Biometrika (2000), **87**, 1, *pp*. 157–171

Printed in Great Britain

Nonparametric estimation of weighted proportions with application to particle size measurement

By ALEXANDER GOLDENSHLUGER AND DAVID FARAGGI
Department of Statistics, University of Haifa, 31905 Haifa, Israel
goldensh@rstat.haifa.ac.il faraggi@stat.haifa.ac.il

Summary

Weighted proportions arise naturally in particle size measurement where one is interested in estimating the proportion of the total volume or surface corresponding to specific particle diameters. The particles are assumed to be spherical, and observations consist of the particle diameters. We consider the weighted proportion as a nonparametric functional of the diameter distribution, and investigate how well this functional can be estimated. It is shown that the commonly used estimator is asymptotically efficient, but it is very sensitive to sample fluctuations. We propose a modified estimator which is also asymptotically efficient and has better small-sample properties. Numerical experiments with real datasets demonstrate the improved practical behaviour of the proposed estimators.

Some key words: Asymptotic efficiency; Nonparametric functional; Particle size distribution; Robustness; Shrinkage.

1. Introduction

Let $X_1, \ldots, X_n$ be independent identically distributed positive random variables with common distribution function F. Let $\mathscr{M}$ denote the set of all probability distributions defined on measurable subsets of the positive real line $\mathbb{R}_+$. In this paper we deal with estimating the functional $T(F): \mathscr{M} \to [0, 1]$ given by

$$T(F) = \left\{ \int_0^\infty w(x)\, dF(x) \right\}^{-1} \int_\Delta w(x)\, dF(x), \tag{1}$$

where Δ is a fixed subset of $\mathbb{R}_+$, and w is a positive monotone increasing unbounded function on $\mathbb{R}_+$. We call $T(F)$ the weighted proportion, since it measures a relative mass assigned to Δ by the probability distribution F.

Estimating the functional $T(F)$ is motivated by a real life problem that arises in particle size measurement. Suppose that a population of particles of a spherical shape is dispersed through a medium. The population is sampled, and diameters of the particles are recorded. Our goal is to estimate the proportion of the total volume or surface corresponding to specific particle diameters. This is a standard problem of conversion from number to volume or surface in the particle size measurement (Allen, 1990, p. 145). If x is the particle diameter and $w(x) = x^3$ then $T(F)$ is the fraction by volume of particles with diameters in Δ. In practice the raw observations are collected in the form of number distributions and are summarised by histograms. Then the standard estimation method is to weight the estimated bin probabilities by the corresponding cubed or squared diameters (Allen, 1990,

p. 145). For our model this method amounts to substituting the empirical probability distribution $F_n(x)=n^{-1}\sum_{i=1}^{n} I(X_i \leqslant x)$ instead of F into (1). It turns out, however, that this simple and obvious estimator does not perform well in many practical situations. The main drawback is that the weighted proportions are very sensitive to small deviations from the underlying distribution F. As a result, using the estimator $T(F_n)$ for $T(F)$ leads to poor repeatability of experiments, a feature which caused problems in applications and motivated this paper.

A parametric approach to model particle size distributions was proposed by Barndorff-Nielsen (1977). It is shown there that the parametric family of continuous distributions such that the logarithm of the density is a hyperbola is an appropriate family for modelling particle size distributions. In the framework of this approach, $T(F)$ can be estimated using standard parametric techniques; see also Fieller, Flenley & Olbricht (1992). Estimation of the functional $T(F)$ was considered by Nicholson (1970) and Watson (1971) in the context of the Wicksell problem in stereology; for review see e.g. Ripley (1981, Ch. 9). Both Nicholson (1970) and Watson (1971) assumed that the observations consist of intersections of a planar or linear probe with the particles. In this case the probability distribution F relates to the distribution of observations through a certain integral equation. Watson (1971) argues that some standard and natural estimators have poor statistical properties for this problem, and warn against their use. It will be shown here that, even with direct observations from F, the standard estimator $T(F_n)$ will be inaccurate for finite samples because of its high sensitivity to sample fluctuations.

In this paper we estimate the weighted proportions in nonparametric fashion and study their properties. First, we investigate the properties of both the functional $T(F)$ and the related standard estimator $T(F_n)$. A lower bound on the estimation accuracy is established, and it is proved that $T(F_n)$ is an asymptotically efficient estimator of $T(F)$. It is shown, however, that $T(F)$ is very sensitive to small deviations from the underlying distribution F. It is important to emphasise that this sensitivity is an intrinsic property of the functional $T(F)$ itself, rather than a sampling property of the estimation method. In other words, the functional $T(F)$ is non-robust; by robustness here we mean stability of the quantity $T(F)$ to be estimated uniformly under small changes in the underlying distribution. In practice this non-robustness results in the poor repeatability of the experiments. If F'_n and F''_n are empirical distributions based on two samples of size n from the same material, then, even though F'_n is close to F''_n, $T(F'_n)$ can be very different from $T(F''_n)$. Thus, small fluctuations in the data lead to large variability in the estimate. To overcome this problem, we propose a modified estimator which is based on more stable functionals. Construction of our estimator uses the idea of downweighting the standard estimator $T(F_n)$ according to the relative accuracy of the estimator F_n for F. Intuitively, since small changes in F_n can lead to high variability in the corresponding estimator $T(F_n)$, one should take into account the accuracy of F_n when estimating F. We propose to shrink $T(F_n)$ towards zero relatively to the estimated accuracy of F_n. Such a strategy introduces a family of more stable nonparametric shrinkage functionals. We prove that the proposed estimator has the same asymptotic normal distribution as $T(F_n)$. We present evidence that the variance of the estimates is reduced for small samples. In addition, the mean squared error can be reduced, especially when the values of $T(F)$ are small.

The outlined approach is in the spirit of 'efficiency-robust' nonparametric estimation, where it is desired to achieve asymptotic efficiency and to improve small-sample properties. This approach was advocated by many authors, e.g. by Johns (1974) for nonparametric estimation of location; see also a review paper by Hogg (1974) and references therein. It is recognised widely that the small-sample properties of the estimator obtained by substituting the empirical distribution can be poor. For example, Fernholz (1997) argues that using the smoothed empirical distribution instead of F_n may improve the small-sample properties of the estimators, if the functional has a discontinuous influence function. The relationship between sensitivity of nonparametric functionals and some standard concepts in robustness theory is discussed by Tibshirani & Wasserman (1988).

The rest of the paper is organised as follows. In § 2 we study the properties of the standard estimator $T(F_n)$. In § 3 we introduce our nonparametric shrinkage estimator and investigate its properties. In § 4 we consider some numerical examples to illustrate the behaviour of the proposed estimator. Section 5 contains a discussion, and proofs are given in the Appendix.

