Available at
www.ElsevierMathematics.com
POWERED BY SCIENCE DIRECT
Journal of Statistical Planning and Inference 119 (2004) 37–54

journal of statistical planning and inference

www.elsevier.com/locate/jspi

Asymptotic properties of maximum weighted likelihood estimators☆

Xiaogang Wang*, Constance van Eeden[1], James V. Zidek

Department of Statistics, University of British Columbia, Vancouver, BC, Canada, V6T 1Z2

Abstract

The relevance weighted likelihood method was introduced by Hu and Zidek (Technical Report No. 161, Department of Statistics, The University of British Columbia, Vancouver, BC, Canada, 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* that contains those considered by Hu and Zidek (Technical Report No. 161, Department of Statistics, The University of British Columbia, Vancouver, BC, Canada, 1995, in: Ahmed, S.E. Reid, N. (Eds.), Empirical Bayes and Likelihood Inference, Springer, New York, 2001, p. 211). Our results complement those of Hu (Can. J. Stat. 25 (1997) 45). In particular, we invoke a different asymptotic paradigm than that in Hu (Can. J. Stat. 25 (1997) 45). Moreover, our adaptive weights are allowed to depend on the data.

MSC: 62F10; 62H12

Keywords: Asymptotic normality; Consistency; James-Stein; Weighted likelihood

1. Introduction

The weighted likelihood (WL) has been developed for a variety of purposes. The underlying heuristics, in fact, are embraced by many inferential methods such as weighted least squares and kernel smoothers. In particular, they seek to reduce the

☆ This research was supported in part by the Natural Sciences and Engineering Research Council of Canada
* Corresponding author. Tel.: +14167362100; fax: +14167365757.
E-mail addresses: steven@mathstat.yorku.ca (X. Wang), cve@xs4all.nl (Constance van Eeden), jim@stat.ubc.ca (James V. Zidek).
[1] Also for correspondence.

0378-3758/02/$ - see front matter
PII: S0378-3758(02)00410-X

variance of estimators in exchange for increasing their bias, with the goal of reducing their mean-squared-error (MSE), i.e. increasing their precision. Substantial gains in precision are achievable, as evidenced by the celebrated James–Stein estimator, itself a weighted likelihood estimator (WLE) with 'adaptive', i.e. estimated weights.

The inferential method described in this paper can be useful in practice since the samples from some 'surrogate' populations may cost less than those from the population of direct interest. For example, a survey sample drawn previously from the current population, even though biased owing to the evolutionary change in that population, provides relevant information. Since it is already in hand, it will essentially cost nothing. It seems apparent that statisticians should use all relevant information available to them in making statistical inference about a population so as to maximally reduce their uncertainty about it. The WL helps statisticians to do just that.

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.

We present two examples to demonstrate our points. The first one is an over-simplified scenario in which two regression models are available, i.e.,

$$X_i = \theta_1 t_i + \varepsilon_i, \quad \varepsilon_i \sim \mathrm{N}(0, \sigma_1^2), \ \ i = 1, \ldots, n_1, \tag{1}$$

$$Y_i = \theta_2 t_i + \varepsilon_i'; \quad \varepsilon_i' \sim \mathrm{N}(0, \sigma_2^2), \ \ i = 1, \ldots, n_2, \tag{2}$$

where the $\{t_i\}_{i=1}^n$ are fixed. The $\{\varepsilon_i\}$'s are i.i.d. So are the $\{\varepsilon_i'\}$'s, while $\mathrm{Cov}(\varepsilon_i, \varepsilon_j') = 0$ if $i \neq j$; $\mathrm{Cov}(\varepsilon_i, \varepsilon_i') = \rho\sigma_1\sigma_2$ for all i. 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 parameter θ_1 is of primary interest. The question is whether we can integrate the information from the second sample to yield a more reliable estimate for the regression coefficient of the first one. The answer is affirmative. Wang et al. (2002) show that when θ_1 and θ_2 are close, the WLE for θ_1 has a smaller MSE when compared with the traditional MLE,

