Technical Report on Weighted Likelihood Estimation and Asymptotic Properties of Maximum Weighted Likelihood Estimators [1]

Xiaogang Wang, Constance van Eeden [2] and James V. Zidek

Department of Statistics, University of British Columbia, Vancouver, B.C., Canada, V6T 1Z2

Abbreviated title:
Asymptotics for maximum weighted likelihood estimators

Abstract

The Relevance Weighted Likelihood Method was introduced by Hu and Zidek (1995) to formally embrace a variety of statistical procedures for trading bias for precision. Their approach combines all relevant information through a weighted version of the likelihood function. The present paper is concerned with the asymptotic properties of a class of *maximum weighted likelihood estimators* (WLE's) that contains those considered by Hu and Zidek (1995, 2001). Our results complement those of Hu (1997). In particular, we invoke a different asymptotic paradigm than that in Hu (1997). Moreover, our adaptive weights are allowed to depend on the data.

AMS 2000 Classifications: 62F10; 62H12
STMA 2000 Classifications: 04:010; 01:180; 04:170

Keywords: Asymptotic normality; consistency; James-Stein; weighted likelihood

1 Introduction

The weighted likelihood (WL for short) has been developed for a variety of purposes. Simple examples are easily constructed where it arises naturally as the integrated likelihood in a Bayesian framework. To give a precise description of the WL in a reasonably general setting we suppose we observe independent random response vectors $\mathbf{X_1}, ..., \mathbf{X_m}$ with probability density functions, $f_1(.;\theta_1), ..., f_m(.;\theta_m)$, where $\mathbf{X_i} = (X_{i1}, ..., X_{in_i})^t$. Further suppose that only θ_1, an unknown vector of parameters, is of inferential interest. Then for fixed $\mathbf{X} = \mathbf{x}$, the WL is defined as

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

[1]This research was supported in part by the Natural Sciences and Engineering Research Council of Canada
[2]Corresponding author

E-mail addresses: steven@stat.ubc.ca (Steven Wang), cve@xs4all.nl (Constance van Eeden), jim@stat.ubc.ca (James V. Zidek).

where $\lambda = (\lambda_1, ..., \lambda_m)$ is the " weight vector" that must be specified by the analyst.
It follows that

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

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

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

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

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

Note that the uniqueness of the WLE is not assumed.
It can be seen that the WL is an extension of the local likelihood method of Tibshirani and Hastie (1987) since the restriction of the weights to indicators or more generally kernel functions in the local likelihood is relaxed in the WL setting. A detailed discussion of the local likelihood and associated properties can be found in Eguchi and Copas (1998). Versions of the weighted likelihood can be seen in a variety of contexts (*c.f.* Newton and Raftery (1994) and Rao (1991)).
Following Hu (1997), Hu and Zidek (1995, 2001) extend the local likelihood to a more general setting but with the similar aim of combining relevant information in samples from other populations thought to resemble that whose parameters are of interest. They call their WL the REWL. In other words, referring to the definition above, the $f_2, ..., f_m$ are thought to be "similar to" $f_1(.; \theta_1)$. In their extension of the REWL, Hu and Zidek (1995) also consider simultaneous inference for all the θ's.
The results reported in the present paper extend those of Wald (1949). They differ from those of Hu (1997) because we adopt a different asymptotic paradigm. 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 of observations 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 λ.
We also consider in this paper the more general version of the adaptively weighted likelihood in which the weights are allowed to depend on the data. Such a likelihood arises naturally when the responses are measured on a sequence of independent draws 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's factors according to their degree of relevance in estimating the unknown parameter vector.
In Section 2 we present our extension of the classical large sample theory for the maximum likelihood estimator. Both consistency and asymptotic normality are shown under appropriate assumptions. The weights may be "adaptive" that is, allowed to depend on the data. In Section 3 we consider examples that demonstrate how our results may be applied. In particular, we show that our theory applies to the James-Stein estimator. Concluding remarks are given in Section 4.

2 Motivating Examples

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:

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

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

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:

$$\mathrm{WL}(\alpha) = L_1(x_1, x_2, ..., x_n; \alpha)^{\lambda_1} L_1(y_1, y_2, ..., y_n; \alpha)^{\lambda_2}, \tag{3}$$

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 (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 $\lambda_2 \le \lambda_1$ 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

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

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

$$\max_{|\beta-\alpha|\leq C} MSE(\tilde{\alpha}) < MSE(\hat{\alpha}) \text{ iff } \frac{\lambda_1}{\lambda_2} > \frac{C^2 S_t^2 + (\sigma_2^2 - \sigma_1^2)}{2(\sigma_1^2 - \rho\sigma_1\sigma_2)}. \tag{4}$$

If $\sigma_1^2 = \sigma_2^2 = \sigma^2$, then the equation (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 (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.