2. The standard estimator

2·1. *Asymptotic results*

The standard estimator $T_n = T(F_n)$ is obtained by substituting the empirical probability distribution F_n into (1):

$$T_n = T(F_n) = \left\{ \int_0^\infty w(x)\, dF_n(x) \right\}^{-1} \int_\Delta w(x)\, dF_n(x) = \left\{ \sum_{i=1}^n w(X_i) \right\}^{-1} \sum_{i:X_i \in \Delta} w(X_i).$$

Recall that, throughout the paper, $X_1, \ldots, X_n$ are independent positive random variables with common distribution F, and $w(.)$ is a positive monotone increasing unbounded function on $\mathbb{R}_+$.

Let $\mathscr{F}_w$ denote the set of all probability distributions F on $\mathbb{R}_+$ such that

$$\int_0^\infty \{w(x)\}^2\, dF(x) < \infty. \tag{2}$$

Theorem 1. *The estimator T_n is consistent at $F : T_n \to T(F)$ in probability and almost surely as $n \to \infty$. Furthermore, if $F \in \mathscr{F}_w$, then $n^{\frac{1}{2}}\{T_n - T(F)\}$ is asymptotically normal with mean* 0 *and variance*

$$\sigma_F^2 = \left(\int_0^\infty w\, dF \right)^{-2} \left\{ (1 - 2T) \int_\Delta w^2\, dF + T^2 \int_0^\infty w^2\, dF \right\}. \tag{3}$$

For the proof see the Appendix.

Theorem 1 is concerned with a fixed underlying distribution $F \in \mathscr{F}_w$. It is desirable to have the above asymptotic properties hold as uniformly in $\mathscr{F}_w$ as possible. It is clear that this can be achieved only on smaller subsets of $\mathscr{F}_w$. For given positive real numbers $c > 0$ and γ, define the following class of distributions:

$$\mathscr{F}_w(c, \gamma) = \left\{ F \in \mathscr{M} : F(c) = 0,\ \int_0^\infty \{w(x)\}^3\, dF(x) \leqslant \gamma < \infty \right\}. \tag{4}$$

Note that the condition $F(c) = 0$ precludes distributions which put a positive mass in a small neighbourhood of zero.

Theorem 2. *The distribution of $n^{\frac{1}{2}}(T_n - T)$ under F converges to the normal distribution with mean* 0 *and variance σ_F^2 uniformly over $\mathscr{F}_w(c, \gamma)$.*

For the proof see the Appendix.

In Theorem A1 in the Appendix we establish a lower bound on the estimation accuracy. Theorem 2 along with Theorem A1 implies that $T_n = T(F_n)$ is asymptotically efficient uniformly over $\mathscr{F}_w(c, \gamma)$. Thus, the standard estimator possesses the best possible large-sample properties and cannot be improved from the viewpoint of efficiency. However, T_n does not perform well for finite samples, being highly sensitive to sample fluctuations. This is caused by non-robustness of the related functional $T(F)$.

2·2. *Non-robustness of* $T(F)$

Since we refer to robustness as stability of the quantity to be estimated uniformly under small changes in the underlying distribution, robustness is an intrinsic property of the functional itself, rather than a sampling property of the estimation method. It turns out that the functional $T(F)$ is not robust according to the above criterion. This will be illustrated by calculating the commonly used measures of quantitative robustness; for the corresponding definitions, see e.g. Huber (1981, Ch. 1) and He & Simpson (1993).

To investigate the sensitivity of $T(F)$, we consider the Huber contamination discrepancy

$$\delta(G, F) \equiv \inf\{\varepsilon : G(A) \geqslant (1-\varepsilon)F(A) \text{ for all measurable sets } A\}.$$

We measure how much $T(F)$ changes when $\delta(G, F) \leqslant \varepsilon$ by the contamination bias

$$b_T(\varepsilon; F) \equiv \sup_{G:\delta(G,\,F)\leqslant\varepsilon} |T(G) - T(F)|.$$

By definition, $b_T(\varepsilon; F) \leqslant 1$ for all $\varepsilon > 0$. In what follows we assume that Δ contains an infinite interval $[a, \infty)$ for some a. This assumption does not restrict generality; otherwise the same bounds on the contamination bias hold with $T(F)$ replaced by $1 - T(F)$. It is evident that even very small mass at infinity yields $T(G) = 1$, and it implies that the lower bound on the contamination bias is

$$b_T(\varepsilon; F) \geqslant 1 - T(F)$$

for all $\varepsilon > 0$. In addition, if $T(F) \leqslant \frac{1}{2}$, then the lower bound is tight: $b_T(\varepsilon; F) = 1 - T(F)$, for all $\varepsilon > 0$. Thus, any changes in the distribution, no matter how small, can influence drastically the quantity to be estimated, especially when $T(F)$ is close to zero. Qualitative robustness of $T(F)$ refers to the continuity of the contamination bias at $\varepsilon = 0$. In our case $b_T(0+; F) \geqslant 1 - T(F)$, so that $T(F)$ is not qualitatively robust with respect to the Huber contamination discrepancy. Since $b_T(\varepsilon; F)$ is not differentiable at zero, the contamination sensitivity equals infinity, i.e.

$$\gamma_T^* \equiv \lim_{\varepsilon\downarrow 0} \sup\{b_T(\varepsilon; F)/\varepsilon\} = \infty;$$

and the breakdown point equals zero:

$$\varepsilon_T^* \equiv \inf\left\{\varepsilon : b_T(\varepsilon; F) = \sup_\beta b_T(\beta; F)\right\} = 0.$$

This shows that the functional $T(F)$ becomes completely uninformative even under very small changes in F. It is straightforward to show that the influence function of the functional $T(F)$ is equal to

$$\textsc{ic}(x, F, T) = \left(\int_0^\infty w\, dF\right)^{-1} w(x)\{I_\Delta(x) - T(F)\}. \tag{5}$$

Thus, the influence function is both unbounded and discontinuous. Unboundedness of $\textsc{ic}(x, F, T)$ leads to infinite gross-error sensitivity:

$$\gamma_T^{(\mathrm{GE})} \equiv \sup_x |\textsc{ic}(x, F, T)| = \infty.$$

The estimator $T_n = T(F_n)$ inherits the non-robust behaviour of the functional $T(F)$. Unboundedness of $\textsc{ic}(x, F, T)$ implies that large observations affect $T(F_n)$ greatly. Discontinuity of $\textsc{ic}(x, F, T)$ implies that, if F_n puts mass on to the discontinuity points, then even small changes in many observations, resulting for example from rounding or grouping, may produce a larger change in $T(F_n)$ (Huber, 1981, p. 9). In practice this leads to high variability of $T(F_n)$ and poor repeatability in experimentation. In § 3 we propose a modified estimator $\tilde{T}_n$ which is based on more stable functionals. In addition, $\tilde{T}_n$ shares the asymptotically efficient properties of $T_n = T(F_n)$ uniformly over the class $\mathscr{F}_w(c, \gamma)$.

