) + n_1 \, B_1(\theta_1) + Constant.$$

It then follows that

$$\begin{aligned} \frac{\partial \ln L_1(X_{11}, X_{12}, ..., X_{1n_1}; \theta_1)}{\partial \theta_1} &= A_1'(\theta_1) \sum_{i=1}^{n_1} S(x_{1i}) + n_1 B_1'(\theta_1) \\ &= n_1 \left(A_1'(\theta_1) T(\underline{\mathbf{x}}^1) + B_1'(\theta_1) \right) \end{aligned}$$

where $T(\underline{\mathbf{x}}^1) = \frac{1}{n_1} \sum_{j=1}^{n} S(x_{1j})$.

It is known (Lehmann 1983, p 123) that for the exponential family, the necessary and sufficient condition for an unbiased estimator to achieve the Cramer-Rao lower bound is that there exists $D(\theta_1)$ such that

$$\frac{\partial \ln \, L_1(X_{11}, X_{12}, ..., X_{1n_1}; \theta_1)}{\partial \theta_1} = n_1 D_1(\theta_1) \left(T(\underline{\boldsymbol{x}}^1) - \theta_1 \right), \quad \forall \theta_1.$$

Theorem 3.1 *Assume that, for any given i,* $X_{ij} \overset{i.i.d.}{\sim} f(x; \theta_i), j = 1, 2, ..., n_i$. *The WLE of* θ_1 *is a linear combination of the MLE's obtained from the marginals if*

$$\frac{\partial ln \, L_1(x_{i1}, x_{i2}, ..., x_{in_i}; \theta_1)}{\partial \theta_1} = n_i D_1(\theta_1) \left(T(\underline{\boldsymbol{x}}^i) - \theta_1 \right),$$

i.e., the WLE of θ_1 *will be the linear combination of the MLE's if the Cramer-Rao lower bound can be achieved by the MLE's derived from the marginals.*

Proof: Under the condition

$$\frac{\partial \ln \, L_1(x_{i1}, x_{i2}, ..., x_{in_1}; \theta_1)}{\partial \theta_1} = n_i D_1(\theta_1) \left(T(\underline{\boldsymbol{x}}^i) - \theta_1 \right),$$

it can be seen that $T(\underline{\boldsymbol{x}}^1)$ is the traditional MLE for θ_1 which achieves the Cramer-Rao lower bound and is unbiased as well. Then we have

$$\frac{\partial \ln \text{ WL}}{\partial \theta_1} = \sum_{i=1}^{m} \lambda_i n_i \, D_1(\theta_1) \left(T(\underline{\boldsymbol{x}}^i) - \theta_1 \right).$$

Thus the WLE is given by

$$\tilde{\theta}_1 = \sum_{i=1}^{m} t_i \lambda_i \, T(\underline{\boldsymbol{x}}^i)$$

where $t_i = n_i / \sum_{i=1}^{m} \lambda_i n_i$.

Therefore the WLE of θ_1 is a linear combination of the MLE's obtained from the marginals. This completes the proof. $\diamond$

Thus, for normal distributions, Bernoulli, exponential and Poisson distributions the WLE is a linear combination of the MLE's obtained from the marginal distributions.

Theorem 3.2 *For distributions of the exponential family form, suppose the MLE of* θ_1 *has the form of* $g(T(\underline{\boldsymbol{x}}^1))$ *where* $T(\underline{\boldsymbol{x}}^1)$ *is the sufficient statistic. Then the WLE of* θ_1 *takes the form* $g\left(\sum_{i=1}^{m} t_i \lambda_i \, T(\underline{\boldsymbol{x}}^i) \right)$, *where* $t_i = n_i / \sum_{i=1}^{m} \lambda_i n_i$.

Proof: As seen above

$$\frac{\partial \ln L_1(x_{11}, x_{12}, ..., x_{1n_1}; \theta_1)}{\partial \theta_1} = n_1 \left(A_1'(\theta_1) T(\underline{\boldsymbol{x}}^1) + B_1'(\theta_1) \right).$$

Consequently, the WLE satisfies

$$\sum_{i=1}^{m} \lambda_i n_i \left(A_1'(\theta_1) T(\underline{\boldsymbol{x}}^i) + B_1'(\theta_1) \right) = 0$$

which implies that

$$A_1'(\theta_1) \left(\sum_{i=1}^{m} t_i \lambda_i \, T(\underline{\boldsymbol{x}}^i) \right) + B_1'(\theta_1) = 0$$

where $t_i = n_i / \sum_{i=1}^{m} \lambda_i n_i$.

Therefore the WLE of θ_1 takes the form

$$\tilde{\theta}_1 = g\left(\sum_{i=1}^{m} t_i \lambda_i \, T(\underline{\boldsymbol{x}}^i)\right)$$

where $t_i = n_i / \sum_{i=1}^{m} \lambda_i n_i$. This completes the proof. $\diamond$.

Therefore the WLE has the same functional form as the MLE obtained from the marginals if we confine our attention to the one-parameter exponential family. The only modification made by the WLE is to use the linear combination of sufficient statistics from the two samples instead of using the sufficient statistic from a single sample. The advantage of doing this is that it may be a better estimator in terms of variance and MSE.

3.3 WLE On Restricted Parameter Spaces

The estimation of parameters in restricted parameter spaces is an old and difficult problem. An overview of the history and some recent developments can be found in van Eeden (1996). van Eeden and Zidek (1998) consider the problem of combining sample information in estimating ordered normal means. van Eeden and Zidek (2001) also consider estimating one of two normal means when their difference is bounded. We are concerned with combining the sample information when a number of populations in question are known to be related.

Often, prior to gathering of the current data in a given investigation, relevant information in some form is available from past studies or expert opinions. All statisticians must make use of such information in their analysis although they might differ in the degree of combining information. A linear combination of the estimates from each population is straightforward to use. In this section, we assume that the WLE is a linear combination of the individual MLE. We remark that the results of this sec-

tion hold for the general case where the new estimator is a linear combination of the estimator derived from each sample. We need the following Lemma on optimization to prove the major theorem of this section. We emphasize that we do not require the λ_i's to be non-negative.

Lemma 3.1 *Let* $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, ..., \lambda_m)^t$ *and* $\sum_{i=1}^{n} \lambda_i = 1$. *Let* A *be a symmetric invertible* $m \times m$ *matrix. The weight function,* $\boldsymbol{\lambda}$, *which minimizes the objective function,* $\boldsymbol{\lambda}^t A \boldsymbol{\lambda}$, *is given by the following formula:*

$$\boldsymbol{\lambda}^* = \frac{A^{-1}\mathbf{1}}{\mathbf{1}^t A^{-1}\mathbf{1}},$$

provided $\mathbf{1}^t A\ \mathbf{1} > 0$.

Proof: Using the Lagrange method for maximizing an objective function under constraints, we only need to maximize

$$G = \boldsymbol{\lambda}^t A\ \boldsymbol{\lambda} - c(\mathbf{1}^t \boldsymbol{\lambda} - 1),$$

subject to $\sum_{i=1}^{m} \lambda_i = 1$.

Differentiating the function G gives

$$\frac{\partial G}{\partial \boldsymbol{\lambda}} = 2A\boldsymbol{\lambda} - c\mathbf{1}.$$

Setting $\frac{\partial G}{\partial \boldsymbol{\lambda}} = 0$, it follows that

$$\boldsymbol{\lambda} = \frac{c}{2} A^{-1}\mathbf{1}.$$

$$\text{Now } 1 = \mathbf{1}^t\ \boldsymbol{\lambda} = \frac{c}{2}\mathbf{1}^t A^{-1}\mathbf{1}.$$

$$\text{So } c = \frac{2}{\mathbf{1}^t A^{-1}\mathbf{1}};$$

Therefore,

$$\boldsymbol{\lambda}^* = \frac{A^{-1}\mathbf{1}}{\mathbf{1}^t A^{-1}\mathbf{1}}.$$

Since $\boldsymbol{\lambda}^t A\ \boldsymbol{\lambda}$ is a quadratic function of λ_i, therefore, $\boldsymbol{\lambda}^{'} A\boldsymbol{\lambda}$ has its global minimum at the stationary point. This completes the proof.$\diamond$

As before, we would assume that the parameters are not too far apart. The following theorem will show that some benefit could be gained if we take advantage of such information, namely that, sufficient precision may be gained at the expense of bias so as to reduce the MSE. However maximizing that gain may entail negative weights.

Theorem 3.3 *Assume that* $X_{ij}, i = 1, 2, ..., m,\ \ j = 1, 2, ..., n_i$ *are random variables with* $E(X_{ij}) = \theta_i$, *and* m *is the the number of related data sources and* n_i *is the sample size for each source. The marginal distributions are assumed to be known. The joint distribution across those related sources is not assumed. Instead, the covariance structure of the joint distribution is assumed to be known. Furthermore,* $\theta_1, \theta_2, ..., \theta_m$ *are all finite and* $|\theta_i - \theta_1| \leq C_i$ *where* C_i *is a known constant. Let* $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_m)^t$ *where* $\hat{\theta}_i = \hat{\theta}_i(x_{i1}, x_{i2}, ..., x_{in_i})$ *is the MLE for the parameter* θ_i *derived the distribution of the* X_{ij}. *Suppose* $E(\hat{\theta}_i) = \theta_i$ *and* $V = cov(\hat{\boldsymbol{\theta}})$ *are known and* $(V + BB^t)^{-1}$ *is invertible.*

Then the minimax linear WLE for θ_1 *is:*

$$\tilde{\theta_1}^* = \sum_{i=1}^{m} \lambda_i^*\ \hat{\theta}_i,$$

where:

$$\boldsymbol{\lambda}^* = (\lambda_1^*, \lambda_2^*, ..., \lambda_m^*)^t = \frac{(V + BB^t)^{-1}\mathbf{1}}{\mathbf{1}^t(V + BB^t)^{-1}\mathbf{1}};$$

$$V = cov(\hat{\boldsymbol{\theta}})_{m\times m}; \qquad B = (0, C_2, C_3, ..., C_m)^t.$$

Proof: We are seeking the best linear combination of these MLE's derived from the marginal distributions. As before, let us consider the MSE of the WLE.

Writing $\tilde{\theta}_1 = \boldsymbol{\lambda}^t \hat{\boldsymbol{\theta}} = \sum_{i=1}^{m} \lambda_i \hat{\theta}_i$, we can calculate

$$\begin{aligned}
MSE(\tilde{\theta}_1) &= E[\sum_i \lambda_i (\hat{\theta}_i - \theta_1)]^2 \quad (since \sum_{i=1}^{m} \lambda_i = 1) \\
&= E[\sum_{i=1}^{m}\sum_{j=1}^{m} \lambda_i \lambda_j (\hat{\theta}_i - \theta_1)(\hat{\theta}_j - \theta_1)] \\
&= \sum_{i=1}^{m}\sum_{j=1}^{m} \lambda_i \lambda_j E[(\hat{\theta}_i - \theta_1)(\hat{\theta}_j - \theta_1)] \\
&= \sum_{i=1}^{m}\sum_{j=1}^{m} \lambda_i \lambda_j E[(\hat{\theta}_i - \theta_i + \theta_i - \theta_1)(\hat{\theta}_j - \theta_j + \theta_j - \theta_1)] \\
&= \sum_{i=1}^{m}\sum_{j=1}^{m} \lambda_i \lambda_j \Big(cov(\hat{\theta}_i, \hat{\theta}_j) + (\theta_i - \theta_1)(\theta_j - \theta_1)\Big) \\
&= \sum_{i=1}^{m}\sum_{j=1}^{m} \lambda_i \lambda_j \, cov(\hat{\theta}_i, \hat{\theta}_j) + \sum_{i=1}^{m}\sum_{j=1}^{m} \lambda_i \lambda_j (\theta_i - \theta_1)(\theta_j - \theta_1) \\
&\leq \boldsymbol{\lambda}^t cov(\hat{\boldsymbol{\theta}})\boldsymbol{\lambda} + \boldsymbol{\lambda}^t BB^t \boldsymbol{\lambda} \quad (by\ the\ assumptions) \\
&= \boldsymbol{\lambda}^t (V + BB^t)\boldsymbol{\lambda}.
\end{aligned}$$

Let $A = V + BB^t$. By applying Lemma 3.1, we conclude that the optimal linear WLE is:

$$\tilde{\theta}_1^* = \sum_{i=1}^{m} \lambda_i^* \theta_i,$$

where $\boldsymbol{\lambda}^* = (\lambda_1^*, \lambda_2^*, ..., \lambda_m^*)^t = \frac{(V+BB')^{-1}\mathbf{1}}{\mathbf{1}^t(V+BB')^{-1}\mathbf{1}}$. $\diamond$

It can be seen that the *max* MSE function is a quadratic form in $\lambda_1, \lambda_2, ...\lambda_m$ and involves only the first and the second moments of the marginal distributions and the joint distributions. The whole procedure consists of two stages. The first step is to work out the functional form of the MSE. The second step is to work out the optimum weights to construct the optimum WLE. Note that the optimum weights are functions of the matrices V and B.

Let's consider some special cases:

(i) $B = \underline{0}$ and $V = \sigma^2 I$.

This implies that all the data can be pooled together in such a way that equal weights should be attached to the MLE derived from each marginal distribution since now we have $\theta_1 = \theta_2 = ... = \theta_m$. That is, $\lambda_i^* = \frac{1}{m}$ for all i for $i = 1, 2, ..., m$. This is consistent with our intuition for the *i.i.d.* normal case.

(ii) $B = \underline{0}$ and $V = diag(\sigma_1^2, \sigma_2^2, ..., \sigma_n^2)$.

The optimum WLE is then given by:

$$\tilde{\theta}_1 = \frac{\frac{1}{\sigma_1^2}\hat{\theta}_1 + \frac{1}{\sigma_2^2}\hat{\theta}_2 + ... + \frac{1}{\sigma_n^2}\hat{\theta}_m}{\sum_{k=1}^{m}\frac{1}{\sigma_k^2}}.$$

Note that most current pooling methods estimate the population parameter θ is a weighted average of the estimates obtained from each data set:

$$\hat{\theta} = \frac{\sum_{i=1}^{m}\frac{1}{\sigma_i^2}\hat{\theta}_i}{\sum_{i=1}^{m}\frac{1}{\sigma_i^2}},$$

provided that the $Var(\hat{\theta}_i) = \sigma_i^2$ are known. The optimal weights under the assumption are proportional to the precision in the *ith* data set, that is, to the inverse of the variance. This is a simple example of a *weighted least squares estimator.* Therefore the optimum WLE coincides with the weighted least squares estimator under the assumption $B = 0$ and $V = diag(\sigma_1^2, \sigma_2^2, ..., \sigma_n^2)$.

Furthermore, let us apply Theorem 3.3 to our motivating example since it is also a special case of the theorem. To be consistent with the notation used in Section 2, we still use a and b to denote the parameters of interest. We need to work out the matrix V and BB^t before applying the theorem since both are assumed to be known. As before, we have

$$Var(\hat{a}) = \frac{\sigma^2}{S_t^2}, \quad Var(\hat{b}) = \frac{\sigma^2}{S_t^2} \quad \text{and} \quad cov(\hat{a}, \hat{b}) = \frac{\rho\sigma^2}{S_t^2}.$$

To write those in a matrix form, we have

$$V = cov((\hat{a}, \hat{b})^t) = \frac{1}{S_t^2} \begin{pmatrix} \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 \end{pmatrix}.$$

As for the matrix BB^t, we have

$$BB^t = \begin{pmatrix} 0 & 0 \\ 0 & C^2 \end{pmatrix}.$$

Adding the above two matrices together gives

$$V + BB^t = \frac{1}{S_t^2} \begin{pmatrix} \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 + C^2\, S_t^2 \end{pmatrix}.$$

Therefore,

$$\begin{aligned} (V + BB^t)^{-1} &= \left[\frac{1}{S_t^2} \begin{pmatrix} \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 + C^2\, S_t^2 \end{pmatrix} \right]^{-1} \\ &= \frac{S_t^2}{|V + BB^t|} \begin{pmatrix} \sigma^2 + C^2 S_t^2 & -\rho\sigma^2 \\ -\rho\sigma^2 & \sigma^2 \end{pmatrix} \\ &= \frac{S_t^2}{\sigma^4 - \rho^2\sigma^4 + C^2 S_t^2 \sigma^2} \begin{pmatrix} \sigma^2 + C^2 S_t^2 & -\rho\sigma^2 \\ -\rho\sigma^2 & \sigma^2 \end{pmatrix} \end{aligned}$$

Thus, we have

$$\begin{aligned} (V + BB^t)^{-1}\mathbf{1} &= \frac{S_t^2}{\sigma^4 - \rho^2\sigma^4 + C^2 S_t^2 \sigma^2} \begin{pmatrix} \sigma^2 + C^2 S_t^2 & -\rho\sigma^2 \\ -\rho\sigma^2 & \sigma^2 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} \\ &= \frac{S_t^2}{\sigma^4 - \rho^2\sigma^4 + C^2 S_t^2 \sigma^2} \begin{pmatrix} (1-\rho)\sigma^2 + C^2 S_t^2 \\ (1-\rho)\sigma^2 \end{pmatrix}. \end{aligned}$$

Furthermore,

$$\mathbf{1^t}(V + BB^t)^{-1}\mathbf{1} = \frac{S_t^2}{\sigma^4 - \rho^2\sigma^4 + C^2 S_t^2 \sigma^2} \left(2(1-\rho)\sigma^2 + C^2 S_t^2\right).$$

Therefore the optimum weight function, $\boldsymbol{\lambda}^* = (\lambda_1^*, \lambda_2^*)^t$, is

$$\boldsymbol{\lambda}^* = \begin{pmatrix} \lambda_1^* \\ \lambda_2^* \end{pmatrix} = \begin{pmatrix} \frac{(1-\rho)\sigma^2+C^2S_t^2}{2(1-\rho)\sigma^2+C^2S_t^2} \\ \frac{(1-\rho)\sigma^2+C^2S_t^2}{2(1-\rho)\sigma^2+C^2S_t^2} \end{pmatrix} = \begin{pmatrix} \frac{1+K}{2+K} \\ \frac{1}{2+K} \end{pmatrix},$$

where $K = \frac{C^2S_t^2}{(1-\rho)\sigma^2}$.

These are exactly the weights we derived in Proposition 2.2. We therefore have shown that Theorem 3.3 can be used to derive the results we found earlier in Proposition 2.2.

3.4 Limits of Optimum Weights

The weights are of fundamental importance. Therefore it seems worthwhile to investigate the behavior of the optimum weights as the sample sizes get large. Since the weights are all fixed numbers for fixed sample sizes, thus the limits are non-stochastic as well. For simplicity, we assume that $C_2 = C_3 = ... = C_m = C$, where C is a constant.

Theorem 3.4 *Let* $B = (0, C, C, ..., C)^t, C > 0$. *Let* $V = cov(\hat{\boldsymbol{\theta}})$, *where* $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_m)^t$. *Re-write* $V = \sigma_1^2 W$, *where the* σ_i^2 *is a function of* n_i *and* W *is a function of* $n_1, n_2, ...n_m$. *Assume* $\lim_{n_1 \to \infty} \sigma_1^2(n_1) = 0$ *and* $\lim_{n_1 \to \infty} W(n_1, n_2, ..., n_m)^{-1} = W_0^{-1}$.

Then

$$\lim_{n_1 \to \infty} \boldsymbol{\lambda}^* = \lim_{n_1 \to \infty} \frac{(V + BB^t)^{-1}\mathbf{1}}{\mathbf{1}^t(V + BB^t)^{-1}\mathbf{1}} = \frac{S_{m\times m}\mathbf{1}}{\mathbf{1}^t S_{m\times m}\mathbf{1}}$$

where $S_{m\times m} = W_0^{-1} - \frac{W_0^{-1}BB^tW_0^{-1}}{B^tW_0^{-1}B}$.

Proof: From the matrix identity (Rao 1965),

$$(A + UV^t)^{-1} = A^{-1} - \frac{A^{-1}U\, V^tA^{-1}}{1 + V^tA^{-1}U},$$

it follows that

$$\begin{aligned}\sigma_1^2(\sigma_1^2 W + BB^t)^{-1} &= \sigma_1^2\left(\frac{1}{\sigma_1^2}W^{-1} - \frac{\frac{1}{\sigma_1^2}W^{-1}BB^t\frac{1}{\sigma_1^2}W^{-1}}{1+\frac{1}{\sigma^2}B^tW^{-1}B}\right)\\ &= W^{-1} - \frac{\frac{1}{\sigma_1^2}W^{-1}BB^tW^{-1}}{1+\frac{1}{\sigma_1^2}B^tW^{-1}B}\\ &= W^{-1} - \frac{W^{-1}BB^tW^{-1}}{\sigma_1^2+B^tW^{-1}B}.\end{aligned}$$

Therefore,

$$\begin{aligned}\lim_{n_1\to\infty}\boldsymbol{\lambda}^* &= \lim_{n_1\to\infty}\frac{(\sigma_1^2W+BB^t)^{-1}\mathbf{1}}{\mathbf{1}^t(\sigma_1^2W+BB^t)^{-1}\mathbf{1}}\\ &= \lim_{n_1\to\infty}\frac{\sigma_1^2(\sigma_1^2W+BB^t)^{-1}\mathbf{1}}{\sigma_1^2\mathbf{1}^t(\sigma_1^2W+BB^t)^{-1}\mathbf{1}}\\ &= \lim_{n_1\to\infty}\frac{\left(W^{-1}-\frac{W^{-1}BB^tW^{-1}}{\sigma_1^2+B^tW^{-1}B}\right)\mathbf{1}}{\mathbf{1}^t\left(W^{-1}-\frac{W^{-1}BB^tW^{-1}}{\sigma_1^2+B^tW^{-1}B}\right)\mathbf{1}}\\ &= \frac{\left(W_0^{-1}-\frac{W_0^{-1}BB^tW_0^{-1}}{B^tW_0^{-1}B}\right)\mathbf{1}}{\mathbf{1}^t\left(W_0^{-1}-\frac{W_0^{-1}BB^tW_0^{-1}}{B^tW_0^{-1}B}\right)\mathbf{1}}\\ &= \frac{S_{m\times m}\mathbf{1}}{\mathbf{1}^tS_{m\times m}\mathbf{1}},\end{aligned}$$

where $S_{m\times m} = W_0^{-1} - \frac{W_0^{-1}BB^tW_0^{-1}}{B^tW_0^{-1}B}$. This completes the proof. $\diamond$

Corollary 3.1 *Under the conditions of Theorem 3.4, assume that* $V = Diag(\sigma_1^2, \sigma_2^2, ..., \sigma_m^2)$. *If* $\sigma_1^2(n_1) \to 0$ *and* $\sigma_1^2(n_1)/\sigma_i^2(n_i) \to \gamma_i \geq 0$, *for* $i \geq 2$, *as* $n_1 \to \infty$ *and* $\sum_{i=2}^{m}\gamma_i > 0$, *then*

$$\lim_{n_1\to\infty}\boldsymbol{\lambda}^* = (1,0,0,...,0)^t.$$

Proof: Consider the following

$$\begin{aligned}
S_{m\times m}\mathbf{1} &= \left(W_0^{-1} - \frac{W_0^{-1}BB^tW_0^{-1}}{B^tW_0^{-1}B}\right)\mathbf{1} \\
&= \left(Diag(1,\gamma_2,...,\gamma_m) - \frac{Diag(1,\gamma_2,...,\gamma_m)BB^tDiag(1,\gamma_2,...,\gamma_m)}{B^tDiag(1,\gamma_2,...,\gamma_m)B}\right)\mathbf{1} \\
&= \left(Diag(1,\gamma_2,...,\gamma_m) - \frac{C^2(0,\gamma_2,...,\gamma_m)^t(0,\gamma_2,...,\gamma_m)}{C^2\sum_{i=2}^{m}\gamma_i}\right)\mathbf{1} \\
&= (1,\gamma_2,...,\gamma_m)^t - \frac{1}{\sum_{i=2}^{m}\gamma_i}\left(0,\gamma_2\sum_{i=2}^{m}\gamma_i,...,\gamma_m\sum_{i=2}^{m}\gamma_i\right)^t \\
&= (1,0,...,0)^t.
\end{aligned}$$

It follows that $\mathbf{1}^tS\mathbf{1} = 1$. This implies that

$$\lim_{n\to\infty}\boldsymbol{\lambda}^* = (1,0,...,0)^t. \quad \diamond$$

Next we will consider the case where the covariance matrix is not a diagonal matrix but has a special form. Suppose that $cov(\hat{\boldsymbol{\theta}}) = \frac{\sigma^2}{n}V_0$ where

$$V_0 = \begin{pmatrix} 1 & \rho & \rho^2 & ... & \rho^{m-1} \\ \rho & 1 & \rho & ... & \rho^{m-2} \\ \vdots & \vdots & & & \vdots \\ \rho^{m-1} & \rho^{m-2} & \rho^{m-3} & ... & 1 \end{pmatrix} \tag{3.2}$$

where $\rho \neq 0$ and $\rho^2 < 1$. This kind of covariance structure is found in first order autoregressive model with lage 1 effect, namely AR(1) model, in time series analysis. If the goal of inference is to predict the average response at the next time point given observations from current and a fixed number of previous time points, then this type of covariance structure is of our interest. By Proposition 2.2, the optimum weights will go to $(1,0)$ if $m = 2$. This is because $K = \frac{C^2}{(1-\rho)\sigma_1^2(n)}$ goes to infinity. As a result, the optimum weights $\boldsymbol{\lambda} = \left(\frac{1+K}{2+k}, \frac{1}{2+K}\right) \to (1,0)$.

Corollary 3.2 *For $m > 2$, if $cov(\hat{\boldsymbol{\theta}}) = \frac{\sigma^2}{n} V_0$, where V_0 takes the form as in (3.2), then*

$$\lim_{n\to\infty} \boldsymbol{\lambda}^* = \left(1, -\rho + \frac{\rho(1-\rho)}{D_0 - \rho^2}, \frac{\rho(1-\rho)^2}{D_0 - \rho^2}, ..., \frac{\rho(1-\rho)^2}{D_0 - \rho^2}, \frac{\rho(1-\rho)}{D_0 - \rho^2}\right)^t,$$

where $D_0 = 1 + (m-2)(1-\rho)^2$.

Proof: It can be verified that

$$W_0^{-1} = \frac{1}{1-\rho^2} \begin{pmatrix} 1 & -\rho & 0 & 0 & ... & 0 \\ -\rho & 1+\rho^2 & -\rho & 0 & ... & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ 0 & ... & 0 & -\rho & 1+\rho^2 & -\rho \\ 0 & 0 & ... & 0 & -\rho & 1 \end{pmatrix}.$$

It follows that

$$\begin{aligned} W_0^{-1}B &= \frac{C}{1-\rho^2}\left(-\rho, 1-\rho+\rho^2, (1-\rho)^2, ..., (1-\rho)^2, 1-\rho\right)^t, \\ B^t W_0^{-1}\mathbf{1} &= \frac{C}{1-\rho^2}(D_0 - \rho), \\ B^t W_0^{-1} B &= \frac{C^2}{1-\rho^2} D_0. \end{aligned}$$

We then have

$$\begin{aligned} \frac{W_0^{-1}BB^tW_0^{-1}\mathbf{1}}{B^tW_0^{-1}B} &= \frac{B^tW_0^{-1}\mathbf{1}}{B^tW_0^{-1}B} W_0^{-1}B \\ &= \frac{D_0 - \rho}{(1-\rho^2)D_0}(-\rho, 1-\rho+\rho^2, (1-\rho)^2, ..., (1-\rho)^2, 1-\rho)^t. \end{aligned}$$

Therefore,

$$S\mathbf{1} = \frac{1}{1-\rho^2}\begin{pmatrix} 1-\rho \\ (1-\rho)^2 \\ (1-\rho)^2 \\ \vdots \\ (1-\rho)^2 \\ 1-\rho \end{pmatrix} - \frac{D_0-\rho}{(1-\rho^2)D_0}\begin{pmatrix} -\rho \\ 1-\rho+\rho^2 \\ (1-\rho)^2 \\ \vdots \\ (1-\rho)^2 \\ 1-\rho \end{pmatrix}$$

$$= \frac{1}{1-\rho^2}\left(1-\frac{\rho^2}{D_0}, -\rho+\frac{\rho(1-\rho+\rho^2)}{D_0}, \frac{\rho(1-\rho)^2}{D_0}, ..., \frac{\rho(1-\rho)^2}{D_0}, \frac{\rho(1-\rho)}{D_0}\right)^t.$$

It can be verified that $\lim_{n_1\to\infty} \mathbf{1}^t S\mathbf{1} = \frac{1}{1-\rho^2}(1-\frac{\rho^2}{D_0})$. This completes the proof. $\diamond$

Note that the limits of λ_2^* and λ_m^* take different form than that of $\lambda_3^*, ..., \lambda_{m-1}^*$. The reason is that $B = (0, C, C, ..., C)$ ignores the first row or column of W_0^{-1} when multiplied with it and the last row of W_0^{-1} is different than the other rows in that matrix.

Positive Weights

If the estimators are uncorrelated, that is, $\rho = 0$, then we show below that asymptotically all weights are positive. Furthermore, all the weights are always positive in this case.

Theorem 3.5 *As before, let V be the covariance matrix of $(\hat{\theta}_1, \hat{\theta}_2, ..., \theta_m)$ and let $B = (0, C, C, ..., C)^t$. Furthermore, assume that $V = diag(\sigma_1^2, \sigma_2^2, ..., \sigma_m^2)$, where the $\sigma_i^2, i = 1, 2, ..., m$, are known. Then $\lambda_i^* > 0$ for all $i = 1, 2, ..., m$ and for all n.*

Proof: Observe that $V^{-1} = diag\left(\frac{1}{\sigma_1^2}, \frac{1}{\sigma_2^2}, ..., \frac{1}{\sigma_m^2}\right)$.

The matrix $(V + BB^t)^{-1}$ can be written as

$$(V + BB^t)^{-1} = V^{-1} - \frac{T_1}{1+t_0},$$

where $t_0 = B^t V^{-1} B$ and $T_1 = (V^{-1}B)(B^t V^{-1})$.

It may be verified that

$$t_0 = \sum_{i=2}^{m} \frac{C^2}{\sigma_i^2},$$

and

$$T_1 = \begin{pmatrix} 0 & 0 & \dots & 0 \\ 0 & \frac{C^2}{\sigma_2^2\sigma_2^2} & \dots & \frac{C^2}{\sigma_2^2\sigma_m^2} \\ \dots & & & \\ 0 & \frac{C^2}{\sigma_m^2\sigma_2^2} & \dots & \frac{C^2}{\sigma_m^2\sigma_m^2} \end{pmatrix}.$$

Therefore, we have

$$\begin{aligned} (V+BB^t)^{-1} &= diag\left(\frac{1}{\sigma_1^2}, \frac{1}{\sigma_2^2}, \dots, \frac{1}{\sigma_m^2}\right) - \frac{1}{1+t_0} T_1 \\ &= \begin{pmatrix} \frac{1}{\sigma_1^2} & 0 & \dots & 0 \\ 0 & \frac{1}{\sigma_2^2} & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & \frac{1}{\sigma_m^2} \end{pmatrix} - \frac{1}{1+t_0} \begin{pmatrix} 0 & 0 & \dots & 0 \\ 0 & \frac{C^2}{\sigma_2^2\sigma_2^2} & \dots & \frac{C^2}{\sigma_2^2\sigma_m^2} \\ \vdots & \vdots & & \vdots \\ 0 & \frac{C^2}{\sigma_m^2\sigma_2^2} & \dots & \frac{C^2}{\sigma_m^2\sigma_m^2} \end{pmatrix}. \end{aligned}$$

Furthermore,

$$\begin{aligned} (V+BB^t)^{-1}\,\mathbf{1} &= diag\left(\frac{1}{\sigma_1^2}, \frac{1}{\sigma_2^2}, \dots, \frac{1}{\sigma_m^2}\right)\mathbf{1} - \frac{1}{1+t_0} T_1\,\mathbf{1} \\ &= \begin{pmatrix} \frac{1}{\sigma_1^2} & 0 & \dots & 0 \\ 0 & \frac{1}{\sigma_2^2} & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & \frac{1}{\sigma_m^2} \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} - \frac{1}{1+t_0} \begin{pmatrix} 0 & 0 & \dots & 0 \\ 0 & \frac{C^2}{\sigma_2^2\sigma_2^2} & \dots & \frac{C^2}{\sigma_2^2\sigma_m^2} \\ \dots & & & \\ 0 & \frac{C^2}{\sigma_m^2\sigma_2^2} & \dots & \frac{C^2}{\sigma_m^2\sigma_m^2} \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \\ &= \begin{pmatrix} \frac{1}{\sigma_1^2} \\ \frac{1}{\sigma_2^2} - \frac{1}{1+t_0}\sum_{i=2}^{m} \frac{C^2}{\sigma_2^2\sigma_i^2} \\ \vdots \\ \frac{1}{\sigma_m^2} - \frac{1}{1+t_0}\sum_{i=2}^{m} \frac{C^2}{\sigma_m^2\sigma_i^2} \end{pmatrix} = \begin{pmatrix} \frac{1}{\sigma_1^2} \\ \frac{1}{\sigma_2^2}\frac{1}{1+t_0} \\ \vdots \\ \frac{1}{\sigma_m^2}\frac{1}{1+t_0} \end{pmatrix}, \end{aligned}$$

since

$$\frac{1}{\sigma_j^2} - \frac{1}{1+t_0}\sum_{i=2}^{m} \frac{C^2}{\sigma_j^2\sigma_i^2} = \frac{1}{\sigma_j^2}\left(1 - \frac{t_0}{1+t_0}\right) = \frac{1}{\sigma_j^2}\frac{1}{1+t_0}.$$

Hence,

$$\mathbf{1}^t \left(V + BB^t\right)^{-1} \mathbf{1} = \frac{1}{\sigma_1^2} + \frac{1}{1+t_0} \sum_{j=2}^{m} \frac{1}{\sigma_j^2}.$$

Therefore,

$$\lambda_1^* = \frac{\frac{1}{\sigma_1^2}}{\frac{1}{\sigma_1^2} + \frac{1}{1+t_0} \sum_{j=2}^{m} \frac{1}{\sigma_j^2}} > 0.$$

Also, for $2 \leq i \leq m$,

$$\lambda_i^* = \frac{\frac{1}{\sigma_i^2} \frac{1}{1+t_0}}{\frac{1}{\sigma_1^2} + \frac{1}{1+t_0} \sum_{j=2}^{m} \frac{1}{\sigma_j^2}} > 0.$$

This completes the proof. $\diamond$

Negative Weights

The above theorem gives conditions that insure positive weights. However, weights will not always be non-negative. In Corollary 3.2 , we have

$$\lim_{n_1 \to \infty} \boldsymbol{\lambda}^* = \left(1, \frac{-\rho D_0 + \rho(1-\rho+\rho^2)}{D_0 - \rho^2}, \frac{\rho(1-\rho)^2}{D_0 - \rho^2}, ..., \frac{\rho(1-\rho)^2}{D_0 - \rho^2}, \frac{\rho(1-\rho)}{D_0 - \rho^2}\right)^t,$$

where $\rho^2 < 1$ and $D_0 = 1 + (m-2)(1-\rho)^2$. It follows that $D_0 > 1$ provided $m > 2$. Hence the sign of those asymptotic weights taking the form $\frac{\rho(1-\rho)^2}{D_0 - \rho^2}$ are determined by the sign of ρ. Notice that $\sum_{i=1}^{m} \lim_{n_1 \to \infty} \lambda_i^* = 1$ and $\lim_{n_1 \to \infty} \lambda_1 = 1^*$. Hence $\lim_{n_1 \to \infty} \lambda_2^* = -\sum_{i=3}^{m} \lim_{n_1 \to \infty} \lambda_i^*$. Therefore, we have the following.

