Ann. Inst. Statist. Math.
Vol. 54, No. 4, 816–826 (2002)

MAXIMUM LIKELIHOOD ESTIMATION OF ASYMMETRIC LAPLACE PARAMETERS

SAMUEL KOTZ[1], TOMASZ J. KOZUBOWSKI[2] AND KRZYSZTOF PODGÓRSKI[3]

[1] *Department of Engineering Management and Systems Engineering, George Washington University, Washington, D.C. 20052, U.S.A.*
[2] *Department of Mathematics, University of Nevada, Reno, NV 89557-0045, U.S.A.*
[3] *Department of Mathematical Sciences, Indiana University - Purdue University Indianapolis, Indianapolis, IN 46202, U.S.A.*

(Received March 26, 2001; revised October 1, 2001)

Abstract. Maximum likelihood estimators (MLE's) are presented for the parameters of a univariate asymmetric Laplace distribution for all possible situations related to known or unknown parameters. These estimators admit explicit form in all but two cases. In these exceptions effective algorithms for computing the estimators are provided. Asymptotic distributions of the estimators are given. The asymptotic normality and consistency of the MLE's for the scale and location parameters are derived directly via representations of the relevant random variables rather than from general sufficient conditions for asymptotic normality of the MLE's.

Key words and phrases: Double exponential distribution, geometric stable law, Laplace distribution, mathematical finance, random summation, skew Laplace distribution.

1. Introduction

In the last several decades, various forms of skewed Laplace distributions have ocassionally appeared in the literature, see, e.g., McGill (1962), Holla and Bhattacharya (1968), Hinkley and Revankar (1977), Lingappaiah (1988), Balakrishnan and Ambagaspitiya (1994), Poiraud-Casanova and Thomas-Agnan (2000), Kozubowski and Podgórski (1999, 2000, 2001), Kotz *et al.* (2001).

Among these distributions, a three-parameter family with the density

$$(1.1)\qquad f_{\theta,\kappa,\sigma}(x)=\frac{\sqrt{2}}{\sigma}\frac{\kappa}{1+\kappa^2}\begin{cases}\exp\left(-\dfrac{\sqrt{2}\kappa}{\sigma}(x-\theta)\right), & \text{for}\quad x\geq\theta,\\ \exp\left(\dfrac{\sqrt{2}}{\sigma\kappa}(x-\theta)\right), & \text{for}\quad x<\theta,\end{cases}$$

obtained by means of converting a symmetric Laplace p.d.f. into a skewed one by incorporating inverse scale factors in the positive and negative orthants (see Fernández and Steel (1998)), stands out as the class of *the asymmetric Laplace* (*AL*) *distributions* (see Kozubowski and Podgórski (1999, 2000, 2001), Kotz *et al.* (2001)). These laws extend naturally all the basic properties of the symmetric Laplace distribution and have properties and features that make them attractive in applications, particularly in financial modeling, see, e.g., Madan *et al.* (1998), Levin and Tchernitser (1999), Kozubowski and Podgórski (1999, 2001). These features include infinite divisibility, finiteness of moments, allowance for asymmetry, simplicity, and natural extensions to multivariate setting (the reader is referred to Kotz *et al.* (2001) for details).

In this paper we present results on maximum likelihood estimation of the parameters of an AL distribution. Explicit formulas for the estimators are obtained for almost all cases (the exceptions being estimation of κ when the values of θ and σ are known and estimation of κ and θ when σ is known). We extend previous results obtained for the cases when the value of θ is known (Hartley and Revankar (1974), Kozubowski and Podgórski (2000)) and when all the three parameters are unknown (Hinkley and Revankar (1977)).

We summarize our results in the next section. We omit proofs, except for a derivation of the asymptotic properties for the case of estimating the location and scale parameters θ and σ, which to the best of our knowledge is established for the first time (see Section 3). A by-product of this derivation is the asymptotic normality and consistency of the MLE's of *symmetric* Laplace parameters ($\kappa = 1$). Details of derivations of other new results are available from the authors and some of them appeared in Kotz *et al.* (2001).

Since the densities of Laplace distributions have a non-differentiable peak at the mode, standard maximum likelihood asymptotic theory is not directly applicable when one of the unknown parameters is θ, and one has to use special results which account for such an irregularity. The conditions for these results are usually quite complex and we found it easier to approach the problem of estimating θ and σ directly. Unlike Hartley and Revankar (1974) and Hinkley and Revankar (1977), who utilized general sufficient conditions, we use distributional representations of the appropriate variables. This direct approach has an additional advantage—it shows the "true" reasons for the asymptotic behavior of asymmetric Laplace random variables. In this work, we present our approach in the proofs for the case of estimation of the scale σ and the location θ when κ is known.

2. Maximum likelihood estimation

Let $X_1, \ldots, X_n$ be an i.i.d. random sample from an AL distribution with the density $f_{\theta,\sigma,\kappa}$ given by (1.1), denoted by $\mathcal{AL}(\theta,\kappa,\sigma)$, $x_1, \ldots, x_n$ being their particular realization. The likelihood function is

$$L(\theta,\kappa,\sigma) = \frac{2^{n/2}}{\sigma^n}\frac{\kappa^n}{(1+\kappa^2)^n}\exp\left\{-\frac{\sqrt{2}n}{\sigma}\left(\kappa\alpha(\theta)+\frac{\beta(\theta)}{\kappa}\right)\right\}, \tag{2.1}$$

where

$$\alpha(\theta) = \frac{1}{n}\sum_{j=1}^{n}(x_j-\theta)^+, \qquad \beta(\theta) = \frac{1}{n}\sum_{j=1}^{n}(x_j-\theta)^-,$$

and $a^+ = \max(0,a)$, $a^- = (-a)^+ = -\min(0,a)$. The Fisher information matrix $I(\theta,\kappa,\sigma)$ corresponding to an $\mathcal{AL}(\theta,\kappa,\sigma)$ distribution is

$$I(\theta,\kappa,\sigma) = \left[E\left\{\frac{\partial}{\partial\gamma_i}\log f_{\theta,\kappa,\sigma}(X)\cdot\frac{\partial}{\partial\gamma_j}\log f_{\theta,\kappa,\sigma}(X)\right\}\right]_{i,j=1,2,3},$$

where X has an $\mathcal{AL}(\theta,\kappa,\sigma)$ distribution with the vector-parameter $\gamma = (\gamma_1,\gamma_2,\gamma_3) \equiv (\theta,\kappa,\sigma)$ and the density $f_{\theta,\kappa,\sigma}$. The Fisher information matrix for the vector of independent identically distributed random variables each with the distribution $\mathcal{AL}(\theta,\kappa,\sigma)$

is given by $nI(\theta,\kappa,\sigma)$. Routine calculations yield

$$I(\theta,\kappa,\sigma)=\begin{bmatrix} \frac{2}{\sigma^2} & -\frac{\sqrt{2}}{\sigma}\frac{2}{1+\kappa^2} & 0 \\ -\frac{\sqrt{2}}{\sigma}\frac{2}{1+\kappa^2} & \frac{1}{\kappa^2}+\frac{4}{(1+\kappa^2)^2} & -\frac{1}{\sigma\kappa}\frac{1-\kappa^2}{1+\kappa^2} \\ 0 & -\frac{1}{\sigma\kappa}\frac{1-\kappa^2}{1+\kappa^2} & \frac{1}{\sigma^2} \end{bmatrix}. \tag{2.2}$$

We summarize our investigations of the seven cases in Tables 1 and 2. Except for the asymptotic properties in Case 4 (κ known) that is discussed in the next section, detailed derivations of the MLE's and their asymptotic properties can be found in Hartley and Revankar (1974), Hinkley and Revankar (1977), Kozubowski and Podgórski (2000), and Kotz *et al.* (2001).

Table 1. Maximum likelihood estimation for asymmetric Laplace distributions—a summary of the results for Cases 1 through 5. All asymptotic distributions are normal with mean zero.

Case	**Parameters**	**Estimators**	**Asympt. variance**
1	θ unknown, (σ, κ known)	$\hat{\theta}_n = X_{j(n):n}$, where $j(n) = [[\frac{n\kappa^2}{1+\kappa^2}]] + 1$ ($[[x]]$ is the integer part of x).	$\sigma^2/2$
2	σ unknown, (θ, κ known)	$\hat{\sigma}_n = \sqrt{2}\kappa\alpha(\theta) + \frac{\sqrt{2}}{\kappa}\beta(\theta)$	σ^2
3	κ unknown, (σ, θ known)	$\hat{\kappa}_n$ is the unique solution to: $1-2\kappa^2/(1+\kappa^2)+\frac{\sqrt{2}}{\sigma}[\beta(\theta)/\kappa-\alpha(\theta)\kappa]=0$, $\alpha(\theta)=\frac{1}{n}\sum_{j=1}^{n}(x_j-\theta)^+$, $\beta(\theta)=\frac{1}{n}\sum_{j=1}^{n}(x_j-\theta)^-$.	$\frac{\kappa^2(1+\kappa^2)^2}{(1+\kappa^2)^2+4\kappa^2}$
4	θ, σ unknown, (κ known)	$\hat{\theta}_n = X_{j(n):n}$, where $j(n)$ as in Case 1, $\hat{\sigma}_n = \sqrt{2}\kappa\alpha(\hat{\theta}_n) + \frac{\sqrt{2}}{\kappa}\beta(\hat{\theta}_n)$	$\mathbf{\Sigma} = \begin{bmatrix} \frac{\sigma^2}{2} & 0 \\ 0 & \sigma^2 \end{bmatrix}$
5	κ, σ unknown, (θ known)	$\hat{\kappa}_n = \sqrt[4]{\frac{\beta(\theta)}{\alpha(\theta)}}$, $\hat{\sigma}_n = \sqrt{2}\sqrt[4]{\alpha(\theta)}\sqrt[4]{\beta(\theta)}(\sqrt{\alpha(\theta)}+\sqrt{\beta(\theta)})$	$\mathbf{\Sigma} = \frac{\sigma^2}{8}(1+\kappa^2)^2\begin{bmatrix} a & c \\ & b \end{bmatrix}$, $a = 1/\sigma^2, c = \frac{1}{\kappa\sigma}\frac{1-\kappa^2}{1+\kappa^2}$ $b = \frac{1}{\kappa^2}\left(1+\frac{4\kappa^2}{(1+\kappa^2)^2}\right)$

Table 2. Continuation of Table 1—cases 6 and 7.

Case	Parameters	Estimators	Asympt. variance
6	θ, κ unknown, (σ known)	**Step 1**: For $i = 1, 2, \dots, n$, solve $1 - \frac{2\kappa^2}{1+\kappa^2} + \frac{\sqrt{2}}{\sigma}[\beta(x_{i:n})/\kappa - \alpha(x_{i:n})\kappa] = 0$, obtaining a unique solution κ_i^0. **Step 2**: Set $\kappa_1 = \begin{cases} \kappa_1^0 & \text{if } \kappa_1^0 \geq 1/(n-1), \\ \frac{1}{n-1} & \text{otherwise,} \end{cases}$ $\kappa_n = \begin{cases} \kappa_n^0 & \text{if } \kappa_n^0 \geq n-1, \\ n-1 & \text{otherwise,} \end{cases}$ and for $i = 2, 3, \dots, n-1$, $\kappa_i = \begin{cases} \frac{i-1}{n-(i-1)} & \text{if } \kappa_i^0 < \frac{i-1}{n-(i-1)}, \\ \kappa_i^0 & \text{if } \frac{i-1}{n-(i-1)} \leq \kappa_i^0 < \frac{i}{n-i}, \\ \frac{i}{n-i} & \text{if } \kappa_i^0 \geq \frac{i}{n-i}. \end{cases}$ **Step 3**: $\hat{\theta}_n$, $\hat{\kappa}_n$ is the pair $x_{i:n}$ and κ_i that maximizes the expression $\log \frac{\kappa_i}{1+\kappa_i^2} + \frac{\sqrt{2}}{\sigma}[\beta(x_{i:n})/\kappa_i - \alpha(x_{i:n})\kappa_i]$.	$\mathbf{\Sigma} = \frac{\sigma^2\kappa^2}{1+\kappa^2}\begin{bmatrix} a & c \\ & b \end{bmatrix}$, $a = \frac{1}{2}\frac{1+\kappa^2}{\kappa^2} + \frac{2}{1+\kappa^2}$, $b = \frac{1+\kappa^2}{\sigma^2}, c = \frac{\sqrt{2}}{\sigma}$
7	θ, κ, σ unknown,	**Step 1**: Find $1 \leq r \leq n$ such that $h(x_{r:n}) \leq h(x_{j:n})$ for $j = 1, 2, \dots, n$, where $h(\theta) = 2\log(\sqrt{\alpha(\theta)} + \sqrt{\beta(\theta)}) + \sqrt{\alpha(\theta)}\sqrt{\beta(\theta)}$. **Step 2**: Set $\hat{\theta}_n = X_{r:n}$, $\hat{\kappa}_n = \sqrt[4]{\beta(\hat{\theta}_n)}/\sqrt[4]{\alpha(\hat{\theta}_n)}$, $\hat{\sigma}_n = \sqrt{2}\sqrt[4]{\alpha(\hat{\theta}_n)}\sqrt[4]{\beta(\hat{\theta}_n)}(\sqrt{\alpha(\hat{\theta}_n)} + \sqrt{\beta(\hat{\theta}_n)})$ where $\alpha(\hat{\theta}_n) = \frac{1}{n}\sum_{j=1}^n (x_j - \hat{\theta}_n)^+$ $\beta(\hat{\theta}_n) = \frac{1}{n}\sum_{j=1}^n (x_j - \hat{\theta}_n)^-$.	$\mathbf{\Sigma} = \frac{\sigma^2}{4}\begin{bmatrix} a & b & c \\ & d & e \\ & & f \end{bmatrix}$, $a = 4, b = \frac{\sqrt{2}}{\sigma}(1+\kappa^2)$, $c = \frac{\sqrt{2}}{\kappa}(1-\kappa^2)$, $d = \frac{(1+\kappa^2)^2}{\sigma^2}$, $e = \frac{1-\kappa^4}{\kappa\sigma}, f = \frac{(1+\kappa^2)^2}{\kappa^2}$

3. Estimation when κ is known

We shall demonstrate that the MLE $(\hat{\theta}_n, \hat{\sigma}_n)$ of θ and σ when κ is known (see Case 4, Table 1) is consistent, asymptotically normal, and efficient, with the asymptotic covariance matrix

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma^2/2 & 0 \\ 0 & \sigma^2 \end{bmatrix}. \tag{3.1}$$

(cf. (2.2)). Observe that both estimators are linear combinations of order statistics. Indeed, $\hat{\theta}_n = X_{j(n):n}$ while $\hat{\sigma}_n$ can be expressed as:

$$\begin{aligned}\hat{\sigma}_n &= \frac{\sqrt{2}}{n}\left\{\kappa \sum_{i=j(n)+1}^{n} (x_{i:n} - x_{j(n):n}) - \frac{1}{\kappa}\sum_{i=1}^{j(n)-1} (x_{i:n} - x_{j(n):n})\right\} \\ &= \frac{\sqrt{2}}{n}\left\{\kappa \sum_{i=j(n)+1}^{n} x_{i:n} - \frac{1}{\kappa}\sum_{i=1}^{j(n)-1} x_{i:n} + \left[\frac{1}{\kappa}(j(n)-1) - \kappa\bar{j}(n)\right] x_{j(n):n}\right\},\end{aligned}$$

where $\bar{j}(n) = n - j(n)$.

Our main result is stated as follows.

PROPOSITION 3.1. *Let $(X_1, \ldots, X_n)$ be i.i.d. $\mathcal{AL}(\theta, \kappa, \sigma)$ random variables with κ assumed to be known. Then, the MLE $(\hat{\theta}_n, \hat{\sigma}_n)$ given in Table* 1 (*Case* 4) *is consistent, asymptotically normal and efficient, with the asymptotic covariance matrix given by* (3.1).

For its proof we use two distributional representations, one for an asymmetric Laplace distribution and the other one for the MLE estimators of the scale and the location.

First, it was shown in Kozubowski and Podgórski (2000) that an $\mathcal{AL}(\theta, \kappa, \sigma)$ random variable X admits the representation

$$X \overset{d}{=} \theta + \frac{\sigma}{\sqrt{2}}\left(\frac{1}{\kappa}\delta_a W_a - \kappa\delta_b W_b\right), \tag{3.2}$$

where W_a and W_b are i.i.d. exponential variables with mean one, the zero-one r.v.'s δ_a and δ_b, $\delta_a + \delta_b = 1$, assume one with probabilities $1/(1+\kappa^2)$ and $\kappa^2/(1+\kappa^2)$, respectively, and are independent of W_a and W_b.

The second representation is stated and proven below.

LEMMA 3.1. *Let (W_i) and (Y_i) be two independent sequences of i.i.d. random variables having an exponential distribution with mean one, and let B_n be a Bernoulli random variable with parameters n and $p = 1/(1+\kappa^2)$, independent of these sequences. Then, the following joint distributional representation of $(\hat{\theta}_n, \hat{\sigma}_n)$ holds:*

$$\hat{\theta}_n \overset{d}{=} \frac{\sigma}{\sqrt{2}}\begin{cases} -\kappa \displaystyle\sum_{i=j(n)}^{\bar{B}_n} \frac{Y_i}{i} & \textit{if} \quad j(n) \le \bar{B}_n, \\ \dfrac{1}{\kappa} \displaystyle\sum_{i=\bar{j}(n)+1}^{B_n} \frac{Y_i}{i} & \textit{if} \quad \bar{j}(n) \le B_n - 1. \end{cases}$$

$$\hat{\sigma}_n \overset{d}{=} \frac{\sigma}{n}\begin{cases} \displaystyle\sum_{j=1}^{B_n} W_j + \sum_{i=1}^{j(n)-1} Y_i + \sum_{i=j(n)}^{\bar{B}_n} \kappa^2\left(\frac{n}{i} - 1\right) Y_i & \textit{if} \quad j(n) \le \bar{B}_n, \\ \displaystyle\sum_{j=1}^{\bar{B}_n} W_j + \sum_{i=1}^{\bar{j}(n)} Y_i + \sum_{i=\bar{j}(n)+1}^{B_n} \frac{1}{\kappa^2}\left(\frac{n}{i} - 1\right) Y_i & \textit{if} \quad \bar{j}(n) \le B_n - 1, \end{cases}$$

where $\bar{B}_n = n - B_n$ and $\overset{d}{=}$ stands for the equality of distributions.

PROOF. Without loss of generality we can assume that $\theta = 0$. It follows from (3.2) that the order statistics of $(X_1, \dots, X_n)$ can be written as follows

$$(X_{1:n}, \dots, X_{n:n}) \overset{d}{=} \frac{\sigma}{\sqrt{2}} \left(-\kappa \bar{W}_{\bar{B}_n:\bar{B}_n}, \dots, -\kappa \bar{W}_{1:\bar{B}_n}, \frac{1}{\kappa} W_{1:B_n}, \dots, \frac{1}{\kappa} W_{B_n:B_n} \right),$$

where $(\bar{W}_i)$ is another sequence of i.i.d. exponential variables with mean one, independent of (W_i). Thus, we have the following representations for $\hat{\theta}_n$ and $\hat{\sigma}_n$:

$$\hat{\theta}_n \overset{d}{=} \frac{\sigma}{\sqrt{2}} \begin{cases} -\kappa \bar{W}_{\bar{B}_n - j(n)+1:\bar{B}_n} & \text{if} \quad j(n) \le \bar{B}_n, \\ \frac{1}{\kappa} W_{B_n - \bar{j}(n):\bar{B}_n} & \text{if} \quad \bar{j}(n) \le B_n - 1. \end{cases}$$

$$\hat{\sigma}_n \overset{d}{=} \frac{\sigma}{n} \begin{cases} \displaystyle\sum_{j=1}^{B_n} W_{j:B_n} + \sum_{i=\bar{B}_n - j(n)+2}^{\bar{B}_n} \bar{W}_{i:\bar{B}_n} - \kappa^2 \sum_{i=1}^{\bar{B}_n - j(n)} \bar{W}_{i:\bar{B}_n} & \\ \qquad - [j(n) - 1 - \kappa^2 \bar{j}(n)] \bar{W}_{\bar{B}_n - j(n)+1:\bar{B}_n} & \text{if} \quad j(n) \le \bar{B}_n, \\ \displaystyle\sum_{j=1}^{\bar{B}_n} \bar{W}_{j:B_n} + \sum_{i=B_n - \bar{j}(n)+1}^{B_n} W_{i:B_n} - \frac{1}{\kappa^2} \sum_{i=1}^{B_n - \bar{j}(n)-1} W_{i:B_n} & \\ \qquad + \left[\frac{1}{\kappa^2} (j(n) - 1) - \bar{j}(n) \right] W_{B_n - \bar{j}(n):B_n} & \text{if} \quad \bar{j}(n) \le B_n - 1. \end{cases}$$

The well-known representation of the order statistics of exponential random variables (see, e.g., Balakrishnan and Cohen (1991)) stating that

$$W_{k:m} \overset{d}{=} \sum_{l=1}^{k} \frac{Y_l}{m - l + 1}$$

leads us (after some algebra) to

$$\hat{\theta}_n \overset{d}{=} \frac{\sigma}{\sqrt{2}} \begin{cases} -\kappa \displaystyle\sum_{l=1}^{\bar{B}_n - j(n)+1} \frac{\bar{Y}_l}{\bar{B}_n - l + 1} & \text{if} \quad j(n) \le \bar{B}_n, \\ \displaystyle\frac{1}{\kappa} \sum_{l=1}^{B_n - \bar{j}(n)} \frac{Y_l}{B_n - l + 1} & \text{if} \quad \bar{j}(n) \le B_n - 1. \end{cases}$$

$$\hat{\sigma}_n \overset{d}{=} \frac{\sigma}{n} \begin{cases} \displaystyle\sum_{j=1}^{B_n} W_j + \sum_{l=\bar{B}_n - j(n)+2}^{\bar{B}_n} \bar{Y}_l + \kappa^2 \sum_{l=1}^{\bar{B}_n - j(n)+1} \left(\frac{n}{\bar{B}_n - l + 1} - 1 \right) \bar{Y}_l & \\ & \text{if} \quad j(n) \le \bar{B}_n, \\ \displaystyle\sum_{j=1}^{\bar{B}_n} \bar{W}_j + \sum_{l=B_n - j(n)+1}^{B_n} Y_l + \frac{1}{\kappa^2} \sum_{l=1}^{B_n - \bar{j}(n)} \left(\frac{n}{B_n - l + 1} - 1 \right) Y_l & \\ & \text{if} \quad \bar{j}(n) \le B_n - 1, \end{cases}$$

where $(\bar{Y}_i)$ is a sequence of i.i.d. exponential variables with mean one, independent of (Y_i) and (W_i). Note that by independence of B_n from all involved sequences of exponential random variables, both $(Y_{\bar{B}_n-i+1})_{i=1}^{B_n}$ and $(\bar{Y}_{B_n-i+1})_{i=1}^{B_n}$ have the same distributions as $(Y_i)_{i=1}^{B_n}$ and $(\bar{Y}_i)_{i=1}^{B_n}$, and we can interchange them in the above representation.

The result now follows from substituting $(\bar{Y}_i)$ by (Y_i) and $(\bar{W}_i)$ by (W_i), which is legitimate since B_n is assumed to be independent of these sequences. □

We now turn to the proof of the main result.

Proof of Propostion 3.1. Without loss of generality, we can assume that $\theta = 0$ and $\sigma = 1$. Let us consider the representation of $\hat{\theta}_n$ and $\hat{\sigma}_n$ given in Lemma 3.1.

It follows from the central limit theorem and the Skorokhod representation theorem that there exists a version of B_n such that $(B_n - np)/\sqrt{npq}$, where $q = 1-p$, is convergent almost surely to a standard normal random variable Z, which is independent of the sequences (Y_i) and (W_i). Note the following relations:

$$\lim_{n\to\infty} \frac{B_n - np}{\sqrt{npq}} \overset{\text{a.s.}}{=} Z, \lim_{n\to\infty} \frac{\bar{B}_n - nq}{\sqrt{npq}} \overset{\text{a.s.}}{=} -Z,$$
$$\lim_{n\to\infty} \frac{B_n - \bar{j}(n)}{\sqrt{npq}} \overset{\text{a.s.}}{=} Z, \lim_{n\to\infty} \frac{\bar{B}_n - j(n)}{\sqrt{npq}} \overset{\text{a.s.}}{=} -Z,$$

where the last two follow from the fact that $j(n)/n - q = O(1/n)$. We need to prove the convergence in distribution:

$$\sqrt{n}\left(\begin{bmatrix} \hat{\theta}_n \\ \hat{\sigma}_n \end{bmatrix} - \begin{bmatrix} 0 \\ 1 \end{bmatrix}\right) \to N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & 1 \end{bmatrix}\right), \quad \text{as} \quad n \to \infty. \tag{3.3}$$

First, we shall show that with probability one,

$$\lim_{n\to 0} \sqrt{n}[E(\hat{\sigma}_n \mid B_n) - 1] = 0. \tag{3.4}$$

To prove this fact, let us compute conditional expectation $\hat{\sigma}_n$ given B_n:

$$E(\hat{\sigma}_n \mid B_n) = \begin{cases} \dfrac{B_n + j(n) - 1}{n} - \kappa^2 \dfrac{\bar{B}_n - j(n) + 1}{n} + \kappa^2 \displaystyle\sum_{k=j(n)}^{\bar{B}_n} \frac{1}{k} & \text{if} \quad j(n) \le \bar{B}_n, \\[2ex] \dfrac{\bar{B}_n + \bar{j}(n)}{n} - \dfrac{B_n - \bar{j}(n)}{n\kappa^2} + \dfrac{1}{\kappa^2} \displaystyle\sum_{k=\bar{j}(n)+1}^{B_n} \frac{1}{k} & \text{if} \quad \bar{j}(n) \le B_n - 1. \end{cases}$$

Now, we have

$$\frac{\bar{B}_n - j(n)}{\bar{B}_n} \le \sum_{k=j(n)}^{\bar{B}_n} \frac{1}{k} \le \frac{\bar{B}_n - j(n) + 1}{j(n)} \tag{3.5}$$

and

$$\frac{B_n - \bar{j}(n)}{B_n} \le \sum_{k=\bar{j}(n)+1}^{B_n} \frac{1}{k} \le \frac{B_n - \bar{j}(n)}{\bar{j}(n)}. \tag{3.6}$$

Consider two disjoint events: $E_1 = \{\lim_{n\to\infty}(\bar{B}_n - j(n))/\sqrt{npq} = -Z > 0\}$ and $E_2 = \{\lim_{n\to\infty}(B_n - \bar{j}(n))/\sqrt{npq} = Z > 0\}$, each having probability equal to 1/2.

On E_1, we shall eventually have $\bar{B}_n \geq j(n)$. Thus, using the inequalities (3.5), we obtain

$$\sqrt{pq}\frac{n}{\bar{B}_n}\frac{\bar{B}_n - j(n)}{\sqrt{npq}} \leq \sqrt{n}\sum_{k=j(n)}^{\bar{B}_n}\frac{1}{k} \leq \sqrt{pq}\frac{n}{j(n)}\frac{\bar{B}_n - j(n)+1}{\sqrt{npq}}.$$

The quantities on the left and the right hand sides of the above inequalities converge almost surely to $-Z\sqrt{p/q}$. Thus, almost surely on E_1, we have

$$\lim_{n\to\infty}\sqrt{n}\sum_{k=j(n)}^{\bar{B}_n}\frac{1}{k} = -Z\sqrt{p/q}. \tag{3.7}$$

Similar arguments applied to E_2 leads to the almost sure limit

$$\lim_{n\to\infty}\sqrt{n}\sum_{k=\bar{j}(n)+1}^{B_n}\frac{1}{k} = Z\sqrt{q/p}. \tag{3.8}$$

We also get almost surely on E_1:

$$\lim_{n\to\infty}\sqrt{n}\left(\frac{B_n + j(n) - 1}{n} - 1\right) = \sqrt{pq}Z$$
$$\lim_{n\to\infty}\sqrt{n}\frac{\bar{B}_n - j(n)+1}{n} = -\sqrt{pq}Z.$$

These relations and (3.7) lead to the almost sure limit on E_1:

$$\lim_{n\to\infty}\sqrt{n}[E(\hat{\sigma}_n \mid B_n) - 1] = \sqrt{pq}Z + \kappa^2\sqrt{pq}Z - \kappa^2\sqrt{p/q}Z = 0.$$

Using the equality (3.8), an analogous argument produces the convergence on E_2, proving (3.4).

Now, it is sufficient to show that we have the following convergence in distribution

$$\sqrt{n}\left(\begin{bmatrix}\hat{\theta}_n \\ \hat{\sigma}_n - E(\hat{\sigma}_n \mid B_n)\end{bmatrix}\right) \xrightarrow{d} N\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}\frac{1}{2} & 0\\ 0 & 1\end{bmatrix}\right).$$

Consider the expression on the left-hand side conditionally on B_n. Assuming first that we are in the event E_2, ultimately for large n we shall have

$$\hat{\theta}_n = \frac{1}{\sqrt{2}}\frac{1}{\kappa}\sum_{i=\bar{j}(n)+1}^{B_n}\frac{Y_i}{i},$$

and thus the following inequalities hold

$$\frac{1}{\sqrt{2}\kappa}\frac{\sqrt{n}}{B_n}\sum_{i=\bar{j}(n)+1}^{B_n} Y_i \leq \sqrt{n}\hat{\theta}_n \leq \frac{1}{\sqrt{2}\kappa}\frac{\sqrt{n}}{\bar{j}(n)}\sum_{i=\bar{j}(n)+1}^{B_n} Y_i. \tag{3.9}$$

The left-hand side is equal in distribution (conditionally on B_n) to

$$\frac{1}{\sqrt{2}\kappa}\frac{\sqrt{n}}{B_n}\sum_{i=1}^{B_n-\bar{j}(n)} Y_i = \frac{\sqrt{pq}}{\sqrt{2}\kappa}\frac{n}{B_n}\frac{B_n-\bar{j}(n)}{\sqrt{npq}}\frac{\sum_{i=1}^{B_n-\bar{j}(n)} Y_i}{B_n-\bar{j}(n)}.$$

The right-hand side of (3.9) is equal to

$$\frac{\sqrt{pq}}{\sqrt{2}\kappa}\frac{n}{\bar{j}(n)}\frac{B_n-\bar{j}(n)}{\sqrt{npq}}\frac{\sum_{i=1}^{B_n-\bar{j}(n)} Y_i}{B_n-\bar{j}(n)}.$$

By the Law of Large Numbers applied to the sample mean of (Y_i) and by the independence of the latter of B_n, we conclude that both sides of (3.9) converge in probability to $Z/\sqrt{2}$.

Similarly, we obtain the convergence to $Z/\sqrt{2}$ on the set E_1. Consequently, conditionally on B_n, we have

$$\lim_{n\to\infty} \sqrt{n}\hat{\theta}_n = Z/\sqrt{2},$$

where convergence is in probability.

Next, consider $\sqrt{n}\,[\hat{\sigma}_n - E(\hat{\sigma}_n \mid B_n)]$ also conditionally on B_n. Assume first that we are in the set E_1. Then, the random variable under consideration can be written as the sum of three independent random variables as follows:

$$\sqrt{n}\,[\hat{\sigma}_n - E(\hat{\sigma}_n \mid B_n)] = L_n + M_n + N_n,$$

where

$$L_n = \frac{\sum_{j=1}^{B_n}(W_j-1)}{\sqrt{n}}, \qquad M_n = \frac{\sum_{k=1}^{j(n)-1}(Y_k-1)}{\sqrt{n}},$$
$$N_n = \kappa^2\frac{\sum_{k=j(n)}^{\bar{B}_n}(n/k-1)(Y_k-1)}{\sqrt{n}}.$$

Note that L_n and M_n are independent on $\hat{\theta}_n$. Thus, $\hat{\sigma}_n - E(\hat{\sigma}_n \mid B_n)$ is dependent on $\hat{\theta}_n$ only through N_n. But the latter variable converges to zero with probability one. Indeed, we have the inequalities:

$$N_n \le \kappa^2\sqrt{pq}\frac{\bar{B}_n - j(n)+1}{\sqrt{npq}}\left[\left(\frac{n}{j(n)}-1\right)\left(\frac{\sum_{k=j(n)}^{\bar{B}_n} Y_k}{\bar{B}_n - j(n)+1}-1\right)\right], \tag{3.10}$$

$$N_n \ge \kappa^2\sqrt{pq}\frac{\bar{B}_n - j(n)+1}{\sqrt{npq}}\left[\left(\frac{n}{\bar{B}_n}-1\right)\left(\frac{\sum_{k=j(n)}^{\bar{B}_n} Y_k}{\bar{B}_n - j(n)+1}-1\right)\right]. \tag{3.11}$$

Since

$$\frac{\sum_{k=j(n)}^{\bar{B}_n} Y_k}{\bar{B}_n - j(n)+1} \overset{d}{=} \frac{\sum_{k=1}^{\bar{B}_n-j(n)+1} Y_k}{\bar{B}_n - j(n)+1}$$

and the right-hand side converges almost surely to 1 on E_1 by the law of large numbers, we conclude that the left-hand side converges in probability to 1 as well. From this and the fact that $j(n)/n \to q$, we obtain after computing all limits in the right hand sides of (3.10) and (3.11) and some elementary algebra that N_n converges in probability to zero.

It remains to compute the distributional limits of L_n and M_n. We have

$$L_n = \frac{\sqrt{B_n}}{\sqrt{n}} \frac{\sum_{j=1}^{B_n}(W_j - 1)}{\sqrt{B_n}},$$
$$M_n = \frac{\sqrt{j(n)-1}}{\sqrt{n}} \frac{\sum_{j=1}^{j(n)-1}(Y_j - 1)}{\sqrt{j(n)-1}}.$$

Since B_n/n converges to p and $j(n)/n$ converges to q, it follows from the central limit theorem that $\lim_{n\to\infty}(L_n+M_n) \overset{d}{=} \sqrt{q}Z_1+\sqrt{p}Z_2$, where Z_1 and Z_2 are independent standard normal variables. Consequently, conditionally on B_n, the sequence $\sqrt{n}\,[\hat{\sigma}_n - E(\hat{\sigma}_n \mid B_n)]$ converges in distribution to a standard normal variable and its distribution is independent of Z. The same arguments apply to the event E_2. Thus, unconditionally, the asymptotic distribution of $\sqrt{n}\hat{\theta}_n$ is independent of that of $\sqrt{n}\,[\hat{\sigma}_n - E(\hat{\sigma}_n \mid B_n)]$.

Consequently, $\sqrt{n}(\hat{\theta}_n, \hat{\sigma}_n - 1) \to (Z/\sqrt{2}, \sqrt{q}Z_1 + \sqrt{p}Z_2)$, where Z, Z_1, and Z_2 are independent standard normal random variables. The above is equivalent to (3.3). □

In the special case $\kappa = 1$, Proposition 3.1 leads to the MLE's (and their asymptotics) of the parameters of the symmetric Laplace distribution, obtained in Kotz *et al.* (2001).

Acknowledgements

The authors wish to thank the referees for their careful reading and valuable comments.

References

Balakrishnan, N. and Ambagaspitiya, R. S. (1994). On skewed-Laplace distributions, Report, McMaster University, Hamilton, Ontario, Canada.

Balakrishnan, N. and Cohen, A. C. (1991). *Order Statistics and Inference: Estimation Methods*, Academic Press, San Diego.

Fernändez, C. and Steel, M. F. J. (1998). On Bayesian modeling of fat tails and skewness, *J. Amer. Statist. Assoc.*, **93**, 359–371.

Hartley, M. J and Revankar, N. S. (1974). On the estimation of the Pareto law from underreported data, *J. Econometrics*, **2**, 327–341.

Hinkley, D. V. and Revankar, N. S. (1977). Estimation of the Pareto law from underreported data, *J. of Econometrics*, **5**, 1–11.

Holla, M. S. and Bhattacharya, S. K. (1968). On a compound Gaussian distribution, *Ann. Inst. Statist. Math.*, **20**, 331–336.

Kotz, S., Kozubowski, T. J. and Podgórski, K. (2001). *The Laplace Distribution and Generalizations: A Revisit with Applications to Communications, Economics, Engineering, and Finance*, Birkhäuser, Boston.

Kozubowski, T. J. and Podgórski, K. (1999). A class of asymmetric distributions, *Actuarial Research Clearing House*, **1**, 113–134.

Kozubowski, T. J. and Podgórski, K. (2000). Asymmetric Laplace distribution, *Math. Sci.*, **25**, 37–46.

Kozubowski, T. J. and Podgórski, K. (2001). Asymmetric Laplace laws and modeling financial data, *Math. Comput. Modelling*, **34**, 1003–1021.

Levin, A. and Tchernitser, A. (1999). Multifactor gamma stochastic variance Value-at-Risk model, Presentation at the Conference *Applications of Heavy Tailed Distributions in Economics, Engineering, and Statistics*, American University, Washington, D.C., June 3–5, 1999.

Lingappaiah, G. S. (1988). On two-piece double exponential distribution, *J. Korean Statist. Soc.*, **17**(1), 46–55.

Madan, D. B., Carr, P. and Chang, E. C. (1998). The variance gamma process and option pricing, *European Finance Review*, **2**, 74–105.

McGill, W. J. (1962). Random fluctuations of response rate, *Psychometrika*, **27**, 3–17.

Poiraud-Casanova, S. and Thomas-Agnan, C. (2000). About monotone regression quantiles, *Statist. Probab. Lett.*, **48**, 101–104.

Ann. Inst. Statist. Math.
Vol. 54, No. 4, 827–839 (2002)

EFFICIENT NON-ITERATIVE AND NONPARAMETRIC ESTIMATION OF HETEROGENEITY VARIANCE FOR THE STANDARDIZED MORTALITY RATIO

DANKMAR BÖHNING[1], UWE MALZAHN[1], JESUS SAROL, JR.[2], SASIVIMOL RATTANASIRI[3] AND ANNIBALE BIGGERI[4]

[1] *Department of Epidemiology, Free University Berlin, Haus 562, Fabeckstr. 60-62, 14195 Berlin, Germany,* e-mail: boehning@zedat.fu-berlin.de
[2] *Department of Epidemiology and Biostatistics, College of Public Health, University of the Philippines Manila, Manila, Philippines,* e-mail: jsarol@nwave.net
[3] *Department of Biostatistics, Faculty of Public Health, Mahidol University, Bangkok, Thailand,* e-mail: r_sasivimol@hotmail.com
[4] *Department of Statistics, University of Florence, 50134 Florence, Italy,* e-mail: abiggeri@stat.ds.unifi.it

(Received December 5, 2000; revised August 6, 2001)

Abstract. In this paper the situation of extra population heterogeneity in the standardized mortality ratio is discussed from the point-of-view of an analysis of variance. First, some simple non-iterative ways are provided to estimate the variance of the heterogeneity distribution without estimating the heterogeneity distribution itself. Next, a wider class of linear unbiased estimators is introduced and their properties investigated. Consistency is shown for a wide sub-class of estimators charactererized by the fact that the associated linear weights are within some positive, finite bounds. Furthermore, it is shown that an efficient estimator is often provided when the weights are proportional to the expected counts.

Key words and phrases: Population heterogeneity, random effects model, moment estimator, variance separation, standardized mortality ratio.

1. Introduction

In a variety of biometric applications the situation of extra-population heterogeneity occurs. This is particularly the case if a good reason exists to model the variable of interest Y through a density of parametric form $p(y \mid \theta)$ with a scalar parameter θ. For a given subpopulation, the density $p(y \mid \theta)$ might be most suitable, but the value of θ cannot cover the whole population of interest. In such situations we speak of extra heterogeneity, which might be caused by unobserved covariates or clustered observations, such as herd clustering when estimating animal infection rates. An introductory discussion can be found in Aitkin *et al.* ((1990), p. 213) and the references given there; see also the review of Pendergast *et al.* ((1996), p. 106). A discussion on extra-binomial variation (i.e. extra-population heterogeneity if $p(y \mid \theta)$ is the binomial) can be found in Williams (1982) and Collet ((1991), p. 192). In this paper, it is understood that extra-population heterogeneity, or in short, population heterogeneity, refers to a situation when the parameter of interest, θ, varies in the population and sampling has not taken this into account (e.g. it has not been observed from which subpopulation (defined by the values of θ) the datum is coming from). As will be clear from equation (1.1) below, inference is affected by the occurrence of extra-population heterogeneity. For example, variances of estimators of interest are often greatly increased, leading to wider confidence intervals as compared to conventional ones. To adjust these variances the estimation of the variance of the distribution associated with the extra-heterogeneity is required. The main objective of this paper is to present a moment estimator for the heterogeneity variance in a simple manner. To be more precise, if θ varies in itself with distribution G and associated density $g(\theta)$, the (unconditional) marginal density of Y can be given as $f(y) = \int_\Theta p(y \mid \theta)g(\theta)d\theta$. Of interest is the separation of the (unconditional) variance of Y (e.g. variance of Y with respect to $f(y)$) into two terms:

$$\text{(1.1)} \qquad \mathrm{Var}(Y) = \int_\Theta \mathrm{Var}(Y \mid \theta)g(\theta)d\theta + \int_\Theta (\mu(\theta) - \mu_Y)^2 g(\theta)d\theta$$

where $\mu(\theta)$ is the $E(Y \mid \theta)$ and $\mu_Y = \int yf(y)dy$ is the marginal mean of Y. Note that $\mu_Y = E_G(\mu(\theta))$. Note that we can also write (1.1) briefly as

$$\mathrm{Var}(Y) = E_G(\sigma^2(\theta)) + \mathrm{Var}_G(\mu(\theta)).$$

In the sequel we will also denote $\mathrm{Var}_G(\mu(\theta))$ by τ_Y^2. Thus, in such instances, it can be said that (1.1) is a partitioning of the variance due to the variation in the subpopulation with parameter value θ (and then averaged over θ) and due to the variance in the heterogeneity distribution G of θ. Also, (1.1) can be taken as an analysis-of-variance partition with a latent factor with distribution G. We have to distinguish carefully between *three* distributional schemes when computing moments. For example, $\mathrm{Var}(Y)$ refers to the unconditional or marginal variance and is computed using the marginal density $f(y)$, $\mathrm{Var}(Y \mid \theta)$ is the *conditional* variance and is computed using the conditional density $p(y \mid \theta)$, and $\mathrm{Var}_G(\mu(\theta))$ refers to the distribution G of θ. The intention is to find an estimate of τ_Y^2 without implying knowledge or estimating the latent heterogeneity distribution G. The idea is very simple: we write (1.1) as

$$\text{(1.2)} \qquad \mathrm{Var}_G(\mu(\theta)) = \tau_Y^2 = \mathrm{Var}(Y) - E_G(\sigma^2(\theta))$$

and replace $\mathrm{Var}(Y)$ and $E_G(\sigma^2(\theta))$ on the right hand side of (1.2) with their respective sample estimates and obtain an estimate for τ_Y^2. In the succeeding text, we will use μ as the mean of θ and τ^2 for its variance.

Example (Poisson). Let $Y_1, Y_2, \ldots, Y_N$ be a random sample of Poisson counts, e.g. $p(y \mid \theta) = \exp(-\theta)\theta^y/y!$. Then, $\sigma^2(\theta) = \theta$, $E_G(\sigma^2(\theta)) = E_G(\theta) = \mu = E(Y)$ and $\tau_Y^2 = \tau^2$. Note that $\mathrm{Var}(Y)$ can simply be estimated by $S^2 = \frac{1}{N-1}\sum_{i=1}^N (Y_i - \bar{Y})^2$ and μ by $\bar{Y}$. Therefore, according to (1.2), an estimator of τ^2 is provided as $\hat{\tau^2} = S^2 - \bar{Y}$. This quantity has also been referred to as a measure of Poisson overdispersion (Böhning 1994). Note, that $E(\hat{\tau^2}) = \tau^2$.

Example (Binomial). Let $Y_1, Y_2, \ldots, Y_N$ be a random sample of Binomial counts, e.g. $p(y \mid \theta) = \binom{n}{y}\theta^y(1-\theta)^{(n-y)}$. Then, $\mu(\theta) = n\theta$ and $\sigma^2(\theta) = n\theta(1-\theta)$. Also, $\tau_Y^2 = n^2\tau^2$. It follows that $E_G(n\theta) = n\mu$, $E_G(\sigma^2(\theta)) = nE_G(\theta - \theta^2) = n(\mu - E_G(\theta^2)) = n(\mu - \tau^2 - \mu^2)$. Since $\mathrm{Var}(Y_i) = E_G(\sigma^2(\theta)) + \tau_Y^2 = n\mu(1-\mu) + n(n-1)\tau^2$, we find $\tau^2 = \frac{1}{n(n-1)}[\mathrm{Var}(Y_i) - n\mu(1-\mu)]$, for $i = 1, \ldots, N$. We can use the estimator, $\hat{\tau^2} = \frac{S^2}{n(n-1)} - [\frac{\bar{Y}}{n}(1 - \frac{\bar{Y}}{n}]/(n-1)$, with $S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \bar{Y})^2$. This estimator has a bias equal to $\mathrm{Var}(Y_i)/[n^2(n-1)N]$ which is practically negligible even for moderate values of n. For example, if $n = 10$ and $N = 10$, then the bias of $\hat{\tau^2}$ is equal to 1/9000 of the variance of Y_i.

The idea to construct a simple moment estimator using equation (1.2) can be found in various instances in the literature including Marshall (1991) and Martuzzi and Elliot (1996). The latter considered the case that $p(y \mid \theta)$ is the binomial. However, the way this moment estimator is constructed is not unique. In this paper, we try to develop a more general framework for these kinds of estimators.

In the next section, we will consider a generalization of this idea to the standardized mortality ratio. In Section 3, we will discuss a more general class of linear unbiased estimators of the heterogeneity variance and provide a closed form expression for its variance. This enables us to provide a closed form expression for the efficient estimator. In Section 4, we will provide simple conditions for consistency. Section 5 considers estimating simultaneously the mean and variance of the heterogneity distribution. Section 6 ends the paper with a discussion of the results.

2. The standardized mortality ratio

We consider a special but important case. Let $Y_1, Y_2, \ldots, Y_N$ be a sample of counts which can be thought of as a sequence of mortality or morbidity cases. For each Y_i there exists a connected non-random number e_i, for $i = 1, \ldots, N$, which is interpreted as an expected number of counts and usually calculated on the basis of an external reference population. With the help of these numbers one can define the standardized mortality ratio as $SMR_i = Y_i/e_i$ and its expected value $E(SMR_i \mid \theta_i) = \theta_i$, for $i = 1, \ldots, N$. Frequently, this sample is coming from N geographic regions or areas. Therefore, this situation is closely related to the so-called field of *disease mapping*. For an introduction to this field see Böhning (2000) or Lawson *et al.* (1999).

Furthermore, conditionally on the value of θ, a Poisson distribution is assumed for $Y \mid \theta$: $p(y_i \mid \theta, e_i) = \exp(-\theta e_i)(\theta e_i)^{y_i}/y_i!$. For this case, the partition of variance (1.1) takes the form

$$\begin{aligned} \mathrm{Var}(Y_i) &= E_G(\sigma_i^2(\theta)) + \mathrm{Var}_G(\mu_i(\theta)) = e_i E_G(\theta) + e_i^2 \mathrm{Var}_G(\theta) \\ &= e_i\mu + e_i^2\tau^2. \end{aligned} \tag{2.1}$$

At this point it is important to understand the consequences of the occurrence of heterogeneity. Suppose μ is estimated using the conventional estimator $\hat{\mu} = \frac{\sum_i Y_i}{\sum_i e_i}$. Then, we have that $\mathrm{Var}(\hat{\mu}) = \mu\frac{1}{\sum_i e_i} + \tau^2\frac{\sum_i e_i^2}{(\sum_i e_i)^2}$, so that, depending on the value of τ^2, its variance might be largely increased. Note also that conventional confidence intervals use the variance formula $\mathrm{Var}(\hat{\mu}) = \mu\frac{1}{\sum_i e_i}$, which might be too small if heterogeneity is present.

We write (2.1) as $E(Y_i - e_i\mu)^2 = e_i\mu + e_i^2\tau^2$ which draws attention to the variate $W_i = \frac{(Y_i - e_i\mu)^2 - e_i\mu}{e_i^2}$. Since $\mathrm{Var}(Y_i) = E(Y_i - e_i\mu)^2$ we note that it follows from (2.1)

$$E(W_i) = \tau^2. \tag{2.2}$$

First, to estimate τ^2, we can replace $\text{Var}(Y_i)$ by its 'estimate' $(Y_i - e_i\mu)^2$ and solve for τ^2 and then average over i:

$$\hat{\tau}_1^2 = \frac{1}{N}\left[\sum_{i=1}^{N}(Y_i - e_i\mu)^2/e_i^2 - \mu\sum_{i=1}^{N}\frac{1}{e_i}\right]. \tag{2.3}$$

Second, in (2.1), we can divide first by e_i and then average over i and solve for τ^2:

$$\hat{\tau}_2^2 = \frac{\sum_{i=1}^{N}(Y_i - e_i\mu)^2/e_i - \mu N}{\sum_{i=1}^{N} e_i}. \tag{2.4}$$

Third, we can also first average over i in (2.1), and then solve for τ^2:

$$\hat{\tau}_3^2 = \frac{\sum_{i=1}^{N}(Y_i - e_i\mu)^2 - \mu\sum_{i=1}^{N} e_i}{\sum_{i=1}^{N} e_i^2}. \tag{2.5}$$

Note that all three estimators are identical if the e_i's are all equal (e.g. if $e_i = e_j$ for all $i, j = 1, \dots, N$). We note in passing that all three estimators are unbiased. In fact, they are special cases of a more general class of *linear unbiased estimators* of τ^2:

$$T(W, \alpha) = \frac{\sum_{i=1}^{N} \alpha_i W_i}{\sum_{i=1}^{N} \alpha_i} \tag{2.6}$$

for any non-random, non-negative numbers $\alpha_1, \alpha_2, \dots, \alpha_N$. It is easy to verify that for $\alpha_i = 1/N$ the estimator $T(W, \alpha) = \hat{\tau}_1^2$, for $\alpha_i = e_i$ the estimator $T(W, \alpha) = \hat{\tau}_2^2$, and for $\alpha_i = e_i^2$ the estimator $T(W, \alpha) = \hat{\tau}_3^2$ is provided. The estimator $\hat{\tau}_1^2$ associated with $\alpha_i = 1/N$ is mentioned in Böhning (2000). The estimator $\hat{\tau}_2^2$ associated with $\alpha_i = e_i$ is suggested by Marshall (1991).

The estimator $T(W, \alpha)$ considered so far requires the knowledge of the overall-mean μ. This assumption is satisfied, if the SMR_is are *indirectly standardized* implying that $\sum_i Y_i / \sum_i e_i = 1$.

2.1 *Example* 1: *Hepatitis B in Berlin*

To illustrate the estimators, we consider two examples. Table 1 gives the observed and expected Hepatitis B cases in the 23 city regions of Berlin for the year 1995. Here, we find that $\sum_i Y_i / \sum_i e_i = 1.019$. A conventional χ^2-test for homogeneity is given by $\chi^2 = \sum_i (Y_i - \mu e_i)^2 / (\mu e_i)$. If μ is replaced with $\hat{\mu} = \sum_i Y_i / \sum_i e_i = 1.019$, we will get $\chi^2 = 193.52$, which clearly indicates heterogeneity. For this illustration, assuming that μ is fixed, the following values for $\hat{\tau}_j^2$ can be achieved: $0.5205 (j = 1)$, $0.4810 (j = 2)$, $0.4226 (j = 3)$. This indicates rather high heterogeneity since $\text{Var}(\widehat{SMR}) = \frac{1}{N-1}\sum_i (SMR_i - \overline{SMR})^2 = 0.6234$. The situation is illustrated in Fig. 1 (using $\hat{\tau}_1^2$ to construct the confidence interval adjusting for heterogeneity). Note that using the "right" estimate of variance leads to an increased length in confidence interval for μ using $\hat{\mu} \pm 1.96\sqrt{\text{Var}(\hat{\mu})}$ for the construction of a 95%-confidence interval where $\hat{\mu}$ corresponds to the pooled estimator.

Table 1. Observed and expected Hepatitis B cases in the 23 city regions of Berlin (1995).

Area i	Y_i	e_i	Area i	Y_i	e_i
1	29	10.7121	13	15	8.3969
2	26	17.9929	143	11	15.6438
3	54	18.1699	15	11	11.8289
4	30	19.2110	16	2	9.9513
5	16	21.9611	17	2	10.8313
6	15	14.6268	18	9	18.3404
7	6	9.6220	19	2	5.1758
8	35	17.2671	20	3	10.9543
9	17	18.8230	21	11	20.0121
10	7	18.2705	22	5	13.8389
11	43	32.1823	23	2	12.7996
12	17	24.5929	-	-	-

Source: Berlin Census Bureau

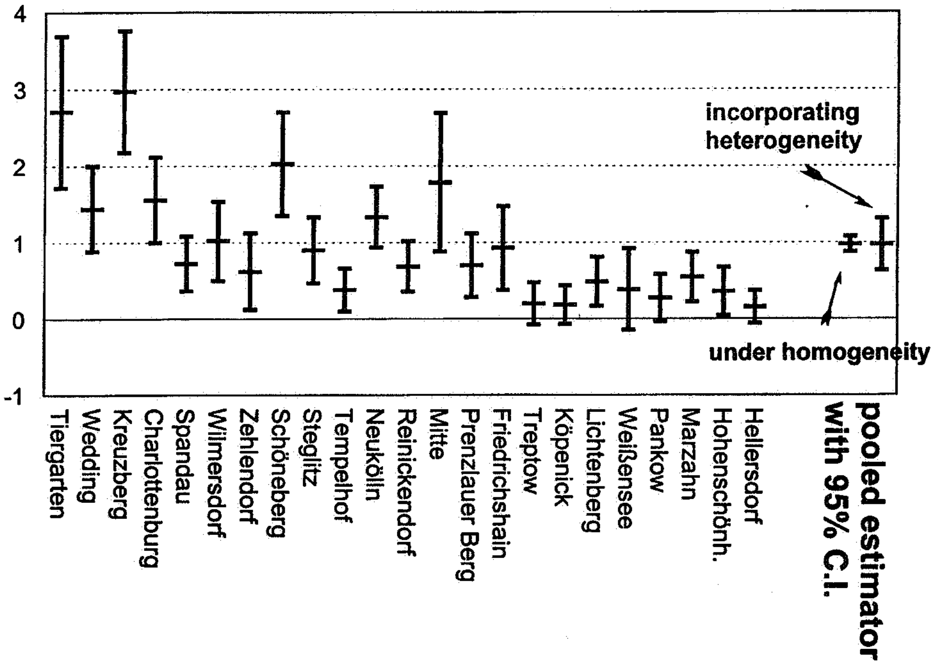

Fig. 1. SMR estimates of Hepatitis B in 23 Berlin city areas with pointwise 95%-confidence intervals.

2.2 *Example 2: Perinatal mortality in the North-West Thames Health Region*

As another realistic data set the small area data of Martuzzi and Hills (1995) on perinatal mortality in the North-West Thames Health Region in England based on the 5-year period 1986-1990 is considered. The region consists of 515 small areas. In this case, $\sum_i Y_i = \sum_i e_i = 2051$. It was found that $\hat{\tau}_1^2 = -0.0272790$ which is truncated to

0, and $\hat{\tau}_2^2 = 0.0167823$ as well as $\hat{\tau}_3^2 = 0.0369576$. There is small heterogeneity present in the data indicated by the ratio $\frac{\hat{\tau}_i^2}{\widehat{\text{Var}(SMR)}}$, where $\widehat{\text{Var}(SMR)} = \frac{1}{N-1}\sum_i(SMR_i - \overline{SMR})^2$, which takes on the values $0, \frac{0.0168}{0.6058}, \frac{0.0370}{0.6058} = 0, 0.0277, 0.0611$ for the 3 estimators, respectively.

3. Efficiency

When investigating the efficiency of the family of estimators $T(W, \alpha)$, we have to consider its variance:

$$\text{Var}(T(W,\alpha)) = \frac{\sum_i \alpha_i^2 \text{Var}(W_i)}{(\sum_i \alpha_i)^2} \tag{3.1}$$

which is completely specified, if $\text{Var}(W_i)$ is known. It is well-known that the efficient estimator (i.e. the one with minimum variance in the family $T(W, \alpha)$) chooses α_i proportional to $\frac{1}{\text{Var}(W_i)}$. Consequently, our interest concentrates on $\text{Var}(W_i)$. We have the following result.

LEMMA 3.1. *Let G be any distribution with finite moments to the power of four. Then*:

$$\begin{aligned}\text{Var}(W_i) = {} & \mu e_i^{-3} + (2\mu^2 + 7\tau^2)e_i^{-2} + 2(3\mu^{(3)} - 7\mu\tau^2 - 3\mu^3)e_i^{-1} + 3\mu^4 \\ & + \mu^{(4)} - \tau^4 + 6\mu^2\tau^2 - 4\mu\mu^{(3)}\end{aligned}$$

with $\mu^{(l)} = E_G(\theta^l)$ for $l = 3, 4$.

PROOF. Note that $W_i = e_i^{-2}(Y_i - e_i\mu)^2 - e_i^{-1}\mu$, where μ is non- random and known. Consequently we have

$$\begin{aligned}\text{Var}(W_i) &= e_i^{-4}\text{Var}\{(Y_i - e_i\mu)^2\} \\ &= e_i^{-4}[E\{(Y_i - e_i\mu)^4\} - (E\{(Y_i - e_i\mu)^2\})^2].\end{aligned} \tag{3.2}$$

Note that for fixed θ_i the random variable Y_i is distributed according to the Poisson distribution with parameter $\theta_i e_i : Y_i \mid \theta_i \sim Po(\theta_i e_i)$. The moments up to the order of four for a Poisson distributed variable Y are needed here to use (3.2). These can be easily derived by the factorial moments. In Haight ((1967), p. 5–6) the moments are given up to the order of ten. In our application it follows

$$\begin{aligned}&E(Y_i \mid \theta_i) = e_i\theta_i, \quad E(Y_i^2 \mid \theta_i) = e_i\theta_i + e_i^2\theta_i^2, \quad E(Y_i^3 \mid \theta_i) = e_i\theta_i + 3e_i^2\theta_i^2 + e_i^3\theta_i^3, \\ &E(Y_i^4 \mid \theta_i) = e_i\theta_i + 7e_i^2\theta_i^2 + 6e_i^3\theta_i^3 + e_i^4\theta_i^4.\end{aligned}$$

Furthermore, for each i the expected value of the SMR, θ_i, is to be interpreted as a realisation of the heterogeneity distribution $G : \theta_i \sim G$. Therefore, we have $E(Y_i^l) = E_G\{E\{Y_i^l \mid \theta_i\}\}, l = 1, 2, 3, 4$. From this fact, the moments of Y_i up to the power of four follow using the notation $\mu^{(l)} = E_G(\theta^l)$, $\mu = \mu^{(1)}$, $\tau^2 = \text{Var}_G(\theta) = \mu^{(2)} - \mu^2$:

$$\begin{aligned}&E(Y_i) = e_i\mu, \quad E(Y_i^2) = e_i\mu + e_i^2(\mu^2 + \tau^2), \\ &E(Y_i^3) = e_i\mu + 3e_i^2(\mu^2 + \tau^2) + e_i^3\mu^{(3)}, \\ &E(Y_i^4) = e_i\mu + 7e_i^2(\mu^2 + \tau^2) + 6e_i^3\mu^{(3)} + e_i^4\mu^{(4)}.\end{aligned}$$

Consequently, we have:

$$E\{(Y_i - e_i\mu)^2\} = e_i\mu + e_i^2\tau^2, \tag{3.3}$$

$$\begin{aligned} E\{(Y_i - e_i\mu)^4\} &= E(Y_i^4 - 4e_i\mu Y_i^3 + 6e_i^2\mu^2 Y_i^2 - 4e_i^3\mu^3 Y_i + e_i^4\mu^4) \\ &= e_i\mu + e_i^2(3\mu^2 + 7\tau^2) + 6e_i^3(\mu^{(3)} - 2\mu\tau^2 - \mu^3) \\ &\quad + e_i^4(\mu^{(4)} - 4\mu\mu^{(3)} + 6\mu^2\tau^2 + 3\mu^4). \end{aligned} \tag{3.4}$$

From (3.2), (3.3) and (3.4), we obtain the expression for $\mathrm{Var}(W_i)$ stated above. This ends the proof.

As a consequence from the expression for the variance of W_i derived above, it follows, that for large e_i, $\mathrm{Var}(W_i)$ behaves like a linear function in e_i^{-1}. To see this, note that

$$\frac{\partial}{\partial e_i^{-1}} \mathrm{Var}(W_i) = 3\mu e_i^{-2} + 2(2\mu^2 + 7\tau^2)e_i^{-1} + 6\mu^{(3)} - 14\mu\tau^2 - 6\mu^3.$$

Consequently, we have

$$\frac{\partial}{\partial e_i^{-1}} \mathrm{Var}(W_i) \to 2(3\mu^{(3)} - 7\mu\tau^2 - 3\mu^3) \quad \text{for} \quad e_i^{-1} \to 0.$$

This fact implies that, if we consider any fixed set of moments $(\mu, \tau^2, \mu^{(3)}, \mu^{(4)})$ and $\mathrm{Var}(W_i)$ as a function in e_i, then $\mathrm{Var}(W_i)$ increases approximately linearly with e_i^{-1} for large e_i. This result can be summarized in the following corollary.

COROLLARY 3.1.

$$\mathrm{Var}(W_i) \approx e_i^{-1} \quad \textit{for large} \quad e_i.$$

A further demonstration of this efficiency result is given below.

Lemma 3.1 above provides a closed form expression for the variance of W_i. However, this variance involves the first 4 moments of G, which are usually unknown. Therefore, it is not possible to give a closed form solution for the efficient estimator. Corollary 3.1 provides support that—for large e_i—$\hat{\tau}_2^2$ should be close to the efficient estimator. However, *largeness* is a vague term and it might be valuable to investigate the efficiency of these estimators for real non-random data sets $\{e_i\}$. Now, given any distribution G we are able to compare any linear unbiased estimator to the efficient estimator avoiding any kind of simulation approach. Below, we compare the three estimators $\hat{\tau}_j^2$, for $j = 1, 2, 3$ to the efficient estimator, where the e_i's stem from the two data sets of Example 1 and Example 2, respectively. We choose as heterogeneity distribution G two cases, namely $G_{(a)} = \binom{0.5\ 1.5}{0.5\ 0.5}$ and $G_{(b)} = \binom{0.8\ 0.9\ 1.1\ 1.2}{0.2\ 0.3\ 0.3\ 0.2}$. Here, the notation $G = \binom{\theta_1\ \cdots\ \theta_k}{p_1\ \cdots\ p_k}$ indicates a discrete probability distribution G giving weights $p_1, \ldots, p_k$ to a finite number k of mass points $\theta_1, \ldots, \theta_k$, respectively. Then, the variance of W_i is computed for each $i, i = 1, \ldots, N$ leading to optimal weights $\alpha_i = \mathrm{Var}(W_i)^{-1}$. These optimal weights are compared with the weights used by the three estimators, namely $1/N$, e_i, and e_i^2 by means of scatterplots α_i versus $1/\mathrm{Var}(W_i)$. The closer this relationship is to a straight line with positive slope, the closer is the associated estimator to the efficient one. The results are provided in Fig. 2 and Fig. 3. There is some evidence that $\hat{\tau}_2^2$ is often close to the efficient estimator, since the relationship between the optimal weights and the weights used by this estimator (e_i) appear to be the most linear. This provides some evidence for using $\hat{\tau}_2^2$.

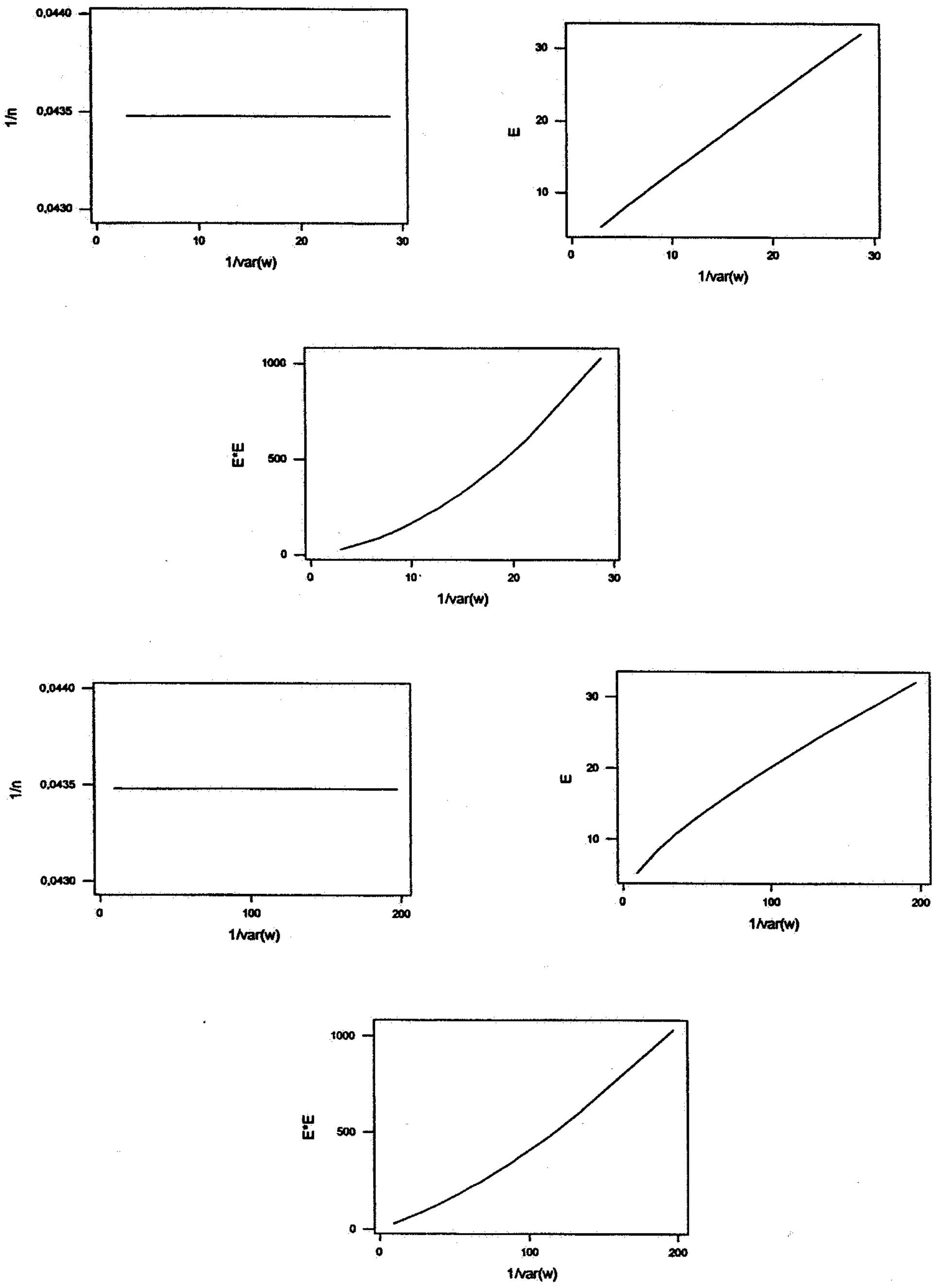

Fig. 2. Scatterplot of the three weighting schemes against $\frac{1}{\mathrm{Var}(W_i)}$ for Hepatits B—data of Example 1 using two different heterogeneity distributions G for computing $\mathrm{Var}(W_i)$; upper page $G_{(a)} = \begin{pmatrix} 0.5 & 1.5 \\ 0.5 & 0.5 \end{pmatrix}$, lower page $G_{(b)} = \begin{pmatrix} 0.8 & 0.9 & 1.1 & 1.2 \\ 0.2 & 0.3 & 0.3 & 0.2 \end{pmatrix}$.