3. The shrinkage estimator

3·1. *Introduction*

Small-sample fluctuations have undue influence on $T(F)$. If Δ contains an infinite interval $[a, \infty)$, then $T(F)$ is particularly non-robust when the value of $T(F)$ is small, corresponding to the case where F assigns a small mass to the set Δ. Note that, if Δ does not contain an infinite interval, then the same effect holds; here F assigns a small mass to the set complementary to Δ. Therefore, observations with small probability of occurrence can have undue influence on the standard estimator $T_n = T(F_n)$. It is important to note that, in the region where such observations appear, F_n typically has large standard deviation in comparison to the value to be estimated. Therefore, it seems reasonable to shrink the standard estimator T_n towards zero according to the estimated relative accuracy of F_n. This idea is similar to the approach to robustness discussed by Lindsay (1994, p. 1082), where the degree to which an observation is unreliable depends on both its probability of occurrence and the sample size.

3·2. *Motivating example*

Assume that the observations $X_1, \ldots, X_n$ are grouped into N bins $B_k = [x_k, x_{k+1}]$, for $k = 1, \ldots, N$. This situation corresponds to the multinomial distribution with parameters $(n; f_1, \ldots, f_N)$. The standard estimators $\hat{f}_k$ for the unknown f_k are

$$\hat{f}_k = \frac{1}{n} \sum_{i=1}^{n} I(X_i \in B_k) \quad (k = 1, \ldots, N).$$

We are interested in estimating the weighted proportions

$$t_k = \frac{f_k w_k}{\sum_{j=1}^{N} f_j w_j} \quad (k = 1, \ldots, N), \tag{6}$$

where the weights w_k can be chosen as $w_k = \{(x_k + x_{k+1})/2\}^3$ or $w_k = (x_k x_{k+1})^{3/2}$. In fact, (6) is a discrete version of the functional $T(F)$; we are estimating N functionals corresponding to $\Delta = B_k$, for $k = 1, \ldots, N$. The standard estimator takes the form

$$\hat{t}_k = \frac{\hat{f}_k w_k}{\sum_{j=1}^{N} \hat{f}_j w_j} \quad (k = 1, \ldots, N). \tag{7}$$

Let $\gamma=(\gamma_1,\ldots,\gamma_N)$ and $\gamma_i \in [0,1]$, for $i=1,\ldots,N$. We take $(\gamma_1\hat{f}_1,\ldots,\gamma_N\hat{f}_N)$ as an estimator for $t=(t_1,\ldots,t_N)$. The risk of the estimator under quadratic loss is

$$R(\gamma,f,t)=E\sum_{i=1}^{N}(\gamma_i\hat{f}_i-t_i)^2=\sum_{i=1}^{N}[\gamma_i^2\{f_i^2+n^{-1}f_i(1-f_i)\}-2\gamma_i f_i t_i+t_i^2].$$

Minimising the naive estimator $R(\gamma,\hat{f},\hat{t})$ of the risk over γ, we obtain the shrinkage estimator

$$\tilde{t}_k=\hat{t}_k\frac{\hat{f}_k^2}{\hat{f}_k^2+n^{-1}\hat{f}_k(1-\hat{f}_k)}\quad (k=1,\ldots,N). \tag{8}$$

If $\hat{f}_k$ is an accurate estimator of f_k, then the shrinkage factor is close to one, and $\tilde{t}_k$ *is* similar to $\hat{t}_k$. On the other hand, if the variance of $\hat{f}_k$ is large in comparison with f_k^2, then the shrinkage factor is small, and $\hat{t}_k$ is shrunk towards zero. It is clear that $\tilde{t}_k$ is first-order asymptotically equivalent to $\hat{t}_k$, but, as we will show, it is less sensitive to sample fluctuations. In § 3·3 we will use a similar idea to define a shrinkage estimator for the general nonparametric setting.

3·3. *Construction of shrinkage estimator*

Let $\{Q(F,n)\}$ be a sequence of functionals from $\mathscr{M}$ to $[0,1]$, and consider the shrinkage functionals

$$\tilde{T}=\tilde{T}(F,n)=\{1-Q(F,n)\}T(F).$$

Define the shrinkage estimator $\tilde{T}_n$ for $T(F)$ as

$$\tilde{T}_n=\tilde{T}(F_n,n)=T_n(1-Q_n),\quad Q_n\equiv Q(F_n,n).$$

We have

$$n^{\frac{1}{2}}\{\tilde{T}_n-T(F)\}=n^{\frac{1}{2}}\{T_n-T(F)\}-n^{\frac{1}{2}}T_nQ_n.$$

If the functionals $Q(F,n)$ are chosen in such a way that $n^{\frac{1}{2}}T_nQ_n$ converges to zero in probability as $n\to\infty$ uniformly over $\mathscr{F}_w(c,\gamma)$, then both $\tilde{T}_n$ and T_n are asymptotically efficient in the sense of Theorem 2. However, we will show that the robustness properties will be improved for a particular choice of $Q(F,n)$, defined by

$$Q_\kappa(F,n)=\frac{\kappa n^{-1}\int_\Delta dF(1-\int_\Delta dF)}{(\int_\Delta dF)^2+\kappa n^{-1}\int_\Delta dF(1-\int_\Delta dF)}, \tag{9}$$

where $\kappa \geqslant 0$ is a number to be chosen. The corresponding estimator $\tilde{T}_n$ is given by

$$\tilde{T}_n=\tilde{T}(F_n,n)=T_n\frac{(\int_\Delta dF_n)^2}{(\int_\Delta dF_n)^2+\kappa n^{-1}\int_\Delta dF_n(1-\int_\Delta dF_n)}. \tag{10}$$

Note that $n\int_\Delta dF_n$ is a binomial random variable with parameters n and $\int_\Delta dF$, so that

$$E_F\left(\int_\Delta dF_n\right)=\int_\Delta dF,\quad \mathrm{var}_F\left(\int_\Delta dF_n\right)=n^{-1}\int_\Delta dF\left(1-\int_\Delta dF\right).$$

Consequently, if $\kappa=1$, then the weight $Q_k(F_n,n)$ is the estimator of the relative accuracy of $\int_\Delta dF_n$, that is the ratio between the variance of $\int_\Delta dF_n$ and the expectation of $(\int_\Delta dF_n)^2$. The estimator (10) is similar to the shrinkage estimator (8).

3·4. *Asymptotic efficiency*

The following theorem shows that the shrinkage estimator $\tilde{T}_n$ is asymptotically efficient uniformly over $\mathscr{F}_w(c, \gamma)$, and possesses the same asymptotic properties as T_n.

THEOREM 3. *For every fixed $\kappa \geqslant 0$, the distribution of $n^{\frac{1}{2}}(\tilde{T}_n - T)$ under F converges to the normal distribution with mean 0 and variance σ_F^2 uniformly over $\mathscr{F}_w(c, \gamma)$.*

For the proof see the Appendix.