1) If $\rho > 0$, then $\lim_{n_1 \to \infty} \lambda_i^* > 0$, $i = 3, ..., m$, and $\lambda_2 < 0$.

2) If $\rho < 0$, then $\lim_{n_1 \to \infty} \lambda_i^* < 0$, $i = 3, ..., m$, and $\lambda_2 > 0$.

A general result on negative weights is shown below.

Theorem 3.6 *Assume* $B = (0, C, C, ..., C)$ *and* $m \geq 3$. *Let* $W_0^{-1} = (a_{ij})_{m \times m}$. *Let* $e_i = \sum_{j=2}^{m} a_{ij}$. *Assume* $a_{11} - \frac{e_1^2}{\sum_{i=2} e_i} > 0$. *Then the asymptotic weight,* $\lim_{n_1 \to \infty} \lambda_i^*$, *is negative if* $\frac{a_{i1}}{e_1} < \frac{e_i}{\sum_{i=2}^{m} e_i}$ *for some* $i > 1$.

Proof:

By Theorem 3.4, we have

$$\lim_{n\to\infty} \boldsymbol{\lambda}^* = \lim_{n\to\infty} \frac{(V+BB^t)^{-1}\mathbf{1}}{\mathbf{1}^t(V+BB^t)^{-1}\mathbf{1}} = \frac{S_{m\times m}\mathbf{1}}{\mathbf{1}^t S_{m\times m}\mathbf{1}}$$

where $S_{m\times m} = W_0^{-1} - \frac{W_0^{-1}BB^tW_0^{-1}}{B^tW_0^{-1}B}$.

We then have

$$W_0^{-1}B = (C\sum_{j=2}^{m} a_{ij}, C\sum_{j=2}^{m} a_{2j}, ..., C\sum_{j=2}^{m} a_{mj})^t = C(e_1, e_2, ..., e_m)^t.$$

Therefore,

$$\begin{aligned}
S_{m\times m}\mathbf{1} &= W_0^{-1}\mathbf{1} - \frac{W_0^{-1}BB^tW_0^{-1}}{B^tW_0^{-1}B}\mathbf{1} \\
&= ((\sum_{j=1}^{m} a_{1j}, \sum_{j=1}^{m} a_{2j}, ..., \sum_{j=1}^{m} a_{mj})^t - \frac{C(e_1, e_2, ..., e_m)^t C(e_1, e_2, ..., e_m)}{(0, C, C, ..., C)C(e_1, e_2, ..., e_m)^t}\mathbf{1} \\
&= (a_{11}+e_1, a_{12}+e_2, ..., a_{m1}+e_m)^t - \frac{C^2(e_1, e_2, ..., e_m)^t(e_1, e_2, ..., e_m)}{C^2\sum_{i=2}^{m} e_i}\mathbf{1} \\
&= \begin{pmatrix} a_{11}+e_1 \\ a_{12}+e_2 \\ \vdots \\ a_{m1}+e_m \end{pmatrix} - \frac{1}{\sum_{i=2}^{m} e_i} \begin{pmatrix} e_1^2 & e_1e_2 & ... & e_1e_m \\ e_2e_1 & e_2^2 & ... & e_2e_m \\ ... & & & \\ e_me_1 & e_me2 & ... & e_m^2 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}
\end{aligned}$$

It follows that

$$
\begin{aligned}
S_{m\times m}\mathbf{1} &= \begin{pmatrix} a_{11}+e_1 \\ a_{12}+e_2 \\ \vdots \\ a_{m1}+e_m) \end{pmatrix} - \frac{1}{\sum_{i=2}^{m} e_i} \begin{pmatrix} e_1 \sum_{i=1}^{m} e_i \\ e_2 \sum_{i=1}^{m} e_i \\ \vdots \\ e_m \sum_{i=1}^{m} e_i \end{pmatrix} \\
&= \begin{pmatrix} a_{11}+e_1 - \frac{e_1(e_1+\sum_{i=2}^{m} e_i}{\sum_{i=2}^{m} e_i} \\ a_{21}+e_2 - \frac{e_2(e_1+\sum_{i=2}^{m} e_i)}{\sum_{i=2}^{m} e_i} \\ \vdots \\ a_{m1}+e_m - \frac{e_m(e_1+\sum_{i=2}^{m} e_i)}{\sum_{i=2}^{m} e_i} \end{pmatrix} \\
&= (a_{11} - \frac{e_1^2}{\sum_{i=2}^{m} e_i}, a_{21} - \frac{e_2\, e_1}{\sum_{i=2}^{m} e_i}, \ldots, a_{m1} - \frac{e_m\, e_1}{\sum_{i=2}^{m} e_i})^t.
\end{aligned}
$$

Note that W_0 is symmetric which implies that the matrix W_0^{-1} is also symmetric. Thus, we have

$$\sum_{j=2}^{m} a_{1j} = \sum_{j=2}^{m} a_{j1}.$$

By the assumption of the theorem, we have

$$\mathbf{1}^t S_{m\times m}\mathbf{1} = a_{11} - \frac{e_1^2}{\sum_{i=2}^{m} e_i} > 0,$$

Therefore, $\lim_{n_1\to\infty} \lambda_i^* < 0$ if $a_{i1} < \frac{e_i\, e_1}{\sum_{i=2}^{m} e_i}$ for any i such that $i \geq 2$. $\diamond$

We can construct a simple example where the non-asymptotic weights can actually be negative. Suppose that information from three populations is available and one single observation is obtained from each population. Further, we assume that the three random variables, X_1, X_2, and X_3, say, follow a multivariate normal distribution

with covariance matrix as follows:

$$V = \begin{pmatrix} 1.0 & 0.7 & 0.3 \\ 0.7 & 1.0 & 0.7 \\ 0.3 & 0.7 & 1.0 \end{pmatrix}.$$

Also, assume that $C_2 = C_3 = 1$. Thus $B = (0, 1, 1)^t$. It follows that

$$V + BB^t = \begin{pmatrix} 1.0 & 0.7 & 0.3 \\ 0.7 & 1.0 & 0.7 \\ 0.3 & 0.7 & 1.0 \end{pmatrix} + \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1.0 & 1.0 \\ 0 & 1.0 & 1.0 \end{pmatrix} = \begin{pmatrix} 1.0 & 0.7 & 0.3 \\ 0.7 & 2.0 & 1.7 \\ 0.3 & 1.7 & 2.0 \end{pmatrix}.$$

Hence, approximately

$$(V + BB^t)^{-1} = \begin{pmatrix} 1.67 & -1.34 & 0.89 \\ -1.34 & 2.88 & -2.24 \\ 0.89 & -2.24 & 2.27 \end{pmatrix}.$$

We then have

$$(V + BB^t)^{-1}\mathbf{1} = (1.22, -0.71, 0.92)^t,$$

and

$$\mathbf{1}^t(V + BB^t)^{-1}\mathbf{1} = 1.43.$$

It follows that

$$\boldsymbol{\lambda}^* = \frac{(V + BB^t)^{-1}\mathbf{1}}{\mathbf{1}^t(V + BB^t)^{-1}\mathbf{1}} = (0.85, -0.49, 0.64).$$

Thus, λ_2 is negative in this example. The negative weights in this example might be due to the collinearity. If we replace 0.7 by 0.3 in the covariance matrix, all the weights will be positive.

Chapter 4

Asymptotic Properties of the WLE

Throughout this chapter, θ_1 is the parameter of primary inferential interest although in their extension of the REWL, Hu and Zidek (1997) consider simultaneous inference for all the θ's. Recall that for fixed $\boldsymbol{X} = \boldsymbol{x}$, the weighted likelihood (WL) is defined as:

$$\prod_{i=1}^{m}\prod_{j=1}^{n_i} f_1(x_{ij};\theta_1)^{\lambda_i}, \tag{4.1}$$

where $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, ..., \lambda_m)^t$ is the "weight vector".

It follows that

$$log\mathrm{WL}(\boldsymbol{x},\theta_1) = \sum_{i=1}^{m}\sum_{j=1}^{n_i} \lambda_i \; log \; f_1(x_{ij};\theta_1).$$

We say that $\tilde{\theta}_1$ is a maximum weighted likelihood estimator (WLE) for θ_1 if

$$\tilde{\theta}_1 = arg \sup_{\theta_1 \in \Theta} \mathrm{WL}(\boldsymbol{x},\theta_1).$$

We will use θ_1^0 to denote the true value of θ_1 in the sequel.

The asymptotic results proved here differ from those of Hu (1997) because a different asymptotic paradigm is used. Hu's paradigm abstracts that of non-parametric regression and function estimation. There information about θ_1 builds up because the number of populations grows with increasingly many in close proximity to that of θ_1.

This is the paradigm commonly invoked in the context of non-parametric regression but it is not always the most natural one. In contrast we postulate a fixed number of populations with an increasingly large number samples from each. Asymptotically, the procedure can rely on just the data from the population of interest alone. These results offer guidance on the difficult problem of specifying $\boldsymbol{\lambda}$.

We also consider the general version of the adaptively weighted likelihood in which the weights are allowed to depend on the data. Such likelihood arises naturally when the responses are measured on a sequence of independent draw on discrete random variables. In that case the likelihood factors into powers of the common probability mass function at successive discrete points in the sample space. (The multinomial likelihood arises in precisely this way for example). The factors in the likelihood may well depend on a vector of parameters deemed to be approximately fixed during the sampling period. The sample itself now "self-weights" the likelihood factors according to their degree of relevance in estimating the unknown parameter.

In Section 4.1, we present our extension of the classical large sample theory for the the asymptotic results for the maximum likelihood estimator. Both consistency and asymptotic normality of the WLE for the fixed weights are shown under appropriate assumptions. The weights may be "adaptive" that is, allowed to depend on the data. In Section 4.2 we present the asymptotic properties of the WLE using adaptive weights.

4.1 Asymptotic Results for the WLE

In this section, establish the consistency and asymptotic normality of the WLE under appropriate conditions.

4.1.1 Weak Consistency

Consistency is a minimal requirement for any good estimate of the parameter of interest. In this and the next sub-section, we will give a general conditions that ensure the consistency of the WLE's. Our analysis concerns σ-finite probability spaces $(\mathcal{X}, \mathcal{F}, \mu_i), i = 1, 2$ under suitable regularity conditions. We assume that the probability measures μ_i is *absolutely continuous* with respect to one another; that is, suppose there exists no set (event) $E \in \mathcal{F}$ for which $\mu_i(E) = 0$ and $\mu_j(E) \neq 0$, or $\mu_i(E) = 0$ and $\mu_j(E) \neq 0$ for $i \neq j$. Let ν be a measure that dominates μ_i, $i = 1, 2$, for example $(\mu_1 + \mu_2)/2$ and f_i, $i = 1, 2$. By the Radon-Nikodym theorem (Royden 1988, p. 276), there exist measurable functions $f_i(x)$, $i = 1, 2$, called *probability densities*, unique up to sets of (probability) measure zero in ν, $0 < f_i(x) < \infty$ $(a.e.\ \nu), i = 1, 2, ..., m$ such that

$$\mu_i(E) = \int_E f_i(x) d\nu(x), \quad i = 1, 2.$$

for all $E \in \mathcal{F}$.

Define the *Kullback-Leibler information number* as:

$$K(f_1, f_2) = E_1\left(log\frac{f_1(X)}{f_2(X)}\right) = \int log\frac{f_1(x)}{f_2(x)} f_1(x) d\nu(x).$$

In this expression, $log(f_1(X)/f_2(X))$ is defined as $+\infty$ if $f_1(x) > 0$ and $f_2(x) = 0$, so the expectation could be $+\infty$. Although $log(f_1(x)/f_2(x))$ is defined as $-\infty$ when $f_1(x) = 0$ and $f_2(x) > 0$, the integrand, $log(f_1(x)/f_2(x))f_1(x)$ is defined as zero in this case. The next lemma gives well known result.

Lemma 4.1 *(Shannon-Kolmogorov Information Inequality) Let $f_1(x)$ and $f_2(x)$ be densities with respect to ν. Then*

$$K(f_1, f_2) = E_1\left(log\frac{f_1(X)}{f_2(X)}\right) = \int log\frac{f_1(x)}{f_2(x)} f_1(x) d\nu(x) \geq 0,$$

with equality if and only if $f_1(x) = f_2(x)$ $(a.e.\ \nu)$.

Proof: (See for example, Ferguson 1996, p. 113).

Let θ_1^0 denote the true value of θ_1 and $\theta^0 = (\theta_1^0, \theta_2, ..., \theta_m)$, for $\theta_i \in \Theta, i = 2, 3, ..., m$. Throughout this chapter, the following assumptions are assumed to hold except where otherwise stated.

Assumption 4.1 *The parameter space Θ is compact and separable.*

Assumption 4.2 *For each $i = 1, .., m$, assume $\{X_{ij} : j = 1, 2, ..., n_i.\}$ are i.i.d. random variables having common probability density functions with respect to ν.*

Assumption 4.3 *Assume $f_1(x, \theta_1) = f_1(x, \theta_1')(a.e.)\ \nu)$ implies that $\theta_1 = \theta_1'$ for any $\theta_1, \theta_1' \in \Theta$ and the densities $f_1(x, \theta)$ have the same support for all $\theta \in \Theta$.*

Assumption 4.4 *For any $\theta_1^0 \in \Theta$ and for any open set $O \subseteq \Theta$, assume*

$$\sup_{\theta_1 \in O} |log(f_1(x; \theta_1^0)/f_1(x; \theta_1))| \quad \inf_{\theta_1 \in O} |log(f_1(x; \theta_1^0)/f_1(x; \theta_1))|$$

$$\sup_{\theta_1 \in \Theta} |log(f_1(x; \theta_1^0)/f_1(x; \theta_1))| \quad \inf_{\theta_1 \in \Theta} |log(f_1(x; \theta_1^0)/f_1(x; \theta_1))|$$

are each measurable in x and

$$E_{\theta_i}[\sup_{\theta_1 \in \Theta} |\log \frac{f_1(X_{ij}; \theta_1^0)}{f_1(X_{ij}; \theta_1)}|]^2 \leq K < \infty,$$

where $K > 0$ is a constant independent of $\theta_i, i = 1, 2, .., m$.

Assumption 4.5 *Let $\boldsymbol{n} = (n_1, n_2, ..., n_m)$. Assume $\boldsymbol{\lambda^{(n)}} = (\lambda_1^{(n)}, \lambda_2^{(n)}, ..., \lambda_m^{(n)})^t$ satisfies*

$$\boldsymbol{\lambda^{(n)}} \to \boldsymbol{w} = (w_1, w_2, ..., w_m)^t \triangleq (1, 0, ..., 0)^t$$

while

$$\max_{1 \leq k \leq m} n_k^2 \max_{1 \leq i \leq m} |w_i - \lambda_i^{(n)}|^2 \leq O\left(n_1^{1-\delta}\right) \quad as \ n_1 \to \infty,$$

for some $\delta > 0$.

Assumption 4.5 will be satisfied if $n_i, i = 2, ..., m$ are in the same order of n_1 and also $|w_i - \lambda_i^{(n)}| = O\left(n^{(1+\delta)/2}\right)$.

In this chapter, we require the density functions to be upper semi-continuous. Let $||.||$ be defined as Euclidean norm; that is

$$||\mathbf{x}|| = (\mathrm{x}^t\mathrm{x})^{1/2} = (\sum_{i=1}^{q} x_i^2)^{1/2},$$

for any $\mathbf{x} = (x_1, x_2, ..., x_q)^t$.

Definition 4.1 *A real-valued function, $g(\theta)$, defined on the parameter space, Θ, is said to be* upper semi-continuous *on Θ, if for all $\theta \in \Theta$ and for any sequence $\theta_n \in \Theta$ such that $\lim_{n\to\infty} ||\theta_n - \theta|| = 0$, we have $g(\theta) \geq \limsup_{n\to\infty} g(\theta_n)$. A function is called* lower semi-continuous *if $g(\theta) \leq \liminf_{n\to\infty} g(\theta_n)$ whenever $\lim_{n\to\infty} ||\theta_n - \theta|| = 0$.*

We need to show that $\sup_{\theta_1 \in O} U(x;\theta_1) = log\frac{f_1(x;\theta_1^0)}{f_1(x;\theta_1)}$ for some open set O is measurable if $f_1(x;\theta_1)$ is upper semi-continuous.

Proposition 4.1 *If $U(x;\theta_1)$ is lower semi-continuous in θ_1 for all x, then $\sup_{||\theta_1-\theta_0||<R} U(x;\theta_1)$ is measurable.*

Proof: For simplicity, let us assume that $\Theta = R$ and $O = \{\theta_1 : ||\theta_1 - \theta_1^0|| < R\}$. We will show that

$$\sup_{|\theta_1-\theta_1^0|<R} U(x;\theta_1) = \sup_{\theta_1 \in D} U(x;\theta_1) \tag{4.2}$$

where $D = N \cap \{\theta_1 : |\theta_1 - \theta_1^0| < R\}$. The set N is the set of all the rational numbers in R.

Let $s = \sup_{|\theta_1-\theta_1^0|<R} U(x;\theta_1)$. It follows that for any $\theta_1 \in D$, $U(x;\theta_1) \leq s$. For any given $\epsilon > 0$, there exist $\theta_1^*(\epsilon)$ such that $|\theta_1^*(\epsilon) - \theta_1^0| < R$ and

$$U(x, \theta_1^*(\epsilon)) > s - \epsilon/2. \tag{4.3}$$

Since D is a dense subset of $\{\theta_1 : |\theta_1-\theta_1^0| < R\}$, then there exist a sequence $\theta_1^{(n)}(\epsilon) \in D$ such that $\lim_{n\to\infty} \theta_1^{(n)}(\epsilon) = \theta_1^*(\epsilon)$. Since $U(x;\theta_1)$ is lower semi-continuous, then, for fixed ϵ and some $\delta > 0$, there exist $\theta_1^{**} \in D$ such that $|\theta_1^{**} - \theta_1^*| < \delta$ and

$$U(x;\theta_1^*) - \epsilon/2 < U(x;\theta_1^{**}). \tag{4.4}$$

Thus, combining equation (4.3) and (4.4), it follows that for any given $\epsilon > 0$, there exist $\theta_1^{**} \in D$ such that $U(x;\theta_1^{**}) > s - \epsilon$. Equation (4.2) is then established.

We then have

$$\begin{aligned}
&\{x : \sup_{|\theta_1-\theta_1^0|<R} U(x;\theta_1) < \alpha\} \\
&= \{x : \sup_{\theta_1 \in D} U(x;\theta_1) < \alpha\} \\
&= \cap_{i=1}^{n} \{x : U(x;\theta_1^{(i)}) < \alpha, \ \ \theta_1^{(i)} \in D\}.
\end{aligned}$$

Since $\{x : U(x;\theta_1) < \alpha\}$ is measurable. Therefore the set $\{x : \sup_{|\theta_1-\theta_1^0|<R} U(x;\theta_1) < \alpha\}$ is measurable for any θ_1. Therefore the set $\{x : \sup_{|\theta_1-\theta_1^0|<R} U(x;\theta_1) < \alpha\}$ is measurable. This completes the proof. $\diamond$

Therefore, if $f_1(x;\theta_1)$ is upper semi-continuous in θ_1 for all x and the open set is defined as $\{\theta_1 : |\theta_1 - \theta_1^0| < R, R > 0\}$, then the Assumption 4.4 is automatically satisfied because $log\frac{f_1(x;\theta_1^0)}{f_1(x;\theta_1)}$ is lower semi-continuous and

$$\sup_{|\theta_1-\theta_1^0|<R} \left| log\frac{f_1(x;\theta_1^0)}{f_1(x;\theta_1)} \right| = \sup_{\theta_1 \in D} \left| log\frac{f_1(x;\theta_1^0)}{f_1(x;\theta_1)} \right|$$

for any denumerable set D dense in $\{\theta_1 : |\theta_1 - \theta_1^0| < R\}$ by Proposition 4.1. The measurability of $\inf_{|\theta_1-\theta_1^0|<R} \left| log\frac{f_1(x,\theta_1^0)}{f_1(x,\theta_1)} \right|$ also follows.

Lemma 4.2 *Let $A_{ij}(x;\theta)$ be measurable function in x for all θ. If $E_{\theta_i}[A(X_{ij};\theta_i)]^2 < K$ for some constant K independent of θ_i, then*

$$\frac{1}{n_1} \sum_{i=1}^{m} \sum_{j=1}^{n_i} (w_i - \lambda_i^{(n)}) A_{ij}(X_{ij};\theta_i) \xrightarrow{P_{\theta^0}} 0$$

for any $\theta_2, \theta_3, ..., \theta_m$, $\theta_i \in \Theta, i = 1, 2, ...m$.

Proof: Put $A_{ij} = A_{ij}(X_{ij})$ and $B_{ij} = (w_i - \lambda_i^{(n)})A_{ij}$. Observe that by the Cauchy-Schwartz Inequality, for any i, i', j, and j', we have

$$E_{\theta^0}|B_{ij}B_{i'j'}| \ \leq \ |w_i - \lambda_i^{(n)}| \, |w_{i'} - \lambda_{i'}^{(n)}| \sqrt{E_{\theta^0}A_{ij}^2 E_{\theta^0}A_{i'j'}^2}$$

in view of the finite second moment condition on the $A_{ij}(X)$.

Further,

$$\begin{aligned} P_{\theta^0}\left(\frac{1}{n_1}\sum_{i=1}^{m}\sum_{j=1}^{n_i}B_{ij} > \epsilon\right) &\leq \frac{1}{n_1^2\epsilon^2}E_{\theta^0}[\sum_{i=1}^{m}\sum_{j=1}^{n_i}|B_{ij}|]^2 \\ &\leq \frac{1}{n_1^2\epsilon^2}\sum_{i=1}^{m}\sum_{i'=1}^{m}\sum_{j=1}^{n_i}\sum_{j'=1}^{n_{i'}}E_{\theta^0}|B_{ij}B_{i'j'}| \\ &\leq O\left(\frac{1}{n_1^2}\right)\max_{1\leq k\leq m} n_k^2 \max_{1\leq i\leq m}|w_i - \lambda_i^{(\boldsymbol{n})}|^2 \\ &\leq O\left(\frac{1}{n_1^2}\right)O\left(n_1^{1-\delta}\right) \\ &\leq O\left(\frac{1}{n_1^{1+\delta}}\right) \longrightarrow 0, \text{ as } n_1 \longrightarrow \infty. \end{aligned}$$

by Assumption 4.5.

It then follows that

$$\sup_{\theta_1\in\Theta}\left|\frac{1}{n_1}\sum_{i=1}^{m}\sum_{j=1}^{n_i}(w_i - \lambda_i^{(\boldsymbol{n})})A_{ij}\right| \leq \frac{1}{n_1}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\left|w_i - \lambda_i^{(\boldsymbol{n})}A_{ij}\right| = \frac{1}{n_1}\sum_{i=1}^{m}\sum_{j=1}^{n_i}|B_{ij}|.$$

We then have

$$\frac{1}{n_1}\sum_{i=1}^{m}\sum_{j=1}^{n_i}(w_i - \lambda_i^{(\boldsymbol{n})})A_{ij}(X_{ij};\theta_i) \xrightarrow{P_{\theta^0}} 0$$

for any $\theta_1, \theta_2, \theta_3, ..., \theta_m, \theta_i \in \Theta, i = 1, 2, ..., m$. $\diamond$

If we set $A_{ij} = log\frac{f_1(X_{ij};\theta_1^0)}{f_1(X_{ij};\theta_1)}$ and

$$\frac{1}{n_1}S_{n_1}(\mathbf{X},\theta_1) = \frac{1}{n_1}\sum_{i=1}^{m}\sum_{j=1}^{n_i}(w_i - \lambda_i^{(\boldsymbol{n})})log\frac{f_1(X_{ij};\theta_1^0)}{f_1(X_{ij};\theta_1)}.$$

It then follows from Lemma 4.2 that

$$\left|\frac{1}{n_1}S_{n_1}(\mathbf{X},\theta_1)\right| \xrightarrow{P_{\theta^0}} 0 \tag{4.5}$$

for any $\theta_2, \theta_3, ..., \theta_m, \theta_i \in \Theta, i = 2, 3, ..., m$.

The above result will be used to establish the following theorem which will be applied to prove the weak consistency and asymptotic normality of the WLE.

Theorem 4.1 *For* $\theta_1 \neq \theta_1^0$,

$$\lim_{n_1 \to \infty} P_{\theta^0} \left(\prod_{i=1}^{m} \prod_{j=1}^{n_i} f_1(X_{ij}; \theta_1^0)^{\lambda_i^{(n)}} > \prod_{i=1}^{m} \prod_{j=1}^{n_i} f_1(X_{ij}; \theta_1)^{\lambda_i^{(n)}} \right) = 1,$$

for any $\theta_2, \theta_3, ..., \theta_m, \theta_i \in \Theta, i = 2, 3, ..., m.$

Proof:

With P_{θ^0} measure 1,

$$\prod_{i=1}^{m} \prod_{j=1}^{n_i} f_1(X_{ij}; \theta_1^0)^{\lambda_i^{(n)}} > \prod_{i=1}^{m} \prod_{j=1}^{n_i} f_1(X_{ij}; \theta_1)^{\lambda_i^{(n)}}$$

if and only if

$$W_{n_1}(\mathbf{X}, \theta_1) > 0.$$

where

$$W_{n_1}(\mathbf{X}, \theta_1) = \frac{1}{n_1} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \lambda_i^{(n)} log \frac{f_1(X_{ij}; \theta_1^0)}{f_1(X_{ij}; \theta_1)}.$$

Observe that

$$\begin{aligned} W_{n_1}(\mathbf{X}, \theta_1) &= \frac{1}{n_1} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \lambda_i^{(n)} \, log \frac{f_1(X_{ij}; \theta_1^0)}{f_1(X_{ij}; \theta_1)} \\ &= \frac{1}{n_1} \sum_{i=1}^{m} \sum_{j=1}^{n_i} w_i \, log \frac{f_1(X_{ij}; \theta_1^0)}{f_1(X_{ij}; \theta_1)} - \frac{1}{n_1} \sum_{i=1}^{m} \sum_{j=1}^{n_i} (w_i - \lambda_i^{(n)}) \, log \frac{f_1(X_{ij}; \theta_1^0)}{f_1(X_{ij}; \theta_1)} \\ &= \frac{1}{n_1} \sum_{j=1}^{n_1} log \frac{f_1(X_{1j}; \theta_1^0)}{f_1(X_{1j}; \theta_1)} - \frac{1}{n_1} S_{n_1}(\mathbf{x}, \theta_1). \end{aligned}$$

By equation (4.5) we have

$$\frac{1}{n_1} S_{n_1}(\mathbf{X}, \theta_1) \xrightarrow{P_{\theta^0}} 0$$

for any $\theta_1 \in \Theta$ and any $\theta_2, \theta_3, ..., \theta_m$.

By the weak law of large numbers,

$$\frac{1}{n_1} \sum_{j=1}^{n_1} log \frac{f_1(X_{1j}; \theta_1^0)}{f_1(X_{1j}; \theta_1)} \xrightarrow{P_{\theta^0}} E_{\theta^0} log \frac{f_1(X_{1j}; \theta_1^0)}{f_1(X_{1j}; \theta_1)} > 0,$$

when $\theta_1 \neq \theta_1^0$ by Lemma 4.1 and Assumption 4.3.

Therefore $\lim_{n_1 \to \infty} P_{\theta^0}(W_{n_1} > 0) = 1$ for all $\theta_1 \neq \theta_1^0, \theta_2, \theta_3, ..., \theta_m$. $\diamond$

For any open set O, let $Z_{ij}(O) = \inf_{\theta_1 \in O} log\frac{f_1(X_{ij};\theta_1^0)}{f_1(X_{ij};\theta_1)}$. We are now in a position to prove the weak consistency of the WLE.

Theorem 4.2 *Suppose that $log f_1(x;\theta)$ is upper semi-continuous in θ for all x. Assume that for every $\theta_1 \neq \theta_1^0$ there is an open set N_{θ_1} such that $\theta_1 \in N_{\theta_1} \subset \Theta$. Then for any sequence of maximum weighted likelihood estimates $\tilde{\theta}_1^{(n_1)}$ of θ_1, and for all $\epsilon > 0$,*

$$\lim_{n_1 \to \infty} P_{\theta^0}\left(||\tilde{\theta}_1^{(n_1)} - \theta_1^0|| > \epsilon\right) = 0,$$

for any $\theta_2, \theta_3, ..., \theta_m, \theta_i \in \Theta, i = 2, 3, ..., m$.

Proof: The proof of this theorem given below resembles the proof of weak consistency of the MLE in Schervish (1995).

For each $\theta_1 \neq \theta_1^0$, let $N_{\theta_1}^{(k)}, k = 1, 2, ...$ be a sequence of open balls centered at θ_1 and of radius at most $1/k$ such that for all k,

$$N_{\theta_1}^{(k+1)} \subseteq N_{\theta_1}^{(k)} \subset \Theta.$$

It follows that $\bigcap_{k=1}^{\infty} N_{\theta_1}^{(k)} = \{\theta_1\}$. Thus, for fixed X_{ij}, $Z_{1j}(N_{\theta_1}^{(k)})$ increases with k and therefore has a limit as $k \to \infty$. Note that $log\frac{f_1(x;\theta_1^0)}{f_1(x;\theta_1)}$ is lower semi-continuous in θ_1 for each x. So,

$$\lim_{k \to \infty} Z_{1j}(N_{\theta_1}^{(k)}) \geq log\frac{f_1(X_{1j};\theta_1^0)}{f_1(X_{1j};\theta_1)}.$$

The limit in the last expression is not required to be finite.

Observe that

$$E_{\theta_1^0}\left|Z_{1j}(N_{\theta_1}^{(k)})\right| = E_{\theta_1^0}\left|\inf_{\theta_1' \in N_{\theta_1}^{(k)}} log\frac{f_1(X_{1j};\theta_1^0)}{f_1(X_{1j};\theta_1')}\right| \leq E_{\theta_1^0} \sup_{\theta_1' \in \Theta}\left|log\frac{f_1(X_{1j};\theta_1^0)}{f_1(X_{1j};\theta_1')}\right| < \infty,$$

by Assumption 4.4. This implies that $Z_{1j}(N_{\theta_1}^{(k)})$ are integrable. Using the monotone convergence theorem, we then have

$$\lim_{k\to\infty} E_{\theta^0} Z_{1j}(N_{\theta_1}^{(k)}) = E_{\theta^0} \lim_{k\to\infty} Z_{1j}(N_{\theta_1}^{(k)}) \geq E_{\theta^0}\left(log\frac{f_1(X_{1j};\theta_1^0)}{f_1(X_{1j};\theta_1)}\right) > 0.$$

Thus, we can choose $k^* = k^*(\theta_1)$ so that $E_{\theta^0} Z_{1j}(N_{\theta_1}^{(k^*)}) > 0$. Let $N_{\theta_1}^*$ be the interior of $N_{\theta_1}^{(k^*)}$ for each $\theta_1 \in \Theta$. Let $\epsilon > 0$ and N_0 be the open ball of radius ϵ around θ_1^0. Now, $\Theta \setminus N_0$ is a compact set since Θ is compact. Also,

$$\{N_{\theta_1}^* : \theta_1 \in \Theta \setminus N_0\}$$

is an open cover of $\Theta \setminus N_0$. Therefore, there exist a finite sub-cover, $N_{\theta_1^1}^*, N_{\theta_1^2}^*, ..., N_{\theta_1^p}^*$ such that $E_{\theta^0} Z_{1j}(N_{\theta_1^l}^*) > 0, \;\; l = 1, 2, ..., p.$

We then have

$$\begin{aligned}
&P_{\theta^0}\left(||\tilde{\theta}_1^{(n_1)} - \theta_1^0|| \geq \epsilon\right) \\
&= P_{\theta^0}\left(\tilde{\theta}_1^{(n_1)} \in N_{\theta_1^l}^* \text{ for some } l\right) \\
&\leq \sum_{l=1}^{p} P_{\theta^0}\left(\tilde{\theta}_1^{(n_1)} \in N_{\theta_1^l}^*\right) \\
&\leq \sum_{l=1}^{p} P_{\theta^0}\left(\inf_{\theta_1' \in N_{\theta_1^l}^*} \frac{1}{n_1}\sum_{i=1}^{m}\sum_{j=1}^{n_i} \lambda_i^{(\boldsymbol{n})} log\frac{f_1(X_{ij};\theta_1^0)}{f_1(X_{ij};\theta_1')} \leq 0\right) \quad \text{(by Theorem 4.1)} \\
&\leq \sum_{l=1}^{p} P_{\theta^0}\left(\frac{1}{n_1}\sum_{i=1}^{m}\sum_{j=1}^{n_i} \lambda_i^{(\boldsymbol{n})} Z_{ij}(N_{\theta_1^l}^*) \leq 0\right) \\
&= \sum_{l=1}^{p} P_{\theta^0}\left(\frac{1}{n_1}\sum_{i=1}^{m}\sum_{j=1}^{n_i} w_i Z_{ij}(N_{\theta_1^l}^*) + \frac{1}{n_1}\sum_{i=1}^{m}\sum_{j=1}^{n_i} (\lambda_i^{(\boldsymbol{n})} - w_i) Z_{ij}(N_{\theta_1^l}^*) \leq 0\right) \\
&= \sum_{l=1}^{p} P_{\theta^0}\left(\frac{1}{n_1}\sum_{j=1}^{n_1} Z_{1j}(N_{\theta_1^l}^*) + \frac{1}{n_1}\sum_{i=1}^{m}\sum_{j=1}^{n_i} (\lambda_i^{(\boldsymbol{n})} - w_i) Z_{ij}(N_{\theta_1^l}^*) \leq 0\right).
\end{aligned}$$

If we show the last expression goes to zero as n_1 goes to infinity, then

$$P_{\theta^0}\left(||\tilde{\theta}_1^{(n_1)} - \theta_1^0|| \geq \epsilon\right) \longrightarrow 0 \;\; as \;\; n_1 \to \infty.$$

Since $E_{\theta^0} Z_{ij}(N^*_{\theta_1^l})^2 \leq E_{\theta_i}\left(\sup_{\theta_1\in\Theta}|log\frac{f_1(X_{ij};\theta_1^0)}{f_1(X_{ij};\theta_1)}|\right)^2 \leq K < \infty$ by Assumption 4.4, it follows from Lemma 4.2 that

$$\frac{1}{n_1}\sum_{i=1}^{m}\sum_{j=1}^{n_i}(\lambda_i^{(\boldsymbol{n})} - w_i)Z_{ij}(N^*_{\theta_1^l}) \xrightarrow{P_{\theta^0}} 0 \ \ as \ \ n_1 \longrightarrow \infty.$$

Also,

$$\frac{1}{n_1}\sum_{j=1}^{n_1} Z_{1j}(N^*_{\theta_1^l}) \xrightarrow{P_{\theta^0}} E_{\theta^0} Z_{1j}(N^*_{\theta_1^l}) > 0, \ \text{ for any } \ \theta_1^l \in \Theta \setminus N_0,$$

by the Weak Law of Large Numbers and the construction of $N^*_{\theta_1^l}$. Thus, for any $\theta_1^l \in \Theta \setminus N_0$,

$$P_{\theta^0}\left(\frac{1}{n_1}\sum_{j=1}^{n_1} Z_{1j}(N^*_{\theta_1^l}) + \frac{1}{n_1}\sum_{i=1}^{m}\sum_{j=1}^{n_i}(\lambda_i^{(\boldsymbol{n})} - w_i)Z_{ij}(N^*_{\theta_1^l}) \leq 0\right) \longrightarrow 0 \ \ as \ \ n_1 \to \infty.$$

This implies that

$$\sum_{l=1}^{p} P_{\theta^0}\left(\frac{1}{n_1}\sum_{j=1}^{n_1} Z_{1j}(N^*_{\theta_1^l}) + \frac{1}{n_1}\sum_{i=1}^{m}\sum_{j=1}^{n_i}(\lambda_i^{(\boldsymbol{n})} - w_i)Z_{ij}(N^*_{\theta_1^l}) \leq 0\right) \longrightarrow 0 \ \ as \ \ n_1 \to \infty.$$

Thus the assertion follows. $\diamond$

In the next theorem we drop Assumption 4.1 which assumes the compactness of the parameter space and replace it with a slightly different condition. At the same time we keep Assumption 4.2- 4.5.