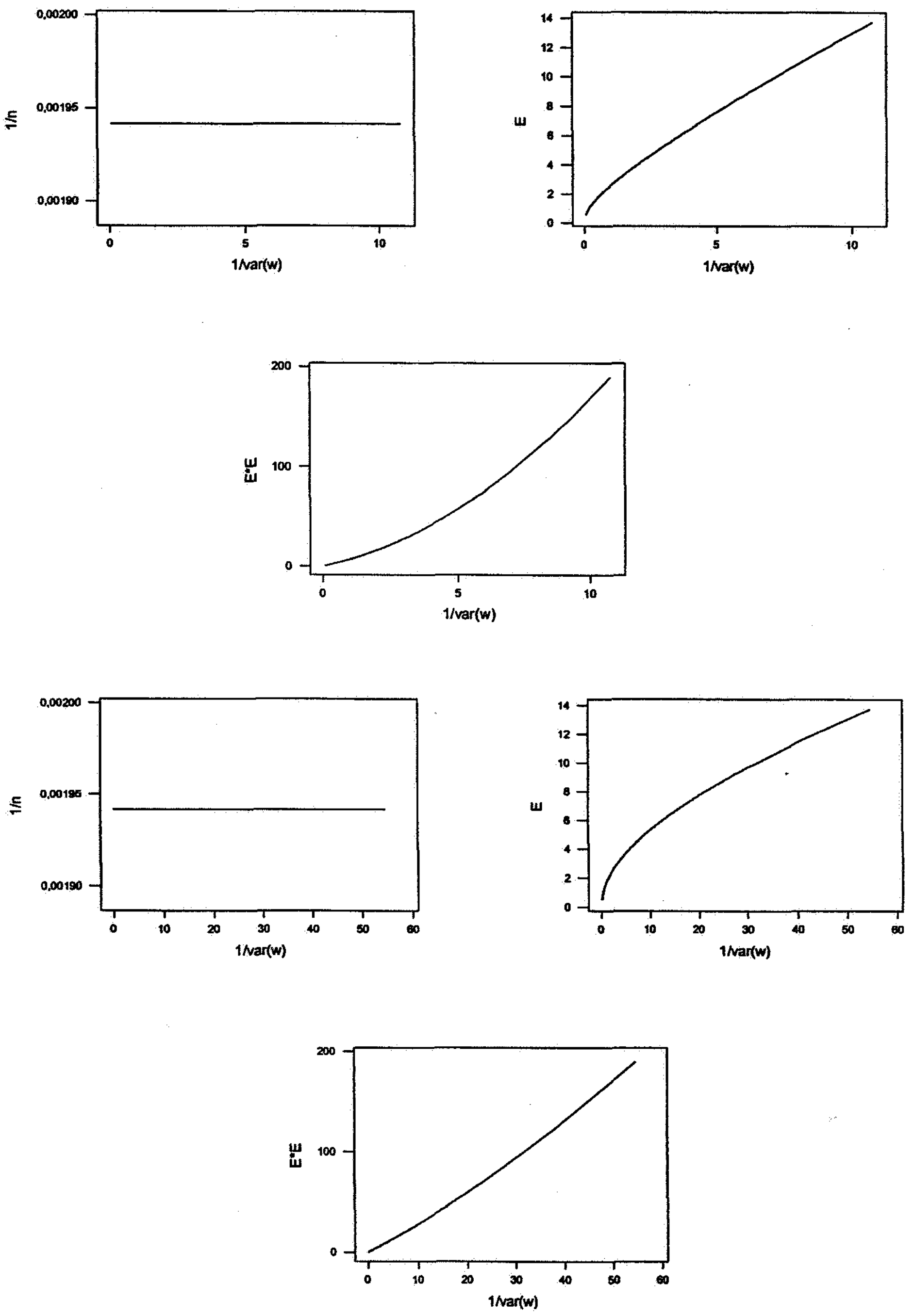

Fig. 3. Scatterplot of the three weighting schemes against $\frac{1}{\mathrm{Var}(W_i)}$ for Perinatal Mortality in the North-West Thames Health Region—data of Example 2 using two different heterogeneity distributions G for computing $\mathrm{Var}(W_i)$; upper page $G_{(a)} = \begin{pmatrix} 0.5 & 1.5 \\ 0.5 & 0.5 \end{pmatrix}$, lower page $G_{(b)} = \begin{pmatrix} 0.8 & 0.9 & 1.1 & 1.2 \\ 0.2 & 0.3 & 0.3 & 0.2 \end{pmatrix}$.

4. Consistency

We are interested in the asymptotic behavior of the estimator $T(W, \alpha)$. For this purpose we require the following two conditions:

(A1) There exists the moment to the power of four for the heterogeneity distribution $G : \mu^{(4)} < \infty$.

(A2) There exist constants $0 < a < A < \infty, 0 < \varepsilon$ such that

$$a \leq \alpha_i \leq A, \qquad \varepsilon \leq e_i \quad \text{for all } i.$$

THEOREM 4.1. *Let* (A1) *and* (A2) *be fulfilled. Then*:

$$T_N(W, \alpha) = \left(\sum_{i=1}^{N} \alpha_i W_i \right) \Big/ \left(\sum_{i=1}^{N} \alpha_i \right) \to \tau^2 \textit{almost surely},$$

in other words, the estimator $T_N(W, \alpha)$ is strongly consistent.

PROOF. We have that

$$E(W_i) = \tau^2 \tag{4.1}$$

and under (A1)

$$\begin{aligned} \text{Var}(W_i) &= \mu e_i^{-3} + (2\mu^2 + 7\tau^2)e_i^{-2} + 2(3\mu^{(3)} - 7\mu\tau^2 - 3\mu^3)e_i^{-1} \\ &\quad +3\mu^4 + \mu^{(4)} - \tau^4 + 6\mu^2\tau^2 - 4\mu\mu^{(3)}. \end{aligned} \tag{4.2}$$

With (A2) it follows, that there exists a finite constant W in such a way, that we have

$$\text{Var}(W_i) \leq W \quad \text{for all} \quad i. \tag{4.3}$$

To obtain W, we have to replace e_i^{-l} by ε^{-l} in (4.2) for $l = 1, 2, 3$. Let us define the following double sequence of random variables:

$$V_i^{(N)} := N \frac{\alpha_i}{\sum_{j=1}^{N} \alpha_j} W_i \quad \text{for} \quad N = 1, 2, \ldots, \quad i = 1, \ldots, N. \text{ Note, that}$$

$$\frac{1}{N} \sum_{i=1}^{N} V_i^{(N)} = T_N(W, \alpha). \tag{4.4}$$

For the variables $V_i^{(N)}$ we have that

$$V_1^{(N)}, \ldots, V_N^{(N)} \text{ are independent for all } N, \tag{4.5}$$

$$E(V_i^{(N)}) = N \left(\frac{\alpha_i}{\sum_{j=1}^{N} \alpha_j} \right) \tau^2 \text{ and with this}$$

$$\sum_{i=1}^{N} E(V_i^{(N)}) = N\tau^2, \tag{4.6}$$

$$\text{Var}(V_i^{(N)}) = N^2 \left(\frac{\alpha_i}{\sum_{j=1}^{N} \alpha_j} \right)^2 \text{Var}(W_i).$$

Consequently, there exists a finite constant, say, $\tilde{W} = (A^2/a^2)W$, such that $\mathrm{Var}(V_i^{(N)}) \leq \tilde{W}$ for all $i = 1, \ldots, N$ and all $N \geq 1$. With this, it follows that

$$\lim_{N\to\infty} \sum_{i=1}^{N} \frac{\mathrm{Var}(V_i^{(N)})}{i^2} \leq \tilde{W} \lim_{N\to\infty} \sum_{i=1}^{N} \frac{1}{i^2} = \tilde{W} \sum_{i=1}^{\infty} \frac{1}{i^2} = \tilde{W}\frac{\pi^2}{6} < \infty. \tag{4.7}$$

According to the strong law of large numbers by *Kolmogorov*, it follows from (4.5) and (4.7)

$$\frac{1}{N}\sum_{i=1}^{N} V_i^{(N)} - \frac{1}{N}\sum_{i=1}^{N} E(V_i^{(N)}) \to 0 \text{ almost surely.}$$

Because of (4.4) and (4.6) this is equivalent to $T_N(W, \alpha) \to \tau^2$ almost surely.

As a consequence we note that $\hat{\tau}_2^2$ and $\hat{\tau}_3^2$ are *strongly* consistent, if there exist positive bounds e, E such that $0 < e \leq e_i \leq E$ for all i. For $\hat{\tau}_1^2$ consistency follows from the fact that in this case we have $V_i^{(N)} = W_i$ as well as (4.3), leading to the inequality (4.7) with W instead of $\tilde{W}$.

5. Estimating heterogeneity mean and variance simultaneously

In many situations, however, it is not appropriate to assume that μ is known. Therefore, we have to replace μ in W_i by some estimate $\hat{\mu}$ leading to

$$W_i(\hat{\mu}) = \frac{(Y_i - e_i\hat{\mu})^2 - e_i\hat{\mu}}{e_i^2}. \tag{5.1}$$

Although only linear unbiased estimators $\hat{\mu}$ might be considered for μ, $W_i(\hat{\mu})$ is *not* necessarily unbiased for τ^2. This fact will cause a bias in $T(W(\hat{\mu}), \alpha)$. The bias will depend on the form of $T(W(\hat{\mu}), \alpha)$ as well as on $\hat{\mu}$ itself. Typically, two mean estimators are considered: the arithmetic mean $\hat{\mu}_1 = \frac{1}{N}\sum_i Y_i/e_i$ and the pooled mean $\hat{\mu}_2 = \frac{\sum_i Y_i}{\sum_i e_i}$. In Böhning (2000), the estimators

$$\hat{\tau}_1^2(\hat{\mu}_j) = \frac{1}{N-1}\left[\sum_{i=1}^{N}(Y_i - e_i\hat{\mu}_j)^2/e_i^2\right] - \hat{\mu}_j\frac{1}{N}\sum_{i=1}^{N}\frac{1}{e_i} \tag{5.2}$$

for $j = 1, 2$ were considered. It was shown that $\hat{\tau}_1^2(\hat{\mu}_1)$ is *unbiased* whereas $\hat{\tau}_1^2(\hat{\mu}_2)$ is biased. This property (unbiasedness) might be one reason to consider $\hat{\tau}_1^2(\hat{\mu}_1)$ at all. For the Hepatitis B data of Berlin we find the results as given in Table 2.

In the light of Section 3, attention is given to the estimator $\hat{\tau}_2^2(\hat{\mu}_j)$ for $j = 1, 2$. It is possible to provide exact expressions for their biases.

Table 2. Estimates of the mean and variance of the SMRs and $\hat{\tau}_1^2$ for Hepatitis B cases in the 23 city regions of Berlin (1995).

Estimator	$\hat{\mu}$	$\widehat{\mathrm{Var}}(SMRs)$	$\hat{\tau}_1^2$	$\hat{\tau}_1^2/\widehat{\mathrm{Var}}(SMRs)$
simple mean	0.9751	0.6214	0.5489	0.883
pooled mean	1.0188	0.6234	0.5470	0.877

THEOREM 5.1. *Let* $\hat{\tau_2^2}(\hat{\mu}_j) = \frac{\sum_{i=1}^N (Y_i - e_i\hat{\mu}_j)^2/e_i - \hat{\mu}_j N}{\sum_{i=1}^N e_i}$ *for* $j = 1, 2$. *Then:*

$$E(\hat{\tau_2^2}(\hat{\mu}_1)) = \left(1 - \frac{1}{n}\right)\tau^2 + \left(\frac{1}{n^2}\sum_i \frac{1}{e_i} - 2\frac{1}{\sum_i e_i}\right)\mu \tag{5.3}$$

$$E(\hat{\tau_2^2}(\hat{\mu}_2)) = \left(1 - \frac{\sum_i e_i^2}{(\sum_i e_i)^2}\right)\tau^2 - \frac{1}{\sum_i e_i}\mu. \tag{5.4}$$

The proof of this theorem is straightforward.

5.1 *Perinatal mortality in the North-West Thames Health Region*

For the data of Example 2, the following values of the biasing constants have been found: $(1 - \frac{1}{n}) = 0.998058$, $(1 - \frac{\sum_i e_i^2}{(\sum_i e_i)^2}) = 0.997376$, and $(\frac{1}{n^2}\sum_i \frac{1}{e_i} - 2\frac{1}{\sum e_i}) = -0.000206377$, $\frac{1}{\sum_i e_i} = 0.000487571$. This example illustrates that the amount of bias involved in expressions (5.3) or (5.4) respectively might be very small.

6. Discussion

The results of this paper can be used for several applications. It was mentioned earlier that the crude *SMR* has several disadvantages including some instability problems for small sample size applications (Lawson *et al.* (1999)). Typical examples are disease mapping and meta-analysis (Böhning (2000)). In these cases, it is more appropriate to use an empirical Bayes estimate of the *SMR*. Often this takes the form $\frac{Y_i + \mu^2/\tau^2}{e_i + \mu/\tau^2}$. It can be shown that this is the *linear Bayes* estimator with respect to the euclidean loss function and it is also the posterior mean if the prior is assumed to be a Gamma distribution (and $Y_i \sim Po(\theta e_i)$) (For details see Böhning (2000)). Clearly, μ and τ^2 need to be replaced by estimates and those that are proposed in this paper might be used for this purpose.

The advantage of the proposed estimators lies in their simple and non-iterative nature. Nevertheless, it should be pointed out that there are many other estimators leading to iterative solutions. One should mention the moment-estimators suggested by Breslow (1984) and Clayton and Kaldor (1987), or the pseudo-maximum-likelihood estimator suggested by Pocock *et al.* (1981), and Breslow (1984). These estimators have been well motivated when they were suggested, and they might be superior in their efficiency to the estimators proposed here. However, a thorough investigation and comparison of these estimators, either in terms of comparing these *iterative* estimators to each other, or comparing the *iterative* estimators to the *non-iterative* estimators suggested here, has not been done yet and is expected to be dealt with in future research.

Acknowledgement

This research is done under support of the German Research Foundation.

REFERENCES

Aitkin, M., Anderson, D., Francis, B. and Hinde, J. (1990). *Statistical Modelling in GLIM*, Clarendon Press, Oxford.

Böhning, D. (1994). A note on a test for Poisson overdispersion, *Biometrika* **81**, 418–419.

Böhning, D. (2000). *Computer-Assisted Analysis of Mixtures and Applications. Meta-Analysis, Disease Mapping, and Others*, Chapman & Hall / CRC, Boca Raton.

Breslow, N. E. (1984). Extra-Poisson variation in log-linear models, *Applied Statistics*, **33**, 38–44.

Clayton, D. and Kaldor, J. (1987). Empirical Bayes estimates of age-standardized relative risks for use in disease mapping, *Biometrics*, **43**, 671–681.

Collet, D. (1991). *Modelling Binary Data*, Chapman & Hall / CRC, Boca Raton.

Haight, F. A. (1967). *The Handbook of the Poisson Distribution*, Wiley, New York.

Lawson, A., Biggeri, A., Böhning, D., Lesaffre, E., Viel, J.-F. and Bertollini, R. (1999). *Disease Mapping and Risk Assessment for Public Health*, Wiley, New York.

Marshall, R. J. (1991). Mapping disease and mortality rates using empirical Bayes estimators, *Applied Statistics*, **40**, 283–294.

Martuzzi, M. and Elliot, P. (1996). Empirical Bayes estimation of small area prevalence of non-rare conditions, *Statistics in Medicine*, **15**, 1867–1873.

Martuzzi, M. and Hills, M. (1995). Estimating the degree of heterogeneity between event rates using likelihood, *American Journal of Epidemiology*, **141**, 369–374.

Pendergast, J. F., Gange, S. J., Newton, M. A., Lindstrom, M. J., Palta, M. and Fisher, M. R. (1996). A survey of methods for analyzing clustered binary response data, *International Statistical Review*, **64**, 89–118.

Pocock, S. J., Cook, D. G. and Beresford, S. A. A. (1981). Regression of area mortality rates on explanatory variables: What weighting is appropriate?, *Applied Statistics*, **30**, 286–295.

Williams, D. A. (1982). Extra-binomial variation in logistic linear models, *Applied Statistics*, **31**, 144–148.

Ann. Inst. Statist. Math.
Vol. 54, No. 4, 840–847 (2002)

ON NEW MOMENT ESTIMATION OF PARAMETERS OF THE GAMMA DISTRIBUTION USING ITS CHARACTERIZATION*

TEA-YUAN HWANG[1] AND PING-HUANG HUANG[2]

[1] *Institute of Statistics, National Tsing Hua University, Hsinchu, 30043, Taiwan, R.O.C.*
[2] *Department of Statistics and Insurance, Aletheia University, Tamsui, Taipei, 25103, Taiwan, R.O.C.*

(Received April 5, 2001; revised July 16, 2001)

Abstract. In this paper, the more convenient estimators of both parameters of the gamma distribution are proposed by using its characterization, and shown to be more efficient than the maximum likelihood estimator and the moment estimator for small samples. Furthermore, the distribution of the square of the sample coefficient of variation is obtained by computer simulation for some various values of the parameters and sample size, and thus the simulated confidence interval of its shape parameter is established.

Key words and phrases: Sample coefficient of variation, shape parameter, moment estimator, gamma distribution.

1. Introduction

The gamma distribution is widely used and plays an important role in the reliability field and the survival analysis, therefore a successful estimation of its parameters will be very important. Unfortunately, there exist some difficulties in present estimation schemes. Maximum likelihood estimation method for its parameters are described in the literature by Johnson and Kotz (1970), Cohen and Norgaard (1977), Cohen and Whitten (1982), Harter and Moore (1965), Bowman *et al.* (1987) and Bowman and Shenton (1988). Also some difficulties and modified MLEs are mentioned in these papers. On the other hand, Bai *et al.* (1991) and Bowman and Shenton (1988) pointed out a high degree of deviation of the estimators from the parent distribution if one uses the methods involving the moments.

Hwang and Hu (1999, 2000) proved the independence of sample coefficient of variation V_n with sample mean $\bar{X}_n$ when random samples are drawn from gamma distribution. In the next section, we use this characterization to derive the expectation and the variance of V_n^2, and then propose the new moment estimators of the shape and the scale parameters of gamma distribution. Furthermore, by simulation, we compare in Section 3 the new estimators with the maximum likelihood estimator and usual moment estimator in term of mean square error.

For finding a simulated confidence interval of the shape parameter, the simulated distribution of V_n^2 will be derived in Section 4. In Hu (1990), a set of non-linear transformations of order statistics was devised to derive the sample distribution of V_n; its explicit probability density function has been obtained only for sample size $n = 3, 4$ and 5. In

*This paper was supported by a grant from the National Sciences Council of R.O.C.

Hwang and Lin (2000), the detailed c.d.f.s of V_n and V_n^2 under exponential population are presented for $n = 3, 4$ and 5 only. Until now it is still difficult to derive explicitly the sample distribution of V_n^2, thus simulation is used to find the sample distribution of normalized V_n^2 for shape parameters $= 0.5, 1.0, 1.5, 2.0$ and scale parameters $= 1, 0.5$ and 0.25 when $n = 5, 10, 15, 20$ and 25 respectively; it looks almost like gamma distribution for the cases mentioned above. Finally, the simulated confidence intervals for shape parameter are established.

2. New moment estimator of parameters of the gamma distribution

For deriving new moment estimator of parameters of the gamma distribution, we need the following theorem taken from Hu (1990) and Hwang and Hu (1999).

THEOREM 2.1. *Let $n \geq 3$ and let X_1, Λ, X_n be n positive i.i.d. random variables having a probability density function $f(x)$. Then the independence of the sample mean $\bar{X}_n$ and the sample coefficient of variation $V_n = S_n/\bar{X}_n$ is equivalent to that f is a gamma density where S_n is the sample standard deviation.*

The next result and Theorem 2.1 are useful in deriving the expectation and the variance of $V_n^2 = (S_n/\bar{X}_n)^2$, where $\bar{X}_n$ and S_n are respectively the sample mean and the sample standard deviation.

THEOREM 2.2. *Let $n \geq 3$ and let X_1, Λ, X_n be drawn from a population having a gamma density*

$$g(x;\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, \qquad x > 0, \alpha > 0, \beta > 0.$$

Then

$$E(\bar{X}_n^2) = \frac{(n\alpha+1)n\alpha}{n^2\beta^2}$$

and

$$E(S_n^2) = \frac{\alpha}{\beta^2}$$

where $\bar{X}_n$ and S_n^2 are respectively their sample mean and sample variance.

PROOF. It is easy to prove that

$$E(X) = \frac{\alpha}{\beta}, \quad \mathrm{Var}(X) = \frac{\alpha}{\beta^2}, \quad V^2 = \frac{\mathrm{Var}(X)}{E^2(X)} = \frac{1}{\alpha},$$

$$E(X^k) = \frac{(\alpha+k-1)\cdots(\alpha+1)\alpha}{\beta^k} \quad \text{for} \quad k \geq 1, \tag{2.1}$$

and that $\bar{X}_n$ has the following p.d.f.

$$g(\bar{x}_n;\alpha,\beta) = \frac{(n\beta)^{n\alpha}}{\Gamma(n\alpha)} \bar{x}_n^{n\alpha-1} e^{-n\beta\bar{x}_n}$$

and moments

$$E(\bar{X}_n^k) = \frac{(n\alpha+k-1)\cdots(n\alpha+1)(n\alpha)}{n^k\beta^k} \quad \text{for} \quad k \geq 1. \tag{2.2}$$

Thus (2.1) and (2.2) together give the following relation:

$$\begin{aligned} E(S_n^2) &= \frac{1}{(n-1)} E\left[\sum_{i=1}^{n}(X_i - \bar{X}_n)^2\right] \\ &= \frac{1}{n-1}[E(X^2) - E(\bar{X}_n^2)] \\ &= \frac{\alpha}{\beta^2} \end{aligned}$$

and Theorem 2.2 is established.

Theorem 2.2 implies that the sample mean $\bar{X}_n$ and the sample variance S_n^2 are respectively the unbiased estimator of population mean α/β and population variance α/β^2, a property also possessed by the normal population. Thus we have the moment estimators $\hat{\alpha}_m$ and $\hat{\beta}_m$ of α and β as follows:

$$\hat{\alpha}_m = \frac{\bar{X}_n^2}{S_n^2}, \qquad \hat{\beta}_m = \frac{\bar{X}_n}{S_n^2}.$$

THEOREM 2.3. *Let $n \geq 3$ and let X_1, Λ, X_n be drawn from a population having a gamma density*

$$g(x;\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, \qquad x > 0, \alpha > 0, \beta > 0.$$

Then

$$E\left(\frac{S_n^2}{\bar{X}_n^2}\right) = \frac{n}{1+n\alpha}$$

where $\bar{X}_n$ and S_n^2 are respectively their sample mean and sample variance.

PROOF. By Theorem 2.1, we have

$$E(S_n^2) = E\left(\frac{S_n^2}{\bar{X}_n^2} \cdot \bar{X}_n^2\right) E\left(\frac{S_n^2}{\bar{X}_n^2}\right) \cdot E(\bar{X}_n^2)$$

and hence

$$E\left(\frac{S_n^2}{\bar{X}_n^2}\right) = \frac{E(S_n^2)}{E(\bar{X}_n^2)}.$$

Applying Theorem 2.2, to the above identity yields that

$$E\left(\frac{S_n^2}{\bar{X}_n^2}\right) = \frac{n}{1+n\alpha}$$

and Theorem 2.3 is established.

Note that $E(S_n^2/\bar{X}_n^2) \to \frac{1}{\alpha}$ as $n \to \infty$ and that $\frac{1}{\alpha}$ is the square of the coefficient of variation. Thus $S_n^2/\bar{X}_n^2$ is an asymptotically unbiased estimator of the square of the coefficient of variation.

By Theorem 2.3 , V_n^2 is the unbiased estimator of $\frac{n}{1+n\alpha}$, thus it seems reasonable to propose $\frac{1}{V_n^2} - \frac{1}{n}$ as the estimator of α, namely

$$\hat{\alpha}_c = \frac{1}{V_n^2} - \frac{1}{n}.$$

It is easy to show that $\hat{\alpha}_c > 0$. Therefore, by the identity $E(\bar{X}_n) = \frac{\alpha}{\beta}$ and moment estimation method approach, it seems also reasonable to propose

$$\hat{\beta}_c = \frac{\hat{\alpha}_c}{\bar{X}_n} = \frac{1}{\bar{X}_n}\left(\frac{1}{V_n^2} - \frac{1}{n}\right).$$

Note that $\hat{\alpha}_c \to \hat{\alpha}_m$ and $\hat{\beta}_c \to \hat{\beta}_m$ as $n \to \infty$, and their differences get bigger when the sample size n gets smaller.

The fact that $\hat{\alpha}_c$ and $\hat{\beta}_c$ are more convenient to be computed than the maximum likelihood estimators $\hat{\alpha}_L$ and $\hat{\beta}_L$ of α and β is quite trivial. For comparing the efficiency of $\hat{\alpha}_c$ and $\hat{\beta}_c$ with $\hat{\alpha}_L$ and $\hat{\beta}_L$, respectively, we apply the next theorem to derive the normalized behevior of V_n.

THEOREM 2.4. *Let $n \geq 3$ and let X_1, Λ, X_n be drawn from a population having a gamma density*

$$g(x;\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, \qquad x > 0, \alpha > 0, \beta > 0.$$

Then

(2.3) $$\operatorname{Var}(S_n^2) = \frac{\alpha}{\beta^4}\left[\frac{2n\alpha}{(n-1)^2} + \frac{6}{n}\right]$$

and

(2.4) $$\operatorname{Var}\left(\frac{S_n^2}{\bar{X}_n^2}\right) \frac{2\alpha(\alpha+1)}{(n-1)\left(\alpha+\frac{1}{n}\right)^2\left(\alpha+\frac{2}{n}\right)\left(\alpha+\frac{3}{n}\right)}.$$

PROOF. Since $M_X(t) = (1 - t/\beta)^{-\alpha}$, we have

$$\begin{aligned}
&E(X) = \alpha/\beta \\
&E(X^2) = \alpha(\alpha+1)/\beta^2 \\
&E(X^3) = \alpha(\alpha+1)(\alpha+2)/\beta^3 \\
&E(X^4) = \alpha(\alpha+1)(\alpha+2)(\alpha+3)/\beta^4
\end{aligned}$$

and $M_{\bar{X}_n}(t) = (1 - \frac{1}{n\beta}t)^{-n\alpha}$ gives

$$\begin{aligned}
&E(\bar{X}_n) = \alpha/\beta \\
&E(\bar{X}_n^2) = \alpha\left(\alpha+\frac{1}{n}\right)\Big/\beta^2 \\
&E(\bar{X}_n^3)\alpha\left(\alpha+\frac{1}{n}\right)\left(\alpha+\frac{2}{n}\right)\Big/\beta^2 \\
&E(\bar{X}_n^4) = \alpha\left(\alpha+\frac{1}{n}\right)\left(\alpha+\frac{2}{n}\right)\left(\alpha+\frac{3}{n}\right)\Big/\beta^4.
\end{aligned}$$

By using the above identities, we obtain

$$E(S_n^4) = \frac{\alpha}{\beta^4(n-1)^2}\left[(n^2-1)\alpha + 6\left(\sqrt{n} - \frac{1}{\sqrt{n}}\right)^2\right]$$

and combing the above identity and Theorem 2.2, we have

$$\mathrm{Var}(S_n^2) = \frac{\alpha}{\beta^4}\left[\frac{2n\alpha}{(n-1)^2} + \frac{6}{n}\right].$$

Next, the independence of $(S_n^2/\bar{X}_n^2)^2$ and $(\bar{X}_n)^2$ gives

$$E\left(\frac{S_n^4}{\bar{X}_n^4}\right) = \frac{E(S_n^4)}{E(\bar{X}_n^4)} \quad \text{and}$$

$$E\left(\frac{S_n^2}{\bar{X}_n^2}\right)^2 = \frac{E(S_n^4)}{E(\bar{X}_n^4)} = \frac{(n^2-1)\alpha + 6\left(\sqrt{n} - \frac{1}{\sqrt{n}}\right)^2}{(n-1)^2\left(\alpha+\frac{1}{n}\right)\left(\alpha+\frac{2}{n}\right)\left(\alpha+\frac{3}{n}\right)}.$$

Thus we have

$$\mathrm{Var}\left(\frac{S_n^2}{\bar{X}_n^2}\right) = \frac{2\alpha(\alpha+1)}{(n-1)\left(\alpha+\frac{1}{n}\right)^2\left(\alpha+\frac{2}{n}\right)\left(\alpha+\frac{3}{n}\right)}$$

and Theorem 2.4 is established.

Theorem 2.4 implies that both $\mathrm{Var}(S_n^2)$ and $\mathrm{Var}(V_n^2)$ tend to zero as $n \to \infty$. Thus S_n^2 and V_n^2 are respectively consistent estimators of $\frac{\alpha}{\beta^2}$ and $\frac{n}{n\alpha+1}$ for large samples. After some computations, we find the following inequality:

$$\frac{\mathrm{Var}(V_n^2)}{\mathrm{Var}(S_n^2)} < \left(\frac{n\beta}{n\alpha+1}\right)^4 = \left(\frac{\beta}{\alpha+\frac{1}{n}}\right)^4$$

$$\mathrm{Var}\,(V_n^2) < \mathrm{Var}(S_n^2), \qquad \beta \leq \alpha + \frac{1}{n}.$$

Furthermore, the fact that $\mathrm{Var}(V_n^2) \to 0$ as $n \to \infty$ also confirms the reason: why V_n can always considered approximately as constant for large samples, and it can be used in checking experiment results and in estimating the standard deviation.

3. The comparison with previous estimators

In this section, the comparison of our estimators $(\hat{\alpha}_c, \hat{\beta}_c)$ with maximum likehihood estimators $(\hat{\alpha}_L, \hat{\beta}_L)$ and moment estimators $(\hat{\alpha}_m, \hat{\beta}_m)$ would be done in terms of mean square error by using the simulation procedures proposed by Greenwood and Durand (1960) which improved Thom (1958). Note that $(\hat{\alpha}_L, \hat{\beta}_L)$ are more difficult to compute than $(\hat{\alpha}_c, \hat{\beta}_c)$ and $(\hat{\alpha}_m, \hat{\beta}_m)$.

We have done more than 100,000 times simulation for $\alpha = 0.5, 1, 1.5, 2$ and $\beta = 1, 2, 4$ when $n = 5, 10, 15, 20$ and 25, and obtain the following conclusions:

(1) $(\hat{\alpha}_c, \hat{\beta}_c)$ is the best estimators of (α, β), $(\hat{\alpha}_L, \hat{\beta}_L)$ the next and $(\hat{\alpha}_m, \hat{\beta}_m)$ the worse for $n \leq 25$, and the smaller n the better $(\hat{\alpha}_c, \hat{\beta}_c)$;

(2) $(\hat{\alpha}_L, \hat{\beta}_L)$ is the best estimators of (α, β), $(\hat{\alpha}_c, \hat{\beta}_c)$ the next and $(\hat{\alpha}_m, \hat{\beta}_m)$ the worse for $n > 25$, and the larger n the better $(\hat{\alpha}_L, \hat{\beta}_L)$.

4. The confidence interval for shape parameter

For deriving the confidence interval of the shape paramter, we need to study the behavior of V_n^2.

By Theorem 2.3 and Theorem 2.4, we construct the normalized distribution of V_n^2 under gamma distribution with various parameters values: $\alpha = 0.5, 1, 1.5, 2$ and $\beta = 1, 2, 4$ when $n = 5, 10, 15$ and 20 by 100,000 simulations . Its simulated c.d.f. are presented in Hwang (2000). Comparing our simulated results with the results presented in Hwang and Lin (2000) for $\alpha = 1$ and $\beta = 1$ when $n = 5$, they are quite same; for example $P(V_5^2 \leq 1.10) = 0.7599$ in Hwang and Lin (2000) while it is equal to 0.7630 in this paper. From the simulated results we conclude that V_n^2 looks almost like a gamma distribution for any α, β and any n. This conclusion is justified by both of the Kolmogorov-Smirnov test and χ^2 test.

Furthermore, we obtain also by simulations the frequencies of V_n^2 falling in one standard deviation; two standard deviation and three standard deviation interval (with its mean as their center) respectively from Hwang (2000) for $\alpha = 0.5, 1.0, 1.5, 2.0$ and $\beta = 1, 2, 4$ when $n = 5, 10, 15, 20, 25$ and 30. The results are presented in Table 1. The

Table 1.

n	α	1σ	2σ	3σ
5	0.5	72.620	94.710	98.870
	1.0	75.680	95.153	98.297
	1.5	76.010	95.313	98.220
	2.0	75.943	95.310	98.300
10	0.5	76.920	95.340	98.287
	1.0	76.323	95.530	98.440
	1.5	75.717	95.467	98.433
	2.0	74.897	95.833	98.707
15	0.5	77.160	95.550	98.443
	1.0	75.890	95.593	98.487
	1.5	74.853	95.773	98.553
	2.0	73.843	95.760	98.627
20	0.5	77.010	95.577	98.353
	1.0	75.047	95.967	98.700
	1.5	74.027	95.983	98.723
	2.0	72.993	95.827	98.657
25	0.5	76.813	95.657	98.410
	1.0	75.037	95.977	98.647
	1.5	73.227	95.680	98.730
	2.0	72.450	95.953	98.957
30	0.5	76.233	95.623	98.460
	1.0	73.780	95.820	98.683
	1.5	72.173	95.647	98.763
	2.0	71.990	95.697	98.803

behavior of sample mean of V_n^2 is also investigated, and the conclusion is the same as central limit theorem; this fact can be justified by any of the Kolmogorov-Smirnov test and χ^2 test.

By Theorem 2.1 and Theorem 2.3, we have the mean and the variance of V_n^2 as follows:

$$E\left(\frac{S_n^2}{\bar{X}_n^2}\right) = \frac{n}{1+n\alpha}$$

and

$$\sigma^2 = \operatorname{Var}\left(\frac{S_n^2}{\bar{X}_n^2}\right) = \frac{2\alpha(\alpha+1)}{(n-1)\left(\alpha+\frac{1}{n}\right)\left(\alpha+\frac{2}{n}\right)\left(\alpha+\frac{3}{n}\right)}.$$

For finding the confidence interval of α, we need to manipulate the following probabilities for various values of α and n,

$$\Pr\left(\frac{n}{n\alpha+1} - k\sigma \le \frac{S_n^2}{\bar{X}_n^2} \le \frac{n}{n\alpha+1} + k\sigma\right).$$

Since it is quite difficult to derive, we present its approximate probabilities in Table 1 and the conclusions would be drawn for some values of $\alpha = 0.5, 1.0, 1.5, 2.0$, and $n = 5, 10, 15, 20, 25$ and 30 as follows:

$$\Pr\left(\frac{n}{n\alpha+1} - \sigma \le \frac{S_n^2}{\bar{X}_n^2} \le \frac{n}{n\alpha+1} + \sigma\right) \cong 0.75,$$

$$\Pr\left(\frac{n}{n\alpha+1} - 2\sigma \le \frac{S_n^2}{\bar{X}_n^2} \le \frac{n}{n\alpha+1} + 2\sigma\right) \cong 0.95$$

and

$$\Pr\left(\frac{n}{n\alpha+1} - 3\sigma \le \frac{S_n^2}{\bar{X}_n^2} \le \frac{n}{n\alpha+1} + 3\sigma\right) \cong 0.98.$$

Here 0.75, 0.95 and 0.98 will be assumed to be the mean probabilities respectively for various α,β and n. Thus the approximated 75.5%, 95% and 98% confidence intervals for α could be concluded respectively as follows:

$$\left(\frac{1}{\frac{S_n^2}{\bar{X}_n^2}+\hat{\sigma}} - \frac{1}{n}, \frac{1}{\frac{S_n^2}{\bar{X}_n^2}-\sigma} - \frac{1}{n}\right), \quad \left(\frac{1}{\frac{S_n^2}{\bar{X}_n^2}+2\hat{\sigma}} - \frac{1}{n}, \frac{1}{\frac{S_n^2}{\bar{X}_n^2}-2\sigma} - \frac{1}{n}\right) \quad \text{and}$$

$$\left(\frac{1}{\frac{S_n^2}{\bar{X}_n^2}+3\hat{\sigma}} - \frac{1}{n}, \frac{1}{\frac{S_n^2}{\bar{X}_n^2}-3\hat{\sigma}} - \frac{1}{n}\right)$$

where $\hat{\sigma}^2 = \frac{2\hat{\alpha}_c(\hat{\alpha}_c+1)}{(n-1)\left(\hat{\alpha}_c+\frac{1}{n}\right)^2\left(\hat{\alpha}_c+\frac{2}{n}\right)\left(\hat{\alpha}_c+\frac{3}{n}\right)}$, and $\hat{\alpha}_c$ is the new moment estimator of α proposed by using Theorem 2.3.

After simplification of the following probability, we write

$$\Pr\left(\frac{n}{n\alpha+1} - k\sigma \le \frac{S_n^2}{\bar{X}_n^2} \le \frac{n}{n\alpha+1} + k\sigma\right)$$

as

$$\Pr\left(\frac{\bar{X}_n^2}{S_n^2}\left(1-\frac{k}{\sqrt{n-1}}\right)-\frac{1}{n}\leq\alpha\leq\frac{\bar{X}_n^2}{S_n^2}\left(1-\frac{k}{\sqrt{n-1}}\right)-\frac{1}{n}\right)$$

and the approximate 75%, 95% and 98% confidence intervals for α are

$$\left(\frac{\bar{X}_n^2}{S_n^2}\left(1-\frac{k}{\sqrt{n-1}}\right)-\frac{1}{n},\frac{\bar{X}_n^2}{S_n^2}\left(1-\frac{k}{\sqrt{n-1}}\right)-\frac{1}{n}\right),\qquad k=1,2,3,$$

respectively for large sample.

Acknowledgements

We thank the editor, associate editor, and the referees for their thoughtful comments. We also thank L. H. Chen, S. T. Kao and K. E. Chang, master students of Institute of Statistics of Tsing Hua University, for their carefully performing all simulations.

References

Bai, J., Jakeman, A. J. and McAleer, M. (1991). A new approach to maximum likelihood estimation of the three-parameter gamma and Weibull distributions, *Austral. J. Statist.*, **33**, 397–410.

Bowman , K. O. and Shenton, L. R. (1988). Properties of estimators for the gamma distribution, Marcel Dekker, New York.

Bowman, K. O., Shenton, L. R. and Lam, H. K. (1987). Simulation and estimation problems associated with the three-parameter gamma distribution, *Comm. Statist. Simulation Comput.*, **16**, 1147–1188.

Cohen, A. C. and Norgaard, N. J. (1977). Progressively censored sampling in the three-parameter gamma distribution, *Technometrics*, **19**, 333–340.

Cohen, A. C. and Whitten, B. J. (1982). Modified moment and maximum likelihood estimators for parameters of the three-parameter gamma distribution, *Comm. Statist. Simulation. Comput.*, **11**, 197–216.

Greenwood, J. A. and Durand, D. (1960). Aids for fitting the gamma distribution by maximum likelihood, *Technometrics*, **2**, 55–65.

Harter, H. L. and Moore, A. H. (1965). Maximum-likelihood estimation of the parameters of the gamma and Weibull populations from complete and from censored samples, *Technometrics*, **7**, 639–643.

Hu, C. Y. (1990). On signal to noise ratio statistics, Ph. D. Thesis, Institute of Statistics, National Tsing Hua University, Taiwan.

Hwang, T. Y. (2000). On the inference of parameters of gamma distribution, Tech. Report, 89-2118-M-007-014, National Science Council, Taiwan.

Hwang, T. Y. and Hu, C. Y. (1999). On a characterization of the gamma distribution: The independence of the sample mean and the sample coefficient of variation, *Ann. Inst. Statist. Math.*, **51**, 749–753.

Hwang, T. Y. and Hu, C. Y. (2000). On some characterizations of population distribution, *Taiwanese J. Math.*, **4**(3), 427–437.

Hwang, T. Y. and Lin, Y. K. (2000). On the distribution of the sample heterogeneity of molecular polymer, *Tamsui Oxford Journal of Mathematical Sciences*, **16**(2), 133–149.

Johnson, N. J. and Kotz, S. (1970). *Continuous Univariate Distributions-1*, Wiley, New York.

Thom, H. C. S. (1958). A note on the gamma distribution, Washington, *Monthly Weather Review*, **86**(4), 117–121, Office of Climatology, U. S. Weather, Washington D. C.

Ann. Inst. Statist. Math.
Vol. 54, No. 4, 848–860 (2002)
©2002 The Institute of Statistical Mathematics

A COMPARISON OF RESTRICTED AND UNRESTRICTED ESTIMATORS IN ESTIMATING LINEAR FUNCTIONS OF ORDERED SCALE PARAMETERS OF TWO GAMMA DISTRIBUTIONS

YUAN-TSUNG CHANG[1] AND NOBUO SHINOZAKI[2]

[1] *Department of Studies on Contemporary Society, Mejiro University, Iwatsuki, Saitama 339-8501, Japan*
[2] *Department of Administration Engineering, Faculty of Science and Technology, Keio University, Yokohama, Kanagawa 223-8522, Japan*

(Received May 15, 2000; revised September 25, 2001)

Abstract. The problem of estimating linear functions of ordered scale parameters of two Gamma distributions is considered. A necessary and sufficient condition on the ratio of two coefficients is given for the maximum likelihood estimator (MLE) to dominate the crude unbiased estimator (UE) in terms of mean square error. A modified MLE which satisfies the restriction is also suggested, and a necessary and sufficient condition is also given for it to dominate the admissible estimator based solely on one sample. The estimation of linear functions of variances in two sample problem and also of variance components in a one-way random effect model is mentioned.

Key words and phrases: MLE, unbiased estimator, admissible estimator, variance estimation.

1. Introduction

In this paper, we discuss the problem of estimating linear functions of scale parameters of $Gamma(\alpha_i, \lambda_i)$, $i = 1, 2$, when α_i, $i = 1, 2$ are known and the restriction $\lambda_1 \leq \lambda_2$ is given. We note that a special case of this general problem is given in two samples problem with different but ordered variances. Estimation of smaller or larger variance has been discussed by Kushary and Cohen (1989). Among the linear functions of variances estimation of those with positive coefficients is especially important since they are the variances of linear functions of two random variables.

Consider, for another example, a one-way random effects model given by

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \qquad i = 1, \dots, I, \quad j = 1, \dots, J,$$

where $\alpha_i \sim N(0, \sigma_A^2)$ and $\varepsilon_{ij} \sim N(0, \sigma_1^2)$. Letting $S_1 = \sum_i \sum_j (y_{ij} - \bar{y}_{i.})^2$ and $S_2 = J \sum_i (\bar{y}_{i.} - \bar{y}_{..})^2$ for $\bar{y}_{i.} = J^{-1} \sum_j y_{ij}$ and $\bar{y}_{..} = (IJ)^{-1} \sum_i \sum_j y_{ij}$, one has that $S_i/\sigma_i^2 \sim \chi^2_{n_i}, i = 1, 2$, for $n_1 = I(J-1)$, $n_2 = I - 1$ and $\sigma_2^2 = \sigma_1^2 + J\sigma_A^2$. In this situation, of great interest is to estimate the between component of variance σ_A^2, being represented by $\sigma_A^2 = J^{-1}\ (\sigma_2^2 - \sigma_1^2)$, which is a linear function of two ordered Gamma scale parameters σ_1^2 and σ_2^2.

There has been considerable interest in the estimation of the parameters when there are linear restrictions among parameters. Typical types of the restrictions are positivity, simple ordering and simple tree ordering. See, for example, Barlow *et al.* (1972) and Robertson *et al.* (1988). Many papers focus on normal mean estimation and on comparing the maximum likelihood estimator (MLE) which satisfies the order restriction with the unbiased estimator (UE) coordinately (Lee (1981), Kelly (1989)). However, MLE does not always improve UE (Lee (1988)), and it is not always true that every linear function of MLE dominates the one of UE in terms of mean square error (MSE) (see also Hwang and Peddada (1994) and Fernández *et al.* (1999)). In recent years, Rueda and Salvador (1995) have considered the problem of estimating general linear function of normal means when two linear inequality constraints are given, and have shown that MLE gives an improvement for any coefficients. In estimating linear functions of positive normal means, Shinozaki and Chang (1999) have given a necessary and sufficient condition on the coefficients so that linear function of MLE dominates the one of UE in terms of MSE. Thus they show that MLE dominates UE for any choice of coefficients if and only if the number of means is less than 5. Independently, Fernández *et al.* (2000) have discussed the same problem under a symmetric unimodal location model. Other than normal distribution, there are also many papers dealing with the estimation of parameters under order restrictions. Kushary and Cohen (1991) considered the estimation of ordered Poisson parameters. Kaur and Singh (1991) considered the estimation of ordered means of two exponential population with the same sample sizes. They compared MLE with UE coordinately and showed that MLE dominates UE. This is a special case of the estimation problem of Gamma scale parameters when order restriction is given. See Hwang and Peddada (1994) and Kubokawa and Saleh (1994) for general scale parameter estimation under order restriction.

Here we first compare MLE with UE in estimating linear functions of ordered scale parameters of two Gamma distributions. To evaluate the difference of MSE of two estimators we give some useful lemmas in Section 2. We give a necessary and sufficient condition on the ratio of coefficients for MLE to dominate UE in terms of MSE. We also numerically obtain the upper bounds of the ratios for some typical values of $\alpha_i, i = 1, 2$. All these results are given in Section 3. Other than UE, there is another standard estimator of λ_i which we can obtain by replacing α_i by $\alpha_i + 1$ in UE. This estimator is an admissible one based solely on one sample under quadratic loss. In Section 4, we suggest a modified MLE which satisfies the restriction and give a necessary and sufficient condition on the ratio of coefficients for the modified MLE to dominate the unrestricted one. The lower bounds of the ratios are also given for some typical values of $\alpha_i, i = 1, 2$. We give some concluding remarks in Section 5.

2. Preliminaries

Let X_i, $i = 1, 2$ be independent $Gamma(\alpha_i, \lambda_i)$ random variables, having density

$$f_{\lambda_i}(x_i) = x_i^{\alpha_i - 1}\lambda_i^{-\alpha_i}e^{-x_i/\lambda_i}/\Gamma(\alpha_i), \qquad 0 < x_i < \infty \tag{2.1}$$

where $\alpha_i(> 0)$ is known and $\lambda_i(> 0)$ is unknown but satisfying $0 < \lambda_1 \leq \lambda_2 < \infty$. We note that even if we have more than one observations, we can reduce the case to the above one by considering the sufficient statistics which also follow Gamma distributions. The MLE of λ_i is given by

$$\hat{\lambda}_i = \frac{X_i}{\alpha_i} + (-1)^i \frac{(\alpha_2 X_1 - \alpha_1 X_2)^+}{\alpha_i(\alpha_1 + \alpha_2)}, \qquad i = 1, 2,$$

where $a^+ = \max(0, a)$ and X_i/α_i is the unbiased estimator (UE) of λ_i.

The best estimator of λ_i of the form cX_i under squared error loss is $X_i/(\alpha_i + 1)$, which is an admissible estimator of λ_i based solely on X_i. We also consider a modified MLE that satisfies the restriction $0 < \lambda_1 \leq \lambda_2 < \infty$ given by

$$\tilde{\lambda}_i = \frac{X_i}{\alpha_i + 1} + (-1)^i \frac{((\alpha_2 + 1)X_1 - (\alpha_1 + 1)X_2)^+}{(\alpha_i + 1)(\alpha_1 + \alpha_2 + 2)}, \qquad i = 1, 2,$$

which we can obtain by replacing α_i by $\alpha_i + 1$ in the MLE $\hat{\lambda}_i$. We note that Kubokawa and Saleh (1994) have proposed another improving estimator of λ_1 by their general argument.

Let c_1, c_2 be given constants and we want to estimate $c_1\lambda_1 + c_2\lambda_2$. We first compare two estimators, UE, $\sum_{i=1}^2 c_i X_i/\alpha_i$ and, MLE, $\sum_{i=1}^2 c_i \hat{\lambda}_i$ by their mean square error (MSE) and give a condition on c_1 and c_2 for MLE to dominate UE. We also compare $\sum_{i=1}^2 c_i X_i/(\alpha_i + 1)$ with modified MLE $\sum_{i=1}^2 c_i \tilde{\lambda}_i$, and give a condition on c_1 and c_2 for the modified MLE to dominate the competitor.

We should first mention that the domination depends only on the ratio c_2/c_1. This is generally true so far as we are concerned with estimation of linear functions $\sum_{i=1}^2 c_i \theta_i$ of parameters θ_1 and θ_2 and compare two estimators of the form $\sum_{i=1}^2 c_i \hat{\theta}_i$ by their MSE, since MSE is a quadratic function of c_1 and c_2.

To evaluate the difference of MSE between the estimators, we need the following lemmas. The following Lemma 2.1 is well known and we can show it by applying integration by parts (Berger (1980)).

LEMMA 2.1. *Let X be a Gamma(α, λ) random variable and assume that $g(x)$ is absolutely continuous on $(0, \infty)$ with $g'(x) = \frac{dg(x)}{dx}$ satisfying*

(i) $E[|Xg'(X)|] < \infty \, and \, E[|g(X)|] < \infty$

(ii) $\lim_{x \to 0} g(x) x^\alpha e^{-x/\lambda} = \lim_{x \to \infty} g(x) x^\alpha e^{-x/\lambda} = 0$, *for* $\lambda > 0$.

Then

$$E[Xg(X)] = \lambda \left\{ \alpha E\left[g(X)\right] + E\left[Xg'(X)\right] \right\}.$$

LEMMA 2.2. *Let X_i, $i = 1, 2$ be independent Gamma(α_i, λ_i) random variables having density* (2.1). *For any constant $b \geq 0$, $I_{x_1 \geq bx_2}$ denotes indicator function of the set $\{(x_1, x_2) \mid x_1 \geq bx_2\}$ and $\rho = b/(b+1)$. Then*

$$\frac{E[X_2 I_{X_1 \geq bX_2}]}{E[X_1 I_{X_1 \geq bX_2}]} \geq \frac{E_0[X_2 I_{X_1 \geq bX_2}]}{E_0[X_1 I_{X_1 \geq bX_2}]} = \frac{\alpha_1 + \alpha_2}{\alpha_1} \frac{1 - I_\rho(\alpha_1, \alpha_2)}{1 - I_\rho(\alpha_1 + 1, \alpha_2)} - 1,$$

where $E_0[\cdot]$ denotes the expectation when $\lambda_1 = \lambda_2$ and $I_x(\alpha, \beta) = \int_0^x u^{\alpha-1}(1-u)^{\beta-1} du / B(\alpha, \beta)$, where $B(\alpha, \beta)$ is the beta function.

The proof is rather technical and we give it in Appendix A.1. We note that $E_0[X_2 I_{X_1 \geq bX_2}]/E_0[X_1 I_{X_1 \geq bX_2}]$ is independent of the common value of λ_1 and λ_2.

3. MSE reduction by MLE in estimating linear functions of Gamma scale parameters

Here we evaluate the difference of MSE between MLE and UE in estimating $c_1\lambda_1 + c_2\lambda_2$, where c_1, c_2 are constants. The difference of squared errors between MLE and UE is given by

$$(3.1)\quad \left\{\sum_{i=1}^{2} c_i\left(\frac{X_i}{\alpha_i}-\lambda_i\right)\right\}^2-\left\{\sum_{i=1}^{2} c_i\left(\frac{X_i}{\alpha_i}-\lambda_i\right)-\frac{(\alpha_2X_1-\alpha_1X_2)^+}{\alpha_1+\alpha_2}\left(\frac{c_1}{\alpha_1}-\frac{c_2}{\alpha_2}\right)\right\}^2$$
$$=\left(\frac{\tilde{c}_1-\tilde{c}_2}{\alpha_1+\alpha_2}\right)\left\{2\sum_{i=1}^{2}\tilde{c}_i(X_i-\alpha_i\lambda_i)(\alpha_2X_1-\alpha_1X_2)^+ -\left(\frac{\tilde{c}_1-\tilde{c}_2}{\alpha_1+\alpha_2}\right)\left[(\alpha_2X_1-\alpha_1X_2)^+\right]^2\right\},$$

where $\tilde{c}_i=c_i/\alpha_i$, $i=1,2$. Without loss of generality we assume that $\tilde{c}_1\geq\tilde{c}_2$ and also for simplicity we denote $I_{\alpha_2X_1\geq\alpha_1X_2}$ by I, hereafter.

To evaluate the expected value of (3.1) we use Lemma 2.1 and have

$$E[X_1(\alpha_2X_1-\alpha_1X_2)^+]=\lambda_1\{\alpha_1E[(\alpha_2X_1-\alpha_1X_2)^+]+E[\alpha_2X_1I]\}$$

and

$$(3.2)\quad E[X_2(\alpha_2X_1-\alpha_1X_2)^+]=\lambda_2\{\alpha_2E[(\alpha_2X_1-\alpha_1X_2)^+]-E[\alpha_1X_2I]\}.$$

Thus we see that the expected value of the quantity in the braces of (3.1) is given by

$$(3.3)\quad 2\tilde{c}_1\lambda_1E[\alpha_2X_1I]-2\tilde{c}_2\lambda_2E[\alpha_1X_2I] -\left(\frac{\tilde{c}_1-\tilde{c}_2}{\alpha_1+\alpha_2}\right)\{\alpha_1\alpha_2(\lambda_1-\lambda_2)E[(\alpha_2X_1-\alpha_1X_2)^+] +\alpha_2\lambda_1E[\alpha_2X_1I]+\alpha_1\lambda_2E[\alpha_1X_2I]\}.$$

We first show that (3.3) is negative for sufficiently large λ_2 if $\tilde{c}_2>0$. Since the third term in (3.3) is non-positive we see from Lemma 2.2 that (3.3) is less than or equal to

$$2E[\alpha_2X_1I]\left\{\tilde{c}_1\lambda_1-\tilde{c}_2\lambda_2\frac{\alpha_1}{\alpha_2}\frac{E_0[X_2\mid\alpha_2X_1\geq\alpha_1X_2]}{E_0[X_1\mid\alpha_2X_1\geq\alpha_1X_2]}\right\},$$

which is negative for sufficiently large λ_2 if $\tilde{c}_2>0$. This means that MLE does not improve UE if $\tilde{c}_2>0$. Thus we see that $\tilde{c}_2$ must be non-positive when $\tilde{c}_1\geq\tilde{c}_2$ in order for MLE to dominate UE. In addition to the condition $\tilde{c}_1\geq\tilde{c}_2$ we assume that $\tilde{c}_2\leq 0$ in the following and give a condition on c_1 and c_2 for MLE to dominate UE.

Since (3.2) is non-negative, we have

$$\alpha_2E[(\alpha_2X_1-\alpha_1X_2)^+]\geq E[\alpha_1X_2I],$$

and we see that (3.3) is greater than or equal to

$$(3.4)\quad 2\tilde{c}_1\lambda_1E[\alpha_2X_1I]-2\tilde{c}_2\lambda_2E[\alpha_1X_2I] -\left(\frac{\tilde{c}_1-\tilde{c}_2}{\alpha_1+\alpha_2}\right)\{\alpha_1(\lambda_1-\lambda_2)E[\alpha_1X_2I] +\alpha_2\lambda_1E[\alpha_2X_1I]+\alpha_1\lambda_2E[\alpha_1X_2I]\}$$
$$=\lambda_1\left(2\tilde{c}_1-\frac{(\tilde{c}_1-\tilde{c}_2)\alpha_2}{\alpha_1+\alpha_2}\right)E[\alpha_2X_1I]$$

$$\begin{aligned}&-\left(\frac{\tilde{c}_1-\tilde{c}_2}{\alpha_1+\alpha_2}\alpha_1\lambda_1+2\tilde{c}_2\lambda_2\right)E[\alpha_1X_2I]\\ \geq\ &\frac{\lambda_1}{\alpha_1+\alpha_2}\{(\tilde{c}_1(2\alpha_1+\alpha_2)+\tilde{c}_2\alpha_2)E[\alpha_2X_1I]\\ &\qquad-(\tilde{c}_1\alpha_1+\tilde{c}_2(\alpha_1+2\alpha_2))E[\alpha_1X_2I]\},\end{aligned}$$

since $\lambda_2 \geq \lambda_1$ and $\tilde{c}_2 \leq 0$. We can easily see that if $\tilde{c}_1(2\alpha_1+\alpha_2)+\tilde{c}_2\alpha_2 \geq 0$, then (3.4) is non-negative since $E[\alpha_2X_1I] \geq E[\alpha_1X_2I]$ and $\tilde{c}_1 > \tilde{c}_2$. Even if $\tilde{c}_1(2\alpha_1+\alpha_2)+\tilde{c}_2\alpha_2 < 0$, (3.4) is non-negative if

$$\frac{E[X_2I]}{E[X_1I]} \geq \frac{\alpha_2}{\alpha_1}\frac{(c_1/c_2)(2+\alpha_2/\alpha_1)+1}{(c_1/c_2)+(2+\alpha_1/\alpha_2)}. \tag{3.5}$$

We note that for fixed α_1 and α_2, the right-hand side of (3.5) is an increasing function of c_1/c_2. Thus we see that for fixed α_1 and α_2 if some c_1 and c_2 satisfy (3.5) then any c_1' and c_2' such that $c_1/c_2 > c_1'/c_2'$ satisfy (3.5).

Putting $R = \{1 - I_\rho(\alpha_1+1,\alpha_2)\}/\{1-I_\rho(\alpha_1,\alpha_2)\}$, we have from Lemma 2.2

$$\frac{E_0[X_2I]}{E_0[X_1I]} = \frac{\alpha_1+\alpha_2}{\alpha_1}\frac{1}{R}-1,$$

where $\rho = \frac{\alpha_1}{\alpha_1+\alpha_2}$. Thus the inequality (3.5) is true if

$$\frac{\alpha_1+\alpha_2}{\alpha_1}\frac{1}{R}-1 \geq \frac{\alpha_2}{\alpha_1}\frac{c_1(2+\alpha_2/\alpha_1)+c_2}{c_1+(2+\alpha_1/\alpha_2)c_2},$$

which is equivalent to

$$R \leq \frac{c_1(1-\rho)+c_2(2-\rho)}{c_1\dfrac{1-\rho}{\rho}+c_2}.$$

The above inequality is also equivalent to the one

$$\frac{c_1}{c_2} \leq \frac{\rho}{1-\rho}\frac{2-\rho-R}{R-\rho}.$$

It should be noted that $R \geq 1 > \rho$, since $I_\rho(\alpha_1+1,\alpha_2) < I_\rho(\alpha_1,\alpha_2)$.

Thus we have shown that MLE dominates UE if c_1 and c_2 satisfy $c_1/c_2 \leq \rho(2-\rho-R)/\{(1-\rho)(R-\rho)\}$. Conversely, we see that this conditions is also necessary for MLE to dominate UE by examining each step of the above evaluation for the case $\lambda_1 = \lambda_2$. If we denote the MSE of an estimator φ of $\sum_{i=1}^2 c_i\lambda_i$ by $MSE(\varphi)$, we have the following theorem.

THEOREM 3.1. *$MSE(\sum_{i=1}^2 c_iX_i/\alpha_i) \geq MSE(\sum_{i=1}^2 c_i\hat{\lambda}_i)$ for any $0 < \lambda_1 \leq \lambda_2 < \infty$ if and only if*

$$\frac{c_1}{c_2} \leq \frac{\rho}{1-\rho}\frac{2-\rho-R}{R-\rho} \tag{3.6}$$

including the case $c_2 = 0$.

Table 1. Upper bounds of c_1/c_2.

$\alpha_1\backslash\alpha_2$	0.1	0.3	0.5	0.8	1	1.5	2	2.5	3	5	8	12	100
0.1	−0.277	−0.152	−0.104	−0.070	−0.057	−0.039	−0.030	−0.024	−0.020	−0.012	−0.008	−0.005	−0.001
0.3	−0.275	−0.188	−0.142	−0.103	−0.086	−0.062	−0.048	−0.039	−0.033	−0.021	−0.013	−0.009	−0.001
0.5	−0.208	−0.153	−0.120	−0.090	−0.077	−0.057	−0.045	−0.037	−0.031	−0.020	−0.013	−0.008	−0.001
0.8	−0.086	−0.066	−0.054	−0.042	−0.036	−0.027	−0.022	−0.018	−0.016	−0.010	−0.006	−0.004	−0.001
1	0	0	0	0	0	0	0	0	0	0	0	0	0
1.5	0.220	0.178	0.149	0.120	0.106	0.082	0.067	0.056	0.049	0.031	0.021	0.014	0.002
2	0.444	0.364	0.308	0.250	0.222	0.174	0.143	0.121	0.105	0.069	0.045	0.031	0.004
2.5	0.670	0.552	0.471	0.385	0.344	0.271	0.224	0.191	0.167	0.110	0.073	0.050	0.006
3	0.896	0.743	0.636	0.524	0.469	0.372	0.309	0.264	0.231	0.154	0.102	0.071	0.009
5	1.803	1.512	1.306	1.089	0.981	0.790	0.662	0.571	0.502	0.340	0.230	0.161	0.021
8	3.167	2.673	2.322	1.949	1.764	1.432	1.210	1.049	0.928	0.637	0.436	0.308	0.042
12	4.987	4.225	3.681	3.103	2.815	2.298	1.949	1.697	1.505	1.044	0.721	0.512	0.071
100	45.043	38.403	33.658	28.586	26.058	21.483	18.386	16.131	14.406	10.216	7.220	5.244	0.797

When $c_2 = 0 (c_1 = 0)$ and $\alpha_1 = \alpha_2$ is a positive integer, the above theorem reduces to Theorem 2.1. (a) (Theorem 2.2. (a)) due to Kaur and Singh (1991). See Kushary and Cohen (1989) for another improving estimator of smaller variance and also Hwang and Peddada (1994) for related results.

We have calculated the values of the right-hand side of (3.6) for some typical values of α_1 and α_2 and have given them in Table 1. We see that the range of the value of c_1/c_2 for which MLE dominates UE is rather small. Especially when we are concerned with the case with positive coefficients it is quite small. If $\alpha_1 = \alpha_2 = 2$, we need $c_1/c_2 \leq 0.143$ and MLE does not dominate UE for most of the choice of coefficients with the same sign. We notice that the range of c_1/c_2 for which MLE dominates UE becomes larger if α_1 or α_2 gets larger. Rather than α_2, α_1 seems to be important to make the range larger.

The case when $c_1 = 0$ corresponds to the estimation of λ_2 and is of particular interest. From Table 1 it is almost obvious that MLE dominates UE for $c_1 = 0$ if and only if $\alpha_1 \geq 1$. We formally give it in the following corollary whose proof is given in Appendix A.2.

Corollary 3.1. *$MSE(\sum_{i=1}^{2} c_i X_i/\alpha_i) \geq MSE(\sum_{i=1}^{2} c_i \hat{\lambda}_i)$ for any $0 < \lambda_1 \leq \lambda_2 < \infty$ and for any $c_1 \geq 0$ and $c_2 \leq 0$ (and also for any $c_1 \leq 0$ and $c_2 \geq 0$) if and only if $\alpha_1 \geq 1$.*

4. MSE reduction of an admissible estimator based solely on one sample

In this section, we compare two estimators of $c_1\lambda_1 + c_2\lambda_2$, $\sum_{i=1}^{2} c_i X_i/(\alpha_i + 1)$ and $\sum_{i=1}^{2} c_i \tilde{\lambda}_i$, by their mean square errors and give a condition on c_1 and c_2 for the latter to dominate the former.

The difference of squared errors between $\sum_{i=1}^{2} c_i X_i/(\alpha_i + 1)$ and $\sum_{i=1}^{2} c_i \tilde{\lambda}_i$ is given by

$$\left(\frac{\tilde{c}'_1 - \tilde{c}'_2}{\alpha_1 + \alpha_2 + 2}\right)\left\{2\sum_{i=1}^{2} \tilde{c}'_i (X_i - (\alpha_i + 1)\lambda_i)[(\alpha_2 + 1)X_1 - (\alpha_1 + 1)X_2]^+ \right.$$
$$\left. - \left(\frac{\tilde{c}'_1 - \tilde{c}'_2}{\alpha_1 + \alpha_2 + 2}\right)\left\{[(\alpha_2 + 1)X_1 - (\alpha_1 + 1)X_2]^+\right\}^2 \right\}, \tag{4.1}$$

where $\tilde{c}_i' = c_i/(\alpha_i+1), i = 1,2$. Without loss of generality we assume that $\tilde{c}_1' \geq \tilde{c}_2'$, and also for simplicity we denote $I_{(\alpha_2+1)X_1 \geq (\alpha_1+1)X_2}$ by I', hereafter. By applying Lemma 2.1 we see that the expected value in the braces of (4.1) is given by

$$\begin{aligned}(4.2)\quad & 2\tilde{c}_1'\lambda_1 E[(\alpha_1+1)X_2I'] - 2\tilde{c}_2'\lambda_2 E[(\alpha_2+1)X_1I'] \\ & -\left(\frac{\tilde{c}_1' - \tilde{c}_2'}{\alpha_1+\alpha_2+2}\right)\{(\alpha_2+1)\lambda_1 E[(\alpha_1+1)X_2I'] + (\alpha_1+1)\lambda_2 E[(\alpha_2+1)X_1I'] \\ & \qquad +(\alpha_1+1)(\alpha_2+1)(\lambda_1-\lambda_2)E[\{(\alpha_2+1)X_1 - (\alpha_1+1)X_2\}^+]\}.\end{aligned}$$

Here we notice that (4.2) is negative for sufficiently large λ_2 if $\tilde{c}_2' > 0$, since the third term in (4.2) is non-positive and $E[(\alpha_2+1)X_1I'] \geq E[(\alpha_1+1)X_2I']$. This implies that $\sum_{i=1}^2 c_i\tilde{\lambda}_i$ does not dominate $\sum_{i=1}^2 c_iX_i/(\alpha_i+1)$ if $\tilde{c}_1' \geq \tilde{c}_2'$ and $\tilde{c}_2' > 0$ (or $\tilde{c}_1' \leq \tilde{c}_2'$ and $\tilde{c}_2' < 0$). Therefore in the following we only consider the case where $\tilde{c}_1 \geq \tilde{c}_2$ and $\tilde{c}_2 \leq 0$ to find the conditions on $\tilde{c}_1'$ and $\tilde{c}_2'$ for $\sum_{i=1}^2 c_i\tilde{\lambda}_i$ to dominate $\sum_{i=1}^2 c_iX_i/(\alpha_i+1)$.