3·5. *Robustness properties*

As before, we assume that Δ contains an infinite interval $[a, \infty)$ for some a. For brevity, write $\alpha_0 = \int_\Delta dF$, and suppose that $T(F) < \frac{1}{2}$. Recall that the contamination bias $b_T(\varepsilon; F)$ of $T(F)$ is equal to $1 - T(F)$ for all $\varepsilon > 0$. Let $b_{\tilde{T}}(\varepsilon; F)$ denote the contamination bias of $\tilde{T}(F, n)$ with respect to the Huber contamination discrepancy. To establish an upper bound on $b_{\tilde{T}}(\varepsilon; F)$ we note that

$$\begin{aligned} |\tilde{T}(G, n) - \tilde{T}(F, n)| &= |\{1 - Q_\kappa(F, n)\}\{T(G) - T(F)\} + T(G)\{Q_\kappa(F, n) - Q_\kappa(G, n)\}| \\ &\leqslant \{1 - Q_\kappa(F, n)\}|T(G) - T(F)| + T(G)|Q_\kappa(F, n) - Q_\kappa(G, n)|. \end{aligned}$$

Let $\alpha = \int_\Delta dG$. Then

$$Q_\kappa(G, n) - Q_\kappa(F, n) = \frac{(\alpha_0 - \alpha)\kappa n^{-1}}{\{\alpha_0 + (1 - \alpha_0)\kappa n^{-1}\}\{\alpha + (1 - \alpha)\kappa n^{-1}\}}.$$

Since $\delta(G, F) \leqslant \varepsilon$, we have that $\alpha = (1 - \varepsilon)\alpha_0 + \varepsilon\mu$ for some $\mu \in [0, 1]$, and

$$Q_\kappa(G, n) - Q_\kappa(F, n) = \frac{\varepsilon(\alpha_0 - \mu)\kappa n^{-1}}{A_0\{A_0 - \varepsilon(\alpha_0 - \mu)(1 - \kappa n^{-1})\}},$$

where $A_0 = \alpha_0 + \kappa n^{-1}(1 - \alpha_0)$. Note that the denominator is positive for all $\mu \in [0, 1]$; also we assume that $\kappa n^{-1} < 1$. Therefore

$$\max_{\mu \in [0,1]} |Q_\kappa(G, n) - Q_\kappa(F, n)| \leqslant \frac{\kappa n^{-1}\varepsilon(1 - \alpha_0)}{A_0\{A_0 - \varepsilon\alpha_0(1 - \kappa n^{-1})\}},$$

and we have the following upper bound for the contamination bias:

$$b_{\tilde{T}}(\varepsilon; F) \leqslant \frac{\alpha_0}{\alpha_0 + \kappa n^{-1}(1 - \alpha_0)} b_T(\varepsilon; F) + \frac{\kappa n^{-1}\varepsilon(1 - \alpha_0)}{A_0\{A_0 - \varepsilon\alpha_0(1 - \kappa n^{-1})\}}. \tag{11}$$

Note that $b_{\tilde{T}}(\varepsilon; F)$ depends on the underlying distribution F. In particular, if $\alpha_0 = 0$, that is F assigns zero probability to Δ, then $b_{\tilde{T}}(\varepsilon; F) \leqslant \varepsilon n\kappa^{-1}$, whereas $b_T(\varepsilon; F) = 1$, for all $\varepsilon > 0$. In addition, for this case $b_{\tilde{T}}(0+; F) = 0$, and hence $\tilde{T}(F, n)$ is qualitatively robust at all F satisfying $\alpha_0 = \alpha_0(F) = 0$. Since $b_{\tilde{T}}(\varepsilon; F)$ is a monotone increasing positive function of ε, we infer that

$$\gamma_{\tilde{T}}^* = \limsup_{\varepsilon \downarrow 0} \{b_{\tilde{T}}(\varepsilon; F)/\varepsilon\} \leqslant n\kappa^{-1},$$

i.e. for a given sample size $\tilde{T}(F, n)$ has finite contamination sensitivity. If $\alpha_0 \neq 0$, then the modified functional $\tilde{T}(F, n)$ loses the qualitative robustness property. The upper bound (11) shows, however, that contamination bias of the modified functional $\tilde{T}(F)$ still may be smaller than that of the functional $T(F)$ in a neighbourhood of zero. In particular, the following proposition is an immediate consequence of (11).

PROPOSITION 1. *Let* α_0 *be small enough so that* $T(F) < \frac{1}{2}$. *Then for every* κ *and* n *there exists a constant* $\varepsilon_* = \varepsilon_*(\alpha_0, \kappa, n) \in (0, 1]$ *such that for all* $\varepsilon \leqslant \varepsilon_*$ *one has* $b_{\tilde{T}}(\varepsilon; F) < b_T(\varepsilon; F)$. *In particular,* $\varepsilon_*(0, \kappa, n) = 1$.

The influence function of the modified functional $\tilde{T}(F, n)$ can be found by straightforward but tedious differentiation:

$$\mathrm{IC}(x, F, \tilde{T}) = \{1 - Q_\kappa(F, n)\}\ \mathrm{IC}(x, F;\ T) + \frac{\kappa n^{-1}\alpha_0^2 T(F) I_\Delta(x)}{\{\alpha_0^2 + \kappa n^{-1}\alpha_0(1-\alpha_0)\}^2}, \tag{12}$$

where $\mathrm{IC}(x, F, T)$ is given in (5). As might be expected, $\mathrm{IC}(x, F, \tilde{T})$ is asymptotically equivalent to $\mathrm{IC}(x, F, T)$ but (12) suggests that for small samples the variance of $\tilde{T}_n$ may be smaller than that of T_n.

3·6. *Choice of* κ

The parameter κ controls the sensitivity of the estimator $\tilde{T}(F_n)$ to sample fluctuations. As κ becomes smaller, $\tilde{T}(F_n)$ approaches $T(F_n)$ and the sensitivity of $\tilde{T}(F_n)$ increases. This effect can be explained from the viewpoint of the bias/variance trade-off. For every fixed n, $\tilde{T}_n$ estimates $\tilde{T}(F, n)$ instead of $T(F)$. This incurs additional bias, given by

$$B_n(\kappa) = \tilde{T}(F, n) - T(F) = T(F) Q_\kappa(F, n).$$

Since $Q_\kappa(F, n)$ becomes large as κ increases, the bias $B_n(\kappa)$ also increases. On the other hand, using differential approximation and neglecting the second-order terms we can write

$$\tilde{T}(F_n, n) - \tilde{T}(F, n) \simeq \{1 - Q_\kappa(F, n)\} \int_0^\infty \phi_F\, d(F_n - F),$$

where $\phi_F(x) = \mathrm{IC}(x, F, T)$ is the influence function of the functional $T(F)$. Since

$$\sigma_F^2 = E_F[\phi_F(X_i) - E_F\{\phi_F(X_i)\}]^2,$$

where σ_F^2 is given by (3), we can write

$$V_n(\kappa) = E_F\{\tilde{T}(F_n, n) - \tilde{T}(F, n)\}^2 \simeq \frac{\sigma_F^2}{n}\{1 - Q_\kappa(F, n)\}^2.$$

Therefore, $V_n(\kappa)$ decreases as κ increases.