Theorem 4.3 *Suppose $log f_1(x;\theta)$ is upper semi-continuous in θ for all x. Assume that for every $\theta_1 \neq \theta_1^0$ there is an open set N_{θ_1} such that $\theta_1 \in N_{\theta_1} \subset \Theta$. In addition, assume that there is a compact subset C of Θ such that $\theta_1^0 \in C$ and*

$$0 < E_{\theta^0}\left(\inf_{\theta_1'\in C^c\cap\Theta} log\frac{f_1(X_{ij};\theta_1^0)}{f_1(X_{ij};\theta_1')}\right) \leq K^C < \infty, \tag{4.6}$$

where K^C is a constant independent of $\theta_2, ..., \theta_m$.

Then for any sequence of maximum weighted likelihood estimates $\tilde{\theta}_1^{(n_1)}$ of θ_1 and for all $\epsilon > 0$

$$\lim_{n_1\to\infty} P_{\theta^0}\left(||\tilde{\theta}_1^{(n_1)} - \theta_1^0|| > \epsilon\right) = 0,$$

for any $\theta_2, \theta_3, ..., \theta_m, \theta_i \in \Theta, i = 2, 3, ..., m.$

Proof: Let N_0 and ϵ be as in the proof of Theorem 4.2, and let $N^*_{\theta_1^1}, N^*_{\theta_1^2}, ..., N^*_{\theta_1^p}$ be an open cover of $C \setminus N_0$ with $E_{\theta^0} Z_{1j}(N_k^*) > 0$. Then

$$\begin{aligned} & P_{\theta^0}\left(||\tilde{\theta}_1^{(n_1)} - \theta_1^0|| \geq \epsilon\right) \\ \leq & \sum_{k=1}^{p} P_{\theta^0}\left(\tilde{\theta}_1^{(n_1)} \in N^*_{\theta_1^k}\right) + P_{\theta^0}\left(\tilde{\theta}_1^{(n_1)} \in C^c \cap \Theta\right) \\ \leq & \sum_{k=1}^{p} P_{\theta^0}\left(\tilde{\theta}_1^{(n_1)} \in N^*_{\theta_1^k}\right) \\ & + P_{\theta^0}\left(\frac{1}{n_1}\sum_{j=1}^{n_1} Z_{1j}(C^c \cap \Theta) + \frac{1}{n_1}\sum_{i=1}^{m}\sum_{j=1}^{n_i}(\lambda_i^{(\boldsymbol{n})} - w_i) Z_{ij}(C^c \cap \Theta) \leq 0\right). \end{aligned}$$

It follows from the proof of Theorem 4.2 that the first term of last expression goes to zero as n goes to infinity.

By the Weak Law of Large Numbers, we have

$$\frac{1}{n_1}\sum_{j=1}^{n_1} Z_{1j}(C^c \cap \Theta) \xrightarrow{P_{\theta^0}} E_{\theta^0}\left(\inf_{\theta_1 \in C^c \cap \Theta} log \frac{f_1(X_{1j}; \theta_1^0)}{f_1(X_{1j}; \theta_1)}\right) > 0, \text{ by equation (4.6).}$$

If we show that $\frac{1}{n_1}\sum_{i=1}^{m}\sum_{j=1}^{n_i}(\lambda_i^{(\boldsymbol{n})} - w_i) Z_{ij}(C^c \cap \Theta) \xrightarrow{P_{\theta^0}} 0$, then the second expression goes to zero. Consequently, the result of the theorem will follow.

Observe that

$$\frac{1}{n_1}\sum_{i=1}^{m}\sum_{j=1}^{n_i}(\lambda_i^{(\boldsymbol{n})} - w_i) Z_{ij}(C^c \cap \Theta) = \sum_{i=1}^{m} \frac{n_i}{n_1}(\lambda_i^{(\boldsymbol{n})} - w_i)\frac{1}{n_i}\sum_{j=1}^{n_i} Z_{ij}(C^c \cap \Theta).$$

By the Weak Law of Large Numbers, it follows that

$$\frac{1}{n_i}\sum_{j=1}^{n_i} Z_{ij}(C^c \cap \Theta) \xrightarrow{P_{\theta^0}} E_{\theta^0}\left(\inf_{\theta_1 \in C^c \cap \Theta} log \frac{f_1(X_{ij}; \theta_1^0)}{f_1(X_{ij}; \theta_1)}\right) \tag{4.7}$$

where $E_{\theta^0}\left(\inf_{\theta_1 \in C^c \cap \Theta} log \frac{f_1(X_{ij}; \theta_1^0)}{f_1(X_{ij}; \theta_1)}\right)$ is a finite number by the condition of this theorem.

By Assumption 4.5, it follows that

$$\frac{n_i}{n_1}(\lambda_i^{(\boldsymbol{n})} - w_i) \longrightarrow 0, \quad as \quad n_1 \to \infty. \tag{4.8}$$

Combining equation (4.7) and (4.8), we then have

$$\frac{1}{n_1}\sum_{i=1}^{m}\sum_{j=1}^{n_i}(\lambda_i^{(n)} - w_i)Z_{ij}(C^c \cap \Theta) \xrightarrow{P_{\theta^0}} 0. \ \diamond$$

4.1.2 Asymptotic Normality

To obtain asymptotic normality of the WLE, more restrictive conditions are needed. In particular, some conditions will be imposed upon the first and second derivative of the likelihood function.

For each fixed n, there may be many solutions to the likelihood equation even if the WLE is unique. However, as will be seen in the next theorem, there generally exist a sequence of solutions of this equation that are asymptotically normal.

Assume that θ_1 is a vector defined in R^p with p, a positive integer, *i.e.* $\theta_1 = (\theta_{11}, \theta_{12}, ..., \theta_{1p})$ and the true value of the parameter is $\theta_1^0 = (\theta_{11}^0, \theta_{12}^0, ..., \theta_{1p}^0)$. Write

$$\psi(x;\theta_1) = \frac{\partial}{\partial\theta_1} log f_1(x;\theta_1), \qquad \text{a } p \text{ column vector,}$$

and

$$\dot{\psi} = \frac{\partial}{\partial\theta_1}\psi(x;\theta_1), \qquad \text{a } p \text{ by } p \text{ matrix.}$$

Then, for any j, the *Fisher Information* matrix is defined as

$$I(\theta_1^0) = E_{\theta_1^0}\psi(X_{1j};\theta_1^0)\psi(X_{1j};\theta_1^0)^t.$$

Assuming that the partial derivatives can be passed under the integral sign in $\int f_1(x;\theta_1^0)d\nu(x) = 1$, we find that, for any j,

$$E_{\theta_1^0}\psi(X_{1j};\theta_1^0) = \int \left(\frac{\frac{\partial}{\partial\theta_1}f_1(x;\theta_1^0)}{f_1(x;\theta_1^0)}\right) f_1(x,\theta_1^0)d\nu(x) = \int \frac{\partial}{\partial\theta_1} f_1(x;\theta_1^0)d\nu(x) = 0, \quad (4.9)$$

so that $I(\theta_1^0)$ is in fact the covariance matrix of ψ,

$$I(\theta_1^0) = cov_{\theta_1^0}\psi(X_{1j};\theta_1^0).$$

If the second partial derivatives with respect to θ_1 can also be passed under the integral sign, then $\int(\partial^2/\partial\theta_1^2)f_1(x;\theta_1^0)d\nu(x) = 0$, and

$$I(\theta_1^0) = -E_{\theta_1^0}\dot{\psi}(X_1;\theta_1^0).$$

To simplify the notation, let

$$\dot{WL}_{n_1}(\boldsymbol{x};\theta_1) = \frac{\partial}{\partial\theta_1}log\mathrm{WL}(\boldsymbol{x};\theta_1) \quad and \quad \dot{WL}_{n_1}(\boldsymbol{x};\theta_1^0) = \frac{\partial}{\partial\theta_1}log\mathrm{WL}(\boldsymbol{x};\theta_1)|_{\theta_1=\theta_1^0}$$

In the next theorem we assume that the parameter space is an open subset of R^p.

Theorem 4.4 *Suppose:*

(1) for almost all x the first and second partial derivatives of $f_1(x;\theta)$ with respect to θ exist, are continuous in $\theta \in \Theta$, and may be passed through the integral sign in $\int f_1(x;\theta)d\nu(x) = 1$;

(2) there exist three functions $G_1(x)$, $G_2(x)$ and $G_3(x)$ such that for all $\theta_2, ..., \theta_m$, $E_{\theta^0}|G_l(X_{ij})|^2 \leq K_l < \infty, l = 1, 2, 3,$, $i = 1, \ldots, m$, and in some neighborhood of θ_1^0 each component of $\psi(x)$ (respectively $\dot{\psi}(x)$) is bounded in absolute value by $G_1(x)$ (respectively $G_2(x)$) uniformly in $\theta_1 \in \Theta$. Further,

$$\frac{\partial^3 log f_1(x;\theta_1)}{\partial\theta_{1k_1}\partial\theta_{1k_2}\partial\theta_{1k_3}},$$

$k_1, k_2, k_3 = 1, .., p$, is bounded by $G_3(x)$ uniformly in $\theta_1 \in \Theta$;

(3) $I(\theta_1^0)$ is positive definite.

Then there exists a sequence of roots $\tilde{\theta}_1^{(n_1)}$ of the weighted likelihood equation that is weakly consistent and

$$\sqrt{n_1}(\tilde{\theta}_1^{(n_1)} - \theta_1^0) \xrightarrow{D} N(0, (I(\theta_1^0))^{-1})), \quad as \quad n_1 \to \infty.$$

Proof: *1. Existence of consistent roots.*The proof of existence of consistent roots resembles the proof in Lehmann (1983, p 430-432). Let a be small enough so that

$S_a = \{\theta_1 : ||\theta_1 - \theta_1^0|| < a\} \subset \Theta$ and let

$$\begin{aligned} I_n(a) &= \{\mathbf{x} : log\ WL(\mathbf{x};\theta_1^0) > log\ WL(\mathbf{x},\theta_1^b) \quad \text{for all boundary points } \theta_1^b \ of\ S_a\} \\ &= \{\mathbf{x} : log\ WL(\mathbf{x};\theta_1^0) > \sup_{\theta_1^b \in S_a} log\ WL(\mathbf{x},\theta_1^b)\}. \end{aligned}$$

The set $I_n(a)$ is measurable since $log\ WL(\mathbf{x};\theta_1^0)$ is measurable and $\sup_{\theta_1^b \in S_a} logWL(\mathbf{x},\theta_1^b)$ is measurable by Proposition 4.1.

We will show that $P_{\theta^0}(\mathbf{X} \in I_n(a)) \to 1$ for all sufficiently small enough a. That is, for any given ϵ, there exist N_ϵ such that, for any $n > N_\epsilon$, we have $P_{\theta^0}(\mathbf{X} \in I_n(a)) > 1 - \epsilon$. This implies that $I_n(a)$ is not an empty set when $n > N_\epsilon$. To prove the claim, we expand the log weighted likelihood on the boundary of S_a about the true value θ_1^0 and divide it by n_1 to find

$$\frac{1}{n_1} log\ WL(\mathbf{x};\theta_1^b) - \frac{1}{n_1} log\ WL(\mathbf{x};\theta_1^0) = S_1 + S_2 + S_3$$

where

$$\begin{aligned} S_1 &= \frac{1}{n_1}\sum_{k_1=1}^{p} A_{k_1}(x)(\theta_{1k_1}^b - \theta_{1k_1}^0) \\ S_2 &= \frac{1}{2n_1}\sum_{k_1=1}^{p}\sum_{k_2=1}^{p}(\theta_{1k_1}^b - \theta_{1k_1}^0)B_{k_1k_2}(\theta_{1k_2}^b - \theta_{1k_2}^0) \\ S_3 &= \frac{1}{6n_1}\Big(\sum_{k_1=1}^{p}\sum_{k_2=1}^{p}\sum_{k_3=1}^{p}(\theta_{1k_1}^b - \theta_{1k_1}^0)(\theta_{1k_2}^b - \theta_{1k_2}^0)(\theta_{1k_3}^b - \theta_{1k_3}^0) \\ &\qquad \times \sum_{i=1}^{m}\sum_{j=1}^{n_i}\lambda_i^{(n)}\zeta_{k_1k_2k_3}(x_{ij})G_3(x_{ij})\Big) \end{aligned}$$

and

$$\begin{aligned} A_{k_1}(\mathbf{x}) &= \sum_{i=1}^{m}\sum_{j=1}^{n_i}\lambda_i^{(n)}\frac{\partial log f_1(x_{ij};\theta_1)}{\partial\theta_{1k_1}}|_{\theta_1=\theta_1^0}, \\ B_{k_1k_2}(\mathbf{x}) &= \sum_{i=1}^{m}\sum_{j=1}^{n_i}\lambda_i^{(n)}\frac{\partial log f_1(x_{ij};\theta_1)}{\partial\theta_{1k_1}\partial\theta_{1k_2}}|_{\theta_1=\theta_1^0}, \end{aligned}$$

and $|\zeta_{k_1k_2k_3}(x_{ij})| \leq 1$ by assumption.

By the Weak Law of Large Numbers and (4.9)

$$\frac{1}{n_1}\sum_{j=1}^{n_1}\frac{\partial log f_1(X_{1j};\theta_1)}{\partial\theta_{1k_1}}|_{\theta_1=\theta_1^0} \xrightarrow{P_{\theta^0}} 0 \quad as \quad n_1 \to \infty, \tag{4.10}$$

$$\frac{1}{n_1}\sum_{j=1}^{n_1}\frac{\partial^2 log f_1(X_{1j};\theta_1)}{\partial\theta_{1k_1}\partial\theta_{1k_2}}|_{\theta_1=\theta_1^0} \xrightarrow{P_{\theta^0}} -I_{k_1k_2}(\theta_1^0) \quad as \quad n_1 \to \infty \tag{4.11}$$

where $I_{k_1k_2}(\theta_1^0)$ is the (k_1, k_2) element of the information matrix $I(\theta_1^0)$. By Lemma 4.2, we then have

$$\frac{1}{n_1}\sum_{i=1}^{m}\sum_{j=1}^{n_i}(\lambda_i^{(\boldsymbol{n})} - w_i)\frac{\partial log f_1(X_{ij};\theta_1)}{\partial\theta_{1k_1}}|_{\theta_1=\theta_1^0} \xrightarrow{P_{\theta^0}} 0 \quad as \quad n_1 \to \infty. \tag{4.12}$$

$$\frac{1}{n_1}\sum_{i=1}^{m}\sum_{j=1}^{n_i}(\lambda_i^{(\boldsymbol{n})} - w_i)\frac{\partial^2 log f_1(X_{ij};\theta_1)}{\partial\theta_{1k_1}\partial\theta_{1k_2}}|_{\theta_1=\theta_1^0} \xrightarrow{P_{\theta^0}} 0 \quad as \quad n_1 \to \infty. \tag{4.13}$$

$$\frac{1}{n_1}\sum_{i=1}^{m}\sum_{j=1}^{n_i}(\lambda_i^{(\boldsymbol{n})} - w_i)G_3(X_{ij}) \xrightarrow{P_{\theta^0}} 0 \quad as \quad n_1 \to \infty. \tag{4.14}$$

To prove that the maximum difference $\frac{1}{n_1}log\ WL(\mathbf{x};\theta_1^b) - \frac{1}{n_1}log\ WL(\mathbf{x};\theta_1^0)$ over all the boundary points θ_1^b of S_a is negative with P_{θ^0} probability tending to 1 for sufficiently small a, we will show that, with P_{θ^0} probability tending to 1, the maximum of S_2 for all the boundary points θ_1^b of S_a is negative while $|S_1|$ and $|S_3|$ are small compared to $|S_2|$.

We begin with S_1. Observe that

$$\frac{1}{n_1}A_{k_1}(\mathbf{x}) = \frac{1}{n_1}\sum_{j=1}^{n_1}\frac{\partial log f_1(x_{1j};\theta_1)}{\partial\theta_{1k_1}}|_{\theta_1=\theta_1^0} + \frac{1}{n_1}\sum_{i=1}^{m}\sum_{j=1}^{n_i}(\lambda_i^{(\boldsymbol{n})} - w_i)\frac{\partial log f_1(x_{ij};\theta_1)}{\partial\theta_{1k_1}}|_{\theta_1=\theta_1^0}.$$

By (4.10) and (4.12), it then follows that, for any $\theta_2, ..., \theta_m$,

$$\frac{1}{n_1}A_{k_1}(\mathbf{X}) \xrightarrow{P_{\theta^0}} 0 \quad as \quad n_1 \to \infty. \tag{4.15}$$

Further, for any boundary point θ_1^b such that $||\theta_1^b - \theta_1^0|| = a$, we have

$$|S_1| \leq \frac{a}{n_1}\sum_{k_1=1}^{p}|A_{k_1}(\mathbf{X})|.$$

For any given a, it follows from (4.15) that with P_{θ^0} probability tending to 1

$$\frac{1}{n_1}\sum_{k_1=1}^{p}|A_{k_1}(\mathbf{X})| < pa^2 \tag{4.16}$$

and (4.16) then gives

$$|S_1| < pa^3 \tag{4.17}$$

with P_{θ^0} probability tending to 1.

Next consider

$$\begin{aligned} 2S_2 = &-\sum_{k_1=1}^{p}\sum_{k_2=1}^{p}(\theta_{1k_1}-\theta^0_{1k_1})I_{k_1k_2}(\theta^0_1)(\theta_{1k_2}-\theta^0_{1k_2}) \\ &+\sum_{k_1=1}^{p}\sum_{k_2=1}^{p}(\theta_{1k_1}-\theta^0_{1k_1})\big(\frac{1}{n_1}B_{k_1k_2}+I_{k_1k_2}(\theta^0_1)\big)(\theta_{1k_2}-\theta^0_{1k_2}). \end{aligned} \tag{4.18}$$

For the second term in the above expression, consider

$$\begin{aligned} \frac{1}{n_1}B_{k_1k_2}+I_{k_1k_2}(\theta^0_1) \quad &= \frac{1}{n_1}\sum_{j=1}^{n_1}\frac{\partial^2 log f_1(X_{1j};\theta_1)}{\partial\theta_{1k_1}\partial\theta_{1k_2}}|_{\theta_1=\theta^0_1}+I_{k_1k_2}(\theta^0_1) \\ &+\frac{1}{n_1}\sum_{i=1}^{m}\sum_{j=1}^{n_i}(\lambda_i^{(n)}-w_i)\frac{\partial^2 log f_1(X_{ij};\theta_1)}{\partial\theta_{1k_1}\partial\theta_{1k_2}}|_{\theta_1=\theta^0_1}. \end{aligned}$$

By (4.11) and (4.13), it then follows that, for any k_1 and k_2,

$$\frac{1}{n_1}B_{k_1k_2}+I_{k_1k_2}(\theta^0_1)\xrightarrow{P_{\theta^0}}0 \quad as \;\; n_1\to\infty. \tag{4.19}$$

Thus, for any boundary points such that $||\theta^b_1-\theta^0_1|| = a$, we have

$$|\sum_{k_1=1}^{p}\sum_{k_2=1}^{p}(\theta_{1k_1}-\theta^0_{1k_1})(\theta_{1k_2}-.\theta^0_{1k_2})| < p^2a^2. \tag{4.20}$$

By equation (4.19) and (4.20), it follows that, for given a

$$|\sum_{k_1=1}^{p}\sum_{k_2=1}^{p}(\theta_{1k_1}-\theta^0_{1k_1})\big(\frac{1}{n_1}B_{k_1k_2}+I_{k_1k_2}(\theta^0_1)\big)(\theta_{1k_2}-\theta^0_{1k_2})| < p^2a^3 \tag{4.21}$$

with P_{θ^0} probability tending to 1.

Let us examine the first term in (4.18). Since $I(\theta_1^0)$ is a symmetric and positive definite matrix, there exits a matrix $B_{p\times p}$ such that $B^tB = Diag(1,1,...,1)$ the identity matrix and $-I(\theta_1^0) = B^tDiag(\delta_1,\delta_2,...,\delta_p)B$ where $\delta_i < 0, i = 1,2,...,m$. Let $\boldsymbol{\xi} = (\xi_1,\xi_2,...,\xi_p)^t = B(\theta_{11}^b - \theta_{11}^0, \theta_{12}^b - \theta_{12}^0, ..., \theta_{1p}^b - \theta_{1p}^0)^t$. Then we have $||\boldsymbol{\xi}||^2 = ||\theta_1^b - \theta_1^0||^2 = a^2$. It follows that

$$-\sum_{k_1=1}^{p}\sum_{k_2=1}^{p}(\theta_{1k_1}^b - \theta_{1k_1}^0)I_{k_1k_2}(\theta_1^0)(\theta_{1k_2}^b - \theta_{1k_2}^0) = \sum_{l=1}^{m}\delta_l\xi_l^2 \le \delta^* a^2 \tag{4.22}$$

where $\delta^* = \max\{\delta_1,\delta_2,...,\delta_p\} < 0$. We see that there exist a_0 such that $\delta^* + p^2a_0 \le 0$. Combining (4.21) and (4.22), it follows that with P_{θ^0} probability tending to 1 there exists $c > 0$ such that for $a < a_0$

$$S_2 < -ca^2. \tag{4.23}$$

Note that $|\zeta_{k_1k_2k_3}(x)| \le 1$. Thus for S_3 we only consider

$$\begin{aligned}&\frac{1}{n_1}\sum_{i=1}^{m}\sum_{j=1}^{n_1}\lambda_i^{(\boldsymbol{n})}G_3(x_{ij})\\ &= \frac{1}{n_1}\sum_{j=1}^{n_i}G_3(x_{1j}) + \frac{1}{n_1}\sum_{i=1}^{m}\sum_{j=1}^{n_i}(\lambda_i^{(\boldsymbol{n})} - w_i)G_3(x_{ij}).\end{aligned}$$

By the Weak Law of Large Numbers with probability tending to 1,

$$\left|\frac{1}{n_1}\sum_{j=1}^{n_1}G_3(x_{1j})\right| < 2(1+K_3) \tag{4.24}$$

where we use the inequality $|EZ| \le E|Z| \le 1 + E|Z|^2$ for any random variable Z. By (4.14), it follows that with P_{θ^0} probability tending to 1

$$\left|\frac{1}{n_1}\sum_{i=1}^{m}\sum_{j=1}^{n_i}(\lambda_i^{(\boldsymbol{n})} - w_i)G_3(X_{ij})\right| < 1 + K_3. \tag{4.25}$$

Hence by (4.24) and (4.25) , for any given a and θ_1^b such that $||\theta_1^b - \theta_1^0|| = a$,

$$|S_3| < \frac{p^3}{2}(1+K_3)a^3. \tag{4.26}$$

Finally combining (4.17), (4.23) and (4.26), we then have, with P_{θ^0} probability tending to 1, for $a < a_0$,

$$\max_{\theta_1^b \in S_a}(S_1 + S_2 + S_3) < -ca^2 + \big(p + \frac{p^3}{2}(1 + K_3)\big)a^3$$

which is less than zero if $a < c/\big(p + \frac{p^3}{2}(1 + K_3)\big)$. This completes the proof of our claim that for any sufficiently small a the P_{θ^0} probability tends to 1 that

$$\max_{\theta_1^b \in S_a} log\ WL(\boldsymbol{x}; \theta_1^b) < log\ WL(\boldsymbol{x}, \theta_1^0).$$

For any such a and $\mathbf{x} \in I_n(a)$, it follows that there exists at least one point $\tilde{\theta}_1^{(n_1)}$ with $||\tilde{\theta}_1^{(n_1)} - \theta_1^0|| \leq a$ at which $WL(\boldsymbol{x}; \theta_1)$ has a local maximum, $\frac{\partial}{\partial \theta_1} WL_{n_1}(\boldsymbol{x}; \theta_1)|_{\theta_1 = \tilde{\theta}_1^{(n_1)}} = 0$. Since $P_{\theta^0}\left(\mathbf{X} \in I_n(a)\right) \longrightarrow 1$ as $n_1 \to \infty$ for all sufficiently small a, it then follows that for such fixed $a > 0$, there exists a sequence of roots $\tilde{\theta}_1^{(n_1)}(a)$ such that $P_{\theta^0}\left(||\tilde{\theta}_1^{(n_1)}(a) - \theta_1^0|| < a\right) \to 1$.

It remains to show that we can determine such a sequence of roots, which does not depend on a. Let $\theta_1^{(n_1)^*}$ be the closest root to θ_1^0 among all the roots to the likelihood equation for every fixed n_1. This closest root always exists. The reason for this can be seen as follows. If there is a finite number of roots within the closed ball $||\theta_1 - \theta_1^0|| < a$, we can always find such a root. If there are infinitely many roots in that sphere which is compact, then there exists a convergent sequence of roots inside the sphere $\tilde{\theta}_1^{(k)}$ such that $\lim_{k \to \infty} ||\tilde{\theta}_1^{(k)} - \theta_1^0|| = \inf_{\theta \in \mathcal{V}_{n_1}} ||\theta_1 - \theta_1^0||$, where $\mathcal{V}_{n_1}$ is the set of all the roots to the likelihood equation. Then the closest root exists since the limit of this sequence of roots $\tilde{\theta}_1^{(k)}$ is again a root by the assumed continuity of the $\frac{\partial}{\partial \theta_1} WL_{n_1}(\theta_1)$. Thus $\theta_1^{(n_1)^*}$ does not depend on a and $P_{\theta^0}\left(||\theta_1^{(n_1)^*} - \theta_1^0|| < a\right) \to 1$.

2. Asymptotic Normality. Expand $\frac{\partial}{\partial \theta_1} log\ \mathrm{WL}(\boldsymbol{x}; \theta_1)$as

$$\dot{WL}_{n_1}(\boldsymbol{x}; \theta_1) = \dot{WL}_{n_1}(\boldsymbol{x}; \theta_1^0) + \int_0^1 \sum_{i=1}^{m} \sum_{j=1}^{n_i} \lambda_i^{(n)} \dot{\psi}(x_{ij}; \theta_1^0 + t(\theta_1 - \theta_1^0)) dt (\theta_1 - \theta_1^0),$$

where $\dot{WL}_{n_1}(\boldsymbol{x};\theta_1^0) = \sum_{i=1}^{m}\sum_{j=1}^{n_i}\lambda_i^{(\boldsymbol{n})}\psi(x_{ij};\theta_1^0)$.

Now let $\theta_1 = \tilde{\theta}_1^{(n_1)}$, where $\tilde{\theta}_1^{(n_1)}$ is any weakly consistent sequence of roots satisfying $\dot{WL}_{n_1}(\boldsymbol{x};\tilde{\theta}_1^{(n_1)}) = 0$, and divide by $\sqrt{n_1}$ to get:

$$\frac{1}{\sqrt{n_1}}\dot{WL}_{n_1}(\boldsymbol{x};\theta_1^0) = B_{n_1}\sqrt{n_1}(\tilde{\theta}_1^{(n_1)} - \theta_1^0), \tag{4.27}$$

where

$$B_{n_1} = -\frac{1}{n_1}\int_0^1 \sum_{i=1}^{m}\sum_{j=1}^{n_i}\lambda_i^{(\boldsymbol{n})}\dot{\psi}(x_{ij};\theta_1^0 + t(\tilde{\theta}_1^{(n_1)} - \theta_1^0))dt.$$

Note that

$$\begin{aligned}\dot{WL}_{n_1}(\boldsymbol{x};\theta_1^0) &= \sum_{i=1}^{m}\sum_{j=1}^{n_i} w_i\psi(X_{ij};\theta_1^0) + \sum_{i=1}^{m}\sum_{j=1}^{n_i}(\lambda_i^{(\boldsymbol{n})} - w_i)\psi(X_{ij};\theta_1^0)\\ &= \sum_{j=1}^{n_1}\psi(X_{1j};\theta_1^0) + \sum_{i=1}^{m}\sum_{j=1}^{n_i}(\lambda_i^{(\boldsymbol{n})} - w_i)\psi(X_{ij};\theta_1^0).\end{aligned}$$

By (4.27), it follows that

$$\frac{1}{\sqrt{n_1}}\sum_{j=1}^{n_1}\psi(X_{1j};\theta_1^0) + \frac{1}{\sqrt{n_1}}\sum_{i=1}^{m}\sum_{j=1}^{n_i}(\lambda_i^{(\boldsymbol{n})} - w_i)\psi(X_{ij};\theta_1^0) = B_{n_1}\sqrt{n_1}(\tilde{\theta}_1^{(n_1)} - \theta_1^0).$$

From the Central Limit Theorem, because $E_{\theta^0}\psi(X_{1j};\theta_1^0) = 0$ and $cov_{\theta_1^0}\psi(X_{1j};\theta_1^0) = I(\theta_1^0)$, we find that

$$\frac{1}{\sqrt{n_1}}\sum_{j=1}^{n_1}\psi(X_{1j};\theta_1^0) \xrightarrow{D} Z^* \sim N(\underline{0}, I(\theta_1^0)) \ [P_{\theta^0}].$$

If we show $\frac{1}{\sqrt{n_1}}\sum_{i=1}^{m}\sum_{j=1}^{n_i}(\lambda_i^{(\boldsymbol{n})} - w_i)\psi(X_{ij};\theta_1^0) \xrightarrow{P_{\theta^0}} 0$ and $B_{n_1} \xrightarrow{P_{\theta^0}} I(\theta_1^0)$, then by the multivariate version of Slutsky's theorem (see for example, Sen and Singer (1993), p. 130.) we have

$$\frac{1}{\sqrt{n_1}}(\tilde{\theta}_1^{(n_1)} - \theta_1^0) = B_{n_1}^{-1}\frac{1}{\sqrt{n_1}}\dot{WL}_{n_1}^1 \xrightarrow{D} I(\theta_1^0)^{-1}Z^* \sim N(\underline{0}, I(\theta_1^0)^{-1})$$

Now we prove

(i) $\frac{1}{\sqrt{n_1}} \sum_{i=1}^{m} \sum_{j=1}^{n_i} (\lambda_i^{(\boldsymbol{n})} - w_i)\psi(X_{ij}; \theta_1^0) \xrightarrow{P_{\theta^0}} \underline{0}$.

Let $\dot{V}_{n_1} = \sum_{i=1}^{m} \sum_{j=1}^{n_i} (\lambda_i^{(\boldsymbol{n})} - w_i)\psi(X_{ij}; \theta_1^0)$.

We then have

$$\begin{aligned} P_{\theta^0}\left(\frac{1}{\sqrt{n_1}}||\dot{V}_{n_1}|| > \epsilon\right) &\leq \frac{4K_1^2}{\epsilon^2 n_1} \sum_{i=1}^{m}\sum_{i'=1}^{m}\sum_{j=1}^{n_i}\sum_{j'=1}^{n_{i'}} |\lambda_i^{(\boldsymbol{n})} - w_i|\,|\lambda_{i'}^{(\boldsymbol{n})} - w_{i'}| \\ &\leq O\left(\frac{1}{n_1^{\delta}}\right) \longrightarrow 0, \quad as\ n_1 \to \infty, \end{aligned}$$

by hypothesis (2) of this theorem.

(ii) $B_{n_1} \xrightarrow{P_{\theta^0}} I(\theta_1^0)$ *as* $n_1 \longrightarrow \infty$.

Let $B_{n_1} = B_{n_1}^I + B_{n_1}^{II}$, where

$$\begin{aligned} B_{n_1}^I &= -\frac{1}{n_1}\int_0^1 \sum_{j=1}^{n_1} \dot{\psi}(X_{1j}; \theta_1^0 + t(\tilde{\theta}_1^{(n_1)} - \theta_1^0))dt, \\ B_{n_1}^{II} &= -\frac{1}{n_1}\int_0^1 \sum_{i=1}^{m}\sum_{j=1}^{n_i} (\lambda_i^{(\boldsymbol{n})} - w_i)\dot{\psi}(X_{ij}; \theta_1^0 + t(\tilde{\theta}_1^{(n_1)} - \theta_1^0))dt. \end{aligned}$$

First, we prove $B_{n_1}^I \xrightarrow{P_{\theta^0}} I(\theta_1^0)$ *as* $n_1 \longrightarrow \infty$.

Let $S_\rho = \{\theta_1 : ||\theta_1 - \theta_1^0|| < \rho\}$. Note that $E_{\theta_1^0}\dot{\psi}(X_{1j}, \theta_1)$ is continuous in θ_1 by condition (1), so there is a $\rho > 0$ such that

$$||\theta_1 - \theta_1^0|| < \rho \Rightarrow |E_{\theta^0}\dot{\psi}(X_{1j}; \theta_1) + I(\theta_1^0)| < \epsilon. \tag{4.28}$$

For any $t \in \mathcal{R}$ such that $0 \leq t \leq 1$, then

$$||\tilde{\theta}_1^{(n_1)} - \theta_1^0|| < \rho \Rightarrow ||\theta_1^0 + t(\tilde{\theta}_1^{(n_1)} - \theta_1^0) \ - \ \theta_1^0|| < \rho. \tag{4.29}$$

By equation (4.28) and (4.29), we then have

$$P_{\theta^0}\big(||\tilde{\theta}_1^{(n_1)} - \theta_1^0|| < \rho\big) \leq P_{\theta^0}\big(|E_{\theta^0}\dot{\psi}\big(X_{1j}; \theta_1^0 + t(\tilde{\theta}_1^{(n_1)} - \theta_1^0)\big) + I(\theta_1^0)| < \epsilon\big).$$

Note that $P_{\theta^0}\big(||\tilde{\theta}_1^{(n_1)} - \theta_1^0|| < \rho\big) \longrightarrow 1$. We then have

$$P_{\theta^0}\big(|E_{\theta^0}\dot{\psi}\big(X_{1j}; \theta_1^0 + t(\tilde{\theta}_1^{(n_1)} - \theta_1^0)\big) + I(\theta_1^0)| < \epsilon\big) \longrightarrow 1 \ \ as \ \ n_1 \longrightarrow \infty. \tag{4.30}$$

This result implies that

$$P_{\theta^0}\big(\theta_1^0 + t(\tilde{\theta}_1^{(n_1)} - \theta_1^0) \in S_\rho\big) \longrightarrow 1 \quad as \ \ n_1 \to \infty. \tag{4.31}$$

From the Uniform Strong Law of Large Numbers, Theorem 16(a) in Ferguson (1991), with P_{θ^0} probability 1, there is an integer N_I such that

$$n_1 > N_I \ \ \Rightarrow \ \ \sup_{\theta_1 \in S_\rho} |\frac{1}{n_1}\sum_{j=1}^{n_1} \dot{\psi}(X_{1j}, \theta_1) - E_{\theta^0}\dot{\psi}(X_{11}, \theta_1)| < \epsilon. \tag{4.32}$$

Then, assuming N is so large that $n_1 > N_I \Rightarrow ||\tilde{\theta}_1^{(n_1)} - \theta_1^0|| < \rho$, then

$$\begin{aligned}
\big|B_{n_1}^I - I(\theta_1^0)\big| \ &\leq \ \int_0^1 \left|\frac{1}{n_1}\sum_{j=1}^{n_1} \dot{\psi}\left(X_{1j}; \theta_1^0 + t(\tilde{\theta}_1^{(n_1)} - \theta_1^0)\right) + I(\theta_1^0)\right| dt \\
&= \ \int_0^1 \Big|\frac{1}{n_1}\sum_{j=1}^{n_1} \dot{\psi}\left(X_{1j}; \theta_1^0 + t(\tilde{\theta}_1^{(n_1)} - \theta_1^0)\right) - E_{\theta^0}\dot{\psi}\left(X_{1j}; \theta_1^0 + t(\tilde{\theta}_1^{(n_1)} - \theta_1^0)\right) \\
&\qquad + E_{\theta^0}\dot{\psi}\big(X_{1j}; \theta_1^0 + t(\tilde{\theta}_1^{(n_1)} - \theta_1^0)\big) + I(\theta_1^0)\Big| dt \\
&\leq \ \int_0^1 \Big(\sup_{\theta_1 \in S_\rho} \left|\frac{1}{n_1}\sum_{j=1}^{n_1} \dot{\psi}(X_{1j}; \theta_1) - E_{\theta^0}\dot{\psi}(X_{1j}; \theta_1)\right| \\
&\qquad + \sup_{0 \leq t \leq 1} \left|E_{\theta^0}\dot{\psi}\left(X_{1j}; \theta_1^0 + t(\tilde{\theta}_1^{(n_1)} - \theta_1^0)\right) + I(\theta_1^0)\right| \Big) dt \\
&\xrightarrow{P_{\theta^0}} \ 0 \ \ as \ \ n_1 \longrightarrow \infty
\end{aligned}$$

by equation (4.30), (4.31) and (4.32).