We first show that (4.2) is non-negative if $(\alpha_1+1)\tilde{c}_1' + (\alpha_1+2\alpha_2+3)\tilde{c}_2' \leq 0$. Since $\tilde{c}_1' \geq \tilde{c}_2'$ and $\lambda_1 \leq \lambda_2$, (4.2) is greater than or equal to

$$\begin{aligned}(4.3)\qquad & \left(\frac{(2\alpha_1+\alpha_2+3)\tilde{c}_1' + (\alpha_2+1)\tilde{c}_2'}{\alpha_1+\alpha_2+2}\right)\lambda_1 E[(\alpha_1+1)X_2I'] \\ & -\left(\frac{(\alpha_1+1)\tilde{c}_1' + (\alpha_1+2\alpha_2+3)\tilde{c}_2'}{\alpha_1+\alpha_2+2}\right)\lambda_2 E[(\alpha_2+1)X_1I'].\end{aligned}$$

Since $\lambda_1 \leq \lambda_2$, $E[(\alpha_1+1)X_2I'] \leq E[(\alpha_2+1)X_1I']$ and $(2\alpha_1+\alpha_2+3)\tilde{c}_1' + (\alpha_2+1)\tilde{c}_2' \geq (\alpha_1+1)\tilde{c}_1' + (\alpha_1+2\alpha_2+3)\tilde{c}_2'$, we see that (4.2) is non-negative if $(\alpha_1+1)\tilde{c}_1' + (\alpha_1+2\alpha_2+3)\tilde{c}_2' \leq 0$.

In the following we assume that $(\alpha_1+1)\tilde{c}_1' + (\alpha_1+2\alpha_2+3)\tilde{c}_2' > 0$. Using the inequality

$$E[\{(\alpha_2+1)X_1 - (\alpha_1+1)X_2\}^+] \geq E[X_1I'],$$

we see that (4.2) is greater than or equal to

$$\begin{aligned}(4.4)\qquad & \frac{\lambda_1}{\alpha_1+\alpha_2+2}\{[(2\alpha_1+\alpha_2+3)\tilde{c}_1' + (\alpha_2+1)\tilde{c}_2']E[(\alpha_1+1)X_2I'] \\ & -[(\alpha_1+1)\tilde{c}_1' + (\alpha_1+2\alpha_2+3)\tilde{c}_2']E[(\alpha_2+1)X_1I']\}.\end{aligned}$$

(4.4) is non-negative if and only if

$$(4.5)\qquad \frac{E[X_2I']}{E[X_1I']} \geq \frac{\alpha_2+1}{\alpha_1+1}\frac{c_1 + \{2+(\alpha_1+1)/(\alpha_2+1)\}c_2}{\{2+(\alpha_2+1)/(\alpha_1+1)\}c_1 + c_2}.$$

Now we denote $\rho' = (\alpha_1+1)/(\alpha_1+\alpha_2+2)$ and $R' = (1 - I_{\rho'}(\alpha_1+1,\alpha_2))/(1 - I_{\rho'}(\alpha_1,\alpha_2))$. Then from Lemma 2.2, we see that the inequality (4.5) is true if

$$(4.6)\qquad \frac{\alpha_1+\alpha_2}{\alpha_1R'} \geq \frac{2(c_1+c_2)}{(\rho'+1)c_1 + \rho'c_2}.$$

Since the right-hand side of (4.6) is a decreasing function of c_1/c_2, we see that if $\alpha_1R'/(\alpha_1+\alpha_2) \leq (\rho'+1)/2$, then $\sum_{i=1}^2 c_i\tilde{\lambda}_i$ dominates $\sum_{i=1}^2 c_iX_i/(\alpha_i+1)$ for any c_1 and c_2 such that $-\infty < c_1/c_2 \leq (\alpha_1+1)/(\alpha_2+1)$ including the case $c_2 = 0$. Similarly

Table 2. Lower bounds of c_1/c_2. (A blank means that lower bound does not exist).

$\alpha_1\backslash\alpha_2$	0.1	0.3	0.5	0.8	1	1.5	2	2.5
0.1	−3.610	−5.315	−9.856					
0.3	−3.784	−5.312	−8.806	−302.000				
0.5	−3.966	−5.384	−8.320	−43.602				
0.8	−4.251	−5.564	−8.037	−23.911				
1	−4.444	−5.714	−8.000	−20.000				
1.5	−4.938	−6.145	−8.159	−16.197	−47.793			
2	−5.438	−6.619	−8.500	−15.000	−31.000			
2.5	−5.943	−7.118	−8.929	−14.654	−25.934			
3	−6.450	−7.630	−9.407	−14.692	−23.800			
5	−8.490	−9.746	−11.541	−16.242	−22.594			
8	−11.565	−12.993	−14.972	−19.779	−25.500	−105.36		
12	−15.673	−17.363	−19.663	−25.021	−30.974	−83.649		
100	−106.18	−114.15	−124.71	−147.56	−170.11	−291.02	−1292.8	

in the case when $\alpha_1 R'/(\alpha_1+\alpha_2) > (\rho'+1)/2$, $\sum_{i=1}^2 c_i\tilde{\lambda}_i$ dominates $\sum_{i=1}^2 c_i X_i/(\alpha_i+1)$ for any c_1 and c_2 such that $\{(\alpha_1+\alpha_2)\rho' - 2\alpha_1 R'\}/\{2\alpha_1 R' - (\alpha_1+\alpha_2)(\rho'+1)\} < c_1/c_2 \le (\alpha_1+1)/(\alpha_2+1)$.

By examining each step of the above evaluation for the case $\lambda_1 = \lambda_2$ we see that this condition is also necessary. Thus we have shown the following theorem.

Theorem 4.1. *$MSE(\sum_{i=1}^2 c_i X_i/(\alpha_i+1)) \ge MSE(\sum_{i=1}^2 c_i\tilde{\lambda}_i)$ for any $0 < \lambda_1 \le \lambda_2 < \infty$ if and only if*

$$\frac{\rho' - 2\rho R'}{2\rho R' - (\rho'+1)} \le \frac{c_1}{c_2} \le \frac{\alpha_1+1}{\alpha_2+1} \quad \textit{when} \quad 2\rho R' > \rho' + 1$$

and

$$-\infty < \frac{c_1}{c_2} \le \frac{\alpha_1+1}{\alpha_2+1} \quad \textit{when} \quad 2\rho R' \le \rho' + 1$$

including the case $c_2 = 0$.

We have calculated the lower bounds of c_1/c_2 if they exist for some typical values of α_1 and α_2 and have given them in Table 2.

The case when $c_2 = 0$ corresponds to the estimation of λ_1 and $\tilde{\lambda}_1$ dominates $X_1/(\alpha_1+1)$ if and only if $2\rho R' \le \rho' + 1$. Although it seems clear from Table 2 for what values of α_1 and α_2 this condition is satisfied, we give the following analytical result which is not the best possible in any sense.

Corollary 4.1. *$MSE(X_1/(\alpha_1+1)) \ge MSE(\tilde{\lambda}_1)$ for any $0 < \lambda_1 \le \lambda_2 < \infty$ if $\alpha_1 \le \alpha_2$ and $\alpha_2 \ge 1$.*

The proof is given in Appendix A.3.

From Table 2 it seems that $\alpha_2 \ge 2.5$ is sufficient for $\tilde{\lambda}_1$ to dominate $X_1/(\alpha_1+1)$ for any α_1, although by Corollary 4.1 we show that $\tilde{\lambda}_1$ dominates $X_1/(\alpha_1+1)$ if $\alpha_1 \le \alpha_2$ and $\alpha_2 \ge 1$. The range of positive coefficients for which $\sum_{i=1}^2 c_i\tilde{\lambda}_i$ dominates $\sum_{i=1}^2 c_i X_i/(\alpha_i+1)$ is completely determined by the ratio $(\alpha_1+1)/(\alpha_2+1)$. If $(\alpha_1+1)/(\alpha_2+1)$ gets larger, the range gets larger. Thus if α_1 is large compared with α_2 we can get the uniform improvement for wide range of positive coefficients.

5. Concluding remarks

A comparison of the results given by Theorems 3.1 and 4.1 (or Tables 1 and 2) may be in order. Although we cannot give clear explanation, we will also point out possible reason of the difference of the two regions of c_1/c_2.

(i) For any c_1 and c_2 with opposite sign both MLE and modified MLE give uniform improvement over their competitors except for the case when α_1 is quite small (in case of MLE) or α_2 is quite small (in case of modified MLE). This implies that we can use these estimators safely to estimate between component of variance in a one-way random effects model.

(ii) Both MLE and modified MLE have larger MSE than their competitors for larger c_1/c_2 ($c_1/c_2 > \alpha_1/\alpha_2$ in case of MLE and $c_1/c_2 > (\alpha_1+1)/(\alpha_2+1)$ in case of modified MLE) when λ_2/λ_1 is sufficient large.

(iii) MLE has larger MSE than UE for the case $\lambda_1 = \lambda_2$ if $\rho(2-\rho-R)/\{(1-\rho)(R-\rho)\} < c_1/c_2 < \alpha_1/\alpha_2$. We note that MLE expands UE in this case, but this does not explain the possible improvement for the case $c_1 = 0$.

(iv) Modified MLE has larger MSE than its competitor for the case $\lambda_1 = \lambda_2$ if $-\infty \le c_1/c_2 < (\rho' - 2\rho R')/\{2\rho R' - (\rho'+1)\}$ when $2\rho R' > \rho' + 1$. We note that modified MLE shrinks $\sum_{i=1}^2 c_i X_i/(\alpha+1)$ although $X_i/(\alpha_i+1)$ itself is a shrinkage of the UE X_i/α_i.

Next, we give some results on the comparison of the two estimators $\sum_{i=1}^2 c_i\hat{\lambda}_i$ and $\sum_{i=1}^2 c_i\tilde{\lambda}_i$ without proof. We have restricted ourselves to the case $\alpha_1 = \alpha_2 = \alpha$ because of a technical difficulty in evaluating the risk difference by the same sort of calculations given in Sections 3 and 4.

(i) $MSE(\sum_{i=1}^2 c_i\hat{\lambda}_i) \ge MSE(\sum_{i=1}^2 c_i\tilde{\lambda}_i)$ for any $0 < \lambda_1 \le \lambda_2$ if $|c_1/c_2| \le 1$.

(ii) For $\lambda_1 = \lambda_2, MSE(\sum_{i=1}^2 c_i\hat{\lambda}_i) < MSE(\sum_{i=1}^2 c_i\tilde{\lambda}_i)$ if and only if

$$\{-(4\alpha^2+2\alpha-1)c_1^2 - 2(2\alpha-1)c_1c_2 + (4\alpha^2+6\alpha+5)c_2^2\}$$
$$+\frac{E_0[X_2I]}{E_0[X_1I]}\{(4\alpha^2+6\alpha+5)c_1^2 - 2(2\alpha-1)c_1c_2 - (4\alpha^2+2\alpha-1)c_2^2\} < 0.$$

In particular $MSE(\hat{\lambda}_1) < MSE(\tilde{\lambda}_1)$ for $\lambda_1 = \lambda_2$ if and only if $E_0(X_2I)/E_0(X_1I) < (4\alpha^2+2\alpha-1)/(4\alpha^2+6\alpha+5)$. By numerical evaluation we have found that this inequality is satisfied for $\alpha_1 = \alpha_2 > 1$. Thus we see that $\sum_{i=1}^2 c_i\tilde{\lambda}_i$ does not improve $\sum_{i=1}^2 c_i\hat{\lambda}_i$ if $|c_1/c_2|$ is sufficiently large and $\alpha_1 = \alpha_2$ is moderately large.

(iii) For any c_1 and c_2, $MSE(\sum_{i=1}^2 c_i\hat{\lambda}_i) > MSE(\sum_{i=1}^2 c_i\tilde{\lambda}_i)$ if λ_1/λ_2 is sufficiently small. Thus $\sum_{i=1}^2 c_i\hat{\lambda}_i$ does not improve $\sum_{i=1}^2 c_i\tilde{\lambda}_i$ for any c_1 and c_2.

Finally, we should mention the case of more than two populations. In case of two populations we have partitioned the sample space into two subregions and have given the expressions of the estimators. Even in case of three populations we have to partition the sample space into six subregions and the expressions of the estimators become much more complicated. Although we believe that the technique used in this paper will be useful, we have not succeeded in obtaining explicit results unfortunately.

Appendix

A.1. *Proof of Lemma* 2.2.

Let

$$W = \frac{X_1}{\lambda_1} + \frac{X_2}{\lambda_2} \quad \text{and} \quad Z = \frac{\frac{X_1}{\lambda_1}}{\frac{X_1}{\lambda_1} + \frac{X_2}{\lambda_2}}.$$

Then W and Z are independent random variables having $Gamma(\alpha_1+\alpha_2, 1)$ distribution and $Beta(\alpha_1, \alpha_2)$ one, respectively. The random variables X_1 and X_2 can be expressed as

$$X_1 = \lambda_1 WZ, \quad \text{and} \quad X_2 = \lambda_2 W(1-Z)$$

respectively.

We first note that $X_1 \geq bX_2$ if and only if $Z \geq b\lambda_2/(b\lambda_2+\lambda_1)$. If we set $\gamma = b\lambda_2/(b\lambda_2+\lambda_1)$, we see that $\lambda_1 \leq \lambda_2$ if and only if $\gamma \geq b/(b+1)$.

Thus we have

$$\begin{aligned} E[bX_2 \mid X_1 \geq bX_2] &= b\lambda_2 E[W(1-Z) \mid Z \geq \gamma] \\ &= (\alpha_1+\alpha_2) b\lambda_2 E[1-Z \mid Z \geq \gamma] \quad \text{and} \\ E[X_1 \mid X_1 \geq bX_2] &= \lambda_1 E[WZ \mid Z \geq \gamma] \\ &= (\alpha_1+\alpha_2)(b\lambda_2+\lambda_1)(1-\gamma) E[Z \mid Z \geq \gamma]. \end{aligned}$$

Therefore

$$\frac{E[bX_2 I_{X_1 \geq bX_2}]}{E[X_1 I_{X_1 \geq bX_2}]} = \frac{E[bX_2 \mid X_1 \geq bX_2]}{E[X_1 \mid X_1 \geq bX_2]} = \frac{\gamma}{1-\gamma} \frac{E[1-Z \mid Z \geq \gamma]}{E[Z \mid Z \geq \gamma]} \equiv T(\gamma).$$

Since we show that $T(\gamma)$ is an increasing function of γ it is minimal when $\gamma = b/(b+1)$ or $\lambda_1 = \lambda_2$ and

$$\frac{E[bX_2 \mid X_1 \geq bX_2]}{E[X_1 \mid X_1 \geq bX_2]} \geq \frac{E_0[bX_2 \mid X_1 \geq bX_2]}{E_0[X_1 \mid X_1 \geq bX_2]} = b\frac{E[1-Z \mid Z \geq \rho]}{E[Z \mid Z \geq \rho]}.$$

Since Z is random variable with Beta distribution $Beta(\alpha_1, \alpha_2)$, we have

$$\begin{aligned} E[Z \mid Z \geq \rho] &= \frac{\int_\rho^1 z^{\alpha_1}(1-z)^{\alpha_2-1} dz}{\int_\rho^1 z^{\alpha_1-1}(1-z)^{\alpha_2-1} dz} \\ &= \frac{B(\alpha_1+1, \alpha_2)}{B(\alpha_1, \alpha_2)} \frac{\frac{1}{B(\alpha_1+1,\alpha_2)} \int_\rho^1 z^{\alpha_1+1-1}(1-z)^{\alpha_2-1} dz}{\frac{1}{B(\alpha_1,\alpha_2)} \int_\rho^1 z^{\alpha_1-1}(1-z)^{\alpha_2-1} dz} \\ &= \frac{\alpha_1}{\alpha_1+\alpha_2} \frac{1 - I_\rho(\alpha_1+1, \alpha_2)}{1 - I_\rho(\alpha_1, \alpha_2)}. \end{aligned}$$

To show that $T(\gamma)$ is an increasing function of γ, we express it as

$$T(\gamma) = \frac{\displaystyle\int_\gamma^1 \frac{\gamma}{1-\gamma} z^{\alpha_1 - 1}(1-z)^{\alpha_2} dz}{\displaystyle\int_\gamma^1 z^{\alpha_1}(1-z)^{\alpha_2 - 1} dz}.$$

In both integrals we make the change of variable $v = (1-z)/(1-\gamma)$ and have

$$T(\gamma) = \frac{\displaystyle\int_0^1 \frac{\gamma v}{1-(1-\gamma)v}\{1-(1-\gamma)v\}^{\alpha_1} v^{\alpha_2 - 1} dv}{\displaystyle\int_0^1 \{1-(1-\gamma)v\}^{\alpha_1} v^{\alpha_2 - 1} dv}.$$

If we put

$$f(v;\gamma) = \frac{\{1-(1-\gamma)v\}^{\alpha_1} v^{\alpha_2 - 1}}{\displaystyle\int_0^1 \{1-(1-\gamma)v\}^{\alpha_1} v^{\alpha_2 - 1} dv}$$

then $f(v;\gamma)$ is a density function with parameter γ, and $T(\gamma)$ is the expected value of $\varphi(v;\gamma) = \frac{\gamma v}{1-(1-\gamma)v}$, and we denote it as $E_\gamma[\varphi(V;\gamma)]$. We show that $f(v;\gamma)$ has monotone likelihood ratio in v. Suppose that $\gamma > \gamma'$. Then

$$\frac{f(v;\gamma)}{f(v;\gamma')} \sim \left(\frac{1-(1-\gamma)v}{1-(1-\gamma')v}\right)^{\alpha_1}$$

is an increasing function of v. Furthermore, since $\varphi(v;\gamma)$ is an increasing function of γ, we have

$$T(\gamma) = E_\gamma[\varphi(V;\gamma)] \geq E_{\gamma'}[\varphi(V;\gamma)] > E_{\gamma'}[\varphi(V;\gamma')] = T(\gamma').$$

This completes the proof.

A.2. *Proof of Corollary* 3.1.

From Theorem 3.1 we see that it is enough for us to show that

$$\text{(A.1)} \qquad \frac{\rho}{1-\rho}\frac{2-\rho-R}{R-\rho} \geq 0$$

or $R \leq 2-\rho$ if and only if $\alpha_1 \geq 1$, where $\rho = \alpha_1/(\alpha_1+\alpha_2)$ and $R = \{1 - I_\rho(\alpha_1 + 1, \alpha_2)\}/\{1 - I_\rho(\alpha_1, \alpha_2)\}$.

By applying an integration by parts we can easily show that

$$I_\rho(\alpha_1 + 1, \alpha_2) = I_\rho(\alpha_1, \alpha_2) - \frac{\rho^{\alpha_1}(1-\rho)^{\alpha_2}}{(\alpha_1+\alpha_2)B(\alpha_1+1,\alpha_2)}.$$

Thus we see that (A.1) is equivalent to

$$\text{(A.2)} \qquad \frac{1}{1-I_\rho(\alpha_1,\alpha_2)}\frac{\rho^{\alpha_1}(1-\rho)^{\alpha_2}}{(\alpha_1+\alpha_2)B(\alpha_1+1,\alpha_2)} \leq \frac{\alpha_2}{\alpha_1+\alpha_2}.$$

We note that

$$
\begin{aligned}
&(\alpha_1+\alpha_2)B(\alpha_1+1,\alpha_2)\{1-I_\rho(\alpha_1,\alpha_2)\}\\
&\quad=\alpha_1\int_\rho^1 x^{\alpha_1-1}(1-x)^{\alpha_2-1}dx\\
&\quad=(\alpha_1+\alpha_2)\rho^{\alpha_1}(1-\rho)^{\alpha_2}\int_0^1\left(1+\frac{1-\rho}{\rho}u\right)^{\alpha_1-1}(1-u)^{\alpha_2-1}du,
\end{aligned}
$$

if we make the change of variable

$$\frac{x-\rho}{1-\rho}=u.$$

Therefore we see that (A.2) is equivalent to the condition

$$\text{(A.3)}\qquad \int_0^1\left(1+\frac{1-\rho}{\rho}u\right)^{\alpha_1-1}(1-u)^{\alpha_2-1}du\ge\frac{1}{\alpha_2}.$$

Since $(1+\frac{1-\rho}{\rho}u)^{\alpha_1-1}\ge 1$ if and only if $\alpha_1\ge 1$ and since

$$\int_0^1(1-u)^{\alpha_2-1}du=\frac{1}{\alpha_2},$$

we see that (A.3) is true if and only if $\alpha_1\ge 1$.

A.3. *Proof of Corollary* 4.1.

We need only to show that if $\alpha_2\ge\alpha_1$ and $\alpha_2\ge 1$, then

$$\text{(A.4)}\qquad \rho R'\le(\rho'+1)/2.$$

By the same argument given in Appendix A.2 we can show that the inequality (A.4) is equivalent to the one

$$\text{(A.5)}\qquad \int_0^1\left(1+\frac{1-\rho'}{\rho'}u\right)^{\alpha_1-1}(1-u)^{\alpha_2-1}du\ge\frac{2(\alpha_1+1)}{\alpha_2(\alpha_1+\alpha_2+3)-\alpha_1}.$$

If we express the left-hand side of (A.5) as $\frac{1}{\alpha_2}E[(1+\frac{1-\rho'}{\rho'}U)^{\alpha_1-1}]$, where U is a random variable having Beta distribution $Beta(1,\alpha_2)$, then we see that the inequality (A.5) is equivalent to the one

$$\text{(A.6)}\qquad E\left[\left(1+\frac{1-\rho'}{\rho'}U\right)^{\alpha_1-1}\right]\ge\frac{2\alpha_2(\alpha_1+1)}{\alpha_2(\alpha_1+\alpha_2+3)-\alpha_1}.$$

When $\alpha_1\ge 1$, the left-hand side of (A.6) is greater or equal to 1, and the right-hand side of (A.6) is less than or equal to 1, if $\alpha_2\ge\alpha_1$. When $\alpha_1<1$, we first note that $(1+\frac{1-\rho'}{\rho'}u)^{\alpha_1-1}$ is a decreasing function of u. Thus we see that for $\alpha_2\ge 1$ the left-hand side of (A.6) is minimized when $\alpha_2=1$. Since the right-hand side of (A.6) is a decreasing function of α_2, we need only to show the inequality (A.6) for the case $\alpha_2=1$. In this case it reduces to the one $\{(\alpha_1+3)/(\alpha_1+1)\}^{\alpha_1}\ge\alpha_1+1$ which is true for $0<\alpha_1\le 1$.

Acknowledgements

The authors are grateful to an associate editor and the referees for their helpful comments which improved the presentation of the paper. One referee kindly suggested the application of the results to a one-way random effect model.

References

Barlow, R. E., Bartholomew, D. J., Bremner, J. M. and Brunk, H. D. (1972). *Statistical Inference under Order Restrictions*, Wiley, New York.

Berger, J. (1980). Improving on inadmissible estimators in continuous exponential families with applications to simultaneous estimation of gamma scale parameters, *Ann. Statist.*, **8**, 545–571.

Fernández, M. A., Rueda, C. and Salvador, B. (1999). The loss of efficiency estimating linear functions under restrictions, *Scand. J. Statist.*, **26**, 579–592.

Fernández, M. A., Rueda, C. and Salvador, B. (2000). Parameter estimation under orthant restrictions, *Canad. J. Statist.*, **28**, 171–181.

Hwang, J. T. G. and Peddada, S. D. (1994). Confidence interval estimation subject to order restrictions, *Ann. Statist.*, **22**, 67–93.

Kelly, R. E. (1989). Stochastic reduction of loss in estimating normal means by isotonic regression, *Ann. Statist.*, **17**, 937–940.

Kaur, A. and Singh, H. (1991). On the estimation of ordered means of two exponential populations, *Ann. Inst. Statist. Math.*, **43**, 347–356.

Kubokawa, T. and Saleh, A. K. MD. E. (1994). Estimation of location and scale parameters under order restrictions, *J. Statist. Res.*, **28**, 41–51.

Kushary, D. and Cohen, A. (1989). Estimating ordered location and scale parameters, *Statist. Decisions,* **7**, 201–213.

Kushary, D. and Cohen, A. (1991). Estimation of ordered Poisson parameters, *Sankhyā Ser. A*, **53**, 334–356.

Lee, C. I. C. (1981). The quadratic loss of isotonic regression under normality, *Ann. Statist.*, **9**, 686–688.

Lee, C. I. C. (1988). The quadratic loss of order restricted estimators for several treatment means and a control mean, *Ann. Statist.*, **16**, 751–758.

Robertson, T., Wright, F. T. and Dykstra, R. L. (1988). *Order Restricted Statistical Inference*, Wiley, New York.

Rueda, C. and Salvador, B. (1995). Reduction of risk using restricted estimators, *Comm. Statist. Theory Methods*, **24**(4), 1011–1022.

Shinozaki, N. and Chang, Y.-T. (1999). A comparison of maximum likelihood and best unbiased estimators in the estimation of linear combinations of positive normal means, *Statist. Decisions*, **17**, 125–136.

Ann. Inst. Statist. Math.
Vol. 54, No. 4, 861–878 (2002)

ESTIMATION OF THE COMMON MEAN OF A BIVARIATE NORMAL POPULATION*

PHILIP L. H. YU[1], YIJUN SUN[2]** AND BIMAL K. SINHA[2]

[1] *Department of Statistics and Actuarial Science, The University of Hong Kong, Pokfulam Road, Hong Kong, CHINA*
[2] *Department of Mathematics and Statistics, University of Maryland, Baltimore County, 1000 Hilltop Circle, Baltimore, MD 21228-5398, U.S.A.*

(Received December 14, 2000; revised June 1, 2001)

Abstract. In this paper we discuss the problem of estimating the common mean of a bivariate normal population based on paired data as well as data on one of the marginals. Two double sampling schemes with the second stage sampling being either a simple random sampling (SRS) or a ranked set sampling (RSS) are considered. Two common mean estimators are proposed. It is found that under normality, the proposed RSS common mean estimator is always superior to the proposed SRS common mean estimator and other existing estimators such as the RSS regression estimator proposed by Yu and Lam (1997, *Biometrics*, **53**, 1070–1080). The problem of estimating the mean Reid Vapor Pressure (RVP) of regular gasoline based on field and laboratory data is considered.

Key words and phrases: Ranked set sampling, relative precision, REML, simple random sampling.

1. Introduction

The problem discussed in this paper is motivated by the following practical issue in the context of the attempt by the Environmental Protection Agency (EPA) of the United States to evaluate the gasoline quality based on what is known as Reid Vapor Pressure (RVP). Occasionally, an EPA inspector would visit gas pumps in a city, take samples of gasoline of a particular brand, and measure RVP right at the spot which produces cheap and quick measurements. Once in a while, the inspector after measuring RVP at the spot will ship a gasoline sample to the laboratory for a measurement of presumably higher precision at a higher cost, thus getting the pair (field, lab). Since usually laboratory measurements (Y) are much more expensive than field measurements (X) because of special packaging to be used to ship a gasoline sample from a field to a laboratory, not all the gasoline samples will be shipped to the laboratory and hence the resulting data would consist of many field measurements with occasional paired measurements obtained from both the field and laboratory. Therefore, it never happens at least in our context that we have lab data without field data.

*The research of Philip L. H. Yu was substantially supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region (Project No. HKU7171/97H). The research of Bimal K. Sinha was funded by a grant under 'Presidential Research Professorship' at UMBC.

**Now at Quintiles Inc., 1801 Rockville Pike, Suite 300, Rockville, MD 20852, U.S.A.

As both field measurement X and lab measurement Y are referred to the same chemical (RVP), it is reasonable to assume that the measurements X and Y have the common mean, denoted by μ, but with possibly unequal variances σ^2 and η^2. Moreover, when a paired measurement (X, Y) is observed, it is natural that X and Y are correlated so that (X, Y) is distributed with mean vector $\mu\mathbf{1}_2$ and variance-covariance matrix Σ, where

$$\Sigma = \begin{bmatrix} \sigma^2 & \xi \\ \xi & \eta^2 \end{bmatrix}, \quad \xi = \rho\sigma\eta, \quad \mathbf{1}_2 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

Here, ρ is the correlation coefficient between X and Y. Of course, when only a field measurement X is observed, X is marginally distributed with mean μ and variance σ^2. The goal here is to efficiently estimate the mean RVP μ in gasoline consumed by the public when X and Y follows a bivariate normal distribution.

In practice, a two-phase or double sampling is usually used to collect the above data. This involves the drawing of a random sample of gas pumps in the first phase, in which a crude RVP measurement X is obtained from each gas pump (field); and the drawing of a subsample from the original units in the second phase, in which a more precise RVP measurement Y is obtained from the laboratory. In this case, this is a classical double sampling scheme. Recently, Yu and Lam (1997) demonstrated that the regression estimator is always more efficient when the data are collected using a double sampling with its second-phase sampling being a ranked set sampling (RSS) rather than a simple random sampling (SRS). Therefore, it is worthwhile to consider the problem of point estimation of the common mean μ under two double sampling methods where the first-phase sampling is always simple random sampling and the second-phase sampling is either simple random sampling or ranked set sampling. Hereafter, we refer these two sampling methods as SRS-SRS double sampling and SRS-RSS double sampling.

In this paper, we first consider the case of SRS-SRS double sampling scheme. In Section 2, we discuss the problem of estimating μ when Σ is known. When Σ is unknown, various estimators for Σ are proposed. In Section 3, we discuss the problem of estimating μ when the data are collected using a SRS-RSS double sampling scheme. Other plausible estimators are proposed in Section 4. Numerical comparisons of the relative precision of the proposed common mean estimators under the two sampling schemes and other estimators are discussed in Section 5. We apply the proposed methods to the above practical EPA problem in Section 6. Section 7 gives some concluding remarks.

2. Estimation of μ using SRS-SRS double sampling

In this section we discuss the problem of estimation of μ based on the data collected using a SRS-SRS double sampling scheme. Suppose that a simple random sample of size $n + m$ is drawn in the first phase (field level) and a subsample of size m is drawn in the second phase (lab level). After collecting the measurements at the field and lab, we have two sets of data: the "field only" data $\{z_i, i = 1, \ldots, n\}$, and the paired "(field,lab)" data $\{\boldsymbol{w}_j = (x_j y_j)', j = 1, \ldots, m\}$. They are summarized by a vector $\boldsymbol{t} = (z_1, z_2, \ldots, z_n, x_1, y_1, x_2, y_2, \ldots, x_m, y_m)'$. It is easily seen that the vector $\boldsymbol{t}$ has mean $\mu\mathbf{1}_{n+2m}$ and variance-covariance matrix V, where

$$V = \begin{bmatrix} \sigma^2 \boldsymbol{I}_n & \mathbf{0} \\ \mathbf{0} & \boldsymbol{I}_m \otimes \Sigma \end{bmatrix}.$$

Here $\otimes$ denotes the Kronecker product, $\boldsymbol{I}_m$ and $\boldsymbol{I}_n$ are identity matrices of orders m and n, respectively.

2.1 *Estimation of μ when Σ is known*

When Σ is known, V is also known. Without any distribution assumption, a natural estimator for μ is to use the generalized least squares (GLS) method which minimizes $(\boldsymbol{t}-\mu\mathbf{1}_{n+2m})'V^{-1}(\boldsymbol{t}-\mu\mathbf{1}_{n+2m})$, leading to the GLS estimator $\hat{\mu}_{srs}$:

$$\begin{aligned}\hat{\mu}_{srs} &= \frac{\mathbf{1}'_{n+2m}V^{-1}\boldsymbol{t}}{\mathbf{1}'_{n+2m}V^{-1}\mathbf{1}_{n+2m}}\\ &= \frac{\frac{n}{\sigma^2}\bar{z}+m\mathbf{1}_2'\Sigma^{-1}\bar{\boldsymbol{w}}}{\frac{n}{\sigma^2}+m\mathbf{1}_2'\Sigma^{-1}\mathbf{1}_2}\\ &= \frac{\frac{n}{\sigma^2}\bar{z}+m\left(\frac{\eta^2-\xi}{\sigma^2\eta^2-\xi^2}\bar{x}+\frac{\sigma^2-\xi}{\sigma^2\eta^2-\xi^2}\bar{y}\right)}{\frac{n}{\sigma^2}+m\frac{\sigma^2+\eta^2-2\xi}{\sigma^2\eta^2-\xi^2}}\end{aligned} \tag{2.1}$$

where $\bar{z}=n^{-1}\sum_{i=1}^n z_i$, $\bar{\boldsymbol{w}}=(\bar{x},\bar{y})'$ with $\bar{x}=m^{-1}\sum_{j=1}^m x_j$ and $\bar{y}=m^{-1}\sum_{j=1}^m y_j$.

Obviously $\hat{\mu}_{srs}$ is also the MLE of μ under normality assumption, and is always unbiased with variance

$$\mathrm{Var}(\hat{\mu}_{srs})=\frac{1}{\frac{n}{\sigma^2}+m\mathbf{1}_2{}'\Sigma^{-1}\mathbf{1}_2}=\frac{1}{\frac{n}{\sigma^2}+m\frac{\sigma^2+\eta^2-2\xi}{\sigma^2\eta^2-\xi^2}}. \tag{2.2}$$

2.2 *Estimation of Σ*

Let $s_z^2=\frac{\sum_{i=1}^n(z_i-\bar{z})^2}{(n-1)}$ and

$$A=\sum_{j=1}^m(\boldsymbol{w}_j-\bar{\boldsymbol{w}})(\boldsymbol{w}_j-\bar{\boldsymbol{w}})'=\begin{bmatrix}a_{11} & a_{12}\\ a_{12} & a_{22}\end{bmatrix}=(m-1)\begin{bmatrix}s_x^2 & s_{xy}\\ s_{xy} & s_y^2\end{bmatrix}$$

where

$$s_x^2=\frac{\sum_{i=1}^m(x_i-\bar{x})^2}{(m-1)},\quad s_y^2=\frac{\sum_{i=1}^m(y_i-\bar{y})^2}{(m-1)}\quad\text{and}\quad s_{xy}=\frac{\sum_{i=1}^m(x_i-\bar{x})(y_i-\bar{y})}{(m-1)}.$$

A simple unbiased estimator for Σ is the sample variance-covariance matrix based only on the paired data $\boldsymbol{w}$'s, i.e. $A/(m-1)$. As both s_z^2 and $s_x^2(=a_{11}/(m-1))$ are unbiased for σ^2, a natural unbiased estimator for Σ based on all the data is given by

$$\hat{\sigma}_1^2=\frac{(n-1)s_z^2+(m-1)s_x^2}{n+m-2},\quad \hat{\eta}_1^2=s_y^2,\quad \hat{\xi}_1=s_{xy}. \tag{2.3}$$

However, it does not guarantee that the resulting $\hat{\Sigma}_1$ is always nonnegative definite(nnd). Below we provide some other estimators for Σ.

2.2.1 *REML and ML estimators of* Σ

Under normality, the well known REML of Σ is obtained by maximizing the marginal likelihood of s_z^2 and A, which is given by:

$$L_1 \propto \frac{1}{(\sigma^2)^{n-1/2}|\Sigma|^{m-1/2}} \exp\left\{-\frac{1}{2}\operatorname{tr}(A\Sigma^{-1}) - \frac{(n-1)s_z^2}{2\sigma^2}\right\}. \tag{2.4}$$

Equating the first derivatives of $\ln L_1$ with respect to the components of Σ to zero and solving the resultant equations lead to the following REML estimator $\hat{\Sigma}_2$ for Σ:

$$\hat{\sigma}_2^2 = \frac{(n-1)s_z^2 + (m-1)s_x^2}{n+m-2} \tag{2.5}$$

$$\hat{\eta}_2^2 = \hat{\sigma}_2^2 \frac{s_{xy}^2}{s_x^4} + \frac{s_x^2 s_y^2 - s_{xy}^2}{s_x^2} \tag{2.6}$$

$$\hat{\xi}_2 = \hat{\sigma}_2^2 \frac{s_{xy}}{s_x^2}. \tag{2.7}$$

Now we discuss the ML estimator for Σ. Let

$$Z(\mu) = \sum_{i=1}^{n}(z_i - \mu)^2, \qquad B(\mu) = \sum_{j=1}^{m}(\boldsymbol{w}_j - \mu\boldsymbol{1}_2)(\boldsymbol{w}_j - \mu\boldsymbol{1}_2)' = \begin{bmatrix} b_{11} & b_{12} \\ b_{12} & b_{22} \end{bmatrix}$$

where $b_{11} = \sum_{i=1}^{m}(x_i - \mu)^2$, $b_{12} = \sum_{i=1}^{m}(x_i - \mu)(y_i - \mu)$ and $b_{22} = \sum_{i=1}^{m}(y_i - \mu)^2$. Then under normality, the likelihood function is given by

$$L_2 \propto \frac{1}{(\sigma^2)^{n/2}|\Sigma|^{m/2}} \exp\left\{-\frac{1}{2}\operatorname{tr}(B(\mu)\Sigma^{-1}) - \frac{Z(\mu)}{2\sigma^2}\right\}. \tag{2.8}$$

Note that if we replace n, m, $Z(\mu)$ and $B(\mu)$ in (2.8) by $(n-1)$, $(m-1)$, $(n-1)s_z^2$ and A respectively, L_2 in (2.8) becomes L_1 in (2.4). So, applying the same steps for L_1 to L_2, we obtain the following equations:

$$\hat{\sigma}_3^2 = \frac{Z(\mu) + b_{11}}{n+m} \tag{2.9}$$

$$\hat{\eta}_3^2 = \hat{\sigma}_3^2 \frac{b_{12}^2}{b_{11}^2} + \frac{b_{11}b_{22} - b_{12}^2}{mb_{11}} \tag{2.10}$$

$$\hat{\xi}_3 = \hat{\sigma}_3^2 \frac{b_{12}}{b_{11}}. \tag{2.11}$$

Equations (2.9), (2.10), (2.11) along with the solution to $\frac{\partial \ln L_2}{\partial \mu} = 0$, i.e. (2.1), are the final equations to be used for solving the ML estimators of μ and Σ. To obtain the MLE, we may plug (2.9)–(2.11) into (2.1) to obtain the MLE for μ first, then obtain the MLE of Σ. However, by doing so, it will result in a complicated fifth degree polynomial in μ. Thus closed form expression for the MLE of μ seems impossible and subsequent inference based on it is indeed a difficult task. Hence, we will not consider the ML estimator for Σ in this paper.

2.2.2 *Properties of the REML Estimator for* Σ

In the following we discuss some properties of the REML estimator $\hat{\Sigma}_2$ for Σ and compare it with the ad hoc estimator $\hat{\Sigma}_1$ in (2.3).

Property 1. Validity. It is easy to see from (2.5)–(2.7) that $\hat{\sigma}_2^2$, $\hat{\eta}_2^2$ and $|\hat{\Sigma}_2| = \hat{\sigma}_2^2 \frac{s_x^2 s_y^2 - s_{xy}^2}{s_x^2}$ are positive with probability 1, thus making $\hat{\Sigma}_2$ a valid estimator for Σ.

Property 2. Bias. Clearly, $\hat{\sigma}_2^2$ and $\hat{\xi}_2$ are unbiased but $\hat{\eta}_2^2$ is not. Using the properties of normality and applying simple algebra, it can be shown that

$$E(\hat{\eta}_2^2) = \eta^2 + \eta^2 \frac{2(n-1)(1-\rho^2)}{(n+m-2)(m-1)(m-3)}, \qquad m > 3.$$

Therefore, the bias of $\hat{\eta}_2^2$ and hence the bias of $\hat{\Sigma}_2$ will tend to zero for large m.

Property 3. Mean squared error (MSE). To derive the MSE of $\hat{\Sigma}_2$, we first represent Σ, $\hat{\Sigma}_1$ and $\hat{\Sigma}_2$ in vectorized forms:

$$\begin{aligned} \Theta &= (\theta_1, \theta_2, \theta_3)' = (\sigma^2, \eta^2, \xi)' \\ \hat{\Theta}_1 &= (\hat{\theta}_{11}, \hat{\theta}_{12}, \hat{\theta}_{13})' = (\hat{\sigma}_1^2, \hat{\eta}_1^2, \hat{\xi}_1)' \\ \hat{\Theta}_2 &= (\hat{\theta}_{21}, \hat{\theta}_{22}, \hat{\theta}_{23})' = (\hat{\sigma}_2^2, \hat{\eta}_2^2, \hat{\xi}_2)'. \end{aligned}$$

The MSE of $\hat{\Sigma}_1$, denoted by $MSE(\hat{\Theta}_1)$, is defined as $E[(\hat{\Theta}_1 - \Theta)(\hat{\Theta}_1 - \Theta)']$ and the expression for MSE of $\hat{\Sigma}_2$ is similar. It is shown in Appendix I that

$$(2.12) \quad MSE(\hat{\Theta}_2) = \frac{2}{n+m-2} \begin{bmatrix} \sigma^4 & \xi^2 & \sigma^2\xi \\ \xi^2 & \frac{(n+m-2)-(n-1)d(\rho)}{m-1}\eta^4 & (1+\frac{(n-1)(1-\rho^2)}{m-3})\xi\eta^2 \\ \sigma^2\xi & (1+\frac{(n-1)(1-\rho^2)}{m-3})\xi\eta^2 & \frac{(n+m-4)\sigma^2\eta^2+(m-n-2)\xi^2}{2(m-3)} \end{bmatrix}$$

where

$$d(\rho) = \rho^4 - \frac{4}{m-3}\rho^2(1-\rho^2) - \frac{7(m-1)(n+m-2) - 4(n+4m-5)}{(n+m-2)(m-1)(m-3)(m-5)}(1-\rho^2)^2.$$

Of course, it is assumed that $m > 5$.

Comparison of REML estimator $\hat{\Sigma}_2$ *with ad hoc estimator* $\hat{\Sigma}_1$

We now compare the MSE of the REML estimator $\hat{\Sigma}_2$ with that of the ad hoc estimator $\hat{\Sigma}_1$. Although $\hat{\Sigma}_1$ is not always a *valid* estimator in the sense of not being nnd, component-wise comparison however makes sense. It is shown in Appendix I that the MSE of $\hat{\Sigma}_1$ is given by

$$MSE(\hat{\Theta}_1) = \frac{2}{n+m-2} \begin{bmatrix} \sigma^4 & \xi^2 & \sigma^2\xi \\ \xi^2 & \frac{n+m-2}{m-1}\eta^4 & \frac{n+m-2}{m-1}\xi\eta^2 \\ \sigma^2\xi & \frac{n+m-2}{m-1}\xi\eta^2 & \frac{n+m-2}{m-1}(\sigma^2\eta^2+\xi^2) \end{bmatrix}$$

and hence

$$MSE(\hat{\Theta}_1) - MSE(\hat{\Theta}_2) = \frac{2}{n+m-2}\begin{bmatrix} 0 & 0 & 0 \\ 0 & \gamma_{22} & \gamma_{23} \\ 0 & \gamma_{23} & \gamma_{33} \end{bmatrix}$$

where

$$\gamma_{22} = \frac{(n-1)d(\rho)}{m-1}\eta^4, \qquad \gamma_{23} = \frac{(n-1)\xi\eta^2}{m-3}\left(\rho^2 - \frac{2}{m-1}\right),$$
$$\gamma_{33} = \frac{(n-1)(m-2)\sigma^2\eta^2}{(m-1)(m-3)}\left(\rho^2 - \frac{1}{m-2}\right).$$

Therefore, for $m > 5$, we have the following observations:

(a) If $\rho^2 > \frac{1}{m-2}$, then $\gamma_{33} > 0$, i.e., $MSE(\hat{\xi}_1) > MSE(\hat{\xi}_2)$, implying that the REML estimator of ξ is preferred to s_{xy}. Note that both are unbiased for ξ.

(b) If $\rho^2 > \Delta/(1+\Delta)$, where

$$\Delta = \frac{2}{m-3}\left\{1 + \sqrt{\frac{11}{4} + \frac{1}{m-5}\left(\frac{7}{2} - \frac{m-3}{m-1} - \frac{3(m-3)}{n+m-2}\right)}\right\},$$

then $d(\rho) > 0$ and $\rho^2 > \frac{1}{m-2}$, i.e., $\gamma_{22} > 0$, $\gamma_{33} > 0$, implying that the REML estimators of ξ and η^2 are better than s_{xy} and s_y^2, respectively.

(c) To have $MSE(\hat{\Theta}_1) - MSE(\hat{\Theta}_2)$ as nnd, $\gamma_{22}\gamma_{33} - \gamma_{23}^2$ should be positive. It can be shown that

$$\gamma_{22}\gamma_{33} - \gamma_{23}^2 = \frac{4\sigma^2\eta^6(m-2)}{(n+m-2)^2(m-1)^2(m-3)}\left[\rho^6 + O\left(\frac{1}{m}\right)\right].$$

Hence, for large m, we expect it to be positive.

In conclusion, we note that, for large m, the REML estimator $\hat{\Sigma}_2$ for Σ has a smaller MSE compared to the ad hoc estimator $\hat{\Sigma}_1$. Therefore in our subsequent analysis, we will use the REML estimator $\hat{\Sigma}_2$ with its subscript dropped for notational simplicity.

2.3 *Estimation of μ when Σ is unknown*

When Σ is unknown, substituting the REML estimator $\hat{\Sigma} = \hat{\Sigma}_2$ into (2.1) gives

$$\tilde{\mu}_{srs} = \frac{\frac{n}{\hat{\sigma}^2}\bar{z} + m\mathbf{1}_2'\hat{\Sigma}^{-1}\bar{\boldsymbol{w}}}{\frac{n}{\hat{\sigma}^2} + m\mathbf{1}_2'\hat{\Sigma}^{-1}\mathbf{1}_2} = \frac{\frac{n}{\hat{\sigma}^2}\bar{z} + m\left(\frac{\hat{\eta}^2 - \hat{\xi}}{\hat{\sigma}^2\hat{\eta}^2 - \hat{\xi}^2}\bar{x} + \frac{\hat{\sigma}^2 - \hat{\xi}}{\hat{\sigma}^2\hat{\eta}^2 - \hat{\xi}^2}\bar{y}\right)}{\frac{n}{\hat{\sigma}^2} + m\frac{\hat{\sigma}^2 + \hat{\eta}^2 - 2\hat{\xi}}{\hat{\sigma}^2\hat{\eta}^2 - \hat{\xi}^2}}. \tag{2.13}$$

Since $\hat{\Sigma}_2$ is independent of $\bar{z}$ and $\bar{\boldsymbol{w}}$, $\tilde{\mu}_{srs}$ is unbiased for μ with variance given by

$$\mathrm{Var}(\tilde{\mu}_{srs}) = E\{\Psi(\hat{\Theta}, \Theta)\}$$

where

$$\Psi(\hat{\Theta}, \Theta) = \mathrm{Var}(\tilde{\mu}_{srs} \mid \hat{\Sigma}) = \frac{\frac{n\sigma^2}{\hat{\sigma}^4} + m\mathbf{1}_2'\hat{\Sigma}^{-1}\Sigma\hat{\Sigma}^{-1}\mathbf{1}_2}{\left(\frac{n}{\hat{\sigma}^2} + m\mathbf{1}_2'\hat{\Sigma}^{-1}\mathbf{1}_2\right)^2}. \tag{2.14}$$

An exact expression for $\mathrm{Var}(\tilde{\mu}_{srs})$ is usually very difficult to obtain. However, for inference purpose what is really needed is an estimate of $\mathrm{Var}(\tilde{\mu}_{srs})$, for which some approximation methods described below can be used.

Method 1. A naive estimator for $\mathrm{Var}(\tilde{\mu}_{srs})$ is obtained by plugging an estimator $\hat{\Sigma}$ of Σ in $\mathrm{Var}(\tilde{\mu}_{srs} \mid \hat{\Sigma})$ given by (2.14), leading to

$$\dot{M}(\hat{\Sigma}) = \frac{1}{\frac{n}{\hat{\sigma}^2} + m\mathbf{1}_2{}'\hat{\Sigma}^{-1}\mathbf{1}_2}. \tag{2.15}$$

As pointed out by many investigators, this method is likely to underestimate $\mathrm{Var}(\tilde{\mu}_{srs})$, a phenomenon discussed later in this section.

Method 2. Here we first approximate $\Psi(\hat{\Theta}, \Theta)$ by a second-order Taylor expansion:

$$\Psi(\hat{\Theta}, \Theta) \approx \Psi(\Theta, \Theta) + (\hat{\Theta} - \Theta)' \left(\frac{\partial \Psi}{\partial \hat{\Theta}}\right)\Bigg|_{\hat{\Theta}=\Theta} + \frac{1}{2}(\hat{\Theta} - \Theta)'\Phi(\hat{\Theta} - \Theta)$$

where $\Phi = (\frac{\partial^2 \Psi}{\partial \hat{\Theta} \partial \hat{\Theta}'})|_{\hat{\Theta}=\Theta}$, the matrix of second derivatives of Ψ with respect to $\hat{\Theta}$ evaluated at Θ. It can be shown by direct derivation that

$$\Psi(\Theta, \Theta) = \dot{M}(\Sigma) = \frac{1}{\frac{n}{\sigma^2} + m\mathbf{1}_2{}'\Sigma^{-1}\mathbf{1}_2}, \qquad \left(\frac{\partial \Psi}{\partial \hat{\Theta}}\right)|_{\hat{\Theta}=\Theta} = 0$$

and $\Phi = (\alpha_{ij})_{3\times3}$ where

$$\begin{aligned}
\alpha_{11} &= \frac{2m}{h(\Theta)}\{n[\sigma^2(\sigma^2\eta^2 - \xi^2)(\sigma^2\eta^2 - 2\xi^2) + \xi^4(\sigma^2 + \eta^2 - 2\xi)] + m\sigma^6(\eta^2 - \xi)^2\} \\
\alpha_{12} &= \frac{2m\sigma^2(\xi - \sigma^2)}{h(\Theta)}\{n[\sigma^4\eta^2 - 2\sigma^2\xi^2 + \xi^3] + m\sigma^4(\eta^2 - \xi)\} \\
\alpha_{13} &= \frac{2m\sigma^2(\xi - \eta^2)}{h(\Theta)}\{n[\sigma^4\eta^2 - 3\sigma^2\xi^2 + 2\xi^3] + m\sigma^4(\eta^2 - \sigma^2)\} \\
\alpha_{22} &= \frac{2m(n + m)\sigma^6(\sigma^2 - \xi)^2}{h(\Theta)} \\
\alpha_{23} &= \frac{2m\sigma^4(\sigma^2 - \xi)}{h(\Theta)}\{n[\sigma^2\eta^2 - 2\sigma^2\xi + \xi^2] + m\sigma^2(\eta^2 - \sigma^2)\} \\
\alpha_{33} &= \frac{2m\sigma^4}{h(\Theta)}\{n[\sigma^2\eta^2(\sigma^2 + \eta^2 - 6\xi) + \xi^2(3\sigma^2 + 3\eta^2 - 2\xi)] + m\sigma^2(\sigma^2 - \eta^2)^2\} \\
h(\Theta) &= [n(\sigma^2\eta^2 - \xi^2) + m\sigma^2(\sigma^2 + \eta^2 - 2\xi)]^3.
\end{aligned}$$

Thus we get

$$\mathrm{Var}(\tilde{\mu}_{srs}) \approx \dot{M}(\Sigma) + \frac{1}{2}E\{(\hat{\Theta} - \Theta)'\Phi(\hat{\Theta} - \Theta)\} = \dot{M}(\Sigma) + \frac{1}{2}\,\mathrm{tr}\{\Phi[MSE(\hat{\Theta})]\}. \tag{2.16}$$

It is obvious that (2.16) will always give a larger estimator for $\mathrm{Var}(\tilde{\mu}_{srs})$ than $\dot{M}(\Sigma)$. In fact in a general mixed linear model setup, which covers our linear model for $\boldsymbol{t}$ as a special case, Kackar and Harville (1984) proposed a similar approximation expression for the variance of estimators of fixed and random effects. It is evidenced by their simulations that (2.16) approximates well the actual variance of $\tilde{\mu}_{srs}$. Therefore, $\mathrm{Var}(\tilde{\mu}_{srs})$ is estimated by

$$\widehat{\mathrm{Var}}(\tilde{\mu}_{srs}) = \dot{M}(\hat{\Sigma}) + \frac{1}{2}\,\mathrm{tr}\{\hat{\Phi}[\widehat{MSE}(\hat{\Theta})]\} \tag{2.17}$$

where $\dot{M}(\hat{\Sigma})$ is given by (2.15), $\hat{\Phi}$ and $\widehat{MSE}(\hat{\Theta})$ respectively refer to Φ and $MSE(\hat{\Theta}_2)$ in (2.12), with elements of Σ replaced by $\hat{\Sigma} = \hat{\Sigma}_2$.

3. Estimation of μ using SRS-RSS double sampling

In this section we explore the use of a ranked set sampling (RSS) in place of a simple random sampling in the second-phase of a double sampling. RSS, originally introduced by McIntyre (1952) for efficient estimation of a population mean in a purely nonparametric setup, has been found to be fairly effective in various problems of parametric estimation (see Chuiv and Sinha (1998) and Patil *et al.* (1994) and references therein). In our specific problem, we propose to use the field-only data and paired (field, lab) data in a modified form described as follows.

For a simple random sample of size r units (gas pumps), we collect X-values (field) from all the units. We identify the unit with the smallest X-value and send the corresponding sample to the laboratory to record the Y-value (lab). We next draw another simple random sample of r units, and collect their X-values (field). We identify the unit with the second smallest X-value and send the corresponding sample to the laboratory to record the Y-value (lab). This process is continued in r steps and at the last stage after collecting X-values (field) from all the r units, we identify the unit with the largest X-value and send it to the laboratory to record the Y-value (lab). At the end of this process, what we have collected is a sample of r^2 field measurements and a suitably selected RSS of r lab measurements. The entire process is now repeated N cycles to yield eventually a sample of Nr^2 field measurements and a suitably selected RSS of Nr lab measurements. Denote the measurements recorded in the i-th cycle by

$$\begin{array}{l} X^{(i)}_{(11)}, \ldots, X^{(i)}_{(1r)}, Y^{(i)}_{[11]} \\ X^{(i)}_{(21)}, \ldots, X^{(i)}_{(2r)}, Y^{(i)}_{[22]} \\ \cdots \\ X^{(i)}_{(r1)}, \ldots, X^{(i)}_{(rr)}, Y^{(i)}_{[rr]} \end{array}$$

where $X^{(i)}_{(jk)}$ is the k-th order statistic (field measurement) in a simple random sample of size r arising out of the j-th sample in the i-th cycle, $i = 1, \ldots, N$, $j, k = 1, \ldots, r$ and $Y^{(i)}_{[kk]}$ is the lab measurement corresponding to the field measurement $X^{(i)}_{(kk)}$ obtained in the i-th cycle, $i = 1, \ldots, N$, $k = 1, \ldots, r$.

Denote the overall mean of X by $\bar{X} = \sum_{i=1}^{N}\sum_{j=1}^{r}\sum_{k=1}^{r} X^{(i)}_{(jk)}/(Nr^2)$, and the sample means of X and Y based on the ranked set sample obtained in the second phase by $\bar{X}_{rss} = \sum_{i=1}^{N}\sum_{j=1}^{r} X^{(i)}_{(jj)}/(Nr)$ and $\bar{Y}_{rss} = \sum_{i=1}^{N}\sum_{j=1}^{r} Y^{(i)}_{[jj]}/(Nr)$ respectively. Note that $\bar{X}$ and $\bar{X}_{rss}$ are always unbiased for μ but $\bar{Y}_{rss}$ may be biased. Suppose X and Y follows the linear model (see David (1973) and Stokes (1977)):

$$Y = \mu + \beta(X - \mu) + \varepsilon \tag{3.1}$$

where $\beta = \rho\eta/\sigma = \xi/\sigma^2$ and ε has zero mean and variance $\eta^2(1-\rho^2)$ and is independent of X. It is easy to show that $\bar{Y}_{rss}$ is unbiased for μ. Since X and Y follow a bivariate normal distribution, the linear model in (3.1) is satisfied and hence $\bar{Y}_{rss}$ is unbiased.

3.1 *Estimation of μ when Σ is known*

As discussed in Section 2, the SRS-SRS double sampling involves drawing of a large random sample of size $n+m$ and a subsample of size m. Based on this sampling method, we derived the MLE for μ when Σ is known as shown in (2.1). Under a SRS-RSS double sampling setting, $n = Nr(r-1)$ and $m = Nr$. After making some obvious changes:

$$\bar{z} = \frac{\sum_{i=1}^{N} \sum \sum_{j \neq k} X_{(jk)}^{(i)}}{Nr(r-1)} = \frac{Nr^2\bar{X} - Nr\bar{X}_{rss}}{Nr(r-1)}, \qquad \bar{x} = \bar{X}_{rss} \qquad \text{and} \qquad \bar{y} = \bar{Y}_{rss},$$

we propose the following estimator for μ when Σ is known:

$$\begin{aligned} \hat{\mu}_{rss} &= \frac{\dfrac{Nr^2\bar{X} - Nr\bar{X}_{rss}}{\sigma^2} + Nr\left[\dfrac{\eta^2 - \xi}{\sigma^2\eta^2 - \xi^2}\bar{X}_{rss} + \dfrac{\sigma^2 - \xi}{\sigma^2\eta^2 - \xi^2}\bar{Y}_{rss}\right]}{\dfrac{Nr(r-1)}{\sigma^2} + Nr\dfrac{\sigma^2 + \eta^2 - 2\xi}{\sigma^2\eta^2 - \xi^2}} \\ &= \frac{\dfrac{r}{\sigma^2}\bar{X} + \dfrac{\sigma^2 - \xi}{\sigma^2\eta^2 - \xi^2}(\bar{Y}_{rss} - \dfrac{\xi}{\sigma^2}\bar{X}_{rss})}{\dfrac{1}{\sigma^2}\left(r + \dfrac{(\sigma^2 - \xi)^2}{\sigma^2\eta^2 - \xi^2}\right)}. \end{aligned} \tag{3.2}$$

Of course, the above estimator for μ is far from being the MLE under a SRS-RSS sampling. Interestingly enough, it is shown in Appendix II that when (3.1) is satisfied, $\hat{\mu}_{rss}$ given by (3.2) is the best linear unbiased estimator (BLUE) for μ based on $\bar{X}$, $\bar{X}_{rss}$ and $\bar{Y}_{rss}$, and the variance of $\hat{\mu}_{rss}$ is given by

$$\operatorname{Var}(\hat{\mu}_{rss}) = \frac{\sigma^2}{Nr} \cdot \frac{1}{r + \dfrac{(\sigma^2 - \xi)^2}{\sigma^2\eta^2 - \xi^2}}. \tag{3.3}$$

Therefore when Σ is known, $\hat{\mu}_{rss}$ is more efficient than $\bar{X}$, $\bar{X}_{rss}$ and $\bar{Y}_{rss}$.

3.2 *Estimation of μ when Σ is unknown*

When Σ is unknown, a standard practice is to start from $\hat{\mu}_{rss}$ given in (3.2) and use a suitable estimator for Σ. In the context of SRS-SRS double sampling discussed in Section 2, it is found that the REML estimator of Σ has some nice properties than other estimators. It is clear that in our context, due to the complicated nature of the likelihood function (due primarily to RSS nature), it is extremely difficult to derive the REML estimator for Σ. In what follows, we adopt the REML estimator for Σ even in our context. Define

$$S_z^2 = \frac{\sum_{i=1}^{N} \sum_{j=1}^{r} \sum_{k=1_{j \neq k}}^{r} (X_{(jk)}^{(i)} - \bar{X})^2}{Nr(r-1) - 1}$$

$$S_x^2 = \frac{\sum_{i=1}^{N} \sum_{k=1}^{r} (X_{(kk)}^{(i)} - \bar{X}_{rss})^2}{Nr - 1}$$

$$S_y^2 = \frac{\sum_{i=1}^N \sum_{k=1}^r (Y_{[kk]}^{(i)} - \bar{Y}_{rss})^2}{Nr-1}$$

$$S_{xy} = \frac{\sum_{i=1}^N \sum_{k=1}^r [(X_{(kk)}^{(i)} - \bar{X}_{rss})(Y_{[kk]}^{(i)} - \bar{Y}_{rss})]}{Nr-1}.$$

Then our proposed estimator for Σ is given by $\hat{\Sigma}_{rss}$ where

$$\hat{\sigma}_{rss}^2 = S_x^2$$

$$\hat{\eta}_{rss}^2 = \hat{\sigma}_{rss}^2 \frac{S_{xy}^2}{S_x^4} + \frac{S_x^2 S_y^2 - S_{xy}^2}{S_x^2}$$

$$\hat{\xi}_{rss} = \hat{\sigma}_{rss}^2 \frac{S_{xy}}{S_x^2}.$$

It may be noted that these estimates are well-defined and valid in the sense of the estimated dispersion matrix being nnd, irrespective of the underlying model. After substituting $\hat{\Sigma}_{rss}$ into (3.2), the resultant estimator of μ is denoted by $\tilde{\mu}_{rss}$.

To prove the unbiasedness of $\tilde{\mu}_{rss}$, we first notice that $\tilde{\mu}_{rss}$ can be expressed as $\mu + \tilde{\mu}^*$ where $\tilde{\mu}^*$ is the $\tilde{\mu}_{rss}$ with X and Y replaced by $X^* = X - \mu$ and $Y^* = Y - \mu$, respectively. Since $\hat{\Sigma}_{rss}$ is an even function of X^* and Y^*, replacing X^* and Y^* by $-X^*$ and $-Y^*$ in $\tilde{\mu}^*$ implies $E[\tilde{\mu}^*] = E[-\tilde{\mu}^*]$. It follows that $E[\tilde{\mu}^*] = 0$ and hence $\tilde{\mu}_{rss}$ is unbiased.

It is clear that the exact variance of $\tilde{\mu}_{rss}$ is difficult to obtain, and in what follows we therefore employ the variance of $\hat{\mu}_{rss}$ given in (3.3) as a large sample approximate of $\text{Var}(\tilde{\mu}_{rss})$ for large N.

4. Other estimators for μ

Note that when the data are collected using a double sampling scheme, a regression estimator is usually used to estimate the population mean of Y based on a covariate X no matter X and Y have common mean or not. Recently, Yu and Lam (1997) proposed a RSS regression estimator based on a SRS-RSS double sampling scheme mentioned in Section 3:

$$\tilde{\mu}_{reg} = \bar{Y}_{rss} + \hat{\beta}(\bar{X} - \bar{X}_{rss}) \tag{4.1}$$

where

$$\hat{\beta} = \hat{\xi}_{rss}/\hat{\sigma}_{rss}^2 = \frac{S_{xy}}{S_x^2} \tag{4.2}$$

is an estimator for the slope β in (3.1). If (3.1) is satisfied and hence normality holds, Yu and Lam (1997) showed that $\tilde{\mu}_{reg}$ is unbiased and its variance is given by:

$$\text{Var}(\tilde{\mu}_{reg}) = \frac{\sigma^2\eta^2 - \xi^2}{\sigma^2 Nr}[1 + \Delta] + \frac{\xi^2}{\sigma^2 Nr^2} \tag{4.3}$$

where

$$\Delta = E\left[\frac{Nr(\bar{X}_{rss} - \bar{X})^2}{(Nr-1)S_x^2}\right] \tag{4.4}$$

and we take $\mu = 0$ and $\sigma = 1$ in the computation of Δ. Obviously, Δ is a fixed constant depending only on N and r.

Of course, a similar SRS regression estimator based on a SRS-SRS double sampling scheme can also be proposed here. However, Yu and Lam (1997) found that under normality, the RSS regression estimator is always superior to the SRS regression estimator for all ρ.

Finally, since $\bar{X}$, $\bar{X}_{rss}$ and $\bar{Y}_{rss}$ do not utilize all the available data and they are inferior than $\hat{\mu}_{rss}$ when Σ is known, we do not intend to consider these estimators although they are unbiased.

In next section, we will compare the two proposed common mean estimators with the RSS regression estimator.

5. Numerical comparisons

Assuming that (X, Y) follows a bivariate normal distribution with common mean $\mu = 0$, we compute the variances of the two proposed common mean estimators $\tilde{\mu}_{srs}, \tilde{\mu}_{rss}$ and the RSS regression estimator $\tilde{\mu}_{reg}$. Since these three estimators are unbiased, we use the variance ratio as a measure of relative precision (RP). The set size examined is $r = 3$, the numbers of cycles are $N = 5, 10$, and the values of ρ are $0, 0.1, 0.2, \ldots, 0.9$. It is easy to see that the RP can be expressed as a function of η/σ and ρ. Without loss of generality, we assume $\sigma = 1$ and consider various choices of $\theta = \eta/\sigma$. As the lab data is expected to be more precise than the field data, θ is usually less than 1. Here, we consider four values of θ: 0.9,0.7,0.3,0.1. The variance of $\tilde{\mu}_{reg}$ is evaluated using (4.3). Because the variances of $\tilde{\mu}_{srs}$ and $\tilde{\mu}_{rss}$ have no exact analytical expressions, their variances are evaluated by a simulation of size 100,000.

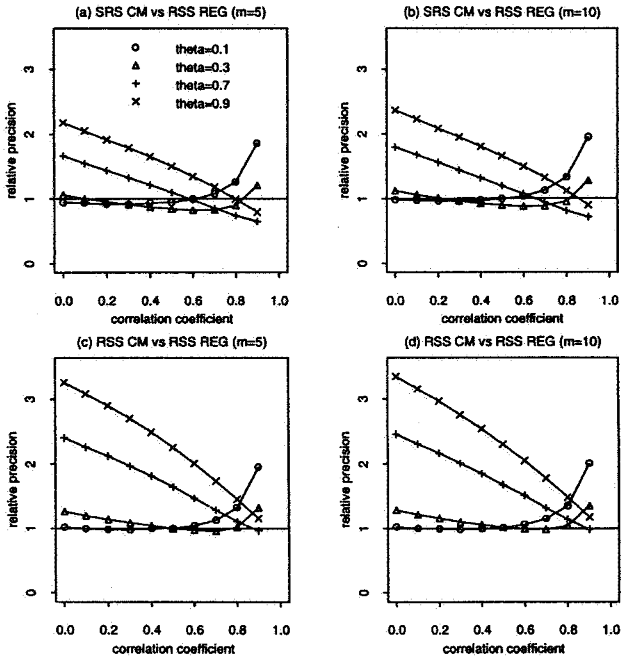

Fig. 1. The relative precision of SRS and RSS common mean estimators relative to RSS regression estimator.