We will base the choice of κ on minimising the mean squared error

$$\mathrm{MSE}(\kappa) = \{B_n(\kappa)\}^2 + V_n(\kappa) \simeq \{T(F) Q_\kappa(F, n)\}^2 + \frac{\sigma_F^2}{n}\{1 - Q_\kappa(F, n)\}^2. \tag{13}$$

Minimising $\mathrm{MSE}(\kappa)$ with respect to $Q_\kappa(F, n)$ we obtain

$$Q_\kappa^*(F, n) = \frac{n^{-1}\sigma_F^2}{\{T(F)\}^2 + n^{-1}\sigma_F^2},$$

and taking (9) into account we find the optimal value $\kappa_* = \alpha\sigma_F^2(1-\alpha)^{-1}\{T(F)\}^{-2}$. However, this choice is not practicable because it depends on unknown quantities $T(F)$, σ_F^2 and $\alpha = \int_\Delta dF$. We suggest the following plug-in approximation based on (3):

$$\kappa_+ = \frac{\hat{\sigma}_F^2 \int_\Delta dF_n}{T_n^2(1 - \int_\Delta dF_n)}, \tag{14}$$

where

$$\hat{\sigma}_F^2 = \frac{(1-2T_n)\int_\Delta w^2\, dF_n + T_n^2 \int_0^\infty w^2\, dF_n}{(\int_0^\infty w\, dF_n)^2}, \tag{15}$$

and $T_n = T(F_n)$ as before. It can be proved that the estimator based on $\kappa = \kappa_+$ is asymptotically efficient in the sense of Theorem 3.

The problem of choosing κ is similar to the problem of choosing the biasing parameter in ridge regression. The literature on this subject is vast, see for example Hoerl & Kennard (1970), Draper & Van Nostrand (1979), Golub, Heath & Wahba (1979) and references therein, and many of the ideas there can be implemented for choosing κ. We are currently investigating this issue, but here we restrict ourselves to the choice (14).

4. Numerical examples

4·1. *Simulated example*

Assume that the underlying distribution is exponential, that is $F(x) = 1 - e^{-x}$. We wish to estimate

$$T = T(F) = \frac{\int_a^\infty x^3 \exp(-x)\, dx}{\int_0^\infty x^3 \exp(-x)\, dx} = \frac{1}{6}(a^3 + 3a^2 + 6a + 6)\exp(-a).$$

Both the standard and shrinkage estimates are computed for different values of a using 1000 simulated samples of 100 observations each. For the shrinkage estimator we choose $\kappa = \kappa_+$, given by (14). The results are reported in Table 1. For each a Table 1 displays $T(F)$, the bias squared, the variance and the mean squared error.

Table 1. *Performance of the standard and shrinkage estimates for the exponential example of* § 4·1, *assessed from* 1000 *simulations of* 100 *observations each*

		Standard			Shrinkage		
a	$T(F)$	Bias^2	Variance	MSE	Bias^2	Variance	MSE
3·00	0·642	0·003	0·030	0·033	0·015	0·021	0·036
3·25	0·591	0·004	0·036	0·040	0·015	0·025	0·040
3·50	0·537	0·004	0·042	0·046	0·015	0·029	0·044
3·75	0·484	0·004	0·047	0·051	0·014	0·033	0·047
4·00	0·433	0·005	0·051	0·056	0·013	0·036	0·049
4·25	0·386	0·005	0·051	0·056	0·013	0·036	0·049
4·50	0·342	0·005	0·051	0·056	0·011	0·036	0·047
4·75	0·302	0·004	0·051	0·055	0·011	0·035	0·046
5·00	0·265	0·004	0·050	0·054	0·010	0·034	0·044

MSE, mean squared error.

As expected, the standard estimator is biased and slightly underestimates $T(F)$. By construction, the bias of the shrinkage estimator is larger than that of the standard one. The variance of the shrinkage estimator is smaller than that of the standard estimator. As $T(F)$ decreases, mean squared error of the shrinkage estimator becomes smaller than that for the standard estimator. This can be explained by the fact that, since F assigns a small mass to Δ, F_n is not an accurate estimator of F, and small-sample fluctuations lead to high variability in the standard estimator.

4·2. *Real data examples*

Example 1. The context is that of particle size measurement. Raw measurements on particles size are collected in the form of a distribution of counts. Figure 1 presents the results obtained for probes of graphite and some other test material. Figures 1(a),(b) display the histograms of diameters constructed from 7500 particles of the test material and graphite respectively. The histograms are based on equal bins, and the bin width is determined by the resolution of a sampling equipment.

Our setting is the same as in the example in § 3, and we are estimating the functionals t_k corresponding to every bin $\Delta = B_k$, for $k = 1, \ldots, N$. The standard estimate is given by (7), and the shrinkage estimate is a slight modification of (8):

$$\tilde{t}_k = \frac{\hat{t}_k \hat{f}_k^2}{\hat{f}_k^2 + \kappa n^{-1} \hat{f}_k (1 - \hat{f}_k)} \quad (k = 1, \ldots, N). \tag{16}$$

Figures 1(c)–(d) show the estimates $\hat{t}_1, \ldots, \hat{t}_N$ of the particle volume distribution corresponding to the count histograms given in Figs 1(a),(b).

For selecting κ in (16) we use the approximation as in (13) to obtain

$$\text{MSE}(\kappa) \simeq \sum_{i=1}^{N} \left\{ \frac{n^{-1} \sigma_i^2 \lambda_i^2}{(\lambda_i + \kappa)^2} + \frac{\kappa^2 t_i^2}{(\lambda_i + \kappa)^2} \right\} = \sum_{i=1}^{N} S_i(\kappa), \tag{17}$$

where $\lambda_i = nf_i/(1 - f_i)$ and σ_i^2 stands for the limit variance in estimating t_i, comparable to σ_F^2. The choice of κ, based on minimisation of (17), is again not practicable because it depends on the unknown parameters t_i, σ_i^2 and f_i. Note, however, that (17) is similar to the expressions for the mean squared error in generalised ridge regression (Hoerl & Kennard, 1970). Therefore, it seems that algorithms for selecting the biasing parameter in ridge regression can be adopted here. We choose to follow the procedure discussed in Montgomery & Peck (1992, p. 348). Let κ_i $(i = 1, \ldots, N)$ minimise $S_i(\kappa)$, that is $\kappa_i = \sigma_i^2 f_i t_i^{-2} (1 - f_i)^{-2}$. We use the same approximation to κ_i as in (14) and (15), which in this case takes the form

$$\kappa_i^+ = \frac{(1 - 2\hat{t}_i) w_i^2 \hat{f}_i + \hat{t}_i^2 (\sum_{j=1}^{N} w_j^2 \hat{f}_j)}{(\sum_{j=1}^{N} w_j \hat{f}_j)^2} \frac{\hat{f}_i}{\hat{t}_i^2 (1 - \hat{f}_i)}.$$