Secondly, we prove $B_{n_1}^{II} \xrightarrow{P_{\theta^0}} \underline{0} \ \ as \ \ n_1 \longrightarrow \infty$. By Lemma 4.2, every component of $B_{n_1}^{II}$ goes to 0 in probability. Thus

$$\big|B_{n_1}^{II}\big| \leq \int_0^1 \frac{1}{n_1}\sum_{i=1}^{m}\sum_{j=1}^{n_i} \left|(w_i - \lambda_i^{(n)})\dot{\psi}(X_{ij}; \theta_1^0 + t(\tilde{\theta}_1^{(n_1)} - \theta_1^0))\right| dt \xrightarrow{P_{\theta^0}} 0 \ \ as \ \ n_1 \longrightarrow \infty.$$

This completes the proof. $\diamond$

REMARK: If there is a unique root of the weighted likelihood equation for every n, as in many applications, this sequence of roots will be consistent and asymptotically normal. Small *et. al.* (2000) discuss the multiple root problems in estimation and propose various methods for selecting among the roots if the solution to the estimating equation is not unique.

4.1.3 Strong Consistency

We prove strong consistency of the WLE in this subsection. Recall that

$$\frac{1}{n_1}S_{n_1}(\mathbf{X},\theta_1)=\frac{1}{n_1}\sum_{i=1}^{m}\sum_{j=1}^{n_i}(w_i-\lambda_i^{(\boldsymbol{n})})log\frac{f_1(X_{ij};\theta_1^0)}{f_1(X_{ij};\theta_1)}.$$

To prove the strong consistency, we prove the following lemma:

Lemma 4.3 *Under Assumptions 4.1- 4.5*

$$\sup_{\theta_1\in\Theta}\left|\frac{1}{n_1}S_{n_1}(\mathbf{x},\theta_1)\right|\longrightarrow 0,\quad a.s.\ [P_{\theta^0}], \tag{4.33}$$

for any $\theta_2,\theta_3,...,\theta_m,\theta_i\in\Theta, i=2,3,...,m.$

Proof: By Lemma 4.2, we then have

$$P_{\theta^0}\left(\frac{1}{n_1}\sum_{i=1}^{m}\sum_{j=1}^{n_i}B_{ij}>\epsilon\right)\leq O\left(\frac{1}{n_1^{1+\delta}}\right).$$

where $B_{ij}=(w_i-\lambda_i^{(\boldsymbol{n})})A_{ij}$ and $A_{ij}=\sup_{\theta_1\in\Theta}\left|log\frac{f_1(X_{ij},\theta_1^0)}{f_1(X_{ij},\theta_1)}\right|$.

We then have

$$\sum_{n_1=1}^{\infty}P_{\theta^0}\left(\frac{1}{n_1}\sum_{i=1}^{m}\sum_{j=1}^{n_i}|w_i-\lambda_i^{(\boldsymbol{n})}|A_{ij}>\epsilon\right)<\infty.$$

It follows by the Borel-Cantelli Lemma that,

$$\frac{1}{n_1}\sum_{i=1}^{m}\sum_{j=1}^{n_i}|w_i-\lambda_i^{(\boldsymbol{n})}|A_{ij}\longrightarrow 0,\quad a.s.\ [P_{\theta^0}].$$

By the definition of A_{ij}, it follows that

$$\sup_{\theta_1\in\Theta}\left|\frac{1}{n_1}S_{n_1}(\mathbf{x};\theta_1)\right|\leq\frac{1}{n_1}\sum_{i=1}^{m}\sum_{j=1}^{n_i}|w_i-\lambda_i^{(\boldsymbol{n})}|A_{ij}.$$

Therefore, we have

$$\sup_{\theta_1 \in \Theta} \left| \frac{1}{n_1} S_{n_1}(\mathbf{x}; \theta_1) \right| \longrightarrow 0, \quad a.s. \ [P_{\theta^0}]$$

for any $\theta_1 \neq \theta_1^0, \theta_2, \theta_3, ..., \theta_m, \theta_i \in \Theta, i = 1, 2, ..., m.$ $\diamond$

Theorem 4.5 *Suppose:*

(1) Θ *is compact;*

(2) $log f_1(x; \theta)$ *is upper semi-continuous in* θ *for all* x*;*

(3) there exists a function $K(x)$ *such that* $E_{\theta_1^0}|K(X_1)| < \infty$ *and* $log \frac{f_1(X_{1j};\theta_1)}{f_1(X_{1j};\theta_1^0)} \leq K(x)$*, for all* x *and* $\theta \in \Theta$*;*

Then for any sequence of maximum weighted likelihood estimates $\tilde{\theta}_1^{(n_1)}$ *of* θ_1*,*

$$\tilde{\theta}_1^{(n_1)} \longrightarrow \theta_1^0 \ \ a.s. \ [P_{\theta_1^0}]$$

for any $\theta_2, \theta_3, ..., \theta_m, \theta_i \in \Theta, i = 2, 3, ..., m.$

Proof: Let θ_1 be the parameter of interest and let

$$\begin{aligned} W'_{n_1} &= \frac{1}{n_1} log \mathrm{WL}(\theta_1) - \frac{1}{n_1} log \mathrm{WL}(\theta_1^0) \\ &= \frac{1}{n_1} \sum_{j=1}^{n_1} \sum_{i=1}^{m} \lambda_i^{(\boldsymbol{n})} (\, log f_1(X_{ij}; \theta_1) - log f_1(X_{ij}; \theta_1^0) \,) \\ &= \frac{1}{n_1} \sum_{j=1}^{n_1} U(X_{1j}, \theta_1) + \frac{1}{n_1} S_{n_1}(\mathbf{X}, \theta_1) \end{aligned}$$

where $U(X_{1j}, \theta_1) = log \frac{f_1(X_{1j};\theta_1)}{f_1(X_{1j};\theta_1^0)}$.

Let $\rho > 0$ and $D = (\theta_1 \in \Theta : ||\theta_1 - \theta_1^0|| \geq \rho)$. Then D is compact by condition (1). We then have, (c.f. Ferguson 1996, p. 109)

$$P_{\theta^0} \left(\limsup_{n_1 \to \infty} \sup_{\theta_1 \in D} \frac{1}{n_1} \sum_{j=1}^{n_1} U(X_{1j}, \theta_1) \leq \sup_{\theta_1 \in D} \mu(\theta_1) \right) = 1, \tag{4.34}$$

where $\mu(\theta_1) = \int log \frac{f_1(x;\theta_1)}{f_1(x;\theta_1^0)} f_1(x; \theta_1^0) d\nu(x) < 0$ for $\theta_1 \in D$ by Lemma 4.1.

Furthermore, $\mu(\theta_1)$ is upper semi-continuous since

$$\mu(\theta_1) \geq \int \limsup_{\theta_1^{(n)} \to \theta_1} log \frac{f_1(x;\theta_1^{(n)})}{f_1(x;\theta_1^0)} f_1(x;\theta_1^0) d\nu(x) \geq \limsup_{\theta_1^{(n)} \to \theta_1} \int log \frac{f_1(x;\theta_1)}{f_1(x;\theta_1^0)} f_1(x;\theta_1^0) d\nu(x).$$

Hence $\mu(\theta_1)$ achieves its maximum value on D. Let $\delta = \sup_{\theta_1 \in D} \mu(\theta_1)$; then $\delta < 0$ by Lemma 4.1 and Assumption 4.3. Then by Lemma 4.3, with P_{θ^0} measure 1, there exists an N_1 such that, for all $n_1 > N_1$,

$$\sup_{\theta_1 \in D} \left| \frac{1}{n_1} S_{n_1}(\mathbf{X}, \theta_1) \right| < -\delta/2.$$

Observe that

$$\sup_{\theta_1 \in D} \left(\frac{1}{n_1} \sum_{j=1}^{n_1} U(X_{1j};\theta_1) + \frac{1}{n_1} S_{n_1}(\mathbf{X},\theta_1) \right) \leq \sup_{\theta_1 \in D} \frac{1}{n_1} \sum_{j=1}^{n_1} U(X_{1j};\theta_1) + \sup_{\theta_1 \in D} \left| \frac{1}{n_1} S_{n_1}(\mathbf{X},\theta_1) \right|.$$

It follows that, with P_{θ^0} measure 1, there exists an N such that for all $n_1 > N$,

$$\sup_{\theta_1 \in D} \left(\frac{1}{n_1} \sum_{j=1}^{n_1} U(X_{1j};\theta_1) + \frac{1}{n_1} S_{n_1}(\mathbf{X},\theta_1) \right) \leq \delta/2 < 0.$$

But, for all $n_1 > N$,

$$\frac{1}{n_1} \sum_{j=1}^{n_1} U(X_{1j};\tilde{\theta}_1^{(n)}) + \frac{1}{n_1} S_{n_1}(\mathbf{X};\tilde{\theta}_1^{(n)}) = \sup_{\theta_1 \in \Theta} \left(\frac{1}{n_1} \sum_{j=1}^{n_1} U(X_{1j};\theta_1) + \frac{1}{n_1} S_{n_1}(\mathbf{X},\theta_1) \right) \geq 0.$$

since

$$\frac{1}{n_1} \sum_{j=1}^{n_1} U(X_{1j};\theta_1^0) + \frac{1}{n_1} S_{n_1}(\mathbf{X};\theta_1^0) = 0.$$

This implies that the WLE, $\tilde{\theta}_1^{(n_1)} \in D^c$ for $n_1 > N$; that is $||\tilde{\theta}_1^{(n_1)} - \theta_1|| < \rho$. Since ρ is arbitrary, the theorem follows. $\diamond$.

The proof of the above theorem resembles the famous proof given by Wald (1949) which established the strong consistency of MLE except that we have to deal with the extra term, $\frac{1}{n_1} S_{n_1}(\mathbf{X},\theta_1)$.

Again, a slightly different condition is required if Θ is not compact.

Theorem 4.6 *Suppose:*

(1) there is a compact subset C of Θ such that $\theta_1^0 \in C$ and

$$E_{\theta^0} \sup_{\theta_1 \in C^c \cap \Theta} log \frac{f_1(X_{ij};\theta_1)}{f_1(X_{ij};\theta_1^0)} < 0;$$

(2) there exist a function $K(x)$ such that $E_{\theta_1^0}|K(X)| < \infty$ and $log \frac{f_1(x;\theta_1)}{f_1(x;\theta_1^0)} \leq K(x)$, for all x and $\theta \in C$;

(3) $log f_1(x;\theta)$ is upper semi-continuous in θ for all x;

Then for any sequence of maximum weighted likelihood estimates $\tilde{\theta}_1^{(n_1)}$ of θ_1,

$$\tilde{\theta}_1^{(n_1)} \longrightarrow \theta_1^0, \quad a.s. \ [P_{\theta^0}]$$

for any $\theta_2, \theta_3, ..., \theta_m, \theta_i \in \Theta, i = 2, 3, ..., m$.

Proof:

Let $D = (\theta_1 : ||\theta_1 - \theta_1^0|| \geq \rho)$ as in the proof of Theorem 4.5 such that $C \cap D \neq \phi$. It follows that $C \cap D$ is also compact. It follows from the proof of Theorem 4.5 that, with P_{θ^0} measure 1, there exits an N_1 such that, for all $n_1 > N_1$,

$$\sup_{\theta \in C \cap D} \left(\frac{1}{n_1} \sum_{j=1}^{n_1} U(X_{1j};\theta_1) + \frac{1}{n_1} S_{n_1}(\mathbf{X};\theta_1) \right) \leq \delta/2 < 0,$$

where $U(X_{1j};\theta_1) = \frac{f_1(X_{1j};\theta_1)}{f_1(X_{1j};\theta_1^0)}$.

Also, with P_{θ^0} measure 1, there exits an N_2 such that, for all $n_1 > N_2$,

$$\frac{1}{n_1} \sum_{j=1}^{n_1} \sup_{\theta_1 \in C^c \cap \Theta} U(X_{1j};\theta_1) < \delta \tag{4.35}$$

by the Strong Law of Large Numbers and the fact that $E_{\theta_1^0 \in C^c \cap \Theta} U(X_{1j};\theta_1) < 0$.

As in the proof of Lemma 4.3, it can be shown that

$$\frac{1}{n_1} \sup_{\theta_1 \in C^c \cap \Theta} S_{n_1}(\mathbf{X};\theta_1) \longrightarrow 0, \quad a.s. \ [P_{\theta^0}].$$

It follows that, with P_{θ^0} measure 1, there exist an N_3, such that, for all $n_1 > N_3$,

$$\frac{1}{n_1}\sum_{j=1}^{n_1}\sup_{\theta_1\in C^c\cap\Theta} U(X_{1j};\theta_1) + \frac{1}{n_1}\sup_{\theta_1\in C^c\cap\Theta} S_{n_1}(\mathbf{X};\theta_1) < \delta/2 < 0.$$

It implies that, with P_{θ^0} measure 1, for all $n_1 > N_3$,

$$\sup_{\theta_1\in C^c\cap\Theta}\frac{1}{n_1}\sum_{j=1}^{n_1} U(X_{1j};\theta_1) + \sup_{\theta_1\in C^c\cap\Theta}\frac{1}{n_1}S_{n_1}(\mathbf{X};\theta_1) < 0. \tag{4.36}$$

Therefore, it follows that, with P_{θ^0} measure 1, there exist an $N^* = max(N_2, N_3)$, such that for all $n > N^*$,

$$\sup_{\theta_1\in S}\left(\frac{1}{n_1}\sum_{j=1}^{n_1} U(X_{1j};\theta_1) + \frac{1}{n_1}S_{n_1}(\mathbf{X};\theta_1)\right) \le \delta < 0.$$

But

$$\frac{1}{n_1}\sum_{j=1}^{n_1} U(X_{1j};\tilde{\theta}_1^{(n_1)}) + \frac{1}{n_1}S_{n_1}(\mathbf{X};\tilde{\theta}_1^{(n_1)}) = \sup_{\theta_1\in\Theta}\left(\frac{1}{n_1}\sum_{j=1}^{n_1} U(X_{1j};\theta_1) + \frac{1}{n_1}S_{n_1}(\mathbf{X};\theta_1)\right) \ge 0.$$

since the sum is equal to 0 for $\theta_1 = \theta_1^0$. This implies that the WLE, $\tilde{\theta}_1 \in D^c$ for $n_1 > N^*$; that is, $||\tilde{\theta}_1^{(n_1)} - \theta_1|| < \rho$. Since ρ is arbitrary, the theorem follows. $\diamond$.

4.2 Asymptotic Properties of Adaptive Weights

At the practical level, we might want to let the relevant weight vector to be a function of the data. This section is concerned with the asymptotic properties of the WLE using adaptive weights. Assumption 4.1-4.4 are assumed to hold in this section.

4.2.1 Weak Consistency and Asymptotic Normality

In this subsection, we adopt the following additional condition:

Assumption 4.6 *(Weak Convergence Condition). Assume:*

(i) $\lim_{n_1\to\infty}\frac{n_i}{n_1} < \infty$, *for* $i = 1, 2, ..., m$;

(ii) the adaptive relevant weight vector $\lambda^{(n)}(\mathbf{X}) = (\lambda_1^{(n)}(\mathbf{X}), \lambda_2^{(n)}(\mathbf{X}), ..., \lambda_m^{(n)}(\mathbf{X}))^t$ *satisfies, for any* $\epsilon > 0$,

$$\lambda_i^{(n)}(\mathbf{X}) \xrightarrow{P_{\theta^0}} w_i, \quad as \ \ n_1 \to \infty,$$

where $(w_1, w_2, ..., w_m)^t \triangleq (1, 0, ..., 0)^t$.

Let

$$\frac{1}{n_1} S_{n_1}^A(\mathbf{X}, \theta_1) = \frac{1}{n_1} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \left(\lambda_i^{(n)}(\mathbf{X}) - w_i \right) log \frac{f_1(X_{ij}; \theta_1^0)}{f_1(X_{ij}; \theta_1)}.$$

We then have the following lemma.

Lemma 4.4 *If the adaptive relevance weight vector satisfies Assumption 4.6, then*

$$\frac{1}{n_1} S_{n_1}^A(\mathbf{X}, \theta_1) \xrightarrow{P_{\theta^0}} 0, \quad as \ \ n_1 \to \infty$$

for any $\theta_2, \theta_3, ..., \theta_m, \theta_i \in \Theta, i = 2, 3, ..., m$.

Proof: Let $T_i = \sum_{j=1}^{n_i} log \frac{f_1(X_{ij}; \theta_1^0)}{f_1(X_{ij}; \theta_1)}$, for $i = 1, 2, ..., m$. Then

$$\begin{aligned} \frac{1}{n_1} S_{n_1}^A &= \frac{1}{n_1} \sum_{i=1}^{m} \left(\lambda_i^{(n)}(\mathbf{X}) - w_i \right) T_i \\ &= \sum_{i=1}^{m} \frac{n_i}{n_1} \left(\lambda_i^{(n)}(\mathbf{X}) - w_i \right) \frac{1}{n_i} T_i. \end{aligned}$$

By the weak law of large numbers, for any $i = 1, 2, ..., m$,

$$\frac{1}{n_i} T_i = \frac{1}{n_i} \sum_{j=1}^{n_i} log \frac{f_1(X_{ij}; \theta_1^0)}{f_1(X_{ij}; \theta_1)} \xrightarrow{P_{\theta_0}} E_{\theta_0} log \frac{f_1(X_{ij}; \theta_1^0)}{f_1(X_{ij}; \theta_1)} < \infty.$$

It then follows that, for any $i = 1, 2, ..., m$,

$$\frac{n_i}{n_1} \left(\lambda_i^{(n)}(\mathbf{X}) - w_i \right) \frac{1}{n_i} T_i \xrightarrow{P_{\theta_0}} 0,$$

for any $\theta_2, \theta_3, ..., \theta_m, \theta_i \in \Theta, i = 2, 3, ..., m$, by Assumption 4.6. $\diamond$

We then have the following theorems:

Theorem 4.7 *For each θ_1^0, the true value of θ_1, and each $\theta_1 \neq \theta_1^0$,*

$$\lim_{n_1 \to \infty} P_{\theta^0} \left(\prod_{i=1}^{m} \prod_{j=1}^{n_i} f_1(X_{ij}; \theta_1^0)^{\lambda_i^{(n)}(\mathbf{X})} > \prod_{i=1}^{m} \prod_{j=1}^{n_i} f_1(X_{ij}; \theta_1)^{\lambda_i^{(n)}(\mathbf{X})} \right) = 1,$$

for any $\theta_2, \theta_3, ..., \theta_m, \theta_i \in \Theta, i = 2, 3, ..., m$.

Theorem 4.8 *Suppose that the conditions of Theorem 4.2 are satisfied. Then for any sequence of maximum weighted likelihood estimates $\tilde{\theta}_1^{(n_1)}$ of θ_1 constructed with adaptive weights $\lambda_i^{(n)}(\mathbf{X})$, and for all $\epsilon > 0$,*

$$\lim_{n_1 \to \infty} P_{\theta^0} \left(||\tilde{\theta}_1^{(n_1)} - \theta_1^0|| > \epsilon \right) = 0,$$

for any $\theta_2, \theta_3, ..., \theta_m, \theta_i \in \Theta, i = 2, 3, ..., m$.

Theorem 4.9 *Suppose that the conditions of Theorem 4.3 are satisfied. Then for any sequence of maximum weighted likelihood estimates $\tilde{\theta}_1^{(n_1)}$ of θ_1 constructed with adaptive weights $\lambda_i(\mathbf{X})$, and for all $\epsilon > 0$*

$$\lim_{n_1 \to \infty} P_{\theta^0} \left(||\tilde{\theta}_1^{(n_1)} - \theta_1^0|| > \epsilon \right) = 0,$$

for any $\theta_2, \theta_3, ..., \theta_m, \theta_i \in \Theta, i = 2, 3, ..., m$.

We remark that the proofs of Theorem 4.7 - 4.9 are identical to the proofs of Theorem 4.1 - 4.3 except that the fixed weights are replaced by adaptive weights and the utilization of Lemma 4.2 is replaced everywhere by Lemma 4.4.

We are now in a position to establish the asymptotic normality for the WLE constructed by adaptive weights. We assume that the parameter space is an open subset of R^p.

Theorem 4.10 *(Multi-dimensional) Suppose that the conditions of Theorem 4.4 are satisfied. Then there exists a sequence of roots of the weighted likelihood function based on adaptive weights $\tilde{\theta}_1^{(n_1)}$ that is weakly consistent and*

$$\sqrt{n_1} \left(\tilde{\theta}_1^{(n_1)} - \theta_1^0 \right) \xrightarrow{D} N\left(0, I(\theta_1^0)\right), \quad as \quad n_1 \to \infty.$$

4.2.2 Strong Consistency by Using Adaptive Weights

To establish the strong consistency of the WLE constructed by the adaptive weights, we need a condition that is stronger than Assumption 4.6. We hence assume the following condition:

Assumption 4.7 *(Strong Convergence Condition) Assume that:*

(i) $\lim_{n_1\to\infty} \frac{n_i}{n_1} < \infty$, *for* $i = 1, 2, ..., m$;

(ii) the adaptive relevant weight vector $\lambda^{(\boldsymbol{n})}(\mathbf{X}) = (\lambda_1^{(\boldsymbol{n})}(\mathbf{X}), \lambda_2^{(\boldsymbol{n})}(\mathbf{X}), ..., \lambda_m^{(\boldsymbol{n})}(\mathbf{X}))^t$ *satisfies*

$$\lambda_i^{(\boldsymbol{n})}(\mathbf{X}) \longrightarrow w_i, \quad a.s. \ [P_{\theta_1^0}],$$

where $(w_1, w_2, ..., w_m)^t \triangleq (1, 0, ..., 0)^t$.

Lemma 4.5 *If the adaptive relevance weight vector satisfies Assumption 4.7, then*

$$\sup_{\theta_1\in\Theta} \left| \frac{1}{n_1} S_{n_1}^A(\mathbf{x}, \theta_1) \right| \longrightarrow 0, \quad a.s. \ [P_{\theta_1^0}],$$

for any $\theta_2, \theta_3, ..., \theta_m, \theta_i \in \Theta, i = 2, 3, ..., m$.

Proof: Let $A_{ij} = \sup_{\theta_1\in\Theta} \left| log \frac{f_1(X_{ij};\theta_1^0)}{f_1(X_{ij};\theta_1)} \right|$, for $i = 1, 2, ..., m$.
By the Strong Law of Large Numbers, for any $i = 1, 2, ..., m$,

$$\frac{1}{n_i} \sum_{j=1}^{n_i} A_{ij} \longrightarrow E_{\theta_0} A_{i1} \ \ a.s. \ [P_{\theta_1^0}]$$

where $E_{\theta_0} A_{i1} = E_{\theta_0} \sup_{\theta_1\in\Theta} \left| log \frac{f_1(X_{1j};\theta_1^0)}{f_1(X_{i1};\theta_1)} \right| < \infty$. This implies that, for any $i = 1, 2, ..., m$,

$$\frac{n_i}{n_1} \left| \lambda_i^{(\boldsymbol{n})}(\mathbf{X}) - w_i \right| \left(\frac{1}{n_i} \sum_{j=1}^{n_i} A_{ij} \right) \longrightarrow 0, \quad a.s. \ [P_{\theta_1^0}]$$

by Assumption 4.7. Since

$$\sup_{\theta_1\in\Theta} \left| \frac{1}{n_1} S_{n_1}^A(\mathbf{x}, \theta_1) \right| \leq \sum_{i=1}^{m} \frac{n_i}{n_1} \left| \lambda_i^{(\boldsymbol{n})} - w_i \right| \left(\frac{1}{n_i} \sum_{j=1}^{n_i} A_{ij} \right),$$

it then follows that

$$\sup_{\theta_1 \in \Theta} \left| \frac{1}{n_1} S^A_{n_1}(\mathbf{x}, \theta_1) \right| \longrightarrow 0, \ \ a.s. \ \ [P_{\theta_1^0}].$$

This completes the proof.$\diamond$

We then have the following theorems:

Theorem 4.11 *Suppose the conditions of Theorem 4.5 are satisfied. Then for any sequence of maximum weighted likelihood estimates $\tilde{\theta}_1^{(n_1)}$ of θ_1 constructed by adaptive weights $\lambda_i(\mathbf{X})$,*

$$\tilde{\theta}_1^{(n_1)} \longrightarrow \theta_1^0 \ \ a.s. \ \ [P_{\theta_1^0}],$$

for any $\theta_2, \theta_3, ..., \theta_m, \theta_i \in \Theta, i = 2, 3, ..., m$.

Theorem 4.12 *Suppose the conditions of Theorem 4.6 are satisfied. Then for any sequence of maximum weighted likelihood estimates $\tilde{\theta}_1^{(n_1)}$ of θ_1 constructed by adaptive weights $\lambda_i(\mathbf{X})$,*

$$\tilde{\theta}_1^{(n_1)} \longrightarrow \theta_1^0, \ \ a.s. \ \ [P_{\theta^0}],$$

for any $\theta_2, \theta_3, ..., \theta_m, \theta_i \in \Theta, i = 2, 3, ..., m$.

4.3 Examples.

In this section we demonstrate the use of our theory in some examples.

4.3.1 Estimating a Univariate normal Mean.

Suppose X_{ij} are independent random variables that follow a normal distribution with mean θ_i and variance 1. Assume $\Theta = (-\infty, \infty)$ and $C = [-M, M]$.

We need to verify the condition that, for $\theta_1^0 \in C$, $0 < E_{\theta^0}\left(\inf_{\theta_1' \in C^c \cap \Theta} log \frac{f_1(X_{ij};\theta_1^0)}{f_1(X_{ij};\theta_1')}\right) \leq K^C < \infty$ for some constants M and K^C.

We then have

$$\inf_{\theta_1' \in C^c \cap \Theta} log \frac{f_1(x;\theta_1^0)}{f_1(x;\theta_1')} = \begin{cases} -\frac{1}{2}(x-\theta_1^0)^2 & if \ \ |x| > M, \\ -\frac{1}{2}(x-\theta_1^0)^2 + \frac{1}{2}(x-M)^2 & if \ \ 0 < x \leq M, \\ -\frac{1}{2}(x-\theta_1^0)^2 + \frac{1}{2}(x+M)^2 & if \ \ -M \leq x \leq 0. \end{cases}$$

It then follows that

$$E_{\theta^0} \inf_{\theta_1' \in C^c \cap \Theta} log \frac{f_1(X_{ij};\theta_1^0)}{f_1(X_{ij};\theta_1')} \ = \ I_{i1} + I_{i2} + I_{i3}, \quad i = 1, 2, ..., m,$$

where

$$\begin{aligned} I_{1i} &= -\int_{|x|>M} \frac{1}{2}(x-\theta_1^0)^2 \frac{1}{\sqrt{2\pi}} exp^{-(x-\theta_i)^2/2} dx, \\ I_{2i} &= \int_0^M \left(-\frac{1}{2}(x-\theta_1^0)^2 + \frac{1}{2}(x-M)^2 \right) \frac{1}{\sqrt{2\pi}} exp^{-(x-\theta_i)^2/2} dx, \\ I_{3i} &= \int_{-M}^0 \left(-\frac{1}{2}(x-\theta_1^0)^2 + \frac{1}{2}(x+M)^2 \right) \frac{1}{\sqrt{2\pi}} exp^{-(x-\theta_i)^2/2} dx. \end{aligned}$$

The first term I_1 goes to zero as M goes to infinity. It can be verified that $I_{2i} + I_{3i} = M^2 + o(M^2)$. It then follows that there exist $M_0 > 0$ such that $I_{1i} + I_{2i} + I_{3i} > 0$ for $M > M_0$. If we choose $K^C = 2M_0^2$, it then follows that, for $i = 1, 2, ..., m, j = 1, 2, ..., n_i$,

$$0 < E_{\theta^0} \left(\inf_{\theta_1' \in C^c \cap \Theta} log \frac{f_1(X_{ij};\theta_1^0)}{f_1(X_{ij};\theta_1')} \right) \leq 2M^2 < \infty, \quad for \ \ all \ \ M > M_0.$$

4.3.2 Restricted Normal Means.

A simple but important example is presented in this subsection. That problem is treated by van Eeden and Zidek (2001). Let $X_{11}, ..., X_{1n_1}$ be *i.i.d.* normal random variables each with mean θ_1 and variance σ^2. We now introduce a second random sample drawn independently of the first one from a second population: $X_{21}, ..., X_{2n_2}$,

i.i.d. normal random variables each with mean θ_2 and variance σ^2. Population 1 is of inferential interest while Population 2 is the relevant population. However, $|\theta_2 - \theta_1| \leq C$ for a known constant $C > 0$. Assumptions 4.2 and 4.3 are obviously satisfied for this example. The condition (4.6) in Theorem 4.3 is satisfied as shown in the previous example. If we show that Assumption 4.5 is also satisfied, then all the conditions assumed will be satisfied for this example.

To verify the final assumption, an explicit expression for the weight vector is needed. Let $n_i\bar{X}_{i\cdot} = \sum_{j=1}^{n_i} X_{ij}$, $i = 1, \ldots, m$, $V = Cov((\bar{X}_{1\cdot}, \bar{X}_{2\cdot})^t)$ and $B = (0, C)^t$. It follows that

$$V + BB^t = \begin{pmatrix} \frac{\sigma^2}{n_1} & 0 \\ 0 & \frac{\sigma^2}{n_2} + C \end{pmatrix}.$$

It can be shown that the "optimum" WLE in this case, the one that minimizes the maximum MSE over the restricted parameter space, takes the following form

$$\tilde{\theta}_1^{(n)} = \lambda_1^* \bar{X}_{1\cdot} + \lambda_2^* \bar{X}_{2\cdot},$$

where

$$(\lambda_1^*, \lambda_2^*)^t = \frac{(V + BB^t)^{-1}\mathbf{1}}{\mathbf{1}^t(V + BB^t)^{-1}\mathbf{1}}.$$

We find that

$$(V + BB^t)^{-1} = \begin{pmatrix} \frac{1}{\sigma^2/n_1} & 0 \\ 0 & \frac{1}{\sigma^2/n_2 + C} \end{pmatrix}.$$

It follows that

$$\mathbf{1}^t(V + BB^t)^{-1}\mathbf{1} = \frac{1}{\sigma^2/n_1} + \frac{1}{\sigma^2/n_2 + C}.$$

Thus, we have

$$\lambda_2^* = \frac{\frac{1}{\sigma^2/n_2 + C}}{\frac{1}{\sigma^2/n_1} + \frac{1}{\sigma^2/n_2 + C}}$$

Finally

$$\begin{aligned}\lambda_1^* &= 1 - \lambda_2^* \\ \lambda_2^* &= \left(\frac{C}{\sigma^2} + \frac{1}{n_2} + \frac{1}{n_1}\right)^{-1}\end{aligned}$$

Estimators of this type are considered by van Eeden and Zidek (2000).

It follows that $|\lambda_i^{(n_1)} - w_i| = O(\frac{1}{n_1})$, $i = 1, 2$. If we have $n_2 = O(n_1^{2-\delta})$, then Assumption 4.5 will b e satisfied. Therefore, we do not require that the two sample sizes approach to infinity at the same rate for this example in order to obtain consistency and asymptotic normality. The sample size of the relevant sample might go to infinity at a much higher rate.

Under the assumptions made in the subsection, it can be shown that the conditions of Theorem 4.4 are satisfied. The maximum of the likelihood estimator in this example is unique for any fixed sample size. Therefore, we have

$$\sqrt{n}(\tilde{\theta}_1^{(n)} - \theta_1) \xrightarrow{D} N(0, \sigma_1^2).$$

4.3.3 Multivariate Normal Means.

Let $\mathbf{X} = (\bar{X}_1, ..., \bar{X}_m)$, where for $i = 1, \ldots, m$,

$$\bar{X}_i = \sum_{j=1}^{n_i} X_{ij}/n_i \overset{ind.}{\sim} N(\theta_i, 1/n_i).$$

Assume that the θ_i are "close" to each other. The objective is to obtain a reasonable good estimate of θ_1. If the sample size from the first population is relatively small, we choose WLE as the estimator. In the normal case, the WLE, $\tilde{\theta}_1$, takes the following form:

$$\tilde{\theta}_1 = \sum_{i=1}^{m} \lambda_i \bar{X}_i.$$

Note that the James-Stein estimator of the parameter $\theta = (\theta_1, ..., \theta_m)$ is given by $\zeta(\mathbf{X}) = (\zeta_1(\mathbf{X}), ..., \zeta_m(\mathbf{X}))$, where

$$\zeta_i(\mathbf{X}) = \left(1 - \frac{m-2}{\sum_{i=1}^{p} \bar{X}_i^2}\right) \bar{X}_i.$$

The quantity,

$$1 - \frac{p-2}{\sum_{i=1}^{m} \bar{X}_i^2},$$

can be viewed as a weight function derived from the weight in the James-Stein estimator.

Consider the following choice of weights of James-Stein type :

$$\begin{aligned} \lambda_1(\mathbf{X}) &= 1 - \frac{1}{n_1^{1+\delta}} \frac{m-2}{\sum_{i=1}^{m} \bar{X}_i{}^2 + c}, \\ \lambda_i(\mathbf{X}) &= \frac{1}{m-1}\left(\frac{1}{n_1^{1+\delta}} \frac{m-2}{\sum_{i=1}^{m} \bar{X}_i^2 + c}\right), \quad i = 2, 3, ..., m, \end{aligned}$$

for some $\delta \geq 0$ and $c > 0$. It can be verified that $\sum_{i=1}^{m} \lambda_i = 1$ and $\lambda_i \geq 0, i = 2, 3, ..., m$. It follows that

$$\begin{aligned} P_{\theta^0}\left(\left|\frac{1}{n_1^{1+\delta}} \frac{m-2}{\sum_{i=1}^{m} \bar{X}_i{}^2 + c}\right| > \epsilon\right) &\leq \frac{m-2}{n_1^{1+\delta}\epsilon} E_{\theta^0}\left|\frac{1}{\sum_{i=1}^{m} \bar{X}_i{}^2 + c}\right| \\ &\leq \frac{m-2}{n_1^{1+\delta}\epsilon} E_{\theta^0} \frac{1}{c} \quad (\text{since } \bar{X}_i^2 \geq 0) \\ &= O\left(\frac{1}{n_1^{1+\delta}}\right). \end{aligned}$$

We then have

$$P_{\theta^0}\left(\left|\frac{1}{n_1^{1+\delta}} \frac{m-2}{\sum_{i=1}^{m} \bar{X}_i^2 + c}\right| > \epsilon\right) = O\left(\frac{1}{n_1^{1+\delta}}\right); \tag{4.37}$$

$$P_{\theta^0}(\lambda_1 < 0) = O\left(\frac{1}{n_1^{1+\delta}}\right). \tag{4.38}$$

We consider the following two scenarios

(i) If we set $\delta = 0$, it follows that

$$\lambda_i(\mathbf{X}) - w_i \xrightarrow{P_{\theta 0}} 0,$$

for $i = 1, 2, ..., m$. Assumption 4.6 is then satisfied. Therefore, the weak consistency and asymptotic normality of the WLE constructed by this set of weights will follow.