The second example is more realistic and involves an important topic in disease mapping. Wang et al. (2002) apply the *maximum WL* approach to parallel time series of hospital-based health data. Specifically, the *WL* approach 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. We assume that the total number of hospital admissions of a week for a particular CSD follows a Poisson distribution, i.e., for year q, CSD i and week j,

$$Y_{ij}^q \overset{\text{ind.}}{\sim} \mathscr{P}(\theta_i^q), \quad j = 1, 2, \ldots, 17; \ \ i = 1, 2, \ldots, 733; \ \ q = 1, 2, \ldots, 6.$$

The raw estimate of θ_i^q is highly unreliable due to the nature of disease data. Extra variation in disease can arise from a variety of causes. In the simplest case, it may be

that there are many underlying geographical factors that are unknown to us. Extra variation leads to increased differences between estimates or measures at different locations. By combining information from adjacent CSD's, this type of variation will be reduced in the mapping. The WLE with adaptive weights has shown advantages over the traditional MLE in the study detailed in Wang et al. (2002). The estimated MSE for WLE is significantly smaller than that of the MLE. More importantly, the WLE down-weights those CSDs which have similar pattern or large correlation with the current one since the estimator realizes that there is not much to be gained by incorporating almost redundant information.

To give a precise description of the WL in a reasonably general setting we suppose we observe independent random response vectors $\mathbf{X_1},\ldots,\mathbf{X_m}$ with probability density functions $f_1(.;\theta_1),\ldots,f_m(.;\theta_m)$, where $\mathbf{X_i}=(X_{i1},\ldots,X_{in_i})^{\mathrm{t}}$. Further suppose that only population 1, in particular θ_1, an unknown vector of parameters, is of inferential interest. The classical likelihood would be

$$L_1(\mathbf{x_1},\theta_1)=\prod_{j=1}^{n_1} f(x_{1j};\theta_1).$$

However, assume that the remaining parameters, $\theta_2,\ldots,\theta_m$, are close to θ_1. This suggests a WL defined as

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

for fixed $\mathbf{X}=\mathbf{x}$, where $\lambda=(\lambda_1,\ldots,\lambda_m)$ is the 'weight vector' that must be specified by the analyst. Note that the remaining parameters $\theta_2,\ldots,\theta_m$ do not appear in the WL defined as above since the inferential interest is on θ_1, the parameter of population 1. Instead, the samples generated from all the populations are incorporated into the WL.

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 estimator WLE for θ_1 if

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

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) for non-parametric regression. In that method the weights

are set to 0 or 1 according as the remaining x's are near to the regressor of interest, x, or not. More generally, kernel functions in the local likelihood are used. In other words, the weights are merely indicator functions of proximity to x. However, restrictions to such weights are relaxed when the WL is applied in that setting. A detailed discussion of the local likelihood and associated properties can be found in Eguchi and Copas (1998). Versions of the WL can be seen in a variety of contexts (cf. Newton and Raftery, 1994; 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. Let $\mathbf{X}=(X_1,X_2,\ldots,X_n)$ be random variables with probability density functions $f_1,f_2,\ldots,f_n$. The density of interest is $f(.,\theta),\theta\in\Theta$ of a study variable X, θ being an unknown parameter. At least in some qualitative sense, the $f_1,f_2,\ldots,f_n$ are thought to be 'like' $f(.,\theta)$. For fixed $\mathbf{X}=\mathbf{x}$, the relevance weighted likelihood (REWL) function is defined as

$$\prod_{i=1}^{n} f^{\lambda_{ni}}(x_i,\theta).$$

Here the λ_{ni} are the so-called relevance weights which depend on the relationship between f_1 and $f(.,\theta)$. 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 WL 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. Asymptotic results for the WLE

In this section we establish the existence of a consistent and asymptotically normal sequence of WL estimators under appropriate conditions. In particular, throughout this section Assumptions 2.1–2.5 stated below are assumed to hold except where otherwise stated. Proofs of our results can be found in the appendix.

2.1. 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 $\theta^0 = (\theta_1^0, \theta_2, \ldots, \theta_m)$, for $\theta_2, \ldots, \theta_m \in \Theta$. Furthermore, we impose the assumptions stated next. We will then show that consistency obtains.