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*

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

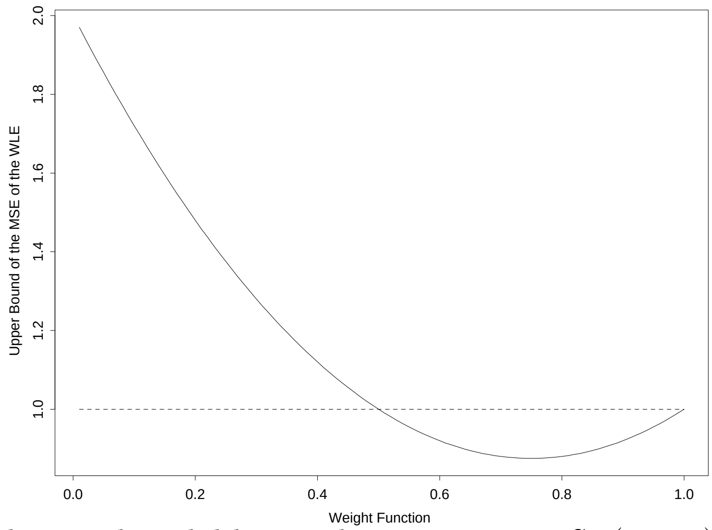

Figure 1: *A special case: the solid line is the* $max_{|\beta-\alpha|\leq C}MSE(WLE)$*, and the broken line represents the* $MSE(MLE)$*. The X-axis represents the value of* λ_1*. The parameters have been set to be the following:* $\frac{S_t^2}{\sigma^2}=1, \rho=0.5$ *and* $C=1$.

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

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 $\{(\alpha,\beta) : |\alpha - \beta| \le C\}$.

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 $|\mu_X - \mu_Y| \le C$ *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,

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

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

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

$$Var(\tilde{\mu}_Y) < Var(\bar{Y}) = \frac{\sigma^2}{n};$$
$$Cov(\tilde{\mu}_X, \tilde{\mu}_Y) > Cov(\bar{X}, \bar{Y}).$$

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,*

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

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

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

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

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

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.

3 Asymptotic Results for the WLE

In this section we establish the existence of a consistent and asymptotically normal sequence of weighted likelihood estimators under appropriate conditions. In particular, throughout this section Assumptions 2.1-2.5 stated below are assumed to hold except where otherwise stated.

3.1 Weak Consistency

Consistency, a minimal requirement for any good estimator, is explored in this subsection. To that end, we let $\theta_1^0 \in \Theta$ denote the true value of θ_1. Let $\theta^0 = (\theta_1^0, ..., \theta_m)$, for $\theta_2, ..., \theta_m \in \Theta$. Furthermore, we impose the assumptions stated next. We will then show that that consistency obtains.

Assumption 3.1 *The parameter space Θ is compact.*

Assumption 3.2 *For each i=1,...,m assume $\{X_{ij} : j = 1, ..., n_i\}$ are i.i.d. random variables having a common probability density function $f_i(x;\theta_i)$ with respect to a σ-finite measure ν.*

Assumption 3.3 *Assume $f_1(x;\theta_i) = f_1(x;\theta_i^{'})$(a.e. ν) implies that $\theta_i = \theta_i^{'}$ for any θ_i, $\theta_i^{'} \in \Theta$ and that the densities $f_1(x;\theta)$ have the same support for all $\theta \in \Theta$.*

Assumption 3.4 *For any $\theta_1^0 \in \Theta$ and for any open set $O \subseteq \Theta$, assume*

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

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

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 3.5 *Assume $\lambda^{(n_1)} = (\lambda_1^{(n_1)}, ..., \lambda_m^{(n_1)})^t$ satisfies*

$$\lambda^{(n_1)} \to (w_1, ..., w_m)^t \triangleq (1, 0, ..., 0)^t$$

while

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

for some $\delta > 0$.

To prove the measurability asserted in Assumption 2.4 in the sequel, we may be able to rely on the easily established fact that if, for all x, $U(x;\theta_1)$ is an upper semi-continuous function of θ_1 in Euclidian space, then $\sup_{\theta_1\in O} U(x;\theta_1)$ is measurable for any open set O. In particular, if $f_1(x;\theta_1)$ is upper semi-continuous in θ_1 and the open set is defined as $\{\theta_1 : |\theta_1-\theta_1^0| < R\}$ for some $R > 0$, then the Assumption 3.4 is automatically satisfied. The measurability of

$$\inf_{|\theta_1-\theta_1^0|<R} |log\frac{f_1(x,\theta_1^0)}{f_1(x,\theta_1)}|$$

also follows.

Lemma 3.1 *Let the function $A_{ij}(x)$ be measurable in x. If, for $i = 1,\dots,m$, $E_{\theta_i}[A_{ij}(X_{ij}]^2 < K_o$ for some positive constant K_o independent of the θ_i, then*

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

for any $\theta = (\theta_1,...,\theta_m)$, $\theta_i \in \Theta$, $i = 1,\dots,m$.

Proof: Put $A_{ij} = A_{ij}(X_{ij})$ and let $B_{ij} = (w_i-\lambda_i^{(n_1)})A_{ij}$. Observe that by the Cauchy-Schwartz inequality, for any i, i', j, and j', we have

$$E_\theta|B_{ij}B_{i'j'}| \;\; \leq \;\; |w_i-\lambda_i^{(n_1)}|\,|w_{i'}-\lambda_{i'}^{(n_1)}|\sqrt{E_\theta A_{ij}^2 E_\theta A_{i'j'}^2}$$

in view of the finite second moment condition on the A_{ij}.
It follows that

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

by Assumption 3.5.
Further,

$$|\frac{1}{n_1}\sum_{i=1}^{m}\sum_{j=1}^{n_i}(w_i-\lambda_i^{(n_1)})A_{ij}(x_{ij})| \leq \frac{1}{n_1}\sum_{i=1}^{m}\sum_{j=1}^{n_i}|w_i-\lambda_i^{(n_1)}||A_{ij}| = \frac{1}{n_1}\sum_{i=1}^{m}\sum_{j=1}^{n_i}|B_{ij}|.$$

We then have

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

for any θ. $\diamond$

Let

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

Under Assumptions 3.1- 3.5, it then follows from Lemma 3.1 that

$$|\frac{1}{n_1}S_{n_1}(\mathbf{X},\theta_1)| \xrightarrow{P_{\theta^0}} 0 \tag{5}$$

for any $\theta_1, ..., \theta_m \in \Theta$.