(ii) If $\delta > 0$,

$$\lambda_i(\mathbf{X}) - w_i \longrightarrow 0, \ \ a.s. \ [P_{\theta 0}],$$

for $i = 1, 2, ..., m$, then this leads to strong consistency. Since strong consistency implies weak consistency, asymptotic normality of the WLE using adaptive weights will follow in this case.

4.4 Concluding Remarks

In this chapter we have shown how classical large sample theory for the maximum likelihood estimator can be extended to the adaptively weighted likelihood estimator. In particular, we have proved the weak consistency of the latter and of the roots of the likelihood equation under more restrictive conditions. The asymptotic normality of the WLE is also proved. Observations from the same population are assumed to be independent although the observations from different populations obtained at the same time can be dependent.

In practice weights will sometimes need to be estimated. Assumption 4.6 states conditions that insure the large sample results obtain. In particular, they obtain as long as the samples drawn from populations different from that of inferential interest are of the same order as that of the drawn from the latter.

This finding could have useful practical implications since often there will be a differential cost of drawing samples from the various populations. The overall cost of sampling may be reduced by judiciously collecting a relatively larger amount of inexpensive data, that although biased, nevertheless increases the accuracy of the estimator. Our theory suggests that as long as the amount of that other data is about the same as obtained from the population of direct interest (and the weights are chosen appropriately), the asymptotic theory will hold.

Chapter 5

Choosing Weights by Cross-Validation

5.1 Introduction

This chapter is concerned with the application of the *cross-validation* criterion to the choice of optimum weights. This concept is an old one. In its most primitive but nevertheless useful form, it consists of controlled and uncontrolled division of the data sample into two subsamples. For example, the subsample can be selected by deleting one or a few observations or it can be a random sample from the dataset. Stone (1974) conducts a complete study of the cross-validatory choice and assessment of statistical predictions. Stone (1974) and Geisser (1975) discuss the application of cross-validation to the so-called *K-group* problem which uses a linear combinations of the sample means from different groups to estimate a common mean. Breiman and Friedman (1997) also demonstrate the benefit of using cross-validation to obtain the linear combination of predictions to achieve better estimation in the context of multivariate regression.

Although there are many ways of dividing the entire sample into subsamples such

as a random selection technique, we use the simplest *leave-one-out* approach in this chapter since the analytic forms of the optimum weights are then completely tractable for the linear WLE. We will denote the vector of parameters and the weight vector by $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_m)$ and $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, ..., \lambda_m)$ respectively. Assume that $||\boldsymbol{\theta}|| < \infty$ for $i = 1, 2, ..., m$. Let $\boldsymbol{\lambda}_e^{opt}$ and $\boldsymbol{\lambda}_u^{opt}$ be the optimum weight vector for samples with equal and unequal sizes. We require that $\sum_{i=1}^{m} \lambda_i = 1$ in this chapter.

Suppose that we have m populations which might be related to each other. The probability density functions or probability mass functions are of the form $f_i(x; \theta_i)$ with θ_i as the parameter for population i. Assume that

$$\begin{array}{llllll} X_{11}, & X_{12}, & X_{13}, \ldots, & X_{1n_1} & \overset{i.i.d}{\sim} & f_1(x;\theta_1) \\ X_{21}, & X_{22}, & X_{23}, \ldots, & X_{2n_2} & \overset{i.i.d}{\sim} & f_2(x;\theta_2) \\ \vdots & & & \vdots & & \\ X_{m1}, & X_{m2}, & X_{m3}, \ldots, & X_{mn_m} & \overset{i.i.d.}{\sim} & f_m(x;\theta_m) \end{array}$$

where, for fixed i, the $\{X_{ij}\}$ are observations obtained from population i and so on. Assume that observations obtained from each population are independent of those from other populations and $E(X_{ij}) = \phi(\theta_i), j = 1, 2, ..., n_i$. The population parameter of the first population, θ_1, is of inferential interest. Taking the usual path, we predict X_{1j} by $\phi(\tilde{\theta}_1^{(-j)})$, the WLE of its mean without using the X_{1j}. Note that $\phi(\tilde{\theta}_1^{(-j)})$ is a function of the weight vector $\boldsymbol{\lambda}$ by the construction of the WLE. A natural measure for the discrepancy of the WLE is the following:

$$D(\boldsymbol{\lambda}) = \sum_{j=1}^{n_1} \left(X_{1j} - \phi(\tilde{\theta}_1^{(-j)}) \right)^2 . \tag{5.1}$$

The optimum weights are derived such that the minimum of $D(\boldsymbol{\lambda})$ is achieved for fixed sample sizes $n_1, n_2, ..., n_m$ and $\sum_{i=1}^{m} \lambda_i = 1$.

We will study the linear WLE by using cross-validation when $E(X_{ij}) = \theta_i, j = 1, 2, ...n_i$ for any fixed i. The asymptotic properties of the WLE are established in

this chapter. The results of simulation studies are shown later in this chapter.

5.2 Linear WLE for Equal Sample Sizes

Stone (1974) and Geisser (1975) discuss the application of the the cross-validation approach to the so-called *K-group* problem. Suppose that the data set S consists of n observations in each of K groups. The prediction of the mean for the ith group is constructed as:

$$\hat{\mu}_i = \alpha \overline{X}_{i.} + (1-\alpha)\overline{X}_{..}$$

where $\overline{X}_{..} = \frac{1}{Kn}\sum_{i=1}^{m}\sum_{j=1}^{n} X_{ij}$ and $\overline{X}_{i.} = \frac{1}{n}\sum_{j=1}^{n} X_{ij}$. If we are interested in group 1, then the prediction for group 1 becomes

$$\hat{\mu}_1 = \left(1 - \frac{K-1}{K}\alpha\right)\overline{X}_{1.} + \sum_{i=2}^{m}\frac{\alpha}{K}\overline{X}_{i.}.$$

We remark that the above formula is a special form of linear combination of the sample means. The cross-validation procedure is used by Stone (1974) to derive the value of α.

We consider general linear combinations. Let $\tilde{\theta}_1^{(e)}$ denote the WLE by using the cross-validation rule when the sample sizes are equal. If $\phi(\theta) = \theta$, the linear WLE for θ_1 is then defined as

$$\tilde{\theta}_1^{(e)} = \sum_{i=1}^{m}\lambda_i \overline{X}_{i.}$$

where $\sum_{i=1}^{m}\lambda_i = 1$. We assume $n_1 = n_2 = ... = n_m = n$ in this section.

In this section, we will use cross-validation by simultaneously deleting $X_{1j}, X_{2j}, ..., X_{mj}$ for each fixed j. That is, we delete one data point from each sample at each step. This might be appropriate if these data points are obtained at the same time point and strong associations exist among these observations. By simultaneously deleting

$X_{1j}, X_{2j}, ..., X_{mj}$ for each fixed j, we might achieve numerical stability of the cross-validation procedure. An alternative approach is to delete a data point from only the first sample at each step. It will be studied in the next section.

Let $\overline{X}_{i.}^{(-j)}$ be the sample mean of the ith sample with jth element in that sample excluded. A natural measure for the discrepancy of $\tilde{\theta}_1$ might be:

$$\begin{aligned} D_e^{(m)} &= \sum_{j=1}^{n} \left(X_{1j} - \sum_{i=1}^{m} \lambda_i \overline{X}_{i.}^{(-j)} \right)^2 \\ &= \sum_{j=1}^{m} X_{1j}^2 - 2 \sum_{j=1}^{n} X_{1j} \sum_{i=1}^{m} \lambda_i \overline{X}_{i.}^{(-j)} + \sum_{j=1}^{n} \sum_{i=1}^{m} \sum_{k=1}^{m} \lambda_i \lambda_k \overline{X}_{i.}^{(-j)} \overline{X}_{k.}^{(-j)} \\ &= \sum_{j=1}^{n} X_{1j}^2 - 2 \sum_{i=1}^{m} \lambda_i \sum_{j=1}^{n} X_{1j} \overline{X}_{i.}^{(-j)} + \sum_{i=1}^{m} \sum_{k=1}^{m} \lambda_i \lambda_k \sum_{j=1}^{n} \overline{X}_{i.}^{(-j)} \overline{X}_{k.}^{(-j)} \\ &= c(\underline{\boldsymbol{X}}) - 2\boldsymbol{\lambda}^t b_e(\underline{\boldsymbol{X}}) + \boldsymbol{\lambda}^t A_e(\underline{\boldsymbol{X}}) \boldsymbol{\lambda} \end{aligned}$$

where $c(\underline{\boldsymbol{X}}) = \sum_{j=1}^{n} X_{1j}^2$, $(b_e(\underline{\boldsymbol{X}}))_i = \sum_{j=1}^{n} X_{1j} \overline{X}_{i.}^{(-j)}$, and $(A_e(\underline{X}))_{ik} = \sum_{j=1}^{n} \overline{X}_{i.}^{(-j)} \overline{X}_{k.}^{(-j)}$, $i = 1, 2, ..., n, k = 1, 2, ..., m$.

An optimum weight vector by using the cross-validation rule is defined to be a weight vector which minimizes the objective function, $D_e^{(m)}$ and satisfies $\sum_{i=1}^{m} \lambda_i = 1$. For expository simplicity, let $b_e = b_e(\underline{X})$ and $A_e = A_e(\underline{X})$ in this chapter.

5.2.1 Two Population Case

For simplicity, first consider a simple case of two populations, *i.e.*

$$\begin{aligned} X_{11}, \quad X_{12}, \quad X_{13}, \quad \ldots, \quad X_{1n} &\overset{i.i.d.}{\sim} f_1(x; \theta_1) \\ X_{21}, \quad X_{22}, \quad X_{23}, \quad \ldots, \quad X_{2n} &\overset{i.i.d.}{\sim} f_2(x; \theta_2) \end{aligned}$$

with $E(X_{1j}) = \theta_1$ and $E(X_{2j}) = \theta_2$. Let σ_1^2 and σ_2^2 denote the variances of X_{1j} and X_{2j} respectively. Let $\rho = cor(X_{1j}, X_{2j})$. Let $\theta^0 = (\theta_1^0, \theta_2^0)$ where θ_1^0 and θ_2^0 are the true values for θ_1 and θ_2 respectively.

We seek the optimum weights such that $\lambda_1 + \lambda_2 = 1$ and they minimize the objective function which is defined as follows:

$$D_e^{(2)} = \sum_{j=1}^{n} \left(X_{1j} - \lambda_1 \overline{X}_{1.}^{(-j)} - \lambda_2 \overline{X}_{2.}^{(-j)} \right)^2 - \gamma(\lambda_1 + \lambda_2 - 1).$$

Differentiating $D_e^{(2)}$ with respect to λ_1 and λ_2, we have

$$\frac{\partial D_e^{(2)}}{\partial \lambda_1} = -\sum_{j=1}^{n} \overline{X}_{1.}^{(-j)} \left(X_{1j} - \lambda_1 \overline{X}_{1.}^{(-j)} - \lambda_2 \overline{X}_{2.}^{(-j)} \right) - \gamma = 0,$$

$$\frac{\partial D_e^{(2)}}{\partial \lambda_2} = -\sum_{j=1}^{n} \overline{X}_{2.}^{(-j)} \left(X_{1j} - \lambda_1 \overline{X}_{1.}^{(-j)} - \lambda_2 \overline{X}_{2.}^{(-j)} \right) - \gamma = 0.$$

It follows that

$$\sum_{j=1}^{n} \left(\overline{X}_{1.}^{(-j)} - \overline{X}_{2.}^{(-j)} \right) \left(X_{1j} - \lambda_1 \overline{X}_{1.}^{(-j)} - \lambda_2 \overline{X}_{2.}^{(-j)} \right) = 0.$$

Note that $\lambda_1 + \lambda_2 = 1$. We then have

$$\begin{cases} \lambda_1^{opt}(\mathbf{X}) = 1 - \dfrac{\sum_{j=1}^{n} \left(\overline{X}_{1.}^{(-j)} - \overline{X}_{2.}^{(-j)} \right) \left(\overline{X}_{1.}^{(-j)} - X_{1j} \right)}{\sum_{j=1}^{n} (\overline{X}_{1.}^{(-j)} - \overline{X}_{2.}^{(-j)})^2}, \\ \lambda_2^{opt}(\mathbf{X}) = \dfrac{\sum_{j=1}^{n} \left(\overline{X}_{1.}^{(-j)} - \overline{X}_{2.}^{(-j)} \right) \left(\overline{X}_{1.}^{(-j)} - X_{1j} \right)}{\sum_{j=1}^{n} \left(\overline{X}_{1.}^{(-j)} - \overline{X}_{2.}^{(-j)} \right)^2}. \end{cases} \tag{5.2}$$

Lemma 5.1 *The following identity holds:*

$$\lambda_1^{opt} = 1 - \lambda_2^{opt} \quad \textit{and} \quad \lambda_2^{opt} = S_2^e / S_1^e,$$

where

$$\begin{aligned} S_1^e &= \frac{n(n-2)}{(n-1)^2} \left(\overline{X}_{1.} - \overline{X}_{2.} \right)^2 + \frac{1}{n(n-1)^2} \sum_{j=1}^{n} (X_{1j} - X_{2j})^2, \\ S_2^e &= \frac{n}{(n-1)^2} \left(\hat{\sigma}_1^2 - \widehat{cov} \right) \end{aligned}$$

where $\hat{\sigma_1}^2 = \frac{1}{n} \sum_{j=1}^{n} \left(X_{1j} - \overline{X}_{1.} \right)^2$ *and* $\widehat{cov} = \frac{1}{n} \sum_{j=1}^{n} \left(X_{1j} - \overline{X}_{1.} \right) \left(X_{2j} - \overline{X}_{2.} \right)$.

Proof: Observe that

$$\begin{aligned}
\overline{X}_{i.}^{(-j)} &= \frac{1}{n-1}(X_{i1} + X_{i2} + ...X_{ij-1} + X_{ij+1} + ... + X_{in}) \\
&= \frac{1}{n-1}\sum_{j=1}^{n} X_{ij} - \frac{1}{n-1}X_{ij} \\
&= \frac{n}{n-1}\overline{X}_{i.} - \frac{1}{n-1}X_{ij} \\
&= e_n\overline{X}_{i.} - \frac{1}{n-1}X_{ij}.
\end{aligned}$$

where $e_n = \frac{n}{n-1}$.

Let $S_1^e = \frac{1}{n}\sum_{j=1}^{n}\left(\overline{X}_{1.}^{(-j)} - \overline{X}_{2.}^{(-j)}\right)^2$. It then follows that

$$\begin{aligned}
S_1^e &= \frac{1}{n}\sum_{j=1}^{n}\left((e_n\overline{X}_{1.} - \frac{1}{n-1}X_{1j}) - (e_n\overline{X}_{2.} - \frac{1}{n-1}X_{2j})\right)^2 \\
&= \frac{1}{n}\sum_{j=1}^{n}\left(e_n(\overline{X}_{1.} - \overline{X}_{2.}) - \frac{1}{n-1}(X_{1j} - X_{2j})\right)^2 \\
&= \frac{1}{n}\Big(ne_n^2(\overline{X}_{1.} - \overline{X}_{2.})^2 - 2\frac{e_n}{n-1}(\overline{X}_{1.} - \overline{X}_{2.})\sum_{j=1}^{n}(X_{1j} - X_{2j}) \\
&\qquad +(\frac{1}{n-1})^2\sum_{j=1}^{n}(X_{1j} - X_{2j})^2\ \Big) \\
&= e_n^2(\overline{X}_{1.} - \overline{X}_{2.})^2 - 2e_n\frac{1}{n-1}(\overline{X}_{1.} - \overline{X}_{2.})^2 + \frac{1}{n(n-1)^2}\sum_{j=1}^{n}(X_{1j} - X_{2j})^2 \\
&= e_n\left(e_n - \frac{2}{n-1}\right)(\overline{X}_{1.} - \overline{X}_{2.})^2 + \frac{1}{n(n-1)^2}\sum_{j=1}^{n}(X_{1j} - X_{2j})^2 \\
&= \frac{n(n-2)}{(n-1)^2}(\overline{X}_{1.} - \overline{X}_{2.})^2 + \frac{1}{n(n-1)^2}\sum_{j=1}^{n}(X_{1j} - X_{2j})^2.
\end{aligned}$$

Let $S_2^e = \frac{1}{n}\sum_{j=1}^{n}(\overline{X}_{1.}^{(-j)} - X_{2.}^{(-j)})(\overline{X}_{1.}^{(-j)} - X_{1j})$. It follows that

$$\begin{aligned}
S_2^e &= \frac{1}{n}\sum_{j=1}^{n}\left(e_n(\overline{X}_{1.} - \overline{X}_{2.}) - \frac{1}{n-1}(X_{1j} - X_{2j})\right)\left((e_n\overline{X}_{1.} - \frac{1}{n-1}X_{1j}) - X_{1j}\right) \\
&= \frac{1}{n}\sum_{j=1}^{n}\left(e_n(\overline{X}_{1.} - \overline{X}_{2.}) - \frac{1}{n-1}(X_{1j} - X_{2j})\right)\left(e_n\overline{X}_{1.} - e_nX_{1j}\right) \\
&= \frac{e_n^2}{n}(\overline{X}_{1.} - \overline{X}_{2.})\sum_{j=1}^{n}(\overline{X}_{1.} - X_{1j}) - \frac{e_n}{n(n-1)}\sum_{j=1}^{n}(X_{1j} - X_{2j})(\overline{X}_{1.} - X_{1j}) \\
&= -\frac{e_n}{n(n-1)}\sum_{j=1}^{n}(X_{1j} - X_{2j})(\overline{X}_{1.} - X_{1j}) \qquad (since \sum_{j=1}^{n}(\overline{X}_{1.} - X_{1j}) = 0) \\
&= -\frac{e_n}{n(n-1)}\left(\overline{X}_{1.}\sum_{j=1}^{n}(X_{1j} - X_{2j}) - \sum_{j=1}^{n}X_{1j}^2 + \sum_{j=1}^{n}X_{1j}X_{2j}\right) \\
&= -\frac{e_n}{n-1}\left(\overline{X}_{1.}(\overline{X}_{1.} - \overline{X}_{2.}) - \frac{1}{n}\sum_{j=1}^{n}X_{1j}^2 + \frac{1}{n}\sum_{j=1}^{n}X_{1j}X_{2j}\right) \\
&= \frac{n}{(n-1)^2}\left(\hat{\sigma}_1^2 - \widehat{cov}\right)
\end{aligned}$$

This completes the proof. $\diamond$

The value of the λ_2^{opt} can be seen as some sort of measure of relevance between the two samples. If that measure is almost zero, the formula for λ_2^{opt} will reflect that by assigning a very small value to λ_2^{opt}. This implies that there is no need to combine the two populations if the difference between the two sample means is relatively large or the second sample has little information of relevance to the first. The weights chosen by the cross-validation rule will then guard against the undesirable scenario in which too much bias might be introduced into the estimation procedure. On the other hand, if the second sample does contain valuable information about the parameter of interest, the cross-validation procedure will recognize that by assigning a non-zero value to λ_2^{opt}.

Proposition 5.1 *If* $\rho < \frac{\sigma_1}{\sigma_2}$, *then*

$$P_{\theta^0}(\lambda_2^{opt} > 0) \xrightarrow{P_{\theta^0}} 1.$$

Proof: By the Weak Law of Large Numbers, it follows that

$$\begin{aligned}
&\overline{X}_{1\cdot} \xrightarrow{P_{\theta^0}} \theta_1^0; \quad \overline{X}_{2\cdot} \xrightarrow{P_{\theta^0}} \theta_2^0, \\
&\frac{1}{n}\sum_{j=1}^{n} X_{1j}^2 \xrightarrow{P_{\theta^0}} \sigma_1^2 + (\theta_1^0)^2, \\
&\frac{1}{n}\sum_{j=1}^{n} X_{1j}X_{2j} \xrightarrow{P_{\theta^0}} \rho\sigma_1\sigma_2 + \theta_1^0\theta_2^0.
\end{aligned}$$

Therefore,

$$\hat{\sigma}_1^2 - \widehat{cov} \longrightarrow \sigma_1^2 - \rho\sigma_1\sigma_2.$$

Thus condition $\rho < \sigma_1/\sigma_2$ implies that $\hat{\sigma}_1^2 > \widehat{cov}$ for sufficiently large n. Thus, λ_2^{opt} eventually will be positive. $\diamond$

We remark that the condition $\rho < \sigma_1/\sigma_2$ is satisfied if $\sigma_2 < \sigma_1$ or $\rho < 0$. If the condition $\rho < \sigma_1/\sigma_2$ is not satisfied, then λ_2^{opt} will have negative sign for sufficiently large n. However, the value of λ_2^{opt} will converge to zero as shown in the next Proposition.

Proposition 5.2 *If* $\theta_1^0 \neq \theta_2^0$, *then, for any given* $\epsilon > 0$,

$$P_{\theta^0}(|\lambda_1^{opt} - 1| \leq \epsilon) \longrightarrow 1 \quad \textit{and} \quad P_{\theta^0}(|\lambda_2^{opt}| < \epsilon) \longrightarrow 1.$$

Proof: From Lemma 5.1, it follows that the second term of S_1 goes to zero in probability as n goes to infinity while the first term converges to $(\theta_1^0 - \theta_2^0)^2$ in probability.

Therefore we have

$$S_1^e \xrightarrow{P_{\theta^0}} (\theta_1^0 - \theta_2^0)^2 \ \ as \ \ n \to \infty,$$

where $(\theta_1^0 - \theta_2^0)^2 \neq 0$ by assumption.

Moreover, we see that $S_2^e = O_P(\frac{1}{n})$. By definition of λ_2^{opt}, it follows that

$$|\lambda_2^*| = |\frac{S_2^e}{S_1^e}| \xrightarrow{P_{\theta^0}} 0 \quad as \ \ n \to \infty.$$

This completes the proof. $\diamond$

The asymptotic limit of the weights will not exist if θ_1^0 equals θ_2^0. This is because the cross-validation procedure will not be able to detect the difference of the two populations involved since there is none. This can be rectified by defining $\lambda_2^{opt} = \frac{S_2^e}{S_1^e + c}$ where $c > 0$. We remark that the knowledge of the variances and covariances is not assumed.

5.2.2 Alternative Matrix Representation of A_e and b_e

To study the case of more than two populations, it is necessary to derive an alternative matrix representation of λ^{opt}. It can be verified that

$$\begin{aligned} \overline{x}_{i.}^{(-j)}\overline{x}_{k.}^{(-j)} &= (e_n\overline{x}_{i.} - \frac{1}{n-1}x_{ij})(e_n\overline{x}_{k.} - \frac{1}{n-1}x_{kj}) \\ &= e_n^2\overline{x}_{i.}\overline{x}_{k.} - \frac{e_n}{n-1}x_{ij}\overline{x}_{k.} - \frac{e_n}{n-1}x_{kj}\overline{x}_{i.} + (\frac{1}{n-1})^2 x_{ij}x_{kj} \end{aligned}$$

where $e_n = \frac{n}{n-1}$.

Thus, we have

$$\begin{aligned}
\sum_{j=1}^{n} \overline{x}_{i.}^{(-j)} \overline{x}_{k.}^{(-j)} &= \sum_{j=1}^{n} \left(e_n^2 \overline{x}_{i.} \overline{x}_{k.} - \frac{e_n}{n-1} x_{ij} \overline{x}_{k.} - \frac{e_n}{n-1} x_{kj} \overline{x}_{i.} + (\frac{1}{n-1})^2 x_{ij} x_{kj} \right) \\
&= n e_n^2 \overline{x}_{i.} \overline{x}_{k.} - \frac{e_n}{n-1} \overline{x}_{k.} \sum_{j=1}^{n} x_{ij} - \frac{e_n}{n-1} \overline{x}_{i.} \sum_{j=1}^{n} x_{kj} + (\frac{1}{n-1})^2 \sum_{j=1}^{n} x_{ij} x_{kj} \\
&= n e_n^2 \overline{x}_{i.} \overline{x}_{k.} - e_n^2 \overline{x}_{i.} \overline{x}_{k.} - e_n^2 \overline{x}_{i.} \overline{x}_{k.} + \frac{n}{(n-1)^2} \frac{1}{n} \sum_{j=1}^{n} x_{ij} x_{kj} \\
&= e_n^2 (n-2) \overline{x}_{i.} \overline{x}_{k.} + \frac{e_n}{n-1} \frac{1}{n} \sum_{j=1}^{n} x_{ij} x_{kj} \\
&= e_n^2 (n-2) \overline{x}_{i.} \overline{x}_{k.} + \frac{e_n}{n-1} \frac{1}{n} \sum_{j=1}^{n} (x_{ij} - \overline{x}_{i.})(x_{kj} - \overline{x}_{k.}) + \frac{e_n}{n-1} \overline{x}_{i.} \overline{x}_{k.} \\
&= \left(e_n^2 (n-2) + \frac{e_n}{n-1} \right) \hat{\theta}_i \hat{\theta}_k \; + \; \frac{e_n}{n-1} \widehat{cov}_{ik}^2,
\end{aligned}$$

where

$$\begin{aligned}
\hat{\theta}_i &= \overline{x}_{i.}, \quad i = 1, 2, ..., m; \\
\widehat{cov}_{ik} &= \frac{1}{n} \sum_{j=1}^{n} (x_{ij} - \overline{x}_{i.})(x_{kj} - \overline{x}_{k.}).
\end{aligned}$$

Recall that, for $1 \leq i \leq m$ and $1 \leq k \leq m$,

$$A_{e(ik)} = \sum_{j=1}^{n} \overline{x}_{i.}^{(-j)} \overline{x}_{k.}^{(-j)}.$$

It follows that

$$A_e = \frac{e_n}{n-1} \hat{\Sigma} \; + \; \left(e_n^2 (n-2) + \frac{e_n}{n-1} \right) \hat{\boldsymbol{\theta}} \hat{\boldsymbol{\theta}}^t \tag{5.3}$$

where $\Sigma_{ik} = \widehat{cov}_{ik}$ and $\hat{\boldsymbol{\theta}} = (\overline{x}_{1.}, ..., \overline{x}_{m.})$.

We also have

$$\begin{aligned}
b_{e(i)}(\mathbf{x}) &= \sum_{j=1}^{n} x_{1j}\ \overline{x}_{i.}^{(-j)} \\
&= \sum_{j=1}^{n} (x_{1j} + \overline{x}_{1.}^{(-j)} - \overline{x}_{1.}^{(-j)})\ \overline{x}_{i.}^{(-j)} \\
&= \sum_{j=1}^{n} \overline{x}_{1.}^{(-j)}\ \overline{x}_{i.}^{(-j)} + \sum_{j=1}^{n} (x_{1j} - \overline{x}_{1.}^{(-j)})\ \overline{x}_{i.}^{(-j)} \\
&= A_{1i} + \sum_{j=1}^{n} \left(x_{1j} - (e_n \overline{x}_{1.} - \frac{1}{n-1} x_{1i}) \right) \left(e_n \overline{x}_{i.} - \frac{1}{n-1} x_{ij} \right) \\
&= A_{1i} + \sum_{j=1}^{n} (e_n x_{1j} - e_n \overline{x}_{1.}) \left(e_n \overline{x}_{i.} - \frac{1}{n-1} x_{ij} \right) \\
&= A_{1i} - \frac{e_n}{n-1} \sum_{j=1}^{n} (x_{1j} - \overline{x}_{1.}) x_{ij}.
\end{aligned}$$

It then follows that

$$b_e(\mathbf{x}) = A_1 - e_n^2 \widehat{\Sigma}_1. \tag{5.4}$$

where A_1 is the first column of A_e and $\widehat{\Sigma}_1$ is the first column of the sample covariance matrix $\widehat{\Sigma}$.

5.2.3 Optimum Weights $\boldsymbol{\lambda}_e^{opt}$ By Cross-validation

We are now in a position to derive the optimum weights when sample sizes are equal.

Proposition 5.3 *The optimum weight vector which minimizes $D_e^{(m)}$ takes the following form*

$$\boldsymbol{\lambda}_e^{opt} = (1, 0, 0, ..., 0)^t - e_n^2 \left(A_e^{-1} \widehat{\Sigma}_1 - \frac{\mathbf{1}^t A_e^{-1} \widehat{\Sigma}_1}{\mathbf{1}^t A_e^{-1} \mathbf{1}} A_e^{-1} \mathbf{1} \right).$$

Proof:

By differentiating $D_e^{(m)} - \nu\,(\mathbf{1}^t \boldsymbol{\lambda} - 1)$ and setting the result to zero, it follows that

$$\frac{\partial D_e^{(m)} - \nu\,(\mathbf{1}^t \boldsymbol{\lambda} - 1)}{\partial \lambda} = -2b_e + 2A_e \boldsymbol{\lambda}_e^{opt} - \nu \mathbf{1} = 0.$$

It then follows that

$$\boldsymbol{\lambda}_e^{opt} = A_e^{-1}\left(b_e + \frac{\nu}{2}\mathbf{1}\right).$$

We then have

$$1 = \mathbf{1}^t\boldsymbol{\lambda}_e^{opt} = \mathbf{1}^t A_e^{-1}\left(b_e + \frac{\nu\mathbf{1}}{2}\right).$$

Thus,

$$\nu = \frac{2}{\mathbf{1}^t A_e^{-1}\mathbf{1}}\left(1 - \mathbf{1}^t A_e^{-1} b_e\right).$$

Therefore,

$$\boldsymbol{\lambda}_e^{opt} = A_e^{-1}\left(b_e + \frac{1 - \mathbf{1}^t A_e^{-1} b_e}{\mathbf{1}^t A_e^{-1}\mathbf{1}}\mathbf{1}\right).$$

Since $D_e^{(m)}$ is a quadratic function of $\boldsymbol{\lambda}$ and $A \geq 0$, the minimum is achieved at the point $\boldsymbol{\lambda}_e^{opt}$. Furthermore, by equation (5.3) and (5.4), we have

$$A_e^{-1} b_e = A_e^{-1}\left(A_1 - e_n^2\widehat{\Sigma}_1\right) = (1, 0, 0, ..., 0)^t - e_n^2 A_e^{-1}\widehat{\Sigma_1}.$$

Denote the optimum weight vector by λ^{opt}. It follows that

$$\boldsymbol{\lambda}_e^{opt} = (1, 0, 0, ..., 0)^t - e_n^2\left(A_e^{-1}\widehat{\Sigma}_1 - \frac{\mathbf{1}^t A_e^{-1}\widehat{\Sigma}_1}{\mathbf{1}^t A_e^{-1}\mathbf{1}} A_e^{-1}\mathbf{1}\right).$$

This completes the proof. $\diamond$

We remark that A_e is invertible since $\hat{\Sigma}$ is invertible. We remark that the expression of the weight vector in the two population case can also be derived by using the matrix representation given as above. The detailed calculation is quite similar to that given in the previous subsection.

5.3 Linear WLE for Unequal Sample Sizes

In the previous section, we discussed choosing the optimum weights when the samples sizes are equal. In this section, we propose cross-validation methods for choosing

adaptive weights for unequal sample sizes. If the sample sizes are not equal, it is not clear whether the *delete-one-column* approach is a reasonable one. For example, suppose that there are 10 observations in first sample and there are 5 observations in the second. Then there is no observation to delete for the second sample for half of the cross-validation steps. Furthermore, we might lose accuracy in prediction deleting one column for small sample sizes. Therefore we propose alternative method which delete only one data point from the first sample and keep all the data points from the rest of samples if the sample sizes are not equal.

5.3.1 Two Population Case

Let us again consider the two population case in which only two populations are considered. The optimum weights $\boldsymbol{\lambda}_u^{opt}$ are obtained by minimizing the following objective function:

$$D_u^{(2)}(\boldsymbol{\lambda}) = \sum_{j=1}^{n_1} \left(X_{1j} - \lambda_1 \overline{X}_{1.}^{(-j)} - \lambda_2 \overline{X}_{2.} \right)^2,$$

where $\sum_{i=1}^{m} \lambda_i = 1$ and $\overline{X}_{1.}^{(-j)} = \frac{1}{n_1 - 1} \sum_{k \neq j}^{n_1} X_{1k}$. We remark that the major difference between $D_e^{(2)}$ and $D_u^{(2)}$ is that only the jth data point of the first sample is left out for the jth term in $D_u^{(2)}$.

Under the condition that $\lambda_1 + \lambda_2 = 1$, we can rewrite $D_u^{(2)}$ as a function of λ_1:

$$\begin{aligned} D_u^{(2)} &= \sum_{j=1}^{n_1} \left(X_{1j} - \lambda_1 \overline{X}_{1.}^{(-j)} - (1 - \lambda_1) \overline{X}_{2.} \right)^2 \\ &= \sum_{j=1}^{n_1} \left((X_{1j} - \overline{X}_{2.}) + \lambda_1 (\overline{X}_{2.} - \overline{X}_{1.}^{(-j)}) \right)^2. \end{aligned}$$

We differentiate $D_u^{(2)}$ with respect to λ_1. It then follows that

$$\frac{\partial D_u^{(2)}}{\partial \lambda_1} = \sum_{j=1}^{n_1} \left((X_{1j} - \overline{X}_{2.}) + \lambda_1 (\overline{X}_{2.} - \overline{X}_{1.}^{(-j)}) \right) \left(\overline{X}_{2.} - \overline{X}_{1.}^{(-j)} \right).$$

We then have

$$\lambda_1^{opt} = \frac{\sum_{j=1}^{n_1}(\overline{X}_{2.} - \overline{X}_{1.}^{(-j)})(\overline{X}_{2.} - X_{1j})}{\sum_{j=1}^{n_1}(\overline{X}_{2.} - \overline{X}_{1.}^{(-j)})^2} = \frac{S_1^u}{S_2^u}.$$

Consider

$$\begin{aligned}
S_1^u &= \sum_{j=1}^{n_1}(\overline{X}_{2.} - \overline{X}_{1.}^{(-j)})(\overline{X}_{2.} - X_{1j}) \\
&= \sum_{j=1}^{n_1}\left(\overline{X}_{2.} - \overline{X}_{1.} - \frac{1}{n_1 - 1}(\overline{X}_{1.} - X_{1j})\right)(\overline{X}_{2.} - X_{1j}) \\
&= \sum_{j=1}^{n_1}(\overline{X}_{2.} - \overline{X}_{1.})(\overline{X}_{2.} - X_{1j}) - \frac{1}{n_1 - 1}\sum_{j=1}^{n_1}(\overline{X}_{1.} - X_{1j})(\overline{X}_{2.} - X_{1j}) \\
&= n_1(\overline{X}_{1.} - \overline{X}_{2.})^2 + \frac{1}{n_1 - 1}\sum_{j=1}^{n_1}(\overline{X}_{1.} - X_{1j})X_{1j} \\
&= n_1(\overline{X}_{1.} - \overline{X}_{2.})^2 - \frac{n_1}{n_1 - 1}\hat{\sigma}_1^2
\end{aligned}$$

and

$$\begin{aligned}
S_2^u &= \sum_{j=1}^{n_1}(\overline{X}_{2.} - \overline{X}_{1.}^{(-j)})^2 \\
&= \sum_{j=1}^{n_1}[\overline{X}_{2.} - \overline{X}_{1.} - \frac{1}{n_1 - 1}(\overline{X}_{1.} - X_{1j})]^2 \\
&= \sum_{j=1}^{n_1}(\overline{X}_{1.} - \overline{X}_{2.})^2 + 2\frac{1}{n_1 - 1}(\overline{X}_{1.} - \overline{X}_{2.})\sum_{j=1}^{n_1}(\overline{X}_{1.} - X_{1j}) + (\frac{1}{n_1 - 1})^2\sum_{j=1}^{n_1}(\overline{X}_{1.} - X_{1j})^2 \\
&= n_1(\overline{X}_{1.} - \overline{X}_{2.})^2 + \frac{n_1}{(n_1 - 1)^2}\hat{\sigma}_1^2.
\end{aligned}$$

We then have

$$\lambda_1^{opt} = \frac{n_1(\overline{X}_{1.} - \overline{X}_{2.})^2 - \frac{n_1}{n_1 - 1}\hat{\sigma}_1^2}{n_1(\overline{X}_{1.} - \overline{X}_{2.})^2 + \frac{n_1}{(n_1 - 1)^2}\hat{\sigma}_1^2}; \quad \lambda_2 = 1 - \lambda_1^{opt}. \tag{5.5}$$

Proposition 5.4 *If* $\theta_1^0 \neq \theta_2^0$, *then* $\lambda_1^{opt} \xrightarrow{P_{\theta^0}} 1$ *and* $\lambda_2^{opt} \xrightarrow{P_{\theta^0}} 0$.