5.1 *Comparison of common mean estimators with RSS regression estimator*

Figure 1 depicts the relative precisions of the two proposed common mean estimators $\tilde{\mu}_{srs}$, $\tilde{\mu}_{rss}$ relative to the RSS regression estimator $\tilde{\mu}_{reg}$. It can be seen that the RSS common mean estimator is almost superior to the RSS regression estimator but not for the SRS common mean estimator. However when θ is large (≥ 0.7 say) and ρ is not too large, both common mean estimators perform significantly better than the RSS regression estimator.

It is not surprising that the RPs of $\tilde{\mu}_{rss}$ to $\tilde{\mu}_{reg}$ are close to 1 when θ is close to 0. Note that when θ is close to 0, X is too variable and becomes nearly useless in estimating μ. Therefore the RSS regression estimator, which aims to estimate the mean of Y, will perform similarly to the RSS common mean estimator. In fact, it can be shown that the RSS common mean estimator $\tilde{\mu}_{rss}$ can be rewritten as a weighted sum of two unbiased estimators $\bar{X}$ and $\tilde{\mu}_{reg}$ with random weights:

$$(5.1) \qquad \tilde{\mu}_{rss} = (1-\hat{a})\bar{X} + \hat{a}\tilde{\mu}_{reg} \quad \text{where} \quad \hat{a} = \frac{1-\hat{\rho}\hat{\theta}}{r\hat{\theta}^2(1-\hat{\rho}^2)+(1-\hat{\rho}\hat{\theta})^2}$$

with $\hat{\theta} = \hat{\eta}_{rss}/\hat{\sigma}_{rss}$ and $\hat{\rho} = \hat{\xi}_{rss}/(\hat{\eta}_{rss}\hat{\sigma}_{rss}) = \hat{\beta}/\hat{\theta}$. Note that $\hat{a} = 1$ if and only if $\hat{\theta} = 0$ or $\hat{\theta} = \hat{\rho}/[r(1-\hat{\rho}^2)+\hat{\rho}^2] \equiv \theta_0$. Table 1 lists the values of θ_0 for various choices of $\hat{\rho}$ and $r = 3$. Thus if $\hat{\theta}$ is close to θ_0, $\hat{a}$ is close to 1 and hence the RSS regression estimator is approximately equivalent to the RSS common mean estimator.

As analogy to $\tilde{\mu}_{rss}$ in (5.1), $\tilde{\mu}_{srs}$ can also be expressed as a weighted sum of $\bar{X}$ and the SRS regression estimator with weight $\hat{b}$ having the similar form to $\hat{a}$. Therefore, when θ is close to 0, $\hat{b}$ is likely close to 1 and hence the SRS common mean estimator is close to the SRS regression estimator. Since Yu and Lam (1997) showed that the SRS regression estimator is always less precise than the RSS regression estimator, the SRS common mean estimator perform poorer than the RSS regression estimator when θ is close to 0.

Table 1. The values of θ_0 for various choices of $\hat{\rho}$ and $r = 3$.

$\hat{\rho}$	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	0.95	0.99
θ_0	0	0.034	0.068	0.106	0.149	0.200	0.263	0.347	0.465	0.652	0.795	0.952

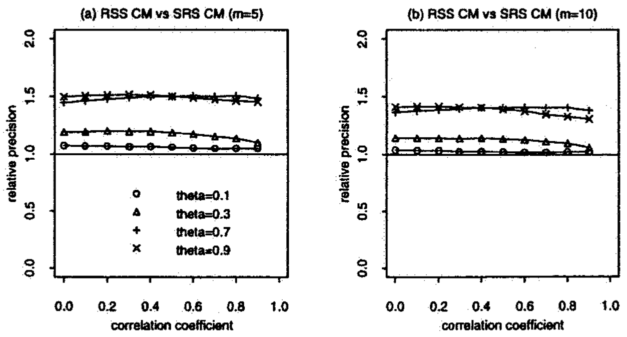

Fig. 2. The relative precision of RSS common mean estimator relative to SRS common mean estimator.

Table 2. The ratios of the approximate variance to the actual variance of $\tilde{\mu}_{rss}$.

		$N = 5$				$N = 10$				$N = 15$			
r	ρ	θ				θ				θ			
		0.1	0.3	0.7	0.9	0.1	0.3	0.7	0.9	0.1	0.3	0.7	0.9
3	0.0	0.97	0.96	0.95	0.95	0.99	0.98	0.98	0.98	0.99	0.99	0.99	0.99
	0.3	0.97	0.95	0.95	0.95	0.98	0.97	0.97	0.97	0.99	0.98	0.99	0.99
	0.6	0.96	0.95	0.95	0.95	0.98	0.98	0.98	0.98	0.99	0.98	0.99	0.99
	0.9	0.96	0.96	0.95	0.96	0.99	0.98	0.97	0.98	1.00	0.98	0.98	0.99
5	0.0	0.98	0.97	0.97	0.98	0.99	0.98	0.99	0.99	0.99	0.99	0.99	1.00
	0.3	0.99	0.98	0.98	0.99	1.00	0.99	0.99	0.99	1.00	1.00	1.00	1.00
	0.6	0.98	0.97	0.98	0.99	1.00	0.99	0.99	1.00	0.99	0.99	0.99	1.00
	0.9	0.99	0.98	0.98	0.99	0.98	0.99	1.00	1.00	1.00	0.99	0.99	1.00

5.2 *Comparison of RSS common mean estimator with SRS common mean estimator*

Figure 2 depicts the relative precision of the RSS common mean estimator $\tilde{\mu}_{rss}$ relative to the SRS common mean estimator $\tilde{\mu}_{srs}$. It is easily seen that the RSS common mean estimator always performs better than the SRS common mean estimator. It should be noted that the values of RPs mainly depends on the value of θ only and they are significantly greater than 1 for large θ. This indicates that when the variances of X and Y are close, a double sampling scheme with its second stage being a ranked set sampling can provide a more precise common mean estimator than the one with its second stage being a simple random sampling.

5.3 *Comparison of the approximate variance and the actual variance for RSS common mean estimator*

Table 2 presents the ratios of the approximate variance to the actual variance for the RSS common mean estimator $\tilde{\mu}_{rss}$ for various combinations of θ and ρ. The approximate variance is computed by using (3.3) while the actual variance is obtained from the above-mentioned simulation based on a bivariate normal distribution. The set size examined is $r = 3, 5$ and the number of cycles is $N = 5, 10, 15$. It can be seen from Table 1 that although the ratios are all less than 1, they vary in a very narrow range from 0.95 to 1.00. This indicates that the approximate variance a little bit underestimates the actual variance of $\tilde{\mu}_{rss}$. The ratios are very close to 1 when the ranked set sample size is moderately large, says $Nr > 30$. This concludes that the approximate variance expression given in (3.3) provides a robust and close-form expression for the variance of $\tilde{\mu}_{rss}$ even the ranked set sample is of moderate size.

6. Application to an EPA data set

In this section, we return to the practical problem of estimating the mean of Reid Vapor Pressure (RVP) of the new reformulated gasoline in the U.S. Since the laboratory analyses are costly, a SRS-RSS double sampling scheme is adopted to reduce the quantity of laboratory analyses and hence save cost. Here a SRS-RSS double sampling scheme with set size $r = 3$ and number of cycles $N = 5$ is used to draw the sample and the field (X) and lab measurements (Y) in the sample are then collected. Table 3 presents the data on X and Y and their summary statistics are shown in Table 4.

Table 3. The field and lab data on RVP for new reformulated gasoline* (bold numbers indicate the selected X in the second phase).

X			Y
8.03	8.09	8.46	8.28
7.37	**8.64**	8.80	8.63
7.59	8.62	**9.14**	9.28
7.86	7.88	7.98	7.85
7.47	**8.70**	8.90	8.62
8.51	8.69	**9.28**	9.14
7.86	7.93	7.96	7.86
7.45	**7.83**	8.02	7.90
7.32	7.45	**8.60**	8.52
7.83	7.86	7.88	7.92
7.39	**7.88**	8.03	7.89
7.31	7.44	**8.56**	8.48
7.83	7.95	7.92	7.95
7.53	**7.99**	8.01	8.32
7.16	7.31	**7.56**	7.60

* Data Source: Private Communication

Table 4. Summary statistics for the crude RVP measurement X and the accurate RVP measurement Y.

r	N	$\bar{X}$	$\bar{Y}_{rss}$	$\bar{X}_{rss}$	S_z^2	S_x^2	S_y^2	S_{xy}
3	5	7.997	8.283	8.239	0.252150	0.284778	0.245392	0.256838

Table 5. Point estimates, standard errors and relative precisions of estimators for μ.

	Benchmark estimators			RSS regression	RSS common mean
	$\bar{X}_{rss}$	$\bar{Y}_{rss}$	$\bar{X}$	estimator, $\tilde{\mu}_{reg}$	estimator, $\tilde{\mu}_{rss}$
Point estimate	8.239	8.283	7.997	8.064	8.035
Standard error	0.0937	0.0898	0.0749	0.0741	0.0727
RP*	100%	109.0%	156.8%	160.0%	166.0%

* RP = relative precision with $\bar{X}_{rss}$ as the base

Using the summary statistics in Table 4, we have $\hat{\sigma}^2_{rss} = 0.252$, $\hat{\eta}^2_{rss} = 0.219$, $\hat{\xi}_{rss} = 0.227$, $\hat{\beta} = 0.902$, $\hat{\theta} = 0.932$ and $\hat{\rho} = 0.968$. Based on these statistics, we can compare the performance of RSS common mean estimator $\tilde{\mu}_{rss}$ and the RSS regression estimators $\tilde{\mu}_{reg}$. Three unbiased estimators $\bar{X}$, $\bar{Y}_{rss}$ and $\bar{X}_{rss}$ are also considered as benchmarks. Table 5 shows their point estimates, standard errors, and relative precisions.

It can be seen from Table 5 that the RSS common mean estimator $\tilde{\mu}_{rss}$ attains the smallest precisions (about 66% increase over the worst benchmark and 6% increase over the best benchmark). This result is not surprising because since in this example

$\hat{a} = 0.566$, $\tilde{\mu}_{rss}$ is approximately an average of $\bar{X}$ and $\tilde{\mu}_{reg}$. Simply using either $\bar{X}$ or $\tilde{\mu}_{reg}$ cannot beat $\tilde{\mu}_{rss}$.

7. Concluding remarks

In this paper, we proposed two common mean estimators and showed that the proposed RSS common mean estimator is more precise than the other estimators including Yu and Lam's (1997) RSS regression estimator, McIntyre's (1952) RSS naive estimator and the proposed SRS common mean estimator. Simulation study performed in Section 4 shows that the approximate variance expression given in (3.3) provides a robust estimate for the actual variance of the RSS common mean estimator even when the sample size is moderate large.

Apart from the problem of estimating the common mean μ, it is also of interest to consider the problems of constructing hypotheses testing and a confidence interval (CI) for μ. As long as the tests and confidence intervals based separately on the 'field-only' data and the paired data are available, we can adopt various combination techniques described in Yu *et al.* (1999) to combine the tests and hence construct a confidence interval for μ by converting the acceptance region of the combined test. For example using the sample drawn by a SRS-SRS double sampling scheme as in Section 3, it is well known that based on $(\bar{z}, s_z^2)$ only, we can use the one-sample t-test to test for $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$, where μ_0 is a given constant, and its test statistic is given by

$$t_1 = \frac{\bar{z} - \mu_0}{\sqrt{s_z^2/n}}$$

which follows a t distribution with $n-1$ d.f. under H_0 and its associated $100(1-\alpha)\%$ CI for μ is

$$\{\mu_0 : |t_1| < t_{\alpha/2,n-1}\} = \left(\bar{z} - t_{\alpha/2,n-1}\sqrt{\frac{s_z^2}{n}}, \bar{z} + t_{\alpha/2,n-1}\sqrt{\frac{s_z^2}{n}}\right)$$

where $t_{\alpha/2,n-1}$ is the upper $\alpha/2$-point of the t_{n-1} distribution. Similarly, based on the paired data (x_i, y_i)'s, we can derive a likelihood ratio test (LRT) for H_0 and the equivalent test statistic is given by

$$t_2 = \frac{\bar{u} - \mu_0 - \dfrac{\bar{v} s_{uv}}{s_v^2}}{\sqrt{h}}$$

where

$$\bar{u} = \frac{\bar{x} + \bar{y}}{2}, \quad \bar{v} = \frac{\bar{y} - \bar{x}}{2}, \quad s_u^2 = \frac{1}{4}(s_x^2 + 2s_{xy} + s_y^2), \; s_v^2 = \frac{1}{4}(s_x^2 - 2s_{xy} + s_y^2),$$

$$s_{uv} = \frac{1}{4}(s_x^2 - s_y^2) \quad \text{and} \quad h = \frac{1}{m-2}\left(\frac{m-1}{m} + \frac{\bar{v}^2}{s_v^2}\right)\frac{(s_u^2 s_v^2 - s_{uv}^2)}{s_v^4},$$

which follows a t distribution with $m-2$ d.f. under H_0 and its associated the $100(1-\alpha)\%$ CI of μ is

$$\{\mu_0 : |t_2| < t_{\alpha/2,m-2}\} = \left(\bar{u} - \frac{\bar{v} s_{uv}}{s_v^2} - t_{\alpha/2,m-2}\sqrt{h}, \bar{u} - \frac{\bar{v} s_{uv}}{s_v^2} + t_{\alpha/2,m-2}\sqrt{h}\right).$$

Let $F_1 = t_1^2$ and $F_2 = t_2^2$ so that $F_1 \sim F_{1,n-1}$ and $F_2 \sim F_{1,m-2}$. Define the p-values based on the two F-statistics as $P_1 = \int_{F_1}^{\infty} f_{1,n-1}(x)dx$ and $P_2 = \int_{F_2}^{\infty} f_{1,m-2}(x)dx$, where $f_{1,k}$ denotes the pdf of the $F_{1,k}$ distribution. Following Yu *et al.* (1999), we can combine these two t_i's or F_i's or P_i's to test for H_0 and hence construct confidence intervals for μ.

Appendix I: MSEs of $\hat{\Sigma}_1$ and $\hat{\Sigma}_2$

MSEs of $\hat{\Sigma}_1$

Following the notations in Section 2, we first note that

$$\hat{\Theta}_1 = \begin{bmatrix} \hat{\sigma}_1^2 \\ \hat{\eta}_1^2 \\ \hat{\xi}_1 \end{bmatrix} = \begin{bmatrix} \dfrac{(n-1)s_z^2}{n+m-2} \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} \dfrac{a_{11}}{n+m-2} \\ \dfrac{a_{22}}{m-1} \\ \dfrac{a_{12}}{m-1} \end{bmatrix}.$$

Since s_z^2 is independent of A and $\hat{\Theta}_1$ is unbiased,

$$MSE(\hat{\Theta}_1) = \mathrm{Var}(\hat{\Theta}_1) = \mathrm{Var}\left(\begin{bmatrix} \dfrac{(n-1)s_z^2}{n+m-2} \\ 0 \\ 0 \end{bmatrix}\right) + \mathrm{Var}\left(\begin{bmatrix} \dfrac{a_{11}}{n+m-2} \\ \dfrac{a_{22}}{m-1} \\ \dfrac{a_{12}}{m-1} \end{bmatrix}\right).$$

Using the result from Muirhead ((1982), p. 90) that if $H = (h_{ij}) \sim W_p(\Sigma, m)$, then $\mathrm{Cov}(h_{ij}, h_{kl}) = m(\sigma_{ik}\sigma_{jl} + \sigma_{il}\sigma_{jk})$, where σ_{ij} is the ij-th element of Σ and the fact that $\mathrm{Var}(s_z^2) = 2\sigma^4/(n-1)$ and $A \sim W_2(\Sigma, m-1)$, the expression for $MSE(\hat{\Theta}_1)$ is then obtained.

MSEs of $\hat{\Sigma}_2$

Using Theorem 3.2.10 of Muirhead (1982) and basic properties on conditional moments, we first obtain some preliminary results:

(a) $a_{12} \mid a_{11} \sim N(\frac{\rho\eta}{\sigma}a_{11}, \eta^2(1-\rho^2)a_{11})$.

(i) $E(a_{12}^2 \mid a_{11}) = a_{11}\eta^2(1-\rho^2) + a_{11}^2\dfrac{\rho^2\eta^2}{\sigma^2}$.

(ii) $\mathrm{Var}(a_{12}^2 \mid a_{11}) = 2a_{11}^2\eta^4\left[(1-\rho^2)^2 + 2a_{11}\dfrac{\rho^2(1-\rho^2)}{\sigma^2}\right]$.

(b) $a_{22} - \frac{a_{12}^2}{a_{11}} \sim \eta^2(1-\rho^2)\chi_{m-2}^2$ and is independent of a_{11} and a_{22}.

(i) $E(a_{22} \mid a_{11}) = (m-1)\eta^2(1-\rho^2) + a_{11}\dfrac{\rho^2\eta^2}{\sigma^2}$.

(ii) $\mathrm{Cov}(a_{12}^2, a_{22} \mid a_{11}) = \mathrm{Var}(a_{12}^2 \mid a_{11})/a_{11} = 2a_{11}\eta^4\left[(1-\rho^2)^2 + 2a_{11}\dfrac{\rho^2(1-\rho^2)}{\sigma^2}\right]$.

(iii) $\mathrm{Var}(a_{22} \mid a_{11}) = 2\eta^4\left[(m-1)(1-\rho^2)^2 + 2a_{11}\dfrac{\rho^2(1-\rho^2)}{\sigma^2}\right]$.

To derive the MSE of $\hat{\Sigma}_2$, we first note that

$$E\hat{\sigma}_2^2 = \sigma^2, \qquad E\hat{\xi}_2 = \xi, \quad E\hat{\eta}_2^2 = \eta^2 \left(1 + \frac{2(n-1)(1-\rho^2)}{(n+m-2)(m-1)(m-3)}\right).$$

So,

- $MSE(\hat{\sigma}_2^2) = \text{Var}(\hat{\sigma}_2^2) = \text{Var}(\hat{\sigma}_1^2) = \frac{2\sigma^4}{n+m-2}$.
- $MSE(\hat{\eta}_2^2) = \text{Var}(\hat{\eta}_2^2) + [E(\hat{\eta}_2^2) - \eta^2]^2$.
- $MSE(\hat{\xi}_2) = \text{Var}(\hat{\xi}_2)$.
- $E(\hat{\sigma}_2^2 - \sigma^2)(\hat{\eta}_2^2 - \eta^2) = E\hat{\sigma}_2^2\hat{\eta}_2^2 - \sigma^2 E\hat{\eta}_2^2$.
- $E(\hat{\sigma}_2^2 - \sigma^2)(\hat{\xi}_2 - \xi) = E\hat{\sigma}_2^2\hat{\xi}_2 - \sigma^2\xi$.
- $E(\hat{\eta}_2^2 - \eta^2)(\hat{\xi}_2 - \xi) = \text{Cov}(\hat{\eta}_2^2, \hat{\xi}_2)$.

The rest of the derivation of $MSE(\hat{\Theta}_2)$ follows by using the previous preliminary results.

Appendix II: The BLUE of μ based on $\bar{X}$, $\bar{X}_{rss}$ and $\bar{Y}_{rss}$

Consider a linear estimator of μ:

$$L = a\bar{X} + b\bar{X}_{rss} + c\bar{Y}_{rss} \quad \text{with} \quad a + b + c = 1. \tag{A.1}$$

We can write $\text{Var}(L) = \text{Var}[E(L \mid X)] + E[\text{Var}(L \mid X)]$. Under (3.1), we get (taking $\mu = 0$ without any loss of generality)

$$\begin{aligned} E[L \mid X] &= a\bar{X} + b\bar{X}_{rss} + c\beta\bar{X}_{rss} \\ &= a\bar{X} + (b + c\beta)\bar{X}_{rss} \\ \text{Var}[L \mid X] &= c^2\eta^2(1-\rho^2)/(Nr). \end{aligned} \tag{A.2}$$

To compute $\text{Var}(E[L \mid X])$, we first condition on all X, denoted as S, and treat the selection of RSS as random and then uncondition on X. Since, given S, $\bar{X}$ is fixed, we get $\text{Var}(E[L \mid X] \mid S) = (b + c\beta)^2 \text{Var}(\bar{X}_{rss} \mid S)$ and hence

$$\begin{aligned} \text{Var}(E[L \mid X]) &= \text{Var}\{E(E[L \mid X] \mid S)\} + E[\text{Var}(E[L \mid X] \mid S)] \\ &= (1 - c + c\beta)^2 \frac{\sigma^2}{Nr^2} + (b + c\beta)^2 E[\text{Var}(\bar{X}_{rss} \mid S)]. \end{aligned} \tag{A.3}$$

Combining (A.2) and (A.3), we get $\text{Var}(L)$. Clearly, for a given c, $\text{Var}(L)$ is minimized when $b = -c\beta$ and $\text{Var}(L)$ becomes

$$\text{Var}(L) = (1 - c + c\beta)^2 \frac{\sigma^2}{Nr^2} + c^2 \frac{\eta^2(1-\rho^2)}{Nr}. \tag{A.4}$$

The optimum value of c is easily obtained by minimizing the above quadratic function in c and turns out to be $c_{opt} = \frac{1-\beta}{(1-\beta)^2 + r(\theta^2 - \beta^2)}$. Substituting $a = 1 - b_{opt} - c_{opt}$, $b_{opt} = -c_{opt}\beta$ and c_{opt} into (A.1) and (A.4), the resulting optimum linear unbiased estimator of μ is precisely $\hat{\mu}_{rss}$ with variance as shown in (3.3).

References

Chuiv, N. and Sinha, B. K. (1998). On some aspects of ranked set sampling in parameter estimation, *Handbook of Statistics 17* (eds. N. Balakrishnan and C. R. Rao), 337–377, Elsevier, North-Holland, Amsterdam.

David, H. A. (1973). Concomitants of order statistics, *Bulletin of the International Statistical Institute*, **45**(1), 295–300.

Kackar, R. N. and Harville, D. A. (1984). Approximations for standard errors of fixed and random effects in mixed linear models, *J. Amer. Statist. Assoc.*, **79**, 853–862.

McIntyre, G. A. (1952). A method of unbiased selective sampling using ranked sets, *Australian Journal of Agricultural Research*, **3**, 385–390.

Muirhead, R. J.(1982). *Aspects of Multivariate Statistical Theory*, Wiley, New York.

Patil, G. P., Sinha, A. K. and Taillie, C. (1994). Ranked set sampling, *Handbook of Statistics 12* (eds. G. P. Patil and C. R. Rao), 167–200, Elsevier, North-Holland, Amsterdam.

Stokes, S. L. (1977). Ranked set sampling with concomitant variables, *Comm. Statist. Theory Methods*, **6**, 1207–1211.

Yu, P. L. H. and Lam, K. (1997). Regression estimator in ranked set sampling, *Biometrics*, **53**, 1070–1080.

Yu, P. L. H., Sun, Y. and Sinha, B. K. (1999). On exact confidence intervals for the common mean of several normal populations, *J. Statist. Plann. Inference*, **81**(2), 263–277.

Ann. Inst. Statist. Math.
Vol. 54, No. 4, 879–899 (2002)

UNIVERSAL CONSISTENCY OF LOCAL POLYNOMIAL KERNEL REGRESSION ESTIMATES*

MICHAEL KOHLER

Mathematisches Institut A, Universität Stuttgart, Pfaffenwaldring 57, D-70569 Stuttgart, Germany,
e-mail: kohler@mathematik.uni-stuttgart.de

(Received December 18, 2000; revised May 25, 2001)

Abstract. Regression function estimation from independent and identically distributed data is considered. The L_2 error with integration with respect to the design measure is used as an error criterion. It is shown that suitably defined local polynomial kernel estimates are weakly and strongly universally consistent, i.e., it is shown that the L_2 errors of these estimates converge to zero almost surely and in L_1 for all distributions.

Key words and phrases: Local polynomial kernel estimates, regression estimates, weak and strong universal consistency.

1. Introduction

1.1 *Nonparametric regression function estimation*

Let (X,Y), (X_1,Y_1), $(X_2,Y_2),\dots$ be independent identically distributed $\mathbb{R}^d \times \mathbb{R}$-valued random vectors with $\boldsymbol{E}Y^2 < \infty$. In regression analysis we want to estimate Y after having observed X, i.e. we want to determine a function f with $f(X)$ "close" to Y. If "closeness" is measured by the mean squared error, then one wants to find a function f^* such that

$$\boldsymbol{E}\{|f^*(X) - Y|^2\} = \min_f \boldsymbol{E}\{|f(X) - Y|^2\}. \tag{1.1}$$

Let $m(x) := \boldsymbol{E}\{Y \mid X = x\}$ be the regression function and denote the distribution of X by μ. The well-known relation which holds for each measurable function f

$$\boldsymbol{E}\{|f(X) - Y|^2\} = \boldsymbol{E}\{|m(X) - Y|^2\} + \int |f(x) - m(x)|^2 \mu(dx) \tag{1.2}$$

implies that m is the solution of the minimization problem (1), and for an arbitrary f, L_2 error $\int |f(x) - m(x)|^2\mu(dx)$ is the difference between $\boldsymbol{E}\{|f(X) - Y|^2\}$ and $\boldsymbol{E}\{|m(X) - Y|^2\}$—the minimum of (1.2).

In the regression estimation problem the distribution of (X,Y) (and consequently m) is unknown. Given a sequence $\mathcal{D}_n = \{(X_1,Y_1),\dots,(X_n,Y_n)\}$ of independent observations of (X,Y), our goal is to construct an estimate $m_n(x) = m_n(x, \mathcal{D}_n)$ of $m(x)$ such that the L_2 error $\int |m_n(x) - m(x)|^2\mu(dx)$ is small.

*Research supported by DAAD.

1.2 *Universal consistency*

A sequence of estimators $(m_n)_{n\in\mathbb{N}}$ is called **weakly universally consistent** if $\mathbf{E}\int|m_n(x)-m(x)|^2\mu(dx)\to 0$ $(n\to\infty)$ for all distributions of (X,Y) with $\mathbf{E}Y^2<\infty$. It is called **strongly universally consistent** if $\int|m_n(x)-m(x)|^2\mu(dx)\to 0$ $(n\to\infty)$ a.s. for all distributions of (X,Y) with $\mathbf{E}Y^2<\infty$.

Stone (1977) first pointed out that there exist weakly universally consistent estimators. He considered *k_n-nearest neighbor estimates*

$$m_n(x)=\sum_{i=1}^{n}W_{n,i}(x)\cdot Y_i \tag{1.3}$$

where

$$W_{n,i}(x)=W_{n,i}(x,X_1,\dots,X_n) \tag{1.4}$$

is one if X_i is among the k_n-nearest neighbors of x in $\{X_1,\dots,X_n\}$ and zero otherwise, and where $k_n\to\infty$ and $k_n/n\to 0$ $(n\to\infty)$. The strong universal consistency of nearest neighbor estimates has been shown in Devroye *et al.* (1994).

Estimates of the form (1.3) with weight functions (1.4) are called local averaging estimates. *Kernel estimates* belong to the class of these estimates. There

$$W_{n,i}(x)=\frac{K\left(\dfrac{x-X_i}{h_n}\right)}{\sum_{j=1}^{n}K\left(\dfrac{x-X_j}{h_n}\right)}$$

$(0/0=0$ by definition) for some kernel function $K:\mathbb{R}^d\to\mathbb{R}_+$ and bandwidth $h_n>0$. Another example of local averaging estimates are *partitioning estimates*, which depend on a partition $\mathcal{P}_n=\{A_{n,1},A_{n,2},\dots\}$ of $\mathbb{R}^d$. There the weights (1.4) are defined by

$$W_{n,i}(x)=\frac{I_{A_n(x)}(X_i)}{\sum_{j=1}^{n}I_{A_n(x)}(X_j)},$$

where $A_n(x)=A_{n,j}$ if $x\in A_{n,j}$ and $I_{A_{n,j}}$ denotes the indicator function of $A_{n,j}$.

The weak universal consistency of kernel estimates has been shown under certain conditions on h_n and K independently by Devroye and Wagner (1980) and Spiegelman and Sachs (1980). The corresponding result for partitioning estimates has been obtained by Györfi (1991). The strong universal consistency of kernel and partitioning estimates for suitably defined kernels, sequences of bandwidths and sequences of partitions has been shown by Walk (2002). Various results concerning consistency of variants of kernel and partitioning estimates can be found in Devroye and Krzyżak (1989), Nobel (1996), Györfi and Walk (1996, 1997) and Györfi *et al.* (1998).

It is easy to see that the partitioning estimate minimizes the so-called empirical L_2 risk

$$\frac{1}{n}\sum_{i=1}^{n}|f(X_i)-Y_i|^2 \tag{1.5}$$

over the class of all real-valued functions f which are piecewise constant with respect to $\mathcal{P}_n$. *Least squares estimates* are defined by minimizing the empirical L_2 risk over general classes of functions (consisting e.g. of piecewise polynomials). The weak and strong universal consistency of various least squares estimates has been shown in Lugosi and Zeger (1995) and Kohler (1997, 1999).

Instead of minimizing the empirical L_2 risk (1.5) over some small class of functions one can also add a penalty term to (1.5) which penalizes the roughness of a function (e.g. a constant times the squared integral of the second derivative of f) and minimize the resulting sum over basically all functions (see Eubank (1988) or Wahba (1990) for details). The strong universal consistency of such *smoothing spline estimates* has been shown in Kohler and Krzyżak (2001).

1.3 *Local polynomial kernel estimates*

It is easy to see that the kernel estimate

$$m_n(x) = \frac{\sum_{i=1}^n K\left(\frac{x-X_i}{h_n}\right) Y_i}{\sum_{j=1}^n K\left(\frac{x-X_j}{h_n}\right)}$$

satisfies for each $x \in \mathbb{R}^d$

$$\frac{1}{n}\sum_{i=1}^n |m_n(x) - Y_i|^2 K\left(\frac{x-X_i}{h_n}\right) = \min_{a\in\mathbb{R}} \frac{1}{n}\sum_{i=1}^n |a - Y_i|^2 K\left(\frac{x-X_i}{h_n}\right).$$

Instead of fitting locally a constant to the data, the *local polynomial kernel estimate* fits locally a polynomial of some fixed degree M to the data, i.e., it is defined by

$$m_n(x) = \hat{p}_x(x) \tag{1.6}$$

where

$$\begin{aligned} \hat{p}_x(\cdot) &= \hat{p}_x(\cdot, \mathcal{D}_n) \in \mathcal{F}_M \\ &= \left\{ \sum_{0\le j_1,\ldots,j_d\le M} a_{j_1,\ldots,j_d} \cdot (x^{(1)})^{j_1} \cdot \ldots \cdot (x^{(d)})^{j_d} : a_{j_1,\ldots,j_d} \in \mathbb{R} \right\} \end{aligned} \tag{1.7}$$

satisfies

$$\frac{1}{n}\sum_{i=1}^n |\hat{p}_x(X_i) - Y_i|^2 K\left(\frac{x-X_i}{h_n}\right) = \min_{p\in\mathcal{F}_M} \frac{1}{n}\sum_{i=1}^n |p(X_i) - Y_i|^2 K\left(\frac{x-X_i}{h_n}\right). \tag{1.8}$$

Local polynomial kernel estimates have been considered by many authors, see e.g. the monographs Härdle (1990), Korostelev and Tsybakov (1993) and Fan and Gijbels (1996) and the literature cited therein.

1.4 *Main results*

As defined in the previous subsection, local polynomial kernel estimates are in general not weakly consistent, even if the regression function is smooth and the distribution of X is nice (Devroye (1998), personal communication): Let X be uniformly distributed on $[0, 1]$, Y be uniformly distributed on $\{-1, 1\}$ and assume that X and Y are independent. Then it can been shown that the local linear estimate m_n defined by (1.6)–(1.8) with $M = 1$ and $K = I_{[-1,1]}$ satisfies $\mathbf{E} \int |m_n(x) - m(x)|^2 \mu(dx) = \infty$ for all n and all $h_n > 0$. The proof of this fact uses that if an interval of length h_n contains exactly two of the X_i's, if the corresponding Y_i's are different and if all other X_j's are more than h_n away from this interval, then the estimate will be on this interval equal to the line which interpolates the two data points with x-values in this interval. This line can have an arbitrary large slope and therefore also the estimate can take arbitrary large values on this interval.

In this paper we modify the definition (1.6)–(1.8). We minimize in (1.8) only over those polynomials whose coefficients are bounded in absolute value by some constant which depends on n and tends to infinity for n tending to infinity. We show that this modified local polynomial kernel estimate is, under some mild conditions on the kernel and the bandwidths, weakly and strongly consistent for all distributions of (X, Y) with X bounded and Y square integrable. Furthermore we show, that if we set this estimate to zero outside of some cube which depends on n and tends to $\mathbb{R}^d$ for n tending to infinity, then the resulting estimate is weakly and strongly universally consistent.

1.5 *Main idea in the proof*

Let $g : \mathbb{R}^d \to \mathbb{R}$ be a square integrable function. Under some regularity conditions on the kernel the generalized Lebesgue density theorem implies that for μ-almost all x the pointwise error $|g(x) - m(x)|^2$ can be approximated for sufficiently small $h > 0$ by

$$\frac{\int |g(z) - m(z)|^2 \frac{1}{h^d} K\left(\frac{x-z}{h}\right) \mu(dz)}{\int \frac{1}{h^d} K\left(\frac{x-z}{h}\right) \mu(dz)}.$$

The nominator in the above integral is equal to

$$\begin{aligned}
&\mathbf{E}\left\{|g(X) - m(X)|^2 \frac{1}{h^d} K\left(\frac{x-X}{h}\right)\right\} \\
&\quad = \mathbf{E}\left\{|Y - g(X)|^2 \frac{1}{h^d} K\left(\frac{x-X}{h}\right)\right\} - \mathbf{E}\left\{|Y - m(X)|^2 \frac{1}{h^d} K\left(\frac{x-X}{h}\right)\right\}.
\end{aligned}$$

By the strong law of large numbers this term is close to

$$\frac{1}{n} \sum_{i=1}^{n} (|Y_i - g(X_i)|^2 - |Y_i - m(X_i)|^2) \frac{1}{h^d} K\left(\frac{x - X_i}{h}\right),$$

if n is large. In the definition of the local polynomial kernel estimate the function g is chosen such that the last term is small.

The main difficulty in the proof is to show that the previous approximations also hold if g is chosen in some data-dependent way from some fixed set of polynomials.

To prove that in this case the Lebesgue density theorem still holds we use that in the definition of the estimate we consider only polynomials, whose coefficients are bounded by some data independent constant. This implies that these polynomials satisfy some Lipschitz condition for some constant, which doesn't depend on the data.

To prove that in this case also something similar to the strong law of large numbers holds, we use techniques from empirical process theory.

1.6 *Notation*

$\mathbb{N}$, $\mathbb{R}$ and $\mathbb{R}_+$ are the sets of natural, real and nonnegative real numbers, respectively. I_A denotes the indicator function, $card(A)$ the cardinality of a set A. The natural logarithm is denoted by $\log(\cdot)$.

The euclidean norm of $x \in \mathbb{R}^d$ is denoted by $\|x\|$, the components of x are denoted by $x^{(1)}, \dots, x^{(d)}$. For a function $f : \mathbb{R}^d \to \mathbb{R}$ set

$$\|f\|_\infty = \sup_{x \in \mathbb{R}^d} |f(x)| \quad \text{and} \quad \|f\|^2 = \int_{\mathbb{R}^d} |f(x)|^2 \mu(dx).$$

For $h > 0$, $z \in \mathbb{R}^d$ and $K : \mathbb{R}^d \to \mathbb{R}$ define

$$K_h(z) = \frac{1}{h^d} K\left(\frac{z}{h}\right).$$

$C_0^\infty(\mathbb{R}^d)$ is the set of all real-valued functions on $\mathbb{R}^d$ which are infinitely often differentiable and have compact support, $supp(X)$ is the support of the distribution of the random variable X.

1.7 *Outline*

The main results are stated in Section 2 and proven in Section 3. In the appendix a list of some results of empirical process theory, which are used in the proofs, is given.

2. Main results

Let $M \in \mathbb{N}_0$ and β_n, $h_n > 0$. Set

$$\mathcal{F}_M(\beta_n) = \left\{ \sum_{0 \le j_1, \dots, j_d \le M} a_{j_1,\dots,j_d} \cdot (x^{(1)})^{j_1} \cdot \ldots \cdot (x^{(d)})^{j_d} : |a_{j_1,\dots,j_d}| \le \beta_n \right\}.$$

For given data $\mathcal{D}_n$ and $x \in \mathbb{R}^d$ choose

$$\hat{p}_x(\cdot) = \hat{p}_x(\cdot, \mathcal{D}_n) \in \mathcal{F}_M(\beta_n) \tag{2.1}$$

such that

$$\begin{aligned} &\frac{1}{n} \sum_{i=1}^n |Y_i - \hat{p}_x(X_i)|^2 K_{h_n}(x - X_i) \\ &\le \inf_{p \in \mathcal{F}_M(\beta_n)} \left(\frac{1}{n} \sum_{i=1}^n |Y_i - p(X_i)|^2 K_{h_n}(x - X_i) + \frac{1}{n} \right), \end{aligned} \tag{2.2}$$

and set

$$m_n(x) = \hat{p}_x(x). \tag{2.3}$$

THEOREM 2.1. *Let $\tilde{K} : \mathbb{R}_+ \to \mathbb{R}_+$ be a monotone decreasing and left—continuous function which satisfies*

$$b \cdot I_{[0,r^2]}(v) \le \tilde{K}(v) \le B \cdot I_{[0,R^2]}(v) \qquad (v \in \mathbb{R}_+)$$

for some $0 < r \leq R < \infty$, $0 < b \leq B < \infty$. *Define the kernel* $K : \mathbb{R}^d \to \mathbb{R}$ *by*

$$K(u) = \tilde{K}(\|u\|^2) \qquad (u \in \mathbb{R}^d).$$

Let $M \in \mathbb{N}_0$. *For* $n \in \mathbb{N}$ *choose* β_n, $h_n > 0$ *such that*

$$\beta_n \to \infty \qquad (n \to \infty), \tag{2.4}$$

$$h_n \cdot \beta_n^2 \to 0 \qquad (n \to \infty) \tag{2.5}$$

and

$$\frac{n \cdot h_n^d}{\beta_n^2 \cdot \log(n)} \to \infty \qquad (n \to \infty). \tag{2.6}$$

Let the estimate m_n *be defined by* (2.1)–(2.3). *Then*

$$\int |m_n(x) - m(x)|^2 \mu(dx) \to 0 \qquad (n \to \infty) \quad a.s.$$

and

$$\mathbf{E}\left\{\int |m_n(x) - m(x)|^2 \mu(dx)\right\} \to 0 \qquad (n \to \infty)$$

for every distribution of (X, Y) *with* $\|X\|$ *bounded a.s. and* $\mathbf{E}Y^2 < \infty$.

In Theorem 2.1 we need boundedness of $\|X\|$ to ensure that the estimate is weakly and strongly consistent. This assumption can be avoided, if we set the estimate to zero outside of a cube which depends on the sample size n and tends to $\mathbb{R}^d$ for n tending to infinity:

THEOREM 2.2. *Let* $\tilde{K} : \mathbb{R}_+ \to \mathbb{R}_+$ *be a monotone decreasing and left-continuous function which satisfies*

$$b \cdot I_{[0,r^2]}(v) \leq \tilde{K}(v) \leq B \cdot I_{[0,R^2]}(v) \qquad (v \in \mathbb{R}_+)$$

for some $0 < r \leq R < \infty$, $0 < b \leq B < \infty$. *Define the kernel* $K : \mathbb{R}^d \to \mathbb{R}$ *by*

$$K(u) = \tilde{K}(\|u\|^2) \qquad (u \in \mathbb{R}^d).$$

Let $M \in \mathbb{N}_0$. *For* $n \in \mathbb{N}$ *choose* A_n, β_n, $h_n > 0$ *such that*

$$A_n \to \infty \qquad (n \to \infty), \tag{2.7}$$

$$\beta_n \to \infty \qquad (n \to \infty), \tag{2.8}$$

$$h_n \cdot \beta_n^2 \cdot A_n^{2M \cdot d} \to 0 \quad (n \to \infty) \tag{2.9}$$

and

$$\frac{n \cdot h_n^d}{A_n^d \cdot \beta_n^2 \cdot \log(n)} \to \infty \qquad (n \to \infty). \tag{2.10}$$

Define m_n *by* (2.1)–(2.3) *and set* $\bar{m}_n(x) = m_n(x) \cdot I_{[-A_n, A_n]^d}(x)$. *Then*

$$\int |\bar{m}_n(x) - m(x)|^2 \mu(dx) \to 0 \qquad (n \to \infty) \quad a.s.$$

and

$$E\left\{\int |\bar{m}_n(x) - m(x)|^2 \mu(dx)\right\} \to 0 \quad (n \to \infty)$$

for every distribution of (X, Y) with $EY^2 < \infty$, i.e., $\bar{m}_n$ is weakly and strongly universally consistent.

Remark 1. We want to stress that in Theorem 2.2 there is no assumption on the underlying distribution of (X, Y) besides $EY^2 < \infty$. In particular it is not required that X have a density with respect to the Lebesgue-Borel measure or that m be (in some sense) smooth.

Remark 2. It is well-known that one cannot derive a non-trivial rate of convergence result for the L_2 error of any estimate without restricting the class of distributions considered, e.g. by assuming some smoothness property on m (see, e.g., Theorem 7.2 in Devroye *et al.* (1996) and Section 3 in Devroye and Wagner (1980)). Stone (1982) showed that local polynomial kernel estimates achieve, in probability, the optimal rate of convergence if the regression function is k-times continuously differentiable, $M \geq k$ and and the distribution of X has a density with respect to the Lebesgue-Borel measure which is bounded away from zero and infinity.

Remark 3. It follows from the proofs given below that Theorems 2.1 and 2.2 also hold if the bandwidth h of the estimate is chosen in an arbitrary data-driven way from some deterministic interval $[h_{min}(n), h_{max}(n)]$, where $h_{min}(n), h_{max}(n) \in \mathbb{R}_+$ satisfy (2.5) and (2.9) with h_n replaced by $h_{max}(n)$ and (2.6) and (2.10) with h_n replaced by $h_{min}(n)$.

Remark 4. Let $M = 0$. Then the kernel estimate satisfies (1.6)–(1.8). It is easy to see that if one truncates the kernel estimates at height $\pm\beta_n$, then this truncated kernel estimate satisfies (2.1)–(2.3). Hence Theorem 2.2 implies that a modified kernel estimate, which is truncated at height $\pm\beta_n$ and is set equal to zero outside of some cube tending to $\mathbb{R}^d$ for n tending to infinity, is weakly and strongly universally consistent. It follows from Devroye and Wagner (1980) and Spiegelman and Sachs (1980) that these modifications are not necessary in order to get weak universal consistency. Walk (2001) shows that under suitable assumptions on the kernel and the bandwidth (including the assumption that the bandwidth doesn't change for every n) these modifications are also not necessary to prove strong universal consistency.

3. Proofs

In the proof of Theorems 2.1 and 2.2 we will apply the following lemma.

LEMMA 3.1. *Assume that the kernel K satisfies the assumptions of Theorem 2.1. Then there exists a constant $c_1 \in \mathbb{R}_+$ such that for all $h > 0$ and all distributions μ of X the following three inequalities are valid:*

a) *For all $z \in \mathbb{R}^d$:*

$$\int \frac{K_h(x - z)}{E\{K_h(x - X)\}} \mu(dx) \leq c_1.$$

b) *For all $A \geq 1$:*

$$\int_{[-A,A]^d} \frac{1}{E\{K_h(x - X)\}} \mu(dx) \leq c_1 \cdot A^d.$$

c) *For all* $f : \mathbb{R}^d \to \mathbb{R}_+$:

$$\int \frac{\mathbf{E}\{f(X)K_h(x-X)\}}{\mathbf{E}\{K_h(x-X)\}}\mu(dx) \le c_1 \cdot \int f(x)\mu(dx).$$

PROOF. a) follows from Lemma 1 in Devroye and Wagner (1980). In order to prove b) choose $z_1, \dots, z_K \in \mathbb{R}^d$ such that the union of all balls $S_{r\cdot h}(z_i)$ of radius $r \cdot h$ around z_i cover $[-A, A]^d$ and $K \le c \cdot A^d \cdot h^{-d}$ for some constant c which depends only on d. Then

$$\begin{aligned}\int_{[-A,A]^d} \frac{1}{\mathbf{E}\{K_h(x-X)\}}\mu(dx) &\le \sum_{i=1}^K \int_{S_{r\cdot h}(z_i)} \frac{1}{\mathbf{E}\{K_h(x-X)\}}\mu(dx)\\ &\le \frac{1}{b}\cdot h^d \sum_{i=1}^K \int_{S_{r\cdot h}(z_i)} \frac{K_h(x-z_i)}{\mathbf{E}\{K_h(x-X)\}}\mu(dx).\end{aligned}$$

This together with a) implies the assertion of b). c) follows from a) and

$$\int \frac{\mathbf{E}\{f(X)K_h(x-X)\}}{\mathbf{E}\{K_h(x-X)\}}\mu(dx) = \int f(z) \int \frac{K_h(x-z)}{\mathbf{E}\{K_h(x-X)\}}\mu(dx)\mu(dz). \qquad \square$$

PROOF OF THEOREM 2.1. Choose $A \in \mathbb{R}_+$, $A > 1$ such that $supp(X) \subseteq [-A, A]^d$. Let L, $\epsilon > 0$ be arbitrary. Then there exists $\bar{m}_\epsilon \in C_0^\infty(\mathbb{R}^d)$ such that $\int |\bar{m}_\epsilon(x) - m(x)|^2\mu(dx) < \epsilon$. For $z \in \mathbb{R}$ set

$$T_L z = \begin{cases} L & \text{if } z > L, \\ z & \text{if } -L \le z \le L, \\ -L & \text{if } z < -L. \end{cases}$$

Set $Y_L = T_L Y$ and $Y_{i,L} = T_L Y_i$ $(i = 1, \dots, n)$. Without loss of generality we assume that n is so large that $\|\bar{m}_\epsilon\|_\infty \le \beta_n$ and $L \le \beta_n$.

In the first step of the proof we show

$$\begin{aligned}(3.1) \int &|m_n(x) - m(x)|^2\mu(dx)\\ &\le 4 \cdot \int \frac{\mathbf{E}\{|\hat{p}_x(X) - m(X)|^2 K_{h_n}(x-X) \mid \mathcal{D}_n\}}{\mathbf{E}\{K_{h_n}(x-X)\}}\mu(dx) + c_2 \cdot (\epsilon + A^{2M\cdot d}\beta_n^2 \cdot h_n)\end{aligned}$$

for some constant c_2 which depends only on M and d.

We have

$$\begin{aligned}(3.2) \int &|m_n(x) - m(x)|^2\mu(dx)\\ &\le 2\int |\hat{p}_x(x) - \bar{m}_\epsilon(x)|^2\mu(dx) + 2\int |\bar{m}_\epsilon(x) - m(x)|^2\mu(dx)\\ &\le 2\epsilon + 2\int \left(|\hat{p}_x(x) - \bar{m}_\epsilon(x)|^2 - \frac{\mathbf{E}\{|\hat{p}_x(X) - \bar{m}_\epsilon(X)|^2 K_{h_n}(x-X) \mid \mathcal{D}_n\}}{\mathbf{E}\{K_{h_n}(x-X)\}}\right)\mu(dx)\\ &\quad +4\int \frac{\mathbf{E}\{|\hat{p}_x(X) - m(X)|^2 K_{h_n}(x-X) \mid \mathcal{D}_n\}}{\mathbf{E}\{K_{h_n}(x-X)\}}\mu(dx)\\ &\quad +4\int \frac{\mathbf{E}\{|m(X) - \bar{m}_\epsilon(X)|^2 K_{h_n}(x-X) \mid \mathcal{D}_n\}}{\mathbf{E}\{K_{h_n}(x-X)\}}\mu(dx).\end{aligned}$$

By Lemma 3.1 c) the last integral is bounded by $c_1 \cdot \int |m(x) - \bar{m}_\epsilon(x)|^2 \mu(dx) \leq c_1 \epsilon$. In order to bound the first integral on the right-hand side of (3.2) observe that the first derivative of any $f \in \mathcal{F}_M(\beta_n)$ is on the cube $[-A, A]^d$ bounded (with respect to the euclidean norm) by $d \cdot M \cdot (M+1)^d A^{M \cdot d} \beta_n$. Hence by mean value theorem $|f(x) - f(u)| \leq c_3 \cdot A^{M \cdot d} \beta_n \cdot \|x - u\|$ for all $f \in \mathcal{F}_M(\beta_n)$ and all x, $u \in [-A, A]^d$. Here c_3 is a constant which depends only on M and d. Furthermore by definiton of $\mathcal{F}_M(\beta_n)$

$$\sup_{x \in [-A,A]^d} |f(x)| \leq (M+1)^d \cdot A^{M \cdot d} \cdot \beta_n \qquad (f \in \mathcal{F}_M(\beta_n)).$$

Because of $\bar{m}_\epsilon \in C_0^\infty(\mathbb{R}^d)$ we can assume without loss of generality that these two relations also hold for $f = \bar{m}_\epsilon$. We conclude that for all x, $u \in [-A, A]^d$ with $\|x - u\| \leq R \cdot h_n$ and all $f \in \mathcal{F}_M(\beta_n)$

$$\begin{aligned}
&\left| |f(x) - \bar{m}_\epsilon(x)|^2 - |f(u) - \bar{m}_\epsilon(u)|^2 \right| \\
&\quad = |(f(x) - f(u)) + (\bar{m}_\epsilon(u) - \bar{m}_\epsilon(x))| \cdot |f(x) + f(u) - \bar{m}_\epsilon(u) - \bar{m}_\epsilon(x)| \\
&\quad \leq 2 \cdot c_3 \cdot A^{M \cdot d} \beta_n \cdot \|x - u\| \cdot 4 \cdot (M+1)^d \cdot A^{M \cdot d} \cdot \beta_n \\
&\quad \leq c_4 \cdot A^{2M \cdot d} \beta_n^2 \cdot h_n.
\end{aligned}$$

From this, together with $K_{h_n}(x - u) = 0$ for $\|x - u\| > R \cdot h_n$, we get

$$\begin{aligned}
&\int \left(|\hat{p}_x(x) - \bar{m}_\epsilon(x)|^2 - \frac{\mathbf{E}\{|\hat{p}_x(X) - \bar{m}_\epsilon(X)|^2 K_{h_n}(x - X) \mid \mathcal{D}_n\}}{\mathbf{E}\{K_{h_n}(x - X)\}} \right) \mu(dx) \\
&\quad = \int \frac{\mathbf{E}\{(|\hat{p}_x(x) - \bar{m}_\epsilon(x)|^2 - |\hat{p}_x(X) - \bar{m}_\epsilon(X)|^2) K_{h_n}(x - X) \mid \mathcal{D}_n\}}{\mathbf{E}\{K_{h_n}(x - X)\}} \mu(dx) \\
&\quad \leq c_4 A^{2M \cdot d} \cdot \beta_n^2 \cdot h_n \cdot \int \frac{\mathbf{E}\{K_{h_n}(x - X) \mid \mathcal{D}_n\}}{\mathbf{E}\{K_{h_n}(x - X)\}} \mu(dx) \\
&\quad = c_4 A^{2M \cdot d} \cdot \beta_n^2 \cdot h_n.
\end{aligned}$$

This proves (3.1).

In the second step of the proof we bound $\mathbf{E}\{|\hat{p}_x(X) - m(X)|^2 K_{h_n}(x - X) \mid \mathcal{D}_n\}$ by a sum of several terms. For $x \in \mathbb{R}^d$ define $\bar{p}_x \in \mathcal{F}_M(\beta_n)$ by $\bar{p}_x(u) = \bar{m}_\epsilon(x)$ $(u \in \mathbb{R}^d)$. Then

$$\begin{aligned}
&\mathbf{E}\{|\hat{p}_x(X) - m(X)|^2 K_{h_n}(x - X) \mid \mathcal{D}_n\} \\
&\quad = \mathbf{E}\{|Y - \hat{p}_x(X)|^2 K_{h_n}(x - X) \mid \mathcal{D}_n\} - \mathbf{E}\{|Y - m(X)|^2 K_{h_n}(x - X)\} \\
&\quad = \mathbf{E}\{|Y - \hat{p}_x(X)|^2 K_{h_n}(x - X) \mid \mathcal{D}_n\} - (1+\epsilon)^3 \frac{1}{n} \sum_{i=1}^n |Y_i - \hat{p}_x(X_i)|^2 K_{h_n}(x - X_i) \\
&\qquad + (1+\epsilon)^3 \cdot \left(\frac{1}{n} \sum_{i=1}^n |Y_i - \hat{p}_x(X_i)|^2 K_{h_n}(x - X_i) - \frac{1}{n} \sum_{i=1}^n |Y_i - \bar{p}_x(X_i)|^2 K_{h_n}(x - X_i) \right) \\
&\qquad + (1+\epsilon)^3 \frac{1}{n} \sum_{i=1}^n |Y_i - \bar{p}_x(X_i)|^2 K_{h_n}(x - X_i) - (1+\epsilon)^5 \mathbf{E}\{|Y - m(X)|^2 K_{h_n}(x - X)\} \\
&\qquad + ((1+\epsilon)^5 - 1) \mathbf{E}\{|Y - m(X)|^2 K_{h_n}(x - X)\} \\
&\quad = \sum_{j=1}^4 T_{j,n}(x).
\end{aligned}$$

In the next steps we give an upper bound for

$$\int \frac{T_{j,n}(x)}{\boldsymbol{E}\{K_{h_n}(x-X)\}}\mu(dx) \tag{3.3}$$

($j \in \{1,2,3,4\}$).

In the third step of the proof we show

$$\int \frac{T_{4,n}(x)}{\boldsymbol{E}\{K_{h_n}(x-X)\}}\mu(dx) \leq \left((1+\epsilon)^5-1\right)\cdot c_1 \boldsymbol{E}\left\{|Y-m(X)|^2\right\}. \tag{3.4}$$

By Lemma 3.1 c) we get

$$\begin{aligned}
&\int \frac{T_{4,n}(x)}{\boldsymbol{E}\{K_{h_n}(x-X)\}}\mu(dx) \\
&\quad = ((1+\epsilon)^5-1)\int \frac{\boldsymbol{E}\{\boldsymbol{E}\{|Y-m(X)|^2 \mid X\}K_{h_n}(x-X)\}}{\boldsymbol{E}\{K_{h_n}(x-X)\}}\mu(dx) \\
&\quad \leq ((1+\epsilon)^5-1)\cdot c_1 \int \boldsymbol{E}\{|Y-m(X)|^2 \mid X=x\}\mu(dx) \\
&\quad = ((1+\epsilon)^5-1)\cdot c_1\boldsymbol{E}\{|Y-m(X)|^2\},
\end{aligned}$$

which proves (3.4).

In the fourth step of the proof we show

$$\begin{aligned}
&\limsup_{n\to\infty} \int \frac{T_{3,n}(x)}{\boldsymbol{E}\{K_{h_n}(x-X)\}}\mu(dx) \\
&\quad \leq 2c_1\left(1+\frac{1}{\epsilon}\right)\cdot(1+\epsilon)^4\boldsymbol{E}\{|Y-Y_L|^2\}+c_1(1+\epsilon)^5\epsilon \quad \text{a.s.}
\end{aligned} \tag{3.5}$$

and

$$\begin{aligned}
&\limsup_{n\to\infty} \boldsymbol{E}\int \frac{T_{3,n}(x)}{\boldsymbol{E}\{K_{h_n}(x-X)\}}\mu(dx) \\
&\quad \leq 2c_1\left(1+\frac{1}{\epsilon}\right)\cdot(1+\epsilon)^4\boldsymbol{E}\left\{|Y-Y_L|^2\right\}+c_1(1+\epsilon)^5\epsilon.
\end{aligned} \tag{3.6}$$

We use the decomposition

$$\begin{aligned}
&T_{3,n}(x) \\
&= (1+\epsilon)^3\frac{1}{n}\sum_{i=1}^n |Y_i-\bar{m}_\epsilon(x)|^2K_{h_n}(x-X_i)-(1+\epsilon)^5\boldsymbol{E}\{|Y-m(X)|^2K_{h_n}(x-X)\} \\
&= (1+\epsilon)^3\left(\tfrac{1}{n}\sum_{i=1}^n |Y_i-\bar{m}_\epsilon(x)|^2K_{h_n}(x-X_i)-(1+\epsilon)\cdot\tfrac{1}{n}\sum_{i=1}^n |Y_{i,L}-\bar{m}_\epsilon(x)|^2K_{h_n}(x-X_i)\right) \\
&\quad +(1+\epsilon)^4\left(\tfrac{1}{n}\sum_{i=1}^n |Y_{i,L}-\bar{m}_\epsilon(x)|^2K_{h_n}(x-X_i)-\tfrac{1}{n}\sum_{i=1}^n |Y_{i,L}-\bar{m}_\epsilon(X_i)|^2K_{h_n}(x-X_i)\right) \\
&\quad +(1+\epsilon)^4\left(\frac{1}{n}\sum_{i=1}^n |Y_{i,L}-\bar{m}_\epsilon(X_i)|^2K_{h_n}(x-X_i)-\boldsymbol{E}\{|Y_L-\bar{m}_\epsilon(X)|^2K_{h_n}(x-X)\}\right) \\
&\quad +(1+\epsilon)^4(\boldsymbol{E}\{|Y_L-\bar{m}_\epsilon(X)|^2K_{h_n}(x-X)\}-(1+\epsilon)\boldsymbol{E}\{|Y-\bar{m}_\epsilon(X)|^2K_{h_n}(x-X)\})
\end{aligned}$$

$$+(1+\epsilon)^5(\boldsymbol{E}\{|Y-\bar{m}_\epsilon(X)|^2 K_{h_n}(x-X)\}-\boldsymbol{E}\{|Y-m(X)|^2 K_{h_n}(x-X)\})$$
$$=\sum_{j=5}^{9} T_{j,n}.$$

Using $(a+b)^2 \le (1+\frac{1}{\epsilon})a^2+(1+\epsilon)b^2$ $(a,b \in \mathbb{R})$ we get

$$T_{5,n}(x) \le \left(1+\frac{1}{\epsilon}\right)\cdot(1+\epsilon)^3\frac{1}{n}\sum_{i=1}^{n}|Y_i-Y_{i,L}|^2 K_{h_n}(x-X_i)$$

and

$$T_{8,n}(x) \le \left(1+\frac{1}{\epsilon}\right)\cdot(1+\epsilon)^4\boldsymbol{E}\left\{|Y_L-Y|^2 K_{h_n}(x-X)\right\}.$$

Hence by Lemma 3.1 a)

$$\text{(3.7)}\qquad \int\frac{T_{5,n}(x)}{\boldsymbol{E}\{K_{h_n}(x-X)\}}\mu(dx)$$
$$\le \left(1+\frac{1}{\epsilon}\right)\cdot(1+\epsilon)^3\frac{1}{n}\sum_{i=1}^{n}|Y_i-Y_{i,L}|^2\int\frac{K_{h_n}(x-X_i)}{\boldsymbol{E}\{K_{h_n}(x-X)\}}\mu(dx)$$
$$\le \left(1+\frac{1}{\epsilon}\right)\cdot(1+\epsilon)^3 c_1\frac{1}{n}\sum_{i=1}^{n}|Y_i-Y_{i,L}|^2,$$

and by Lemma 3.1 c)

$$\text{(3.8)}\qquad \int\frac{T_{8,n}(x)}{\boldsymbol{E}\{K_{h_n}(x-X)\}}\mu(dx)$$
$$\le \left(1+\frac{1}{\epsilon}\right)\cdot(1+\epsilon)^4\int\frac{\boldsymbol{E}\left\{\boldsymbol{E}\{|Y_L-Y|^2 \mid X\}K_{h_n}(x-X)\right\}}{\boldsymbol{E}\{K_{h_n}(x-X)\}}\mu(dx)$$
$$\le \left(1+\frac{1}{\epsilon}\right)\cdot(1+\epsilon)^4 c_1\boldsymbol{E}\left\{|Y-Y_L|^2\right\}.$$

Furthermore

$$T_{6,n}(x)$$
$$=(1+\epsilon)^4\frac{1}{n}\sum_{i=1}^{n}(Y_{i,L}-\bar{m}_\epsilon(x)-(Y_{i,L}-\bar{m}_\epsilon(X_i)))$$
$$\cdot(Y_{i,L}-\bar{m}_\epsilon(x)+Y_{i,L}-\bar{m}_\epsilon(X_i))\,K_{h_n}(x-X_i)$$
$$\le (1+\epsilon)^4(2L+2\|\bar{m}_\epsilon\|_\infty)\cdot\sup_{\|u-v\|\le R\cdot h_n}|\bar{m}_\epsilon(u)-\bar{m}_\epsilon(v)|\cdot\frac{1}{n}\sum_{i=1}^{n}K_{h_n}(x-X_i),$$

which together with Lemma 3.1 a) implies

$$\int\frac{T_{6,n}(x)}{\boldsymbol{E}\{K_{h_n}(x-X)\}}\mu(dx) \le (1+\epsilon)^4(2L+2\|\bar{m}_\epsilon\|_\infty)\cdot\sup_{\|u-v\|\le R\cdot h_n}|\bar{m}_\epsilon(u)-\bar{m}_\epsilon(v)|\cdot c_1.$$

Because of $\bar{m}_\epsilon \in C_0^\infty(\mathbb{R}^d)$ this together with $h_n \to 0$ $(n \to \infty)$ implies

$$\limsup_{n\to\infty} \int \frac{T_{6,n}(x)}{\boldsymbol{E}\{K_{h_n}(x-X)\}} \mu(dx) \leq 0 \tag{3.9}$$

and

$$\limsup_{n\to\infty} \boldsymbol{E} \int \frac{T_{6,n}(x)}{\boldsymbol{E}\{K_{h_n}(x-X)\}} \mu(dx) \leq 0. \tag{3.10}$$

Next, we observe

$$\begin{aligned}
&\int \frac{T_{7,n}(x)}{\boldsymbol{E}\{K_{h_n}(x-X)\}} \mu(dx) \\
&\quad = (1+\epsilon)^4 \left(\frac{1}{n} \sum_{i=1}^n |Y_{i,L} - \bar{m}_\epsilon(X_i)|^2 \cdot \int \frac{K_{h_n}(x-X_i)}{\boldsymbol{E}\{K_{h_n}(x-X)\}} \mu(dx) \right. \\
&\qquad\qquad \left. - \boldsymbol{E}\left\{ |Y_L - \bar{m}_\epsilon(X)|^2 \cdot \int \frac{K_{h_n}(x-X)}{\boldsymbol{E}\{K_{h_n}(x-X)\}} \mu(dx) \right\} \right) \\
&\quad = (1+\epsilon)^4 \left(\frac{1}{n} \sum_{i=1}^n Z_{i,n} - \boldsymbol{E}\{Z_{1,n}\} \right).
\end{aligned}$$

The random variables $Z_{1,n}, \ldots, Z_{n,n}$ are independent and identically distributed. It follows from Lemma 3.1 a) that they take, with probability one, only values in an interval of length $c_1(2L^2 + 2\|\bar{m}_\epsilon\|_\infty^2)$. Hence Hoeffding's inequality together with Borel-Cantelli lemma imply

$$\frac{1}{n} \sum_{i=1}^n Z_{i,n} - \boldsymbol{E}\{Z_{1,n}\} \to 0 \quad (n \to \infty) \quad \text{a.s.}$$

This proves

$$\limsup_{n\to\infty} \int \frac{T_{7,n}(x)}{\boldsymbol{E}\{K_{h_n}(x-X)\}} \mu(dx) = 0 \quad \text{a.s.} \tag{3.11}$$

Furthermore, independence and identical distribution of $Z_{1,n}, \ldots, Z_{n,n}$ imply

$$\boldsymbol{E} \int \frac{T_{7,n}(x)}{\boldsymbol{E}\{K_{h_n}(x-X)\}} \mu(dx) = 0 \quad (n \in \mathbb{N}). \tag{3.12}$$

Finally by Lemma 3.1 c) and definition of $\bar{m}_\epsilon$ we get

$$\begin{aligned}
&\int \frac{T_{9,n}(x)}{\boldsymbol{E}\{K_{h_n}(x-X)\}} \mu(dx) \\
&\quad = (1+\epsilon)^5 \int \frac{\boldsymbol{E}\{|\bar{m}_\epsilon(X) - m(X)|^2 K_{h_n}(x-X)\}}{\boldsymbol{E}\{K_{h_n}(x-X)\}} \mu(dx) \\
&\quad \leq (1+\epsilon)^5 c_1 \int |\bar{m}_\epsilon(x) - m(x)|^2 \mu(dx) \\
&\quad \leq (1+\epsilon)^5 c_1 \epsilon.
\end{aligned}$$

This together with (3.7)–(3.12) and the strong law of large numbers implies (3.5) and (3.6).

In the fifth step of the proof we show

$$\begin{aligned}&\limsup_{n\to\infty}\int \frac{T_{2,n}(x)}{\boldsymbol{E}\{K_{h_n}(x-X)\}}\mu(dx)\le 0 \quad \text{and}\\ &\limsup_{n\to\infty}\boldsymbol{E}\int \frac{T_{2,n}(x)}{\boldsymbol{E}\{K_{h_n}(x-X)\}}\mu(dx)\le 0.\end{aligned}\tag{3.13}$$

By definition of $\hat{p}_x$

$$T_{2,n}(x)\le (1+\epsilon)^3\frac{1}{n}.$$

This together with Lemma 3.1 b) and $supp(X)\subseteq[-A,A]^d$ implies

$$\int \frac{T_{2,n}(x)}{\boldsymbol{E}\{K_{h_n}(x-X)\}}\mu(dx)\le (1+\epsilon)^3\frac{1}{n}\cdot c_1A^d,$$

which in turn implies (3.13).