Theorem 3.1 *For each* $\theta_1 \neq \theta_1^0$,

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

for any $\theta_2, ..., \theta_m \in \Theta$.

Proof:

With P_{θ^0} measure 1,

$$\prod_{i=1}^{m}\prod_{j=1}^{n_i} f_1(X_{ij};\theta_1^0)^{\lambda_i^{(n_1)}} > \prod_{i=1}^{m}\prod_{j=1}^{n_i} f_1(X_{ij};\theta_1)^{\lambda_i^{(n_1)}}$$

if and only if

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

where

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

Observe that

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

By equation (5) we have $|\frac{1}{n_1}S_{n_1}(\mathbf{X},\theta_1)| \xrightarrow{P_{\theta^0}} 0$ for any $\theta_1, ..., \theta_m \in \Theta$.

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 $\theta_1 \neq \theta_o$, by the Shannon-Kolmogorov Information Inequality and Assumption 3.3. Therefore $\lim_{n_1\to\infty} P_{\theta^0}(W_{n_1}(\mathbf{X},\theta_1) > 0) = 1$ for all $\theta_1 \neq \theta_1^0$ and all $\theta_2..., \theta_m$ thus implying the conclusion asserted in the theorem. $\diamond$

In the sequel, we will let $||.||$, be the Euclidean norm,

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

for any $\mathbf{x} = (x_1, ..., x_q)^t$. Θ, Furthermore, for any open set O, let

$$Z_{ij}(O) = \inf_{\theta_1' \in O} log(f_1(X_{ij};\theta_1^0)/f_1(X_{ij};\theta_1')).$$

Theorem 3.2 *Suppose* $logf_1(x;\theta)$ *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* $for\ all$ $\epsilon > 0$,

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

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 closed 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 $logf_1(x;\theta_1^0)/f_1(x;\theta_1)$ 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

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

by Assumption 3.4. This implies that $Z_{1j}(N_{\theta_1}^{(k)})$ are integrable. Using the monotone convergence theorem, we then have

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

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

$$\begin{aligned}
& P_{\theta^0}(||\tilde{\theta}_1^{(n_1)} - \theta_1^0|| \geq \epsilon) \\
= \ & P_{\theta^0}(\tilde{\theta}_1^{(n_1)} \in N_{\theta_1^l}^* \ \text{ for some } l) \\
\leq \ & \sum_{l=1}^p P_{\theta^0}(\tilde{\theta}_1^{(n_1)} \in N_{\theta_1^l}^*) \\
\leq \ & \sum_{l=1}^p P_{\theta^0}\Big(\frac{1}{n_1}\sum_{i=1}^m\sum_{j=1}^{n_i}\lambda_i^{(n_1)} Z_{ij}(N_{\theta_1^l}^*) \leq 0\Big) \\
= \ & \sum_{l=1}^p P_{\theta^0}\Big(\frac{1}{n_1}\sum_{j=1}^{n_1} Z_{1j}(N_{\theta_1^l}^*) + \frac{1}{n_1}\sum_{i=1}^m\sum_{j=1}^{n_i}(\lambda_i^{(n_1)} - w_i) Z_{ij}(N_{\theta_1^l}^*) \leq 0\Big).
\end{aligned}$$

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

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

Since $E_{\theta^0}Z_{ij}(N_{\theta_1^l}^*)^2 \leq E_{\theta_i}[\sup_{\theta_1\in\Theta}|log(f_1(X_{ij};\theta_1^0)/f_1(X_{ij};\theta_1))|]^2 \leq K < \infty$ by Assumption 3.4, it follows from Lemma 3.1 that

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

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 $\theta_1^l \in \Theta \setminus N_0$,

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

This implies that

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

Thus the assertion follows. $\diamond$

In the next theorem we drop Assumption 3.1 and replace it with a slightly different condition. At the same time we keep Assumption 3.2- 3.5.

Theorem 3.3 *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*

$$0 < E_{\theta^0}\big\{ \inf_{\theta_1' \in C^c \cap \Theta} log \frac{f_1(X_{ij};\theta_1^0)}{f_1(X_{ij};\theta_1')} \big\} \leq K^C < \infty, \tag{6}$$

where K^C is a constant independent of the θ_i. Then for any sequence of maximum weighted likelihood estimates $\tilde{\theta}_1^{(n_1)}$ of θ_1^o and for all $\epsilon > 0$

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

for any $\theta_2, ..., \theta_m \in \Theta$.