Proof: From equation (5.5), it follows that

$$\lambda_1^{opt} = 1 - \frac{\left(\frac{n_1}{n_1 - 1}\right)^2\hat{\sigma}_1^2}{n_1(\overline{X}_{1.} - \overline{X}_{2.})^2 + \frac{1}{n_1 - 1}\hat{\sigma}_1^2}.$$

By the Weak Law of Large Numbers, we have

$$\hat{\sigma}_1^2 \xrightarrow{P_{\theta^0}} \sigma_1^2$$
$$(\overline{X}_{1.} - \overline{X}_{2.})^2 \xrightarrow{P_{\theta^0}} (\theta_1^0 - \theta_2^0)^2 \neq 0.$$

It then follows that

$$\frac{\left(\frac{n_1}{n_1-1}\right)^2 \hat{\sigma}_1^2}{n_1(\overline{X}_{1.} - \overline{X}_{2.})^2 + \frac{1}{n_1-1}\hat{\sigma}_1^2} \xrightarrow{P_{\theta^0}} 0.$$

We then have

$$\lambda_1^{opt} \xrightarrow{P_{\theta^0}} 1.$$

The last assertion of the theorem follows by the fact that $\lambda_1 + \lambda_2 = 1$. $\diamond$

5.3.2 Optimum Weights By Cross-Validation

We derive the general formula for the optimum weights by cross-validation when the sample sizes are not all equal.

The objective function is defined as follows:

$$\begin{aligned} D_u^{(m)} &= \sum_{j=1}^{n_1} \left(X_{1j} - \lambda_1 \overline{X}_{1.}^{(-j)} - \sum_{i=2}^{m} \lambda_i \overline{X}_{i.} \right)^2 \\ &= \sum_{j=1}^{n_1} X_{1j}^2 - 2\sum_{j=1}^{n_1} X_{1j} \left(\lambda_1 \left(\overline{X}_{1.} + \frac{1}{n_1 - 1}(\overline{X}_{1.} - X_{1j}) \right) + \sum_{i=2}^{m} \lambda_i \overline{X}_{i.} \right) \\ &\quad + \sum_{j=1}^{n_1} \left(\lambda_1 \left(\overline{X}_{1.} + \frac{1}{n_1 - 1}(\overline{X}_{1.} - X_{1j}) \right) + \sum_{i=2}^{m} \lambda_i \overline{X}_{i.} \right)^2 \\ &= c(\underline{X}) - 2\boldsymbol{b}(\underline{X})\boldsymbol{\lambda}_u + \boldsymbol{\lambda}_u^t A(\underline{X})\boldsymbol{\lambda}_u \end{aligned}$$

where

$$\begin{aligned} b_1 &= \sum_{j=1}^{n_1} X_{1j} \left(\overline{X}_{1.} + \frac{1}{n_1 - 1}(\overline{X}_{1.} - X_{1j}) \right) = n_1 \overline{X}_{1.}^2 - \frac{n_1}{n_1 - 1}\hat{\sigma}_1^2 \\ b_i &= n_1 \overline{X}_{1.} \overline{X}_{i.}, \qquad i = 2, ..., m; \end{aligned}$$

and

$$\begin{aligned} a_{11} &= \sum_{j=1}^{n_1}\left(\overline{X}_{1.}+\frac{1}{n_1-1}(\overline{X}_{1.}-X_{1j})\right)^2 = n_1\overline{X}_{1.}^2+\frac{n_1}{(n_1-1)^2}\hat{\sigma}_1^2 \\ a_{ij} &= n_1\overline{X}_{i.}\overline{X}_{j.}, \qquad i\neq 1 \ \ or \ \ j\neq 1. \end{aligned}$$

It then follows that

$$A = n_1\left(\hat{\theta}_1,\hat{\theta}_2,...,\hat{\theta}_m\right)^t\left(\hat{\theta}_1,\hat{\theta}_2,...,\hat{\theta}_m\right)+D$$

where

$$\begin{aligned} d_{11} &= \frac{n_1}{(n_1-1)^2}\hat{\sigma}_1^2 \\ d_{ij} &= 0, \qquad i\neq 1 \ \ or \ \ j\neq 1. \end{aligned}$$

By the elementary rank inequality, it follows

$$rank\,(A)\leq rank\left(\hat{\theta}^t\hat{\theta}\right)+rank(D)=2.$$

Therefore, we have

$$rank(A)<m \quad if \quad m>2.$$

It then follows that A is not invertible for $m>2$. Thus the Lagrange method will not work in this case since it involves the inversion of the matrix A.

To solve this problem, we can rewrite the objective function $D_u^{(m)}$ in terms of $\lambda_2,\lambda_3,...,\lambda_m$ only, that is, we replace λ_1 by $1-\sum_{i=2}^{m}\lambda_i$. The original minimization problem is then transformed into a minimization problem without any constraint. As we will see in the following derivation, the new objective function is a quadratic function of $\lambda_2,\lambda_3,...,\lambda_m$. Thus, the minimum of the new objective function exists and is unique.

By replacing λ_1 by $1 - \sum_{i=2}^{m} \lambda_i$, we then have

$$\begin{aligned} b(\underline{X})^t \boldsymbol{\lambda}_u &= b_1\lambda_1 + \sum_{i=2}^{m} b_i\lambda_i \\ &= \left(1 - \sum_{i=2}^{m} \lambda_i\right) b_1 + \sum_{i=2}^{m} b_i\lambda_i \\ &= b_1 + \sum_{i=2}^{m} (b_i - b_1)\lambda_i \\ &= b_1 + n_1 \sum_{i=2}^{m} (\overline{X}_{1.}\overline{X}_{i.} - \overline{X}_{1.}^2 + \frac{1}{n_1 - 1}\hat{\sigma}_1^2)\lambda_i \\ &= b_1 + n_1 \sum_{i=2}^{m} \left(\hat{\theta}_1(\hat{\theta}_i - \hat{\theta}_1) + \frac{1}{n_1 - 1}\hat{\sigma}_1^2\right)\lambda_i. \end{aligned}$$

Since A is symmetric, we also have

$$\begin{aligned} \boldsymbol{\lambda}_u^t A \boldsymbol{\lambda}_u &= \lambda_1^2 a_{11} + 2\lambda_1 \sum_{i=2}^{m} \lambda_i a_{1i} + \sum_{i=2}^{m}\sum_{k=2}^{m} \lambda_i a_{ik} \lambda_k \\ &= \left(1 - \sum_{i=2}^{m} \lambda_i\right)^2 a_{11} + 2\left(1 - \sum_{i=2}^{m} \lambda_i\right) \sum_{i=2}^{m} \lambda_i a_{1i} + \sum_{i=2}^{m}\sum_{k=2}^{m} \lambda_i a_{ij} \lambda_k \\ &= \left(a_{11} - 2a_{11} \sum_{i=2}^{m} \lambda_i + \sum_{i=2}^{m}\sum_{k=2}^{m} \lambda_i a_{11} \lambda_k\right) \\ &\quad +2\left(\sum_{i=2}^{m} \lambda_i a_{1i} - \sum_{i=2}^{m}\sum_{k=2}^{m} \lambda_i a_{1i} \lambda_k\right) + \sum_{i=2}^{m}\sum_{k=2}^{m} \lambda_i a_{ij} \lambda_k \\ &= a_{11} - 2\sum_{i=2}^{m} (a_{11} - a_{1i})\lambda_i + \sum_{i=2}^{m}\sum_{k=2}^{m} \lambda_i (a_{ij} + a_{11} - 2a_{i1})\lambda_k. \end{aligned}$$

We then have

$$\begin{aligned} D_u^{(m)} &= c(\underline{X}) - 2b_1 - 2n_1 \sum_{i=2}^{m} \left(\hat{\theta}_1(\hat{\theta}_i - \hat{\theta}_1) + \frac{1}{n_1 - 1}\hat{\sigma}_1^2 \right) \lambda_i \\ &\quad + a_{11} - 2\sum_{i=2}^{m}(a_{11} - a_{1i})\lambda_i + \sum_{i=2}^{m}\sum_{k=2}^{m} \lambda_i(a_{ij} + a_{11} - 2a_{1i})\lambda_k \\ &= a_{11} - 2b_1 + c(\underline{X}) - 2n_1 \sum_{i=2}^{m} \left(\hat{\theta}_1(\hat{\theta}_i - \hat{\theta}_1) + \frac{1}{n_1 - 1}\hat{\sigma}_1^2 + \frac{1}{n_1}(a_{11} - a_{1i}) \right) \lambda_i \\ &\quad + \sum_{i=2}^{m}\sum_{k=2}^{m} \lambda_i(a_{ij} + a_{11} - 2a_{i1})\lambda_k \\ &= a_{11} - 2b_1 + c(\underline{X}) - 2n_1 \sum_{i=2}^{m} \frac{n_1}{(n_1 - 1)^2}\hat{\sigma}_1^2\lambda_i \\ &\quad + n_1 \sum_{i=2}^{m}\sum_{k=2}^{m} \lambda_i \left(\hat{\theta}_i\hat{\theta}_j + \hat{\theta}_1^2 - 2\hat{\theta}_i\hat{\theta}_1 + \frac{1}{(n_1 - 1)^2}\hat{\sigma}_1^2 \right) \lambda_k. \end{aligned}$$

If C is invertible, it then follows that

$$\boldsymbol{\lambda}_u^{opt(-1)} = (\lambda_2, \lambda_3, ..., \lambda_m)^t = \frac{n_1\hat{\sigma}_1^2}{(n_1 - 1)^2} C^{-1}\mathbf{1}$$

where C is a $m-1$ by $m-1$ matrix, and for $i = 1, 2, ...m-1,\ j = 1, 2, ..., m-1$,

$$C_{ij} = \hat{\theta}_{i+1}\hat{\theta}_{j+1} + \hat{\theta}_1^2 - 2\hat{\theta}_{i+1}\hat{\theta}_1 + \frac{1}{(n_1 - 1)^2}\hat{\sigma}_1^2.$$

We then have

$$\boldsymbol{\lambda}_u^{opt} = (\lambda_1, (\boldsymbol{\lambda}_u^{opt(-1)})^t)^t.$$

where $\lambda_1^{opt} = 1 - \mathbf{1}^t\boldsymbol{\lambda}_u^{opt(-1)}$. We remark that C is indeed invertible for $m = 2$ and $m = 3$. It is not clear whether C is invertible for $m > 3$. Therefore, the g-inverse of the matrix A should be considered in order find the optimum weight vector.

5.4 Asymptotic Properties of the Weights

In this section, we derive the asymptotic properties of the cross-validated weights. Let $\hat{\theta}_1^{(n_1)}$ be the MLE based on the first sample of size n_1. Let $\hat{\theta}_1^{(-j)}$ and $\tilde{\theta}_1^{(-j)}$ respectively

be the MLE and WLE based on m samples without the jth data point from the first sample. This generalizes the two cases where either only the jth data point is deleted from the first sample or jth data point from each sample is deleted. Note that $\tilde{\theta}_1^{(-j)}$ is a function of the weight function $\boldsymbol{\lambda}$. Let $\frac{1}{n_1}D_{n_1}$ be the average discrepancy in the cross-validation which is defined as

$$\frac{1}{n_1}D_{n_1}(\boldsymbol{\lambda}) = \frac{1}{n_1}\sum_{j=1}^{n_1}\left(X_{1j} - \phi(\tilde{\theta}_1^{(-j)})\right)^2.$$

Let $\boldsymbol{\lambda}^{(cv)}$ be the optimum weights chosen by using the cross-validation. Let $\theta^0 = (\theta_1^0, \theta_2, ..., \theta_m))$ where θ_1^0 is the true values for θ_1. We then have the following theorem.

Theorem 5.1 *Assume that*

(1) $\frac{1}{n_1}D_{n_1}$ *has a unique minimum for any fixed* n_1*;*

(2) $\frac{1}{n_1}\sum_{j=1}^{n_1}\left(\phi(\hat{\theta}_1^{(-j)}) - \phi(\theta_1^0)\right) \xrightarrow{P_{\theta^0}} 0$ *as* $n_1 \to \infty$*;*

(3) $P_{\theta^0}\left(\frac{1}{n_1}\sum_{j=1}^{n_1}\left(X_{1j} - \phi(\hat{\theta}_1^{(-j)})\right)^2 < K\right) \xrightarrow{P_{\theta^0}} 1$ *for some constant* $0 < K < \infty$*;*

(4) $P_{\theta^0}\left(\left|\phi(\hat{\theta}_1^{n_1}) - \phi(\tilde{\theta}_1^{n_1})\right| > M\right) = o(\frac{1}{n_1})$ *for some constant* $0 < M < \infty$*;*

then

$$\boldsymbol{\lambda}^{(cv)} \xrightarrow{P_{\theta^0}} \boldsymbol{w}_0 = (1, 0, 0, ..., 0)^t. \tag{5.6}$$

Proof: Consider

$$\begin{aligned}
\frac{1}{n_1}D_{n_1}(\boldsymbol{\lambda}) &= \frac{1}{n_1}\sum_{j=1}^{n_1}\left(X_{1j} - \phi(\tilde{\theta}_1^{(-j)})\right)^2 \\
&= \frac{1}{n_1}\sum_{j=1}^{n_1}\left(\left(X_{1j} - \phi(\hat{\theta}_1^{(-j)})\right) + \left(\phi(\hat{\theta}_1^{(-j)}) - \phi(\tilde{\theta}_1^{(-j)})\right)\right)^2 \\
&= \frac{1}{n_1}\sum_{j=1}^{n_1}\left(X_{1j} - \phi(\hat{\theta}_1^{(-j)})\right)^2 + \frac{1}{n_1}\sum_{j=1}^{n_1}\left(\phi(\hat{\theta}_1^{-j}) - \phi(\tilde{\theta}_1^{(-j)})\right)^2 \\
&\quad + \frac{2}{n_1}\sum_{j=1}^{n_1}\left(X_{1j} - \phi(\hat{\theta}_1^{(-j)})\right)\left(\phi(\hat{\theta}_1^{(-j)}) - \phi(\tilde{\theta}_1^{(-j)})\right).
\end{aligned}$$

Note that

$$\begin{aligned}
&\frac{1}{n_1}\sum_{j=1}^{n_1}\left(X_{1j}-\phi(\hat{\theta}_1^{(-j)})\right)\left(\phi(\hat{\theta}_1^{(-j)})-\phi(\tilde{\theta}_1^{(-j)})\right)\\
&=\frac{1}{n_1}\sum_{j=1}^{n_1}\left(X_{1j}-\phi(\theta_1^0)\right)\left(\phi(\hat{\theta}_1^{(-j)})-\phi(\tilde{\theta}_1^{(-j)})\right)\\
&\quad+\frac{1}{n_1}\sum_{j=1}^{n_1}\left(\phi(\theta_1^0)-\phi(\hat{\theta}_1^{(-j)})\right)\left(\phi(\hat{\theta}_1^{(-j)})-\phi(\tilde{\theta}_1^{(-j)})\right)\\
&=S_1+S_2
\end{aligned}$$

where

$$\begin{aligned}
S_1 &= \frac{1}{n_1}\sum_{j=1}^{n_1}\left(X_{1j}-\phi(\theta_1^0)\right)\left(\phi(\hat{\theta}_1^{(-j)})-\phi(\tilde{\theta}_1^{(-j)})\right),\\
S_2 &= \frac{1}{n_1}\sum_{j=1}^{n_1}\left(\phi(\theta_1^0)-\phi(\hat{\theta}_1^{(-j)})\right)\left(\phi(\hat{\theta}_1^{(-j)})-\phi(\tilde{\theta}_1^{(-j)})\right).
\end{aligned}$$

We first show that $S_1 \xrightarrow{P_{\theta^0}} 0$.

Consider

$$\begin{aligned}
&P_{\theta^0}\left(|S_1|>\epsilon\right)\\
=\;&P_{\theta^0}\left(\epsilon<|S_1| \;\; and \;\; \left|\phi(\hat{\theta}_1^{(-j)})-\phi(\tilde{\theta}_1^{(-j)})\right|<M \;\; for \;\; all \;\; j\right)\\
&+P_{\theta^0}\left(\epsilon<|S_1| \;\; and \;\; \left|\phi(\hat{\theta}_1^{(-l)})-\phi(\tilde{\theta}_1^{(-l)})\right|\geq M \;\; for \;\; some \;\; l\right)\\
\leq\;&P_{\theta^0}\left(\epsilon<|S_1|<\frac{M}{n_1}\sum_{j=1}^{n_1}\left|X_{1j}-\phi(\theta_1^0)\right|\right)+\sum_{l=1}^{n_1}P_{\theta^0}\left(\left|\phi(\hat{\theta}_1^{(-l)})-\phi(\tilde{\theta}_1^{(-l)})\right|\geq M\right)\\
\leq\;&P_{\theta^0}\left(\frac{\epsilon}{M}<\left|\frac{1}{n_1}\sum_{j=1}^{n_1}(X_{1j}-\phi(\theta_1^0))\right|\right)+n_1P_{\theta^0}\left(|\phi(\hat{\theta}_1^{(-1)})-\phi(\tilde{\theta}_1^{(-1)})|\geq M\right)\\
=\;&P_{\theta^0}\left(\left|\frac{1}{n_1}\sum_{j=1}^{n_1}\left(X_{1j}-\phi(\theta_1^0)\right)\right|>\frac{1}{M}\epsilon\right)+n_1P_{\theta^0}\left(|\phi(\hat{\theta}_1^{(n_1-1)})-\phi(\tilde{\theta}_1^{(n_1-1)})|\geq M\right)
\end{aligned}$$

The first term goes to zero by the Weak Law of Large Numbers. The second term also goes to zero by assumption (4). We then have

$$P_{\theta^0}(|S_1|>\epsilon)\longrightarrow 0 \;\; as \;\; n_1 \;\; \to\infty. \tag{5.7}$$

We next show that $S_2 \xrightarrow{P_{\theta^0}} 0$ as $n_1 \to \infty$.

Consider

$$\begin{aligned}
& P_{\theta^0}(|S_2| > \epsilon) \\
= \; & P_{\theta^0}\left(\epsilon < |S_2| \;\; and \;\; \left|\phi(\hat{\theta}_1^{(-j)}) - \phi(\tilde{\theta}_1^{(-j)})\right| < M \;\; for \;\; all \;\; j\right) \\
& + P_{\theta^0}\left(\epsilon < |S_2| \;\; and \;\; \left|\phi(\hat{\theta}_1^{(-l)}) - \phi(\tilde{\theta}_1^{(-l)})\right| \geq M \;\; for \;\; some \;\; l\right) \\
\leq \; & P_{\theta^0}\left(\epsilon < |S_2| < \frac{M}{n_1}\left|\sum_{j=1}^{n_1}(\phi(\hat{\theta}_1^{(-j)}) - \phi(\theta_1^0))\right|\right) + \sum_{l=1}^{n_1} P_{\theta^0}\left(\left|\phi(\hat{\theta}_1^{(-l)}) - \phi(\tilde{\theta}_1^{(-l)})\right| \geq M\right) \\
\leq \; & P_{\theta^0}\left(\frac{1}{M}\epsilon < \left|\frac{1}{n_1}\sum_{j=1}^{n_1}(\phi(\hat{\theta}_1^{(-j)}) - \phi(\theta_1^0))\right|\right) + n_1 P_{\theta^0}\left(\left|\phi(\hat{\theta}_1^{(-1)}) - \phi(\tilde{\theta}_1^{(-1)})\right| \geq M\right) \\
= \; & P_{\theta^0}\left(\left|\frac{1}{n_1}\sum_{j=1}^{n_1}\left(\phi(\hat{\theta}_1^{(-j)}) - \phi(\theta_1^0)\right)\right| > \frac{1}{M}\epsilon\right) + n_1 P_{\theta^0}\left(\left|\phi(\hat{\theta}_1^{(n_1-1)}) - \phi(\tilde{\theta}_1^{(n_1-1)})\right| \geq M\right)
\end{aligned}$$

The first term goes to zero by assumption (2). The second term also goes to zero by assumption (4). We then have

$$P_{\theta^0}(|S_2| > \epsilon) \longrightarrow 0 \;\; as \;\; n_1 \;\; \to \infty. \tag{5.8}$$

It then follows that

$$\frac{1}{n_1}D_{n_1}(\boldsymbol{\lambda}) = \frac{1}{n_1}\sum_{j=1}^{n_1}\left(X_{1j} - \phi(\hat{\theta}_1^{(-j)})\right)^2 + \frac{1}{n_1}\sum_{j=1}^{n_1}\left(\phi(\hat{\theta}_1^{-j}) - \phi(\tilde{\theta}_1^{(-j)})\right)^2 + R_n \tag{5.9}$$

where $R_n \xrightarrow{P_{\theta^0}} 0$. Observe that the first term is independent of $\boldsymbol{\lambda}$. Therefore the second term must be minimized with respect to $\boldsymbol{\lambda}$ to obtain the minimum of $\frac{1}{n_1}D_{n_1}(\boldsymbol{\lambda})$. We see that the second term is always non-negative. It then follows that, with probability tending to 1,

$$\frac{1}{n_1}D_{n_1}(\boldsymbol{\lambda}) \geq \frac{1}{n_1}\sum_{j=1}^{n_1}\left(X_{1j} - \phi(\hat{\theta}_1^{(-j)})\right)^2 = \frac{1}{n_1}D_{n_1}(\boldsymbol{w}),$$

since $\phi(\hat{\theta}_1^{(-j)}) = \phi(\tilde{\theta}_1^{(-j)})$ for $\boldsymbol{\lambda}^{(cv)} = \boldsymbol{w}_0 = (1, 0, 0, ..., 0)^t$ for fixed n_1.

Finally, we will show that

$$\boldsymbol{\lambda}^{(cv)} \xrightarrow{P_{\theta^0}} \boldsymbol{w}_0, \quad as \;\; n_1 \to \infty.$$

Suppose that $\boldsymbol{\lambda}^{(cv)} \xrightarrow{P_{\theta^0}} \boldsymbol{w}_0 + \boldsymbol{d}$ where d is a non-zero vector. Then there exist n_0 such that for $n_1 > n_0$,

$$\frac{1}{n_1} D_{n_1}\left(\boldsymbol{\lambda}^{(cv)}\right) \geq \frac{1}{n_1} D_{n_1}(\boldsymbol{w}).$$

This is a contradiction because $\boldsymbol{\lambda}^{(cv)}$ is the vector which minimizes $\frac{1}{n_1} D_{n_1}$ for any fixed n_1 and the minimum of $\frac{1}{n_1} D_{n_1}(\boldsymbol{\lambda})$ is unique by assumption. $\diamond$

To check the assumptions of the above theorem, let us consider the linear WLE for two samples with equal sample sizes. Assumption (1) is satisfied since $\frac{1}{n_1} D_{n_1}(\boldsymbol{\lambda})$ is a quadratic form in $\boldsymbol{\lambda}$ and its minimum is indeed unique for each fixed n_1. To check Assumption (2), consider

$$\begin{aligned}
\frac{1}{n_1}\sum_{j=1}^{n_1}\left(\phi(\hat{\theta}_1^{(-j)}) - \phi(\theta_1^0)\right) &= \frac{1}{n_1}\sum_{j=1}^{n_1}\left(\overline{X}_{1.}^{(-j)} - \theta_1^0\right) \\
&= \frac{1}{n_1}\sum_{j=1}^{n_1}\left(\frac{1}{n_1-1}\sum_{l\neq j} X_{1l} - \theta_1^0\right) \\
&= \frac{1}{n_1}\sum_{j=1}^{n_1}\left(\frac{n_1}{n_1-1}\overline{X}_{1.} - \frac{1}{n_1-1}X_{1j}\right) - \theta_1^0 \\
&= \overline{X}_{1.} - \theta_1^0 \xrightarrow{P_{\theta^0}} 0 \;\; as \;\; n_1 \to \infty.
\end{aligned}$$

Next we consider

$$\begin{aligned}
\frac{1}{n_1}\sum_{j=1}^{n_1}\left(X_{1j} - \phi(\hat{\theta}_1^{(-j)})\right)^2 &= \frac{1}{n_1}\sum_{j=1}^{n_1}\left(X_{1j} - \overline{X}_{1.}^{(-j)}\right)^2 \\
&= \frac{1}{n_1}\sum_{j=1}^{n_1}\left(X_{1j} - \left(\frac{n_1}{n_1-1}\overline{X}_{1.} - \frac{1}{n_1-1}X_{1j}\right)\right)^2 \\
&= \left(\frac{n_1}{n_1-1}\right)^2 \frac{1}{n_1}\sum_{j=1}^{n_1}\left(X_{1j} - \overline{X}_{1.}\right)^2 \xrightarrow{P_{\theta^0}} var(X_{11}) < \infty \quad as \;\; n_1 \to \infty.
\end{aligned}$$

For the last assumption of the previous theorem, consider

$$\begin{aligned}\left|\phi(\hat{\theta}_1^{(n_1)}) - \phi(\tilde{\theta}_1^{(n_1)})\right| &= \left|\overline{X}_{1.} - (\lambda_1^{(cv)}\overline{X}_{1.} + \lambda_2^{(cv)}\overline{X}_{2.})\right| \\ &= \left|\lambda_2^{(cv)}\right| \left|\overline{X}_{1.} - \overline{X}_{2.}\right|.\end{aligned}$$

It then follows from Lemma 5.1 that

$$\begin{aligned}P_{\theta^0}\left(\left|\phi(\hat{\theta}_1^{(n_1)}) - \phi(\tilde{\theta}_1^{(n_1)})\right| > \epsilon\right) &= P_{\theta^0}\left(\frac{1}{n-2}\left|\frac{(\hat{\sigma}_1^2 - \widehat{cov})^2(\bar{X}_{1.} - \bar{X}_{2.})^2}{[(\bar{X}_{1.} - \bar{X}_{2.})^2 + \frac{1}{n^2}\sum_{i=1}^{n}(X_{1j} - X_{2j})^2]^2}\right| > \epsilon\right) \\ &\leq \frac{1}{(n-2)^2\epsilon^2} E_{\theta^0}\left|\frac{(\hat{\sigma}_1^2 - \widehat{cov})^2(\bar{X}_{1.} - \bar{X}_{2.})^2}{(\bar{X}_{1.} - \bar{X}_{2.})^4}\right| \\ &\leq \frac{1}{(n-2)^2\epsilon^2} E_{\theta^0}\left|\frac{\hat{\sigma}_1^2 - \widehat{cov}}{\bar{X}_{1.} - \bar{X}_{2.}}\right|^2 = o(n)\end{aligned}$$

since $\bar{X}_{1.} - \bar{X}_{2.} \xrightarrow{a.s.} \theta_1^0 - \theta_2^0 \neq 0$ and $\hat{\sigma}_1^2 - \widehat{cov} \xrightarrow{a.s.} \sigma_1^2 - cov(X_{11}, X_{21})$. Thus the assumptions of the theorem are all satisfied. We then have

$$\left|\lambda_2^{(cv)}\right| \xrightarrow{P_{\theta^0}} 1; \qquad \left|\lambda_2^{(cv)}\right| \xrightarrow{P_{\theta^0}} 0.$$

This is consistent with the result of Proposition 5.2.

Since the cross-validated weight function converges in probability to $\boldsymbol{w}_0$. Therefore the asymptotic normality of $\tilde{\theta}_1$ of using cross-validated weights follows by Theorem 4.10.

5.5 Simulation Studies

To demonstrate and verify the benefits of using cross-validation procedures described in previous sections, we perform simulations according to the following algorithm which deletes jth point from each sample, i.e. delete-one-column approach. *Step 1: Draw random samples of size n from* $f_1(x;\theta_1^0)$ *and* $f_2(x;\theta_2^0)$;

n	MSE(MLE)	SD of (MLE-θ_1^0)^2	MSE(WLE)	SD of (WLE-θ_1^0)^2	$\frac{MSE(WLE)}{MSE(MLE)}$
5	0.203	0.451	0.130	0.360	0.638
10	0.100	0.317	0.075	0.274	0.751
15	0.069	0.262	0.057	0.238	0.826
20	0.051	0.227	0.042	0.204	0.809
25	0.041	0.203	0.035	0.187	0.843
30	0.035	0.187	0.031	0.177	0.895
35	0.030	0.173	0.028	0.166	0.931
40	0.025	0.159	0.023	0.153	0.932
45	0.023	0.151	0.023	0.151	0.997
50	0.020	0.141	0.020	0.141	1.007
55	0.018	0.135	0.019	0.139	1.066
60	0.017	0.129	0.018	0.133	1.057

Table 5.1: MSE of the MLE and the WLE and standard deviations of the squared errors for samples with equal sizes for $N(0,1)$ and $N(0.3,1)$.

Step 2: Calculate the cross-validated optimum weights by using (5.2);

Step 3. Calculate the (MLE- θ_1^0)^2 and (WLE- θ_1^0)^2;

Repeat Step 1 - 3 for 1000 times. Calculate the averages and standard deviations of the squared differences of MLE and WLE to the value of the true parameter θ_1^0 respectively. Calculate the averages and standard deviations of the optimum weights.

We generate random samples from $N(0,\sigma_1^2)$ and $N(c,\sigma_2^2)$. For simplicity, we set $\sigma_1 = \sigma_2 = 1$. For the purpose of the demonstration, we set $c = 0.3$ which is 30% of the variance. Other values were tried. In general, the larger the value of c, the less improvement in the MSE. For example, if we set $\sigma_1 = \sigma_2 = 1$ and $c = 1$, the ratio of the MSE for MLE and WLE will almost be 1 which means that the cross-validation procedure will not consider the second sample useful in that case. Some of the result

for $c = 0.3$ is shown Table 5.1.

n	AVE. of λ_1	AVE. of λ_2	SD of λ_1 and λ_2
5	0.710	0.290	0.027
10	0.725	0.275	0.053
15	0.734	0.266	0.064
20	0.750	0.250	0.075
25	0.755	0.245	0.080
30	0.765	0.235	0.085
35	0.777	0.223	0.089
40	0.785	0.216	0.092
45	0.785	0.215	0.092
50	0.788	0.220	0.094
55	0.792	0.208	0.094
60	0.807	0.193	0.097

Table 5.2: Optimum weights and their standard deviations for samples with equal sizes from $N(0,1)$ and $N(0.3,1)$.

It is obvious from the table that the MSE of WLE is much smaller than that of the MLE for small and moderate sample sizes. The standard deviations of the squared differences for the WLE is less or equal to that of the MLE. This suggests that not only does WLE achieve smaller MSE but also its MSE has less variation than that of MLE. Intuitively, as the sample size increases, the importance of the second sample diminishes. As indicated by Table 5.2, the cross-validation procedure begins to realize this by assigning a larger value to λ_1 as the sample size increases. Although quite slowly, the optimum weights do increase towards the asymptotic weights $(1,0)$ for the normal case.

We repeat the algorithm for Poisson distribution with $\mathcal{P}(3)$ and $\mathcal{P}(3.6)$. Some of the results are shown in Table 5.3 and Table 5.4. The result for Poisson distributions is somewhat different than that from the normal. The striking difference can be seen from the ratio of the average MSE of WLE and average of MSE of WLE. The WLE achieves smaller MSE on average when the sample sizes are less than 30. Once it is over 30, it seems that we should not combine the two samples. This is not the case for the normal until the sample size reaches 45.

We remark that the reduction in MSE will disappear if we set $c = 1.5$ in the above case. Thus the cross-validation procedure will not combine two samples if the second sample does not help to predict the behavior of the first. We also emphasize that the value c in both cases are not assumed to be known to the cross-validation procedure.

n	MSE(mle)	SD	MSE(wle)	SD	$\frac{MSE(wle)}{MSE(mle)}$
5	0.312	0.558	0.235	0.484	0.753
15	0.142	0.377	0.127	0.357	0.896
25	0.120	0.347	0.114	0.338	0.950
30	0.104	0.323	0.104	0.323	1.000
35	0.077	0.277	0.081	0.284	1.054
40	0.074	0.272	0.076	0.275	1.025
45	0.072	0.268	0.075	0.274	1.045
50	0.057	0.238	0.065	0.255	1.141
55	0.054	0.233	0.060	0.245	1.098
60	0.046	0.215	0.052	0.229	1.132

Table 5.3: MSE of the MLE and the WLE and their Standard deviations for samples with equal sizes from $\mathcal{P}(3)$ and $\mathcal{P}(3.6)$.

n	AVE. of λ_1	AVE. of λ_2	SD of λ_1 and λ_2
5	0.710	0.289	0.027
10	0.729	0.270	0.057
15	0.738	0.261	0.065
20	0.754	0.245	0.077
25	0.754	0.245	0.078
30	0.768	0.231	0.086
35	0.777	0.222	0.091
40	0.789	0.210	0.093
45	0.797	0.202	0.097
50	0.799	0.200	0.095
55	0.812	0.187	0.097
60	0.820	0.179	0.096

Table 5.4: Optimum weights and their standard deviations for samples with equal sizes from $\mathcal{P}(3)$ and $\mathcal{P}(3.6)$

We remark that simulations of using the *delete-one-point* approach have also been done. They give quite similar results.

Chapter 6

Derivations of the Weighted Likelihood Functions

In this chapter, we shall develop the theoretical foundation for using weighted likelihood. Hu and Zidek (1997) discuss the connection between relevance weighted likelihood and maximum entropy principle for the discrete case. We also show that the weighted likelihood function can be derived from maximum entropy principle advocated by Akaike for the continuous case.

6.1 Introduction

Akaike (1985) reviewed the historical development of entropy and discussed the importance of the maximum entropy principle. Hu and Zidek (1997) discovered the connection between relevance weighted likelihood and maximum entropy principle for the discrete case. We offer a proof for the continuous case.