Assumption 2.1. The parameter space Θ is compact.

Assumption 2.2. For each $i = 1, \ldots, m$ assume $\{X_{ij} : j = 1, \ldots, n_i\}$ are i.i.d. random variables having a common probability density function $f_i(x; \theta_i)$ with respect to a σ-finite measure v.

Assumption 2.3. Assume $f_1(x; \theta_i) = f_1(x; \theta_i')$ (a.e. v) implies that $\theta_i = \theta_i'$ for any $\theta_i, \theta_i' \in \Theta$ and that the densities $f_1(x; \theta)$ have the same support for all $\theta \in \Theta$.

Assumption 2.4. For any $\theta_1^0 \in \Theta$ and for any open set $O \subseteq \Theta$, assume

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

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

are each measurable in x and

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

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

Assumption 2.5. Assume $\lambda^{(n_1)} = (\lambda_1^{(n_1)}, \ldots, \lambda_m^{(n_1)})^{\mathrm{t}}$ satisfies

$$\lambda^{(n_1)} \to (w_1, \ldots, w_m)^{\mathrm{t}} = (1, 0, \ldots, 0)^{\mathrm{t}}$$

while

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

for some $\delta > 0$.

Assumption 2.1 is relaxed in Theorem 2.3. 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 Euclidean 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 2.4 is automatically satisfied. 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. Assumption 2.5 is the most important assumption. It implies that $\lambda_k^{(n_1)} \leqslant M n_1^{(1-\delta)/2}/n_k$ for some constant M and $k = 2,\ldots,m$. Thus it governs the degree of combining information from other populations to yield a more reliable estimate of the parameter of interest without losing weak consistency and asymptotic normality. For our proofs we need the following lemma.

Lemma 2.1. *Let the functions $A_{ij}(x)$ be measurable in $x, x\in R^p$. If, for $i = 1,\ldots,m$; $j = 1,2,\ldots,n_i$, $E_{\theta_i}[A_{ij}(X_{ij})]^2 < K_o$ 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 $\theta = (\theta_1,\ldots,\theta_m)$, $\theta_i\in\Theta$, $i = 1,\ldots,m$.

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 2.1–2.5, it then follows from Lemma 2.1 that

$$\left|\frac{1}{n_1} S_{n_1}(\mathbf{X},\theta_1)\right| \xrightarrow{P_{\theta^0}} 0 \tag{3}$$

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

Theorem 2.1. *For each $\theta_1 \neq \theta_1^0$,*

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

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

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

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

for any $\mathbf{x} = (x_1, \ldots, x_q)^{\mathrm{t}}$. 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 2.2. *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 $\theta_1 \in N_{\theta_1} \subset \Theta$. *Then for any sequence of maximum WLE* $\tilde{\theta}_1^{(n_1)}$ *of* θ_1, *and for all* $\varepsilon > 0$,

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

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

In the next theorem we drop Assumption 2.1 and replace it with a slightly different condition. At the same time we keep Assumptions 2.2–2.5.

Theorem 2.3. *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*

$$0 < E_{\theta^0}\left\{\inf_{\theta_1' \in C^c \cap \Theta} \log \frac{f_1(X_{ij}; \theta_1^0)}{f_1(X_{ij}; \theta_1')}\right\} \leqslant K^C < \infty, \tag{4}$$

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

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

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

2.2. 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. $\theta_1 = (\theta_{11},\ldots,\theta_{1p})$ and the true value of the parameter is $\theta_1^0 = (\theta_{11}^0,\ldots,\theta_{1p}^0)$. 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