Proof: Let N_0 and ϵ be as in the proof of Theorem 3.2, 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

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

It follows from the proof of Theorem 3.2 that the first term of last expression goes to zero as n goes to infinity.
By the Weak Law of Large Numbers, we have

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

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

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

By the Weak Law of Large Numbers, it follows that

$$\frac{1}{n_i}\sum_{j=1}^{n_i} Z_{ij}(C^c\cap\Theta) \xrightarrow{P_{\theta^0}} E_{\theta^0}\{\inf_{\theta_1\in C^c\cap\Theta} log\frac{f_1(X_{ij};\theta_1^0)}{f_1(X_{ij};\theta_1)}\} \tag{7}$$

where $E_{\theta^0}\{\inf_{\theta_1\in C^c\cap\Theta} log(f_1(X_{ij};\theta_1^0)/f_1(X_{ij};\theta_1))\}$ is a finite number by the hypotheses of this theorem. By Assumption 3.5, it follows that

$$\frac{n_i}{n_1}(\lambda_i^{(n_1)} - w_i) \longrightarrow 0, \quad as \ \ n_1\to\infty. \tag{8}$$

Combining equation (7) and (8), we then have

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

The hard part of using the theorem is verifying the conditions. We illustrate their use in Section 3.

3.2 Asymptotic Normality

In practice, the WLE will usually be found by computing the roots of the likelihood equation. In this subsection, we turn our attention to these roots and to that end restrict our attention to vector valued parameters with real valued co-ordinates. We are then able to address both the consistency and asymptotic normality of those roots.

To obtain the asymptotic normality of WLE, more restrictive conditions are needed. In particular, some conditions will be imposed upon the first and second derivatives of the likelihood function.
For each fixed sample size, 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.* $\theta_1 = (\theta_{11}, ..., \theta_{1p})$ and the true value of the parameter is $\theta_1^0 = (\theta_{11}^0, ..., \theta_{1p}^0)$. As a notational convenience we will in the sequel take $\partial/\partial\theta_1$ to mean the column gradient vector obtained by co-ordinate-wise differentiation with respect to θ_1. Write

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

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

$$E_{\theta_1^0}\psi(X_{1j};\theta_1^0) = \int \frac{\partial}{\partial\theta_1} f_1(x;\theta_1^0)d\nu(x) = 0, \tag{9}$$

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

$$\begin{aligned} E_{\theta_1^0}\dot{\psi}(X_{1j},\theta_1^0) &= \int [\frac{\partial}{\partial\theta_1}\frac{(\partial/\partial\theta_1)f_1(x;\theta_1^0)}{f_1(x;\theta_1^0)}] f_1(x;\theta_1^0)d\nu(x) \\ &= 0 - \int \psi(x;\theta_1^0)\psi(x;\theta_1^0)^t f_1(x;\theta_1^0)d\nu(x). \end{aligned}$$

Thus $I(\theta_1^0) = -E_{\theta_1^0}\dot{\psi}(X_1;\theta_1^0)$.
To simplify notation, let

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

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

Theorem 3.4 *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

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

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

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

We will show that $P_{\theta_1^0}(\mathbf{X} \in I_n(a)) \to 1$ for all sufficiently small a. That is, for any given ϵ, there exist N_ϵ such that, for any $n > N_\epsilon$, we have $P_{\theta_1^0}(\mathbf{X} \in I_n(a)) > 1 - \epsilon$. 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

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

and

$$
\begin{aligned}
A_{k_1}(\mathbf{x}) &= \sum_{i=1}^{m}\sum_{j=1}^{n_i}\lambda_i^{(n_1)}\frac{\partial log f_1(x_{ij};\theta_1)}{\partial\theta_{1k_1}}|_{\theta_1=\theta_1^0}, \\
B_{k_1k_2}(\mathbf{x}) &= \sum_{i=1}^{m}\sum_{j=1}^{n_i}\lambda_i^{(n_1)}\frac{\partial log f_1(x_{ij};\theta_1)}{\partial\theta_{1k_1}\partial\theta_{1k_2}}|_{\theta_1=\theta_1^0},
\end{aligned}
$$

and $|\zeta_{k_1k_2k_3}(x_{ij})|\le 1$ by hypothesis (2) of this theorem.
By the Weak Law of Large Numbers and (9)

$$
\frac{1}{n_1}\sum_{j=1}^{n_1}\frac{\partial log f_1(X_{1j};\theta_1)}{\partial\theta_{1k_1}}|_{\theta_1=\theta_1^0}\xrightarrow{P_{\theta_1^0}} 0 \quad as \;\; n_1\to\infty, \tag{10}
$$

$$
\frac{1}{n_1}\sum_{j=1}^{n_1}\frac{\partial^2 log f_1(X_{1j};\theta_1)}{\partial\theta_{1k_1}\partial\theta_{1k_2}}|_{\theta_1=\theta_1^0}\xrightarrow{P_{\theta_1^0}} -I_{k_1k_2}(\theta_1^0) \quad as \;\; n_1\to\infty \tag{11}
$$

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 3.1, we then have

$$
\frac{1}{n_1}\sum_{i=1}^{m}\sum_{j=1}^{n_i}(\lambda_i^{(n_1)}-w_i)\frac{\partial log f_1(X_{ij};\theta_1)}{\partial\theta_{1k_1}}|_{\theta_1=\theta_1^0}\xrightarrow{P_{\theta 0}} 0 \quad as \;\; n_1\to\infty; \tag{12}
$$

$$
\frac{1}{n_1}\sum_{i=1}^{m}\sum_{j=1}^{n_i}(\lambda_i^{(n_1)}-w_i)\frac{\partial^2 log f_1(X_{ij};\theta_1)}{\partial\theta_{1k_1}\partial\theta_{1k_2}}|_{\theta_1=\theta_1^0}\xrightarrow{P_{\theta 0}} 0 \quad as \;\; n_1\to\infty; \tag{13}
$$

$$
\frac{1}{n_1}\sum_{i=1}^{m}\sum_{j=1}^{n_i}(\lambda_i^{n_1}-w_i)G_3(X_{ij})\xrightarrow{P_{\theta 0}} 0 \quad as \;\; n_1\to\infty. \tag{14}
$$