We first state the maximum entropy principle: *all statistical activities are directed to maximize the expected entropy of the predictive distribution in each particular application.* When a parametric model $\{p(x;\theta);\theta \in \Theta\}$ of the distribution of a future

observation x is given, the goodness of a particular model $\{p(x;\theta);\theta \in \Theta\}$ as the predictive distribution of x is evaluated by the relative entropy

$$B(f;p(.;\theta)) = -I(f,p(.;\theta)) = -\int f(x)log\frac{f(x)}{p(x;\theta)}dx.$$

where $f(x)$ is the true distribution. In this expression, $logf(x)/p(x;\theta)$ is defined as $+\infty$ if $f(x) > 0$ and $p(x;\theta)$, so the expectation could be $+\infty$. Although $logf(x)/p(x;\theta)$ is defined as $-\infty$ when $f(x) = 0$ and $p(x;\theta) > 0$, the integrand, $f(x)logf(x)/p(x;\theta)$ is defined as zero in this case. The relative entropy is a natural measure of discrepancy between two probability functions.

Hence, maximizing $B(f;p(.;\theta))$ is equivalent to minimizing $I(f,p(.;\theta))$ with respect to θ. Without any restrictions, the desired density function is $f(x)$ itself. However, the true density function $f(x)$ is indeed unknown. We therefore use a set of density functions, $f_1(x), f_2(x), ..., f_m(x)$, say, to represent the true density function. The density function $f_1(x)$ represents the density function which is thought to be the "closest" to the true density function $f(x)$. This resembles the idea of compromised MLE proposed by Easton (1991).

To be consistent with our use of relative entropy, we use it in interpreting "resemblance" of any density function $g(x)$ to the $f_i(x)$ and define that term to mean

$$\int f_i(x)logf_i(x)/g(x)dx \leq a_i, \quad i = 1,2,3...,m.$$

The a_i reflects the maximum allowable deviation from the density function $f_i(x)$. If a_i is set to be zero, then $g(x)$ takes exact the same functional form of $f_i(x)$.

For a given set of density functions, we seek a probability density function which minimizes $I(f_1,g) = \int f_1(x)log\,\frac{f_1(x)}{g(x)}dx$ over all probability densities g satisfying

$$\int f_i(x)log\frac{f_i(x)}{g(x)}dx \leq a_i, \quad i = 2,...,m, \tag{6.1}$$

where a_i are finite non-negative constants. This idea of minimizing the relative entropy under certain constraint is also similar to the approach outlined in Kullback(1959, Chapter 3) for the hypothesis testing.

6.2 Existence of the Optimal Density

Let V be a reflexive Banach space. Let $\mathcal{E}$ be a non-empty closed convex subset of V. Define a function $I(g)$ on $\mathcal{E}$ into $\mathcal{R}$ where g is a continuous function.

We are concerned with minimization problem:

$$\inf_{g\in\mathcal{E}} I(g).$$

To avoid trivial cases, we assume that the function $I(g)$ is proper, *i.e.* it does not take the value $-\infty$ and is not identically equal to $+\infty$. We then have the following theorem.

Theorem 6.1 *Assume that $I(g)$ is convex, lower semi-continuous and proper with respect to g. In addition, assume that the set $\mathcal{D}$ is bounded, i.e. there exist a constant M, say, such that*

$$\sup_{g\in\mathcal{D}} I(g) < M.$$

Then the minimization problem has at least one solution. The solution is unique if the function $I(g)$ is strictly convex on $\mathcal{D}$.

Proof: (See, for example, Ekeland 1976, p 35.) $\diamond$

Consider L^p spaces $(1 < p < \infty)$. It is known that L^p $(1 < p < \infty)$ is reflexive (Royden 1988, 227). Define $I(g) = \int f_1(x) log \frac{f_1(x)}{g(x)} dx$ for some given density $f_1(x)$. It can be seen that $I(g)$ is a proper convex function and also continuous in g on L^p. We also assume that $I(g) < \infty$.

Define

$$\mathcal{E}_i = \{g : \int f_i(x) log \frac{f_i(x)}{g(x)} dx \leq a_i, \int g(x)dx = 1, g(x) \geq 0, g \in L^p\}, \quad i = 2, ..., m,$$

and

$$\mathcal{E} = \cap_{i=2}^{m-1} \mathcal{E}_i.$$

Lemma 6.1 *The following inequality holds:*

$$V^2(f_1, f_2)/2 \leq I(f_1, f_2)$$

where $V(f_1, f_2) = \int |f_1(x) - f_2(x)|dx$.

Proof: (See Kullback, 1954) $\diamond$

Lemma 6.2 *If* $|g(x) - f_i(x)|^{p-1} \leq B_i$, *for all* x, $1 < p < \infty$ *and* B_i *is a constant,* i=2,...,m, *then the set* $\mathcal{E}$ *is bounded subset of* L^p.

Proof: It suffices to show that each $\mathcal{E}_i$ is bounded in L^p. For any density in $\mathcal{E}_i, i = 2, .., m$, we have

$$\begin{aligned} \int |g(x) - f_i(x)|^p dx &= \int |g(x) - f_i(x)||g(x) - f_i(x)|^{p-1} dx \\ &\leq B_i \int |g(x) - f_i(x)|dx \\ &\leq B_i \sqrt{2I(f_i, g)} \\ &= B_i \sqrt{2a_i}. \end{aligned}$$

This implies that $||g - f_i||_p < C_i$.

By the Mankowski Inequality for L^p spaces with $1 < p < \infty$, we have

$$||g||_p \leq ||g - f_i||_p + ||f_i||_p < C_i + ||f_i||_p.$$

It then follows that $\mathcal{E}_i$ is bounded in L^p. $\diamond$

We are now in a position to establish the existence of the optimal solution.

Theorem 6.2 *For a given set of density functions $f_1, f_2, ..., f_m$, the minimization problem (6.1) has a unique solution if the admissible density functions satisfy the conditions*

$$|g(x) - f_i(x)|^{p-1} \leq B_i, \quad for \ all \ \ x \ \ and \ \ g \in L^p, 1 < p < \infty.$$

Proof: Note that *log* function is strictly convex. The assertion of this theorem follows from Theorem 6.1 and Lemma 6.2. $\diamond$

6.3 Solution to the Isoperimetric Problem

In order to find the solution of the problem, it is useful to state certain results in the calculus of variations. In 1744, Euler formulated the general *Isoperimetric problem* as the following:

Find a vector function $(y_1, y_2, ..., y_m)$ which minimizes the functional

$$\mathcal{I}[y_1, y_2, ..., y_m] = \int_a^b f(x, y_1(x), ..., y_m(x), y_1'(x), ..., y_m'(x))dx, \tag{6.2}$$

and satisfies the initial conditions

$$y_i(a) = y_i^a, \quad i = 1, 2, .., m, \tag{6.3}$$

as well as the terminal conditions

$$y_i(b) = y_i^b, \quad i = 1, 2, .., m, \tag{6.4}$$

while

$$\int_a^b f_i(x, y_1(x), ..., y_m(x), y_1'(x), ..., y_m'(x))dx = l_i, \quad i = 1, 2, ..., m, \tag{6.5}$$

where $l_1, l_2, ..., l_m$ are given numbers.

We now state a fundamental theorem in the calculus of variations.

Theorem 6.3 *For* $(y_1, y_2, ...y_m) \in C^1[a,b]^m$ *to be a solution of the Isoperimetric problem [(6.2)-(6.5)], it is necessary that there exist* $m+1$ *constants* $\boldsymbol{t} = (t_0, t_1, ..., t_m) \neq (0, ..., 0)$ *such that, for* $k = 1, 2, ..., m$,

$$h^I_{y_k}(x, y_1, y_2, ..., y_m, y_1', y_2', ..., y_m', \boldsymbol{t}) - \frac{\partial}{\partial x} h^I_{y_k'}(x, y_1, y_2, ..., y_m, y_1', y_2', ..., y_m', \boldsymbol{t}) = 0 \quad (6.6)$$

where $h^I_{y_k} = \frac{\partial h^I}{\partial y_k}$ *and*

$$\begin{aligned} h^I(x, y_1, y_2, ..., y_m, y_1', y_2', ..., y_m', \boldsymbol{t}) &= t_0 f(x, y_1, y_2, ..., y_m, y_1', y_2', ..., y_m') \\ &\quad + \sum_{\rho=1}^{m} t_\rho f_\rho(x, y_1, y_2, ..., y_m, y_1', y_2', ..., y_m'). \end{aligned} \quad (6.7)$$

Proof: See, for example, Theorem 2 on p.90 in Giaquinta and Hildebrant (1996). $\diamond$

Note that the values of a and b are arbitrary. They can take the value of $+\infty$ and $-\infty$. The original proof is given in Bliss (1930). Detailed discussions can also be found in that paper.

6.4 Derivation of the WL Functions

We now establish the necessary condition for the optimal solution to the minimization problem. Assume that the density functions $f_1, f_2, ..., f_m$ are continuous and twice differentiable.

Theorem 6.4 *(Necessary Condition) For* g^* *to be the optimal solution, it is necessary that it is a mxiture distribution, i.e., there exist constants* $t_1^*, t_2^*, ..., t_m^*$ *such that* $\sum_{i=1}^{m} t_i^* = 1$, *and*

$$g^*(x) = \sum_{k=1}^{m} t_k^* f_k(x) \geq 0. \quad (6.8)$$

Proof: There exist constants $b_2, b_3, ..., b_m$ such that $I(f_i, g) = b_i \leq a_i$, $i = 2, 3, ..., m$ for any particular choice of g satisfying the constraints (6.1). We seek the optimal

solution which is a point in a certain manifold in the function spaces. Thus the optimization problem can be re-formulated in the context of calculus of variations as follows

$$\min_{g \in E} I(g) = \min_{g} \int f_1(x)\ log \frac{f_1(x)}{g(x)}\ dx$$

such that g satisfies the following constraints:

$$\int f_i(x)\ log \frac{f_1(x)}{g(x)}\ dx = b_i,\ \ i = 2, ..., m;$$
$$\int g(x)dx = 1\ \ and\ \ g(x) \geq 0.$$

Define $\psi(x, g) = f_1(x) log \frac{f_1(x)}{g(x)} + l_0 g(x) + \sum_{k=1}^{m} l_k f_k(x) log \frac{f_k(x)}{g(x)}$. By Theorem 6.3, it follows that the necessary condition for g^* to be the optimal solution is that it has to satisfy the *Euler-Lagrange* equation, i.e.

$$\frac{\partial \psi}{\partial g} - \frac{\partial}{\partial x}(\frac{\partial \psi}{\partial g'}) = 0, \tag{6.9}$$

where $g' = \frac{\partial g}{\partial x}$.

Notice that $\psi(x, g)$ is not a function of g'. It implies that $\frac{\partial \psi}{\partial g'} = 0$. The Euler equation then becomes

$$\frac{\partial \psi}{\partial g} = 0.$$

It follows that

$$-\frac{f_1(x)}{g(x)} + l_0 - \sum_{k=1}^{m} l_k \frac{f_k(x)}{g(x)} = 0.$$

We then have

$$g^*(x) = \sum_{k=1}^{m} t_k^* f_k(x),$$

where $t_1^* = 1/l_0, t_i^* = l_k/l_0, k = 2, ..., m$.

Since we seek a density function, it then follows that the sum of the t_i must be 1 since $\int g^*(x)dx = \sum_{k=1}^{m} t_k^* = 1$. It also follows that $g^*(x) = \sum_{k=1}^{m} t_k^* f_k(x) \geq 0$.

Since if $g^*(x) = \sum_{k=1}^{m} t_k^* f_k(x) < 0$ for all $x \in \mathcal{K}$ with $P_i(\mathcal{K}) > 0$, the constraints $\int f_i(x) log(f_i(x)/g^*(x))dx = b_i \geq 0$ then will not be satisfied. This completes the proof. $\diamond$

Consider the minimization problem without any constraint. We then seek the optimal density function g^* such that it minimizes $I(f_1, g)$ for any given $f_1(x)$. According to Theorem 6.4, the necessary condition of the optimal function g^* is that

$$g^*(x) = t_1^* f_1(x).$$

Since $t_i^* = 0, i = 2, 3, .., m$, then $t_1^* = 1$. It then follows that $g^*(x) = f_1(x)$, $a.e.$. Furthermore, we have $I(f_1, g) \geq I(f_1, g^*) = I(f_1, f_1) = 0$ for any density function g. This result is also known as the *Shannon-Kolmogorov Information Inequality.*

We establish the uniqueness of the optimal solution in the next theorem.

Theorem 6.5 *(Uniqueness) Suppose* $g^* = \sum_{i=1}^{m} t_i^* f_i(x)$ *with the* t_i^* *chosen so that* g^* *satisfies the constraints (6.1) and* $\sum_{i=1}^{m} t_i^* = 1, 0 \leq t_i^* \leq 1, \ i = 1, 2, ..., m.$ *Then* g^* *uniquely minimizes* $I(f_1, g)$ *over all probability densities* g *satisfying constraints (6.1).*

Proof: Suppose that there exist a probability density function g_0 such that

$$\int f_1(x)\ log \frac{f_1(x)}{g_0(x)}\ dx \leq \int f_1(x)\ log\ \frac{f_1(x)}{g^*(x)}\ dx,$$

while

$$\int f_i(x) log\ \frac{f_i(x)}{g_0(x)}\ dx \leq a_i, \ \ i = 2, ..., m.$$

It follows that

$$\int f_1(x)\ log\ g^*(x)\ dx \leq \int f_1(x)\ log\ g_0(x)\ dx$$

while

$$\int f_i(x)\ log\ g^*(x)\ dx \leq \int f_i(x)\ log\ g_0(x)\ dx, \ \ i = 2, ..., m.$$

We then have

$$\begin{aligned}
\sum_{i=1}^{m} t_i^* \int f_i(x)\ log\ g^*(x)\ dx &\leq \sum_{i=1}^{m} t_i^* \int f_i(x)\ log\ g_0(x)\ dx \\
\int \sum_{i=1}^{m} t_i^* f_i(x)\ log\ g^*(x)\ dx &\leq \int \sum_{i=1}^{m} t_i^* f_i(x)\ log\ g_0(x)\ dx \\
\int g^*(x)\ log\ g^*(x)\ dx &\leq \int g^*(x)\ log\ g_0(x)\ dx
\end{aligned}$$

It follows that $I(g^*, g_0) \leq 0$. However, we know that $I(g^*, g_0) \geq 0$.

Therefore, $g^*(x) = g_0(x)$ for all x. This completes the proof. $\diamond$

The weights t_i^* are functions of $a_1, a_2, ..., a_m$. To describe the relationships between t_i^* and a_i, we have the next theorem.

Theorem 6.6 *Suppose there exists* $\mathbf{a}^0 = (a_1, a_2, ..., a_m)^t$ *and* $\delta^0 = (\delta_1, \delta_2, ..., \delta_m)^t$ *such that there exists* $g_0 = \sum_{i=1}^{m} t_i f_i(x)$ *with* t_i *chosen so that* g_0 *achieves the equalities in the constraints (6.1) and* $\sum_{i=1}^{m} t_i = 1, 0 \leq t_i \leq 1,\ i = 1, 2, ..., m$ *for any* $\mathbf{a}$ *such that* $|a_i - a_i^0| < \delta_i^0$. *Then* t_i *are monotone functions of* a_i. *Moreover,*

$$\begin{aligned}
\frac{\partial t_i}{\partial a_i} &\leq 0,\ \ i = 2, ..., m, \\
\frac{\partial}{\partial a_i} \sum_{k \neq i} t_k &\geq 0,\ i = 2, ..., m.
\end{aligned}$$

Proof: Let $\phi_i(x) = f_i(x) - f_1(x)$. Therefore,

$$\begin{aligned}
g^*(x) &= f_1(x) + \sum_{k=2}^{m} t_k \phi_k(x) \\
and \int \phi_i(x) dx &= 0, \qquad i = 2, ..., m.
\end{aligned}$$

It also follows that

$$f_i(x) = g^*(x) + \phi_i(x) - \sum_{k=2}^{m} t_k \phi_k(x) \geq 0.$$

This implies that

$$-[\phi_i(x) - \sum_{k=2}^{m} t_k \phi_k(x)] \leq g^*(x). \tag{6.10}$$

Since g^* satisfies the constraints (6.1), it follows that, for $2 \leq i \leq m$

$$\begin{aligned}
\frac{\partial a_i}{\partial t_i} &= \frac{\partial}{\partial t_i}[\int f_i(x)\ log\ \frac{f_i(x)}{g^*(x)}\ dx] \\
&= \frac{\partial}{\partial t_i}[\int f_i(x)\ log\ \frac{f_i(x)}{\sum_{k=1}^{m} t_k f_k(x)}\ dx\ ] \\
&= -\int f_i(x) \frac{\phi_i(x)}{g^*(x)}\ dx \\
&= -\int [g^*(x) + \phi_i(x) - \sum_{k=2}^{m} t_k \phi_k(x)] \frac{\phi_i(x)}{g^*(x)}\ dx \\
&= -\int g^*(x)\ \frac{\phi_i(x)}{g^*(x)}\ dx - \int\ [\phi_i(x) - \sum_{k=2}^{m} t_k \phi_k(x)]\ \frac{\phi_i(x)}{g^*(x)}\ dx \\
&\leq -\int \phi_i(x) dx + \int g^*(x)\ \frac{\phi_i(x)}{g^*(x)}\ dx \qquad by\ \ (6.10) \\
&= 0.
\end{aligned}$$

Therefore, it follows that, for $i = 2, ..., m$,

$$\frac{\partial t_i}{\partial a_i} = \frac{1}{\frac{\partial a_i}{\partial t_i}} \leq 0.$$

It also follows that

$$\frac{\partial}{\partial a_i} \sum_{k \neq i} t_k \geq 0$$

since $t_1 + t_2 + ... + t_m = 1$. This completes the proof. $\diamond$

Theorem 6.7 *The weights t_i are all between 0 and 1.*

Proof: Note that if we set $a_i = 0$, then $t_i = 1$; if $a_i = \infty$, then $t_i = 0$. Since t_i is a monotone function of a_i for any fixed $a_j, i \neq j$, it follows that $0 \leq t_i \leq 1$. $\diamond$

The distributions functions $f_1, f_2, ..., f_m$ are, in fact, unknown. We have to derive the optimum function by using samples from different populations. The derivation of

the weighted likelihood function for the discrete case is given in Hu and Zidek (1994). We now generalize their derivation of the weighted likelihood function.

Theorem 6.8 *Assume that the optimal distribution takes the functional form $f(x)$. Given $X_{ij}, i = 1, 2, ..., m, j = 1, 2, ..., n_i$, the optimization problem considered in the this section is equivalent to optimizing the weighted likelihood function.*

Proof: By the proof of Theorem 6.4 and the Lagrange theorem, we need to choose the optimal density function g^* which minimizes

$$\begin{aligned}
&\int f_1(x) log \frac{f_1(x)}{f(x)} dx + l_0 (\int f(x)dx - 1) + \sum_{i=2}^{m} l_i (\int f_i(x) log \frac{f_i(x)}{f(x)} dx - a_i) \\
&= \int f_1(x) log \frac{f_1(x)}{f(x)} dx + l_0 \left(\int f(x)dx - 1 \right) + \\
&\quad \sum_{i=2}^{m} l_i \left(\int f_i(x) log\ f_i(x)dx - a_i - \int f_i(x) log f(x)dx \right) \\
= &- \left(\int f_1(x) log f(x)dx + \sum_{i=2}^{m} l_i f_i(x) log f(x) \right) + \\
&l_0 \left(\int f(x)dx - 1 \right) + \left(\int f_1(x) log f_1(x)dx + \sum_{i=2}^{m} l_i \left(\int f_i(x) log f_i(x)dx - a_i \right) \right).
\end{aligned}$$

The minimization problem considered is then equivalent to maximizing first term in the above equation, *i.e*

$$\begin{aligned}
&\max_f \sum_{i=1}^{m} t_i \int f_i(x) log f(x)dx \\
&= \max_{\theta \in \Theta} \sum_{i=1}^{m} t_i \int log\ f(x;\theta) dF_i(x).
\end{aligned}$$

However distributions $f_1(x), f_2(x), ..., f_m(x)$ are unknown. Their natural estimate in non-parametric context would replace F_i by its empirical counterpart. Assume that the the optimal density function takes the same functional form of f_1. The estimate of the parameter of the optimal distribution would be found as

$$\arg\max_{\theta \in \Theta} \prod_{i=1}^{m} \prod_{j=1}^{n_i} f_1(X_{ij};\theta)^{t_i/n_i}.$$

This implies that the estimate of parameter of the optimal density is equivalent to finding the WLE derived from the weighted likelihood function if the functional form of the optimal density function is known. This completes the proof. $\diamond$

We have shown that the optimal function takes the form

$$g^*(x) = \sum_{k=1}^{m} t_k f_k(x).$$

However the density function g^* does not exist if the constraints define an empty set. Let us consider a relative simple situation where three populations are involved. Recall that the t_i need to satisfy the following condition:

$$t_1 + t_2 + t_3 = 1,$$

$$0 \leq t_i \leq 1, \quad i = 1, 2, 3.$$

The above condition is equivalent to the following:

$$0 \leq t_2 + t_3 \leq 1;$$

$$0 \leq t_2 \leq 1; 0 \leq t_3 \leq 1.$$

If we set $a_2 = a_3 = 0$, then there is no probability distribution satisfying the constraints since a probability density function can not take the functional form of f_2 and f_3 at the same time if $f_2 \neq f_3$. The reason is as follows. In order to satisfy the condition $a_2 = 0$, the weight t_2 must be set to 1. We must also have $t_3 = 1$ for the same reason. Clearly, this set of weights no longer satisfies the condition $t_1 + t_2 + t_3 = 1$.

Lemma 6.3 *The following inequality holds:*

$$I(f_i, \sum_{k=1}^{m} t_k f_k) \leq \sum_{k=1}^{m} t_k I(f_i, f_k), \quad i = 2, 3, ..., m. \tag{6.11}$$

Proof: The function $-log\ g$ is a convex function of g. It then follows that

$$-log\big(\sum_{k=1}^{m} t_k f_k(x)\big) \leq -\sum_{k=1}^{m} t_k log f_k(x).$$

We then have

$$\begin{aligned} I(f_i, \sum_{k=1}^{m} t_k f_k) &= \int log \frac{f_i(x)}{\sum_{k=1}^{m} t_k f_k(x)} f_i(x)dx \\ &\leq \int [log f_i(x) - \sum_{k=1}^{m} t_k log f_k(x)] f_i(x)dx \\ &= \int [\sum_{k=1}^{m} t_k log f_i(x) - \sum_{k=1}^{m} t_k log f_k(x)] f_i(x)dx \\ &= \sum_{k=1}^{m} t_k I(f_i, f_k). \end{aligned}$$

This completes the proof. ◇

Let $D = (d_{ij})$ where

$$d_{ij} = \begin{cases} 1 & if \ \ i = 1 \\ I(f_i, f_j) & if \ \ i = 2, 3, ...m. \end{cases}$$

and $\boldsymbol{a}_{m\times 1} = (1, a_2, ..., a_m)^t$.

Theorem 6.9 *(Existence) The optimal solution does not exist if*

$$rank(D) < rank(B) \tag{6.12}$$

where $B_{m\times(m+1)} = [D_{m\times m} : \boldsymbol{a}]$.

Proof: Note that $I(f_i, \sum_{k=1}^{m} t_k f_k)$ is bounded by $\sum_{k=1}^{m} t_k I(f_i, f_k)$ by Lemma 6.3. Set $a_i = \sum_{k=1}^{m} t_k I(f_i, f_k)$, $i = 2, 3, ..., m$. Note that $\sum_{k=1}^{m} t_k = 1$. We then have the following simultaneous linear equations in t_i:

$$D\boldsymbol{t} = \boldsymbol{a}.$$

By a result from elementary linear algebra, the assertion of the Theorem follows. ◇

Chapter 7

Saddlepoint Approximation of the WLE

7.1 Introduction

In the context of weighted likelihood estimation, the *i.i.d.* assumption is no longer valid. Furthermore, the sample sizes are usually moderate or even very small. The powerful saddlepoint technique is applied to derive the approximate distribution of WLE from exponential family. The saddlepoint approximation for estimating equations proposed in Daniels (1983) is further generalized to derive the approximate density of the WLE derived from an estimating equation.

7.2 Review of the Saddlepoint Approximation

In a pioneering paper, Daniels (1954) introduced a new idea into statistics by applying saddlepoint techniques to derive a very accurate approximation to the distribution of the sample mean for the *i.i.d.* case. It is a general technique which allows us to

compute asymptotic expansions of integrals of the form

$$\int_{\mathcal{P}} e^{vw(z)}\phi(z)dz \tag{7.1}$$

when the real parameter v is large and positive. Here w and ϕ are analytic functions of z in a domain of the complex plane which contains the path of integration $\mathcal{P}$. This technique is called the *method of steepest descent* and is used to derive saddlepoint approximations to density function of a general statistic.

Consider the integral (7.1). In order to find the approximation we deform arbitrarily the path of integration $\mathcal{P}$ provided we remain the domain where w and ϕ are analytic. We deform the path $\mathcal{P}$ such that

(i) the new path of integration passes through a zero of the derivative $w'(z)$;

(ii) the imaginary part of w, $\mathcal{I}w(z)$ is constant on the new path.

If we write

$$\begin{aligned} z &= x+iy, \qquad z_0 = x_0+iy_0, \\ w(z) &= u(x,y)+iv(x,y), \qquad w'(z_0)=0, \end{aligned}$$

and denote by S the surface $(x,y) \to u(x,y)$, then by Cauchy-Riemann differential equations

$$\frac{\partial u}{\partial x} = \frac{\partial v}{\partial y}, \qquad \frac{\partial u}{\partial y} = -\frac{\partial v}{\partial x},$$

it then follows that the point (x_0, y_0) can not be a maximum or minimum but a saddlepoint on the surface S. Moreover, the orthogonal trajectories to the level curves $u(x,y) = \textit{constant}$ are given the the curves $v(x,y) = \textit{constant}$. It follows that the paths on S corresponding to the orthogonal trajectories of the level curves are paths of steepest descent. We will truncate the integration at certain point on the paths of steepest descent. The major part of the integration is then used to approximate

the integration on the complex plane. Detailed discussions can be found in Daniels (1954).

Suppose that $X_1, X_2, ..., X_n$ are *i.i.d.* real-valued random variables with density f and $T_n(X_1, X_2, ..., X_n)$ is real-valued statistic with density f_n. Let $M_n(\alpha) = \int e^{\alpha t} f_n(t)dt$ be the moment-generating function, and $K_n(\alpha) = logM_n(\alpha)$ be the cumulant-generating function. Further suppose that the moment generating function $M_n(\alpha)$ exists for real α in some non-vanishing interval that contains the origin. By Fourier inversion,

$$\begin{aligned} f_n(x) &= \frac{1}{2\pi}\int_{-\infty}^{\infty} M_n(ir)e^{-irx}dr \\ &= \frac{n}{2\pi i}\int_{\mathcal{I}} M_n(nz)e^{-nzx}dz \\ &= \frac{n}{2\pi i}\int_{\tau-i\infty}^{\tau+i\infty} exp\Big(n(R_n(z)-zx)\Big)dz, \end{aligned}$$

where $\mathcal{I}$ is the imaginary axis in the complex plane, τ is any real number in the interval where the moment generating exists, and

$$R_n(z) = K_n(nz)/n.$$

Applying the saddlepoint approximation to the last integral gives the approximation of f_n with uniform error of order n^{-1}:

$$g_n(x) = \left(\frac{n}{2\pi\tilde{R}_n''(\alpha_0)}\right)^{1/2} exp\big(n[R_n(\alpha_0)-\alpha_0 x]\big), \tag{7.2}$$

where α_0 is the saddlepoint determined by the equation

$$R_n'(\alpha_0) = t,$$

where R_n' and R_n'' denote the first two derivatives of R_n. Detail discussions of the saddlepoint can be found in Daniels (1954) and Field and Ronchetti (1990).

n	$n!$	Stirling	Saddlepoint	r.e. of Stirling	r.e. of saddlepoint
1	1	0.92	0.99	0.07786	0.00102
2	2	1.92	1.99	0.04050	0.00052
3	6	5.83	5.99	0.02730	0.00028
4	24	23.50	24.00	0.02058	0.00017
5	120	118.02	119.99	0.01651	0.00016
6	720	710.08	719.94	0.01378	0.00008
7	5040	4980.40	5039.69	0.01183	0.00006
8	40320	39902.87	40318.05	0.01036	0.00005
9	362880	359537.62	362865.91	0.00921	0.00004

Table 7.1: Saddlepoint Approximations of $\Gamma(n+1) = n!$

It is known that the *Stirling* formula serves as a very good approximation to the Gamma function. The comparison of accuracies between the *Stirling* formula and the saddlepoint approximation based on the the first two terms is given in the above table with the last two columns for the relative error for using Stirling formula and saddlepoint approximation respectively. The saddlepoint approximation is given by $\sqrt{2\pi}n^{n+1/2}e^{-n}\left(1+\frac{1}{12n}\right)$. Notice that the expression before $\left(1+\frac{1}{12n}\right)$ is exactly the *Stirling Formula.*

7.3 Results for Exponential Family

The saddlepoint approximation stated in the last section is for a general statistic constructed by a series of *i.i.d.* random variables. In this section, we derive the saddlepoint approximation to the distribution of a sum of a finite number of random variables that are independent but not identically distributed.

Suppose that we consider the distribution of the following statistic:

$$S(\mathbf{X}) = \sum_{i=1}^{m} T_i(X_{i1}, X_{i2}, ..., X_{in_i}),$$

where X_{ij} are $i.i.d$ for any given i. But (X_{ij}) and $(X_{i'j})$ with $i \neq i'$ do not follow the same distribution in general.

Theorem 7.1 *The saddlepoint approximation to the density function of the random variable defined by the convolution is given by:*

$$\tilde{f}_S(x) = \left(\frac{n_1}{2\pi \sum_{i=1}^{m} R_i''(n_1\alpha_0^*)/n_1} \right)^{1/2} exp\left(n_1\Big(\sum_{i=1}^{m} R_i(n_1\alpha_0^*)/n_1 - \alpha_0^* x \Big) \right) \qquad (7.3)$$

where α_0^* satisfying the following equation

$$\sum_{i=1}^{m} R_i'(\alpha_0^*) = x.$$

Proof: The moment generating function of $S(\mathbf{X})$ is $M_1 \times M_2 ... \times M_m$, where M_m is the moment generating function of $T_i(X_{i1}, X_{i2}, ..., X_{in_i})$. By Fourier inversion,

$$\begin{aligned} f_{S_{(n1,n2,..,n_m)}}(x) &= \frac{1}{2\pi} \int_{-\infty}^{\infty} M_1(ir)\ M_2(ir)...M_m(ir)e^{-irx}dr \\ &= \frac{n_1}{2\pi i} \int_{\mathcal{I}} M_1(n_1 z)\ M_2(n_1 z)...M_m(n_1 z)e^{-n_1 zx}dz \\ &= \frac{n_1}{2\pi i} \int_{\tau-i\infty}^{\tau+i\infty} exp\left(n_1\Big(\sum_{i=1}^{m} R_i(n_1 z)/n_1 - zx \Big) \right) dz, \end{aligned}$$

where R_i is the cumulant generating funtion of T_i.

Applying the saddlepoint technique we derive the approximate density for $S(\mathbf{X})$:

$$\tilde{f}_{S_{(n_1,n_2,...,n_m)}}(x) = \left(\frac{n}{2\pi \sum_{i=1}^{m} R_i''(n_1\alpha_0^*)/n_1} \right)^{1/2} exp\left(n_1\Big(\sum_{i=1}^{m} R_i(n_1\alpha_0^*)/n_1 - \alpha_0^* x \Big) \right)$$

where α_0^* is the root of the equation $\sum_{i=1}^{m} R_i'(n_1\alpha_0^*)/n_1 = x$. $\diamond$

Example 7.1 (*The Sum of Sample Means*) Let us examine the distribution of W_n where

$$W_n = \frac{1}{n_1}(X_{11} + X_{12} + ... + X_{1n_1}) + \frac{1}{n_2}(X_{21} + X_{22} + ... + X_{2n_2}).$$

The moment generating function of W_n is $M_1(\frac{t}{n_1})^{n_1} \times M_2(\frac{t}{n_2})^{n_2}$ where M_1 and M_2 are the moment generating functions of X_{11} and X_{21} respectively. Let K_1 and K_2 be the cumulant generating function of X_{11} and X_{21} respectively. It then follows

$$R_i(n_1 z)/n_1 = n_i K_i\left(\frac{z n_1}{n_i}\right)/n_1 = \frac{n_i}{n_1} K_i\left(\frac{n_1}{n_i} z\right), \quad i = 1, 2.$$

The saddlepoint in this case is then a root of the equation

$$\frac{\partial}{\partial z} K_1(z) + \frac{n_2}{n_1}\frac{\partial}{\partial z} K_2\left(\frac{n_1}{n_2} z\right) = x.$$

The saddlepoint approximate density of W_n is

$$\begin{aligned} \tilde{f}_{W_n}(x) &= \left(\frac{n_1}{2\pi\left(\frac{\partial^2}{\partial z^2} K_1(\alpha_0^*) + \frac{n_2}{n_1}\frac{\partial^2}{\partial z^2} K_2(\alpha_0^*)\right)}\right)^{1/2} \\ &\quad exp\left(n_1\left(K_1(\alpha_0^*) + \frac{n_2}{n_1} K_2(\frac{n_1}{n_2}\alpha_0^*) - \alpha_0^* t\right)\right). \end{aligned}$$

Assume that $n_1 = n_2 = n$ and $K^{(1)} = K^{(2)} = K$. We then have

$$\hat{f}_{W_n}(x) = \left(\frac{n}{2\pi 2\frac{\partial^2}{\partial z^2} K(\alpha_0^*)}\right)^{1/2} exp\left(n\big(2K(\alpha_0^*) - \alpha_0^* x\big)\right).$$

The sample average of the combined sample is $W_n/2$. It then follows that

$$\begin{aligned} \tilde{f}_{W_n/2}(y) &= \left(\frac{n_1}{2\pi 2\frac{\partial^2}{\partial z^2} K(\alpha_0^*)}\right)^{1/2} exp\left(n\big(2K(\alpha_0^*) - \alpha_0^* 2y\big)\right) 2 \\ &= \left(\frac{2n}{2\pi\frac{\partial^2}{\partial z^2} K(\alpha_0^*)}\right)^{1/2} exp\left(2n\big(K(\alpha_0^*) - \alpha_0^* y\big)\right) \end{aligned}$$

where α_0^* is the root of the equation

$$2\frac{\partial}{\partial z} K(z) = x = 2y.$$

It then implies that α_0^* indeed satisfies the following equation

$$\frac{\partial}{\partial z}K(z) = y.$$

This is exactly the saddlepoint α_0 for an *i.i.d.* sample with size $2n$. Thus, the saddlepoint approximation of the density of $W_n/2$ by Theorem 7.1 when the random variables from both samples are indeed *i.i.d.* is exactly the saddlepoint approximation of the sample mean of a *i.i.d.* sample with size $2n$.

Example 7.2 (*Spread Density for the Exponential*) If $Y_1, Y_2, ..., Y_{m+1}$ are independent, exponentially distributed random variables with common mean 1, then the spread, $Y_{(m+1)} - Y_{(1)}$, the difference between maxima and minima of the sample, has the density

$$m\sum_{k=1}^{m}(-1)^{(k-1)}\binom{m-1}{k-1}e^{-kx}.$$

This is also the density of the sum $X_1 + X_2 + ... + X_m$ of independent, exponentially distributed random variables Y_j with respective means $1, 1/2, ..., 1/m$. A proof of this claim is sketched in Problem I.13.13 in Feller (1971).