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

and

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

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

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

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

$$E_{\theta_1^0}\psi(X_{1j};\theta_1^0) = \int \frac{\partial}{\partial\theta_1} f_1(x;\theta_1^0)\,\mathrm{d}v(x) = 0, \tag{5}$$

so that $I(\theta_1^0)$ is in fact the covariance matrix of ψ, $I(\theta_1^0) = \mathrm{cov}_{\theta_1^0}\psi(X_{1j};\theta_1^0)$. If the second partial derivatives with respect to θ_1 can also be passed under the integral sign then $\int(\partial^2/\partial\theta_1^2)f_1(x;\theta_1^0)\,\mathrm{d}v(x) = 0$, and

$$E_{\theta_1^0}\dot{\psi}(X_{1j},\theta_1^0) = \int\left[\frac{\partial}{\partial\theta_1}\frac{(\partial/\partial\theta_1)f_1(x;\theta_1^0)}{f_1(x;\theta_1^0)}\right] f_1(x;\theta_1^0)\,\mathrm{d}v(x)$$

$$= 0 - \int \psi(x;\theta_1^0)\psi(x;\theta_1^0)^{\mathrm{t}} f_1(x;\theta_1^0)\,\mathrm{d}v(x).$$

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

To simplify notation, let

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

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

Theorem 2.4. *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* $\int f_1(x;\theta)\,\mathrm{d}v(x) = 1$;

(2) *there exist three functions* $G_1(x)$, $G_2(x)$ *and* $G_3(x)$ *such that for all* $\theta_2,\ldots,\theta_m$, $E_{\theta^0}|G_l(X_{ij})|^2 \leqslant K_l < \infty$, $l = 1,2,3$, $i = 1,\ldots,m$, *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*,

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

$k_1,k_2,k_3 = 1,\ldots,p$, *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 WL equation that is weakly consistent and*

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

Remark. (1) If there is a unique root of the WL 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 be estimated in practice. We then refer to the WL as 'adaptively weighted'. It turns out that our results on consistency and asymptotic normality are easily extended to this more general case. Assumptions 2.1–2.4 are required along with an additional condition:

Assumption 2.6. Assume:

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

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

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

where $(w_1, w_2, \ldots, w_m)^{\mathrm{t}} \stackrel{\triangle}{=} (1, 0, \ldots, 0)^{\mathrm{t}}$.

(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.

3. Examples

3.1. Restricted normal means

A simple but important example considered by van Eeden and Zidek (2001) is presented in this subsection. Casella and Strawderman (1981) consider an estimation problem of the same type. Let $X_{11}, \ldots, X_{1n_1}$ 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: $X_{21}, \ldots, X_{2n_2}$, 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, $|\theta_2 - \theta_1| \leqslant C$ for a known constant $C > 0$. In practice, finding such C will not prove difficult. But it is not unique and the smallest allowable C, $|\theta_2 - \theta_1|$ itself, is unknown. Assumptions 2.2 and 2.3 are obviously satisfied for this example. The condition (4) in Theorem 2.3 is satisfied as shown in Wang et al. (2002). If we show that Assumption 2.5 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 $n_i \bar{X}_{i.} = \sum_{j=1}^{n_i} X_{ij}$, $i = 1, \ldots, m$, $V = \mathrm{Cov}((\bar{X}_{1.}, \bar{X}_{2.})^{\mathrm{t}})$ and $B = (0, C)^{\mathrm{t}}$. It

follows that:

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

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

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

We find that

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

It follows that

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

Thus, we have

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

Finally

$$\lambda_1^* = 1 - \lambda_2^*$$

$$\lambda_2^* = \frac{1}{n_1}\left(\frac{C}{\sigma^2} + \frac{1}{n_2} + \frac{1}{n_1}\right)^{-1}.$$

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

It follows that $|\lambda_i^{(n_1)} - w_i| = \mathrm{O}(1/n_1)$, $i = 1, 2$. If we have $n_2 = \mathrm{O}(n_1^{2-\delta})$, then Assumption 2.5 will be 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. This fact is obtained in this example because we have been able to choose the weights judiciously, we do not know how such high rate can be achieved in general. Under the assumptions made in the subsection it can be shown that the conditions of Theorem 2.4 are satisfied. The maximum likelihood estimator in this example is unique for any fixed sample size. Therefore, we have

$$\sqrt{n_1}(\tilde{\theta}_1^{(n_1)} - \theta_1) \xrightarrow{\mathrm{D}} \mathrm{N}(0, \sigma^2).$$