Finally we set $\kappa = \kappa_+$ as the harmonic mean of the κ_i^+:

$$\kappa_+ = \frac{N}{\sum_{i=1}^{N} (1/\kappa_i^+)}. \tag{18}$$

Figures 1(e),(f) show the shrinkage estimates of the volume distribution for particles of the test material and graphite respectively. For the test material the algorithm chooses $\kappa_+ = 6{\cdot}13$, while for graphite $\kappa_+ = 4{\cdot}877$. It is important to note that the data contain few large particles with diameters larger than 40 μm. These data are not visible in the counts histograms in Figs 1(a),(b), but they greatly influence the standard estimates of the volume distribution, see Figs 1(c),(d). As discussed before, the counts histograms are not accurate in this region, so that the shrinkage estimates in Figs 1(e),(f) are quite different from the standard ones there. The shrinkage estimates also differ slightly from the standard ones in the range where many particles occurred.

Example 2. Finally we present an example to demonstrate that the shrinkage estimator improves the repeatability of the experiments. From a fixed probe of a material 10 different

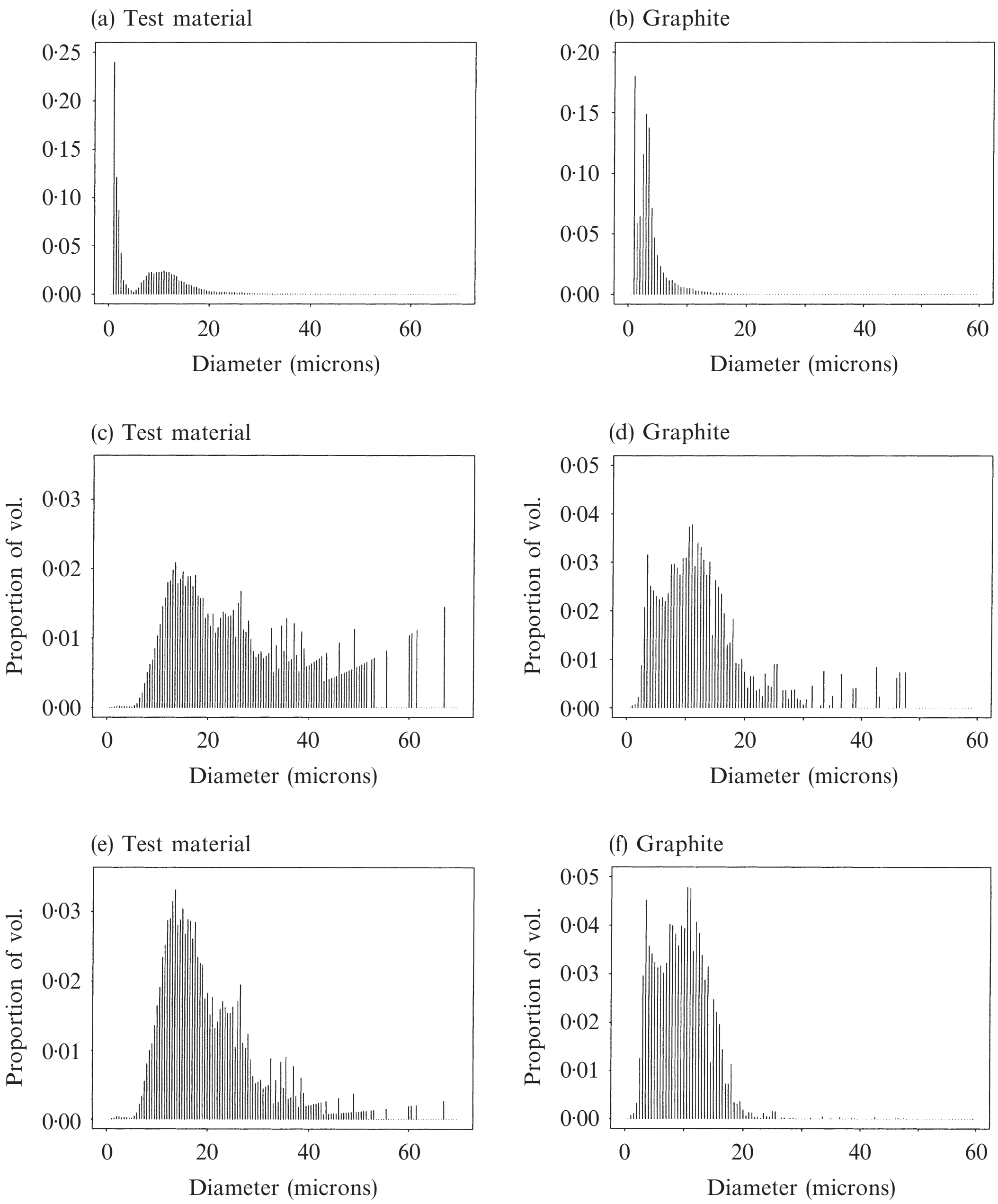

Fig. 1: Example 1. Histograms for a test material, and graphite: (a), (b) counts histograms; (c), (d) standard estimates of the volume distribution; (e), (f) shrinkage estimates of the volume distribution.

samples consisting of 7500 particles of graphite were taken. For each sample we computed the standard and shrinkage estimates of the volume distributions, and present them in the form of cumulative volume distributions. We took $\kappa = \kappa_+$ as in (18). In this example κ_+ varied over the 10 samples from 4·234 to 5·619. Figure 2(a),(b) shows the standard and shrinkage estimates, respectively, of the cumulative volume distributions for the 10 independent samples. Clearly, the shrinkage estimates demonstrate better repeatability.

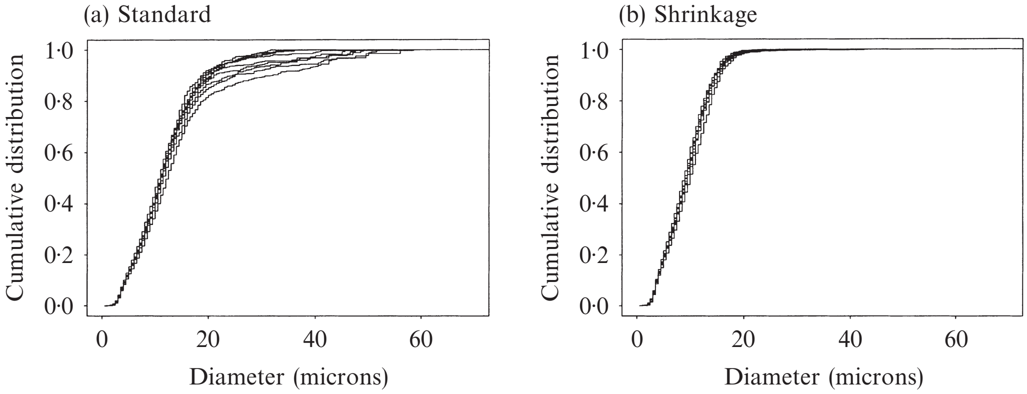

Fig. 2: Example 2. Estimates of cumulative volume distributions for 10 independent samples of graphite particles

