Math. Proc. Camb. Phil. Soc. (1994), **115**, 335
Printed in Great Britain

Optimal choice between parametric and non-parametric bootstrap estimates

BY STEPHEN MAN SING LEE

Statistical Laboratory, University of Cambridge, 16 Mill Lane, Cambridge CB2 1SB

(*Received* 23 *July* 1993; *revised* 12 *October* 1993)

Abstract

A parametric bootstrap estimate (PB) may be more accurate than its non-parametric version (NB) if the parametric model upon which it is based is, at least approximately, correct. Construction of an optimal estimator based on both PB and NB is pursued with the aim of minimizing the mean squared error. Our approach is to pick an empirical estimate of the optimal tuning parameter $\epsilon \in [0, 1]$ which minimizes the mean square error of $\epsilon\text{NB} + (1-\epsilon)\,\text{PB}$. The resulting hybrid estimator is shown to be more reliable than either PB or NB uniformly over a rich class of distributions. Theoretical asymptotic results show that the asymptotic error of this hybrid estimator is quite close in distribution to the smaller of the errors of PB and NB. All these errors typically have the same convergence rate of order $O(n^{-\frac{1}{2}})$. A particular example is also presented to illustrate the fact that this hybrid estimate can indeed be strictly better than either of the pure bootstrap estimates in terms of minimizing mean squared error. Two simulation studies were conducted to verify the theoretical results and demonstrate the good practical performance of the hybrid method.

1. *Introduction*

The idea of bootstrapping in statistical estimation was put forward by Efron [2]. The bootstrap method can be applied in both parametric and non-parametric fashions. In the context of estimation of a statistical functional $\alpha(F)$ based on a random sample of size n from an unknown distribution F, the parametric bootstrap estimate (PB) and the non-parametric bootstrap estimate (NB) amount to replacing F by $F_{\hat{\boldsymbol{\theta}}_n}$ and $\hat{F}_n$ respectively, where $\hat{\boldsymbol{\theta}}_n$ is some estimate of the index $\boldsymbol{\theta}$ parametrizing a family $\{F_{\boldsymbol{\theta}}\}$ thought to contain F, and $\hat{F}_n$ is the usual empirical distribution with equal probability placed on each data point.

In the literature PB and NB have been discussed mainly as two quite separate procedures. Their relative performance depends largely on correctness of our assumption of the parametric model. The estimate PB is generally more accurate if our proposed model is sufficiently close to F, whereas NB is the better choice if the parametric model is far from correct. The question of which one to use in circumstances where knowledge of the underlying distribution is vague remains unanswered. A straightforward criterion to guide our choice between the two bootstrap estimates is necessary.

Hjort [5] looks briefly at an intermediate bootstrap estimate (IB) obtained by replacing F with a mixture distribution $\epsilon F_{\hat{\boldsymbol{\theta}}_n} + (1-\epsilon)\hat{F}_n$, where $\epsilon \in [0, 1]$ is a pseudo-

https://doi.org/10.1017/S0305004100072121 Published online by Cambridge University Press

parameter to be determined in some optimal sense. Inspired by Hjort's suggestion we propose a slightly different ϵ-hybrid estimator, defined as

$$T_{n,\epsilon} \equiv \epsilon \mathrm{NB} + (1-\epsilon)\,\mathrm{PB},$$

where $\epsilon \in [0,1]$ is, in the terminology of [7], a 'tuning parameter'. The asymptotic properties of the ϵ-hybrid estimator are, unlike Hjort's approach, easy to study. Moreover, the optimal ϵ for $T_{n,\epsilon}$ is known explicitly so that it can be computed directly, whereas Hjort's estimate (IB) generally requires resampling from a class of distributions indexed by ϵ in order to determine the approximate value of the optimal ϵ.

Section 2 gives a full theoretical account of the asymptotic properties of our ϵ-hybrid estimator. It will be shown that $T_{n,\epsilon}$ combines the merits of PB and NB in a robust way. The tuning parameter ϵ, if picked to minimize the mean squared error of $T_{n,\epsilon}$, results in an optimal hybrid estimator. In practice, ϵ is determined by minimizing the empirical mean squared error of $T_{n,\epsilon}$. An exact expression of the optimal ϵ is readily available and its sample version can either be computed directly or approximated via bootstrap resampling. The optimal properties of $T_{n,\epsilon}$ are demonstrated through two simple examples in Section 3. The second example shows that $T_{n,\epsilon}$ can even be strictly better than either PB or NB, uniformly over a rich class of distributions. Simulation results are obtained and closely support our theory. A few open questions are raised in Section 4 about possible generalization of our procedure.