In the sixth step of the proof we show

$$\begin{aligned}&\int \frac{T_{1,n}(x)}{\boldsymbol{E}\{K_{h_n}(x-X)\}}\mu(dx)\\ &\le c_1\left(1+\frac{1}{\epsilon}\right)\boldsymbol{E}\{|Y-Y_L|^2\}+c_1(1+\epsilon)^2\left(1+\frac{1}{\epsilon}\right)\frac{1}{n}\sum_{i=1}^n|Y_{i,L}-Y_i|^2\\ &+c_1(1+\epsilon)A^dT_{10,n},\end{aligned}\tag{3.14}$$

where

$$\begin{aligned}T_{10,n}=\sup_{f\in\mathcal{F}_M(\beta_n),z\in\mathbb{R}^d}\Bigg(&\boldsymbol{E}\{|Y_L-f(X)|^2K_{h_n}(z-X)\}\\ &-(1+\epsilon)\cdot\frac{1}{n}\sum_{i=1}^n|Y_{i,L}-f(X_i)|^2K_{h_n}(z-X_i)\Bigg).\end{aligned}\tag{3.15}$$

We use the decomposition

$$\begin{aligned}&T_{1,n}(x)\\ &=\boldsymbol{E}\{|Y-\hat{p}_x(X)|^2K_{h_n}(x-X)\mid\mathcal{D}_n\}-(1+\epsilon)\boldsymbol{E}\{|Y_L-\hat{p}_x(X)|^2K_{h_n}(x-X)\mid\mathcal{D}_n\}\\ &\quad+(1+\epsilon)\boldsymbol{E}\{|Y_L-\hat{p}_x(X)|^2K_{h_n}(x-X)\mid\mathcal{D}_n\}-(1+\epsilon)^2\tfrac{1}{n}\sum_{i=1}^n|Y_{i,L}-\hat{p}_x(X_i)|^2K_{h_n}(x-X_i)\\ &\quad+(1+\epsilon)^2\tfrac{1}{n}\sum_{i=1}^n|Y_{i,L}-\hat{p}_x(X_i)|^2K_{h_n}(x-X_i)-(1+\epsilon)^3\tfrac{1}{n}\sum_{i=1}^n|Y_i-\hat{p}_x(X_i)|^2K_{h_n}(x-X_i).\end{aligned}$$

Bounding the first and third terms on the right hand side as in the fourth step (cf. proof of (3.7) and (3.8)) we get

$$\begin{aligned}&\int \frac{T_{1,n}(x)}{\boldsymbol{E}\{K_{h_n}(x-X)\}}\mu(dx)\\ &\le c_1\left(1+\frac{1}{\epsilon}\right)\boldsymbol{E}\{|Y-Y_L|^2\}+c_1(1+\epsilon)^2\left(1+\frac{1}{\epsilon}\right)\frac{1}{n}\sum_{i=1}^n|Y_{i,L}-Y_i|^2\end{aligned}$$

$$+\int \left(\frac{(1+\epsilon)\boldsymbol{E}\{|Y_L - \hat{p}_x(X)|^2 K_{h_n}(x-X) \mid \mathcal{D}_n\}}{\boldsymbol{E}\{K_{h_n}(x-X)\}} - \frac{(1+\epsilon)^2 \frac{1}{n}\sum_{i=1}^n |Y_{i,L} - \hat{p}_x(X_i)|^2 K_{h_n}(x-X_i)}{\boldsymbol{E}\{K_{h_n}(x-X)\}} \right) \mu(dx).$$

The difference of the nominators in the integral above is bounded by $(1+\epsilon)$ times $T_{10,n}$. $T_{10,n}$ doesn't depend on x, hence the whole integral can be bounded by $T_{10,n}$ times

$$(1+\epsilon) \cdot \int \frac{1}{\boldsymbol{E}\{K_{h_n}(x-X)\}} \mu(dx).$$

Applying Lemma 3.1 b) to the last term yields (3.14).

In the seventh step of the proof we show

$$\limsup_{n\to\infty} T_{10,n} \le 0 \quad \text{a.s.} \quad \text{and} \quad \limsup_{n\to\infty} \boldsymbol{E}T_{10,n} \le 0. \tag{3.16}$$

To this end let $t > 0$ be arbitray. Then

$$\begin{aligned}
&\boldsymbol{P}\{T_{10,n} > t\} \\
&= \boldsymbol{P}\Bigg\{ \exists f \in \mathcal{F}_M(\beta_n), z \in \mathbb{R}^d : \\
&\qquad \boldsymbol{E}\{|Y_L - f(X)|^2 K_{h_n}(z-X)\} - (1+\epsilon)\cdot \frac{1}{n}\sum_{i=1}^n |Y_{i,L} - f(X_i)|^2 K_{h_n}(z-X_i) > t \Bigg\} \\
&\le \boldsymbol{P}\Bigg\{ \exists f \in \mathcal{F}_M(\beta_n), z \in \mathbb{R}^d : \\
&\qquad \frac{\boldsymbol{E}\{|Y_L - f(X)|^2 K_{h_n}(z-X)\} - \frac{1}{n}\sum_{i=1}^n |Y_{i,L} - f(X_i)|^2 K_{h_n}(z-X_i)}{t + \epsilon\cdot \boldsymbol{E}\{|Y_L - f(X)|^2 K_{h_n}(z-X)\}} > \frac{1}{1+\epsilon} \Bigg\} \\
&= \boldsymbol{P}\Bigg\{ \exists f \in \mathcal{F}_M(\beta_n), z \in \mathbb{R}^d : \\
&\qquad \frac{\boldsymbol{E}\left\{|Y_L - f(X)|^2 K\left(\frac{z-X}{h_n}\right)\right\} - \frac{1}{n}\sum_{i=1}^n |Y_{i,L} - f(X_i)|^2 K\left(\frac{z-X_i}{h_n}\right)}{\frac{t\cdot h_n^d}{\epsilon} + \boldsymbol{E}\left\{|Y_L - f(X)|^2 K\left(\frac{z-X}{h_n}\right)\right\}} > \frac{\epsilon}{1+\epsilon} \Bigg\}.
\end{aligned}$$

By Lemma A.1 in the Appendix, which uses the notion of covering numbers introduced in Definition A.1 in the Appendix, the last probability is bounded by

$$4 \cdot \boldsymbol{E}\mathcal{N}_1\left(\frac{t \cdot h_n^d}{8(1+\epsilon)}, \mathcal{G}, (X,Y)_1^n\right) \cdot \exp\left(-\frac{n \cdot \frac{t \cdot h_n^d}{\epsilon} \cdot \left(\frac{\epsilon}{1+\epsilon}\right)^2}{64B\beta_n^2}\right),$$

where

$$\mathcal{G} = \left\{g : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R} : g(x,y) = |T_L y - f(x)|^2 K\left(\frac{u-x}{h_n}\right) ((x,y) \in \mathbb{R}^d \times \mathbb{R}) \right.$$
$$\left. \text{for some } u \in \mathbb{R}^d, f \in \mathcal{F}_M(\beta_n)\right\}.$$

We will show in the eighth step of the proof that

$$\mathcal{N}_1\left(\frac{t \cdot h_n^d}{8(1+\epsilon)}, \mathcal{G}, (X,Y)_1^n\right) \le \left(c_5 \frac{(1+\epsilon) \cdot \beta_n^2 A^{2M \cdot d}}{t \cdot h_n^d}\right)^{c_6} \tag{3.17}$$

for some constants c_5 and c_6 which depend only on M, B and d. This implies

(3.18) $\boldsymbol{P}\{T_{10,n} > t\}$

$$\le 4\left(c_5 \frac{(1+\epsilon) \cdot \beta_n^2 A^{2M \cdot d}}{t \cdot h_n^d}\right)^{c_6} \exp\left(-\frac{n \cdot \frac{t \cdot h_n^d}{\epsilon} \cdot \left(\frac{\epsilon}{1+\epsilon}\right)^2}{64B\beta_n^2}\right)$$

$$= 4 \cdot \exp\left(-\log(n^2) \cdot \frac{nh_n^d}{\beta_n^2 2\log(n)} \cdot \left(\frac{t \cdot \epsilon}{64B(1+\epsilon)^2} - \frac{c_6 \log\left(c_5(1+\epsilon)\frac{\beta_n^2 A^{2Md}}{t \cdot h_n^d}\right)\beta_n^2}{nh_n^d}\right)\right).$$

The assumptions of Theorem 2.1 imply

$$\frac{nh_n^d}{\beta_n^2 2\log(n)} \to \infty \quad (n \to \infty)$$

and for n sufficiently large

$$\frac{c_6 \log\left(c_5(1+\epsilon)\frac{\beta_n^2 A^{2Md}}{t \cdot h_n^d}\right)\beta_n^2}{nh_n^d} \le \frac{c_6 \log(n)\,\beta_n^2}{nh_n^d} \to 0 \quad (n \to \infty).$$

It follows that the right-hand side of (3.18) is summable for each $t > 0$, hence the Borel-Cantelli lemma yields the first part of (3.16). In order to prove the second part, let $\delta > 0$ be arbitrary. Then

$$\boldsymbol{E}T_{10,n} \le \int_0^\infty \boldsymbol{P}\{T_{10,n} > t\}dt$$

$$
\begin{aligned}
&\leq \delta + \int_{\delta}^{\infty} 4\left(c_5 \frac{(1+\epsilon)\cdot\beta_n^2 A^{2M\cdot d}}{\delta\cdot h_n^d}\right)^{c_6} \exp\left(-\frac{n\cdot\frac{t\cdot h_n^d}{\epsilon}\cdot\left(\frac{\epsilon}{1+\epsilon}\right)^2}{64B\beta_n^2}\right) dt \\
&= \delta + 4\left(c_5 \frac{(1+\epsilon)\cdot\beta_n^2 A^{2M\cdot d}}{\delta\cdot h_n^d}\right)^{c_6} \cdot \frac{64B\beta_n^2(1+\epsilon)^2}{n\cdot h_n^d\epsilon} \exp\left(-\frac{n\cdot h_n^d}{\beta_n^2}\cdot\frac{\delta\cdot\epsilon}{64B(1+\epsilon)^2}\right) \\
&\to \delta \quad (n\to\infty)
\end{aligned}
$$

by the assumptions of Theorem 2.1. With $\delta \to 0$ the second part of (3.16) follows.

In the eighth step of the proof we show (3.17). Therefore we use arguments from the proof of Theorem 2 in Krzyżak *et al.* (1996). We have $\mathcal{G} = \{g_1 \cdot g_2 : g_1 \in \mathcal{G}_1, g_2 \in \mathcal{G}_2\}$, where

$$\mathcal{G}_1 = \{g : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R} : g(x,y) = |T_L y - f(x)|^2 ((x,y) \in \mathbb{R}^d \times \mathbb{R}) \text{ for some } f \in \mathcal{F}_M(\beta_n)\}$$

and

$$\mathcal{G}_2 = \left\{g : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R} : g(x,y) = K\left(\frac{u-x}{h_n}\right) ((x,y) \in \mathbb{R}^d \times \mathbb{R}) \text{ for some } u \in \mathbb{R}^d\right\}.$$

The functions in $\mathcal{G}_1$ and $\mathcal{G}_2$ are bounded on $[-A,A]^d \times \mathbb{R}$ in absolute value by

$$(2L^2 + 2(\beta_n(M+1)^d A^{M\cdot d})^2) \leq 4\beta_n^2(M+1)^{2d}A^{2M\cdot d}$$

and B, respectively. Hence by Lemma A.2 in the Appendix we get

$$
\begin{aligned}
\mathcal{N}_1\left(\frac{t\cdot h_n^d}{8(1+\epsilon)}, \mathcal{G}, (X,Y)_1^n\right) &\leq \mathcal{N}_1\left(\frac{t\cdot h_n^d}{16(1+\epsilon)B}, \mathcal{G}_1, (X,Y)_1^n\right) \\
&\quad \cdot \mathcal{N}_1\left(\frac{t\cdot h_n^d}{64(1+\epsilon)\beta_n^2(M+1)^{2d}A^{2M\cdot d}}, \mathcal{G}_2, (X,Y)_1^n\right).
\end{aligned}
$$

If $h_i(x,y) = |f_i(x) - T_L y|^2$ for some $f_i : [-A,A]^d \to \mathbb{R}$ bounded in absolute value by $\beta_n(M+1)^d A^{M\cdot d}$, then

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n |h_1(X_i,Y_i) - h_2(X_i,Y_i)|^2 \\
&\quad = \frac{1}{n}\sum_{i=1}^n |f_1(X_i) - T_L Y_i + f_2(X_i) - T_L Y_i| \cdot |f_1(X_i) - f_2(X_i)| \\
&\quad \leq (2L + 2\beta_n(M+1)^d A^{M\cdot d}) \cdot \frac{1}{n}\sum_{i=1}^n |f_1(X_i) - f_2(X_i)|
\end{aligned}
$$

which implies

$$
\begin{aligned}
&\mathcal{N}_1\left(\frac{t\cdot h_n^d}{16(1+\epsilon)B}, \mathcal{G}_1, (X,Y)_1^n\right) \\
&\quad \leq \mathcal{N}_1\left(\frac{t\cdot h_n^d}{16(1+\epsilon)B\,(2L + 2\beta_n(M+1)^d A^{M\cdot d})}, \mathcal{F}_M(\beta_n), X_1^n\right).
\end{aligned}
$$

Next we need the notion of VC dimension, which is introduced in Defintion A.2 in the Appendix. $\mathcal{F}_M(\beta_n)$ is a subset of a linear vector space of dimension $(M+1)^d$, hence by Lemma A.4 in the Appendix

$$V_{\mathcal{F}_M(\beta_n)^+} \leq (M+1)^d + 1 \leq (M+2)^d.$$

This together with Lemma A.3 in the Appendix implies

$$\mathcal{N}_1\left(\frac{t \cdot h_n^d}{16(1+\epsilon)B}, \mathcal{G}_1, (X,Y)_1^n\right) \leq 2\left(\frac{4e(M+1)^d\beta_n A^{M\cdot d}}{\frac{t\cdot h_n^d}{16(1+\epsilon)B\left(2L+2\beta_n(M+1)^d A^{M\cdot d}\right)}}\right)^{2(M+2)^d}$$
$$\leq \left(c_7 \cdot \frac{(1+\epsilon)\beta_n^2 A^{2M\cdot d}}{t\cdot h_n^d}\right)^{2(M+2)^d},$$

where c_7 is a constant which depends only on M, B and d.

Next we bound

$$\mathcal{N}_1\left(\frac{t\cdot h_n^d}{64(1+\epsilon)\beta_n^2(M+1)^{2d}A^{2M\cdot d}}, \mathcal{G}_2, (X,Y)_1^n\right).$$

By Lemma A.3 in the Appendix we get

$$\mathcal{N}_1\left(\frac{t\cdot h_n^d}{64(1+\epsilon)\beta_n^2(M+1)^{2d}A^{2M\cdot d}}, \mathcal{G}_2, (X,Y)_1^n\right)$$
$$\leq 2\left(\frac{4eB}{\frac{t\cdot h_n^d}{64(1+\epsilon)\beta_n^2(M+1)^{2d}A^{2M\cdot d}}}\right)^{2V_{\mathcal{G}_2^+}}$$
$$\leq \left(c_8\cdot\frac{(1+\epsilon)\beta_n^2 A^{2M\cdot d}}{t\cdot h_n^d}\right)^{2V_{\mathcal{G}_2^+}},$$

where c_8 is a constant which depends only on M, B and d. Hence it suffices to derive a bound on the VC dimension of the class of all subgraphs of

$$\mathcal{G}_2 = \left\{g : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R} : g(x,y) = \tilde{K}\left(\frac{\|u-x\|^2}{h_n^2}\right) ((x,y) \in \mathbb{R}^d \times \mathbb{R}) \text{ for some } u \in \mathbb{R}^d\right\}.$$

Since $\tilde{K}$ is left continuous and monotone decreasing we have

$$\tilde{K}\left(\frac{\|u-x\|^2}{h_n^2}\right) \geq t \text{ if and only if } \frac{\|u-x\|^2}{h_n^2} \leq \phi(t)$$

where $\phi(t) = \sup\{z : \tilde{K}(z) \geq t\}$. Equivalently, (x,y,t) must satisfy

$$x^T x - 2u^T x + u^T u - h_n^2\phi(t) \leq 0.$$

Consider now the set of real functions

$$\begin{aligned}\mathcal{G}_3 = \{g_{\alpha,\beta,\gamma,\delta} : \mathbb{R}^d \times \mathbb{R} \times \mathbb{R} \to \mathbb{R} : g_{\alpha,\beta,\gamma,\delta}(x,y,s) = \alpha x^T x + \beta^T x + \gamma s + \delta \\ ((x,y,s) \in \mathbb{R}^d \times \mathbb{R} \times \mathbb{R}) \text{ for some } \alpha, \gamma, \delta \in \mathbb{R}, \beta \in \mathbb{R}^d.\}.\end{aligned}$$

If for a given collection of points $\{(x_i, y_i, t_i)\}_{i=1,\dots,n}$ a set $\{(x,y,t) : g(x,y) \geq t\}$, $g \in \mathcal{G}_2$ picks out the points $\{(x_{i_1}, y_{i_1}, t_{i_1}), \dots, (x_{i_l}, y_{i_l}, t_{i_l})\}$ then there exist α, β, γ, δ such that $\{(x,y,s) : g_{\alpha,\beta,\gamma,\delta}(x,y,s) \geq 0\}$ picks out exactly $\{(x_{i_1}, y_{i_1}, \phi(t_{i_1})), \dots, (x_{i_l}, y_{i_l}, \phi(t_{i_l}))\}$ from $\{(x_1, y_1, \phi(t_1)), \dots, (x_n, y_n, \phi(t_n))\}$. This shows $V_{\mathcal{G}_2^+} \leq V_{\{\{(x,y,s): g(x,y,s) \geq 0\} : g \in \mathcal{G}_3\}}$. $\mathcal{G}_3$ is a linear vector space of dimension $d+3$, hence we can conclude from Lemma A.4 in the Appendix $V_{\mathcal{G}_2^+} \leq d+3$. Summarizing the above results we get

$$\begin{aligned}&\mathcal{N}_1\left(\frac{t \cdot h_n^d}{8(1+\epsilon)}, \mathcal{G}, (X,Y)_1^n\right) \\ &\leq \left(c_7 \cdot \frac{(1+\epsilon)\beta_n^2 A^{2M\cdot d}}{t \cdot h_n^d}\right)^{2(M+2)^d} \cdot \left(c_8 \cdot \frac{(1+\epsilon)\beta_n^2 A^{2M\cdot d}}{t \cdot h_n^d}\right)^{2(d+3)} \\ &\leq \left(c_5 \cdot \frac{(1+\epsilon)\beta_n^2 A^{2M\cdot d}}{t \cdot h_n^d}\right)^{c_6}\end{aligned}$$

for constants c_5 and c_6 which depend only on M, B and d.

In the ninth and last step of the proof we finish the proof by summarizing the above results. By the results of the first and second step we have

$$\begin{aligned}&\int |m_n(x) - m(x)|^2 \mu(dx) \\ &\leq c_2(\epsilon + A^{2M\cdot d}\beta_n^2 h_n) + 4\sum_{j=1}^{4} \int \frac{T_{j,n}(x)}{\mathbf{E}\{K_{h_n}(x-X)\}} \mu(dx).\end{aligned}$$

Using the results of steps three to seven and $\beta_n^2 h_n \to 0$ $(n \to \infty)$ one gets

$$\begin{aligned}&\limsup_{n\to\infty} \int |m_n(x) - m(x)|^2 \mu(dx) \\ &\quad \leq c_2\epsilon + 4((1+\epsilon)^5 - 1)c_1 \mathbf{E}|Y - m(X)|^2 \\ &\quad\quad + 8c_1(1+1/\epsilon)(1+\epsilon)^4 \mathbf{E}|Y - Y_L|^2 + 4c_1(1+\epsilon)^5\epsilon \\ &\quad\quad + 4c_1(1+1/\epsilon)\mathbf{E}|Y - Y_L|^2 + 4c_1(1+\epsilon)^2(1+1/\epsilon)\mathbf{E}|Y - Y_L|^2 \quad \text{a.s.}\end{aligned}$$

With $L \to \infty$ and $\epsilon \to 0$ this implies $\int |m_n(x) - m(x)|^2 \mu(dx) \to 0$ $(n \to \infty)$ a.s. The proof of $\mathbf{E} \int |m_n(x) - m(x)|^2 \mu(dx) \to 0$ $(n \to \infty)$ follows in an analogous way from the previous results. □

PROOF OF THEOREM 2.2. By definition of $\bar{m}_n$

$$\begin{aligned}&\int |\bar{m}_n(x) - m(x)|^2 \mu(dx) \\ &\quad = \int_{[-A_n, A_n]^d} |m_n(x) - m(x)|^2 \mu(dx) + \int_{\mathbb{R}^d \setminus [-A_n, A_n]^d} |m(x)|^2 \mu(dx).\end{aligned}$$

Because of $A_n \to \infty$ $(n \to \infty)$ and $\int |m(x)|^2 \mu(dx) < \infty$ we have

$$\int_{\mathbb{R}^d \setminus [-A_n, A_n]^d} |m(x)|^2 \mu(dx) \to 0 \quad (n \to \infty).$$

Hence it suffices to show

$$\int_{[-A_n, A_n]^d} |m_n(x) - m(x)|^2 \mu(dx) \to 0 \quad (n \to \infty)$$

a.s. and *in* L_1. This can be done by replacing in the proof of Theorem 2.1 A by A_n and $\int \ldots$ by $\int_{[-A_n,A_n]^d} \ldots$ Then one has to show in the seventh step

$$\limsup_{n\to\infty} A_n^d \cdot T_{10,n} \leq 0 \quad \text{a.s.} \qquad \text{and} \qquad \limsup_{n\to\infty} A_n^d \cdot \boldsymbol{E} T_{10,n} \leq 0.$$

To this end one uses

$$\begin{aligned} &\boldsymbol{P}\{A_n^d \cdot T_{10,n} > t\} \\ &= \boldsymbol{P}\left\{T_{10,n} > \frac{t}{A_n^d}\right\} \\ &\leq 4 \left(c_5 \frac{(1+\epsilon)\cdot \beta_n^2 A_n^{2M\cdot d}}{(t/A_n^d)\cdot h_n^d} \right)^{c_6} \exp\left(- \frac{n \cdot \frac{t \cdot h_n^d}{A_n^d \epsilon} \cdot \left(\frac{\epsilon}{1+\epsilon}\right)^2}{64 B \beta_n^2} \right) \end{aligned}$$

and proceeds otherwise as before. □

Acknowledgements

The author thanks two anonymous referees for several helpful comments.

Appendix

A. Some results of empirical process theory

In this section we list the definitions and results of empirical process theory which we have used in Section 3. An excellent introduction to most of these results can be found in Devroye *et al.* (1996).

We start with the definition of covering numbers of classes of functions.

Definition A.1 Let $\mathcal{F}$ be a class of functions $f : \mathbb{R}^d \to \mathbb{R}$. The covering number $\mathcal{N}_1(\epsilon, \mathcal{F}, z_1^n)$ is defined for any $\epsilon > 0$ and $z_1^n = (z_1, \ldots, z_n) \in \mathbb{R}^{d\cdot n}$ as the smallest integer k such that there exist functions $g_1, \ldots, g_k : \mathbb{R}^d \to \mathbb{R}$ with

$$\min_{1 \leq i \leq k} \frac{1}{n} \sum_{j=1}^{n} |f(z_j) - g_i(z_j)| \leq \epsilon$$

for each $f \in \mathcal{F}$.

If $Z_1^n = (Z_1, \ldots, Z_n)$ is a sequence of $\mathbb{R}^d$-valued random variables, then $\mathcal{N}_1(\epsilon, \mathcal{F}, Z_1^n)$ is a random variable with expected value $\boldsymbol{E}\mathcal{N}_1(\epsilon, \mathcal{F}, Z_1^n)$.

LEMMA A.1 (Haussler (1992), Th. 2)) *Let $\mathcal{F}$ be a class of functions $f : \mathbb{R}^d \to [0, B]$, and let $Z_1^n = (Z_1, \ldots, Z_n)$ be $\mathbb{R}^d$-valued i.i.d. random variables. Then for any α, $\epsilon > 0$*

$$\boldsymbol{P}\left[\sup_{f\in\mathcal{F}} \frac{\left|\frac{1}{n}\sum_{i=1}^n f(Z_i) - \boldsymbol{E}f(Z_1)\right|}{\alpha + \boldsymbol{E}f(Z_1)} > \epsilon\right] \leq 4\boldsymbol{E}\left(\mathcal{N}_1\left(\frac{\alpha\epsilon}{8}, \mathcal{F}, Z_1^n\right)\right) \exp\left(-\frac{n\alpha\epsilon^2}{16B}\right).$$

The following lemma is useful for bounding covering numbers of products of functions.

LEMMA A.2 (Devroye *et al.* (1996), Th. 29.7) *Let $\mathcal{G}_1$ and $\mathcal{G}_2$ be two families of real functions on $\mathbb{R}^d$ with $|g_1(z)| \leq B_1$ and $|g_2(z)| \leq B_2$ for all $z \in \mathbb{R}^d$, $g_1 \in \mathcal{G}_1$ and $g_2 \in \mathcal{G}_2$. Then for any $z_1^n \in \mathbb{R}^{d\cdot n}$ and $\epsilon > 0$ we have*

$$\mathcal{N}_1(\epsilon, \{g_1 \cdot g_2 : g_1 \in \mathcal{G}_1, g_2 \in \mathcal{G}_2\}, z_1^n) \leq \mathcal{N}_1\left(\frac{\epsilon}{2B_2}, \mathcal{G}_1, z_1^n\right) \cdot \mathcal{N}_1\left(\frac{\epsilon}{2B_1}, \mathcal{G}_2, z_1^n\right).$$

To bound covering numbers we use the following definition of the VC dimension.

DEFINITION A.2 Let $\mathcal{D}$ be a class of subsets of $\mathbb{R}^d$ and let $F \subseteq \mathbb{R}^d$. One says that $\mathcal{D}$ shatters F if each subset of F has the form $D \cap F$ for some D in $\mathcal{D}$. The VC dimension $V_{\mathcal{D}}$ of $\mathcal{D}$ is defined as the largest integer k for which a set of cardinality k exists which is shattered by $\mathcal{D}$.

A connection between covering numbers and VC dimensions is given by the following lemma, which uses the notation $V_{\mathcal{F}^+}$ for the VC dimension of the set

$$\mathcal{F}^+ := \{\{(x,t) \in \mathbb{R}^d \times \mathbb{R} : t \leq f(x)\} : f \in \mathcal{F}\}$$

of all subgraphs of functions of $\mathcal{F}$.

LEMMA A.3 (Haussler (1992), Th. 6) *Let $\mathcal{F}$ be a class of functions $f : \mathbb{R}^d \to [-B, B]$. Then one has for any $z_1^n \in \mathbb{R}^{d\cdot n}$ and any $\epsilon > 0$*

$$\mathcal{N}_1(\epsilon, \mathcal{F}, z_1^n) \leq 2\left(\frac{4eB}{\epsilon}\log\left(\frac{4eB}{\epsilon}\right)\right)^{V_{\mathcal{F}^+}}.$$

The following result is often useful for bounding the VC dimension.

LEMMA A.4 (Dudley (1978)) *Let $\mathcal{F}$ be a k-dimensional vector space of functions $f : \mathbb{R}^d \to \mathbb{R}$. Then the class of sets of the form $\{x \in \mathbb{R}^d : f(x) \geq 0\}$, $f \in \mathcal{F}$, has VC dimension less than or equal to k.*

REFERENCES

Devroye, L. and Krzyżak, A. (1989). An equivalence theorem for L_1 convergence of the kernel regression estimate, *J. Statist. Plann. Inference*, **23**, 71–82.

Devroye, L. P. and Wagner, T. J. (1980). Distribution–free consistency results in nonparametric discrimination and regression function estimation, *Ann. Statist.*, **8**, 231–239.

Devroye, L., Györfi, L., Krzyżak, A. and Lugosi, G. (1994). On the strong universal consistency of nearest neighbor regression function estimates, *Ann. Statist.*, **22**, 1371–1385.

Devroye, L., Györfi, L. and Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition*, Springer, New York.

Dudley, R. (1978). Central limit theorems for empirical measures, *Ann. Probab.*, **6**, 899–929.

Eubank, R. L. (1988). *Spline Smoothing and Nonparametric Regression*, Marcel Dekker, New York.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications*, Chapman & Hall, London.

Györfi, L. (1991). Universal consistency of a regression estimate for unbounded regression functions, *Nonparametric Functional Estimation and Related Topics* (ed. G. Roussas), NATO ASI Series, 329–338, Kluwer, Dordrecht.

Györfi, L. and Walk, H. (1996). On the strong universal consistency of a series type regression estimate, *Math. Methods Statist.*, **5**, 332–342.

Györfi, L. and Walk, H. (1997). On the strong universal consistency of a recursive regression estimate by Pál Révész, *Statist. Probab. Lett.*, **31**, 177–183.

Györfi, L., Kohler, M. and Walk, H. (1998). Weak and strong universal consistency of semi-recursive partitioning and kernel regression estimates, *Statist. Decisions*, **16**, 1–18.

Härdle, H. (1990). *Applied Nonparametric Regression*, Cambridge University Press, Cambridge, Massachusetts.

Haussler, D. (1992). Decision theoretic generalizations of the PAC model for neural net and other learning applications, *Inform. and Comput.*, **100**, 78–150.

Kohler, M. (1997). On the universal consistency of a least squares spline regression estimator, *Math. Methods Statistics*, **6**, 349–364.

Kohler, M. (1999). Universally consistent regression function estimation using hierarchical B-splines, *J. Multivariate Anal.*, **67**, 138–164.

Kohler, M. and Krzyżak, A. (2001). Nonparametric regression estimation using penalized least squares, in *IEEE Transactions on Information Theory*, **47**, 3054–3058.

Korostelev, A. P. and Tsybakov, A. B. (1993). *Minimax Theory of Image ReConstruction*, Springer, Berlin.

Krzyżak, A., Linder, T. and Lugosi, G. (1996). Nonparametric estimation and classification using radial basis function nets and empirical risk minimization, *IEEE Transactions on Neural Networks*, **7**, 475–487.

Lugosi, G. and Zeger, K. (1995). Nonparametric estimation via empirical risk minimization, *IEEE Trans. Inform. Theory*, **41**, 677–687.

Nobel, A. (1996). Histogram regression estimation using data-dependent partitions, *Ann. Statist.*, **24**, 1084–1105.

Spiegelman, C. and Sacks, J. (1980). Consistent window estimation in nonparametric regression, *Ann. Statist.*, **8**, 240–246.

Stone, C. J. (1977). Consistent nonparametric regression, *Ann. Statist.*, **5**, 595–645.

Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression, *Ann. Statist.*, **10**, 1040–1053.

Wahba, G. (1990). *Spline Models for Observational Data*, SIAM, Philadelphia, Pennsylvania.

Walk, H. (2002). Almost sure convergence properties of Nadaraya–Watson regression estimates, *Essays on Uncertainty-S. Yakowitz Memorial Volume* (eds. M. Dror, P. L'Ecuyer and F. Szidarovszky), 201–223, Kluwer, Dordrecht.

Ann. Inst. Statist. Math.
Vol. 54, No. 4, 900–917 (2002)
©2002 The Institute of Statistical Mathematics

IMPROVING PENALIZED LEAST SQUARES THROUGH ADAPTIVE SELECTION OF PENALTY AND SHRINKAGE

RUDOLF BERAN

Department of Statistics, University of California, Davis, Davis, CA 95616, U.S.A.

(Received February 7, 2001; revised July 2, 2001)

Abstract. Estimation of the mean function in nonparametric regression is usefully separated into estimating the means at the observed factor levels—a one-way layout problem—and interpolation between the estimated means at adjacent factor levels. Candidate penalized least squares (PLS) estimators for the mean vector of a one-way layout are expressed as shrinkage estimators relative to an orthogonal regression basis determined by the penalty matrix. The shrinkage representation of PLS suggests a larger class of candidate monotone shrinkage (MS) estimators. Adaptive PLS and MS estimators choose the shrinkage vector and penalty matrix to minimize estimated risk. The actual risks of shrinkage-adaptive estimators depend strongly upon the economy of the penalty basis in representing the unknown mean vector. Local annihilators of polynomials, among them difference operators, generate penalty bases that are economical in a range of examples. Diagnostic techniques for adaptive PLS or MS estimators include basis-economy plots and estimates of loss or risk.

Key words and phrases: Nonparametric regression, one-way layout, adaptation, loss estimator, risk estimator, economical basis, orthogonal polynomial, local annihilator.

1. Introduction

The regression model that motivates statistical procedures studied in this paper is

$$y_i = m(t_i) + e_i, \qquad 1 \le i \le n. \tag{1.1}$$

The nonrandom design points are ordered so that $t_1 \le t_2 \le \cdots \le t_n$. The errors $\{e_i\}$ are independent, identically distributed, each having a $N(0, \sigma^2)$ distribution. Both the function m and the variance σ^2 are unknown. Estimation of m from the observed $\{y_i, t_i\}$ is the task undertaken. This probabilistic formulation serves for the derivation and initial study of estimators for m. Asymptotic theory developed under the model is supplemented with computational experiments on real and artificial data that respect the fundamental distinction between data and probability model and bring out additional aspects of estimator performance. These experiments also explore the use of estimated losses and certain diagnostic plots to assess the performance of competing estimators on particular data.

Let $y = \{y_i\}$, $\mu = \{m(t_i)\}$, and $e = \{e_i\}$ be $n \times 1$ vectors with the stated components. Nonparametric regression as just described can be separated logically into two problems. The first is to estimate the values $\{m(t_i) : 1 \le i \le n\}$. This amounts to estimation of the vector μ in the possibly unbalanced one-way layout

$$y = \mu + e, \tag{1.2}$$

where e has a multivariate $N(0, \sigma^2 I_n)$ distribution. It follows from Stein (1956) that the least squares estimator of μ is inadmissible under quadratic loss whenever the number of factor levels exceeds 2. As will be seen, the least squares estimator can have high quadratic risk when compared with alternative estimators less prone to overfitting the data.

Given an efficient estimator of μ, the second problem is interpolation among its components so as to estimate the function m. This is a problem in approximation theory that is highly sensitive to assumptions on the nature of m. The observed $\{y_i, t_i\}$ will not tell us how many derivatives m has. In the absence of strong prior information about the smoothness of m, we may settle for straightforward linear interpolation or spline interpolation between the estimated components of μ. At a minimum, such interpolation is a convenient visual device for displaying estimators of m at the design points. To consider separately the estimation at design points and the interpolation between design points clarifies what can be done in nonparametric regression. Examples presented in this paper support the claim that efficient estimation of the mean function at the design points is often more important for data analysis than sophisticated interpolation between adjacent estimates.

Suppose that the design points $\{t_i\}$ contain $p \leq n$ distinct values $s_1 < s_2 < \cdots < s_p$, which are the factor levels. Let X denote the $n \times p$ incidence matrix defined as follows: row i contains a 1 in the column j such that $s_j = t_i$ and has zeroes in the other $p-1$ positions. Let $\beta = (m(s_1), m(s_2), \ldots, m(s_p))'$ denote the mean responses at the factor levels. The mean vector of the one-way layout (1.2) is then

$$\mu = X\beta \tag{1.3}$$

and the least squares estimator of μ is $\hat{\mu}_{LS} = X(X'X)^{-1}X'y$.

Let D be any matrix with p columns, let ν be an element of the extended non-negative reals $[0, \infty]$, and let $|\cdot|$ denote quadratic norm. The candidate *penalized least squares* (PLS) estimator of μ is

$$\hat{\mu}_{PLS}(D, \nu) = X\hat{\beta}_{PLS}(D, \nu) \tag{1.4}$$

where

$$\hat{\beta}_{PLS}(D, \nu) = \underset{\beta \in R^p}{\operatorname{argmin}}[|y - X\beta|^2 + \nu|D\beta|^2]. \tag{1.5}$$

It is understood that $\hat{\beta}_{PLS}(D, \infty) = \lim_{\nu \to \infty} \hat{\beta}_{PLS}(D, \nu)$. Explicitly,

$$\hat{\mu}_{PLS}(D, \nu) = X(X'X + \nu D'D)^{-1}X'y. \tag{1.6}$$

In this form, $\hat{\mu}_{PLS}(D, \nu)$ may be viewed as a generalized ridge estimator.

Effective choice of penalty matrix D and of the non-negative penalty weight ν are central issues. When ν is zero, the candidate PLS estimator reduces to the least squares estimator $\hat{\mu}_{LS}$. For very large ν, the PLS estimator effectively minimizes the residual sum of squares subject to the constraint that $|D\beta|^2$ is approximately zero. To guide the choice of D and ν, we will assess the quality of any estimator $\hat{\mu}$ through normalized quadratic loss and corresponding risk

$$L(\hat{\mu}, \mu) = p^{-1}|\hat{\mu} - \mu|^2, \qquad R(\hat{\mu}, \mu, \sigma^2) = \mathrm{E}L(\hat{\mu}, \mu). \tag{1.7}$$

Let

$$S(D,\nu) = X(X'X + \nu D'D)^{-1}X' \tag{1.8}$$

and let $|\cdot|$ denote Euclidean matrix norm. That is, $|C|^2 = \mathrm{tr}(CC') = \mathrm{tr}(C'C)$ for any matrix C. The risk of the candidate estimator $\hat{\mu}_{PLS}(D,\nu)$ is then

$$R(\hat{\mu}_{PLS}(D,\nu),\mu,\sigma^2) = p^{-1}[\sigma^2|S(D,\nu)|^2 + |\mu - S(D,\nu)\mu|^2]. \tag{1.9}$$

For the least squares estimator $\hat{\mu}_{LS} = \hat{\mu}_{PLS}(D,0)$, this risk reduces to σ^2.

Let $\hat{\sigma}^2$ be a trustworthy estimator of σ^2. Customary when n substantially exceeds p is the variance estimator $\hat{\sigma}^2_{LS} = (n-p)^{-1}|y - \hat{\mu}_{LS}|^2$. The derivation of the Mallows (1973) C_L criterion yields the risk estimator

$$\hat{R}(D,\nu) = p^{-1}[|y - S(D,\nu)y|^2 + \{2\,\mathrm{tr}[S(D,\nu)] - n\}\hat{\sigma}^2]. \tag{1.10}$$

In particular, when $\hat{\sigma}^2 = \hat{\sigma}^2_{LS}$, the estimated risk for the least squares estimator of μ is $\hat{R}(D,0) = \hat{\sigma}^2_{LS}$. We propose to choose both the penalty weight ν and the penalty matrix D so as to minimize the estimated risk $\hat{R}(D,\nu)$.

When represented with respect to the orthogonal penalty basis for the regression space that is defined in the next section, PLS estimators suggest a larger class of candidate monotone shrinkage (MS) estimators for μ. The themes of this paper are: asymptotic theory to support the strategy of choosing the candidate estimator that minimizes estimated risk; the advantages of adaptive MS over adaptive PLS; methods for designing effective penalty matrices; and the use of estimated loss/risk and of diagnostic plots to assess the performance of adaptive PLS or MS estimators on given data.

The need for asymptotic analysis and for restrictions on the extent of adaptation is indicated by an example. Suppose that S is permitted to vary over all $n \times n$ symmetric matrices that have a specified set of eigenvectors and that σ^2 is known. The symmetric matrix S that minimizes the right side of (1.10) over the class just described then generates an estimator of μ whose risk is dominated by that of the least squares estimator $\hat{\mu}_{LS}$. This may be seen from Remark A on p. 1829 of Beran and Dümbgen (1998).

For fixed penalty matrix D, the shrinkage-adaptive PLS estimator is defined to be $\hat{\mu}_{PLS}(D,\hat{\nu})$, where $\hat{\nu}$ minimizes the estimated risk $\hat{R}(D,\nu)$ over all ν in $[0,\infty]$. We will call this the PLS(D) estimator. Section 2.3 describes how to compute it effectively. Under the probability model described there, the risk of the adaptive estimator PLS(D) converges to the risk of the unrealizable candidate PLS estimator with smallest risk. Thus, the asymptotic risk of the PLS(D) estimator cannot exceed that of the least squares estimator. In practice, it is often far smaller and the shrinkage-adaptive MS(D) estimator to be defined in Subsection 2.2 typically reduces risk further. Subsection 3.2 develops possibilities for adaptation through choice of the penalty matrix D in addition to ν.

Though valuable in exploring the scope of adaptation and the overall behavior of an estimator, ensemble results such as asymptotic minimaxity or rates of convergence do not indicate the adequacy of a particular estimator on particular data. Section 3 addresses the use of estimated loss and of diagnostic basis-economy and shrinkage-vector plots to assess adaptive PLS and MS estimators on given data.

Figure 1 exhibits penalized least squares estimates on three sets of artificial data. The smooth case was suggested by the Canadian earnings data that was analyzed, with further background, in Chu and Marron (1991). The respective mean functions are:

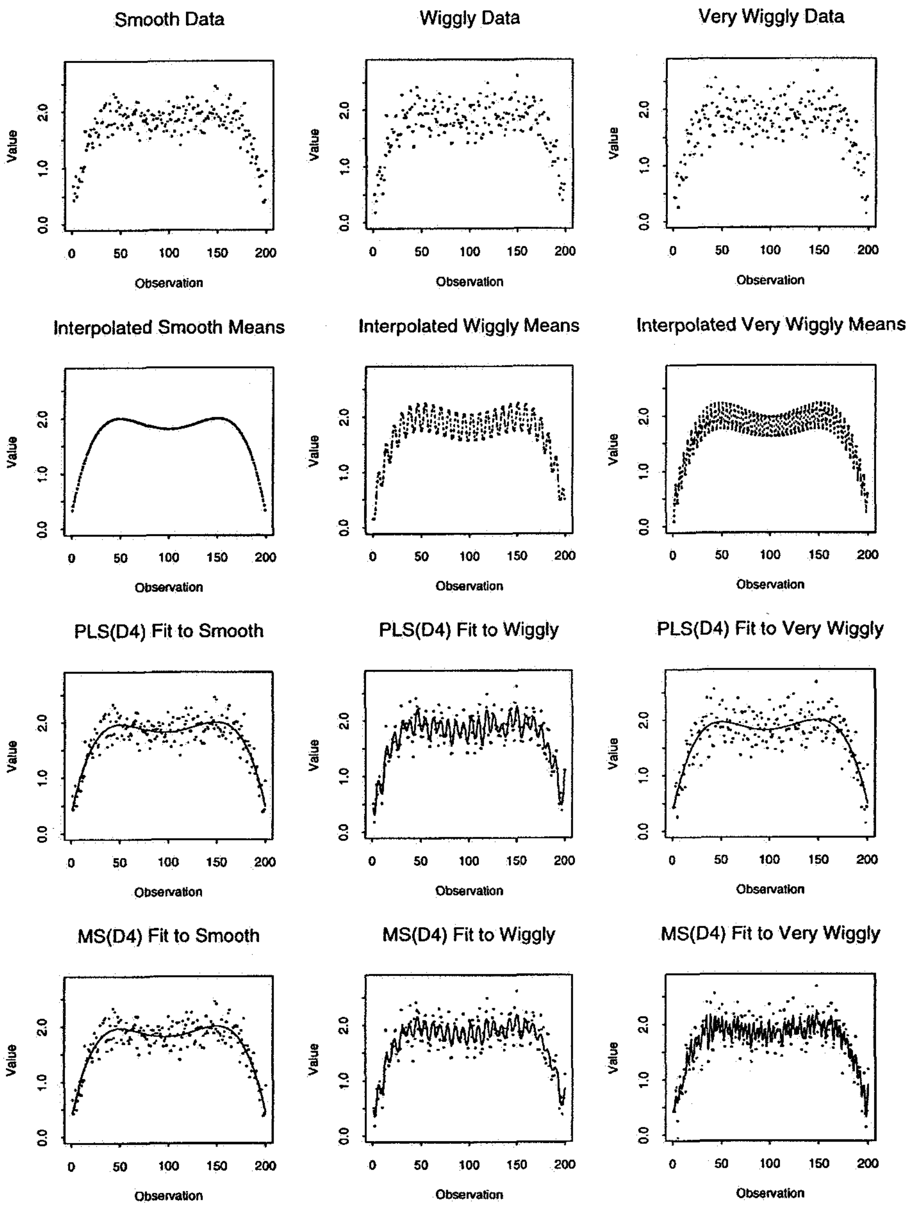

Fig. 1. Each column displays the artificial data, the true mean vector, the PLS(D_4) estimate, and the MS(D_4) estimate.

Smooth: $m_1(t) = 2 - 50((t-25)(t-75))^2$.
Wiggly: $m_2(t) = m_1(t) - .25\sin(50\pi t)$.
Very Wiggly: $m_3(t) = m_1(t) - .25\sin(100\pi t)$.

The design points are $\{t_i = i/(n+1)\colon 1 \le i \le n\}$ with $n = 200$. The j-th artificial data set is $\{m_j(i/201)\colon 1 \le i \le 200\} + e$, where e is a single pseudo-random sample drawn from the $N(0, \sigma^2 I_{200})$ distribution and $\sigma = .2$. In this design, $p = n$. The variance σ^2 is estimated by the high component estimator defined in (2.13), with $q = .75p$.

As penalty matrix we use the $(p-4)\times p$ fourth difference matrix D_4. The first row of D_4 consists of entries 1, -4, 6, -4, 1 followed by zeros. The second row shifts the non-zero entries one place to the right and puts a zero in the first column. Construction of subsequent rows continues the shift of nonzero entries to the right. The d-th difference penalty matrix D_d, defined formally in Section 3.2, is particularly appropriate when the components of β are equally spaced values on a curve whose local behavior mimics a polynomial of degree $d-1$. In the present example, either $d=4$ or $d=5$ works well. Computations for this and other examples in the paper were done with S-Plus 2000 for Windows.

Column j in Fig. 1 plots the j-th artificial sample in the first row and the linearly interpolated (dashed line) components of the mean vector $\mu=\{m_j(i/201)\}$ in the second row. The function `rnorm`, initialized with `set.seed(2)`, produced the pseudo-Gaussian errors that are added to the means in the second row to obtain the artificial samples. Any sinusoidal wiggles present in μ are not apparent to the eye in the scatterplots of this data.

The third row in the figure superposes on the data the linearly interpolated (solid line) components of estimator PLS(D_4). This shrinkage-adaptive PLS estimator recovers the means well from the Smooth sample and detects the sinusoid underlying the Wiggly sample, even though it distorts that sinusoid's amplitude and regularity. However, on the Very Wiggly sample, PLS(D_4) fails utterly, like the eye, to detect the sinusoid and settles for estimating the smooth component of the trend. The fourth row of Fig. 1 plots the adaptive MS(D_4) generalization of PLS(D_4) that is defined in Section 2.2. This estimator succeeds in handling the Very Wiggly sample as well as the other two.

2. Estimated risk and shrinkage adaptation

A canonical representation assists both theoretical study and numerical computation of the candidate PLS estimators $\hat{\mu}_{PLS}(D,\nu)$. These and the candidate MS estimators defined in Section 2.2 are particular symmetric linear smoothers in the sense of Buja *et al.* (1989) and are candidate REACT estimators in the sense of Beran (2000).

2.1 *The penalty basis*

The replication matrix $R=X'X$ is a $p\times p$ diagonal matrix whose k-th diagonal element indicates the number of $\{t_i\}$ that equal s_k. For any matrix C, let $\mathcal{M}(C)$ denote the subspace spanned by its columns. The columns of the matrix $U_0=XR^{-1/2}$ provide an orthonormal basis for the regression problem: $U_0'U_0=I_p$ and $\mathcal{M}(U_0)=\mathcal{M}(X)$. Let $B=R^{-1/2}D'DR^{-1/2}$. Because $X=U_0R^{1/2}$, equation (1.6) is equivalent to

$$\hat{\mu}_{PLS}(D,\nu)=U_0(I_p+\nu B)^{-1}U_0'y. \tag{2.1}$$

The symmetric matrix B has spectral representation $B=\Gamma\Lambda\Gamma'$ where the eigenvector matrix satisfies $\Gamma'\Gamma=\Gamma\Gamma'=I_p$ and the diagonal matrix $\Lambda=\operatorname{diag}\{\lambda_i\}$ gives the ordered eigenvalues with $0\le\lambda_1\le\lambda_2\le\cdots\le\lambda_p$. This eigenvalue ordering, the reverse of the customary, is used here because the eigenvectors associated with the smallest eigenvalues largely determine the value and performance of candidate estimator $\hat{\mu}_{PLS}(D,\nu)$. Let $U=U_0\Gamma$. It follows from (2.1) that

$$\hat{\mu}_{PLS}(D,\nu)=U(I_p+\nu\Lambda)^{-1}U'y. \tag{2.2}$$

The columns of the matrix U constitute the orthonormal *penalty basis* for the regression space determined by the penalty matrix $D : U'U = I_p$ and $\mathcal{M}(U) = \mathcal{M}(X)$.

Variational characterization of U. Alternatively, the successive columns $u_1, u_2, \ldots,$ u_p of the penalty basis matrix U may be defined through their variational properties:

- As above, let $U_0 = XR^{-1/2}$ provide an initial orthonormal basis matrix for the regression space $\mathcal{M}(X)$.
- Find a unit vector $u_1 = U_0\gamma$ in $\mathcal{M}(X)$ that minimizes the penalty $|D(X'X)^{-1} X'u_1|^2$. This reduces to finding the $p \times 1$ unit vector γ that minimizes $|DR^{-1/2}\gamma|^2 = \gamma' B\gamma$. The desired minimum penalty vector is thus $u_1 = U_0\gamma_1$, where γ_j is the j-th column of the eigenvector matrix Γ.
- Find a unit vector $u_2 = U_0\gamma$ in $\mathcal{M}(X)$ that minimizes the penalty $|D(X'X)^{-1} X'u_2|^2$ subject to the constraint that u_2 is orthogonal to u_1. This reduces to finding the $p \times 1$ unit vector γ orthogonal to γ_1 that minimizes $|DR^{-1/2}\gamma|^2 = \gamma' B\gamma$. The desired minimum penalty vector is thus $u_2 = U_0\gamma_2$.
- Continue sequential constrained minimization to obtain the penalty basis matrix

$$U = (U_0\gamma_1, U_0\gamma_2, \ldots, U_0\gamma_p) = U_0\Gamma. \tag{2.3}$$

In the one-way layout under consideration, $(X'X)^{-1}X'u_k$ extracts the components of basis vector u_k that are associated with the p factor levels. The penalty for this extracted vector is

$$|D(X'X)^{-1}X'u_k|^2 = |DR^{-1/2}\gamma_k|^2 = \gamma_k' B\gamma_k = \lambda_k. \tag{2.4}$$

When the penalty matrix is a d-th difference operator, the preceding variational characterization of U explains intuitively why its successive column vectors are increasingly wiggly.

2.2 *From PLS to monotone shrinkage estimators*

Fix D so that the penalty basis U is determined. Let $z = U'y$ and let $f(\nu)$ denote the column vector $(1/(1+\nu\lambda_1), 1/(1+\nu\lambda_2), \ldots, 1/(1+\nu\lambda_p))'$, with the understanding that $f(\infty) = \lim_{\nu\to\infty} f(\nu)$. The distribution of z is $N_p(\xi, \sigma^2 I_p)$, where $\xi = U'\mu$. The PLS estimator of ξ implied by expression (2.2) is

$$\hat{\xi}_{PLS}(D,\nu) = U'\hat{\mu}_{PLS}(D,\nu) = f(\nu)z, \tag{2.5}$$

where the multiplication of vectors in the expression to the right is performed componentwise as in the S language. Equivalently,

$$\hat{\mu}_{PLS}(D,\nu) = U\hat{\xi}_{PLS}(D,\nu) = U\,\mathrm{diag}\{f(\nu)\}U'y. \tag{2.6}$$

The structure of representation (2.6) suggests a larger family of candidate estimators for μ. Let

$$\mathcal{F}_{MS} = \{f \in [0,1]^p : f_1 \geq f_2 \geq \cdots f_p\} \tag{2.7}$$

and let

$$\hat{\xi}_{MS}(D,f) = fz \quad \text{for} \quad f \in \mathcal{F}_{MS}. \tag{2.8}$$

The candidate *monotone shrinkage* (MS) estimators for μ associated with penalty matrix D are defined by

$$\hat{\mu}_{MS}(D,f) = U\hat{\xi}_{MS}(D,f) = U\,\mathrm{diag}\{f\}U'y \quad \text{for} \quad f \in \mathcal{F}_{MS}. \tag{2.9}$$

It follows from (2.6) that the candidate PLS estimators are a proper subset of the MS family in which the shrinkage vector f is restricted to the form $\{f(\nu): \nu \in [0,\infty]\}$.

The next section develops three good reasons for considering monotone shrinkage estimators. First, for every candidate PLS estimator there is an MS estimator whose risk is at least as small. Second, minimizing the *estimated* risk of candidate MS or PLS estimators over all shrinkage vectors permitted by their definitions turns out to minimize asymptotic risk over the respective classes of candidate estimators. Third, computation of adaptive MS estimators is faster than computation of their adaptive PLS counterparts.

2.3 *Estimated risks and shrinkage adaptation*

For any vector h, let ave(h) denote the average of its components. Define the function

$$\rho(f,\xi^2,\sigma^2) = \mathrm{ave}[f^2\sigma^2 + (1-f)^2\xi^2] \quad \text{for} \quad f \in [0,1]^p. \tag{2.10}$$

Because $|\hat{\mu}_{MS}(D,f) - \mu|^2 = |fz - \xi|^2$, it follows that the normalized quadratic risk of the candidate MS estimator is

$$R(\hat{\mu}_{MS}(D,f),\mu,\sigma^2) = \rho(f,\xi^2,\sigma^2) \quad \text{for} \quad f \in \mathcal{F}_{MS}. \tag{2.11}$$

In particular, the risk $R(\hat{\mu}_{PLS}(D,\nu),\mu,\sigma^2)$ of the candidate PLS estimator, expressed in the original coordinate system by equation (1.9), is simply $\rho(f(\nu),\xi^2,\sigma^2)$.

The risk function $\rho(f,\xi^2,\sigma^2)$ depends on the unknown parameters ξ^2 and σ^2. Having obtained a variance estimator $\hat{\sigma}^2$, we may estimate ξ^2 by $z^2 - \hat{\sigma}^2$ and hence $\rho(f,\xi^2,\sigma^2)$ by

$$\hat{\rho}(D,f) = \mathrm{ave}[f^2\hat{\sigma}^2 + (1-f)^2(z^2 - \hat{\sigma}^2)]. \tag{2.12}$$

Expression (2.12) expresses in canonical form the Mallows risk estimator (1.10).

The following definitions carry out several strategies for estimating the variance σ^2:

- *The least squares variance estimator.* The least squares variance estimator $\hat{\sigma}^2_{LS} = (n-p)^{-1}|y - \hat{\mu}_{LS}|^2$ is unbiased and is consistent for σ^2 when $n - p$ tends to infinity.
- *The first-difference estimator.* This estimator, $\hat{\sigma}^2_{D1} = [2(n-1)]^{-1}\sum_{i=2}^n (y_i - y_{i-1})^2$, was treated by Rice (1984). It is consistent for σ^2 when n tends to infinity and the bias $\lim_{n\to\infty}[2(n-1)]^{-1}\sum_{i=2}^n(\mu_i - \mu_{i-1})^2 = 0$. Similar estimators may be constructed from higher-order differences of y.

The next two variance estimators make use of the penalty basis U. Choose $\bar{U}$ so that the concatenated matrix $(U \mid \bar{U})$ is orthogonal. Set $\bar{z} = \bar{U}'y$ in analogy to the earlier $z = U'y$.

- *The high-component variance estimator.* The strategy of pooling sums of squares in analysis of variance suggests

$$\hat{\sigma}^2_H = (n-q)^{-1}\left[\sum_{i=q+1}^{p} z_i^2 + |\bar{z}|^2\right] = (n-q)^{-1}\left[\sum_{i=q+1}^{p} z_i^2 + |y - \hat{\mu}_{LS}|^2\right], \tag{2.13}$$

where $q \leq \min\{p, n-1\}$. The bias of $\hat{\sigma}_H^2$ is $(n-q)^{-1}\sum_{i=q+1}^{p}\xi_i^2$. Consistency of $\hat{\sigma}_H^2$ is assured when this bias tends to zero as $n-q$ tends to infinity. When $q = p < n$, the estimator $\hat{\sigma}_H^2$ reduces to $\hat{\sigma}_{LS}^2$.

- *The robust high-component variance estimator.* Let w denote the vector obtained by concatenating $\{z_i : q+1 \leq i \leq p\}$ with $\bar{z}$. Robustness theory suggests the estimator

$$\hat{\sigma}_{RH} = \text{median}\{|w_j| : 1 \leq j \leq n-q\}/\Phi^{-1}(.75) \tag{2.14}$$

for σ, where Φ^{-1} is the standard normal quantile function. Under model (1.2), $\hat{\sigma}_{RH}^2$ approaches σ^2 in probability when $n-q$ is large and the high order components of ξ are small.

Let $\hat{g} = (z^2 - \hat{\sigma}^2)/z^2$. The risk estimator $\hat{\rho}(D, f)$ in (2.12) can be rewritten in the form

$$\hat{\rho}(D, f) = \text{ave}[(f-\hat{g})^2 z^2] + \hat{\sigma}^2\,\text{ave}(\hat{g}). \tag{2.15}$$

For fixed penalty matrix D, the *shrinkage-adaptive* PLS(D) estimator is defined to be $\hat{\mu}_{MS}(D, \hat{\nu})$, where

$$\hat{\nu} = \underset{\nu\in[0,\infty]}{\text{argmin}}\,\hat{\rho}(D, f(\nu)) = \underset{\nu\in[0,\infty]}{\text{argmin}}\,\text{ave}[(f(\nu)-\hat{g})^2 z]. \tag{2.16}$$

Computation of $\hat{\nu}$ is thus a weighted least squares problem that can be solved with the S-Plus function nls in the manner exhibited on p. 244 of Venables and Ripley (1999). The PLS fits plotted in the third row of Fig. 1 were obtained in this fashion.

Similarly, for fixed penalty matrix D, the *shrinkage-adaptive* MS(D) estimator is defined to be $\hat{\mu}_{MS}(D, \hat{f}_{MS})$, where

$$\hat{f}_{MS} = \underset{f\in\mathcal{F}_{MS}}{\text{argmin}}\,\hat{\rho}(D, f) = \underset{f\in\mathcal{F}_{MS}}{\text{argmin}}\,\text{ave}[(f-\hat{g})^2 z]. \tag{2.17}$$

To facilitate this minimization, let $\mathcal{H} = \{h \in R^p : h_1 \geq h_2 \geq \cdots \geq h_p\}$ and let

$$\hat{h} = \underset{h\in\mathcal{H}}{\text{argmin}}\,\text{ave}[(h-\hat{g})^2 z]. \tag{2.18}$$

Then $\hat{f}_{MS} = \hat{h}_+$. That is, each component of $\hat{f}_{MS}$ is the positive part of the corresponding component of $\hat{h}$. For a proof, see Beran and Dümbgen (1998). Computation of $\hat{h}$ is a weighted isotonic least squares problem that can be solved with the pool-adjacent-violators algorithm (cf. Robertson *et al.* (1988)). The MS fits plotted in the last row of Fig. 1 were obtained in this fashion. Computation is faster for MS(D) than for PLS(D). S-Plus code for the examples in this paper is available from the author.

The following theorem shows that adaptation works in the sense that minimizing estimated risk over either the MS or PLS shrinkage class for fixed D succeeds in minimizing risk asymptotically over that class. The result makes no smoothness assumptions on the unknown mean vector μ and follows from Theorems 2.1 and 2.2 in Beran and Dümbgen (1998).

THEOREM 2.1. *Let $\mathcal{F}$ be any subset of $\mathcal{F}_{MS}$ that is closed in $[0,1]^p$. In particular, $\mathcal{F}$ could be either the PLS shrinkage class $\{f(\nu) : \nu \in [0,\infty]\}$ or the monotone shrinkage class $\mathcal{F}_{MS}$. Suppose that $\hat{\sigma}^2$ is consistent in that, for every $r > 0$ and $\sigma^2 > 0$,*

$$\lim_{p\to\infty}\sup_{\text{ave}(\xi^2)\leq\sigma^2 r} E|\hat{\sigma}^2 - \sigma^2| = 0. \tag{2.19}$$

Let $V(f)$ denote either the loss $L(\hat{\mu}(D,f),\mu)$ or the estimated risk $\hat{\rho}(D,f)$. Then, for every penalty matrix D, every $r>0$, and every $\sigma^2>0$,

$$\lim_{p\to\infty}\sup_{\text{ave}(\xi^2)\le\sigma^2 r} E\sup_{f\in\mathcal{F}}|V(f)-\rho(f,\xi^2,\sigma^2)|=0. \tag{2.20}$$

Moreover, if $\hat{f}=\text{argmin}_{f\in\mathcal{F}}\,\hat{\rho}(D,f)$, then

$$\lim_{p\to\infty}\sup_{\text{ave}(\mu^2)/\sigma^2\le r}|R(\hat{\mu}(D,\hat{f}),\mu,\sigma^2)-\min_{f\in\mathcal{F}}R(\hat{\mu}(D,f),\mu,\sigma^2)|=0. \tag{2.21}$$

By (2.20), the loss, risk and estimated risk of a candidate estimator converge together asymptotically. Uniformity of this convergence over $\mathcal{F}$ makes the estimated risk of candidate estimators a reasonable surrogate for true risk or loss. By (2.21), the risk of the shrinkage-adaptive estimator $\hat{\mu}(D,\hat{f})$ converges to that of the best candidate estimator. These conclusions break down when the class of shrinkage vectors $\mathcal{F}$ is too large in a covering number sense. In particular, it does not hold if $\mathcal{F}=[0,1]^p$, as shown in Beran and Dümbgen (1998).

Remarks. Condition (2.19) holds for the variance estimator $\hat{\sigma}^2_{LS}$ if $n-p$ tends to infinity with p. Asymptotic results for other variance estimators are given in Beran (1996) and Beran and Dümbgen (1998). The quantity $\text{ave}(\mu^2)/\sigma^2=\text{ave}(\xi^2)/\sigma^2$ in (2.21) measures the signal to noise ratio. Limits (2.20) and (2.21) both hold without any restrictions on the smoothness of μ. Because the monotone shrinkage class $\mathcal{F}_{MS}$ is strictly larger than the generating PLS shrinkage class $\{f(\nu):\nu\in[0,\infty]\}$, the asymptotic risk of MS(D) cannot exceed that of PLS(D).

COROLLARY 2.1. *Under the conditions for Theorem* 2.1,

$$\lim_{p\to\infty}\sup_{\text{ave}(\xi^2)\le\sigma^2 r}\mathrm{E}\mid\hat{\rho}(D,\hat{f})-W)\mid=0 \tag{2.22}$$

for W equal to either $L(\hat{\mu}(D,\hat{f}),\mu)$ or $R(\hat{\mu}(D,\hat{f}),\mu,\sigma^2)$.

PROOF. Equation (2.20) implies that

$$\lim_{p\to\infty}\sup_{\text{ave}(\xi^2)\le\sigma^2 r} E\sup_{f\in\mathcal{F}}|\hat{\rho}(D,f)-L(\hat{\mu}(D,f),\mu)|=0, \tag{2.23}$$

which yields (2.22) for the first choice of W. Because $\hat{f}$ minimizes $\hat{\rho}(D,f)$ over $f\in\mathcal{F}$, equation (2.20) also implies that

$$\lim_{p\to\infty}\sup_{\text{ave}(\xi^2)\le\sigma^2 r}\mathrm{E}|\hat{\rho}(D,\hat{f})-\min_{f\in\mathcal{F}}\rho(f,\xi^2,\sigma^2)|=0. \tag{2.24}$$

Combining this with (2.21) yields (2.22) for the second choice of W.

That the *plug-in* loss/risk estimator $\hat{\rho}(D,\hat{f})$ converges asymptotically to the actual loss/risk of PLS(D) or MS(D) is useful when comparing adaptive estimators on specific data. For the examples of Fig. 1, the plug-in loss/risk estimates and actual losses for

	MS loss	MS plug-in	PLS loss	PLS plug-in	LS loss	LS plug-in
Smooth	.0011	$-.0068$	.0013	$-.0066$	.0372	.0455
Wiggly	.0111	.0015	.0138	.0072	.0372	.0456
Very Wiggly	.0127	.0092	.0326	.0290	.0372	.0454

PLS(D_4), MS(D_4), and the least squares estimator are shown in Table 1. In scrutinizing this table, we observe that:

- The plug-in estimated losses for the shrinkage-adaptive MS and PLS estimates are noticeably smaller than the true losses.
- The plug-in losses indicate correctly the ordering of the true losses for the MS, PLS and LS estimates.
- The loss of the LS estimator in each of the three examples is .0372, a value reasonably close to the LS risk $\sigma^2 = .04$. The high-component variance estimator used in this experiment overestimates the true variance modestly.

3. Penalty matrix adaptation

Section 3.1 analyzes the manner in which the penalty matrix D affects the asymptotic risks of adaptive estimators MS(D) and PLS(D). The economy of the penalty basis in representing the unknown mean vector μ is a key factor. Section 3.2 develops candidate penalty matrices for equally and unequally spaced factor levels and considers adaptation over both penalty matrix and shrinkage vector. Section 3.3 discusses diagnostic plots that display the empirical economy of candidate penalty bases and considers an alternative to plug-in estimates for the loss or risk of adaptive estimators.

3.1 *Role of an economical penalty basis*

As will be seen, the risk of the shrinkage-adaptive PLS or MS estimator for μ is relatively small if all but the first few components of $\xi = U'\mu$ are very nearly zero. In this event, we say that the columns of the matrix U provide an *economical* basis for the regression space $\mathcal{M}(X)$. The benefit of using an economical regression basis is clear heuristically. In that case, we need only identify and estimate from the data the relatively few non-zero components of ξ, using the naive estimate zero for the remaining components. The quadratic risk then accumulates small squared biases from ignoring the nearly zero components of ξ but does not accumulate the many variances that would arise from an attempt to estimate these unbiasedly.

An idealized formulation of basis economy enables mathematical analysis of how economy affects risk. For every $b \in [0,1]$, every $r > 0$, and every $\sigma^2 > 0$, consider the projected ball

$$B(r, b, \sigma^2) = \{\xi\colon \operatorname{ave}(\xi^2)/\sigma^2 \le r \text{ and } \xi_i = 0 \text{ for } i > bp\}. \tag{3.1}$$

Suppose that the regression basis U associated with penalty matrix D is economical in the formal sense that the transformed mean vector ξ lies in $B(r, b, \sigma^2)$ for some small value of b and some finite positive value of r. Though this description is too simple to serve as a complete definition of basis economy, it yields the following quantitative results about the effect of basis economy on the risk of estimators of μ.

THEOREM 3.1. *Fix the penalty basis U by choice of D. For every $b \in [0,1]$, every $r > 0$, and every $\sigma^2 > 0$, the asymptotic minimax quadratic risk over all estimators of μ is*

$$\lim_{p\to\infty} \inf_{\hat{\mu}} \sup_{\xi\in B(r,b,\sigma^2)} R(\hat{\mu},\mu,\sigma^2) = \sigma^2 rb/(r+b). \tag{3.2}$$

The shrinkage-adaptive estimator $\hat{\mu}_{MS}(D,\hat{f}_{MS})$ achieves asymptotic minimax bound (3.2) in that

$$\lim_{p\to\infty} \sup_{\xi\in B(r,b,\sigma^2)} R(\hat{\mu}_{MS}(D,\hat{f}_{MS}),\mu,\sigma^2) = \sigma^2 rb/(r+b) \tag{3.3}$$

for every possible b, r, and σ^2.