3.2. Multivariate normal means

Let $\mathbf{X} = (\bar{X}_1, \ldots, \bar{X}_m)$, where for $i = 1, \ldots, m,\ m > 2$,

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

Assume that the θ_i are 'close' to each other. The objective is to obtain a reasonably good estimate of θ_1 using all the $\bar{X}_i$'s. 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:

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

Strawderman (2000) considers the James–Stein estimator of the parameter $\theta = (\theta_1, \ldots, \theta_m)$ for the unequal variance case $\zeta(\mathbf{X}) = (\zeta_1(\mathbf{X}), \ldots, \zeta_m(\mathbf{X}))$, where

$$\zeta_i(\mathbf{X}) = \left(1 - \frac{1}{n_i} \frac{m-2}{\sum_{j=1}^{m} \bar{X}_j^2 n_j^2}\right) \bar{X}_i.$$

The quantity,

$$1 - \frac{1}{n_i} \frac{m-2}{\sum_{j=1}^{m} \bar{X}_j^2 n_j^2},$$

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

Hu and Zidek (2001) consider simultaneous estimation of the parameters $\theta_1, \ldots, \theta_m$. They derive James–Stein type weights in their paper. Since we are combining means to yield a more reliable estimate of θ_1, it is natural to choose weights of James–Stein type since they are controlled by $n_i \sum_{j=1}^{m} \bar{X}_j^2 n_j^2$, which measures the overall similarity of the populations and possible different sample sizes. Consider the following weights:

$$\lambda_1(\mathbf{X}) = 1 - \frac{1}{n_i} \frac{m-2}{\sum_{j=1}^{m} \bar{X}_j^2 n_j^2},$$

$$\lambda_i(\mathbf{X}) = \frac{1}{m-1} \left(\frac{1}{n_i} \frac{m-2}{\sum_{j=1}^{m} \bar{X}_j^2 n_j^2}\right), \quad i = 2, 3, \ldots, m$$

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

Assume that $\lim_{n_1 \to \infty} n_1/n_k < \infty$, it follows that:

$$P_{\theta^0}\left(\left|\frac{1}{n_1} \frac{m-2}{\sum_{j=1}^{m} \bar{X}_j^2 n_j^2}\right| > \varepsilon\right) \leqslant \frac{m-2}{n_1^3} E_{\theta^0}\left|\frac{1}{\sum_{i=1}^{m} \bar{X}_i^2 (n_i/n_j)^2}\right| = \mathrm{O}\left(\frac{1}{n_1^3}\right).$$

Thus asymptotic normality of the WLE using adaptive weights will follow in this case.

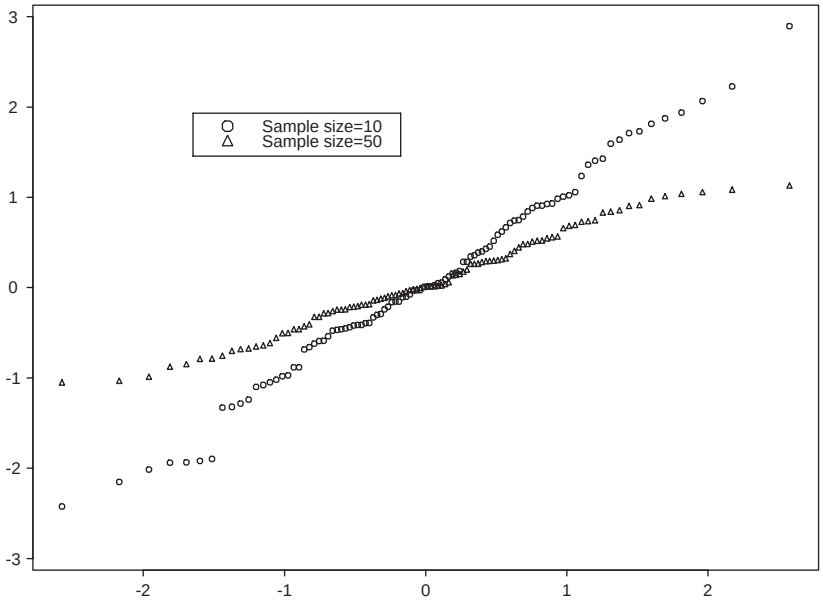

Fig. 1. Q–Q plots for sample size of 10 and 50.