2. *Theory of ϵ-hybrid estimator*

2·1. *Problem specification.* A random sample of n observations

$$\mathcal{X} = (X_1, \ldots, X_n)$$

is drawn from an unknown distribution F. We suspect F belongs to a parametric family of distributions,

$$\mathcal{F} \equiv \{F_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \boldsymbol{\Theta}\},$$

where $\boldsymbol{\Theta} \subseteq \mathbf{R}^p$ is a non-empty parameter subspace. We wish to estimate a real-valued statistical functional $\alpha(F)$ based on $\mathcal{X}$.

Define $\hat{F}_n$ to be the empirical distribution of $\mathcal{X}$, that is, the distribution that places an equal mass of n^{-1} on each data point in $\mathcal{X}$. Define $\hat{\boldsymbol{\theta}}_n$ to be the maximum likelihood estimator (MLE) of $\boldsymbol{\theta}$ based on $\mathcal{X}$ assuming $F \in \mathcal{F}$. Then the parametric bootstrap estimator (PB) and non-parametric bootstrap estimator (NB) are respectively $\alpha(F_{\hat{\boldsymbol{\theta}}_n})$ and $\alpha(\hat{F}_n)$. Define an ϵ-hybrid estimator $T_{n,\epsilon}$ to be a convex combination of PB and NB, namely,

$$T_{n,\epsilon} \equiv \epsilon \alpha(\hat{F}_n) + (1-\epsilon)\,\alpha(F_{\hat{\boldsymbol{\theta}}_n}),$$

for $\epsilon \in [0,1]$.

Suppose now $\epsilon_n(F)$ minimizes the mean squared error (MSE) of $T_{n,\epsilon}$, that is,

$$\begin{aligned} \mathrm{MSE}_F(T_{n,\epsilon_n(F)}) &= \inf_{0 \leqslant \epsilon \leqslant 1} \mathrm{MSE}_F(T_{n,\epsilon}) \\ &= \inf_{0 \leqslant \epsilon \leqslant 1} \mathbf{E}_F\{[T_{n,\epsilon} - \alpha(F)]^2\}. \end{aligned}$$

Note that the optimal $\epsilon_n(F)$ depends on the unknown distribution F and is therefore

https://doi.org/10.1017/S0305004100072121 Published online by Cambridge University Press

not generally available for evaluation. It may, however, be estimated by its non-parametric bootstrap estimate $\epsilon_n(\hat{F}_n)$. Write

$$\hat{\epsilon}_n \equiv \epsilon_n(\hat{F}_n)$$

for simplicity. We shall study the asymptotic properties of both the *theoretical* optimal ϵ-hybrid estimator $T_{n,\epsilon_n(F)}$ and the *empirical* optimal ϵ-hybrid estimator $T_{n,\hat{\epsilon}_n}$.

2·2. *Notation and assumptions.* Before laying down the basic assumptions underlying the theoretical results to be derived later, we first introduce some notation:

1. $\|\cdot\|$ is the usual Euclidean norm defined on the space of real matrices, so that

$$\|A\| \equiv \left(\sum_{i=1}^{p}\sum_{j=1}^{q} a_{ij}^2\right)^{\frac{1}{2}},$$

where $A = (a_{ij})$ is any $p \times q$ matrix. Thus, in particular,

$$\|\mathbf{x}\| = \left(\sum_{i=1}^{p} x_i^2\right)^{\frac{1}{2}} \quad \forall \mathbf{x} \in \mathbf{R}^p.$$

2. For any differentiable function $g: \mathbf{\Omega} \to \mathbf{R}$ with $\mathbf{\Omega} \subseteq \mathbf{R}^p$, write

$$\nabla g(\mathbf{x}) = \left(\frac{\partial g}{\partial x_1}, \ldots, \frac{\partial g}{\partial x_p}\right)^{\mathrm{T}} \quad \forall \mathbf{x} \in \mathbf{\Omega}.$$

3. For any twice differentiable function $g: \mathbf{\Omega} \to \mathbf{R}$ with $\mathbf{\Omega} \subseteq \mathbf{R}^p$, let $\mathscr{H}g$ denote its Hessian matrix, that is,

$$\mathscr{H}g(x_1, \ldots, x_p) \equiv \begin{bmatrix} \dfrac{\partial^2 g}{\partial x_1^2} & \dfrac{\partial^2 g}{\partial x_1 \partial x_2} & \cdots & \dfrac{\partial^2 g}{\partial x_1 \partial x_p} \\ \dfrac{\partial^2 g}{\partial x_2 \partial x_1} & \dfrac{\partial^2 g}{\partial x_2^2} & \cdots & \\ \vdots & \vdots & & \vdots \\ \dfrac{\partial^2 g}{\partial x_p \partial x_1} & \dfrac{\partial^2 g}{\partial x_p \partial x_2} & \cdots & \dfrac{\partial^2 g}{\partial x_p^2} \end{bmatrix}.$$