Limit (3.3) follows from Theorem 4 in Beran (2000). As discussed in that paper, equation (3.2) is a special case of Pinsker's (1980) asymptotic minimax bound. Note that (3.3) establishes more than formal asymptotic minimaxity of shrinkage-adaptive estimator MS(D). When b is small, the right side of (3.3) is much smaller than the risk σ^2 of the least squares estimator $\hat{\mu}_{LS}$. To the extent that estimator PLS(D) approximates estimator MS(D), its performance also benefits strongly from economy of the penalty basis. This phenomenon underlies the very similar appearance of PLS(D_4) and MS(D_4) in the first column of Fig. 1.

3.2 *Candidate penalty matrices and adaptation*

The ideal choice of penalty basis U would have its first column proportional to the unknown mean vector μ so that only the first component of ξ would be nonzero. Though unrealizable, this ideal choice indicates that prior information or conjecture about μ can be exploited in devising the penalty matrix D that generates the penalty basis. The discussion in this section relates prior notions about the local behavior of the mean function m to the construction of reasonable candidate penalty matrices.

Difference operators. Consider the important case when the factor level vector $s = (s_1, s_2, \ldots, s_p)'$ has equally spaced components. To define the d-th difference matrix D_d, consider the $(p-1)\times p$ matrix $\Delta(p) = \{\delta_{i,j}\}$ in which $\delta_{i,i} = 1$, $\delta_{i,i+1} = -1$ for every i and all other entries are zero. Then,

$$D_1 = \Delta(p) \quad \text{and} \quad D_d = \Delta(p-d+1)D_{d-1} \quad \text{for } 2 \le d < p. \tag{3.4}$$

It may be verified that the $(p-d)\times p$ matrix D_d annihilates powers of s up to power $d-1$ in the sense that

$$D_d s^k = 0 \quad \text{for} \quad 0 \le k \le d-1. \tag{3.5}$$

Moreover, in row i of D_d, the elements not in columns i, $i+1, \ldots i+d$ are zero.

The penalty term in (1.5) is proportional to $|D\beta|^2$ where $\beta = m(s)$. When m behaves locally like a polynomial of degree $d-1$, property (3.5) and the subsequent remark about zeros entail that $|D_d\beta|$ is small. We may therefore expect that both PLS(D_d) and MS(D_d) will favor fits with local polynomial behavior of degree $d-1$. This implicit preference is appropriate whenever m has such local polynomial behavior. The success of fits based on penalty matrix D_4 in the first column of Fig. 1 illustrates the point. We note that normalizing the row vectors of D_d to have unit length does not change the corresponding penalty basis U. However, (3.5) breaks down for $k \geq 1$ when the components of s are not equally spaced.

Local annihilators. To devise useful candidate penalty matrices for arbitrary factor levels $s \in R^p$ and for other notions about m, we draw on the mathematical interpretation of (3.5) as an orthogonality property. Let g_0, $g_1, \ldots, g_{d-1}$ be a given set of real-valued functions defined on the real line. We hypothesize that the mean function m behaves locally like a linear combination of the $\{g_k : 0 \leq k \leq d-1\}$. Local polynomial behavior is the special case where $g_k(s_i) = s_i^k$ for every k.

For each i such that $1 \leq i \leq p-d$, assume that the d vectors $\{(g_k(s_i), \ldots, g_k(s_{i+d})$: $0 \leq k \leq d-1\}$ are linearly independent in R^{d+1}. This is a condition on the functions $\{g_k\}$ that is satisfied, for instance, when $g_k(s_i) = s_i^k$. Let $\mathcal{G}_i$ denote the d-dimensional subspace of R^{d+1} that is spanned by these vectors. Define the $(p-d) \times p$ *local annihilator* matrix $A_d = \{a_{i,j}\}$ as follows: In the i-th row of A_d, the subvector $\{a_{i,j} : i \leq j \leq i+d\}$ is the unit vector in R^{d+1}, unique up to sign, that is orthogonal to $\mathcal{G}_i$. The remaining elements of A_d are zero.

THEOREM 3.2. *Let* $g_k(s) = (g_k(s_1), g_k(s_2), \ldots, g_k(s_p))'$. *Each row vector of the local annihilator matrix* A_d *has unit length and*

$$A_d g_k(s) = 0 \quad \textit{for} \quad 0 \leq k \leq d-1. \tag{3.6}$$

PROOF. The definition of A_d ensures that its rows have unit length and

$$\sum_{j=1}^{p} a_{i,j} g_k(s_j) = \sum_{j=i}^{i+d} a_{i,j} g_k(s_j) = 0 \quad \text{for} \quad 0 \leq k \leq d-1. \tag{3.7}$$

Of particular utility as the generalization of D_d for unequally spaced factor levels is the *local polynomial annihilator*. This is obtained by setting $g_k(s_i) = s_i^k$ in the definition of A_d. Thus, in the i-th row of the local polynomial annihilator, the subvector $\{a_{i,j} : i \leq j \leq i+d\}$ is the basis vector of degree d in the orthonormal polynomial basis on the factor levels $(s_i, \ldots, s_{i+d})$. All other elements in the row are zero. The S-Plus function `poly` enables computation of the local polynomial annihilator in a numerically stable way for d up to 50 or so. When the components of s are equally spaced, the local polynomial annihilator A_d becomes a scalar multiple of the d-th difference matrix D_d. Of course, local polynomial A_1 is proportional to D_1 for every factor level vector s.

Remark. A referee kindly pointed out that the foregoing discussion of annihilators can be linked to the algorithm for L-splines described at the end of Heckman and Ramsay (2000). Let $m^{(j)}$ denote the j-th derivative of m and let L be a differential operator such that $Lm = \sum_{j=0}^{d-1} a_j m^{(j)}$. The set of all m such that $Lm = 0$ is a linear space of dimension d. Let g_0, $g_1, \ldots, g_{d-1}$ denote a basis for this space. The construction of the sparse matrix A_d in Theorem 3.2 follows from the Heckman and Ramsay algorithm by setting Q', D and U in their notation to A_d, identity matrix and $(g_0(s), g_1(s), \ldots, g_{d-1}(s))$ in the present setting.

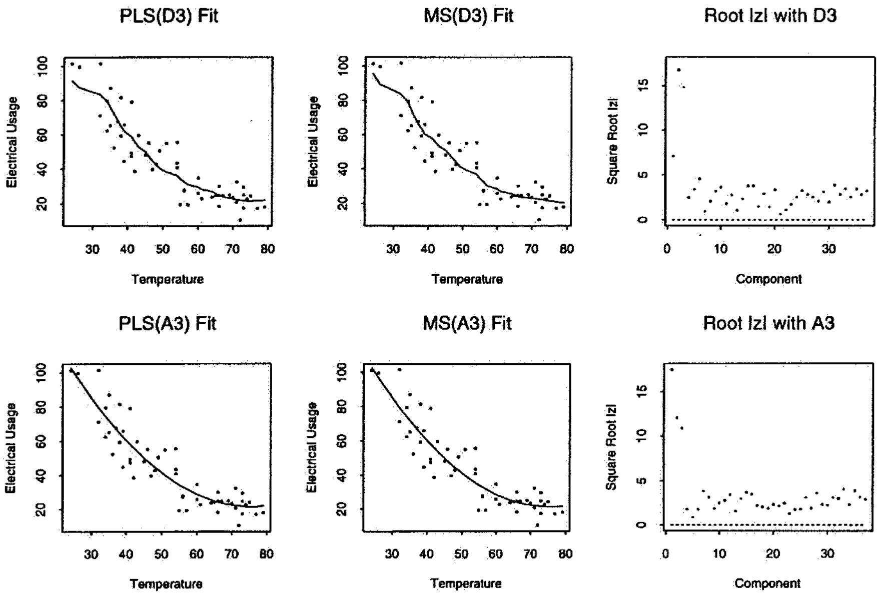

Fig. 2. Using penalty matrices D_3 and local polynomial A_3 respectively, each row displays adaptive PLS and MS estimates for conditional mean electrical usage and the associated basis economy plot.

	D_3 penalty	A_3 penalty
PLS plug-in loss	−33.55	−42.95
MS plug-in loss	−35.57	−42.99

Figure 2 exhibits competing PLS and MS estimates for mean electrical usage as a function of temperature. The data is described in Simonoff (1996). We estimate mean electrical usage conditional on the observed temperatures, whose distinct values are not equally spaced. The variance is estimated by $\hat{\sigma}^2_{LS}$. Because the trend in the data appears to be roughly quadratic, we expect that MS and PLS fits generated with the local polynomial annihilator A_3 as penalty matrix will have relatively low estimated risks. This turns out to be the case. The first row of Fig. 2 gives the PLS and MS fits when the penalty matrix is D_3 while the second row gives the corresponding fits when the penalty matrix is local polynomial A_3. The plug-in loss/risk estimates for these competing fits are shown in Table 2. In sharp contrast, the loss/risk estimate for the least squares estimator of μ is 129.70.

The negativity of the risk estimates in this table is an artifact of the small regression space dimension, $p = 37$. The ordering of the estimated risks matches the visual quality of the competing fits in Fig. 2. In this example, MS does not improve significantly upon PLS. However, choosing the penalty matrix to handle unequal spacing of the design points is clearly beneficial. The basis-economy plots in the third column of Fig. 2 exhibit the superior empirical economy of the local polynomial A_3 penalty basis relative to the D_3 basis.

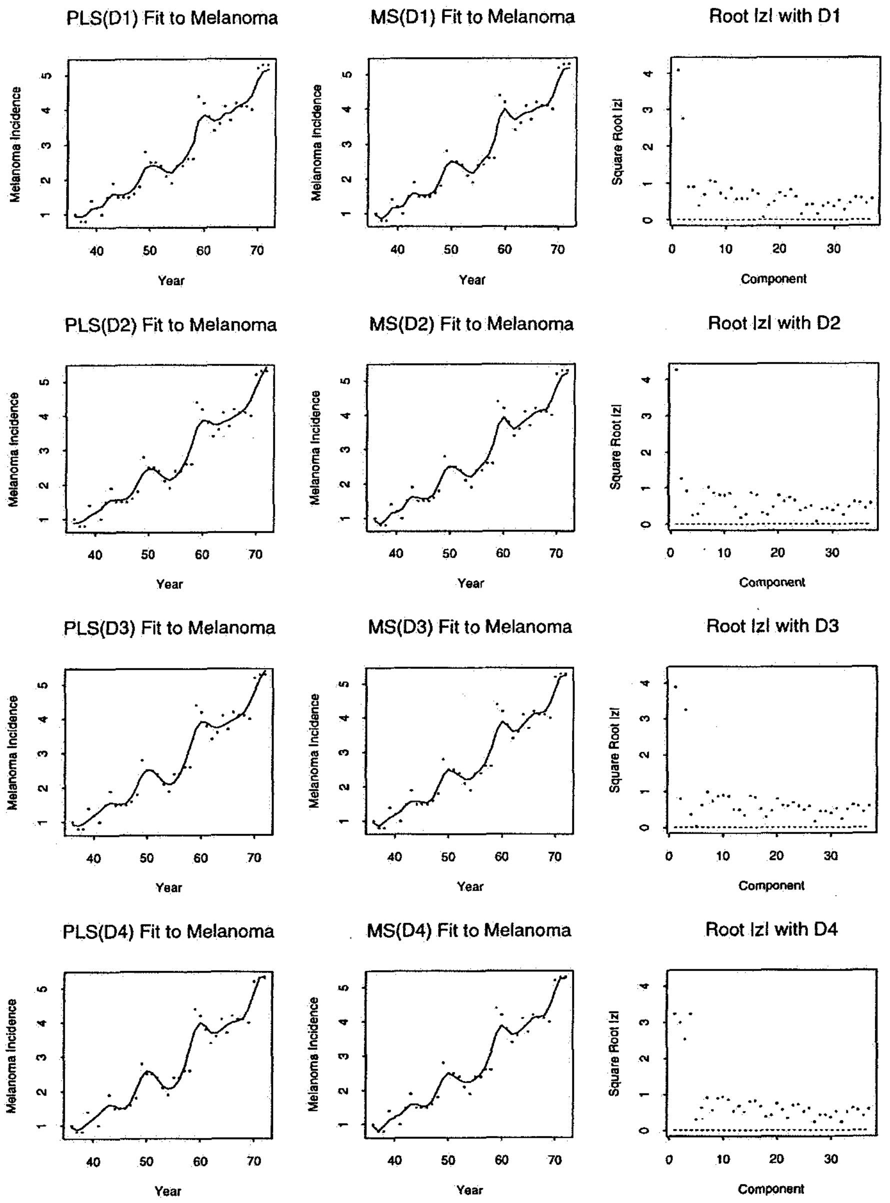

Fig. 3. Row d displays the shrinkage-adaptive PLS(D_d) and MS(D_d) estimates for mean melanoma incidence and, in the third column, the associated basis-economy plot.

Adaptation over penalty bases. Having devised a set $\mathcal{D}$ of candidate penalty matrices, we may use estimated risk to select an empirically best PLS or MS estimator by extending the adaptation method described in Section 2. Over shrinkage class $\mathcal{F}$ and penalty matrix class $\mathcal{D}$, the fully adaptive estimator of μ is defined to be $\hat{\mu}_{\mathcal{D},\mathcal{F}} = \hat{\mu}(\hat{D}, \hat{f})$, where

$$(\hat{D}, \hat{f}) = \operatorname*{argmin}_{D \in \mathcal{D}, f \in \mathcal{F}} \hat{\rho}(D, f). \tag{3.8}$$

	$d=1$	$d=2$	$d=3$	$d=4$
PLS(D_d) plug-in loss	.0310	.0294	.0326	.0349
MS(D_d) plug-in loss	.0165	.0166	.0194	.0230

If the cardinality of $\mathcal{D}$ is $o(p^{1/2})$, $\mathcal{F}$ is a closed subset of $\mathcal{F}_{MS}$, and $E|\hat{\sigma}^2 - \sigma^2| = O(p^{1/2})$, then Theorem 2.1 and Corollary 2.1 may be extended to justify the simultaneous adaptation in (3.8) over both f and D. The extension follows from the error bounds established in Theorems 2.1 and 2.2 of Beran and Dümbgen (1998). Justifying adaptation over larger classes of penalty matrices is an open question. Because local polynomials of degree up to 6 or so approximate a wide range of smooth mean vectors, adaptation over large $\mathcal{D}$ need not be advantageous.

Figure 3 exhibits competing adaptive PLS and MS estimates for mean melanoma incidence in males based on measurements for the years 1936 to 1972 and using the first-difference variance estimator $\hat{\sigma}^2_{D1}$. The data is given on pp. 199–201 of Andrews and Herzberg (1985). The first two columns in Fig. 3 display linearly interpolated PLS(D_d) and MS(D_d) fits to the data, the candidate penalty matrices being $\{D_d: 1 \leq d \leq 4\}$. The plug-in loss/risk estimates for these competing fits are shown in Table 3.

The loss/risk estimate for the least squares estimator of μ, which coincides here with the raw data, is .1165. It is not too surprising that the PLS(D_2) and MS(D_2) estimators have relatively low estimated risk among this group of competing shrinkage-adaptive estimators because the underlying trend in the melanoma data is roughly linear. The plotted shrinkage-adaptive estimators capture ripples in melanoma incidence that are associated with the sunspot cycle. It is notable that the competing adaptive fits in Fig. 3 are visually similar, even though their estimated risks differ. Heckman and Ramsay (2000) obtained similar fits to this data with continuous-spline penalized least squares, using differential penalty operators analogous to D_d and choosing penalty weight by generalized cross-validation or by equivalent degrees-of-freedom. Their treatment also considered a penalty differential operator that annihilates sinusoids of specified frequency.

The third column in Fig. 3 plots the components $\{|z_i|^{1/2}\}$ against i for each of the four penalty bases considered. Such diagnostic plots will be called *basis-economy plots.* The square root transformation reduces the vertical range and makes more visible the values near zero. The purpose of a basis-economy plot is to approximate the unobservable ideal plot of the $\{|\xi_i|^{1/2}\}$ against i so as to assess the economy of the penalty basis. For the melanoma data, the penalty basis generated by D_2 is empirically the most economical in Fig. 3. This finding is consistent with the ranking of estimated risks described above. At the same time, all four penalty matrices $\{D_d: 1 \leq d \leq 4\}$ yield similar looking fits.

3.3 *Diagnostic tools*

The foregoing theory and examples have identified two key factors that govern the risk of PLS and MS estimators. The first and more important factor is the economy of the basis U generated by the penalty matrix D. The second factor is the extent to which adaptive monotone shrinkage or penalized least squares shrinkage is able to exploit whatever economy exists in the chosen basis U. Flexibility in the shrinkage strategy becomes particularly important when, as columns two and three of Fig. 1, high-frequency details in the unknown mean entail that strict economy does not hold.

For a given penalty matrix, a comparative shrinkage-vector plot displays, with linear

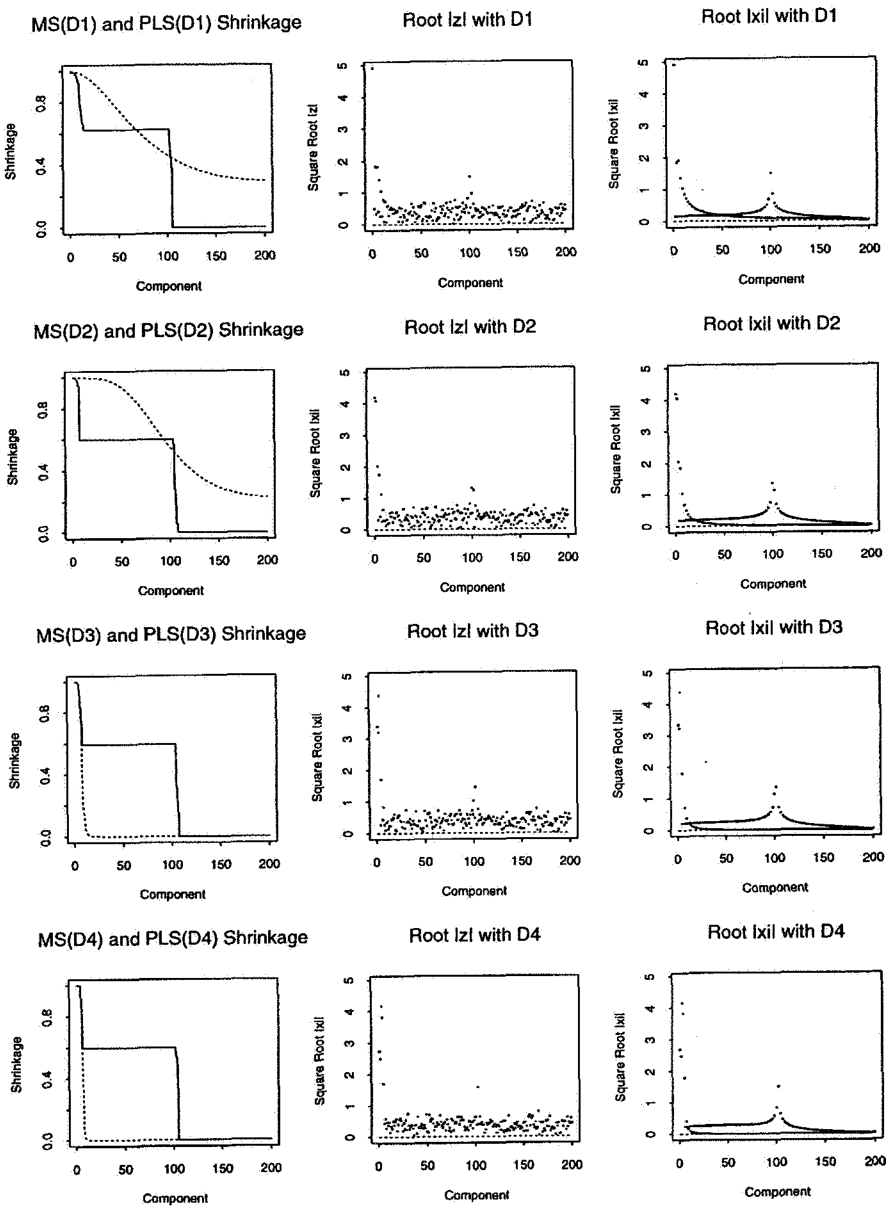

Fig. 4. On the Very Wiggly data, row d displays the shrinkage vectors for estimators PLS(D_d) (solid interpolation) and MS(D_d) (dashed interpolation), the basis economy plot, and the ideal basis economy plot.

interpolation for visibility, the components of the adaptively chosen shrinkage vectors $\hat{f}_{PLS}$ and $\hat{f}_{MS}$. Setting such a plot next to the basis-economy plot enables one to assess how well adaptive MS or PLS estimation exploits the degree of economy present in the penalty basis. For the Very Wiggly data described in the Introduction, the first two columns in Fig. 4 display the shrinkage-vectors and basis-economy plots generated by penalty matrices D_1 through D_4. The D_4 basis appears more economical than the other three penalty bases, but not much more than the D_3 basis. However, for either the D_3 or D_4 basis, adaptive PLS does a poor job of mimicking adaptive MS. This phenomenon underlies the inability of estimate PLS(D_4) in Fig. 1 to recover the sinusoidal component of trend. The third column of Fig. 4 plots the actual components of ξ, which are available here because the data is artificial and μ is known. It is gratifying that the empirical basis-economy plots in the middle column capture the essential features found in the ideal plots of the third column.

Feedback about which nonparametric regression procedure to use in a particular data analysis can come from estimated performance summaries as well as from diagnostic plots. A broadband diagnostic approach is surely more effective than any single tool. Plug-in estimated losses sharpen our scrutiny of the fits and diagnostic plots in Figs. 1 to 4. However, the discussion accompanying Fig. 1 indicated that plug-in estimated loss/risk for an adaptive PLS or MS estimate tends to underestimate true loss. We therefore consider another approach to estimating the loss or risk of a general estimator $\hat{\mu} = \hat{\mu}(y)$. Let $g(y) = \hat{\mu}(y) - y$. If the function g satisfies assumptions detailed in Stein (1981), then the risk of $\hat{\mu}$ under the Gaussian model described in the Introduction is

$$R(\hat{\mu}, \mu, \sigma^2) = \sigma^2 + E\left[2\sigma^2 n^{-1} \sum_{i=1}^{n} \partial g_i(y)/\partial y_i + n^{-1}|g(y)|^2\right]. \tag{3.9}$$

The implied estimator of loss or risk is

$$\hat{L}(\hat{\mu}) = \hat{\sigma}^2 + 2\hat{\sigma}^2 n^{-1} \sum_{i=1}^{n} \partial g_i(y)/\partial y_i + n^{-1}|g(y)|^2. \tag{3.10}$$

When $\hat{\mu}(y)$ lacks a tractable closed form, the partial derivatives needed in (3.10) may be approximated numerically. Let v_i denote the vector in R^n whose i-th component is 1 and whose other components are 0. Then, for small real values of δ,

$$\partial g_i(y)/\partial y_i \approx \delta^{-1}[g_i(y + \delta v_i) - g_i(y)], \qquad 1 \le i \le n. \tag{3.11}$$

Computing these difference quotients requires computing $\hat{\mu}(y) = y + g(y)$ and the n perturbed estimators $\{\hat{\mu}(y + \delta v_i): 1 \le i \le n\}$.

Sometimes the Stein loss/risk estimator in (3.10) has a closed form expression. For the candidate estimators $\hat{\mu}_{PLS}(D, \nu)$ or $\hat{\mu}_{MS}(D, f)$, the estimator (3.10) reduces to $\hat{\rho}(D, f(\nu))$ or $\hat{\rho}(D, f)$ respectively. For either PLS(D) or MS(D), the loss, the risk, and the plug-in loss/risk estimator converge together as p tends to infinity; Theorem 2.1 and Corollary 2.1 give the details. However, the experiment reported in Section 2.3 indicates that the rate of convergence may not be swift and that plug-in loss/risk estimators may underestimate true loss.

Alternatively, we can construct by numerical approximation the Stein loss/risk estimator (3.11) for the shrinkage-adaptive estimators PLS(D) and MS(D). Does this approach produce better estimates of loss than the plug-in method? For the examples of Fig. 1, where the penalty matrix is D_4, the approximate Stein loss/risk estimate obtained from (3.11) with $\delta = .0001$ may be compared with their plug-in counterparts and the true losses (Table 4). In this table, the Stein and the plug-in estimates for the loss of MS(D_4) and PLS(D_4) are close; their ranking is the same; and the former is only slightly closer to the true loss in most cases. There is no compelling reason in this experiment to prefer the Stein loss/risk estimates over their computationally simpler plug-in counterparts.

	MS loss	MS Stein	MS plug-in	PLS loss	PLS Stein	PLS plug-in
Smooth	.0011	−.0061	−.0068	.0013	−.0061	−.0066
Wiggly	.0111	.0031	.0015	.0138	.0076	.0072
Very Wiggly	.0127	.0100	.0092	.0326	.0294	.0290

Acknowledgements

This research was supported in part by National Science Foundation Grant DMS99–70266.

References

Andrews, D. F. and Herzberg, A. M. (1985). *Data: A Collection of Problems from Many Fields for the Student and Research Worker*, Springer, New York.

Beran, R. (1996). Confidence sets centered at C_p estimators, *Ann. Inst. Statist. Math.*, **48**, 1–15.

Beran, R. (2000). REACT scatterplot smoothers: Superefficiency through basis economy, *J. Amer. Statist. Assoc.*, **63**, 155–171.

Beran, R. and Dümbgen, L. (1998). Modulation of estimators and confidence sets, *Ann. Statist.*, **26**, 1826–1856.

Buja, A., Hastie, T. and Tibshirani, R. (1989). Linear smoothers and additive models (with discussion), *Ann. Statist.*, **17**, 453–555.

Chu, C.-K. and Marron, J. S. (1991). Choosing a kernel regression estimator, *Statist. Sci.*, **6**, 404–436.

Heckman, N. E. and Ramsay, J. O. (2000). Penalized regression with model-based penalties, *Canad. J. Statist.*, **28**, 241–258.

Mallows, C. L. (1973). Some comments on C_p, *Technometrics*, **15**, 661–676.

Pinsker, M. S. (1980). Optimal filtration of square-integrable signals in Gaussian noise, *Problems Inform. Transmission*, **16**, 120–133.

Rice, J. (1984). Bandwidth choice for nonparametric regression, *Ann. Statist.*, **12**, 1215–1230.

Robertson, T., Wright, F. T. and Dykstra, R. L. (1988). *Order Restricted Statistical Inference*, Wiley, New York.

Simonoff, J. S. (1996). *Smoothing Methods in Statistics*, Springer, New York.

Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution, *Proc. Third Berkeley Symp. on Math. Statist. Prob.*, Vol. 1 (ed. J. Neyman), 197–206, University of California Press, Berkeley.

Stein, C. (1981). Estimation of the mean of a multivariate normal distribution, *Ann. Statist.*, **9**, 1135–1151.

Venables, W. N. and Ripley, B. D. (1999). *Modern Applied Statistics with S-PLUS*, 3rd ed., Springer, New York.

Ann. Inst. Statist. Math.
Vol. 54, No. 4, 918–933 (2002)
©2002 The Institute of Statistical Mathematics

INDEPENDENCE OF LIKELIHOOD RATIO CRITERIA FOR HOMOGENEITY OF SEVERAL POPULATIONS

TAKESI HAYAKAWA

Faculty of Economics, Hitotsubashi University, Kunitachi, Tokyo 186-8601, Japan

(Received October 17, 2000; revised November 8, 2001)

Abstract. Let Π_i be an i-th population with a probability density function $f(\cdot \mid \theta_i)$ with one dimensional unknown parameter θ_i, $i = 1, 2, \dots, k$. Let n_i sample be drawn from each Π_i. The likelihood ratio criteria $\lambda_{j|(j-1)}$ for testing hypothesis that the first j parameters are equal against alternative hypothesis that the first $(j-1)$ parameters are equal and the j-th parameter is different with the previous ones are defined, $j = 2, 3, \dots, k$. The paper shows the asymptotic independence of $\lambda_{j|(j-1)}$'s up to the order $1/n$ under a hypothesis of equality of k parameters, where n is a number of total samples.

Key words and phrases: Likelihood ratio criterion, asymptotic expansion, homogeneity of parameters, asymptotic independence.

1. Introduction

Bartlett (1937) dealt with the case of homogeneity of variances of k normal populations. As the exact distribution of a test statistic was unknown, he considered to give a good approximation based on a knowledge of the moments of it. The method consisted in multiplying $-2\log$ (likelihood ratio criterion) by a scalor factor which results in a statistic having the same moments as chi-square random variable ignoring quantities of order $1/n^2$, where n is the size of the total sample. This correction factor was known as Bartlett correction factor in the sense of the moment.

For a case of a general population Lawley (1956) considered an asymptotic behavior of the likelihood ratio criterion for testing a composite hypothesis and obtained Bartlett correction factor in the sense of the moment. He decomposed a log-likelihood ratio criterion into a sum of log-likelihood ratio criteria corresponding to a sequence of a nested hypothesis and he showed that each log-likelihood ratio criteria has a Bartlett correction factor in the sense of the moment.

Hayakawa (1977, 1987) gave an asymptotic expansion of the distribution function of a likelihood ratio criterion for testing a simple hypothesis up to the order $1/n$ and showed that Bartlett correction factor in the sense of the moment yields a statistic having a chi-square distribution ignoring quantities of the order $1/n^2$. This implies that Bartlett correction factor in the sense of the moment is same as a Bartlett correction factor in the sense of the distribution for a likelihood ratio criterion. For other statistic it is usually hard to claim this fact, for example, this does not hold for Rao's score statistic. Thus the concept of Bartlett correctness in the sense of the distribution is stronger than that of Bartlett correctness in the sense of the moment. Thus if a statistic has Bartlett correction factor in the sense of the distribution, we call that it is Bartlett correctable.

Harris (1986) and Cordeiro (1987) pointed out an incompleteness of Hayakawa's 1977 result for the case of composite hypothesis testing.

Bickel and Ghosh (1990) considered independence of a sequence of signed likelihood ratio criterion which corresponds to Lawley's decomposition of log-likelihood ratio criterion from Bayesian point of view, and Takemura and Kuriki (1996) also handled a similar problem from a frequentist point of view. Takemura and Kuriki introduced a new parameter transformation which makes some higher order cross moments of derivatives of log-likelihood ratio criterion vanish.

Let $X_i = [x_{i1}, x_{i2}, \ldots, x_{in_i}]$ be a random sample from the i-th population Π_i with probability density function (pdf) $f(x \mid \theta_i)$, $i = 1, 2, \ldots, k$ and θ_i's are one dimensional parameters. For testing a hypothesis of homogeneity of parameters

$$H : \theta_1 = \theta_2 = \cdots = \theta_k (= \theta, \text{say})$$

against the alternative

$$K : \text{violation of at least one equality},$$

the likelihood ratio criterion λ is defined as

$$\lambda = \frac{\prod_{i=1}^{k} \prod_{\alpha=1}^{n_i} f(x_{i\alpha} \mid \tilde{\theta})}{\prod_{i=1}^{k} \prod_{\alpha=1}^{n_i} f(x_{i\alpha} \mid \hat{\theta}_i)}, \tag{1.1}$$

where $\tilde{\theta}$ is the maximum likelihood estimator of θ based on $n = \sum_{i=1}^{k} n_i$ observations under H and $\hat{\theta}_i$ is the maximum likelihood estimator of θ_i based on n_i observations X_i.

This is a general set up of Bartlett's homogeneity of variances of k normal populations. Hayakawa (1993) studied the asymptotic behavior of the distribution of λ, and Hayakawa (1994) dealt with the case of p dimensional parameter and showed that Bartlett correction factor is closely related to the corresponding expression given by Hayakawa (1977) in the context of a one-sample problem. Hayakawa (2001) considered this problem by use of Rao's score statistic, and Hayakawa and Doi (1999) also considered this by use of Wald statistic.

Consider a sequence of hypotheses $H_{j|(j-1)}$ and $K_{j|(j-1)}$ defined as

$$H_{j|(j-1)} : \theta_1 = \cdots = \theta_{j-1} = \theta_j \quad \text{vs.} \quad K_{j|(j-1)} : \theta_1 = \cdots = \theta_{j-1} \neq \theta_j, \quad j = 2, 3, \ldots, k.$$

The likelihood rato criterion for testing $H_{j|(j-1)}$ is given by

$$\lambda_{j|(j-1)} = \frac{\prod_{i=1}^{j} \prod_{\alpha=1}^{n_i} f(x_{i\alpha} \mid \tilde{\theta}_i)}{\prod_{i=1}^{j-1} \prod_{\alpha=1}^{n_i} f(x_{i\alpha} \mid \tilde{\theta}_{j-1}) \prod_{\alpha=1}^{n_j} f(x_{j\alpha} \mid \hat{\theta}_j)}, \quad j = 2, 3, \ldots, k, \tag{1.2}$$

where $\tilde{\theta}_i$ is the maximum likelihood estimator of $\theta_1 = \cdots = \theta_i$ based on $\tilde{n}_i = \sum_{j=1}^{i} n_j$ observations $[X_1, \ldots, X_i]$. The likelihood ratio criterion λ for testing H against K is decomposed as

$$\lambda = \lambda_{2|1}\lambda_{3|2}\cdots\lambda_{k|(k-1)}. \tag{1.3}$$

The purpose of this paper is to show the independence of $\lambda_{j|(j-1)}$, $j = 2, 3, \dots, k$ up to the order $1/n$.

It should be noted that by use of an appropriate parameter transformation this testing problem is reduced to the case of one sample problem dealt with by Bickel and Ghosh (1990) and Takemura and Kuriki (1996). However, it is not self-evident to have a concrete expression because of a complexity of a parameter structure even though our problem is included in a general set up, and it would be worth to express a final result by original parameters and to be able to handle this problem without any choice of an appropriate prior probability density function.

2. Asymptotic independence of LRC

Let θ_i, $i = 1, 2, \dots, k$ be a univariate parameter. Defining the log-likelihood function based on independent random sample $x_{i1}, \dots, x_{in_i}$ by

$$L_i(\theta_i) = \sum_{\alpha=1}^{n_i} \log f(x_{i\alpha} \mid \theta_i), \qquad i = 1, 2, \dots, k,$$

the following notations and convensions will be used. We assume that each $L_i(\theta_i)$ is regular with respect to θ_i derivatives.

(i) $y_i^{(l)} = n_i^{-l/2} \sum_{\alpha=1}^{n_i} \dfrac{\partial^l L_i(\theta_i)}{\partial \theta_i^l}$, $l = 1, 2, 3, 4, \qquad y_i \equiv y_i^{(1)}$, $i = 1, 2, \dots, k$,

(ii) $m_{(r_1^{\alpha_1}, r_2^{\alpha_2}, \dots, r_l^{\alpha_l})}(\theta_i)$

$$= \int \left\{ \frac{\partial^{r_1} \log f(x \mid \theta_i)}{\partial \theta_i^{r_1}} \right\}^{\alpha_1} \cdots \left\{ \frac{\partial^{r_l} \log f(x \mid \theta_i)}{\partial \theta_i^{r_l}} \right\}^{\alpha_l} f(x \mid \theta_i)\, dx.$$

Bartlett identities (Barndorff-Nielsen and Cox (1994)) hold

$$m_{(2)}(\theta_i) + m_{(1^2)}(\theta_i) = 0,$$
$$m_{(3)}(\theta_i) + 3m_{(21)}(\theta_i) + m_{(1^3)}(\theta_i) = 0,$$
$$m_{(4)}(\theta_i) + 4m_{(31)}(\theta_i) + 3m_{(2^2)}(\theta_i) + 6m_{(21^2)}(\theta_i) + m_{(1^4)}(\theta_i) = 0,$$

(iii) $\rho_i = n_i/n (> 0)$, $\quad \sum_{i=1}^k \rho_i = 1$.

Under the hypothesis $H : \theta_1 = \cdots = \theta_k = \theta$ (say) all moments are expressed as $m_{(3)}(\theta) = m_{(3)}$, $m_{(21)}(\theta) = m_{(21)}$, $m_{(31)}(\theta) = m_{(31)}$, etc.

By noting

$$\lambda_{j|(j-1)} = \frac{\prod_{i=1}^{j}\prod_{\alpha=1}^{n_i} f(x_{i\alpha} \mid \tilde{\theta}_j)}{\prod_{i=1}^{j}\prod_{\alpha=1}^{n_i} f(x_{i\alpha} \mid \hat{\theta}_i)} \Bigg/ \frac{\prod_{i=1}^{j-1}\prod_{\alpha=1}^{n_i} f(x_{i\alpha} \mid \tilde{\theta}_{j-1})}{\prod_{i=1}^{j-1}\prod_{\alpha=1}^{n_i} f(x_{i\alpha} \mid \hat{\theta}_i)} = \lambda_{12\cdots j}/\lambda_{12\cdots(j-1)},$$

we have

$$-2\log \lambda_{j|(j-1)} = 2\log \lambda_{12\cdots(j-1)} - 2\log \lambda_{12\cdots j}.$$

By use of the asymptotic expansion (2) in Hayakawa (1994) we have

$$-2\log\lambda_{j|(j-1)} = w_0^{(j)} + w_1^{(j)} + w_2^{(j)} + o_p\left(\frac{1}{n}\right), \qquad j = 2, 3, \dots, k, \tag{2.1}$$

where

$$\begin{aligned}
w_0^{(j)} &= -\frac{y_j^2}{y_j^{(2)}} + \frac{\left(\sum_{i=1}^{j}\sqrt{\rho_i}y_i\right)^2}{\sum_{i=1}^{j}\rho_i y_i^{(2)}} - \frac{\left(\sum_{i=1}^{j-1}\sqrt{\rho_i}y_i\right)^2}{\sum_{i=1}^{j-1}\rho_i y_i^{(2)}},\\
w_1^{(j)} &= -\frac{1}{3}\frac{y_j^{(3)}y_j^3}{(y_j^{(2)})^3} - \frac{1}{3}\sum_{i=1}^{j-1}\rho_i\sqrt{\rho_i}y_i^{(3)}\frac{\left(\sum_{i=1}^{j-1}\sqrt{\rho_i}y_i\right)^3}{\left(\sum_{i=1}^{j-1}\rho_i y_i^{(2)}\right)^3}\\
&\quad + \frac{1}{3}\sum_{i=1}^{j}\rho_i\sqrt{\rho_i}y_i^{(3)}\frac{\left(\sum_{i=1}^{j}\sqrt{\rho_i}y_i\right)^3}{\left(\sum_{i=1}^{j}\rho_i y_i^{(2)}\right)^3},\\
w_2^{(j)} &= -\frac{1}{4}\frac{(y_j^{(3)})^2 y_j^4}{(y_j^{(2)})^5} + \frac{1}{12}\frac{y_j^{(4)}y_j^4}{(y_j^{(2)})^4}\\
&\quad - \frac{1}{4}\frac{\left(\sum_{i=1}^{j-1}\rho_i\sqrt{\rho_i}y_i^{(3)}\right)^2\left(\sum_{i=1}^{j-1}\sqrt{\rho_i}y_i\right)^4}{\left(\sum_{i=1}^{j-1}\rho_i y_i^{(2)}\right)^5} + \frac{1}{4}\frac{\left(\sum_{i=1}^{j}\rho_i\sqrt{\rho_i}y_i^{(3)}\right)^2\left(\sum_{i=1}^{j}\sqrt{\rho_i}y_i\right)^4}{\left(\sum_{i=1}^{j}\rho_i y_i^{(2)}\right)^5}\\
&\quad + \frac{1}{12}\sum_{i=1}^{j-1}\rho_i^2 y_i^{(4)}\frac{\left(\sum_{i=1}^{j-1}\sqrt{\rho_i}y_i\right)^4}{\left(\sum_{i=1}^{j-1}\rho_i y_i^{(2)}\right)^4} - \frac{1}{12}\sum_{i=1}^{j}\rho_i^2 y_i^{(4)}\frac{\left(\sum_{i=1}^{j}\sqrt{\rho_i}y_i\right)^4}{\left(\sum_{i=1}^{j}\rho_i y_i^{(2)}\right)^4}.
\end{aligned}$$

To find the moment generating function of these statistics we need to use the Edgeworth type expansion of the joint density function of y_i, $y_i^{(2)}$, $y_i^{(3)}$, $y_i^{(4)}$, $i = 1, 2, \dots, k$, which is stated as follows.

$$f = f_0\left[1 + \frac{1}{\sqrt{n}}F_1 + \frac{1}{n}F_2\right] + o\left(\frac{1}{n}\right), \tag{2.2}$$

where

$$f_0 = \prod_{i=1}^{k}(2\pi m_{(1^2)}(\theta_i))^{-1/2}\exp\{-y_i^2/2m_{(1^2)}(\theta_i)\}\prod_{l=2}^{4}\delta_{li},$$

$$F_1 = \frac{1}{6}\sum_{i=1}^{k} \frac{1}{\sqrt{\rho_i}} m_{(1^3)}(\theta_i) H_3(y_i) - \sum_{i=1}^{k} \frac{1}{\sqrt{\rho_i}} m_{(21)}(\theta_i) H_1(y_i) d_{2i}^{(1)},$$

$$\begin{aligned}
F_2 = {} & \frac{1}{2}\sum_{i=1}^{k} \frac{1}{\rho_i}\{m_{(2^2)}(\theta_i) - m_{(2)}^2(\theta_i)\} d_{2i}^{(2)} \\
& - \frac{1}{2}\sum_{i=1}^{k} \frac{1}{\rho_i}\{m_{(21^2)}(\theta_i) - m_{(2)}(\theta_i) m_{(1^2)}(\theta_i)\} H_2(y_i) d_{2i}^{(1)} \\
& - \sum_{i=1}^{k} \frac{1}{\rho_i} m_{(31)}(\theta_i) H_1(y_i) d_{3i}^{(1)} + \frac{1}{2}\sum_{i=1}^{k} \frac{1}{\rho_i} m_{(21)}^2(\theta_i) H_2(y_i) d_{2i}^{(2)} \\
& + \frac{1}{24}\sum_{i=1}^{k} \frac{1}{\rho_i}\{m_{(1^4)}(\theta_i) - 3m_{(1^2)}^2(\theta_i)\} H_4(y_i) \\
& - \frac{1}{6}\sum_{i=1}^{k} \frac{1}{\rho_i} m_{(21)}(\theta_i) m_{(1^3)}(\theta_i) H_4(y_i) d_{2i}^{(1)} \\
& + \frac{1}{72}\sum_{i=1}^{k} \frac{1}{\rho_i} m_{(1^3)}^2(\theta_i) H_6(y_i) \\
& + \frac{1}{2}\sum_{i\neq j} \frac{1}{\sqrt{\rho_i \rho_j}} m_{(21)}(\theta_i) m_{(21)}(\theta_j) H_1(y_i) H_1(y_j) d_{2i}^{(1)} d_{2j}^{(1)} \\
& - \frac{1}{6}\sum_{i\neq j} \frac{1}{\sqrt{\rho_i \rho_j}} m_{(21)}(\theta_i) m_{(1^3)}(\theta_j) H_1(y_i) H_3(y_j) d_{2i}^{(1)} \\
& + \frac{1}{72}\sum_{i\neq j} \frac{1}{\sqrt{\rho_i \rho_j}} m_{(1^3)}(\theta_i) m_{(1^3)}(\theta_j) H_3(y_i) H_3(y_j),
\end{aligned}$$

$$\delta_{li} = \delta(y_i^{(l)} - m_{(l)}(\theta_i)/n_i^{(l-2)/2}),$$

$$d_{li}^{(r)} = \delta^{(r)}(y_i^{(l)} - m_{(l)}(\theta_i)/n_i^{(l-2)/2})/\delta(y_i^{(l)} - m_{(l)}(\theta_i)/n_i^{(l-2)/2})$$

and $\delta^{(r)}$ is the r-th derivative of Dirac delta function δ. $H_r(y)$ is defined by

$$\text{(2.3)} \qquad \frac{d^n}{dy^n}\exp\left(-\frac{y^2}{2m_{(1^2)}}\right) = (-1)^n H_n(y) \exp\left(-\frac{y^2}{2m_{(1^2)}}\right).$$

PROPOSITION. *Under the hypothesis H, $-2\log\lambda_{j|(j-1)}$, $j = 2, 3, \ldots, k$ are mutually independent in the limit.*

PROOF. With help of law of large numbers $y_i^{(2)}$ converges to $m_{(2)}(\theta_i) = -m_{(1^2)}(\theta_i)$ in the limit and $y_i^{(l)}$, $l \geq 3$ converges to zero in the limit, respectively. Thus we have

$$\begin{aligned}
\text{(2.4)} \quad \tilde{w}_0^{(2)} &= \underset{\substack{n_i\to\infty \\ i=1,\cdots,k}}{p\lim} \{-2\log\lambda_{2|1}\} \\
&= \frac{1}{m_{(1^2)}(\theta)}\left\{\sum_{i=1}^{2} y_i^2 - \frac{1}{\hat{\rho}_2}\left(\sum_{i=1}^{2}\sqrt{\rho_i} y_i\right)^2\right\},
\end{aligned}$$

$$(2.5)\quad \tilde{w}_0^{(j)} = \underset{\substack{n_i\to\infty \\ i=1,\cdots,k}}{p\lim}\,\{-2\log\lambda_{j|(j-1)}\}$$

$$= \frac{1}{m_{(1^2)}(\theta)}\left\{y_j^2 - \frac{1}{\hat{\rho}_j}\left(\sum_{i=1}^{j}\sqrt{\rho_i}y_i\right)^2 + \frac{1}{\hat{\rho}_{j-1}}\left(\sum_{i=1}^{j-1}\sqrt{\rho_i}y_i\right)^2\right\}, \qquad 3\le j\le k,$$

where $\hat{\rho}_j = \sum_{i=1}^{j}\rho_i$.

Define $P_j = \sqrt{\rho_j^*}\sqrt{\rho_j^*}'$, $I_{jj} = e_je_j'$ and $y = (y_1, y_2, \dots, y_k)'$, where $\sqrt{\rho_j^*} = (\sqrt{\rho_1}, \dots, \sqrt{\rho_j}, 0, \dots, 0)'$ and $e_j' = (0, \dots, 0, \overset{j}{\check{1}}, 0, \dots, 0)$. Then (2.4), (2.5) are expressed as

$$(2.6)\qquad \frac{1}{m_{(1^2)}(\theta)}y'Q_jy, \qquad Q_j = I_{jj} - P_j + P_{j-1}, \quad j = 2, 3, \dots, k$$

and y is k dimensional random vector with mean 0 and covariance matrix $m_{(1^2)}(\theta)I_k$.

By noting

$$(2.7)\qquad Q_jQ_l = \delta_{jl}Q_j, \qquad rank\,Q_j = 1$$

and by use of Craig theorem (e.g. Ogawa (1949), and others) we have that $-2\log\lambda_{j|(j-1)}$, $j = 2, 3, \dots, k$ are mutually independent and these have chi-square distribution with one degree of freedom in the limit, respectively.

THEOREM. *Under the hypothesis H, the joint moment generating function of $-2\log\lambda_{j|(j-1)}$, $j = 2, 3, \dots, k$ is expressed as*

$$(2.8)\quad M(t_2, \dots, t_k) = E\left[\exp\left\{\sum_{j=2}^{k} t_j(-2\log\lambda_{j|(j-1)})\right\}\right]$$

$$= \prod_{j=2}^{k}\frac{1}{(1-2t_j)^{1/2}}\cdot\left\{1 + \frac{1}{n}A_j\left(\frac{1}{1-2t_j} - 1\right) + o\left(\frac{1}{n}\right)\right\},$$

where

$$(2.9)\qquad A_j = a\frac{1}{\hat{\rho}_j}\left(\frac{\hat{\rho}_{j-1}}{\rho_j} + 1 + \frac{\rho_j}{\hat{\rho}_{j-1}}\right),$$

$$a = \frac{1}{8}\frac{1}{(m_{(1^2)})^2}\left\{m_{(2^2)} - m_{(1^4)} - 2m_{(21^2)}\right\}$$

$$+\frac{1}{24}\frac{1}{(m_{(1^2)})^3}\left\{m_{(3)}m_{(21)} - 5m_{(3)}m_{(1^3)} - 8m_{(21)}m_{(1^3)}\right\}.$$

This implies that $-2\log\lambda_{j|(j-1)}$, $j = 2, 3, \dots, k$ are mutually independent up to the order $1/n$.

PROOF. The proof is given in Section 3.

Note. If we set $t_2 = t_3 = \cdots = t_k = t$, then the moment generating function is reduced to the one of $-2\log\lambda$ for testing H against K. By noting $\hat{\rho}_1 = \rho_1$ and $\hat{\rho}_k = 1$, and

$$\sum_{j=2}^{k} \frac{1}{\hat{\rho}_j}\left(\frac{\hat{\rho}_{j-1}}{\rho_j} + 1 + \frac{\rho_j}{\hat{\rho}_{j-1}}\right) = \sum_{j=1}^{k} \frac{1}{\rho_j} - 1,$$

the moment generating function is expressed as

$$(2.10) \qquad (1-2t)^{-1/2(k-1)}\left[1 + \frac{a}{n}\left(\sum_{i=1}^{k}\frac{1}{\rho_i} - 1\right)\left(\frac{1}{1-2t} - 1\right) + o\left(\frac{1}{n}\right)\right].$$

This is the one given by Hayakawa (1994).

Note. $-2\log\lambda_{j|(j-1)}$, $j = 2, 3, \ldots, k$ are Bartlett correctable.

3. Proof of theorem

The joint moment generating function of $-2\log\lambda_{j|(j-1)}$, $j = 2, 3, \ldots, k$ is expressed as

$$(3.1) \qquad M(t_2, \ldots, t_k) = \int \exp\left\{\sum_{j=2}^{k} t_j w_0^{(j)} + \sum_{j=2}^{k} t_j w_1^{(j)} + \sum_{j=2}^{k} t_j w_2^{(j)}\right\}$$
$$\times f_0\left[1 + \frac{1}{\sqrt{n}}F_1 + \frac{1}{n}F_2\right] dy dy^{(2)} dy^{(3)} dy^{(4)} + o\left(\frac{1}{n}\right),$$

where $y^{(l)} = (y_1^{(l)}, y_2^{(l)}, \ldots, y_k^{(l)})'$, $l = 2, 3, 4$.

The limit of the moment generation function is expressed as

$$(3.2) \qquad \int \exp\left\{\sum_{j=2}^{k} t_j \tilde{w}_0^{(j)}\right\} \prod_{i=1}^{k} n(0, m_{(1^2)}(\theta)) dy$$
$$= |\Omega|^{1/2} \int \frac{1}{(2\pi m_{(1^2)}(\theta))^{k/2}|\Omega|^{1/2}} \exp\left\{-\frac{1}{2m_{(1^2)}(\theta)} y'\Omega^{-1}y\right\} dy,$$

where

$$(3.3) \qquad \Omega = \sum_{j=2}^{k} c_j Q_j + \sqrt{\rho}\sqrt{\rho}', \qquad \sqrt{\rho}' = (\sqrt{\rho_1}, \sqrt{\rho_2}, \ldots, \sqrt{\rho_k}),$$
$$c_j = (1-2t_j)^{-1}, \qquad j = 2, 3, \ldots, k, \quad |\Omega| = \prod_{j=2}^{k} c_j.$$

This implies that y is dealt with as a normal random vector with mean zero and covariance matrix $m_{(1^2)}(\theta)\Omega$.

We use following different expectation notations according to the order situations.

$$(3.4)\quad |\Omega|^{1/2}\hat{E}[g] = \int g\cdot\exp\left\{\sum_{j=2}^{k} t_j w_0^{(j)}\right\} f_0 dy dy^{(2)} dy^{(3)} dy^{(4)},$$

$$(3.5)\quad |\Omega|^{1/2}\hat{\hat{E}}[g] = \int g\cdot\exp\left\{\sum_{j=2}^{k} t_j w_0^{(j)} + \sum_{j=2}^{k} t_j w_1^{(j)}\right\} f_0 dy dy^{(2)} dy^{(3)} dy^{(4)},$$

$$(3.6)\quad |\Omega|^{1/2}\hat{\hat{\hat{E}}}[g] = \int g\cdot\exp\left\{\sum_{j=2}^{k} t_j w_0^{(j)} + \sum_{j=2}^{k} t_j w_1^{(j)} + \sum_{j=2}^{k} t_j w_2^{(j)}\right\} f_0 dy dy^{(2)} dy^{(3)} dy^{(4)}.$$

Thus we have variance and covariance of y with respect to an operator $\hat{E}$ as follows.

$$(3.7)\quad \hat{E}[1] = 1,$$

$$(3.8)\quad \hat{E}[y_1^2/m_{(1^2)}] = w_{11} = \rho_1\left[c_2\frac{\rho_2}{\rho_1\hat{\rho}_2} + c_3\frac{\rho_3}{\hat{\rho}_2\hat{\rho}_3} + \cdots + c_k\frac{\rho_k}{\hat{\rho}_{k-1}\hat{\rho}_k} + 1\right],$$

$$(3.9)\quad \hat{E}[y_j^2/m_{(1^2)}] = w_{jj} = \rho_j\left[c_j\frac{\hat{\rho}_{j-1}}{\rho_j\hat{\rho}_j} + c_{j+1}\frac{\rho_{j+1}}{\hat{\rho}_j\hat{\rho}_{j+1}} + \cdots + c_k\frac{\rho_k}{\hat{\rho}_{k-1}\hat{\rho}_k} + 1\right],\quad j = 2, 3, \ldots, k,$$

$$(3.10)\quad \hat{E}[y_j y_l/m_{(1^2)}] = w_{jl} = \sqrt{\rho_j\rho_l}\left[c_l\left(-\frac{1}{\hat{\rho}_l}\right) + c_{l+1}\frac{\rho_{l+1}}{\hat{\rho}_l\hat{\rho}_{l+1}} + \cdots + c_k\frac{\rho_k}{\hat{\rho}_{k-1}\hat{\rho}_k} + 1\right],\quad 1 \le j < l \le k.$$

where $c_j = (1-2t_j)^{-1}$, $j = 2, 3, \ldots, k$.

Hereafter we give several moments.

(I) The integration of the first term in (2.2).

$$\int \exp\left\{\sum_{j=2}^{k} t_j w_0^{(j)} + \sum_{j=2}^{k} t_j w_1^{(j)} + \sum_{j=2}^{k} t_j w_2^{(j)}\right\} f_0 dy dy^{(2)} dy^{(3)} dy^{(4)}$$
$$= \int \exp\left\{\sum_{j=2}^{k} t_j \tilde{w}_0^{(j)}\right\} n(0, m_{(1^2)}(\theta) I)$$
$$\cdot\left[1 + \frac{1}{\sqrt{n}}\sum_{j=2}^{k} t_j\tilde{w}_1^{(j)} + \frac{1}{n}\left\{\sum_{j=2}^{k} t_j\tilde{w}_2^{(j)} + \frac{1}{2}\left(\sum_{j=2}^{k} t_j\tilde{w}_1^{(j)}\right)^2\right\}\right] dy + o\left(\frac{1}{n}\right),$$

where

$$\tilde{w}_1^{(2)} = \frac{1}{3}\frac{m_{(3)}}{(m_{(1^2)})^3}\left[\sum_{i=1}^{2}\frac{y_i^3}{\sqrt{\rho_i}} - \frac{1}{\hat{\rho}_2^2}\left(\sum_{i=1}^{2}\sqrt{\rho_i}y_i\right)^3\right],$$

$$\tilde{w}_1^{(j)} = \frac{1}{3}\frac{m_{(3)}}{(m_{(1^2)})^3}\left[\frac{y_j^3}{\sqrt{\rho_j}} - \frac{1}{\hat{\rho}_j^2}\left(\sum_{i=1}^{j}\sqrt{\rho_i}y_i\right)^3 + \frac{1}{\hat{\rho}_{j-1}^2}\left(\sum_{i=1}^{j-1}\sqrt{\rho_i}y_i\right)^3\right],\quad 3 \le j \le k,$$

and

$$\tilde{w}_2^{(2)} = \left\{ \frac{1}{4} \frac{(m_{(3)})^2}{(m_{(1^2)})^5} + \frac{1}{12} \frac{m_{(4)}}{(m_{(1^2)})^4} \right\} \left[\sum_{i=1}^{2} \frac{y_i^4}{\rho_i} - \frac{1}{\hat{\rho}_2^3} \left(\sum_{i=1}^{2} \sqrt{\rho_i} y_i \right)^4 \right],$$

$$\tilde{w}_2^{(j)} = \left\{ \frac{1}{4} \frac{(m_{(3)})^2}{(m_{(1^2)})^5} + \frac{1}{12} \frac{m_{(4)}}{(m_{(1^2)})^4} \right\}$$
$$\times \left[\frac{y_j^4}{\rho_j} - \frac{1}{\hat{\rho}_j^3} \left(\sum_{i=1}^{j} \sqrt{\rho_i} y_i \right)^4 + \frac{1}{\hat{\rho}_{j-1}^3} \left(\sum_{i=1}^{j-1} \sqrt{\rho_i} y_i \right)^4 \right], \quad 3 \le j \le k.$$

Thus by setting $u_j = c_j - 1$, $j = 2, 3, \ldots, k$, we have

$$(3.11) \quad \hat{E}\left[\sum_{j=2}^{k} t_j \tilde{w}_1^{(j)} \right] = 0,$$

$$(3.12) \quad \hat{E}\left[\sum_{j=2}^{k} t_j \tilde{w}_2^{(j)} \right] = \left\{ \frac{3}{8} \frac{(m_{(3)})^2}{(m_{(1^2)})^3} + \frac{1}{8} \frac{m_{(4)}}{(m_{(1^2)})^2} \right\} \left[\sum_{j=2}^{k} u_j^2 \frac{1}{\hat{\rho}_j} \left(\frac{\hat{\rho}_{j-1}}{\rho_j} - 1 + \frac{\rho_j}{\hat{\rho}_{j-1}} \right) \right.$$
$$\left. + \sum_{j=2}^{k} u_j \frac{1}{\hat{\rho}_j} \left(\frac{\hat{\rho}_{j-1}}{\rho_j} + 1 + \frac{\rho_j}{\hat{\rho}_{j-1}} \right) + 2 \sum_{2 \le p < q \le k} u_p u_q \frac{\rho_q}{\hat{\rho}_{q-1} \hat{\rho}_q} \right],$$

$$(3.13) \quad \hat{E}\left[\frac{1}{2} \left\{ \sum_{j=2}^{k} t_j \tilde{w}_1^{(j)} \right\}^2 \right] = \frac{(m_{(3)})^2}{(m_{(1^2)})^3} \left[\frac{5}{24} \sum_{j=2}^{k} u_j^3 \frac{1}{\hat{\rho}_j} \left(\frac{\hat{\rho}_{j-1}}{\rho_j} - 2 + \frac{\rho_j}{\hat{\rho}_{j-1}} \right) \right.$$
$$+ \frac{1}{24} \sum_{j=2}^{k} u_j^2 \frac{1}{\hat{\rho}_j} \left(5 \frac{\hat{\rho}_{j-1}}{\rho_j} - 1 + 5 \frac{\rho_j}{\hat{\rho}_{j-1}} \right) + \frac{1}{4} \sum_{2 \le p < q < r \le k} u_p u_q u_r \frac{\rho_r}{\hat{\rho}_{r-1} \hat{\rho}_r}$$
$$+ \frac{3}{8} \sum_{2 \le p < q \le k} u_p^2 u_q \frac{\rho_q}{\hat{\rho}_{q-1} \hat{\rho}_q} + \frac{1}{4} \sum_{2 \le p < q \le k} u_p u_q^2 \frac{(\rho_q - \hat{\rho}_{q-1})}{\hat{\rho}_{q-1} \hat{\rho}_q}$$
$$\left. + \frac{1}{4} \sum_{2 \le p < q \le k} u_p u_q \frac{\rho_q}{\hat{\rho}_{q-1} \hat{\rho}_q} \right].$$

(II) The integration of the second term in (2.2).