We perform a simple simulation study to verify the asymptotic results. Consider three populations while the first is of primary interest. First, we set the mean of the first population to 0 and pick two values between -1 and 1 and assign them to the means of the other two populations. Then for any fixed sample size n, we generate n, $2n$ and $3n$ random variables from $N(0,1), N(\mu_2,1)$ and $N(\mu_3,1)$, respectively. We then compute the WLE. This process is repeated for 100 times. The Q–Q plots of WLE for sample size of 10 and 50 are shown in Fig. 1. It can be seen that the Q–Q plot for sample size of 50 is very close to a straight line while the Q–Q plot for sample size of 10 has very heavy tails.

4. Concluding remarks

In this paper we have shown how classical large sample theory for the maximum likelihood estimator can be extended to the adaptively WLE. 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.

Acknowledgements

We thank the associate editor and two anonymous referees for their valuable suggestions and comments.

Appendix A. Proofs of main results

Proof of Lemma 2.1. The proof can be found in Wang et al. (2002). □

Proof of Theorem 2.1. The proof can be found in Wang et al. (2002). □

Proof of Theorem 2.2. The proof of this theorem given below resembles the proof of weak consistency of the MLE in Schervish (1995, p. 239). For each $\theta_1 \neq \theta_1^0$, let $N_{\theta_1}^{(k)}, k = 1, 2, \ldots$ 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

$$\bigcap_{k=1}^{\infty} N_{\theta_1}^{(k)} = \{\theta_1\}.$$

We then have

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

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 $\varepsilon > 0$ 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, $N_{\theta_1^1}^*, N_{\theta_1^2}^*, \ldots, N_{\theta_1^p}^*$ such that $E_{\theta^0} Z_{1j}(N_{\theta_1^l}^*) > 0$, $l = 1, 2, \ldots, p$.

We then have

$$\begin{aligned} &P_{\theta^0}(\|\tilde{\theta}_1^{(n_1)} - \theta_1^0\| \geqslant \varepsilon) \\ &\quad = P_{\theta^0}(\tilde{\theta}_1^{(n_1)} \in N_{\theta_1^l}^* \text{ for some } l) \\ &\quad \leqslant \sum_{l=1}^{p} P_{\theta^0}(\tilde{\theta}_1^{(n_1)} \in N_{\theta_1^l}^*) \end{aligned}$$

$$\leqslant \sum_{l=1}^{p} P_{\theta^0}\left(\frac{1}{n_1}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\lambda_i^{(n_1)} Z_{ij}(N^*_{\theta_1^l}) \leqslant 0\right)$$

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

Since $E_{\theta^0} Z_{ij}(N^*_{\theta_1^l})^2 \leqslant E_{\theta_i}[\sup_{\theta_1\in\Theta}|\log(f_1(X_{ij};\theta_1^0)/f_1(X_{ij};\theta_1))|]^2 \leqslant K < \infty$ by Assumption 2.4, it follows from Lemma 2.1 that

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

Also,

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

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

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

This implies that

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

Thus the assertion follows. □

Proof of Theorem 2.3. Let N_0 and ε be as in the proof of Theorem 2.2, and let $N^*_{\theta_1^1}, N^*_{\theta_1^2}, \ldots, N^*_{\theta_1^p}$ be an open cover of $C \setminus N_0$ with $E_{\theta^0} Z_{1j}(N_k^*) > 0$. Then

$$P_{\theta^0}(\|\tilde{\theta}_1^{(n_1)} - \theta_1^0\| \geqslant \varepsilon)$$
$$\leqslant \sum_{k=1}^{p} P_{\theta^0}(\tilde{\theta}_1^{(n_1)} \in N^*_{\theta_1^k}) + P_{\theta^0}(\tilde{\theta}_1^{(n_1)} \in C^c \cap \Theta)$$

$$\leqslant \sum_{k=1}^{p} P_{\theta^0}(\tilde{\theta}_1^{(n_1)} \in N^*_{\theta_1^k})$$

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

It follows from the proof of Theorem 2.2 that the first term of last expression goes to zero as n goes to infinity.