4. Let Y_n be a random variable such that along almost all sample sequences $X_1, X_2, \ldots$, the conditional distribution of Y_n given $(X_1, \ldots, X_n)$ converges weakly to a limit distribution G corresponding to a random variable Y say. Then we write

$$Y_n \,|\, X_1, \ldots, X_n \xrightarrow{\mathscr{D}} Y \text{ (or } G) \quad \text{a.s.}$$

where 'a.s.' stands for 'almost surely' as usual.

5. Similarly, if Y_n converges in conditional probability to Y given $(X_1, \ldots, X_n)$ along almost all random sequences $X_1, X_2, \ldots$, which means given any $\epsilon > 0$,

$$\mathbf{P}[|Y_n - Y| > \epsilon \,|\, X_1, \ldots, X_n] \to 0 \quad \text{a.s.},$$

we then write $$Y_n \,|\, X_1, \ldots, X_n \xrightarrow{\mathrm{p}} Y \quad \text{a.s.}$$

https://doi.org/10.1017/S0305004100072121 Published online by Cambridge University Press

6. If Y_n is uniformly integrable with respect to the conditional distribution given $(X_1, \ldots, X_n)$ along almost all random sequences $X_1, X_2, \ldots$, that is,

$$\lim_{c\to\infty} \sup_n \mathbf{E}[|Y_n| \mathbf{1}\{|Y_n| > c\} | X_1, \ldots, X_n] = 0 \quad \text{a.s.}$$

where $\mathbf{1}\{E\}$ is the indicator function of the event E, then we write

$$Y_n | X_1, X_2, \ldots, X_n \quad \text{is u.i. a.s.}$$

7. A superscript '*' is used to indicate the bootstrap version of any statistic based on a bootstrap resample $\mathscr{X}^*$ drawn randomly from $\mathscr{X}$ with replacement. For example, the empirical distribution of $\mathscr{X}^*$ is denoted by $\hat{F}_n^*$.

We now make the following assumptions:

(A 1) For all distribution functions G, H,

$$\alpha(H) = \alpha(G) + \int \psi_G(x)\, dH(x) + \text{Rem}\,(H, G),$$

with

$$\int \psi_G(x)\, dG(x) = 0.$$

(A 2) Under the true distribution F, ψ_F as determined in (A 1) satisfies

$$\mathbf{E}_F[\psi_F^2(X)] < \infty.$$

(A 3) The equation

$$\mathbf{E}_F[\nabla \log f(X; \boldsymbol{\theta})] = \mathbf{0}$$

has a unique solution $\boldsymbol{\theta}_0$ for $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ where $f(x; \boldsymbol{\theta})$ is the density function corresponding to the distribution $F_{\boldsymbol{\theta}}$, so that $F_{\boldsymbol{\theta}_0}$ is the member of $\mathscr{F}$ 'closest' to the true underlying distribution F.

(A 4) There exists a real-valued function H, defined on $\mathbf{R}$ with $\mathbf{E}_F[H(X)] < \infty$, such that for all $x \in \mathbf{R}$ and $i, j, k = 1, 2, \ldots, p$,

$$\left| \frac{\partial^3 \log f(x; \boldsymbol{\theta})}{\partial\theta_i\, \partial\theta_j\, \partial\theta_k} \right| \leqslant H(x).$$

(A 5) $\text{Var}_F[\nabla \log f(X; \boldsymbol{\theta})|_{\boldsymbol{\theta}_0}]$ and $\mathbf{E}_F[\mathscr{H} \log f(X; \boldsymbol{\theta})|_{\boldsymbol{\theta}_0}]$ exist with the latter being negative definite.

(A 6) Define a function $\beta: \boldsymbol{\Theta} \to \mathbf{R}$ by $\beta(\boldsymbol{\theta}) = \alpha(F_{\boldsymbol{\theta}})$. Then β is twice differentiable and $\|\mathscr{H}\beta\|$ is bounded in a neighbourhood of $\boldsymbol{\theta}_0$.

(A 7) $\sqrt{n}\,\text{Rem}\,(\hat{F}_n, F) \xrightarrow{\text{p}} 0$ and $\sqrt{n}\,\text{Rem}\,(\hat{F}_n^*, \hat{F}_n) | X_1, \ldots, X_n \xrightarrow{\text{p}} 0$ a.s.