To have terms of order $1/n$ it is enough only to use up to the second term in the exponent,

$$\int \exp\left\{ \sum_{j=2}^{k} t_j w_0^{(j)} + \sum_{j=2}^{k} t_j w_1^{(j)} \right\} f_0 F_1 dy dy^{(2)} dy^{(3)} dy^{(4)}.$$

$$(\text{II-1}) \quad \int \exp\left\{ \sum_{j=2}^{k} t_j w_0^{(j)} + \sum_{j=2}^{k} t_j w_1^{(j)} \right\} f_0 \frac{m_{(1^3)}}{6} \sum_{i=1}^{k} \frac{1}{\sqrt{\rho_i}} H_3(y_i) dy dy^{(2)} dy^{(3)} dy^{(4)}$$
$$= \frac{1}{\sqrt{n}} \int \exp\left\{ \sum_{j=2}^{k} t_j \tilde{w}_0^{(j)} \right\} n(0, m_{(1^2)}(\theta) I_k) \left\{ \sum_{j=2}^{k} t_j \tilde{w}_1^{(j)} \right\}$$

$$\times \frac{m_{(1^3)}}{6} \sum_{i=1}^{k} \frac{1}{\sqrt{\rho_i}} H_3(y_i) dy + o\left(\frac{1}{\sqrt{n}}\right),$$

$$\begin{aligned}
(3.14)\ \ \hat{E}&\left[\sum_{j=2}^{k} t_j \tilde{w}_1^{(j)} \frac{m_{(1^3)}}{6} \sum_{i=1}^{k} \frac{1}{\sqrt{\rho_i}} H_3(y_i)\right] \\
&= \frac{m_{(3)} m_{(1^3)}}{(m_{(1^2)})^3} \left[\frac{5}{12} \sum_{j=2}^{k} u_j^3 \frac{1}{\hat{\rho}_j} \left(\frac{\hat{\rho}_{j-1}}{\rho_j} - 2 + \frac{\rho_j}{\hat{\rho}_{j-1}}\right)\right. \\
&\quad + \frac{1}{12} \sum_{j=2}^{k} u_j^2 \frac{1}{\hat{\rho}_j} \left(7 \frac{\hat{\rho}_{j-1}}{\rho_j} - 5 + 7 \frac{\rho_j}{\hat{\rho}_{j-1}}\right) + \frac{1}{6} \sum_{j=2}^{k} u_j \frac{1}{\hat{\rho}_j} \left(\frac{\hat{\rho}_{j-1}}{\rho_j} + 1 + \frac{\rho_j}{\hat{\rho}_{j-1}}\right) \\
&\quad + \frac{1}{2} \sum_{2 \le p < q < r \le k} u_p u_q u_r \frac{\rho_r}{\hat{\rho}_{r-1} \hat{\rho}_r} + \frac{3}{4} \sum_{2 \le p < q \le k} u_p^2 u_q \frac{\rho_q}{\hat{\rho}_{q-1} \hat{\rho}_q} \\
&\quad \left. + \frac{1}{2} \sum_{2 \le p < q \le k} u_p u_q^2 \frac{(\rho_q - \hat{\rho}_{q-1})}{\hat{\rho}_{q-1} \hat{\rho}_q} + \sum_{2 \le p < q \le k} u_p u_q \frac{\rho_q}{\hat{\rho}_{q-1} \hat{\rho}_q}\right].
\end{aligned}$$

$$\text{(II-2)} \qquad - \frac{m_{(21)}}{m_{(1^2)}} \int \exp\left\{\sum_{j=2}^{k} t_j w_0^{(j)} + \sum_{j=2}^{k} t_j w_1^{(j)}\right\} f_0 \sum_{i=1}^{k} \frac{y_i}{\sqrt{\rho_i}} d_{2i}^{(1)} dy dy^{(2)} dy^{(3)} dy^{(4)}.$$

Noting the integration with respect to $d_{2i}^{(l)}$,

$$(3.15) \qquad \int h(y_i^{(2)}) f_0 d_{2i}^{(l)} dy_i^{(2)} = (-1)^l \left.\frac{\partial^l h}{\partial (y_i^{(2)})^l}\right|_{y_i^{(2)} = m_{(2)} = -m_{(1^2)}},$$

we have

$$\begin{aligned}
(3.16)\ \ -\frac{m_{(21)}}{m_{(1^2)}} \hat{E}\left[\sum_{i=1}^{k} \frac{y_i}{\sqrt{\rho_i}} d_{2i}^{(1)}\right] &= \frac{1}{\sqrt{n}} \frac{m_{(3)} m_{(21)}}{(m_{(1^2)})^3} \left[\frac{5}{4} \sum_{j=2}^{k} u_j^2 \frac{1}{\hat{\rho}_j} \left(\frac{\hat{\rho}_{j-1}}{\rho_j} - 2 + \frac{\rho_j}{\hat{\rho}_{j-1}}\right)\right. \\
&\quad + \frac{1}{4} \sum_{j=2}^{k} u_j^2 \frac{1}{\hat{\rho}_j} \left(11 \frac{\hat{\rho}_{j-1}}{\rho_j} - 7 + 11 \frac{\rho_j}{\hat{\rho}_{j-1}}\right) + \frac{3}{2} \sum_{j=2}^{k} u_j \frac{1}{\hat{\rho}_j} \left(\frac{\hat{\rho}_{j-1}}{\rho_j} + 1 + \frac{\rho_j}{\hat{\rho}_{j-1}}\right) \\
&\quad + \frac{3}{2} \sum_{2 \le p < q < r \le k} u_p u_q u_r \frac{\rho_r}{\hat{\rho}_{r-1} \hat{\rho}_r} + \frac{9}{4} \sum_{2 \le p < q \le k} u_p^2 u_q \frac{\rho_q}{\hat{\rho}_{q-1} \hat{\rho}_q} \\
&\quad \left. + \frac{3}{2} \sum_{2 \le p < q \le k} u_p u_q^2 \frac{(\rho_q - \hat{\rho}_{q-1})}{\hat{\rho}_{q-1} \hat{\rho}_q} + \frac{9}{2} \sum_{2 \le p < q \le k} u_p u_q \frac{\rho_q}{\hat{\rho}_{q-1} \hat{\rho}_q}\right] + O\left(\frac{1}{n}\right).
\end{aligned}$$

(III) The terms of order $1/n$ in (2.2) are obtained after some lengthy algebra as follows.

$$(3.17) \qquad \hat{E}\left[\frac{1}{2} \sum_{i=1}^{k} \frac{1}{\rho_i} \left\{m_{(2^2)} - (m_{(2)})^2\right\} d_{2i}^{(2)}\right]$$

$$= \frac{(m_{(2^2)} - (m_{(2)})^2)}{(m_{(1^2)})^2} \left[\frac{3}{8} \sum_{j=2}^{k} u_j^2 \frac{1}{\hat{\rho}_j} \left(\frac{\hat{\rho}_{j-1}}{\rho_j} - 1 + \frac{\rho_j}{\hat{\rho}_{j-1}} \right) \right.$$

$$\left. + \frac{1}{2} \sum_{j=2}^{k} u_j \frac{1}{\hat{\rho}_j} \left(\frac{\hat{\rho}_{j-1}}{\rho_j} + 1 + \frac{\rho_j}{\hat{\rho}_{j-1}} \right) + \frac{3}{4} \sum_{2 \le p < q \le k} u_p u_q \frac{\rho_q}{\hat{\rho}_{q-1} \hat{\rho}_q} \right],$$

$$\text{(3.18)} \qquad \hat{E} \left[-\frac{1}{2} (m_{(21^2)} - m_{(2)} m_{(1^2)}) \sum_{i=1}^{k} \frac{1}{\rho_i} H_2(y_i) d_{2i}^{(1)} \right]$$

$$= \frac{(m_{(21^2)} - m_{(2)} m_{(1^2)})}{(m_{(1^2)})^2} \left[\frac{3}{4} \sum_{j=2}^{k} u_j^2 \frac{1}{\hat{\rho}_j} \left(\frac{\hat{\rho}_{j-1}}{\rho_j} - 1 + \frac{\rho_j}{\hat{\rho}_{j-1}} \right) \right.$$

$$\left. + \frac{1}{2} \sum_{j=2}^{k} u_j \frac{1}{\hat{\rho}_j} \left(\frac{\hat{\rho}_{j-1}}{\rho_j} + 1 + \frac{\rho_j}{\hat{\rho}_{j-1}} \right) + \frac{3}{2} \sum_{2 \le p < q \le k} u_p u_q \frac{\rho_q}{\hat{\rho}_{q-1} \hat{\rho}_q} \right],$$

$$\text{(3.19)} \quad -m_{(31)} \hat{\hat{E}} \left[\sum_{i=1}^{k} \frac{1}{\rho_i} H_1(y_i) d_{3i}^{(1)} \right] = \frac{m_{(31)}}{(m_{(1^2)})^2} \left[\frac{1}{2} \sum_{j=2}^{k} u_j^2 \frac{1}{\hat{\rho}_j} \left(\frac{\hat{\rho}_{j-1}}{\rho_j} - 1 + \frac{\rho_j}{\hat{\rho}_{j-1}} \right) \right.$$

$$\left. + \frac{1}{2} \sum_{j=2}^{k} u_j \frac{1}{\hat{\rho}_j} \left(\frac{\hat{\rho}_{j-1}}{\rho_j} + 1 + \frac{\rho_j}{\hat{\rho}_{j-1}} \right) + \sum_{2 \le p < q \le k} u_p u_q \frac{\rho_q}{\hat{\rho}_{q-1} \hat{\rho}_q} \right] + O\left(\frac{1}{\sqrt{n}} \right),$$

$$\text{(3.20)} \quad \frac{1}{24} \left\{ m_{(1^4)} - 3(m_{(1^2)})^2 \right\} \hat{E} \left[\sum_{i=1}^{k} \frac{1}{\rho_i} H_4(y_i) \right] = \frac{m_{(1^4)} - 3(m_{(1^2)})^2}{(m_{(1^2)})^2}$$

$$\times \left[\frac{1}{8} \sum_{j=2}^{k} u_j^2 \frac{1}{\hat{\rho}_j} \left(\frac{\hat{\rho}_{j-1}}{\rho_j} - 1 + \frac{\rho_j}{\hat{\rho}_{j-1}} \right) + \frac{1}{4} \sum_{2 \le p < q \le k} u_p u_q \frac{\rho_q}{\hat{\rho}_{q-1} \hat{\rho}_q} \right],$$

$$\text{(3.21)} \, \frac{1}{2} (m_{(21)})^2 \hat{E} \left[\sum_{i=1}^{k} \frac{1}{\rho_i} H_2(y_i) d_{2i}^{(2)} \right]$$

$$= \frac{(m_{(21)})^2}{(m_{(1^2)})^3} \left[\frac{15}{8} \sum_{j=2}^{k} u_j^3 \frac{1}{\hat{\rho}_j^3} \left\{ \left(\sum_{\alpha=1}^{j-1} \rho_\alpha^2 \right) \frac{\rho_j^3}{\hat{\rho}_{j-1}^3} + \frac{\hat{\rho}_{j-1}^3}{\rho_j} \right\} \right.$$

$$+ \sum_{j=2}^{k} u_j^2 \left[\left(\sum_{\alpha=1}^{j-1} \rho_\alpha^2 \right) \left\{ \frac{3}{2} \frac{\rho_j^3}{\hat{\rho}_{j-1}^3 \hat{\rho}_j^3} + \frac{9}{2} \frac{\rho_j^2}{\hat{\rho}_{j-1}^2 \hat{\rho}_j^3} + \frac{3}{2} \frac{\rho_j^2}{\hat{\rho}_{j-1}^3 \hat{\rho}_j^2} \right\} \right.$$

$$\left. + \frac{15}{8} \frac{\hat{\rho}_{j-1}^3}{\rho_j \hat{\rho}_j^3} + \frac{39}{8} \frac{\hat{\rho}_{j-1}^2}{\hat{\rho}_j^3} + \frac{9}{8} \frac{\hat{\rho}_{j-1}^2}{\rho_j \hat{\rho}_j^2} \right]$$

$$+ \sum_{j=2}^{k} u_j \left[\left(\sum_{\alpha=1}^{j-1} \rho_\alpha^2 \right) \left\{ \frac{\rho_j^2}{\hat{\rho}_{j-1}^3 \hat{\rho}_j^2} - \frac{1}{2} \frac{\rho_j^2}{\hat{\rho}_{j-1}^2 \hat{\rho}_j^3} + \frac{1}{2} \frac{\rho_j}{\hat{\rho}_{j-1} \hat{\rho}_j^3} + \frac{5}{2} \frac{\rho_j}{\hat{\rho}_{j-1}^2 \hat{\rho}_j^2} \right\} \right.$$

$$
+\frac{3}{2}\frac{\hat{\rho}_{j-1}^2}{\rho_j\hat{\rho}_j^2} - \frac{1}{2}\frac{\hat{\rho}_{j-1}\rho_j}{\hat{\rho}_j^3} - \frac{3}{2}\frac{\rho_j^2}{\hat{\rho}_j^3} + \frac{\hat{\rho}_{j-1}}{\hat{\rho}_j^2} + \frac{3}{2}\frac{1}{\hat{\rho}_j} - \frac{1}{2}\frac{\hat{\rho}_{j-1}}{\rho_j\hat{\rho}_j}\Bigg]
$$

$$
+\frac{45}{4}\sum_{2\le p<q<r\le k} u_p u_q u_r \frac{\rho_p}{\hat{\rho}_{p-1}\hat{\rho}_p}\left(\hat{\rho}_{p-1}^2 + \sum_{\alpha=1}^{p-1}\rho_\alpha^2\right)\frac{\rho_q}{\hat{\rho}_{q-1}\hat{\rho}_q}\frac{\rho_r}{\hat{\rho}_{r-1}\hat{\rho}_r}
$$

$$
+\frac{45}{8}\sum_{2\le p<q\le k} u_p^2 u_q \frac{1}{\hat{\rho}_{p-1}^2\hat{\rho}_p^2}\left\{\left(\sum_{\alpha=1}^{p-1}\rho_\alpha^2\right)\rho_p^2 + \hat{\rho}_{p-1}^4\right\}\frac{\rho_q}{\hat{\rho}_{q-1}\hat{\rho}_q}
$$

$$
+\frac{45}{8}\sum_{2\le p<q\le k} u_p u_q^2 \frac{\rho_p}{\hat{\rho}_{p-1}\hat{\rho}_p}\left(\hat{\rho}_{p-1}^2 + \sum_{\alpha=1}^{p-1}\rho_\alpha^2\right)\frac{\rho_q^2}{\hat{\rho}_{q-1}^2\hat{\rho}_q^2}
$$

$$
+\sum_{2\le p<q\le k} u_p u_q \Bigg\{\frac{17}{4}\frac{\rho_p}{\hat{\rho}_{p-1}\hat{\rho}_p^2}\left(\hat{\rho}_{p-1}^2 + \sum_{\alpha=1}^{p-1}\rho_\alpha^2\right)\frac{\rho_q}{\hat{\rho}_{q-1}\hat{\rho}_q} + \frac{3}{4}\frac{\rho_p}{\hat{\rho}_{p-1}^2\hat{\rho}_p}
$$

$$
\times\left(\hat{\rho}_{p-1}^2 - \sum_{\alpha=1}^{p-1}\rho_\alpha^2\right)\frac{\rho_q}{\hat{\rho}_{q-1}\hat{\rho}_q} + 3\frac{\rho_p}{\hat{\rho}_{p-1}\hat{\rho}_p}\left(\hat{\rho}_{p-1}^2 + \sum_{\alpha=1}^{p-1}\rho_\alpha^2\right)\left(\frac{1}{\hat{\rho}_{q-1}^2} - \frac{1}{\hat{\rho}_q^2}\right)
$$

$$
+\frac{3}{4}\frac{1}{\hat{\rho}_{p-1}^2\hat{\rho}_p^2}\left\{\rho_p^2\left(\sum_{\alpha=1}^{p-1}\rho_\alpha^2\right) + \hat{\rho}_{p-1}^4\right\}\frac{\rho_q}{\hat{\rho}_{q-1}\hat{\rho}_q} - \frac{3}{4}\frac{\rho_q}{\hat{\rho}_{q-1}\hat{\rho}_q}
$$

$$
+\frac{1}{2}\left\{\left(\sum_{\alpha=1}^{p-1}\rho_\alpha^2\right)\frac{\rho_p(6\hat{\rho}_p - \hat{\rho}_{p-1})}{\hat{\rho}_{p-1}^2\hat{\rho}_p^2} + \frac{\hat{\rho}_{p-1}(6\hat{\rho}_p - \rho_p)}{\hat{\rho}_p^2}\right\}\frac{\rho_q}{\hat{\rho}_{q-1}\hat{\rho}_q}\Bigg\}\Bigg],
$$

$$
(3.22)\quad -\frac{1}{6}m_{(21)}m_{(1^3)}\hat{E}\left[\sum_{i=1}^{k}\frac{1}{\rho_i}H_4(y_i)d_{2i}^{(1)}\right]
$$

$$
= \frac{m_{(21)}m_{(1^3)}}{(m_{(1^2)})^3}\Bigg[\frac{5}{4}\sum_{j=2}^{k}u_j^3\frac{1}{\hat{\rho}_j^3}\left\{\left(\sum_{\alpha=1}^{j-1}\rho_\alpha^2\right)\frac{\rho_j^3}{\hat{\rho}_{j-1}^3} + \frac{\hat{\rho}_{j-1}^3}{\rho_j}\right\}
$$

$$
+\sum_{j=2}^{k}u_j^2\left\{\left(\sum_{\alpha=1}^{j-1}\rho_\alpha^2\right)\left\{\frac{\rho_j^3}{\hat{\rho}_{j-1}^3\hat{\rho}_j^3} + 2\frac{\rho_j^2}{\hat{\rho}_{j-1}^2\hat{\rho}_j^3}\right\} + \frac{5}{2}\frac{\hat{\rho}_{j-1}^3}{\rho_j\hat{\rho}_j^3} - \frac{3}{2}\frac{\hat{\rho}_{j-1}^2}{\rho_j\hat{\rho}_j^2} + \frac{7}{2}\frac{\hat{\rho}_{j-1}^2}{\hat{\rho}_j^3}\right\}
$$

$$
+\sum_{j=2}^{k}u_j\left\{\frac{5}{4}\frac{\hat{\rho}_{j-1}^3}{\rho_j\hat{\rho}_j^3} - \frac{3}{2}\frac{\hat{\rho}_{j-1}^2}{\rho_j\hat{\rho}_j^2} + \frac{9}{4}\frac{\rho_j\hat{\rho}_{j-1}}{\hat{\rho}_j^3} + \frac{7}{2}\frac{\hat{\rho}_{j-1}^2}{\hat{\rho}_j^3} - \frac{5}{2}\frac{\hat{\rho}_{j-1}}{\hat{\rho}_j^2} + \frac{1}{4}\frac{\hat{\rho}_{j-1}}{\rho_j\hat{\rho}_j}\right\}
$$

$$
+\frac{15}{2}\sum_{2\le p<q<r\le k} u_p u_q u_r \frac{\rho_p}{\hat{\rho}_{p-1}\hat{\rho}_p}\left(\hat{\rho}_{p-1}^2 + \sum_{\alpha=1}^{p-1}\rho_\alpha^2\right)\frac{\rho_q}{\hat{\rho}_{q-1}\hat{\rho}_q}\frac{\rho_r}{\hat{\rho}_{r-1}\hat{\rho}_r}
$$

$$
+\frac{15}{4}\sum_{2\le p<q\le k} u_p^2 u_q \frac{1}{\hat{\rho}_{p-1}^2\hat{\rho}_p^2}\left\{\rho_p^2\left(\sum_{\alpha=1}^{p-1}\rho_\alpha^2\right) + \hat{\rho}_{p-1}^4\right\}\frac{\rho_q}{\hat{\rho}_{q-1}\hat{\rho}_q}
$$

$$
+\frac{15}{4}\sum_{2\le p<q\le k} u_p u_q^2 \frac{\rho_p}{\hat{\rho}_{p-1}\hat{\rho}_p}\left(\hat{\rho}_{p-1}^2 + \sum_{\alpha=1}^{p-1}\rho_\alpha^2\right)\left(\frac{\rho_q}{\hat{\rho}_{q-1}\hat{\rho}_q}\right)^2
$$

$$
\begin{aligned}
&+\sum_{2\leq p<q\leq k} u_p u_q \left\{ 4\frac{1}{\hat{\rho}_{p-1}^2\hat{\rho}_p^2}\left\{ \rho_p^2\left(\sum_{\alpha=1}^{p-1}\rho_\alpha^2\right)+\hat{\rho}_{p-1}^4\right\}\frac{\rho_q}{\hat{\rho}_{q-1}\hat{\rho}_q}\right. \\
&+\frac{\rho_p}{\hat{\rho}_{p-1}\hat{\rho}_p}\left(\hat{\rho}_{p-1}^2+\sum_{\alpha=1}^{p-1}\rho_\alpha^2\right)\left(\frac{1}{\hat{\rho}_{q-1}^2}-\frac{1}{\hat{\rho}_q^2}\right) \\
&+5\frac{\rho_p}{\hat{\rho}_{p-1}\hat{\rho}_p^2}\left(\hat{\rho}_{p-1}^2+\sum_{\alpha=1}^{p-1}\rho_\alpha^2\right)\frac{\rho_q}{\hat{\rho}_{q-1}\hat{\rho}_q} \\
&\left.+3\frac{\rho_p}{\hat{\rho}_{p-1}^2\hat{\rho}_p}\left(\hat{\rho}_{p-1}^2-\sum_{\alpha=1}^{p-1}\rho_\alpha^2\right)\frac{\rho_q}{\hat{\rho}_{q-1}\hat{\rho}_q}-3\frac{\rho_q}{\hat{\rho}_{q-1}\hat{\rho}_q}\right\}\Bigg],
\end{aligned}
$$

$$
\begin{aligned}
&\frac{1}{72}(m_{(1^3)})^2\hat{E}\left[\sum_{i=1}^{k}\frac{1}{\rho_i}H_6(y_i)\right] \\
&=\frac{(m_{(1^3)})^2}{(m_{(1^2)})^3}\left[\frac{5}{24}\sum_{j=2}^{k}u_j^3\frac{1}{\hat{\rho}_j^3}\left\{\left(\sum_{\alpha=1}^{j-1}\rho_\alpha^2\right)\frac{\rho_j^3}{\hat{\rho}_{j-1}^3}+\frac{\hat{\rho}_{j-1}^3}{\rho_j}\right\}\right. \\
&+\frac{5}{4}\sum_{2\leq p<q<r\leq k}u_p u_q u_r\frac{\rho_p}{\hat{\rho}_{p-1}\hat{\rho}_p}\left(\hat{\rho}_{p-1}^2+\sum_{\alpha=1}^{p-1}\rho_\alpha^2\right)\frac{\rho_q}{\hat{\rho}_{q-1}\hat{\rho}_q}\frac{\rho_r}{\hat{\rho}_{r-1}\hat{\rho}_r} \\
&+\frac{5}{8}\sum_{2\leq p<q\leq k}u_p^2 u_q\frac{1}{\hat{\rho}_{p-1}^2\hat{\rho}_p^2}\left\{\rho_p^2\left(\sum_{\alpha=1}^{p-1}\rho_\alpha^2\right)+\hat{\rho}_{p-1}^4\right\}\frac{\rho_q}{\hat{\rho}_{q-1}\hat{\rho}_q} \\
&\left.+\frac{5}{8}\sum_{2\leq p<q\leq k}u_p u_q^2\frac{\rho_p}{\hat{\rho}_{p-1}\hat{\rho}_p}\left(\hat{\rho}_{p-1}^2+\sum_{\alpha=1}^{p-1}\rho_\alpha^2\right)\frac{\rho_q^2}{\hat{\rho}_{q-1}^2\hat{\rho}_q^2}\right],
\end{aligned}
\tag{3.23}
$$

$$
\begin{aligned}
&\frac{1}{72}(m_{(1^3)})^2\hat{E}\left[\sum_{i\neq j}\frac{1}{\sqrt{\rho_i\,\rho_j}}H_3(y_i)H_3(y_j)\right] \\
&=\frac{(m_{(1^3)})^2}{(m_{(1^2)})^3}\left[-\frac{5}{24}\sum_{j=2}^{k}u_j^3\frac{1}{\hat{\rho}_j^3}\left\{\left(\sum_{\alpha=1}^{j-1}\rho_\alpha^2\right)\frac{\rho_j^3}{\hat{\rho}_{j-1}^3}+2\rho_j\hat{\rho}_{j-1}-\frac{\rho_j^3}{\hat{\rho}_{j-1}}\right\}\right. \\
&+\sum_{2\leq p<q<r\leq k}u_p u_q u_r\left\{-\frac{5}{4}\frac{\rho_p}{\hat{\rho}_{p-1}\hat{\rho}_p}\left(\hat{\rho}_{p-1}^2+\sum_{\alpha=1}^{p-1}\rho_\alpha^2\right)\frac{\rho_q}{\hat{\rho}_{q-1}\hat{\rho}_q}\frac{\rho_r}{\hat{\rho}_{r-1}\hat{\rho}_r}\right. \\
&\qquad\qquad\qquad\qquad\left.+\frac{1}{4}\frac{\rho_r}{\hat{\rho}_{r-1}\hat{\rho}_r}\right\} \\
&+\sum_{2\leq p<q\leq k}u_p^2 u_q\left\{-\frac{1}{4}\frac{\rho_p\hat{\rho}_{p-1}}{\hat{\rho}_p^2}\left(\frac{\hat{\rho}_{p-1}}{\rho_p}-3+\frac{\rho_p}{\hat{\rho}_{p-1}}\right)\frac{\rho_q}{\hat{\rho}_{q-1}\hat{\rho}_q}\right. \\
&\qquad\qquad\left.+\frac{5}{8}\frac{\rho_p^2}{\hat{\rho}_{p-1}^2\hat{\rho}_p^2}\left(\hat{\rho}_{p-1}^2-\sum_{\alpha=1}^{p-1}\rho_\alpha^2\right)\frac{\rho_q}{\hat{\rho}_{q-1}\hat{\rho}_q}\right\}
\end{aligned}
\tag{3.24}
$$

$$+\sum_{2\le p<q\le k} u_p u_q^2 \Bigg\{ \frac{1}{4}\frac{(\rho_q-\hat\rho_{q-1})}{\hat\rho_{q-1}\hat\rho_q} - \frac{5}{8}\frac{\rho_p}{\hat\rho_{p-1}\hat\rho_p}\left(\hat\rho_{p-1}^2+\sum_{\alpha=1}^{p-1}\rho_\alpha^2\right)\frac{\rho_q^2}{\hat\rho_{q-1}^2\hat\rho_q^2}\Bigg\}\Bigg],$$

$$(3.25)\ \frac{1}{2}(m_{(21)})^2\hat E\left[\sum_{i\ne j}\frac{1}{\sqrt{\rho_i\,\rho_j}}H_1(y_i)H_1(y_j)d_{2i}^{(1)}d_{2j}^{(1)}\right]$$

$$= \frac{(m_{(21)})^2}{(m_{(1^2)})^3}\Bigg[\sum_{j=2}^{k} u_j^3\Bigg\{-\frac{15}{8}\left(\sum_{\alpha=1}^{j-1}\rho_\alpha^2\right)\frac{\rho_j^3}{\hat\rho_{j-1}^3\hat\rho_j^3}+\frac{15}{8}\frac{\rho_j^3}{\hat\rho_{j-1}\hat\rho_j^3}-\frac{15}{4}\frac{\rho_j\hat\rho_{j-1}}{\hat\rho_j^3}\Bigg\}$$

$$+\sum_{j=2}^{k} u_j^2\Bigg\{-\frac{3}{2}\left(\sum_{\alpha=1}^{j-1}\rho_\alpha^2\right)\Bigg\{\frac{\rho_j^3}{\hat\rho_{j-1}^3\hat\rho_j^3}+3\frac{\rho_j^2}{\hat\rho_{j-1}^2\hat\rho_j^3}+\frac{\rho_j^2}{\hat\rho_{j-1}^3\hat\rho_j^2}\Bigg\}$$

$$+\frac{3}{2}\frac{\rho_j^3}{\hat\rho_{j-1}\hat\rho_j^3}+3\frac{\rho_j\hat\rho_{j-1}}{\hat\rho_j^3}+3\frac{\rho_j^2}{\hat\rho_j^3}-\frac{3}{2}\frac{\hat\rho_{j-1}^2}{\hat\rho_j^3}+\frac{3}{2}\frac{\rho_j^2}{\hat\rho_{j-1}\hat\rho_j^2}\Bigg\}$$

$$+\sum_{j=2}^{k} u_j\Bigg\{\left(\sum_{\alpha=1}^{j-1}\rho_\alpha^2\right)\Bigg\{\frac{1}{2}\frac{\rho_j^3}{\hat\rho_{j-1}^3\hat\rho_j^3}-\frac{\rho_j^2}{\hat\rho_{j-1}^3\hat\rho_j^2}-\frac{1}{2}\frac{\rho_j}{\hat\rho_{j-1}\hat\rho_j^3}-\frac{5}{2}\frac{\rho_j}{\hat\rho_{j-1}^2\hat\rho_j^2}\Bigg\}$$

$$-\frac{3}{2}\frac{\rho_j\hat\rho_{j-1}}{\hat\rho_j^3}-\frac{1}{2}\frac{\rho_j^2}{\hat\rho_j^3}+\frac{\rho_j^2}{\hat\rho_{j-1}\hat\rho_j^2}+\frac{5}{2}\frac{\rho_j}{\hat\rho_j^2}\Bigg\}$$

$$+\sum_{2\le p<q<r\le k} u_p u_q u_r\Bigg\{-\frac{45}{4}\frac{\rho_p}{\hat\rho_{p-1}\hat\rho_p}\left(\hat\rho_{p-1}^2+\sum_{\alpha=1}^{p-1}\rho_\alpha^2\right)\frac{\rho_q}{\hat\rho_{q-1}\hat\rho_q}\frac{\rho_r}{\hat\rho_{r-1}\hat\rho_r}$$

$$+\frac{9}{4}\frac{\rho_r}{\hat\rho_{r-1}\hat\rho_r}\Bigg\}$$

$$+\sum_{2\le p<q\le k} u_p^2 u_q\Bigg\{-\frac{9}{4}\frac{\rho_p\hat\rho_{p-1}}{\hat\rho_p}\left(\frac{\hat\rho_{p-1}}{\rho_p}-3+\frac{\rho_p}{\hat\rho_{p-1}}\right)\frac{\rho_q}{\hat\rho_{q-1}\hat\rho_q}$$

$$+\frac{45}{8}\left(\frac{\rho_p}{\hat\rho_{p-1}\hat\rho_p}\right)^2\left(\hat\rho_{p-1}^2-\sum_{\alpha=1}^{p-1}\rho_\alpha^2\right)\frac{\rho_q}{\hat\rho_{q-1}\hat\rho_q}\Bigg\}$$

$$+\sum_{2\le p<q\le k} u_p u_q^2\Bigg\{-\frac{45}{8}\frac{\rho_p}{\hat\rho_{p-1}\hat\rho_p}\left(\hat\rho_{p-1}^2+\sum_{\alpha=1}^{p-1}\rho_\alpha^2\right)\left(\frac{\rho_q}{\hat\rho_{q-1}\hat\rho_q}\right)^2+\frac{9}{4}\frac{(\rho_q-\hat\rho_{q-1})}{\hat\rho_{q-1}\hat\rho_q}\Bigg\}$$

$$+\sum_{2\le p<q\le k} u_p u_q\Bigg\{-\frac{3}{4}\frac{1}{\hat\rho_{p-1}^2\hat\rho_p^2}\left(\rho_p^2\left(\sum_{\alpha=1}^{p-1}\rho_\alpha^2\right)+\hat\rho_{p-1}^4\right)\frac{\rho_q}{\hat\rho_{q-1}\hat\rho_q}$$

$$-\frac{1}{2}\Bigg\{\frac{\rho_p}{\hat\rho_{p-1}^2\hat\rho_p^2}(6\hat\rho_p-\hat\rho_{p-1})\left(\sum_{\alpha=1}^{p-1}\rho_\alpha^2\right)+\frac{\hat\rho_{p-1}}{\hat\rho_p^2}(6\hat\rho_p-\rho_p)\Bigg\}\frac{\rho_q}{\hat\rho_{q-1}\hat\rho_q}$$

$$-3\frac{\rho_p}{\hat\rho_{p-1}\hat\rho_p}\left(\hat\rho_{p-1}^2+\sum_{\alpha=1}^{p-1}\rho_\alpha^2\right)\left(\frac{1}{\hat\rho_{q-1}^2}-\frac{1}{\hat\rho_q^2}\right)$$

$$+\left(-\frac{3}{4}\frac{\rho_p}{\hat{\rho}_{p-1}^2\hat{\rho}_p}+\frac{17}{4}\frac{\rho_p}{\hat{\rho}_{p-1}\hat{\rho}_p^2}\right)\left(\hat{\rho}_{p-1}^2-\sum_{\alpha=1}^{p-1}\rho_\alpha^2\right)\frac{\rho_q}{\hat{\rho}_{q-1}\hat{\rho}_q}$$

$$\left.\left.+\frac{21}{4}\frac{\rho_q}{\hat{\rho}_{q-1}\hat{\rho}_q}-\frac{17}{2}\frac{\rho_p\hat{\rho}_{p-1}}{\hat{\rho}_p^2}\frac{\rho_q}{\hat{\rho}_{q-1}\hat{\rho}_q}\right\}\right],$$

$$(3.26)\ -\frac{1}{6}m_{(21)}m_{(1^3)}\hat{E}\left[\sum_{i\neq j}\frac{1}{\sqrt{\rho_i\,\rho_j}}H_1(y_i)H_3(y_j)d_{2i}^{(1)}\right]$$

$$=\frac{m_{(21)}m_{(1^3)}}{(m_{(1^2)})^3}\left[-\frac{5}{4}\sum_{j=2}^{k}u_j^3\left\{\left(\sum_{\alpha=1}^{j-1}\rho_\alpha^2\right)\frac{\rho_j^3}{\hat{\rho}_{j-1}^3\hat{\rho}_j^3}-\frac{\rho_j^3}{\hat{\rho}_{j-1}\hat{\rho}_j^3}+2\frac{\rho_j\hat{\rho}_{j-1}}{\hat{\rho}_j^3}\right\}\right.$$

$$+\sum_{j=2}^{k}u_j^2\left\{-\left(\sum_{\alpha=1}^{j-1}\rho_\alpha^2\right)\left(\frac{\rho_j^3}{\hat{\rho}_{j-1}^3\hat{\rho}_j^3}+2\frac{\rho_j^2}{\hat{\rho}_{j-1}^2\hat{\rho}_j^3}\right)+\frac{\rho_j^3}{\hat{\rho}_{j-1}\hat{\rho}_j^3}\right.$$

$$\left.+\frac{1}{4}\frac{\rho_j\hat{\rho}_{j-1}}{\hat{\rho}_j^3}-\frac{1}{2}\frac{\hat{\rho}_{j-1}^2}{\hat{\rho}_j^3}+\frac{3}{4}\frac{\rho_j^2}{\hat{\rho}_j^3}+\frac{3}{4}\frac{\rho_j}{\hat{\rho}_j^2}\right\}$$

$$+\sum_{j=2}^{k}u_j\left\{-\frac{1}{2}\frac{\rho_j\hat{\rho}_{j-1}}{\hat{\rho}_j^3}-\frac{1}{2}\frac{\hat{\rho}_{j-1}^2}{\hat{\rho}_j^3}+\frac{1}{2}\frac{\hat{\rho}_{j-1}}{\hat{\rho}_j^2}\right\}$$

$$+\sum_{2\le p<q<r\le k}u_pu_qu_r\left\{-\frac{15}{2}\frac{\rho_p}{\hat{\rho}_{p-1}\hat{\rho}_p}\left(\hat{\rho}_{p-1}^2+\sum_{\alpha=1}^{p-1}\rho_\alpha^2\right)\frac{\rho_q}{\hat{\rho}_{q-1}\hat{\rho}_q}\frac{\rho_r}{\hat{\rho}_{r-1}\hat{\rho}_r}\right.$$

$$\left.+\frac{3}{2}\frac{\rho_r}{\hat{\rho}_{r-1}\hat{\rho}_r}\right\}$$

$$+\sum_{2\le p<q\le k}u_p^2u_q\left\{-\frac{3}{2}\frac{\rho_p\hat{\rho}_{p-1}}{\hat{\rho}_p^2}\left(\frac{\hat{\rho}_{p-1}}{\rho_p}-3+\frac{\rho_p}{\hat{\rho}_{p-1}}\right)\frac{\rho_q}{\hat{\rho}_{q-1}\hat{\rho}_q}\right.$$

$$\left.+\frac{15}{4}\left(\frac{\rho_p}{\hat{\rho}_{p-1}\hat{\rho}_p}\right)^2\left(\hat{\rho}_{p-1}^2-\sum_{\alpha=1}^{p-1}\rho_\alpha^2\right)\frac{\rho_q}{\hat{\rho}_{q-1}\hat{\rho}_q}\right\}$$

$$+\sum_{2\le p<q\le k}u_pu_q^2\left\{\frac{3}{2}\frac{(\rho_q-\hat{\rho}_{q-1})}{\hat{\rho}_{q-1}\hat{\rho}_q}-\frac{15}{4}\frac{\rho_p}{\hat{\rho}_{p-1}\hat{\rho}_p}\left(\hat{\rho}_{p-1}^2+\sum_{\alpha=1}^{p-1}\rho_\alpha^2\right)\left(\frac{\rho_q}{\hat{\rho}_{q-1}\hat{\rho}_q}\right)^2\right\}$$

$$+\sum_{2\le p<q\le k}u_pu_q\left\{-\frac{\rho_p}{\hat{\rho}_{p-1}\hat{\rho}_p}\left(\hat{\rho}_{p-1}^2+\sum_{\alpha=1}^{p-1}\rho_\alpha^2\right)\left\{\frac{3}{2}\left(\frac{\rho_q}{\hat{\rho}_{q-1}\hat{\rho}_q}\right)^2+\frac{5}{2}\frac{\rho_q}{\hat{\rho}_{q-1}\hat{\rho}_q^2}\right\}\right.$$

$$+\left(\hat{\rho}_{p-1}^2-\sum_{\alpha=1}^{p-1}\rho_\alpha^2\right)\left(\frac{1}{\hat{\rho}_{p-1}^2}-\frac{1}{\hat{\rho}_p^2}\right)\frac{\rho_q}{\hat{\rho}_{q-1}\hat{\rho}_q}$$

$$-\frac{1}{2}\frac{\rho_p}{\hat{\rho}_{p-1}\hat{\rho}_p}\left(\hat{\rho}_{p-1}^2-\sum_{\alpha=1}^{p-1}\rho_\alpha^2\right)\frac{\rho_q}{\hat{\rho}_{q-1}^2\hat{\rho}_q}$$

$$+\frac{1}{2}\frac{\rho_q}{\hat{\rho}_{q-1}\hat{\rho}_q}-2\frac{\rho_p\hat{\rho}_{p-1}}{\hat{\rho}_p^2}\frac{\rho_q}{\hat{\rho}_{q-1}\hat{\rho}_q}+\frac{\rho_p\hat{\rho}_{p-1}}{\hat{\rho}_p}\frac{\rho_q}{\hat{\rho}_{q-1}^2\hat{\rho}_q}\Bigg\}\Bigg]$$

$$+O\left(\frac{1}{\sqrt{n}}\right).$$

Combining (3.12), (3.17), (3.18), (3.19) and (3.20) for the fourth moments and using Bartlett identities, we have

$$\text{(3.27)}\qquad \frac{1}{8}\sum_{j=2}^{k}(c_j-1)\frac{1}{\hat{\rho}_j}\left(\frac{\hat{\rho}_{j-1}}{\rho_j}+1+\frac{\rho_j}{\hat{\rho}_{j-1}}\right)\frac{(m_{(2^2)}-2m_{(21^2)}-m_{(1^4)})}{(m_{(1^2)})^2}.$$

Similary combining the third moments, we have after some lengthy algebra

$$\text{(3.28)}\ \frac{1}{24}\sum_{j=2}^{k}(c_j-1)\frac{1}{\hat{\rho}_j}\left(\frac{\hat{\rho}_{j-1}}{\rho_j}+1+\frac{\rho_j}{\hat{\rho}_{j-1}}\right)\frac{(m_{(3)}m_{(21)}-5m_{(3)}m_{(1^3)}-8m_{(21)}m_{(1^3)})}{(m_{(1^2)})^3},$$

which gives joint moment generating function (2.8).

References

Barndorff-Nielsen, O. E. and Cox, D. R. (1994). *Inference and Asymptotics*, Chapman and Hall, London.

Bartlett, M. S. (1937). Properties of sufficiency and statistical tests, *Proc. Royal Soc. London Ser. A*, **160**, 268–282.

Bickel, P. J. and Ghosh, J. K. (1990). A decomposition for the likelihood ratio statistic and the Bartlett correction—A Bayesian argument, *Ann. Statist.*, **18**, 1070–1090.

Cordeiro, G. M. (1987). On the corrections to the likelihood ratio statistics, *Biometrika*, **74**, 265–274.

Harris, P. (1986). A note on Bartlett adjustments to likelihood ratio tests, *Biometrika*, **73**, 735–737.

Hayakawa, T. (1977). The likelihood ratio criterion and the asymptotic expansion of its distribution, *Ann. Inst. Statist. Math.*, **29**, 359–378 (Correction: ibid. (1987). **39**, P. 681).

Hayakawa, T. (1993). Test of homogeneity of parameters, *Statistical Science and Data Analysis* (eds. K. Matusita, M. L. Puri and T. Hayakawa), 337–343, VSP Zeist.

Hayakawa, T. (1994). Test of homogeneity of multiple parameters, *J. Statist. Plann. Inference*, **38**, 351–357.

Hayakawa, T. (2001). Rao's statistic for homogeneity of multiple parameters, *J. Statist. Plann. Inference*, Special Issue for Rao's Score Statistic (eds. A. K. Bera and R. Mukerjee), **97**, 101–111.

Hayakawa, T. and Doi, M. (1999). Modified Wald statistic for homogeneity of multiple parameters, *Comm. Statist. Theory. Methods*, **28**, 755–771.

Lawley, D. N. (1956). A general method of approximating to the distribution of likelihood ratio criteria, *Biometrika*, **43**, 295–303.

Ogawa, J. (1949). On the independence of bilinear form and quadratic forms of a random sample from a normal population, *Ann. Inst. Statist. Math.*, **1**, 83–108.

Takemura, A. and Kuriki, S. (1996). A proof of independent Bartlett correctability of nested likelihood ratio tests, *Ann. Inst. Statist. Math.*, **48**, 603–620.

Ann. Inst. Statist. Math.
Vol. 54, No. 4, 934–944 (2002)

MULTIVARIATE PERCENTILE TESTS FOR INCOMPLETE DATA

HYO-IL PARK

Department of Statistics, Chong-ju University, Chong-ju, Choong-book, 360-764, Korea,
e-mail: hipark@chongju.ac.kr

(Received May 23, 2001; revised October 19, 2001)

Abstract. In this paper, we consider the percentile test procedures for multivariate and right censored data. Because of the involvement of censoring distribution into the distribution of the proposed test statistic, we study the asymptotic normality using the estimated covariance matrix. Finally, we derive the asymptotic relative efficiency and illustrate our procedures with an example.

Key words and phrases: Asymptotic relative efficiency, noncentrality parameter, Pitman translation parameter, two sample problem.

1. Introduction

Median tests as nonparametric procedures for two sample problem are well known and useful for detecting location translations. Basically there are two kinds of median tests in the univariate case. One is the control median test (e.g. Mathisen (1943)) and the other, the combined median test (e.g. Mood (1950)). The distinction between two kinds of median tests is as follows: the control median test uses a median from control sample whereas the combined median test uses a median from combined sample. From now on, we call simply median test for the combined median test. Two kinds of median tests have been modified or extended to the various directions. As a particular modification of the control median test, Gastwirth (1968) proposed the first median test in order to improve its performance as a two-sided test, which permits the experimenter to reach a decision early. Therefore the first median test would be useful in case of the life trial situation. Also Gastwirth discussed the application of the curtailed sampling to the first median test for the early decision in the same paper. For more detailed discussion of the curtailed sampling, we may refer to Alling (1963). Hettmansperger (1973) further considered a conservative test based on the first median test statistic to cover the Behrens-Fisher problem. Chatterjee and Sen (1964), Hettmansperger (1984) and Babu and Rao (1988) considered extensions of the median test to multivariate data. Recently, Park and Desu (1999) extended the control median test to multivariate data. Brookmeyer and Crowley (1982) modified the median test for right censored data. Gastwirth and Wang (1988) proposed the control median test for right censored data. Also Park and Desu (1998) considered an extension of the control median test to multivariate and right censored data. Therefore one may expect the advent of a median test procedure for multivariate and right censored data. However, for the case of right censoring data, it is not rare that one may not obtain a sample median or medians because of heavy censoring for larger observations or early termination of experiments. In this case, it is impossible to compare treatment effects with sample medians. In order to circumvent this stalemate, we consider a percentile test which uses the corresponding quantlie points instead of using medians. Also Gastwirth and Wang (1988) proposed a control percentile test for the consideration of the efficiency in the univariate case. In the next section, we propose a multivariate percentile test for right censored data. We deal with the large sample approximation in Section 3. Finally, we consider the asymptotic relative efficiency and show an example for illustration of our proposed test procedure.

2. Multivariate percentile test for right censored data

Let $\boldsymbol{X}$ and $\boldsymbol{Y}$ be two independent q-variate random vectors with continuous distribution functions F and G, respectively. It is of our concern to test the hypothesis $H_0 : F = G$. Since the location translation alternatives are of interest, we assume that in general,

$$G(\boldsymbol{x}) = F(\boldsymbol{x} - \boldsymbol{\Delta}) \text{ for all } \boldsymbol{x} \in \boldsymbol{R}^q \text{ and for some } \boldsymbol{\Delta} \in \boldsymbol{R}^q. \tag{2.1}$$

In view of this assumption, the null hypothesis can be restated as $H_0 : \boldsymbol{\Delta} = \boldsymbol{0}$. Usually, a random sample $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ of $\boldsymbol{X}$ and an independent random sample $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_m$ of $\boldsymbol{Y}$ are observed and tests are performed based on these samples. However in some experiments, one can only observe $\{(\boldsymbol{V}_i, \boldsymbol{\delta}_i), i = 1, \ldots, n\}$ and $\{(\boldsymbol{W}_j, \boldsymbol{\tau}_j), j = 1, \ldots, m\}$, where $V_{ki} = \min(X_{ki}, C_{ki})$, $\delta_{ki} = I(V_{ki} = X_{ki})$, $W_{kj} = \min(Y_{kj}, D_{kj})$ and $\tau_{kj} = I(W_{kj} = Y_{kj})$ for $i = 1, \ldots, n$, $j = 1, \ldots, m$ and $k = 1, \ldots, q$. $I(\cdot)$ is the indicator function. It is assumed that $\boldsymbol{C}_1, \ldots, \boldsymbol{C}_n$ is a censoring random sample with distribution function H_F and $\boldsymbol{D}_1, \ldots, \boldsymbol{D}_m$ is an independent censoring random sample with distribution function H_G. Furthermore, it is assumed that $\boldsymbol{X}$'s, $\boldsymbol{Y}$'s, $\boldsymbol{C}$'s, and $\boldsymbol{D}$'s are all independent each other. For each k, $k = 1, \ldots, q$, we denote F_k and G_k as the marginal distribution functions of F and G and $\hat{F}_{kn}$ and $\hat{G}_{km}$, as the corresponding Kaplan-Meier estimates. Also for each k, let $H_{kN} = (n/N)F_k + (m/N)G_k$ and $\hat{H}_{kN} = (n/N)\hat{F}_{kn} + (m/N)\hat{G}_{km}$ with $N = m + n$. Finally, for each p with $0 < p < 1$ and for each k, let $\xi^*_{kN}(p)$ be a p-th quantile of H_{kN} and $\hat{\xi}^*_{kN}(p)$, the corresponding p-th sample quantile of $\hat{H}_{kN}$. Then for any consistent estimate $\hat{\Sigma}_N(p)$ of the limiting null covariance matrix $\Sigma_0(p)$ of

$$\sqrt{n}(\hat{F}_{1n}(\hat{\xi}^*_{1N}(p)), \ldots, \hat{F}_{qn}(\hat{\xi}^*_{qN}(p))),$$

assuming that the inverse $\hat{\Sigma}_N^{-1}(p)$ of $\hat{\Sigma}_N(p)$ exists, we propose a q-variate p-th percentile test statistic $\boldsymbol{M}_N$ as follows:

$$\boldsymbol{M}_N = n \begin{pmatrix} \hat{F}_{1n}(\hat{\xi}^*_{1N}(p)) - p \\ \cdots \\ \hat{F}_{qn}(\hat{\xi}^*_{qN}(p)) - p \end{pmatrix}^T \hat{\Sigma}_N^{-1}(p) \begin{pmatrix} \hat{F}_{1n}(\hat{\xi}^*_{1N}(p)) - p \\ \cdots \\ \hat{F}_{qn}(\hat{\xi}^*_{qN}(p)) - p \end{pmatrix},$$

where T means the transpose of a matrix or a vector. We will identify $\Sigma_0(p)$ and $\hat{\Sigma}_N(p)$ later. Then an α-level test of $H_0 : F = G$ against $H_1 : F \neq G$ is to

$$\text{“reject } H_0 \quad \text{if} \quad \boldsymbol{M}_N \geq C(\alpha)\text{”}.$$

The constant $C(\alpha)$ is chosen so that the size of the test is α. Since the exact null distribution of $\boldsymbol{M}_N$ depends on F, G, H_F and H_G in a complicated manner, it is natural to consider the large sample approximation.

3. Limiting distribution of M_N

For the derivation of the limiting null distribution of M_N, we introduce some more notations about distribution and subdistribution functions. In the following, for each k, H_{F_k} and H_{G_k} denote the marginal distribution functions for H_F and H_G, respectively. For each k, $k = 1, \ldots, q$, let

$$\begin{aligned}
&F_k^*(u) = P(V_{ki} \le u, \delta_{ki} = 1), \qquad G_k^*(u) = P(W_{kj} \le u, \tau_{kj} = 1),\\
&S_{F_k}(u) = P(V_{ki} > u) = (1 - F_k(u))(1 - H_{F_k}(u)),\\
&S_{G_k}(u) = P(W_{kj} > u) = (1 - G_k(u))(1 - H_{G_k}(u)).
\end{aligned}$$

Also for each $1 \le k \ne l \le q$, let

$$\begin{aligned}
&F_{kl}^*(u,v) = P(V_{ki} \le u, V_{li} \le v, \delta_{ki} = \delta_{li} = 1),\\
&G_{kl}^*(u,v) = P(W_{kj} \le u, W_{lj} \le v, \tau_{kj} = \tau_{lj} = 1),\\
&S_{F_{kl}}(u,v) = P(V_{ki} > u, V_{li} > v), \qquad S_{G_{kl}}(u,v) = P(W_{kj} > u, W_{lj} > v),\\
&N_{F_{kl}}(u,v) = P(V_{ki} \le u, V_{li} \ge v, \delta_{ki} = 1), \qquad M_{F_{kl}}(u,v) = P(V_{ki} \ge u, V_{li} \le v, \delta_{li} = 1),\\
&N_{G_{kl}}(u,v) = P(W_{kj} \le u, W_{lj} \ge v, \tau_{kj} = 1),\\
&M_{G_{kl}}(u,v) = P(W_{kj} \ge u, W_{lj} \le v, \tau_{lj} = 1).
\end{aligned}$$

Also we need the following assumptions:

ASSUMPTION 1. As $N \to \infty$, $n/m \to \lambda \in (0, \infty)$.

ASSUMPTION 2. For each k, $k = 1, \ldots, q$, F_k and G_k are continuous and twice differentiable at $\xi_{kN}^*(p)$ with $f_k(\xi_{kN}^*(p)) > 0$ and $g_k(\xi_{kN}^*(p)) > 0$ for each N, where f_k and g_k are the respective densities.

Now we state Bahadur representation of the Kaplan-Meier estimate, which is due to Lo and Singh (1985).

LEMMA 1. *For each k, $k = 1, \ldots, q$ and for each $0 < p < 1$, with Assumptions 1 and 2, we have with probability one (w.p.1), as $N \to \infty$,*

$$\begin{aligned}
&\hat{F}_{kn}(\xi_{kN}^*(p)) - F_k(\xi_{kN}^*(p))\\
&\quad = \frac{1}{n}\sum_{i=1}^{n} \beta(V_{ki}, \delta_{ki}, \xi_{kN}^*(p)) + O(N^{-3/4}(\log N)^{3/4}) \quad \textit{and}\\
&\hat{G}_{km}(\xi_{kN}^*(p)) - G_k(\xi_{kN}^*(p)) = \frac{1}{m}\sum_{j=1}^{m} \gamma(W_{kj}, \tau_{kj}, \xi_{kN}^*(p)) + O(N^{-3/4}(\log N)^{3/4}),
\end{aligned}$$

where

$$\beta(V_{ki}, \delta_{ki}, t) = (1 - F_k(t))\left\{\frac{I(V_{ki} \le t, \delta_{ki} = 1)}{S_{F_k}(V_{ki})} - \int_0^t \frac{I(V_{ki} \ge u)dF_k^*(u)}{S_{F_k}^2(u)}\right\} \quad \textit{and}$$

$$\gamma(W_{kj}, \tau_{kj}, t) = (1 - G_k(t))\left\{\frac{I(W_{kj} \le t, \tau_{kj} = 1)}{S_{G_k}(W_{kj})} - \int_0^t \frac{I(W_{kj} \ge v)dG_k^*(v)}{S_{G_k}^2(v)}\right\}.$$

LEMMA 2. *Under Assumption* 2, *for each* k, $k = 1, \ldots, q$ *and for each* $0 < p < 1$, *w.p.*1, *as* $N \to \infty$,

$$\hat{\xi}^*_{kN}(p) - \xi^*_{kN}(p) = \frac{p - \hat{H}_{kN}(\xi^*_{kN}(p))}{h_{kN}(\xi^*_{kN}(p))} + O(N^{-3/4}(\log N)^{3/4}),$$

where h_{kN} *is the density of* H_{kN}.

PROOF. First of all, we note that

$$\hat{H}_{kN}(t) - H_{kN}(t) = \frac{n}{N}[\hat{F}_{kn}(t) - F_k(t)] + \frac{m}{N}[\hat{G}_{km}(t) - G_k(t)]. \tag{3.1}$$

Then from Lemma 3 in Lo and Singh (1985), we see that w.p.1, as $N \to \infty$,

$$\sup_{0<p<1} |\hat{H}^{-1}_{kN}(p) - H^{-1}_{kN}(p)| = O(N^{-1/2}(\log N)^{1/2}). \tag{3.2}$$

Thus from Taylor's expansion around $\xi^*_{kN}(p)$ and (3.2), we have w.p.1, as $N \to \infty$,

$$H_{kN}(\hat{\xi}^*_{kN}(p)) - H_{kN}(\xi^*_{kN}(p)) = h_{kN}(\xi^*_{kN}(p))(\hat{\xi}^*_{kN}(p) - \xi^*_{kN}(p)) + O(N^{-1}\log N). \tag{3.3}$$

Also from Cheng (1984) with (3.1), we have w.p.1, as $N \to \infty$,

$$\begin{aligned} &\hat{H}_{kN}(\hat{\xi}^*_{kN}(p)) - \hat{H}_{kN}(\xi^*_{kN}(p)) - H_{kN}(\hat{\xi}^*_{kN}(p)) + H_{kN}(\xi^*_{kN}(p)) \\ &\quad = O(N^{-3/4}(\log N)^{3/4}). \end{aligned} \tag{3.4}$$

Since $\hat{H}_{kN}(\hat{\xi}^*_{kN}(p)) = p + O(N^{-1})$, we have w.p.1 from (3.4) with (3.3), we obtain the result.

THEOREM 1. *Under Assumption* 2, *for each* k, $k = 1, \ldots, q$ *and for each* $0 < p < 1$, *w.p.*1, *as* $N \to \infty$,

$$\begin{aligned} &\hat{F}_{kn}(\hat{\xi}^*_{kN}(p)) - F_k(\xi^*_{kN}(p)) \\ &\quad = \left\{\frac{1}{n} - \frac{f_k(\xi^*_{kN}(p))}{h_{kN}(\xi^*_{kN}(p))}\frac{1}{N}\right\} \sum_{i=1}^{n} \beta(V_{ki}, \delta_{ki}, \xi^*_{kN}(p)) \\ &\qquad - \frac{f_k(\xi^*_{kN}(p))}{h_{kN}(\xi^*_{kN}(p))}\frac{1}{N} \sum_{j=1}^{m} \gamma(W_{kj}, \tau_{kj}, \xi^*_{kN}(p)) + O(N^{-3/4}(\log N)^{3/4}). \end{aligned}$$

PROOF.

$$\begin{aligned} &\hat{F}_{kn}(\hat{\xi}^*_{kN}(p)) - F_k(\xi^*_{kN}(p)) \\ &\quad = \{\hat{F}_{kn}(\hat{\xi}^*_{kN}(p)) - F_k(\hat{\xi}^*_{kN}(p)) - \hat{F}_{kn}(\xi^*_{kN}(p)) + F_k(\xi^*_{kN}(p))\} \\ &\qquad + \{F_k(\hat{\xi}^*_{kN}(p)) - F_k(\xi^*_{kN}(p))\} + \{\hat{F}_{kn}(\xi^*_{kN}(p)) - F_k(\xi^*_{kN}(p))\} \\ &\quad = A + B + C, \qquad \text{say}. \end{aligned}$$

Then by Cheng (1984), w.p.1, as $N \to \infty$,

$$A = O(N^{-3/4}(\log N)^{3/4}).$$

From Taylor's expansion and Lemmas 1 and 2, w.p.1, as $N \to \infty$,

$$B = -\frac{f_k(\xi^*_{kN}(p))}{h_{kN}(\xi^*_{kN}(p))}\frac{1}{N}\left\{\sum_{i=1}^{n}\beta(V_{ki},\delta_{ki},\xi^*_{kN}(p)) + \sum_{j=1}^{m}\gamma(W_{kj},\tau_{kj},\xi^*_{kN}(p))\right\}$$
$$+O(N^{-3/4}(\log N)^{3/4}).$$

Therefore this theorem is followed by applying Lemma 1 to C.

We note that under H_0

$$E(\beta(V_{ki},\delta_{ki},\xi^*_{kN}(p))) = E(\gamma(W_{kj},\tau_{kj},\xi^*_{kN}(p))) = 0$$
$$V(\beta(V_{ki},\delta_{ki},\xi^*_{kN}(p))) = (1-F_k(\xi^*_{kN}(p)))^2\int_0^{\xi^*_{kN}(p)}\frac{dF^*_k(u)}{S^2_{F_k}(u)} \quad \text{and}$$
$$V(\gamma(W_{kj},\tau_{kj},\xi^*_{kN}(p))) = (1-G_k(\xi^*_{kN}(p)))^2\int_0^{\xi^*_{kN}(p)}\frac{dG^*_k(u)}{S^2_{G_k}(u)}.$$

Also we obtain by applying Fubini's theorem that

(3.5) $\mathrm{Cov}(\beta(V_{ki},\delta_{ki},\xi^*_{kN}(p)),\beta(V_{li},\delta_{li},\xi^*_{lN}(p)))$

$$= (1-F_k(\xi^*_{kN}(p)))(1-F_l(\xi^*_{lN}(p)))$$
$$\left[\int_0^{\xi^*_{kN}(p)}\int_0^{\xi^*_{lN}(p)}\frac{d^2F^*_{kl}(u,v)}{S_{F_k}(u)S_{F_l}(v)} + \int_0^{\xi^*_{kN}(p)}\int_0^{\xi^*_{lN}(p)}\frac{S_{F_{kl}}(u,v)dF^*_k(u)dF^*_l(v)}{S^2_{F_k}(u)S^2_{F_l}(v)}\right.$$
$$-\int_0^{\xi^*_{lN}(p)}\left\{\int_0^{\xi^*_{kN}(p)}\int_v^{\infty}\frac{d^2N_{F_{kl}}(u,s)}{S_{F_k}(u)}\right\}\frac{dF^*_l(v)}{S^2_{F_l}(v)}$$
$$\left.-\int_0^{\xi^*_{kN}(p)}\left\{\int_0^{\xi^*_{lN}(p)}\int_u^{\infty}\frac{d^2M_{F_{kl}}(s,v)}{S_{F_l}(v)}\right\}\frac{dF^*_k(u)}{S^2_{F_k}(u)}\right]$$
$$= (1-F_k(\xi^*_{kN}(p)))(1-F_l(\xi^*_{lN}(p)))(C_1(F)+C_2(F)-C_3(F)-C_4(F)), \quad \text{say}$$

and

(3.6) $\mathrm{Cov}(\gamma(W_{kj},\tau_{kj},\xi^*_{kN}(p)),\gamma(W_{lj},\tau_{lj},\xi^*_{lN}(p)))$

$$= (1-G_1(\xi^*_{kN}))(1-G_2(\xi^*_{lN}))$$
$$\left[\int_0^{\xi^*_{kN}(p)}\int_0^{\xi^*_{lN}(p)}\frac{d^2G^*_{kl}(u,v)}{S_{G_k}(u)S_{G_l}(v)} + \int_0^{\xi^*_{kN}(p)}\int_0^{\xi^*_{lN}(p)}\frac{S_{G_{kl}}(u,v)dG^*_k(u)dG^*_l(v)}{S^2_{G_k}(u)S^2_{G_l}(v)}\right.$$
$$-\int_0^{\xi^*_{lN}(p)}\left\{\int_0^{\xi^*_{kN}(p)}\int_v^{\infty}\frac{d^2N_{G_{kl}}(u,s)}{S_{G_k}(u)}\right\}\frac{dG^*_l(v)}{S^2_{G_l}(v)}$$
$$\left.-\int_0^{\xi^*_{kN}(p)}\left\{\int_0^{\xi^*_{lN}(p)}\int_u^{\infty}\frac{d^2M_{G_{kl}}(s,v)}{S_{G_l}(v)}\right\}\frac{dG^*_k(u)}{S^2_{G_k}(u)}\right]$$
$$= (1-G_k(\xi^*_{kN}(p)))(1-G_l(\xi^*_{lN}(p)))(C_1(G)+C_2(G)-C_3(G)-C_4(G)), \quad \text{say.}$$

We now return to the subject of the limiting null distribution of $\boldsymbol{M}_N$. From Theorem 1, for each k, $k = 1, \ldots, q$, the limiting distributions of $\sqrt{n}(\hat{F}_{kn}(\hat{\xi}^*_{kN}(p)) - p)$ and

$$\frac{1}{n}\sum_{i=1}^{n}\beta(V_{ki}, \delta_{ki}, \xi^*_{kN}(p)) - \frac{1}{N}\left\{\sum_{i=1}^{n}\beta(V_{ki}, \delta_{ki}, \xi^*_{kN}(p)) + \sum_{j=1}^{m}\gamma(W_{kj}, \tau_{kj}, \xi^*_{kN}(p))\right\}$$

are the same under H_0. Therefore from the central limit theorem with Assumptions 1 and 2, under H_0, we see that for each k, $k = 1, \ldots, q$, $\sqrt{n}(\hat{F}_{kn}(\hat{\xi}^*_{kN}(p)) - p)$ converges in distribution to a normal random variable with mean 0 and variance $\sigma^2_k(p)$,

$$\sigma^2_k(p) = \frac{(1-p)^2}{(1+\lambda)^2}\int_0^{\xi^*_k(p)} \frac{dF^*_k(u)}{S^2_{F_k}(u)} + \frac{\lambda(1-p)^2}{(1+\lambda)^2}\int_0^{\xi^*_k(p)} \frac{dG^*_k(u)}{S^2_{G_k}(u)},$$

where $\xi^*_k(p) = \lim_{N\to\infty}\xi^*_{kN}(p)$. Also from (3.5) and (3.6), we see that the limiting null covariance between

$$\sqrt{n}(\hat{F}_{kn}(\hat{\xi}^*_{kN}(p)) - p) \quad \text{and} \quad \sqrt{n}(\hat{F}_{ln}(\hat{\xi}^*_{lN}(p)) - p)$$

is

$$\begin{aligned}\sigma_{kl}(p) &= \frac{(1-p)^2}{(1+\lambda)^2}(C_1(F) + C_2(F) - C_3(F) - C_4(F)) \\ &+ \frac{\lambda(1-p)^2}{(1+\lambda)^2}(C_1(G) + C_2(G) - C_3(G) - C_4(G))\end{aligned}$$

with substitution of $\xi^*_k(p)$ for $\xi^*_{kN}(p)$ in (3.5) and (3.6). Then by applying Cramér-Wold device (cf. Billingsley (1986)), we obtain the following result.

THEOREM 2. *For any consistent estimate $\hat{\Sigma}_N(p)$ of $\Sigma_0(p)$, under H_0, $\boldsymbol{M}_N$ converges in distribution to a χ^2 random variable with q degrees of freedom, where*

$$\Sigma_0(p) = \begin{pmatrix} \sigma^2_1(p) \cdots \sigma_1q(p) \\ \cdots \\ \sigma_1q(p) \cdots \sigma^2_q(p) \end{pmatrix}.$$

We note that under H_0, the first and the second parts of each variance and covariance term are the same except for λ. Therefore we could have reduced the expression of $\Sigma_0(p)$ to a more concise form. However since we have to obtain $\hat{\Sigma}_N(p)$ from two samples, we do not reduce them in this manner. A consistent estimate $\hat{\Sigma}_N(p)$ for $\Sigma_0(p)$ can be obtained by substituting empirical ones for the quantities, which were introduced at the beginning of this section. Then one can show the consistency of $\hat{\Sigma}_N(p)$ by proving the consistency of each component of $\hat{\Sigma}_N(p)$. For more detailed discussion, we may refer to Park and Desu (1998).

4. Asymptotic relative efficiency and an example

In this section, we study the asymptotic relative efficiency (ARE). For this matter, we only consider comparing two types of the median tests. Let $\boldsymbol{L}_N$ be the control median test statistic which was proposed by Park and Desu (1998). We begin this section by stating the definition of ARE for the multivariate version (cf. Puri and Sen (1985)). For two sequences of test statistics, say $\{\boldsymbol{Q}_N\}$ and $\{\boldsymbol{Q}^*_N\}$, having asymptotically (under a sequence $\{H_{1N}\}$ of alternative hypotheses) noncentral chi-square distributions with q degrees of freedom and noncentrality parameters $\boldsymbol{\Psi}$ and $\boldsymbol{\Psi}^*$, respectively, the ARE of $\{\boldsymbol{Q}_N\}$ relative to $\{\boldsymbol{Q}^*_N\}$ is defined by

$$\mathrm{ARE}(\boldsymbol{Q},\boldsymbol{Q}^*) = \frac{\boldsymbol{\Psi}}{\boldsymbol{\Psi}^*}.$$

Noncentrality parameters of $\{\boldsymbol{M}_N\}$ and $\{\boldsymbol{L}_N\}$ under alternatives depend on the censoring distributions in a complicated manner. Therefore we will assume, in this section, that censoring distributions for two samples are equal. We consider ARE under the Pitman translation alternatives: for each k, $k = 1,\ldots,q$ and for each N,

$$H_{1N} : \boldsymbol{\Delta}_N = (\Delta_{1N},\ldots,\Delta_{qN})^T = (\theta_1/\sqrt{N},\ldots,\theta_q/\sqrt{N})^T,$$

where for each k, θ_k is some nonzero constant. Before we derive the ARE$(\boldsymbol{L},\boldsymbol{M})$, we review a useful relation between the noncentrality parameter and the efficacies of components of test statistics under the Pitman translation alternatives. For this purpose, let $\{\boldsymbol{Z}_N = (Z_{1N},\ldots,Z_{qN})^T\}$ be a sequence of q-variate test statistics such that for each N, $\boldsymbol{Z}_N$ is arbitrarily distributed with mean vector, $\boldsymbol{\mu}_N(\boldsymbol{\Delta}_N)$ and covariance matrix, $\Sigma_N(\boldsymbol{\Delta}_N)$, where $\boldsymbol{\mu}_N(\boldsymbol{\Delta}_N) = E(\boldsymbol{Z}_N \mid \boldsymbol{\Delta}_N)$ and $\Sigma_N(\boldsymbol{\Delta}_N) = V(\boldsymbol{Z}_N \mid \boldsymbol{\Delta}_N)$. We assume that $\boldsymbol{Z}_N$ converges in distribution to $\boldsymbol{Z}$, where $\boldsymbol{Z}$ is normally distributed with mean vector, $\boldsymbol{\mu}$ and covariance matrix, $\boldsymbol{\Sigma}$. Then we note that $(\boldsymbol{Z}_N - \boldsymbol{\mu}_N)^T\Sigma_N^{-1}(\boldsymbol{Z}_N - \boldsymbol{\mu}_N)$ converges in distribution to a chi-square random variable with q degrees of freedom. With those notations and assumptions, we state the following result.

LEMMA 3. *For the sequence $\{\boldsymbol{Z}_N\}$ of test statistics, suppose that*

(1) *for each k and for each N, $\frac{d}{d\Delta}\mu_{kN}(\Delta) = \mu'_{kN}(\Delta)$ is assumed to exist and be continuous in some neighborhood of* 0 *with $\mu'_N(0) \neq 0$,*

(2) $\lim_{N\to\infty} \mu'_{kN}(\Delta_{kN})/\mu'_{kN}(0) = 1$ *and*

(3) $\lim_{N\to\infty} \Sigma_N(\boldsymbol{\Delta}_N) = \boldsymbol{\Sigma}$.

Then under the Pitman translation alternatives, the limiting distribution of $\boldsymbol{Z}_N\Sigma_N^{-1}(\boldsymbol{\Delta}_N)$ $\boldsymbol{Z}_N$ is a noncentral chi-square distribution with q degrees of freedom and the noncentrality parameter

$$\boldsymbol{\Psi} = \begin{pmatrix}\theta_1 e_1\\ \cdots\\ \theta_p e_p\end{pmatrix}^T \boldsymbol{P}^{-1} \begin{pmatrix}\theta_1 e_1\\ \cdots\\ \theta_p e_p\end{pmatrix},$$

where for each k, e_k is the efficacy of the k-th component, Z_{kN} of the test statistic, $\boldsymbol{Z}_N$ and $\boldsymbol{P}$ is the limiting correlation matrix.

PROOF. See Park and Desu (1999).

We note that the conditions (1) and (2) in Lemma 3 with the condition that the sequence $\{\boldsymbol{Z}_N\}$ of test statistics has a limiting distribution, are exactly the same as those for the derivation of the efficacy of $\{\boldsymbol{Z}_N\}$ in Theorem 5.2.7 of Randles and Wolfe (1979) except the existence of the efficacy itself. Assumption 1 in Section 3 implies that $\xi^*_{kN} \to \xi^*_k$, where ξ^*_k is a median of $H_k = (1/(1+\lambda))F_k + (\lambda/(1+\lambda))G_k$. Since the noncentrality parameter contains the expressions of the limiting distribution of the sequence of test statistics, without loss of generality, we use ξ_k^* instead of ξ_{kN}^* in the sequel to obtain the noncentrality parameter. Also we use Δ_k instead of Δ_{kN} for each N when there is no confusion. We note that under H_0 and the Pitman translation alternatives, ξ_k^* becomes also a median of F_k and G_k.

Therefore in view of Lemma 3, it is enough to derive the efficacies of q components and limiting correlation matrix for the noncentrality parameter. For each k, $k = 1, \ldots, q$ and $1 \leq k \neq l \leq q$, define

$$\mu_{kN}(\Delta_k) = \sqrt{n}(G_k(\xi_k^* + \Delta_k) - 1/2),$$

$$\sigma_{kN}^2(\Delta_k) = \left\{1 - \frac{n}{N}\frac{f_k(\xi_k^* + \Delta_k)}{h_{kN}(\xi_k^* + \Delta_k)}\right\}^2 (1 - F_k(\xi_k^* + \Delta_k))^2 \int_0^{\xi_k^* + \Delta_k} \frac{dF_k^*(u)}{S_{F_k}^2(u)}$$
$$+ \frac{mn}{N^2}\left\{\frac{f_k(\xi_k^* + \Delta_k)}{h_{kN}(\xi_k^* + \Delta_k)}\right\}^2 (1 - G_k(\xi_k^* + \Delta_k))^2 \int_0^{\xi_k^* + \Delta_k} \frac{dG_k^*(u)}{S_{G_k}^2(u)}$$

and

$$\sigma_{klN}(\boldsymbol{\Delta}) = \left\{1 - \frac{n}{N}\frac{f_k(\xi_k^* + \Delta_k)}{h_{kN}(\xi_k^* + \Delta_k)}\right\}\left\{1 - \frac{n}{N}\frac{f_l(\xi_l^* + \Delta_l)}{h_{lN}(\xi_l^* + \Delta_l)}\right\}$$
$$\mathrm{Cov}\{\beta(V_{ki}, \delta_{ki}, \xi_k^* + \Delta_k), \beta(V_{li}, \delta_{li}, \xi_l^* + \Delta_l)\}$$
$$+ \frac{mn}{N^2}\frac{f_k(\xi_k^* + \Delta_k) f_l(\xi_l^* + \Delta_l)}{h_{kN}(\xi_k^* + \Delta_k) h_{lN}(\xi_l^* + \Delta_l)}$$
$$\mathrm{Cov}\{\gamma(W_{kj}, \tau_{kj}, \xi_k^* + \Delta_k), \gamma(W_{lj}, \tau_{lj}, \xi_l^* + \Delta_l)\}.$$

Then under the Pitman translation alternatives,

$$n \begin{pmatrix} \hat{F}_{1n}(\hat{\xi}_1^*) - 1/2 - \mu_{1N}(\Delta_1) \\ \cdots \\ \hat{F}_{qn}(\hat{\xi}_q^*) - 1/2 - \mu_{qN}(\Delta_q) \end{pmatrix}^T \boldsymbol{\Sigma}_N^{-1}(\boldsymbol{\Delta}) \begin{pmatrix} \hat{F}_{1n}(\hat{\xi}_1^*) - 1/2 - \mu_{1N}(\Delta_1) \\ \cdots \\ \hat{F}_{qn}(\hat{\xi}_q^*) - 1/2 - \mu_{qN}(\Delta_q) \end{pmatrix}$$

converges in distribution to a chi-square random variable with q degrees of freedom. Therefore we can use the Lemma 3 to derive the noncentrality parameter by checking the three conditions. Assumption 2 in Section 3 guarantees the condition (1). Thus we have

$$\frac{d\mu_{kN}(\Delta_k)}{d\Delta_k} = \sqrt{n} f_k(\xi_k^* + \Delta_k) \quad \text{and} \quad \left.\frac{d\mu_{kN}(\Delta_k)}{d\Delta_k}\right|_{\Delta_k = 0} = \sqrt{n} f_k(\xi_k^*).$$

With the fact that $\Delta_k \to 0$ as $N \to \infty$, we see that

$$\lim_{N \to \infty} \frac{d\mu_{kN}(\Delta_k)/d\Delta_k}{d\mu_{kN}(\Delta_k)/d\Delta_k|_{\Delta_k = 0}} = 1,$$

which confirms the condition (2).