5. Discussion

Tibshirani & Wasserman (1988) called a parameter or functional $T(F): \mathscr{M} \to \Theta$ sensitive with respect to a discrepancy measure δ on $\mathscr{M}$ if, for all $F \in \mathscr{M}$, $\varepsilon > 0$ and $\theta \in \Theta$, there exists a distribution G such that $\delta(F, G) \leqslant \varepsilon$ and $\theta = T(G)$. In other words, the sensitive functional can take any value from the parametric set Θ under any, no matter how small, change in distribution F. It immediately follows that every sensitive functional has zero breakdown point. The converse is not true. Tibshirani & Wasserman (1988) claim that a functional which is not sensitive in the above sense and has zero breakdown has to be somewhat artificial. It is interesting to note that the functional

$$T(F) = \left(\int_0^\infty w \, dF \right)^{-1} \int_\Delta w \, dF$$

is not sensitive with respect to the Huber discrepancy in the sense of the above definition, but it has zero breakdown point. Indeed, in this case $\Theta = [0, 1]$, and for $\varepsilon < \int_\Delta dF$ there is no distribution G such that $\delta(F, G) \leqslant \varepsilon$ and $T(G) = 0$.

Acknowledgement

The authors thank Dr N. Karasikov from Galai Laboratories Ltd for providing the data and motivating our interest in this problem, and the editor and two referees for useful comments that improved the manuscript.

Appendix

Proofs

Proof of Theorem 1. The proof follows from standard results for M-estimators, since $T_n = T(F_n)$ is the M-estimator defined by the equation

$$\sum_{i=1}^{n} \psi(X_i, t) = 0,$$

where $\psi(x, t) = w(x)\{I_\Delta(x) - t\}$, $I_\Delta(x) = 1$ is $x \in \Delta$, and $I_\Delta(x) = 0$ otherwise. Let

$$\lambda_F(t) = E_F \psi(x, t) = \int_\Delta w \, dF - t \int_0^\infty w \, dF.$$

Then $T(F)$ is the unique solution to the equation $\lambda_F(t)=0$. The first statement of the theorem follows immediately from Corollary 2.2 in Huber (1981, p. 48). Note that $\psi(x,t)$ is monotone decreasing in t. Then asymptotic normality is an immediate consequence of Corollary 2.5 in Huber (1981, p. 50): $n^{\frac{1}{2}}\{T_n-T(F)\}$ is asymptotically normal with mean 0 and variance

$$[\lambda_F'\{T(F)\}]^{-2}E_F[\psi\{x,T(F)\}]^2=\sigma_F^2. \qquad \square$$

Proof of Theorem 2. We use the von Mises differential approach to derive the asymptotic distribution (Serfling, 1980, Ch. 6). Note that the functional $T(F)$ is differentiable in the von Mises sense, so that we can write

$$\begin{aligned} T(F_n)-T(F) &= \int_0^\infty \phi_F\, d(F_n-F)+\frac{1}{2}\left.\frac{d^2T\{F+t(F_n-F)\}}{dt^2}\right|_{t=t_*} \\ &\equiv \eta_n(F)+\zeta_n(T,t_*), \end{aligned} \tag{A1}$$

where $t_*\in[0,1]$ and $\phi_F(.)$ is the influence function, which in our case is given by

$$\phi_F(x)=\left(\int_0^\infty w\,dF\right)^{-1}w(x)\{I_\Delta(x)-T(F)\}.$$

The first term $\eta_n(F)$ in the right-hand side of (A1) is expressed as a sum of independent identically distributed random variables,

$$\eta_n(F)=\int_0^\infty \phi_F\,d(F_n-F)=\frac{1}{n}\sum_{i=1}^n\{\phi_F(X_i)-E_F\phi_F(X_i)\}\equiv\frac{1}{n}\sum_{i=1}^n\xi_i(F).$$

Note that

$$E_F\{\xi_i(F)\}=0,\quad E_F\{\xi_i(F)\}^2=E_F\{\phi_F(X_i)\}^2-\{E_F\phi_F(X_i)\}^2=\sigma_F^2<\infty,$$

provided that $\int_0^\infty w^2\,dF<\infty$. Recall that σ_F^2 is defined in (3). Note also that $E_F|\xi_i(F)|^3$ is bounded uniformly over $\mathscr{F}_w(c,\gamma)$. Therefore, by the Berry–Esseen theorem, convergence of the distribution of $n^{\frac{1}{2}}\sigma_F^{-1}\eta_n(F)$ to $\mathscr{N}(0,1)$ is also uniform in $F\in\mathscr{F}_w(c,\gamma)$. Thus, to complete the proof of the theorem it suffices to show that

$$\sup_{F\in\mathscr{F}_w(c,\gamma)}E_F\left\{n^{\frac{1}{2}}\sup_{t_*\in[0,1]}|\zeta_n(F,t_*)|\right\}\to 0, \tag{A2}$$

as $n\to\infty$. By straightforward differentiation we have

$$\zeta_n(F,t_*)=-\frac{A(F_n,F)\int_0^\infty w\,d(F_n-F)}{\{\int_0^\infty w\,dF+t_*\int_0^\infty w\,d(F_n-F)\}^3},$$

where

$$A(F_n,F)=\int_\Delta w\,d(F_n-F)\int_0^\infty w\,dF-\int_0^\infty w\,d(F_n-F)\int_\Delta w\,dF.$$

Now we note that

$$\sup_{t_*\in[0,1]}|\zeta_n(F,t_*)|=\left|\frac{A(F_n,F)\int_0^\infty w\,d(F_n-F)}{\{\min(\int_0^\infty w\,dF,\int_0^\infty w\,dF_n)\}^3}\right|. \tag{A3}$$

Since there exists $c>0$ such that $F(c)=0$, the denominator is bounded away from zero uniformly in $F\in\mathscr{F}_w(c,\gamma)$, and we can write

$$\sup_{t_*\in[0,1]}|\zeta_n(F,t_*)|\leqslant\{w(c)\}^{-3}\left|A(F_n,F)\int_0^\infty w\,d(F_n-F)\right|.$$

Now we bound the expectation of the numerator in the right-hand side of (A3). We have

$$E_F\left\{\int_0^\infty w\, d(F_n - F)\right\}^2 = \frac{1}{n}\left\{\int_0^\infty w^2\, dF - \left(\int_0^\infty w\, dF\right)^2\right\}.$$