(A 8) Almost sure convergence holds as follows:

$$\frac{1}{n} \sum_{i=1}^n [\psi_{\hat{F}_n}(X_i) - \psi_F(X_i)]^2 \to 0 \quad \text{a.s.},$$

$$\frac{1}{n} \sum_{i=1}^n \|\nabla \log f(X_i; \boldsymbol{\theta})|_{\hat{\boldsymbol{\theta}}_n} - \nabla \log f(X_i; \boldsymbol{\theta})|_{\boldsymbol{\theta}_0}\|^2 \to 0 \quad \text{a.s.}$$

and

$$\frac{1}{n} \sum_{i=1}^n \left| \frac{\partial^2}{\partial\theta_j\, \partial\theta_k} \log f(X_i; \boldsymbol{\theta}) \Big|_{\hat{\boldsymbol{\theta}}_n} - \frac{\partial^2}{\partial\theta_j\, \partial\theta_k} \log f(X_i; \boldsymbol{\theta}) \Big|_{\boldsymbol{\theta}_0} \right| \to 0 \quad \text{a.s.}$$

for all $j, k = 1, \ldots, p$.

https://doi.org/10.1017/S0305004100072121 Published online by Cambridge University Press

(A 9) The quantities $n[\alpha(\hat{F}_n)-\alpha(F)]^2$ and $n[\alpha(F_{\hat{\theta}_n})-\alpha(F_{\theta_0})]^2$ are u.i. with respect to F. Also, for any given sample sequence $X_1, X_2, \ldots,$

$$n[\alpha(\hat{F}_n^*)-\alpha(\hat{F}_n)]^2 \,|\, X_1, X_2, \ldots, X_n$$

and

$$n[\alpha(F_{\hat{\theta}_n^*})-\alpha(F_{\hat{\theta}_n})]^2 \,|\, X_1, X_2, \ldots, X_n$$

are u.i. a.s.

Roughly speaking, the above assumptions hold in general if:

1. F has finite moments of sufficiently high orders;

2. all the distributions $F_{\boldsymbol{\theta}} \in \mathscr{F}$ are regular and $\mathscr{F}$ is parametrized by $\boldsymbol{\theta}$ in some smooth manner. Serfling [**8**] provides a thorough account of the regularity conditions required for the theory of parametric maximum likelihood estimation. And:

3. the functional to be estimated, $\alpha(\cdot)$, is *Hadamard* differentiable with a well-behaved influence function ψ. See [**4**] for a definition of Hadamard differentiability.

2·3. *Preliminary results.* Two lemmas are presented here to establish convergence of bootstrap statistics conditional on a given sample of observations $X_1, \ldots, X_n$. They are analogous to the Central Limit Theorem and Weak Law of Large Numbers for ordinary random variables.

LEMMA A. *Suppose* $\mathbf{X}_1, \mathbf{X}_2, \ldots,$ *are a sequence of independent multivariate random variables drawn from a distribution* G *on* $\mathbf{R}^p$. *Let* $q: \mathbf{R}^p \times \mathscr{G} \to \mathbf{R}^k$ *be a statistical functional such that*

$$\mathbf{E}_G[\|q(\mathbf{X}, G)\|^2] < \infty,$$

where $\mathscr{G}$ *is some convex class of distributions containing* G *and all point masses. Let* $\hat{G}_n$ *be the empirical distribution of* $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$. *Given* $\mathbf{X}_1, \ldots, \mathbf{X}_n$, *let* $(\mathbf{X}_1^*, \ldots, \mathbf{X}_n^*)$ *be a random sample drawn from* $\hat{G}_n$. *Then along almost all sample sequences* $\mathbf{X}_1, \mathbf{X}_2, \ldots,$ *as* $n \to \infty$,

$$\text{(i)}\quad \frac{1}{\sqrt{n}}\left\{\sum_{j=1}^n q(\mathbf{X}_j^*, G) - \sum_{j=1}^n q(\mathbf{X}_j, G)\right\} \,|\, \mathbf{X}_1, \ldots, \mathbf{X}_n \xrightarrow{\mathscr{D}} N_k(\mathbf{0}, \mathrm{Var}_G[q(\mathbf{X}, G)]).$$

Furthermore, if

$$\frac{1}{n}\sum_{i=1}^n \|q(\mathbf{X}_i, \hat{G}_n) - q(\mathbf{X}_i, G)\|^2 \to 0 \quad \text{a.s.},$$

then

$$\text{(ii)}\quad \frac{1}{\sqrt{n}}\left\{\sum_{j=1}^n q(\mathbf{X}_j^*, \hat{G}_n) - \sum_{j=1}^n q(\mathbf{X}_j, \hat{G}_n)\right\} \,|\, \mathbf{X}_1, \ldots, \mathbf{X}_n \xrightarrow{\mathscr{D}} N_k(\mathbf{0}, \mathrm{Var}_G[q(\mathbf{X}, G)]),$$

along almost all sample sequences $\mathbf{X}_1, \mathbf{X}_2, \ldots$.