In order to check the condition (3), we take $\boldsymbol{\Sigma}_0$ as $\boldsymbol{\Sigma}$. $\boldsymbol{\Sigma}_0$ was defined in Section 3. Since $h_{kN} = (n/N)f_k + (m/N)g_k$ with the fact that ξ_k^* is a common median of F_k and G_k under the Pitman translation alternatives, we have

$$\lim_{N \to \infty} \frac{f_k(\xi_k^* + \Delta_k)}{h_{kN}(\xi_k^* + \Delta_k)} = 1, \quad \text{and}$$
$$\lim_{N \to \infty} F_k(\xi_k^* + \Delta_k) = \lim_{N \to \infty} G_k(\xi_k^* + \Delta_k) = G_k(\xi_k^*) = 1/2.$$

Thus with the assumption that censoring distributions for the control and the treatment are equal, we can conclude with Assumption 1 that for each k,

$$\lim_{N\to\infty} \sigma_{kN}^2(\Delta_k) = \sigma_k^2.$$

Also with the same arguments used for σ_k^2, we can show that

$$\lim_{N\to\infty} \sigma_{klN}(\boldsymbol{\Delta}) = \sigma_{kl}.$$

Therefore we have shown that all the three conditions in Lemma 3 are satisfied. This means that, in view of Lemma 3, it is enough to consider the efficacies of two components with limiting null correlation matrix to obtain the noncentrality parameter for the median test $\boldsymbol{M}_N$.

The conditions and method for the derivation of the efficacy for tests statistics are well summarized in Randles and Wolfe (1979). Already we have noticed that all the conditions in Theorem 5.2.7 in Randles and Wolfe are satisfied except the existence of the efficacy. Therefore it is enough to check that condition. Then some simple considerations for the efficacy e_k of the k-th component of $\boldsymbol{M}_N$ leads as follows:

$$e_k = 2g_k(\xi_k^*)\left\{\frac{(1+\lambda)^2}{\lambda}\int_0^{\xi_k^*}\frac{dG_k^*(u)}{S_{G_k}^2(u)}\right\}^{-1/2}.$$

Also straightforward calculations produce the limiting null correlation matrix $\boldsymbol{P}$ with

$$P_{11} = \cdots = P_{qq} = 1 \quad \text{and}$$
$$P_{kl} = P_{lk} = (C_1(G) + C_2(G) - C_3(G) - C_4(G))\left\{\int_0^{\xi_k^*}\frac{dG_k(u)}{S_{G_k}^2(u)}\int_0^{\xi_l^*}\frac{dG_l(u)}{S_{G_l}^2(u)}\right\}^{-1/2}.$$

In the following, we denote ξ_k as a median of G_k for each k. Then we note that under H_0 and Pitman translation alternatives, $\xi_k = \xi_k^*$. In order to derive the noncentrality parameter for the control median tests statistics $\{\boldsymbol{L}_N\}$ (cf. Park and Desu (1998)), define for each k, $k = 1, \ldots, q$ and $1 \le k \ne l \le q$,

$$\begin{aligned}
\mu_{kN}(\Delta_k) &= \sqrt{n}(G_k(\xi_k + \Delta_k) - 1/2),\\
\sigma_{kN}^2(\Delta_k) &= (1 - F_k(\xi_k + \Delta_k))^2\int_0^{\xi_k+\Delta_k}\frac{dF_k^*(u)}{S_{F_k}^2(u)}\\
&\quad + \frac{n}{m}(1 - G_k(\xi_k + \Delta_k))^2\frac{f_k^2(\xi_k + \Delta_k)}{g_k^2(\xi_k + \Delta_k)}\int_0^{\xi_k+\Delta_k}\frac{dG_k^*(u)}{S_{G_k}^2(u)}
\end{aligned}$$

and

$$\begin{aligned}
\sigma_{klN}(\boldsymbol{\Delta}) &= \mathrm{Cov}\{\beta(V_{ki}, \delta_{ki}, \xi_k + \Delta_k), \beta(V_{li}, \delta_{li}, \xi_l + \Delta_l)\}\\
&\quad + \frac{n}{m}\frac{f_k(\xi_k + \Delta_k)f_l(\xi_l + \Delta_l)}{g_k(\xi_k + \Delta_k)g_l(\xi_l + \Delta_l)}\\
&\quad \mathrm{Cov}\{\gamma(W_{kj}, \tau_{kj}, \xi_k + \Delta_k), \gamma(W_{lj}, \tau_{lj}, \xi_l + \Delta_l)\}.
\end{aligned}$$

Thus

$$n\begin{pmatrix}\hat{F}_{1n}(\hat{G}_{1m}^{-1}(1/2)) - 1/2 - \mu_{1N}(\Delta_1)\\ \cdots \\ \hat{F}_{qn}(\hat{G}_{qm}^{-1}(1/2)) - 1/2 - \mu_{qN}(\Delta_q)\end{pmatrix}^T \Sigma_N^{-1}(\boldsymbol{\Delta})\begin{pmatrix}\hat{F}_{1n}(\hat{G}_{1m}^{-1}(1/2)) - 1/2 - \mu_{1N}(\Delta_1)\\ \cdots \\ \hat{F}_{qn}(\hat{G}_{qm}^{-1}(1/2)) - 1/2 - \mu_{qN}(\Delta_q)\end{pmatrix}$$

converges in distribution to a chi-square random variable with q degrees of freedom. Therefore by the same arguments used for $\{\boldsymbol{M}_N\}$, we can show that the conditions (1), (2) and (3) in Lemma 3, are all satisfied. Then by checking all the conditions of Theorem 5.4.7 in Randles and Wolfe (1979), we obtain the same efficacies as those of $\{\boldsymbol{L}_N\}$. Also it is easy to show that the limiting correlation matrix for $\boldsymbol{L}_N$ is the same as that of $\boldsymbol{M}_N$. Therefore we conclude that with the fact that $\xi_k^* = \xi_k$ under Pitman translation alternatives, $\mathrm{ARE}(\boldsymbol{L}, \boldsymbol{M}) = 1$.

Finally we illustrate our procedure with the NCGS data considered by Wei and Lachin (1984). The patients are allocated into two groups, i.e. control (placebo) and treatment (high dose) groups with sample sizes $n = 48$ and $m = 65$. The Kaplan-Meier estimate for the second component of high dose group (X_{12}) shows that a sample median cannot be obtained because of the heavy censoring of higher observations. Therefore one can not apply any median test procedure. Since the lower (or first) sample quartile point (25%) can be achieved for all components, we consider applying the 25 percentile test to this example. The necessary statistics for obtaining the 25 percentile test statistic are as follows:

$$\hat{\xi}^*_{1,113}(.25) = 249.23 \quad \text{and} \quad \hat{\xi}_{2,113}(.25) = 640.26$$

$$\hat{F}_{1,48}(249.23) - 0.25 = 0.44 - 0.25 = 0.19$$

$$\hat{F}_{2,48}(640.26) - 0.25 = 0.27 - 0.25 = 0.02$$

$$\Sigma_{113} = \begin{pmatrix} 0.2842884 & 0.1733398 \\ 0.1733398 & 0.1834684 \end{pmatrix} \quad \text{and} \quad \Sigma_{113}^{-1} = \begin{pmatrix} 8.2975131 & -7.839439 \\ -7.839439 & 12.857183 \end{pmatrix}.$$

Then we obtain that $M_{113} = 11.765$, whose p-value is less than 0.005 from the chi-square distribution with 2 degrees of freedom. Therefore we may conclude that the two groups of patients are significantly different for the disease progression.

Acknowledgements

The author would like to express his sincere appreciations to an anonymous referee for suggesting constructive opinions, introducing the necessary materials and pointing out errors to improve the original version of this paper.

References

Alling, D. W. (1963). Early decision in the Wilcoxon two-sample test, *J. Amer. Statist. Assoc.*, **58**, 713–720.

Babu, G. J. and Rao, C. R. (1988). Joint asymptotic distribution of marginal quantile functions in samples from a multivariate population, *J. Multivariate Anal.*, **27**, 15–23.

Billingsley, P. (1986). *Probability and Measure*, 2nd ed., Wiley, New York.

Brookmeyer, R. and Crowley, J. (1982). A k-sample median test for censored data, *J. Amer. Statist. Assoc.*, **77**, 433–440.

Chatterjee, S. K. and Sen, P. K. (1964). Non-parametric tests for the bivariate two-sample location problem, *Calcutta Statist. Assoc. Bull.*, **13**, 18–58.

Cheng, K. F. (1984). On almost sure representations for quantiles of the product-limit estimator with applications, *Sankhyā, Ser. A*, **46**, 426–443.

Gastwirth, J. L. (1968). The first-median test: A two-sided version of the control median test, *J. Amer. Statist. Assoc.*, **63**, 692–706.

Gastwirth, J. L. and Wang, J. L. (1988). Control percentile test procedures for censored data, *J. Statist. Plann. Inference*, **18**, 267–276.

Hettmansperger, T. P. (1973). A large sample conservative test for location with unknown scale parameters, *J. Amer. Statist. Assoc.*, **68**, 466–468.

Hettmansperger, T. P. (1984). *Statistical Inference Based on Ranks*, Wiley, New York.

Lo, S. H. and Singh, K. (1985). The product-limit estimator and the bootstrap: Some asymptotic representations, *Probab. Theory Related Fields*, **71**, 455–465.

Mathisen, H. C. (1943). A method of testing the hypothesis that two samples are from the same population, *Ann. Math. Statist.*, **14**, 188–194.

Mood, A. M. (1950). *Introduction to the Theory of Statistics*, McGraw-Hill, New York.

Park, H. I. and Desu, M. M. (1998). Multivariate control median test for right censored data, *Comm. Statist. Theory Methods*, **27**(8), 1923–1935.

Park, H. I. and Desu, M. M. (1999). A multivariate control median test, *J. Statist. Plann. Inference*, **79**, 123–139.

Puri, M. L. and Sen, P. K. (1985). *Nonparametric Methods in General Linear Models*, Wiley, New York.

Randles, H. R. and Wolfe, D. A. (1979). *Introduction to the Theory of Nonparametric Statistics*, Wiley, New York.

Wei, L. J. and Lachin, J. M. (1984). Two-sample asymptotically distribution-free tests for incomplete multivariate observations, *J. Amer. Statist. Assoc.*, **79**, 653–661.

Ann. Inst. Statist. Math.
Vol. 54, No. 4, 945–959 (2002)

D-OPTIMAL DESIGNS FOR TRIGONOMETRIC REGRESSION MODELS ON A PARTIAL CIRCLE

HOLGER DETTE[1], VIATCHESLAV B. MELAS[2] AND ANDREY PEPELYSHEV[2]

[1] *Fakultät für Mathematik, Lehrstuhl III (Stochastik), Ruhr-Universität Bochum, Gebaude NA 3/72, 44780 Bochum, Germany,* e-mail: holger.dette@ruhr-uni-bochum.de
[2] *Department of Mathematics, St. Petersburg State University, St. Petersburg, Russian Federation,* e-mail: v.melas@pobox.spbu.ru; andrey@ap7236.spb.edu

(Received March 16, 2001; revised August 24, 2001)

Abstract. In the common trigonometric regression model we investigate the D-optimal design problem, where the design space is a partial circle. It is demonstrated that the structure of the optimal design depends only on the length of the design space and that the support points (and weights) are analytic functions of this parameter. By means of a Taylor expansion we provide a recursive algorithm such that the D-optimal designs for Fourier regression models on a partial circle can be determined in all cases. In the linear and quadratic case the D-optimal design can be determined explicitly.

Key words and phrases: Trigonometric regression, D-optimality, implicit function theorem, orthogonal polynomial.

1. Introduction

Trigonometric regession models of the form

$$y = \beta_0 + \sum_{j=1}^{m} \beta_{2j-1} \sin(jt) + \sum_{j=1}^{m} \beta_{2j} \cos(jt) + \varepsilon, \qquad t \in [c, d]; \tag{1.1}$$

$-\infty < c < d < \infty$; are widely used to describe periodic phenomena (see e.g. Mardia (1972), Graybill (1976) or Kitsos *et al.* (1988)) and the problem of designing experiments for Fourier regression models has been discussed by several authors (see e.g. Hoel (1965), Karlin and Studden (1966), p. 347, Fedorov (1972), p. 94, Hill (1978), Lau and Studden (1985), Riccomagno *et al.* (1997)). Most authors concentrate on the design space $(-\pi, \pi]$, but Hill (1978) and Kitsos *et al.* (1988) point out that in many applications it is impossible to take observations on the full circle $[-\pi, \pi]$. We refer to Kitsos *et al.* (1988) for a concrete example, who investigated a design problem in rhythmometry involving circadian rhythm exhibited by peak expiratory flow, for which the design region has to be restricted to a partial cycle of the complete 24-hour period.

In the present paper, we address the question of designing experiments in trigonometric models, where the design space is not necessarily the full circle but an arbitrary interval $[c, d] \subset \mathbb{R}$. Recently, Dette and Melas (2003) considered optimal designs for estimating individual coefficients in this model and gave a partial solution to this problem. In the present paper, we consider the D-optimality criterion, which is a reasonable criterion if efficient estimates of all parameters in the model are desired. It is demonstrated in Section 2 that the structure of the D-optimal design depends only on the length $a = (c - d)/2$ of the design space and that there only exist two types of D-optimal designs (this result seems to be even unknown for the complete circle). Our main result of Section 3 proves that the support points (and weights) of the D-optimal design are analytic functions of the parameter a and that an appropriately scaled version of the D-optimal design converges weakly as $a \to 0$ to a nondegenerate discrete distribution on the interval $[0, 1]$. Following Melas (1978), these results are applied to obtain Taylor expansions for the support points of the D-optimal design (considered as a function of the parameter $a = (d - c)/2$), which allows a complete solution of the D-optimal design problem in the trigonometric regression model (1.1) on the interval $[c, d]$. Finally, some examples are given in Section 4, and in the linear and quadratic trigonometric regression model on the interval $[-a, a]$ D-optimal designs are determined explicitly.

2. Preliminary results for D-optimal designs in trigonometric regression models on a partial circle

Consider the trigonometric regression model (1.1), define $\beta = (\beta_0, \beta_1, \ldots, \beta_{2m})^T$ as the vector of parameters and

$$f(t) = (1, \sin t, \cos t, \ldots, \sin(mt), \cos(mt))^T = (f_0(t), \ldots, f_{2m}(t))^T \tag{2.1}$$

as the vector of regression functions. An approximate design is a probability measure ξ on the design space $[c, d]$ with finite support (see e.g. Kiefer (1974)). The support points of the design ξ give the locations, where observations are taken, while the weights give the corresponding proportions of total observations to be taken at these points. Due to the 2π-periodicity of the regression functions we restrict ourselves without loss of generality to design spaces with length $d - c \leq 2\pi$. For uncorrelated observations (obtained from an approximate design) the covariance matrix of the least squares estimator for the parameter β is approximately proportional to the matrix

$$M(\xi) = \int f(t) f^T(t) d\xi(t) \in \mathbb{R}^{2m+1 \times 2m+1}, \tag{2.2}$$

which is called Fisher information matrix in the design literature. An optimal design minimizes (or maximizes) an appropriate convex (or concave) function of the information matrix and there are numerous criteria proposed in the literature, which can be used for the discrimination between competing designs (see e.g. Fedorov (1972), Silvey (1980) or Pukelsheim (1993)).

In this paper, we are interested in D-optimal designs for the trigonometric regression model (1.1) on the interval $[c, d]$, which maximize the determinant $\det M(\xi)$ of the Fisher information matrix in the space of all approximate designs on the interval $[c, d]$. Note that a D-optimal design minimizes the (approximate) volume of the ellipsoid of concentration for the vector β of the unknown parameters in the model (1.1) (see e.g. Fedorov (1972)) and that optimal designs in the trigonometric regression model (1.1) for the full circle $[c, d] = [-\pi, \pi]$ have been determined by numerous authors (see e.g. Karlin and Studden (1966), Fedorov (1972), Lau and Studden (1985), Pukelsheim (1993) or Dette and Haller (1998) among many others).

Our first preliminary result demonstrates that for the solution of the D-optimal design problem on a partial circle it is sufficient to consider only symmetric design spaces. To be precise, let

$$\eta = \begin{pmatrix} t_0 & \dots & t_n \\ \omega_0 & \dots & \omega_n \end{pmatrix} \tag{2.3}$$

denote a design on the interval $[c,d]$ with different support points $t_0 < \cdots < t_n$ and positive weights $\omega_0, \dots, \omega_n$ adding to one and define its affine transformation onto the symmetric interval $[-a,a]$ by

$$\xi_\eta = \begin{pmatrix} \tilde{t}_0 & \dots & \tilde{t}_n \\ \omega_0 & \dots & \omega_n \end{pmatrix} \tag{2.4}$$

where $a = (d-c)/2$ and $\tilde{t}_i = t_i - (d+c)/2$, $i = 1, \dots, n$.

LEMMA 2.1. *Let $M(\eta)$ and $M(\xi_\eta)$ denote the information matrices in the trigonometric regression model* (1.1) *of the designs η and ξ_η defined by* (2.3) *and* (2.4), *respectively, then*

$$\det M(\xi_\eta) = \det M(\eta). \tag{2.5}$$

PROOF. If the number of support points satisfies $n+1 < 2m+1$, then both sides of the equation (2.5) vanish and the proof is trivial. Next consider the case $n = 2m$, for which we have (see e.g. Karlin and Studden (1966))

$$\det M(\xi_\eta) = (\det F(\xi_\eta))^2 \prod_{i=0}^{2m} \omega_i, \tag{2.6}$$

where the matrix $F(\xi_\eta) \in \mathbb{R}^{2m+1 \times 2m+1}$ is defined by

$$F(\xi_\eta) = \left(f_i(\tilde{t}_j)\right)_{i=0,\dots,2m}^{j=0,\dots,2m}. \tag{2.7}$$

Now it is easy to see that the vector $f(t)$ defined by (2.1) satisfies for any $\alpha \in \mathbb{R}$

$$f(t+\alpha) = Pf(t)$$

where P is a $(2m+1) \times (2m+1)$ diagonal block matrix defined by

$$P = \begin{pmatrix} 1 & & & \\ & Q(\alpha) & & \\ & & \ddots & \\ & & & Q(m\alpha) \end{pmatrix}$$

and $Q(\beta)$ is a 2×2 rotation matrix given by

$$Q(\beta) = \begin{pmatrix} \cos(\beta) & \sin(\beta) \\ -\sin(\beta) & \cos(\beta) \end{pmatrix}.$$

Obviously, we have $\det P = 1$ and obtain from (2.6) and (2.7)

$$\det M(\xi_\eta) = \det M(\eta),$$

which proves the assertion of the lemma in the case $n = 2m$. Finally, in the remaining case $n > 2m$, the assertion follows from the Cauchy Binet formula and the arguments given for the case $n = 2m$. □

From Lemma 2.1 it is clear that it is sufficient to determine the D-optimal designs for symmetric intervals

$$[c, d] = [-a, a], \qquad 0 < a \leq \pi$$

and we will restrict ourselves to this case throughout this paper. For fixed $a \in (0, \pi]$ let ξ_a^* denote a D-optimal design for the trigonometric regression model (1.1) on the interval $[-a, a]$. Note that in general the D-optimal design for the trigonometric regression model is not necessarily unique (see e.g. Fedorov (1972), who considered the case $a = \pi$). However, it is known that the optimal information matrix $M(\xi_a^*)$ is unique and nonsingular (see e.g. Pukelsheim (1993), p. 151). Moreover, due to the equivalence theorem for D-optimality (see Kiefer (1974)) the design ξ_α^* satisfies

$$d(t, \xi_\alpha^*) \leq 0 \quad \text{for all} \quad t \in [-a, a], \tag{2.8}$$

with equality at the support points, where

$$d(t, \xi) = f^T(t) M^{-1}(\xi) f(t) - (2m + 1) \tag{2.9}$$

denotes the directional derivative of the function $\xi \longrightarrow \log\det M(\xi)$ (see Silvey (1980), p. 20). Let $\Xi_a^{(1)}$ denote the set of all designs of the form

$$\xi = \xi(a) = \begin{pmatrix} -t_m & \dots & -t_1 & t_0 & t_1 & \dots & t_m \\ \dfrac{1}{2m+1} & \dots & \dfrac{1}{2m+1} & \dfrac{1}{2m+1} & \dfrac{1}{2m+1} & \dots & \dfrac{1}{2m+1} \end{pmatrix} \tag{2.10}$$

where $0 = t_0 < t_1 < \dots < t_m = a$ and define

$$\Xi^{(2)} = \{\xi \mid supp(\xi) \subset [-a, a],\ d(t, \xi) = 0 \text{ for all } t \in [-a, a]\} \tag{2.11}$$

as the set of all designs on the interval $[-a, a]$ with vanishing directional derivative for all $t \in [-a, a]$, then we obtain the following auxiliary result.

LEMMA 2.2. *Let ξ_a^* denote a D-optimal design on the interval $[-a, a]$, then*

$$\xi_a^* \in \Xi_a^{(1)} \cup \Xi_a^{(2)}.$$

PROOF. Due to the equivalence theorem (2.8) any design $\xi \in \Xi_a^{(2)}$ is D-optimal for trigonometric regression model (1.1) on the interval $[-a, a]$. Now assume that

$$\xi = \begin{pmatrix} u_1 & \dots & u_n \\ \omega_1 & \dots & \omega_n \end{pmatrix}$$

is D-optimal for the trigonometric regression on the interval $[-a, a]$, where the support points satisfy $-a \leq u_1 < \dots < u_n \leq a$. If $\xi \notin \Xi_a^{(2)}$, then $d(t, \xi) \not\equiv 0$, but due the equivalence theorem we have

$$\begin{aligned} & d(u, \xi) \leq 0 \quad \forall u \in [-a, a] \\ & d(u_i, \xi) = 0 \quad \forall i = 1, \dots, n \\ & \frac{d}{du} d(u, \xi)|_{u=u_i} = 0 \quad \forall i = 2, \dots, n-1. \end{aligned} \tag{2.12}$$

If $\tilde{\xi}$ denotes the reflection of ξ at the origin, then it is easy to see that $\det M(\xi) = \det M(\tilde{\xi})$ and consequently $\tilde{\xi}$ is also D-optimal. Moreover, the concavity of the D-criterion implies that the symmetric design $\xi^* = (\xi + \tilde{\xi})/2$ is also D-optimal in the trigonometric regression (1.1) on the interval $[-a, a]$. Note that there exists a permutation matrix $P \in \mathbb{R}^{2m+1 \times 2m+1}$ such that

$$PM(\xi)P^T = \begin{pmatrix} M_1(\xi) & M_2(\xi) \\ M_2^T(\xi) & M_3(\xi) \end{pmatrix}, \tag{2.13}$$

where

$$\begin{aligned} M_1(\xi) &= \int_{-a}^{a} f_c(t) f_c^T(t) d\xi(t) \in \mathbb{R}^{m+1 \times m+1} \\ M_2(\xi) &= \int_{-a}^{a} f_c(t) f_s^T(t) d\xi(t) \in \mathbb{R}^{m+1 \times m} \\ M_3(\xi) &= \int_{-a}^{a} f_s(t) f_s^T(t) d\xi(t) \in \mathbb{R}^{m \times m} \end{aligned} \tag{2.14}$$

and $f_c(t) = (1, \cos(t), \ldots, \cos(mt))^T$, $f_s(t) = (\sin(t), \ldots, \sin(mt))^T$. Because the information matrix of the D-optimal design is unique (see Pukelsheim (1993)), we obtain (note that ξ^* is symmetric)

$$M_2(\xi) = M_2(\tilde{\xi}) = M_2(\xi^*) = 0 \in \mathbb{R}^{m+1 \times m},$$

which implies for the directional derivative in (2.9)

$$\begin{aligned} g(t) = d(t, \xi) &= f_c{}^T(t) M_1^{-1}(\xi) f_c(t) + f_s{}^T(t) M_3^{-1}(\xi) f_s(t) - (2m+1) \\ &= \sum_{i=0}^{2m} \gamma_i \cos(it) \end{aligned} \tag{2.15}$$

for appropriate constants $\gamma_0, \ldots, \gamma_{2m}$ (note that the last representation follows by well known trigonometric formulas). From $\xi \notin \Xi_a^{(2)}$ we obtain that the polynomial $g(t)$ is not identically zero and the equivalence theorem shows that every suppport point is a zero of the function g. Moreover, the functions $\{1, \cos t, \ldots, \cos(2mt)\}$ form a Chebyshev system on the interval $[0, a]$ and a Chebyshev system on the interval $[-a, 0]$. Consequently, g has at most $2m + 1$ roots in the interval $[0, a]$ and at most $2m + 1$ zeros in the interval $[-a, 0]$ (including counting of multiplicities) (see Karlin and Studden (1966)). Consider the case $[0, a]$ and substitute $t = \arccos x$, then it follows, observing the definition of the Chebyshev polynomials of the first kind

$$T_i(x) = \cos(i \arccos x), \tag{2.16}$$

(see Rivlin (1974)) that $g(\arccos x)$ is a nonpositive polynomial of degree $2m$ on the interval $[\cos a, 1]$. Consequently, if $g(\arccos x)$ has exactly $2m$ roots (including counting of multiplicities), the boundary points $\cos a$ and 1 have to be roots of $g(\arccos x)$. Note that a similar argument applies to the interval $[-a, 0]$ and therefore the nonpositive function g defined in (2.15) has at most $4m$ roots (including counting of multiplicities) in the interval $[-a, a]$. Because the number of regression functions is $2m+1$, it therefore follows from (2.12) that any D-optimal design $\eta \notin \Xi_a^{(2)}$ has exactly $2m + 1$ support points in the interval $[-a, a]$ including the boundary points $-a, a$. A standard argument shows that all weights of the D-optimal design have to be equal, i.e. $\omega_j = 1/(2m+1)$, $j = 1, \dots, 2m+1$. If $\xi \notin \Xi_a^{(1)}$, then $\xi \neq \tilde{\xi}$ and consequently $\xi^* = (\xi + \tilde{\xi})/2$ is a D-optimal design for the trigonometric regression model (1.1) on interval $[-a, a]$ with more than $2m+1$ support points, which is impossible, by the above discussion. This shows $\xi \in \Xi_a^{(1)}$ and proves Lemma 2.2. □

3. Analytic properties of D-optimal designs in trigonometric regression models on a partial circle

Lemma 2.2 motivates the consideration of designs of the form (2.10) and our next lemma gives an explicit representation for the determinant of the information matrix of this type of design.

LEMMA 3.1. *Let ξ denote a design of the form* (2.10) *and $x_i = \cos t_i$, $i = 0, \dots, m$, then*

$$\det M(\xi) = \frac{2^{2m^2}}{(2m+1)^{2m+1}} \prod_{i=1}^{m} (1 - x_i^2)(1 - x_i)^2 \prod_{1 \le i < j \le m} (x_j - x_i)^4. \tag{3.1}$$

PROOF. For any design ξ of the form (2.10) we have

$$\det M(\xi) = \det M_1(\xi) \det M_3(\xi),$$

where the matrices $M_1(\xi), M_3(\xi)$ are defined by (2.14) and the matrix $M_2(\xi)$ is the null-matrix, which follows form the discussion in Section 2. Define the design η_ξ by

$$\eta_\xi = \begin{pmatrix} x_0 & x_1 & \dots & x_m \\ \dfrac{1}{2m+1} & \dfrac{2}{2m+1} & \dots & \dfrac{2}{2m+1} \end{pmatrix},$$

then it is straightforward to see, that

$$M_1(\xi) = \left(\int_{-1}^{1} T_i(x) T_j(x) d\eta_\xi(x) \right)_{i,j=0}^{m}, \tag{3.2}$$

$$M_3(\xi) = \left(\int_{-1}^{1} (1 - x^2) U_i(x) U_j(x) d\eta_\xi(x) \right)_{i,j=0}^{m-1} \tag{3.3}$$

where $T_i(x)$ is the Chebyshev polynomial of the first kind defined in (2.16) and

$$U_i(x) = \frac{\sin((i+1)\arccos x)}{\sin(\arccos x)} \tag{3.4}$$

is the Chebyshev polynomial of the second kind (see Rivlin (1974)). Because $T_i(x)$ is a polynomial of degree i with leading coefficient 2^{i-1}, it follows that $M_1(\xi)$ is essentially a Vandermonde determinant, i.e.

$$\begin{aligned} \det M_1(\xi) &= 2^{m(m-1)} \frac{2^m}{(2m+1)^{m+1}} (\det((x_j^i)_{i=0,\dots,m}^{j=0,\dots,m}))^2 \\ &= \frac{2^{m^2}}{(2m+1)^{m+1}} \prod_{i=1}^{m} (1 - x_i)^2 \prod_{1 \le i < j \le m} (x_j - x_i)^2 \end{aligned}$$

(note that $x_0 = 1$). Note that the support point x_0 of η_ξ has a vanishing contribution to the matrix $M_3(\xi)$ and that the leading coefficient of $U_i(x)$ is 2^i. Therefore we have by similar arguments

$$\det M_3(\xi) = \frac{2^{m^2}}{(2m+1)^m} \prod_{i=1}^{m}(1 - x_i^2) \prod_{1 \le i < j \le m} (x_j - x_i)^2$$

and a combination of these formulas yields (3.1), which proves the assertion of Lemma 3.1. □

We are now studying the function

$$\phi(x, a) = \prod_{i=1}^{m}(1 - x_i^2)(1 - x_i)^2 \prod_{1 \le i < j \le m} (x_j - x_i)^4 \tag{3.5}$$

as a function of the length a of the design space. To this end we note that $x_m = \cos(a)$ and introduce the set

$$T = \{(\tau_1, \dots, \tau_{m-1})^T \mid 0 < \tau < \dots < \tau_{m-1} < 1\} \tag{3.6}$$

$$\mathcal{X} = \{(x_1, \dots, x_{m-1})^T \mid x_i = \cos(a\tau_i), i = 1, \dots, m-1, (\tau_1, \dots, \tau_{m-1})^T \in T\}. \tag{3.7}$$

Note that any design $\xi \in \Xi_a^{(1)}$ of the form (2.10) is uniquely determined by a point $\tau = (\tau_1, \dots, \tau_{m-1})^T \in T$ or its corresponding function $x = (x_1, \dots, x_{m-1})^T \in \mathcal{X}$ by the transformation $t_i = a\tau_i = \arccos x_i$, $i = 1, \dots, m-1$ (note that $t_0 = 0, t_m = a$) and by Lemma 3.1 the determinant of $M(\xi)$ is proportional to the function ϕ given in (3.5). By standard arguments it can now be verified that for fixed $a \in (0, \pi]$ the function ϕ in (3.5) is a strictly concave function of $x = (x_1, \dots, x_{m-1})^T \in \mathcal{X}$. Therefore (for fixed a) the function $\phi(x, a)$ has a unique maximum in $\mathcal{X}$, which will be denoted by $x^*(a)$ (because of its dependence on the length of the design space). The function ϕ is obviously differentiable and $x^*(a)$ can be obtained as the unique solution of the equations

$$\frac{\partial}{\partial x}\phi(x, a) = 0 \in \mathbb{R}^{m-1}. \tag{3.8}$$

Moreover, for any $x \in \mathcal{X}$ the matrix of the second partial derivatives

$$G(x, a) = \left(\frac{\partial^2}{\partial x_i \partial x_j}\phi(x, a)\right)_{i,j=1}^{m-1} \tag{3.9}$$

is positive definite and in particular the matrix

$$J(a) = G(x^*(a), a) \tag{3.10}$$

is positive definite for all $a \in (0, \pi]$. It therefore follows from the implicit function theorem (see Gunning and Rossi (1965)) that the function

$$x^* : \begin{cases} (0, \pi] & \rightarrow & \mathcal{X} \\ a & \rightarrow & x^*(a) \end{cases} \tag{3.11}$$

defined as the solution of the equation (3.8) is real analytic. In other words: for any point $a_0 \in (0, \pi]$ there exists a neighbourhood U_0 of a_0, such that the function $x^*|_{U_0}$ can be expanded in a convergent Taylor series. Observing the symmetry $\phi(x,a) = \phi(x,-a)$, it therefore follows that the function

$$\tau^* : \begin{cases} [-\pi,\pi]\backslash\{0\} \to T \\ a \to \tau^*(a) = \left(\dfrac{\arccos x_1^*(|a|)}{a}, \dots, \dfrac{\arccos x_{m-1}^*(|a|)}{a}\right)^T \end{cases} \tag{3.12}$$

is also real analytic. The following result shows that the function τ^* can be extended to a real analytic function on the full circle $[-\pi,\pi]$.

Lemma 3.2. *The function τ^* defined by* (3.12) *can be extended to a real analytic function on the interval $[-\pi,\pi]$, where*

$$\tau^*(0) = \lim_{a\to 0} \tau(a) = (\tau_1^*, \dots, \tau_{m-1}^*)^T,$$

$\tau_1^* < \cdots < \tau_{m-1}^*$ are the positive roots of the polynomial

$$P_{m-1}^{(1,1/2)}(2x^2-1) = \frac{1}{2x} P_{2m-1}^{(1,1)}(x) = \frac{1}{(2m+1)x} P_{2m}'(x)$$

and $P_i^{(\alpha,\beta)}(x)$ denotes the i-th Jacobi polynomial orthogonal with respect to the measure $(1-x)^\alpha(1+x)^\beta dx$ and $P_{2m}(x)$ is the $2m$-th Legendre polynomial orthogonal with respect to the Lebesgue measure on the interval $[-1,1]$.

Proof. The assertion of Lemma 3.2 follows if we prove the existence of $\lim_{a\to 0} \tau^*(a)$ and the claimed form of its components. Let $x_\tau = (\cos(a\tau_1), \dots, \cos(a\tau_{m-1}))^T$, then the expansions $\sin t = t + o(t)$, $\cos t = 1 - t^2/2 + o(t^2)$ show that for $a \to 0$

$$\phi(x_\tau, a) = \frac{a^{2m(2m+1)}}{2^{2m^2}} \prod_{i=1}^m \tau_i^6 \prod_{1\le i<j\le m} (\tau_i^2 - \tau_j^2)^4 (1 + o(a))$$

$(\tau_m = 1)$ and consequently, the limit $\lim_{a\to 0} \tau^*(a)$ exists and can be obtained by maximizing the function

$$\bar\phi(\tau) = \prod_{i=1}^m \tau_i^3 (1-\tau_i^2)^2 \prod_{1\le i<j\le m-1} (\tau_i^2 - \tau_j^2)^2 \tag{3.13}$$

over the set T defined in (3.6). Note that standard arguments show the strict concavity of the function $\bar\phi$ and consequently, the point $\tau^* = (\tau_1^*, \dots, \tau_{m-1}^*)^T$ where the maximum is obtained is unique. Taking partial derivatives of the logarithm of $\bar\phi$ yields the system

$$\frac{3}{\tau_i} + \frac{4\tau_i}{\tau_i^2 - 1} + \sum_{j=1, j\ne i}^{m-1} \frac{4\tau_i}{\tau_i^2 - \tau_j^2} = 0, \qquad i = 1, \dots, m-1 \tag{3.14}$$

and substituting $\tau_i^2 = y_i \in (0,1)$ gives

$$\frac{3}{y_i} + \frac{4}{y_i - 1} + \sum_{j=1, j\ne i}^{m-1} \frac{4}{y_i - y_j} = 0, \qquad i = 1, \dots, m-1. \tag{3.15}$$

Table 1. Values of the components $\tau_1^*(0), \ldots, \tau_{m-1}^*(0)$ of the vector $\tau^*(0)$ defined in Lemma 3.2 and the polynomial solution of the differential equation (3.16) for various values of m.

m	$\psi(y)$ and $\tau_j(0)$
2	$\psi(y) = y - 3/7$
	$\tau_1^*(0) = \sqrt{3/7} \approx 0.6546$
3	$\psi(y) = y^2 - 10/11y + 5/33$
	$\tau_1^*(0) \approx 0.4688, \tau_2^*(0) \approx 0.8302$
4	$\psi(y) = y^3 - 7/5y^2 + 7/13y - 7/143$
	$\tau_1^*(0) \approx 0.3631, \tau_2^*(0) \approx 0.6772, \tau_3^*(0) \approx 0.8998$
5	$\psi(y) = y^4 - 36/19y^3 + 378/323y^2 - 84/323y + 63/4199$
	$\tau_1^*(0) \approx 0.2958, \tau_2^*(0) \approx 0.5652, \tau_3^*(0) \approx 0.7845, \tau_4^*(0) \approx 0.9340$

Similar arguments as given in Karlin and Studden (1966) or Fedorov (1972) show that the polynomial $\psi(y) = \prod_{i=1}^{m-1}(y - y_i)$ satisfies the differential equation

$$y(1-y)\psi''(y) + (3/2 - 7/2y)\psi'(y) + (m-1)(m+3/2)\psi(y) = 0. \tag{3.16}$$

It is well known (see e.g. Szegö (1975), Section 4.21) that the unique polynomial solution of this differential equation is given by the polynomial

$$P_{m-1}^{(1/2,1)}(1-2y)$$

and the assertion of the lemma now follows from transformation $y = \tau^2$ and the equation $P_n^{(\alpha,\beta)}(-x) = (-1)^n P_{m-1}^{(\beta,\alpha)}(x)$ (see Szegö (1975), formula (4.1.3)). The alternative representations of the polynomial $P_{m-1}^{(1,1/2)}(2x^2-1)$ are a consequence of $P_n^{(0,0)}(x) = P_n(x)$ and Theorem 4.1 in Szegö (1975). □

Table 1 shows the polynomial $P_{m-1}^{(1,1/2)}(2y-1)$ (normalized such that the leading coefficient is 1) and the corresponding values $\tau_i^* = \sqrt{y_i}$ for lower degrees $m = 2, 3, 4, 5$. The following result shows that for small designs space, i.e. $a \le \pi(1 - 1/(2m+1))$, the solution of the optimal design problem can be obtained by a Taylor expansion of the function τ^* in (3.12) at the point $a = 0$, where the i-th component $\tau_i^*(0)$ of the vector $\tau^*(0)$ is the i-th positive root of the polynomial $P_{m-1}^{(1,1/2)}(2x^2-1)$.

THEOREM 3.1. *Consider the trigonometric regression model* (1.1) *with design space* $[-a, a]$, *where* $0 < a \le \pi$.

(i) *If* $a \ge \pi(1 - 1/(2m+1))$, *then the design* ξ_a^* *with equal masses at the* $2m+1$ *points*

$$t_i^* = 2\pi \frac{i-1-m}{2m+1}, \quad i = 1, \ldots, 2m+1 \tag{3.17}$$

is a D-optimal design.

(ii) *If* $a < \pi(1 - 1/(2m+1))$, *the D-optimal design is unique and of the form*

$$(3.18)\quad \xi_a^* = \begin{pmatrix} -a & -a\tau_{m-1}^*(a) & \cdots & -a\tau_1^*(a) & 0 & a\tau_1^*(a) & \cdots & a\tau_{m-1}^*(a) & a \\ \frac{1}{2m+1} & \frac{1}{2m+1} & \cdots & \frac{1}{2m+1} & \frac{1}{2m+1} & \frac{1}{2m+1} & \cdots & \frac{1}{2m+1} & \frac{1}{2m+1} \end{pmatrix}$$

where τ^* *is a real analytic function on the interval* $[-\pi,\pi]$ *defined by* (3.12) *and Lemma* 3.2.

PROOF. Recall the definition of the set $\Xi_a^{(2)}$ in (2.11) and assume that the design $\xi^* \in \Xi_a^{(2)}$ is D-optimal for the trigonometric regression model (1.1) on the interval $[-a,a]$. Because $d(t,\xi^*) = 0$ for all $t \in [-a,a]$ it follows from the Chebyshev property of the functions $\{1, \sin t, \cos t, \dots, \sin mt, \cos mt\}$ that the directional derivative $d(t,\xi^*)$ also vanishes on the full circle $[-\pi,\pi]$ (see Karlin and Studden (1966), p. 20). Consequently, ξ^* is also D-optimal for the trigonometric regression on the interval $[-\pi,\pi]$, which implies (by the uniqueness of the D-optimal information matrix) $M(\xi^*) = \operatorname{diag}(1, 1/2, \dots, 1/2)$, $\det M(\xi^*) = 2^{-2m}$. On the other hand we have

$$\lim_{a\to 0} \max_{\xi} \det M(\xi) = 0,$$

and consequently for sufficiently small a the D-optimal design cannot be an element of the set $\Xi_a^{(2)}$. From Lemma 2.2 it follows that the D-optimal design must belong to the set $\Xi_a^{(1)}$ and the discussion in the first part of this section shows that for sufficiently small a the D-optimal design is unique and of the form (3.18). Now let ξ_a^* denote the design defined by (3.18) and

$$(3.19)\quad \begin{aligned} a^* &= \sup\{a \in (0,\pi] \mid \xi_a^* \text{ is } D\text{-optimal}\} \\ &= \sup\{a \in (0,\pi] \mid \det M(\xi^*) < 2^{-2m}\} \end{aligned}$$

(note that the second equality follows by continuity and Lemma 2.2). It is well known (see Fedorov (1972) or Pukelsheim (1993)) that the uniform distribution ξ_u at the $2m+1$ points defined by (3.17) is D-optimal for the trigonometric regression model on the interval $[-\pi,\pi]$. If $\hat{a} = \pi(1 - 1/(2m+1))$ denotes the largest support point of this design, then it follows that $\xi_{\hat{a}}^* = \xi_u$. Consequently, the design $\xi_{\hat{a}}^*$ specified in part (i) of Theorem 3.1 is also D-optimal for the trigonometric regression on the interval $[-\hat{a},\hat{a}]$ and the D-optimality of $\xi_{\hat{a}}^*$ on $[-\pi,\pi]$ shows

$$\xi_{\hat{a}}^* \in \Xi_{\hat{a}}^{(1)} \cap \Xi_{\hat{a}}^{(2)},$$

which implies for the critical bound in (3.18) the inequality $a^* \leq \hat{a}$. Now for any design of the form

$$(3.20)\quad \xi = \xi(a) = \begin{pmatrix} -t_m & \cdots & -t_1 & t_0 & t_1 & \cdots & t_m \\ \frac{1}{2m+1} & \cdots & \frac{1}{2m+1} & \frac{1}{2m+1} & \frac{1}{2m+1} & \cdots & \frac{1}{2m+1} \end{pmatrix}$$

with $0 < t_1 < \cdots < t_m \leq \pi$ it follows from Lemma 3.1 that

$$\det M(\xi) = C \prod_{i=1}^{m} (1 - x_i^2)(1 - x_i)^2 \prod_{1 \leq i < j \leq m} (x_j - x_i)^4 =: h(x_\xi)$$

with $C = 2^{2m^2}/(2m+1)^{2m+1}$, $x_\xi = (x_1, \ldots, x_m)^T$, $x_i = \cos t_i$ $(i = 1, \ldots, m)$. The discussion at the beginning of this section shows that h is strictly concave. Additionally, we have for the design $\xi_{\hat{a}}^*$, $h(x_{\xi_{\hat{a}}^*}) = 2^{-2m}$ and for any other design ξ of the form (3.20) $h(x_\xi) < 2^{-2m}$ (because otherwise a convex combination of $\xi_{\hat{a}}^*$ and ξ_a would have an information matrix with a determinant larger than 2^{-2m}, which is impossible). Consequently, because ξ_u^* is of the form (3.20) it follows for the quantity a^* defined by (3.19) that $a^* = \hat{a}$.

If $a \geq \hat{a}$, the discussion of this proof shows that the design specified by part (i) of Theorem 3.1 is D-optimal. If $a < \hat{a}$, the definition (3.19) shows that the D-optimal design is in the set $\Xi_a^{(1)}$ and Lemmas 3.1 and 3.2 (with their corresponding proofs) imply that the D-optimal design for the trigonometric regression on the interval $[-a, a]$ is of the form (3.18), which completes the proof of the theorem. □

Note that Theorem 3.1 provides a complete solution of the D-optimal design problem. In the case (i) with $a \geq \pi(1 - 1/(2m+1))$ a D-optimal design for the trigonometric regression model (1.1) on the interval $[-a, a]$ is explicitly given by the uniform distribution at the support points specified by (3.17), but is not necessarily unique. If $a < \pi(1 - 1/(2m+1))$ the D-optimal design is unique and specified by (3.18), where the vector $\tau^*(a) = (\tau_1^*(a), \ldots, \tau_{m-1}^*(a))^T$ can be obtained by means of a Taylor expansion at the point $a = 0$

$$\tau^*(a) = \sum_{i=0}^{\infty} \tau_{(i)}^* a^i \tag{3.21}$$

and the vector $\tau_{(0)}^* = \tau^*(0)$ is given in Lemma 3.2. It is shown in Dette *et al.* (2002) that the coefficients in the above expansion can be calculated by the recursive relations

$$\tau_{(s+1)}^* = -\frac{1}{(s+1)!} J^{-1}(0) \left(\frac{d}{da}\right)^{s+1} g(\tau_{<s>}^*(a), a)|_{a=0}$$

$s = 0, 1, 2, \ldots$, where

$$\tau_{<s>}^*(a) = \sum_{i=0}^{s} \tau_{(i)}^* a^i$$

denotes the Taylor polynomial of degree $s \in \{0, 1, 2, \ldots\}$,

$$J(0) = \left(\frac{\partial^2}{\partial \tau_i \partial \tau_j} \phi(x_\tau, a) \right)_{i,j=1}^{m-1} \Bigg|_{\tau=\tau^*(0)} \quad \text{and}$$

$$g(\tau, a) = \frac{\partial}{\partial \tau} \phi(x_\tau, a) \in \mathbb{R}^{m-1}.$$

Note that in general an exact determination of the radius of convergence for the Taylor expansion (3.21) seems to be intractable. In general several re- expansions could be needed to obtain the D-optimal design for any $a \in (0, \pi(1 - 1/(2m+1)))$. However, our numerical calculations in the following section indicate that only one expansion at the point $a = 0$ is required to obtain the D-optimal design for the trigonometric regression model (1.1) on the interval $[-a, a]$ for any $a \in (0, \pi(1 - 1/(2m+1)))$.

Remark 3.1. As pointed out by a referee it might be of interest to obtain similar results for multidimensional models. Unfortunately, it seems to be difficult to obtain such results, because in the multidimensional case the system of regression functions does not satisfy any Chebyshev properties. For interesting work on optimal designs in multidimensional models on the complete circle $(-\pi, \pi]$ we refer to Riccomagno *et al.* (1997) and Dette (1998).

4. Examples

Example 4.1. Our first example considers the linear trigonometric regression model ($m = 1$) on the interval $[-a, a]$, for which the solution is rather obvious. If $a \geq 2\pi/3$, the design

$$\xi_a^* = \begin{pmatrix} -\frac{2\pi}{3} & 0 & \frac{2\pi}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{pmatrix}$$

is D-optimal, while for $a < 2\pi/3$ the D-optimal design for the linear trigonometric regression model on the interval $[-a, a]$ is given by

$$\xi_a^* = \begin{pmatrix} -a & 0 & a \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{pmatrix}.$$

This follows directly from Theorem 3.1. For A-and E-optimal designs in this model see Wu (2002).

Example 4.2. In the quadratic regression model the situation is more complicated. If $a \geq 4\pi/5$, then part (i) of Theorem 3.1 shows that the design

$$\xi_a^* = \begin{pmatrix} -\frac{4\pi}{5} & -\frac{2\pi}{5} & 0 & \frac{2\pi}{5} & \frac{4\pi}{5} \\ \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} \end{pmatrix}$$

is D-optimal. If $a < 4\pi/5$, the D-optimal design can be obtained by means of a Taylor expansion as indicated in the second part of Theorem 3.1. However, in this particular case an explicit solution is possible by a careful inspection of the arguments given in Section 3. Part (ii) of Theorem 3.1 shows that the D-optimal design in the quadratic trigonometric regression model is in the set $\Xi_a^{(1)}$, whenever $a < 4\pi/5$ and consequently only one support point $t_1^* = t_1^*(a)$ has to be determined. This can be done by a direct differentiation of the function $\phi(x, a)$ in (3.5). Note that $m = 2$, $x_2 = \cos a$ and therefore

$\phi(x, a)$ is a function of only one variable, say $x_1 \in (-1, 1)$. Elementary calculus yields that the derivative of ϕ has zeros at the points $x_1 = \cos a$, $x_2 = 1$ and

$$x_{3,4} = \frac{1}{8}[2\cos(a) - 1 \mp \sqrt{33 + 12\cos(a) + 4\cos(a)^2}].$$

It is easy to see that only one of these two points yields to a solution in the interval $[\cos a, 1]$ and consequently the D-optimal design for the quadratic trigonometric regression model on the interval $[-a, a]$ with $0 < a \leq 4\pi/5$ is given by

$$\xi_a^* = \begin{pmatrix} -a & -t_1^*(a) & 0 & t_1^*(a) & a \\ \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} \end{pmatrix}$$

where

$$t_1^*(a) = \arccos\left(\frac{1}{8}[2\cos(a) - 1 + \sqrt{33 + 12\cos(a) + 4\cos(a)^2}]\right).$$

Example 4.3. In the general case $m \geq 3$ the second part of Theorem 3.1 has to be applied if $a \leq \pi(1 - 1/(2m+1))$ (note that in the remaining case a D-optimal design is explicitly given in part (i) of Theorem 3.1). From Table 1 we obtain the values of $\tau_i^*(0)$, $i = 1, \ldots, m-1$ (provided $m \leq 5$) and the nontrivial support points $\tau_i^*(a)$ for $0 < a < \pi(1 - 1/(2m+1))$ can now be calculated by means of a Taylor expansion as indicated at the end of Section 3. Table 2 shows the values of the first coefficients in the expansion

$$\tau_i^*(a) = \sum_{l=0}^{\infty} \tau_{i(l)}^* \left(\frac{a}{\pi}\right)^l, \qquad i = 1, \ldots, m-1 \tag{4.1}$$

for $m = 2, 3, 4, 5$. It can easily be shown that $\tau_i^*(a)$ is an even function of the parameter a and consequently the odd coefficients vanish and only the even coefficients are displayed.

Table 2. Coefficients in the expansion (4.1). The D-optimal design in the trigonometric regression model (1.1) on the interval $[-a, a]$ with $0 < a < \pi(1 - 1/(2m+1))$ has equal masses at the points $-a, -t_{m-1}, \ldots, -t_1, 0, t_1, \ldots, t_{m-1}, a$, where $t_i = a\tau_i^*(a)$, $i = 1, \ldots, m-1$.

	i	0	2	4	6	8	10
$m = 2$	$\tau_{1\,(i)}^*$	.65465	−.21977	−.07747	.04852	.06118	−.02116
$m = 3$	$\tau_{1\,(i)}^*$	.46885	−.19145	−.00875	.02584	−.00184	−.00283
	$\tau_{2\,(i)}^*$	.83022	−.13502	−.10286	−.05465	−.00161	.03946
$m = 4$	$\tau_{1\,(i)}^*$	.36312	−.15556	.00820	.01117	−.00368	−.00011
	$\tau_{2\,(i)}^*$	.67719	−.18093	−.07349	.00094	.02393	.01100
	$\tau_{3\,(i)}^*$	.89976	−.08456	−.07603	−.06025	−.03806	−.01256
$m = 5$	$\tau_{1\,(i)}^*$	.29576	−.12851	.01204	.00501	−.00238	.00036
	$\tau_{2\,(i)}^*$	.56524	−.18316	−.03971	.01585	.01178	−.00245
	$\tau_{3\,(i)}^*$	.78448	−.14366	−.08805	−.03360	.00483	.01980
	$\tau_{4\,(i)}^*$	.93400	−.05677	−.05431	−.04874	−.03965	−.02762

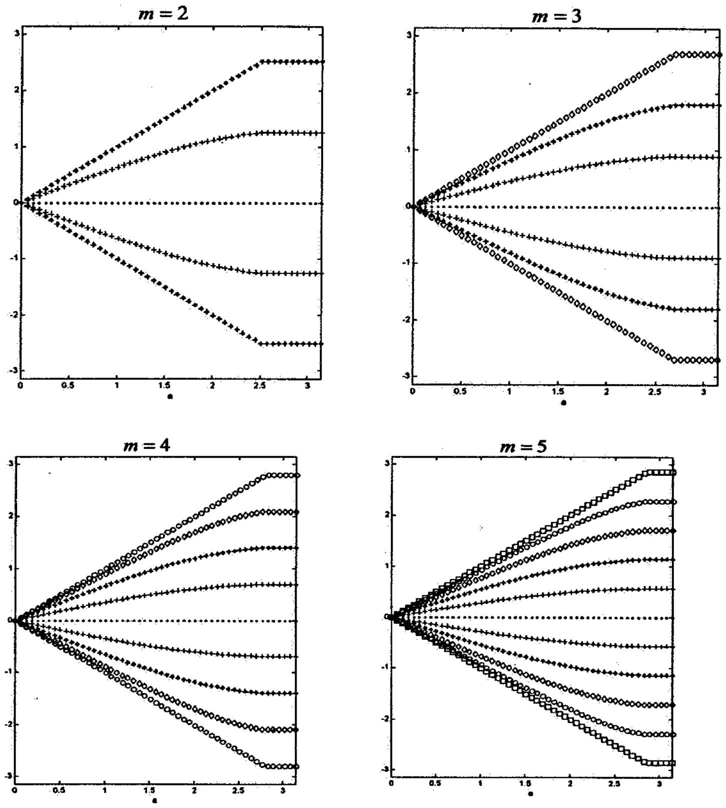

Fig. 1. The support points of the D-optimal design in the trigonometric regression model (1.1) on the interval $[-a, a]$ as a function of the parameter a for various degrees m. The D-optimal design has equal masses at these points.

Consider as a concrete example the case $m = 3$. If $a \geq 6\pi/7$ a D-optimal design for the cubic trigonometric regression model on the interval $[-a, a]$ is given by part (i) of Theorem 3.1, i.e.

$$\xi_a^* = \begin{pmatrix} -\frac{6\pi}{7} & -\frac{4\pi}{7} & -\frac{2\pi}{7} & 0 & \frac{2\pi}{7} & \frac{4\pi}{7} & \frac{6\pi}{7} \\ \frac{1}{7} & \frac{1}{7} & \frac{1}{7} & \frac{1}{7} & \frac{1}{7} & \frac{1}{7} & \frac{1}{7} \end{pmatrix}.$$

If $0 < a < 6\pi/7$ the D-optimal design can be calculated from the expansion (4.1) and Table 2. For example if $a = 1$ we obtain that the D-optimal design for the cubic trigonometric regression model on the interval $[-1, 1]$ is given by

$$\xi_a^* = \begin{pmatrix} -1 & -0.8154 & -0.4494 & 0 & 0.4494 & 0.8154 & 1 \\ \frac{1}{7} & \frac{1}{7} & \frac{1}{7} & \frac{1}{7} & \frac{1}{7} & \frac{1}{7} & \frac{1}{7} \end{pmatrix}.$$

Figure 1 shows the support points of D-optimal designs as a function of the length a of the design space for $m = 2, 3, 4, 5$. The support points have been determined by a Taylor expansion as indicated in Section 3 and the D-optimal design puts equal masses at these points.

Acknowledgements

The work of V. B. Melas and A. N. Pepelyshev was partially supported by the Russian Foundation of Basic Reseach (grant No. 00-01-00495). The work of H. Dette was supported by the Deutsche Forschungsgemeinschaft (SFB 475: Komplexitätsreduktion in multivariaten Datenstrukturen). The authors are grateful to S. Biedermann and to two unknown referees for some very helpful comments on an earlier version of this paper.

References

Dette, H. (1998). Some applications of canonical moments in Fourier regression models, *New Developments and Applications in Experimental Design* (eds. N. Flournoy, W. F. Rosenberger and W. K. Wong), IMS Lecture Notes Monogr. Ser., **34**, 175–185, Hayward, California.

Dette, H. and Haller, G. (1998). Optimal designs for the identification of the order of a Fourier regression, *Ann. Statist.*, **26**, 1496–1521.

Dette, H. and Melas, V. B. (2003). Optimal designs for estimating individual coefficients in Fourier regression models, *Ann. Statist.* (to appear), http://www.ruhr-uni-bochum.de/mathematik3/preprint.htm.

Dette, H. Melas, V. B. and Pepelyshev, A. (2002). Optimal designs for estimating individual coefficients–A functional approach, *J. Statist. Plann. Inference* (to appear), http://www.ruhr-uni- bochum.de/mathematik3/preprint.htm.

Fedorov, V. V. (1972). *Theory of Optimal Experiments*, Academic Press, New York.

Graybill, F. A. (1976). *Theory and Application of the Linear Model*, Wadsworth, Belmont, California.

Gunning, R. C. and Rossi, H. (1965). *Analytical Functions of Several Complex Variables*, Prentice Hall, New York.

Hill, P. D. H. (1978). A note on the equivalence of D-optimal design measures for three rival linear models, *Biometrika*, **65**, 666-667.

Hoel, P. (1965). Minimax design in two-dimensional regression, *Ann. Math. Statist.*, **36**, 1097–1106.

Karlin, S. and Studden, W. J. (1966). *Tchebycheff Systems: With Applications in Analysis and Statistics*, Interscience, New York.

Kiefer, J. C. (1974). General equivalence theory for optimum designs (approximate theory), *Ann. Statist.*, **2**, 849–879.

Kitsos, C. P., Titterington, D. M. and Torsney, B. (1988). An optimal design problem in rhythmometry, *Biometrics*, **44**, 657–671.

Lau, T. S. and Studden, W. J. (1985). Optimal designs for trigonometric and polynomial regression, *Ann. Statist.*, **13**, 383–394.

Mardia, K. (1972). *The statistics of directional data*, Academic Press, New York.

Melas, V. B. (1978). Optimal designs for exponential regression, *Mathematische Operations for schung und Statistik*, *Series Statistics*, **9**, 45–59.

Pukelsheim, F. (1993). *Optimal Design of Experiments*, Wiley, New York.

Riccomagno, E., Schwabe, R. and Wynn, H. P. (1997). Lattice-based D-optimum design for Fourier regression, *Ann. Statist.*, **25**, 2313–2327.

Rivlin, T. J. (1974). *Chebyshev Polynomials*, Wiley, New York.

Silvey, S. D. (1980). *Optimal Design*, Chapman and Hall, London.

Szegö, G. (1975). Orthogonal polynomials, *Amer. Math. Soc. Colloqu. Publ.*, **23**, Providence, Rhode Island.

Wu, H. (2002). Optimal designs for first order trigonometric regression on a partial circle, *Statistica Sinica*, **12**, 917–930.

Acknowledgement

Annals of the Institute of Statistical Mathematics acknowledges the following individuals who served as referees from October 2001 to September 2002.

Ahmad, Khalaf El-Dabe
Ahsanullah, Mohammad
Aki, Sigeo
Amari, Shun-ichi
Anderson, Jon E.
Antzoulakos, Demetrios L.
Aoki, Satoshi
Asadi, Majid
Azzalini, Adelchi
Babu, Gutti Jogesh
Bacry, Emmanuel
Barbour, Andrew D.
Beran, Rudolf J.
Berger, James O.
Bhattacharya, Bhaskar
Bilias, Yannis
Bilker, Warren B.
Brown, Bruce M.
Chanda, Kamal C.
Chaubey, Yogendra P.
Chen, Robert W.
Chen, Rong
Chen, Song Xi
Chen, Zhiqiang
Chernick, Michael R.
Chhikara, Raj S.
Chiaromonte, Francesca
Chib, Siddhartha
Choi, In
Cline, Daren B. H.
Csörgö, Miklós
del Castillo, Joan
Delgado, Miguel A.
Denker, Manfred
Deshpande, Jayant V.
Dey, Dipak K.
Dharmadhikari, Sudhakar W.
Doucet, Arnaud
Drmota, Michael
Dryden, Ian L.
Ebrahimi, Nader B.
Eilers, Paul H. C.
Einmahl, John H. J.
Émery, Michel
Eubank, Randall L.
Falk, Michael
Farin, Gerald E.
Farrington, Paddy
Feng, Yuanhua
Ferentinos, Kosmas
Fernández, Miguel A.
Fernholz, Luisa T.
Fourdrinier, Dominique
Fraser, Donald A. S.
Frieden, B. Roy
Fu, James C.
Fujikoshi, Yasunori
Gangopadhyay, Ashis K.
Gastwirth, Joseph L.
Gelman, Andrew
Ghorai, Jugal K.
Ghosal, Subhashis
Girolami, M.
Goldstein, Larry
Gombay, Edit
Gu, Chong
Györfi, László
Hartigan, John A.
Hayashi, Masato
Hayter, Anthony J.
Hazelton, Martin L.
Hendi, M. I.
Henna, Jōgi
Herbrich, R.
Herschkorn, Stephen J.
Higuchi, Tomoyuki
Hirano, Katuomi
Ho, Hwai-Chung
Honda, Toshio
Horowitz, Joel L.
Hosoya, Yuzo
Hu, Chin-Yuan
Hu, Taizhong
Hu, Xiaoqiong Joan

Huang, Deng Yuan
Hušková, Marie
Ishiguro, Makio
Jammalamadaka, S. Rao
Jarre, Florian
Jensen, Jens Ledet
Joe, Harry
Kagan, Abram
Kakizawa, Yoshihide
Kao, Chih-Hwa Duke
Kawanabe, Motoaki
Keiding, Niels
Keribin, Christine
Klebanov, Lev B.
Klein, John P.
Koehler, Kenneth J.
Kokoszka, Piotr S.
Kolaczyk, Eric D.
Komaki, Fumiyasu
Konishi, Sadanori
Kou, Samuel
Koul, Hira Lal
Koutras, Markos V.
Kubokawa, Tatsuya
Kuboki, Hisataka
Kushary, Debashis
Lahiri, Soumendra Nath
LaRiccia, Vincent N.
Lee, Youngjo
Lengyel, Tamas
Leśkow, Jacek
Liang, Hua
Liang, TaChen
Liese, Friedrich
Lin, Chih-Jen
Lin, Gwo Dong
Luceño, Alberto
Lugosi, Gabor
Lund, Robert B.
Lunn, Mary
Mammen, Enno
Maronna, Ricardo Antonio
Martin, Michael A.
Masry, Elias
McCormick, William P.
McCullagh, Peter
McCulloch, Charles E.
Mckean, Joseph W.
McKenzie, Eddie
Melard, Guy
Menéndez, José A.
Meng, Xiao-Li
Mitra, Murari
Miwa, Tetsuhisa
Móri, Tamás F.
Mukhopadhyay, Nitis
Muramatsu, Masakazu
Nagao, Hisao
Neal, Radford
Nelson, Paul I.
Neumann, Michael H.
Ng, Pin T.
Nikulin, Mikhail S.
Nobel, Andrew B.
Novikov, Alex
Padgett, William J.
Pakes, Anthony G.
Papadatos, Nickos
Pawlitschko, J.
Peligrad, Magda
Peña, Edsel A.
Pensky, Marianna
Penzer, Jeremy
Petruccelli, Joseph D.
Politis, Dimitris N.
Powell, James L.
Puig, Pedro
Qaqish, Bahjat F.
Qin, Jing
Rachev, Svetlozar T.
Raqab, Mohammad Z.
Ritov, Ya'acov
Rojo, Javier
Romano, Joseph P.
Ronchetti, Elvezio M.
Rossini, Anthony J.
Rueda Sabater, Cristina
Sakata, Toshio
Sato, Seisho
Scarsini, Marco
Schott, James R.
Sen, Kanwar
Serfling, Robert J.
Seshadri, Vanamamalai
Shaked, Moshe
Shao, Jun
Shen, Xiaotong
Shi, Ning Zhong
Shimizu, Ryoichi
Sibuya, Masaaki

Silvapulle, Mervyn J.
Siotani, Minoru
Skovgaard, Ib Michael
Small, Christopher G.
Srivastava, Muni S.
Stefanov, Valery
Steutel, F. W.
Stockis, Jean-Pierre
Stone, Charles J.
Strawderman, William E.
Sturmfels, Bernd
Stute, Winfried
Sundberg, Rolf
Suzukawa, Akio
Szarek, Stanislaw J.
Takada, Yoshikazu
Takemura, Akimichi
Tanaka, Yutaka
Tawn, Jonathan
Teugels, Jozef L.
Thavaneswaran, A.
Thompson, Robin
Tipping, Michael E.
Tran, Lanh Tat
Tsao, Min
Tyler, David E.
Van der Laan, Mark
van Dyk, David A.
Van Houwelingen, Hans C.
Vannucci, Marina
Vere-Jones, David
von Rosen, Dietrich
Vos, Paul W.
Wago, Hajime
Wang, Liqun
Wang, Mei Cheng
Warrack, Giles
Wei, William W. S.
Weiss, Robert Erin
Woodroofe, Michael B.
Wu, Colin O.
Wu, Huaiqing
Xia, Yingcun
Yajima, Yoshihiro
Ying, Zhiliang
Yohai, Victor J.
Yoshida, Nakahiro
Yoshihara, Ken-ichi
Yum, Bong-Jin
Zacks, Shelemyahu
Zhang, Biao
Zhang, Shuanglin
Zheng, Zukang