To prove that the maximum difference $\frac{1}{n_1}WL(\mathbf{x};\theta_1^b) - \frac{1}{n_1}WL(\mathbf{x};\theta_1^0)$ over all the boundary point 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 probability tending to 1, the maximum of S_2 for all the boundary point 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

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

By (10) and (12), it then follows that, for any $\theta_2,\ldots,\theta_m \in \Theta$,

$$\frac{1}{n_1}A_{k_1}(\mathbf{X}) \xrightarrow{P_{\theta^0}} 0 \quad as \;\; n_1 \to \infty. \tag{15}$$

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

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

For any given a, it follows from (15) that with P_{θ^0} probability tending to 1

$$\frac{1}{n_1}\sum_{k_1=1}^{p}|A_{k_1}(\mathbf{X})| < pa^2 \tag{16}$$

and (16) then gives

$$|S_1| < pa^3 \tag{17}$$

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

$$\begin{aligned} 2S_2 \;\; = \;\; & - \sum_{k_1=1}^{p}\sum_{k_2=1}^{p}(\theta_{1k_1} - \theta_{1k_1}^0)I_{k_1k_2}(\theta_1^0)(\theta_{1k_2} - \theta_{1k_2}^0) \\ & + \sum_{k_1=1}^{p}\sum_{k_2=1}^{p}(\theta_{1k_1} - \theta_{1k_1}^0)\big(\tfrac{1}{n_1}B_{k_1k_2} + I_{k_1k_2}(\theta_1^0)\big)(\theta_{1k_2} - \theta_{1k_2}^0). \end{aligned} \tag{18}$$

For the second term in the above expression, consider

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

By (11) and (13), it then follows that, for any k_1 and k_2,

$$\frac{1}{n_1}B_{k_1k_2} + I_{k_1k_2}(\theta_1^0) \xrightarrow{P_{\theta^0}} 0 \quad as \;\; n_1 \to \infty. \tag{19}$$

Thus, for any boundary points such that $||\theta_1^b - \theta_1^0|| = a$, we have

$$|\sum_{k_1=1}^{p}\sum_{k_2=1}^{p}(\theta_{1k_1} - \theta_{1k_1}^0)(\theta_{1k_2} - \theta_{1k_2}^0)| < p^2a^2. \tag{20}$$

By equation (19) and (20), it follows that, for given a

$$|\sum_{k_1=1}^{p}\sum_{k_2=1}^{p}(\theta_{1k_1} - \theta_{1k_1}^0)\big(\frac{1}{n_1}B_{k_1k_2} + I_{k_1k_2}(\theta_1^0)\big)(\theta_{1k_2} - \theta_{1k_2}^0)| < p^2a^3 \tag{21}$$

with P_{θ^0} probability tending to 1.
Let us examine the first term in (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 $\xi = (\xi_1, \xi_2, ..., \xi_p)^t = B(\theta_{11}^b - \theta_{11}^0, \theta_{12}^b - \theta_{12}^0, ..., \theta_{1p}^b - \theta_{1p}^0)^t$. Then we have $||\xi||^2 = ||\theta_1^b - \theta_1^0||^2 = a^2$. It follows that

$$-\sum_{k_1=1}^{p}\sum_{k_2=1}^{p}(\theta_{1k_1}^b - \theta_{1k_1}^0)I_{k_1k_2}(\theta_1^0)(\theta_{1k_2}^b - \theta_{1k_2}^0) = \sum_{l=1}^{m}\delta_l\xi_l^2 \le \delta^* a^2 \tag{22}$$

where $\delta^* = \max\{\delta_1, ..., \delta_p\} < 0$. We see that there exist a_0 such that $\delta^* + p^2a_0 \le 0$.
Combining (21) and (22), it follows that with P_{θ^0} probability tending to 1 there exists $c > 0$ such that for $a < a_0$

$$S_2 < -ca^2. \tag{23}$$

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

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

By the Weak Law of Large Numbers with P_{θ_0} probability tending to 1,

$$|\frac{1}{n_1}\sum_{j=1}^{n_1}G_3(x_{1j})| < 2(1 + K_3) \tag{24}$$

where we use the inequality $|EZ| \leq E|Z| \leq 1 + E|Z|^2$ for any random variable Z. By (14), it follows that with P_{θ_0} probability tending to 1

$$|\frac{1}{n_1}\sum_{i=1}^{m}\sum_{j=1}^{n_i}(\lambda_i^{(n_a)1} - w_i)G_3(X_{ij})| < 1 + K_3. \tag{25}$$

Hence by (24) and (25) , for any given a and θ_1^b such that $||\theta_1^b - \theta_1^0|| = a$

$$|S_3| < \frac{p^3}{2}(1 + K_3)a^3. \tag{26}$$

Finally, combining (17), (23) and (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

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

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 $WL(\mathbf{x}; \theta_1)$ has a local maximum, $\frac{\partial}{\partial\theta_1}WL_{n_1}(\mathbf{x}; \theta_1)|_{\theta_1 = \tilde{\theta}_1^{(n_1)}} = 0$. Since $P_{\theta_1^0}\big(\mathbf{X} \in I_n(a)\big) \longrightarrow 1$ 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 $P_{\theta^0}(||\tilde{\theta}_1^{(n_1)}(a) - \theta_1^0|| < a) \to 1$. 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 $\frac{\partial}{\partial\theta_1}WL_{n_1}(\mathbf{x}; \theta_1)$. Thus $\theta_1^{(n_1)^*}$ does not depend on a and $P_{\theta^0}(||\theta_1^{(n_1)^*} - \theta_1^0|| < a) \to 1$.