By the Weak Law of Large Numbers, we have

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

Observe that

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

By the Weak Law of Large Numbers, it follows that

$$\frac{1}{n_i}\sum_{j=1}^{n_i} Z_{ij}(C^c \cap \Theta) \xrightarrow{P_{\theta^0}} E_{\theta^0}\left\{ \inf_{\theta_1 \in C^c \cap \Theta} \log \frac{f_1(X_{ij};\theta_1^0)}{f_1(X_{ij};\theta_1)} \right\}, \tag{6}$$

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 2.5, it follows that

$$\frac{n_i}{n_1}(\lambda_i^{(n_1)} - w_i) \to 0 \quad \text{as } n_1 \to \infty. \tag{7}$$

Combining Eqs. (6) and (7), we then have

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

Proof of Theorem 2.4. (1) *Existence of consistent roots.* The proof of existence of consistent roots resembles the proof in Lehmann (1983, pp. 430–432). It can be found in Wang et al. (2002).

2. Asymptotic normality. The proof in this part resembles that in Ferguson (1996, p. 121). The difference is that we need to prove the convergence of an extra term introduced by the weighted likelihood. Expand $\partial/\partial\theta_1 \log \mathrm{WL}(\mathbf{x};\theta_1)$ as

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

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

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

$$\frac{1}{\sqrt{n_1}} \log \dot{\mathrm{WL}}_{n_1}(\mathbf{x}; \theta_1^0) = B_{n_1} \sqrt{n_1}(\tilde{\theta}_1^{(n_1)} - \theta_1^0), \tag{8}$$

where

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

Note that

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

By (8), 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 $\mathrm{cov}_{\theta_1^0} \psi(X_{1j}; \theta_1^0) = I(\theta_1^0)$, we find that

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

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

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

Now we prove

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

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

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

by hypothesis (2) of this theorem.

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

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

$$B_{n_1}^{\mathrm{I}} = -\frac{1}{n_1} \int_0^1 \sum_{j=1}^{n_1} \dot{\psi}(X_{1j}; \theta_1^0 + t(\tilde{\theta}_1^{(n_1)} - \theta_1^0))\, \mathrm{d}t,$$

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

First, we prove $B_{n_1}^{\mathrm{I}} \xrightarrow{P_{\theta^0}} I(\theta_1^0)$ as $n_1 \to \infty$. The proof can be found in Wang et al. (2002).

Next we prove $B_{n_1}^{\mathrm{II}} \xrightarrow{P_{\theta^0}} (0, \ldots, 0)^{\mathrm{t}}$ as $n_1 \to \infty$.

By Lemma 2.1, every component of $B_{n_1}^{\mathrm{II}}$ goes to 0 in probability. Thus

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

This completes the proof. □

References

Casella, G., Strawderman, W.E., 1981. Estimation of bounded normal mean. Ann. Statist. 9, 870–878.

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

Ferguson, T.S., 1996. A Course in Large Sample Theory. Chapman & 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, Department of Statistics, The University of British Columbia, Vancouver, BC, 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, New York, pp. 211–235.

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

Newton, M.A., Raftery, A.E., 1994. Approximate Bayesian inference with the weighted likelihood bootstrap. J. Roy. 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: the exchangable case. Preliminary Draft

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

Strawderman, W.E., 2000. Lectures Notes of Special Lectures given at the Department of Statistics of University of British Columbia.

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

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 for publication.

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

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. Ph.D. Thesis, Department of Statistics, The University of British Columbia, Vancouver, BC, Canada.

Wang, X., van Eeden, C., Zidek, J.V. 2002. Technical report on weighted likelihood estimation and asymptotic properties of the weighted likelihood estimators. Technical Report No. 201, Department of Statistics, University of British Columbia. http://www.stat.ubc.ca/research/tr02.html.