It follows that the cumulant generating function of the sum $S(\underline{\mathbf{X}}) = X_1 + X_2 ... + X_m$ is $\sum_{i=1}^{m} R_i(z) = -\sum_{j=1}^{m} ln(1 - z/j)$. The equation $\sum_{i=1}^{m} R_i'(z) = x$ in Theorem 7.1 becomes $\sum_{j=1}^{m} 1/(j - z) = x$ which needs to be solved numerically. Due to the unequal variances of Y_j, the normal approximation does not work well. Lange (1999) calculates the saddlepoint approximation for this particular example. Note that the following table is part of Table 17.2 in Lange (1999). The last column is the difference between the exact density and the saddlepoint approximation.

It can be seen that saddlepoint approximation gives a very accurate approximation. Lange (1999) also shows that the saddlepoint approximation out-performs the Edgeworth expansion in this example.

x	Exact	Error
0.5	.00137	−.00001
1.0	.05928	−.00027
1.5	.22998	.00008
2.0	.36563	.00244
2.5	.37874	.00496
3.0	.31442	.00550
3.5	.22915	.00423
4.0	.15508	.00242
4.5	.10046	.00098
5.0	.06340	.00012
5.5	.03939	−.00026
6.0	.02424	−.00037
6.5	.01483	−.00035
7.0	.00904	−.00028
7.5	.00550	−.00021
8.0	.00334	−.00014

Table 7.2: Saddlepoint Approximation of the Spread Density for $m = 10$.

We now consider the saddlepoint approximation to the distribution of the WLE in the general exponential family. It has been shown that the WLE takes the form $g\left(\sum_{i=1}^{m} \lambda_i T_i(X_{i1}, ..., X_{in_i})\right)$ for some smooth function g under fairly general conditions.

Theorem 7.2 *Assume that g is a smooth function. The saddlepoint approximation*

to the density of $g\left(\sum_{i=1}^{m}\lambda_i T_i(X_{i1},....,X_{in_i})\right)$ *is*

$$\tilde{f}_{WLE}(y) = \left(\frac{n_1}{2\pi\sum_{i=1}^{m}R_i''(\lambda_i n_1\alpha_0^w)/n_1}\right)^{1/2} \times exp\left(n_1\left(\sum_{i=1}^{m}R_i(\lambda_i n_1\alpha_0^w)/n_1 - \alpha_0^* g^{-1}(y)\right)\right)\left|\frac{1}{g'(g^{-1}(y))}\right|$$

where α_0^w *satisfies the following equation*

$$\sum_{i=1}^{m}R_i'(\lambda_i n_1 z)(\alpha_0^w)/n_1 = g^{-1}(y).$$

Proof: We first derive the approximate density function for $S=\sum_{i=1}^{m}\lambda_i T_i(X_{i1},....,X_{in_i})$. The cumulant generating function of $\lambda_i T_i$ is $R_i(\lambda_i z)$ where $R_i(z)$ is the cumulant generating function of T_i. By Theorem 7.1, the approximate density function of $S=\sum_{i=1}^{m}\lambda_i T_i$ is given by

$$\tilde{f}_S(x) = \left(\frac{n_1}{2\pi\sum_{i=1}^{m}R_i''(\lambda_i n_1\alpha_0^w)/n_1}\right)^{1/2} exp\left(n_1\left(\sum_{i=1}^{m}R_i(\lambda_i n_1\alpha_0^w)/n_1 - \alpha_0^w x\right)\right)$$

where α_0^* is the root of the equation

$$\sum_{i=1}^{m}R_i'(\lambda_i n_1\alpha_0^w)/n_1 = x.$$

It then follows that the approximate density function of the WLE, $g\left(\sum_{i=1}^{m}\lambda_i T_i(X_{i1},....,X_{in_i})\right)$ is given by

$$\tilde{f}_{WLE}(y) = \left(\frac{n_1}{2\pi\sum_{i=1}^{m}R_i''(\lambda_i n_1\alpha_0^w)/n_1}\right)^{1/2} \times exp\left(n_1\left(\sum_{i=1}^{m}R_i(\lambda_i n_1\alpha_0^w)/n_1 - \alpha_0^* g^{-1}(y)\right)\right)\left|\frac{1}{g'(g^{-1}(y))}\right|$$

where α_0^w satisfies the following equation

$$\sum_{i=1}^{m} R_i'(\lambda_i n_1 z)(\alpha_0^w)/n_1 = g^{-1}(y).$$

This completes the proof.

7.4 Approximation for General WL Estimation

The saddlepoint approximation to estimating equations for the *i.i.d.* case is developed by Daniels (1983). We generalize the techniques to the WLE derived from the estimation equation constructed by the Weighted Likelihood Function. Recall that the WLE is the root of the following estimating equation:

$$\sum_{i=1}^{m} \lambda_i \sum_{j=1}^{n_i} \frac{\partial}{\partial \theta_1} log f(X_{ij}; \tilde{\theta}_1) = 0. \tag{7.4}$$

For simplicity, let $\psi(X_{ij}; \theta_1) = \frac{\partial}{\partial \theta_1} log f(X_{ij}; \theta_1)$. The estimating equation for WLE can be written as

$$\sum_{i=1}^{m} \lambda_i \sum_{j=1}^{n_i} \psi(X_{ij}; \tilde{\theta}_1) = 0. \tag{7.5}$$

Assume that $\psi(X_{ij}; \theta)$ is a monotone decreasing function of θ and $\lambda_i \geq 0, i = 1, 2, ..., m$. Write

$$S(a) = \sum_{i=1}^{m} \lambda_i \sum_{j=1}^{n_i} \psi(X_{ij}; a),$$

where a is a fixed number.

Let α_0^S be the root of the equation

$$\sum_{i=1}^{m} n_i \lambda_i \frac{\partial}{\partial z} K_i(n_1 \lambda_i z, a) = x.$$

We have the following theorem:

Theorem 7.3 *Let $\tilde{\theta}_1$ be the WLE derived from the weighted likelihood equation. Then*

$$P_{\theta_1}(\tilde{\theta}_1 > a) = P(S > 0) \sim \int_0^{\infty} \tilde{f}_S(x, a) dx \tag{7.6}$$

and

$$\tilde{f}_S(x) = \left(\frac{n_1}{2\pi \sum_{i=1}^{m} n_i n_1 \lambda_i^2 \frac{\partial^2}{\partial z^2} K_i(n_1\lambda_i z, a)|_{z=\alpha_0^S}} \right)^{1/2} \exp\left(n_1 \Big(\sum_{i=1}^{m} n_i \lambda_i K_i(n_1\lambda_i z, a)|_{z=\alpha_0^S} - x\alpha_0^S \Big) \right) \tag{7.7}$$

and $exp\left(K_i(t,a)\right) = E_{\theta_i} exp\left(t\psi(X_{ij},a)\right), i = 1,2,...,m.$

Proof: The moment generating function of $S(a)$ is

$$\begin{aligned} M_S(t,a) &= exp\left(K_S(t,a)\right) \\ &= \prod_{i=1}^{m} \left(\int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty} exp\left(t\lambda_i \sum_{j=1}^{n_i} \psi(X_{ij};a) \right) \prod_{j=1}^{n_i} f(x_{ij};\theta_i) dx_{i1}...dx_{in_i} \right) \\ &= \prod_{i=1}^{m} exp\left(K_i(\lambda_i t,a)\right)^{n_i} \\ &= exp\left(\sum_{i=1}^{m} n_i K_i(\lambda_i t,a) \right) \end{aligned}$$

where $exp\left(K_i(t,a)\right) = E_{\theta_i} exp\left(t\psi(X_{ij},a)\right), i = 1,2,...,m.$ The function $M_S(t,a)$ is assumed to converge for real t in a non-vanishing interval containing the origin.

It then follows that

$$\begin{aligned} f_S(x) &= \frac{1}{2\pi i} \int_{-\infty}^{\infty} exp\left(\sum_{i=1}^{m} n_i K_i(ir\lambda_i,a) \right) exp\left(-irx\right) dr \\ &= \frac{n_1}{2\pi i} \int_{\tau-i\infty}^{\tau+i\infty} exp\left(\sum_{i=1}^{m} n_i K_i(n_1\lambda_i z,a) - n_1 xz \right) dz \\ &= \frac{n_1}{2\pi i} \int_{\tau-i\infty}^{\tau+i\infty} exp\left(n_1 \Big(\sum_{i=1}^{m} \frac{n_i}{n_1} K_i(n_1\lambda_i z,a) - xz \Big) \right) dz \end{aligned}$$

The saddlepoint α_0^S is the root of the equation

$$\sum_{i=1}^{m} n_i \lambda_i \frac{\partial}{\partial z} K_i(n_1\lambda_i z,a) = x.$$

It then follows that the saddlepoint approximation to the density $f_S(x)$ is given by

$$\tilde{f}_S(x) = \left(\frac{n_1}{2\pi \Big(\sum_{i=1}^{m} n_i n_1 \lambda_i^2 \frac{\partial^2}{\partial z^2} K_i(n_1 \lambda_i z, a)|_{z=\alpha_0^S} \Big)} \right)^{1/2} exp\left(n_1 \Big(\sum_{i=1}^{m} n_i \lambda_i K_i(n_1 \lambda_i z, a)|_{z=\alpha_0^S} - x\alpha_0^S \Big) \right).$$

We can then deploy the device used in Field and Hampel (1982) and Daniels (1983)

$$P_{\theta_1}(\tilde{\theta}_1 > a) = P_{\theta_1}(S(a) > 0). \tag{7.8}$$

We then have

$$P_{\theta_1}(\tilde{\theta}_1 > a) = P(S > 0) \sim \int_0^\infty \tilde{f}_S(x, a) dx. \quad \diamond$$

In general, the saddlepoint approximation is very computational intensive since the normalizing constant needs to be calculated by numerical integration. We remark that the saddlepoint approximations to the WLE proposed in this chapter are for fixed weights. The saddlepoint approximation to the WLE with adaptive weights needs further study.

Chapter 8

Application to Disease Mapping

8.1 Introduction

In this chapter, we present the results of the application of the *maximum weighted likelihood estimation* to parallel time series of hospital-based health data. Specifically, the *weighted likelihood method* is illustrated on daily hospital admissions of respiratory disease obtained from 733 census sub-division (CSD) in Southern Ontario over the May-to-August period from 1983 to 1988. Our main interest is on the estimation of the rate of weekly hospital admissions of certain densely populated areas. The association between air pollutants and respiratory morbidity are studied in Zidek *et al.* (1998) and Burnett (1994).

For the purpose of our demonstration, we will consider the estimation of the rate of weekly hospital admissions of CSD # 380 which has the largest yearly hospital admissions total among all CSD's from 1983 to 1988. The estimation of the rate of weekly admissions is a challenging task due to the sparseness of the data set. The original data set contains many 0's which represent no hospital admissions on most of the days in the summer. On certain days of the summer, however, quite a number of

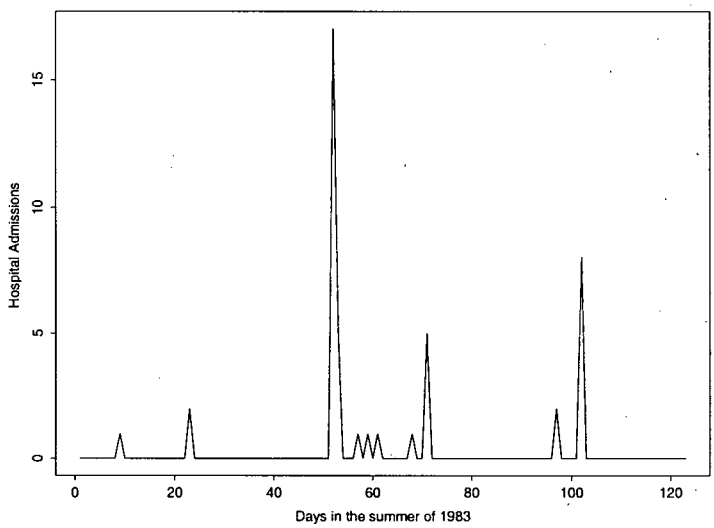

Figure 8.1: Daily hospital admissions for CSD # 380 in the summer of 1983.

people with respiratory disease went to hospital to seek treatments due perhaps to the high level of pollution in the region. For CSD # 380 that has the largest number of hospital admissions among all the CSD's, there are no hospital admission for a total of 112 days out of 123 days in the summer of 1983. However there were 17 hospital admissions on day 51. The daily counts of this CSD are shown in Figure 8.1. The problem of data sparseness and high level of variation is quite obvious. Thus we will consider the estimation of the rate of weekly admissions instead of the daily counts. There are 17 weeks in total. For simplicity, the data obtained in the last few days of each year are dropped from the analysis since they do not constitute a whole week.

8.2 Weighted Likelihood Estimation

We assume that the total number of hospital admissions of a week for a particular CSD follows Poisson distribution, *i.e.*, for year q, CSD i and week j,

$$Y_{ij}^{q} \overset{ind.}{\sim} \mathcal{P}\left(\theta_{ij}^{q}\right), j = 1, 2, ..., 17; i = 1, 2, ..., 733; q = 1, 2, ..., 6.$$

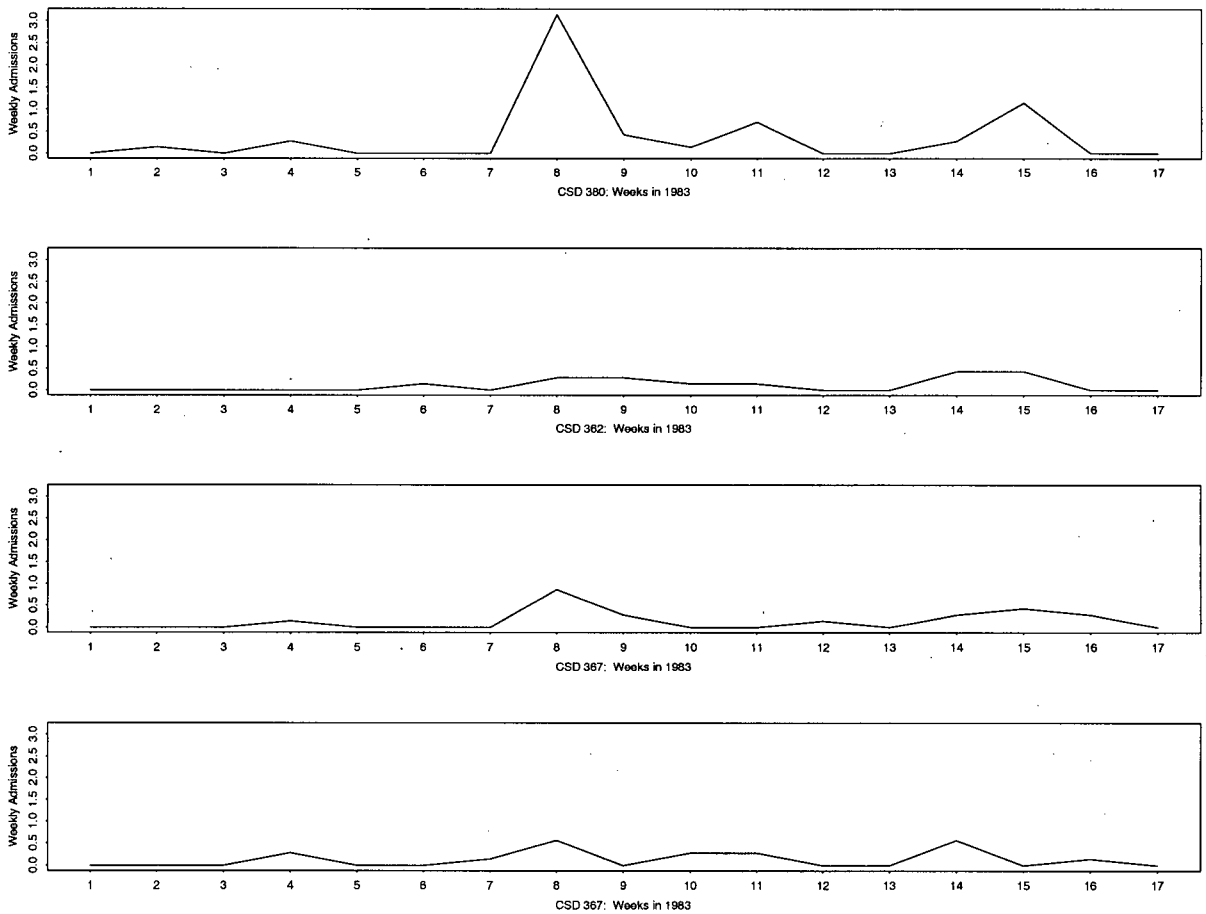

Figure 8.2: Hospital Admissions for CSD # 380, 362, 366 and 367 in 1983.

The raw estimate of θ_{ij}^q which is Y_{ij}^q is highly unreliable. The sample size is only 1 in this case. Also each CSD may contain only a small group of people whose lung conditions are susceptible to the high levels of air pollution. Therefore we think that it is desirable to combine neighboring CSD's in order to produce a more reliable estimate. For any given CSD, the neighboring CSD's are defined to be CSD's that are in close proximity to the CSD of interest, CSD #380 in our analysis. To study the rate of weekly hospital admissions in a particular CSD, we would expect that neighboring subdivisions contain relevant information which might help us to derive

a better estimate than the traditional sample average. The importance of combining disease and exposure data is discussed in Waller *et al.* (1997). The Euclidean distance between the target CSD and any other CSD in the dataset is calculated by using the longitudes and latitudes. CSD's whose Euclidean distances are less than 0.2 from the target CSD are selected as the relevant CSD. For CSD # 380, neighboring CSD's are CSD # 362, 366 and 367. The time series plots of weekly hospital admissions for those CSD's in 1983 are shown in Figure 8.2. It seems that the hospital admissions of these CSD's at a given week might be related since the major peaks in the time series plot occurred at roughly the same time point. However, including data from other CSD's might introduce bias. The weight function defined in the WLE can control the degree of bias introduced by the combination of data from other CSD's.

Ideally, we would assume that $\theta_{ij}^q = \theta_i^q$ for $j = 1, 2, ..., 17$. But this assumption does not hold due to seasonality. For example, week 8 always has the largest hospital admissions for CSD 380. By examining the data more closely, we realize that the $8th$ week for each year is a week with more than normal hospital admissions. In 1983, there are 21 admissions in week 8 while the second largest weekly count is only 7 in week 15. In fact, week 8 is an unusual week through 1983 to 1988. The air pollution level might further explain this phenomenon by using the method proposed in Zidek *et al.* (1998). Thus, the assumption is violated at week 8. One alternative is to perform the analysis on a window of a few weeks and repeat the analysis while we move the window one week forward. This is equivalent of assuming that θ_{ij}^q take the same value for a period of a few weeks instead of the entire summer. In this chapter, we will simply exclude the observations from week 8 and proceed with the assumption that $\theta_{ij}^q = \theta_i^q$ for $j = 1, 2, ...7, 9, ..17$. The fact that the sample means and sample variances of the weekly hospital admissions for those 16 weeks of CSD 380 are quite close to each other supports our assumption.

Let $\overline{Y}_{i.}^{q}$ to be the overall sample average for a particular CSD i for a given year. For Poisson distributions, the MLE of θ_1^q is the sample average of the weekly admissions of CSD #380 and the WLE is a linear combination of the sample average for each CSD according to Theorem 3.1. Thus, the *weighted likelihood estimate* of the average weekly hospital admissions for a CSD, θ_1^q, is

$$WLE^q = \sum_{i=1}^{4} \lambda_i^q \overline{Y}_{i.}^{q}, \quad q = 1, 2.., 6.$$

For our analysis, the weights are selected by the cross-validation procedure proposed in Chapter 5. Recall that the cross-validated weights for equal sample sizes can be calculated as follows:

$$\boldsymbol{\lambda}^q = A_q^{-1} \left(b_q + \frac{1 - \mathbf{1}^t A_q^{-1} b}{\mathbf{1}^t A_q^{-1} \mathbf{1}} A_q^{-1} \mathbf{1} \right)$$

where $b_q(\underline{y}) = \sum_{j=1}^{17} Y_{1j}^q \overline{Y}_{i.}^{q(-j)}$, and $A_q(\underline{y})_{ik} = \sum_{j=1}^{17} \overline{Y}_{i.}^{q(-j)} \overline{Y}_{k.}^{q(-j)}$, $i = 1, 2, 3, 4; k = 1, 2, 3, 4$.

8.3 Results of the Analysis

We assess the performance of the MLE and the WLE by comparing their MSE's. The MSE of the MLE and the WLE are defined as, for $q = 1, 2, .., 6$,

$$\begin{aligned} MSE_M^q(\theta_1^q) &= E_{\theta_1^q} \left(\overline{Y}_{1.}^{q} - \theta_1^q \right)^2 \\ MSE_W^q(\theta_1^q) &= E_{\theta_1^q} \left(\sum_{i=1}^{4} \lambda_i^q \overline{Y}_{i.}^{q} - \theta_1^q \right)^2 . \end{aligned}$$

In fact, the θ_1^q are unknown. We then estimate the MSE_M and MSE_W by replacing θ_1^q by the MLE. Under the assumption of Poisson distributions, the estimated MSE for the MLE is given by:

$$\widehat{MSE_M^q} = \widehat{var(Y_{11})}/16, \quad q = 1, 2, ...6.$$

The estimated MSE for the WLE is give as following:

$$\begin{aligned} MSE_W^q &= E\left(\sum_{i=1}^{m}\lambda_i^q\overline{Y}_{i.}^q - \theta_1^q\right)^2 \\ &= Var\left(\sum_{i=1}^{m}\lambda_i^q\overline{Y}_{i.}^q\right) + \left(E\sum_{i=1}^{m}\lambda_i^q\overline{Y}_{i.}^q - \theta_1^q\right)^2 \\ &\approx \sum_{i=1}^{4}\sum_{k=1}^{4}\lambda_i^q\lambda_j^q\widehat{cov}(\overline{Y}_{i.}^q, \overline{Y}_{j.}^q) + \left(\sum_{i=1}^{m}\lambda_i^q\overline{Y}_{i.}^q - \overline{Y}_{1.}^q\right)^2. \end{aligned}$$

The estimated MSE for the MLE and the WLE are given in the following table. It can be seen from the table that the MSE for the WLE is much smaller than that of the MLE. In fact, the average reduction of the MSE by using WLE is about 25%.

Year	7 MLE	7 WLE	16 $\widehat{MSE}_M^q$	16 $\widehat{MSE}_W^q$	$\widehat{MSE}_W^q/\widehat{MSE}_M^q$
1	.185	.174	.101	.084	0.80
2	.328	.282	.241	.131	0.87
3	.227	.257	.286	.143	0.54
4	.151	.224	.159	.084	0.96
5	.303	.322	.298	.130	0.80
6	.378	.412	.410	.244	0.54

Table 8.1: Estimated MSE for the MLE and the WLE.

Combining information across these CSD's might also help us in the prediction since the patterns exhibited in one neighboring location in a particular year might manifest itself at the location of interest the next year. To assess the performance of the WLE, we also use the WLE derived from one particular year to predict the overall weekly average of the next year. The overall prediction error is defined as the average of those prediction errors. To be more specific, the overall prediction errors

for the WLE and the traditional sample average are defined as follows:

$$PRED_M = \sqrt{\frac{1}{5}\sum_{q=1}^{5}\left(\overline{Y}_{1.}^{q} - \overline{Y}_{1.}^{q+1}\right)^2};$$

$$PRED_W = \sqrt{\frac{1}{5}\sum_{q=1}^{5}\left(WLE^q - \overline{Y}_{1.}^{q+1}\right)^2}.$$

The average prediction error for the MLE, $Pred_M$, is 0.065 while the $Pred_W$, the average prediction error for the WLE, is 0.047 which is about 72% of that of the MLE.

8.4 Discussion

Bayes methods are popular choices in the area of disease mapping. Manton *et al.* (1989) discuss the Empirical Bayes procedures for stabilizing maps of cancer mortality rates. Hierarchical Bayes generalized linear models are proposed for the analysis of disease mapping in Ghosh *et al.* (1999). But it is not obvious how one would specify a neighborhood which needs to be defined in these approaches. The numerical values of the weight functions can be used as a criterion to appropriately define the neighborhood in the Bayesian analysis. We will use the following example to demonstrate how a neighborhood can be defined by using the weight functions derived from the cross-validation procedure for the WLE.

From Table 8.3, we see that there is strong linear association between CSD 380 and CSD 366. However, the weight assigned to CSD 366 is the smallest one. It shows that CSD's with higher correlation contain less information for the prediction since they might have too similar a pattern to the target CSD for a given year to be helpful in the prediction for the next year. Thus CSD 366 which has the smallest weight should not be included the analysis. Therefore, the "neighborhood" of CSD 380 in

	CSD 380	CSD 362	CSD 366	CSD 367	Weights
CSD 380	1.000	0.421	0.906	0.572	0.455
CSD 362	0.421	1.000	0.400	0.634	0.202
CSD 366	0.906	0.400	1.000	0.553	0.128
CSD 367	0.572	0.634	0.553	1.000	0.215

Table 8.2: Correlation matrix and the weight function for 1984.

the analysis should only include CSD 362 and CSD 367.

In general, we might examine those CSD which are in close proximity to the target CSD. We can calculate the weight for each CSD selected by using the cross-validation procedure. The CSD with small weights should be dropped from the analysis since they are not deemed to be helpful or relevant to our analysis according to the cross-validation procedure.

We remark that the weight function can also be helpful in selecting an appropriate distribution that takes into account the spatial structure. Ghosh *et al.* (1999) propose a very general hierarchical Bayes spatial generalized model that is considered broad enough to cover a large number of situations where a spatial structure needs to be incorporated. In particular, they propose the following:

$$\theta_i = q_i = x_i^t \mathbf{b} + u_i + v_i, i = 1, 2, ..., m$$

where the q_i are known constants, x_i are covariates, u_i and v_i are mutually independent with $v_i \overset{i.i.d.}{\sim} N(0, \sigma_v^2)$ and the u_i have joint pdf

$$f(u) \propto (\sigma_u)^2)^{-1/2m} exp\left(-\sum_{i=1}^{m}\sum_{j\neq i}(u_i - u_j)^2 w_{ij}/(2\sigma_u^2)\right)$$

where $w_{ij} \geq 0$ for all $1 \leq i \neq j \leq m$. The above distribution is designed to take into account the spatial structure. In their paper, they propose to use $w_i = 1$ if location i and j are neighbors. They also mention the possibility of using the inverse

of the correlation matrix as the weight function. The weights function derived from the cross-validation procedure might be a better choice since it takes account of the spatial structure without any specific model assumptions.

The predictive distribution for the weekly total will be Poisson (WLE). We can then derive the 95% predictive interval for the weekly average hospital admissions. This might be criticized as failing to take into account the the uncertainty of the unknown parameter. Smith (1998) argues that the traditional plug-in method has a small MSE compared to the posterior mean under certain circumstances. In particular, it has a smaller MSE when the true value of the parameter is not large. Let CI_W and CI_M be the 95% predictive intervals of the weekly averages calculated from the WLE and the MLE respectively. The results are shown in the following table.

Year	CI_M	CI_W
1983	[0,3]	[0, 3]
1984	[0, 5]	[0, 4]
1985	[0, 4]	[0, 4]
1986	[0, 3]	[0, 4]
1987	[0, 4]	[0, 5]
1988	[0, 5]	[0, 6]

Table 8.3: MSE of the MLE and the WLE for CSD 380.

We remark that this chapter is merely a demonstration of the weighted likelihood method. Further analysis is needed if one wants to compare the performances of the WLE, the MLE and the Bayesian estimator in disease mapping.

Bibliography

[1] Akaike, H. (1985). Prediction and entropy, In: *A Celebration of Statistics* 1-24, *Edited by* Atkinson, A. C. and Fienberg, S. E., Springer-Verlag, New York.

[2] Bliss, G. A. (1930). The problem of lagrange in the calculus of variations. *The American Journal of Mathematics* **52** 673-744.

[3] Breiman, L. and Friedman, H. J. (1997). Predicting multivariate responses in multiple regression, *Journal of Royal Statistical Society: Series B* **36** 111-147.

[4] Burnett, R. and Krewski, D. (1994). Air pollution effects on hospital admission rates: A random effects modeling approach. *The Canadian Journal of Statistics* **22** 441-458.

[5] Cox. D. R. (1981). Combination of data. *Encyclopedia of Statistical Sciences* **2** 45-52, John Wiley & Sons, Inc., New York.

[6] Csiszar, I. (1975) I-divergence geometry of probability distributions and minimization problems. *The Annals of Probability* **3** 146-158.

[7] Daniels, H. E. (1954). Saddlepoint approximation in statistics. *The Annals of Mathematical Statistics* **25** 59-63.

[8] Daniels, H. E. (1983). Saddlepoint approximation for estimating equations. *Biometrika* **70** 89-96.

[9] Dickey, J. M. (1971) The weighted likelihood ratio, linear hypotheses on normal location parameters. *The Annals of Mathematical Statistics* **42** 204–223.

[10] Dickey, J. M and Lientz, B. P. (1970) The weighted likelihood ratio, sharp hypotheses about chances, the order of a Markov chain. *The Annals of Mathematical Statistics* **41** 214–226.

[11] Easton, G. (1991). Compromised maximum likelihood estimators for location. *Journal of the American Statistical Association* **86** 1051-1064.

[12] Eguchi, S. and Copas, J. (1998). A class of local likelihood methods and near-parametric asymptotics. *Journal of Royal Statistical Society*, Series B **60** 709-724.

[13] Ekeland, I. and Temam, R. (1976). *Convex Analysis and Variational Problems.* American Elsevier Publishing Company Inc., New York.

[14] Feller, W. (1971) *An Introduction to Probability Theory and Its Applications, Vol 2.* John Wiley & Sons, Inc., New York.

[15] Ferguson, T. S. (1996). *A Course in Large Sample Theory.* Chapman and Hall, New York.

[16] Field, C. A. and Hampel, F. R. (1982). Small Sample Asymptotics Distributions of M-estimators of Location. *Biometrika* **69** 29-46.

[17] Field, C. A. and Ronchetii, E. (1990). *Small Sample Asymptotics.* Institute of Mathematical Statistics, Hayward.

[18] Geisser, S. (1975). The predictive sample reuse method with applications, *Journal of the American Statistical Association* **70** 320-328.

[19] Genest, C. and Zidek, J. V. (1986). Combining probability distributions: a critique and an annotated bibliography. *Statistical Science* **1** 114-148.

[20] Ghosh, M., Natarajan, K., Waller, L. A. and Kim, D. (1999). Hierarchical Bayes GLMs for the analysis of spatial data: An application to disease mapping. *Journal of Statistical Planning and Inference* **75** 305-318.

[21] Giaquinta, M. and Hildebrandt, S. (1996). *Calculus of Variations.* Springer-Verlag Series, New York.

[22] Hardle, W. and Gasser, T. (1984) Robust Nonparametric Function Fitting. *Journal of the Royal Statistical Society, Series B,* **46** 42-51.

[23] Hu, F. (1994). *Relevance Weighted Smoothing and A New Bootstrap Method,* Ph.D. Dissertation, Department of Statistics, University of British Columbia, Canada.

[24] Hu, F. (1997). The asymptotic properties of the maximum-relevance weighted likelihood estimators. *The Canadian Journal of Statistics* **25** 45-59.

[25] Hu, F., Rosenberger, W. F. and Zidek, J. V. (2000). Relevance weighted likelihood for dependent data. *Metrika* **51** 223-243.

[26] Hu, F. and Zidek, J. V. (1997). The relevance weighted likelihood with applications. In: *Empirical Bayes and Likelihood Inference* 211-235, *Edited by* Ahmed, S. E. and Reid, N., Springer, New York.

[27] Hunsberger, S. (1994) Semiparametric regression in likelihood-based models. *Journal of the American Statistical Association* **89** 1354-1365.

[28] Kullback, S. (1954). Certain inequality in information theory and the Cramer-Rao inequality. *The Annals of Mathematical Statistics* **25** 745-751.

[29] Kullback, S. (1959). *Information Theory and Statistics.* Lecture Notes-Monograph Series Volume 21, Institute of Mathematical Statistics.

[30] Lange, K. (1999) *Numerical Analysis for Statisticians.* Springer-Verlag, New York.

[31] Lehmann, E. L. (1983), *Theory of Point Estimation.* John Wiley & Sons Inc., New York.

[32] Markatou, M., Basu, A. and Lindsay, B. (1997). Weighted likelihood estimating equations: The discrete case with applications to logistic regression. *Journal of Statistical Planning and Inference* **92** 215-232.

[33] Markatou, M., Basu, A. and Linday, B. (1998). Weighted likelihood equations with bootstrap root search. *Journal of the American Statistical Association* **93** 740-750.

[34] Manton, K. G., Woodbury, M. A., Stallard, E. Riggan, W. B. Creason, J. P. and Pellon, A. C. (1989). Empirical Bayes procedures for stabilizing maps of U.S. cancer mortality rates. *Journal of the American Statistical Association* **84** 637-650.

[35] National Research Council (1992). *Combining Information: Statistical Issues and Opportunities for Research.* National Academy Press, Washington D.C..

[36] Newton, M. A. and Raftery, A. E. (1994). Approximate Bayesian inference with the weighted likelihood boostrap. *Journal of Royal Statistical Society: Series B* **56** 3-48.

[37] Rao, B. L. S. (1991). Asymptotic theory of weighted maximum likelihood estimation for growth models. In: *Statistical Inference in Stochastic Processes* 183-208, edited by Prabhu, N.U. and Basawa, I. V., Marcel Dekker, Inc., New York.

[38] Rao, C. R. (1965). *Linear Statistical Inference and Its Applications.* John Wiley & Sons, Inc., New York.

[39] Royden, H. L. (1988). *Real Analysis.* Prentice Hall, New York.

[40] Savage, L. J. (1954). *The Foundations of Statistics.* Springer-Verlag, New York.

[41] Schervish, M. J. (1995). *Theory of Statistics*, New York: Springer-Verlag.

[42] Small, C. G., Wang, J. and Yang, Z. (2000). Eliminating multiple root problems in estimation. *Statistical Science* **15**, 313-341.

[43] Smith, R. L. (1998). Bayesian and frequentist approaches to parametric predictive inference. *Bayesian Statistics* **6** 589–612.

[44] Staniswalis, J. G. (1989). The kernel estimate of a regression function in likelihood-based models. *Journal of the American Statistical Association* **89** 276-283.

[45] Stone, M. (1974). Cross-validation choice and assessment of statistical predictions. *Journal of Royal Statistical Society: Series B* **36** 111-147.

[46] Tibshirani, R. and Hastie, T. (1987). Local likelihood estimation. *Journal of the American Statistical Association* **82** 559-567.

[47] van Eeden, C. (1996). Estimation in restricted parameter spaces-Some history and some recent developments. *Statistics & Decisions* **17** 1-30.

[48] van Eeden, C. and Zidek, J. V. (1998). Combining sample information in estimating ordered normal means. *Technical Report # 182, Department of Statistics, University of British Columbia.*

[49] van Eeden, C. and Zidek, J.V. (2001). Estimating one of two normal means when their difference is bounded. *Statistics & Probability Letters* **51** 277-284.

[50] Wald, A. (1949). Note on the consistency of the maximum likelihood estimate. *The Annal of Mathematical Statistics* **15** 358-372.

[51] Waller, L. A., Louis, T. A., and Carlin, B. P. (1997). Bayes methods for combining disease and exposure data in assessing environmental justice. *Environmental and Ecological Statistics* **4** 267-281.

[52] Warm, T. A. (1987). Weighted likelihood estimation of ability in item response theory. *Psychometrika* **54** 427-450.

[53] Zidek, J. V., White, R. and Le, N. D. (1998). Using spatial data in assessing the association between air pollution episodes and respiratory morbidity. *Statistics for the Environment 4: Pollution Assessment and Control* 117 -136.