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

$$log\dot{W}L_{n_1}(\mathbf{x}; \theta_1) = log\dot{W}L_{n_1}(\mathbf{x}; \theta_1^0) + \int_0^1 \sum_{i=1}^{m}\sum_{j=1}^{n_i}\lambda_i^{(n_1)}\dot{\psi}(x_{ij}; \theta_1^0 + t(\theta_1 - \theta_1^0))dt(\theta_1 - \theta_1^0),$$

where $log\dot{W}L_{n_1}(\mathbf{x};\theta_1^0) = \sum_{i=1}^{m}\sum_{j=1}^{n_i}\lambda_i^{(n_1)}\psi(x_{ij};\theta_1^0)$.

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

$$\frac{1}{\sqrt{n_1}}log\dot{W}L_{n_1}(\mathbf{x};\theta_1^0) = B_{n_1}\sqrt{n_1}(\tilde{\theta}_1^{(n_1)} - \theta_1^0), \tag{27}$$

where

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

Note that

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

By (27), it follows that

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

¿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

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

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

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

Now we prove

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

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

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

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

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

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 hypothesis (1) of this theorem, so there is a $\rho > 0$ such that

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

It then follows that

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

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

$$P_{\theta^0}\big(|E_{\theta^0}\dot{\psi}\big(X_{1j};\theta_1^0 + t(\tilde{\theta}_1^{(n_1)} - \theta_1^0)) + I(\theta_1^0)| < \epsilon\big) \longrightarrow 1\ \ as\ \ n_1 \longrightarrow \infty. \tag{28}$$

Furthermore for any $t \in \mathcal{R}$ such that $0 \leq t \leq 1$, if $||\tilde{\theta}_1^{(n_1)} - \theta_1^0|| < \rho$, then

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

This result implies that

$$P_{\theta^0}\big(\theta_1^0 + t(\tilde{\theta}_1^{(n_1)} - \theta_1^0) \in S_\rho\big) \longrightarrow 1 \quad as\ \ n_1 \to \infty. \tag{29}$$

From the Uniform Strong Law of Large Numbers, Theorem 16(a) in Ferguson (1996), with $P_{\theta_1^0}$ probability 1, there is an integer N_I such that

$$n_1 > N_I \quad \Rightarrow \quad \sup_{\theta_1 \in S_\rho} |\frac{1}{n_1}\sum_{j=1}^{n_1} \dot{\psi}(X_{1j}, \theta_1) - E_{\theta^0}\dot{\psi}(X_{11}, \theta_1)| < \epsilon. \tag{30}$$

Assume N is large enough that $n_1 > N_I \Rightarrow ||\tilde{\theta}_1^{(n_1)} - \theta_1^0|| < \rho$. Then

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

by equation (28), (29) and (30).
Next we prove $B_{n_1}^{II} \xrightarrow{P_{\theta^0}} (0, ..., 0)^t \;\; as \;\; n_1 \longrightarrow \infty$.
By Lemma 3.1, every component of $B_{n_1}^{II}$ goes to 0 in probability. Thus

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

This completes the proof. ⋄

REMARKS:

1. 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.

2. Realistically, the weight vector will often have to estimated in practice. We then refer to the weighted likelihood as "adaptively weighted." It turns out that our results on consistency and asymptotic normality are easily extended to this more general case. Assumptions 3.1-3.4 are required along with 1 an additional condition:

   Assumption 3.6 *Assume:*
   (i) $\lim_{n_1 \to \infty} \frac{n_i}{n_1} < \infty$, *for* $i = 1, 2, ..., m$*;*

(ii) the adaptive weight vector $\lambda^{(n_1)}(\mathbf{X}) = (\lambda_1^{(n_1)}(\mathbf{X}), ..., \lambda_m^{(n_1)}(\mathbf{X}))^t$ *satisfies, for any* $\epsilon > 0$,

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

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

3. By strengthening our assumptions one can obtain strong consistency even in the adaptively weighted case (Wang 2001). However, for brevity, these results will not be included in this paper.

4 Examples.

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

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

$$\begin{aligned} I_{1i} &= -\int\limits_{|x|>M} \frac{1}{2}(x-\theta_1^0)^2 \frac{1}{\sqrt{2\pi}} exp^{-(x-\theta_i)^2/2} dx, \\ I_{2i} &= \int\limits_0^M \left(-\frac{1}{2}(x-\theta_1^0)^2 + \frac{1}{2}(x-M)^2\right) \frac{1}{\sqrt{2\pi}} exp^{-(x-\theta_i)^2/2} dx, \\ I_{3i} &= \int\limits_{-M}^0 \left(-\frac{1}{2}(x-\theta_1^0)^2 + \frac{1}{2}(x+M)^2\right) \frac{1}{\sqrt{2\pi}} exp^{-(x-\theta_i)^2/2} dx. \end{aligned}$$

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 $i = 1, 2, ..., m, j = 1, 2, ..., n_i$,