Similarly,

$$E_F|A(F_n, F)|^2 \leqslant 2\left(\int_0^\infty w\, dF\right)^2 E_F\left\{\int_\Delta w\, d(F_n - F)\right\}^2 + 2\left(\int_\Delta w\, dF\right)^2 E_F\left\{\int_0^\infty w\, d(F_n - F)\right\}^2$$
$$\leqslant \frac{4(\int_0^\infty w\, dF)^2 \int_0^\infty w^2\, dF}{n}.$$

Thus, applying the Cauchy–Schwartz inequality we obtain

$$n^{\frac{1}{2}} E_F\left\{\sup_{t_* \in [0,1]} |\zeta_n(F, t_*)|\right\} \leqslant 4\{w(c)\}^{-3} n^{-\frac{1}{2}} \int_0^\infty w\, dF \int_0^\infty w^2\, dF \to 0,$$

as $n \to \infty$, and (A2) follows. This completes the proof. □

Lower bound. We investigate how well the functional $T(F)$ can be estimated. Given an estimator $\hat{T}_n$ of $T(F)$ based on the observations $X_1, \ldots, X_n$, we measure its estimation accuracy by the risk $E_F|\hat{T}_n - T(F)|^2$. In the following theorem we establish a lower bound on the risk. Note that we do not restrict ourselves to a parametric family of probability distributions F.

THEOREM A1. *Let $\{U_N(F)\}$ be a sequence of neighbourhoods of the distribution $F \in \mathscr{F}_w$ in the weak convergence topology such that $U_N(F)$ shrinks to F as $N \to \infty$. Then*

$$\lim_{N\to\infty} \liminf_{n\to\infty} \inf_{\hat{T}_n} \sup_{G \in U_N(F)} E_G\{n|\hat{T}_n - T(G)|^2\} \geqslant \sigma_F^2,$$

where σ_F^2 is given by (3).

Proof. This is based on direct application of lower bounds for estimating nonparametric functionals differentiable in the von Mises sense (Levit, 1975; Ibragimov & Has'minskii, 1981, Ch. IV, § 2). For each $F \in \mathscr{F}_w$, $T(F)$ is the unique solution of the equation

$$\int_0^\infty \psi(x, t)\, dF(x) = 0, \quad \psi(x, t) = w(x)\{I_\Delta(x) - t\},$$

where $I_\Delta(x) = 1$ if $x \in \Delta$, and $I_\Delta(x) = 0$ otherwise. In addition, the function

$$\lambda_F(t) = E_F \psi(x, t) = \int_\Delta w\, dF - t \int_0^\infty w\, dF$$

is differentiable in t and $\lambda'_F\{T(F)\} = -\int_0^\infty w\, dF \neq 0$ for all $F \in \mathscr{F}_w$. Further, it is immediately seen that $E_F|\psi\{x, T(F)\}|^2 < \infty$ for all $F \in \mathscr{F}_w$. By straightforward calculation we obtain that, for any F_1 and F_2 from $\mathscr{F}_w$, one has $T\{F_1 + t(F_2 - F_1)\} \to T(F_1)$ as $t \to 0$. Thus, all conditions of Example 4 of Ibragimov & Has'minskii (1981, p. 224) hold, and the lower bound follows if we note that

$$\sigma_F^2 = [\lambda'_F\{T(F)\}]^{-2} \int_0^\infty [\psi\{x, T(F)\}]^2\, dF(x). \qquad \square$$

Proof of Theorem 3. Let $\Lambda_n = \{(X_1, \ldots, X_n) : \int_\Delta dF_n \neq 0\}$. Note that, on the complementary set $\bar{\Lambda}_n$, $\tilde{T}_n = T_n = 0$. Thus, it suffices to prove that for every $\beta > 0$

$$\sup_{F \in \mathscr{F}_w(c,\gamma)} P_F(n^{\frac{1}{2}}|Q_n T_n| > \beta, \Lambda_n) \to 0, \tag{A4}$$

as $n \to \infty$. It is evident that $n^{-\frac{1}{2}} \int_\Delta dF_n(1 - \int_\Delta dF_n)$ converges in probability to 0, and this convergence is uniform in $F \in \mathscr{F}_w(c, \gamma)$. Similarly, $(\int_\Delta dF_n)^2 + n^{-1} \int_\Delta dF_n (1 - \int_\Delta dF_n)$ converges in probability to $(\int_\Delta dF)^2$, and (A4) follows. Then, using

$$n^{\frac{1}{2}}(\tilde{T}_n - T) = n^{\frac{1}{2}}(T_n - T) + n^{\frac{1}{2}} Q_n T_n$$

and Theorem 2, we complete the proof of the theorem. □

References

Allen, T. (1990). *Particle Size Measurement*. London: Chapman and Hall.

Barndorff-Nielsen, O. (1977). Exponentially decreasing distributions for the logarithm of particle sizes. *Proc. R. Soc. Lond.* A **353**, 401–19.

Draper, N. & Van Nostrand, R. C. (1979). Ridge regression and James-Stein estimation: review and comments. *Technometrics* **21**, 451–66.

Fernholz, L. T. (1997). Reducing the variance by smoothing. *J. Statist. Plan. Infer.* **57**, 29–38.

Fieller, N. R. J., Flenley, E. C. & Olbricht, W. (1992). Statistics of particle size data. *Appl. Statist.* **41**, 127–46.

Golub, G., Heath, M. & Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics* **21**, 215–24.

He, X. & Simpson, D. (1993). Lower bounds for contamination bias: globally minimax versus locally linear estimation. *Ann. Statist.* **21**, 314–37.

Hoerl, A. & Kennard, R. (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* **12**, 55–67.

Hogg, R. (1974). Adaptive robust procedures: a partial review and some suggestions for future applications and theory. *J. Am. Statist. Assoc.* **69**, 909–27.

Huber, P. (1981). *Robust Statistics*. New York: John Wiley.

Ibragimov, I. & Has'minskii, R. (1981). *Statistical Estimation*. New York: Springer-Verlag.

Johns, M. V. (1974). Nonparametric estimation of location. *J. Am. Statist. Assoc.* **69**, 453–60.

Levit, B. (1975). On the efficiency of a class of nonparametric estimates. *Theory Prob. Applic.* **20**, 723–40.

Lindsay, B. G. (1994). Efficiency versus robustness: the case for minimum distance and related methods. *Ann. Statist.* **22**, 1081–114.

Montgomery, D. & Peck, E. (1992). *Introduction to Linear Regression Analysis*. New York: John Wiley.

Nicholson, W. (1970). Estimation of linear properties of particle size distributions. *Biometrika* **57**, 273–98.

Ripley, B. D. (1981). *Spatial Statistics*. New York: John Wiley.

Serfling, R. (1980). *Approximation Theorems of Mathematical Statistics*. New York: John Wiley.

Tibshirani, R. & Wasserman, L. (1988). Sensitive parameters. *Can. J. Statist.* **16**, 185–92.

Watson, G. S. (1971). Estimating functionals of particle size distributions. *Biometrika* **58**, 483–90.

[*Received June* 1998. *Revised August* 1999]