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

4.2 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 3.2 and 3.3 are obviously satisfied for this example. The condition (6) in Theorem 3.3 is satisfied as shown in the previous example. If we show that Assumption 3.5 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 $n_i \bar{X}_{i.} = \sum_{j=1}^{n_i} X_{ij}$, $i = 1, \ldots, m$, $V = Cov((\bar{X}_{1.}, \bar{X}_{2.})^t)$ 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

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

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 3.5 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 3.4 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 Mulivariate 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,

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

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 :

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

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

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

We then have

$$P_{\theta^0}\Big(\Big|\frac{1}{n_1^{1+\delta}}\frac{m-2}{\sum\limits_{i=1}^{m}\bar{X}_i^2+c}\Big|>\epsilon\Big)=O\Big(\frac{1}{n_1^{1+\delta}}\Big); \tag{31}$$

$$P_{\theta^0}(\lambda_1<0)=O\Big(\frac{1}{n_1^{1+\delta}}\Big). \tag{32}$$

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 3.6 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 $i = 1, 2, ..., m$ and Wang (2001) shows this leads to strong consistency. Since strong consistency implies weak consistency, asymptotic normality of the WLE using adaptive weights will follow in this case.

5 Application to Disease Mapping

5.1 Introduction

In this section, 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).

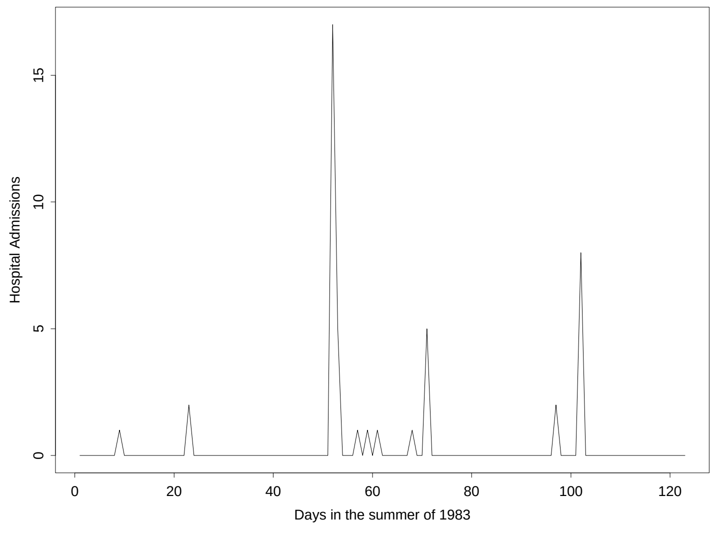

Figure 2: Daily hospital admissions for CSD # 380 in the summer of 1983.

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 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.

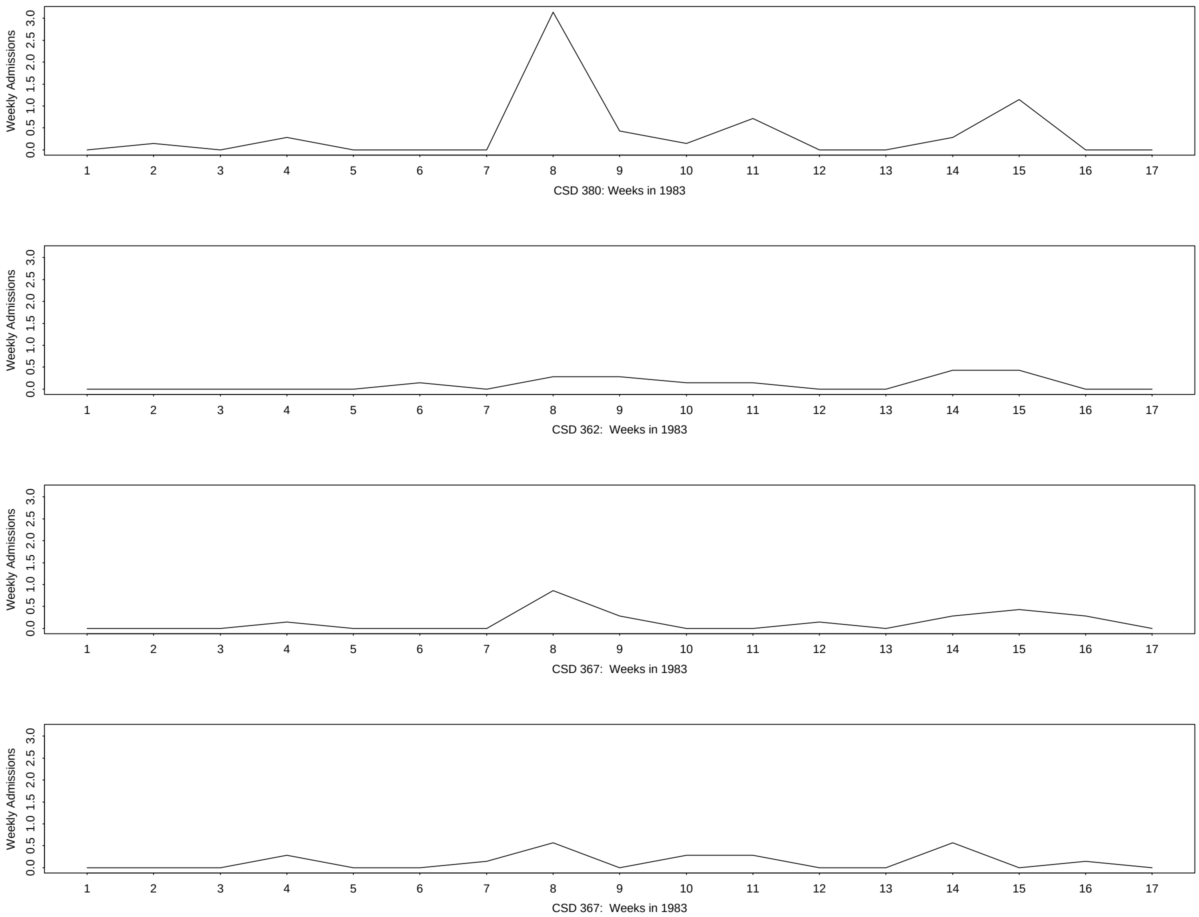

Figure 3: Hospital Admissions for CSD # 380, 362, 366 and 367 in 1983.

5.2 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,

$$Y_{ij}^{q} \overset{ind.}{\sim} \mathcal{P}\left(\theta_{ij}^{q}\right), j = 1, 2, ..., 17; i = 1, 2, ..., 733; q = 1, 2, ..., 6.$$

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 section, 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 $A_q(\underline{y})_{ik} = \sum_{j=1}^{17} \overline{Y}_{i.}^{q(-j)} \overline{Y}_{k.}^{q(-j)}$, $i = 1, 2, 3, 4; k = 1, 2, 3, 4.$

5.3 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:

$$MSE_M^q = \widehat{var(Y_{11})}/16, \quad q = 1, 2, ...6.$$

The estimated MSE for the WLE is give as following:

$$\begin{aligned} MSE_W^q &= E\left(\sum_{i=1}^{m} \lambda_i^q \overline{Y}_{i.}^q - \theta_1^q\right)^2 \\ &= Var\left(\sum_{i=1}^{m} \lambda_i^q \overline{Y}_{i.}^q\right) + \left(E\sum_{i=1}^{m} \lambda_i^q \overline{Y}_{i.}^q - \theta_1^q\right)^2 \\ &\approx \sum_{i=1}^{4}\sum_{k=1}^{4} \lambda_i^q \lambda_j^q \widehat{cov}(\overline{Y}_{i.}^q, \overline{Y}_{j.}^q) + \left(\sum_{i=1}^{m} \lambda_i^q \overline{Y}_{i.}^q - \overline{Y}_{1.}^q\right)^2. \end{aligned}$$

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 1: Estimated MSE for the MLE and the WLE.

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%.

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:

$$
\begin{aligned}
PRED_M &= \sqrt{\frac{1}{5}\sum_{q=1}^{5}\left(\overline{Y}_{1.}^{q} - \overline{Y}_{1.}^{q+1}\right)^2}\,; \\
PRED_W &= \sqrt{\frac{1}{5}\sum_{q=1}^{5}\left(WLE^{q} - \overline{Y}_{1.}^{q+1}\right)^2}\,.
\end{aligned}
$$

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.

5.4 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.

	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 2: Correlation matrix and the weight function for 1984.

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 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 3: MSE of the MLE and the WLE for CSD 380.

We remark that this section 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.

6 Concluding Remarks

In this paper 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 2.6 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.

References

Eguchi, S., Copas, J., 1998. A class of local likelihood methods and near-parametric asymptotics, J. Royal Statist. Soc. Ser. B 60, 709-724.

Ferguson, T.S., 1996. A Course in Large Sample Theory. Chapman and Hall, New York.

Hu, F., 1997. The asymptotic properties of the maximum-relevance weighted likelihood estimators, Canad. J. Statist., 25, 45-59.

Hu, F., Zidek, J.V., 1995. Incorporating relevant sample information using the likelihood. Technical Report No. 161 Dept. of Statistics, The University of British Columbia, Vancouver, B.C., Canada.

Hu, F., Zidek, J.V., 2001. The relevance weighted likelihood with applications. In: Ahmed, S.E., Reid, N. (Eds.), Empirical Bayes and Likelihood Inference, Springer Verlag, New York, pp. 211-235.

Lehmann, E.L., 1983. Theory of Point Estimation. New York: John Wiley & Sons.

Newton, M.A., Raftery, A.E., 1994. Approximate Bayesian inference with the weighted likelihood bootstrap. J. Royal Statist. Soc. Ser. B, 56, 3-48.

Rao, P.B.L.S., 1991. Asymptotic theory of weighted maximum likelihood estimation for growth models. In: Prabhu, N., Vasawa, I.V. (Eds.), Statistical Inference for Stochastic Processes, Dekker, New York, pp. 183-208.

Schervish, M.J., 1995. Theory of Statistics. New York: Springer-Verlag.

Sen, P.K., Singer, J.M., 1993. Large Sample Methods in Statistics. New York: Springer-Verlag.

Serfling, J.R., 1980. *Approximation Theorems of Mathematical Statistics.* Wiley, New York.

Tibshirani, R., Hastie, T., 1987. Local likelihood estimation. J. Amer. Statist. Assoc., 82, 559–567.

van Eeden, C., Zidek, J.V., 2001. Estimating one of two normal means when their difference is bounded. Statist. Probab. Lett., 51, 277-284.

van Eeden, C., Zidek, J.V., 2000. Combining the data from two normal populations to estimate the mean of one when their mean difference is bounded. Submitted.

Wald, A., 1949. Note on the consistency of the maximum likelihood estimate, Ann. Math. Statist., 15, 358-372.

Wang, X., 2001. Maximum Weighted Likelihood Estimation. PhD Thesis. Department of Statistics, The University of British Columbia, Vancouver, B.C., Canada.