Note that the additional assumption that

$$\frac{1}{n}\sum_{i=1}^n \|q(\mathbf{X}_i, \hat{G}_n) - q(\mathbf{X}_i, G)\|^2 \to 0 \quad \text{a.s.}$$

required for result (ii) is usually satisfied by functionals $q(\mathbf{x}, G)$ sufficiently smooth with respect to G under moment conditions on $\mathbf{X}$.

Sketch proof of Lemma A. The first result (i) can be derived easily from the proof

https://doi.org/10.1017/S0305004100072121 Published online by Cambridge University Press

of Theorem 2·1 in [1]. That paper also suggests the second result (ii) for a scalar functional q. For a general k-dimensional setting, we can easily show by Chebyshev's inequality that

$$\frac{1}{\sqrt{n}}\left\{\sum_{j=1}^{n} q(\mathbf{X}_j^*, \hat{G}_n) - \sum_{j=1}^{n} q(\mathbf{X}_j, \hat{G}_n)\right\} - \frac{1}{\sqrt{n}}\left\{\sum_{j=1}^{n} q(\mathbf{X}_j^*, G) - \sum_{j=1}^{n} q(\mathbf{X}_j, G)\right\} \mid \mathbf{X}_1, \ldots, \mathbf{X}_n \xrightarrow{\text{p}} 0 \quad \text{a.s.}$$

By the previous result (i), the second term in the above difference converges in conditional distribution to $N_k(\mathbf{0}, \mathrm{Var}_G[q(\mathbf{X}, G)])$ and hence result (ii) follows. ∎

LEMMA B. *Suppose $X_1, X_2, \ldots,$ are a sequence of independent random variables drawn from G on $\mathbf{R}$. Suppose*

$$\mathbf{E}_G[|X|] < \infty.$$

Let $X_1^*, \ldots, X_n^*$ be a random sample drawn from $\hat{G}_n$, the empirical distribution based on $(X_1, \ldots, X_n)$. Let $\tau: \mathbf{R} \times \mathcal{G} \to \mathbf{R}$ be a statistical functional where $\mathcal{G}$ is a convex class of distributions containing G and all point masses. Suppose τ satisfies

$$\mathbf{E}_G[|\tau(X, G)|] < \infty.$$

Then we have

(i) $$\frac{1}{n}\sum_{i=1}^{n} \tau(X_i^*, G) \mid X_1, \ldots, X_n \xrightarrow{\text{p}} \mathbf{E}_G[\tau(X, G)] \quad \text{a.s.}$$

Further, if

$$\frac{1}{n}\sum_{i=1}^{n} |\tau(X_i, \hat{G}_n) - \tau(X_i, G)| \to 0 \quad \text{a.s.},$$

then

(ii) $$\frac{1}{n}\sum_{i=1}^{n} \tau(X_i^*, \hat{G}_n) \mid X_1, \ldots, X_n \xrightarrow{\text{p}} \mathbf{E}_G[\tau(X, G)] \quad \text{a.s.}$$

As in Lemma A, the additional condition stated for Lemma B (ii) is usually satisfied by sufficiently smooth functionals $\tau(x, G)$ under moment conditions on X.

Sketch proof of Lemma B. Let Γ_1 be the set of distributions H satisfying the condition $\int |x|\, dH(x) < \infty$. Define a metric d_1 on Γ_1 such that $d_1(H_1, H_2)$ is the infimum of $\mathbf{E}|X - Y|$ where X, Y have marginal distributions H_1, H_2 respectively. Without ambiguity we can write $d_1(Y, Z) = d_1(H_1, H_2)$, where $Y \sim H_1$ and $Z \sim H_2$. Let $H^{\langle n \rangle}$ be the distribution of $n^{-1}\sum_{i=1}^{n} Y_i$ with $Y_1, \ldots, Y_n$ independently distributed from H. Then for all H_1, H_2 in Γ_1,

$$d_1(H_1^{\langle n \rangle}, H_2^{\langle n \rangle}) \leqslant d_1(H_1, H_2).$$

Also, we have $$d_1(H_n, H) \to 0 \quad \text{as} \quad n \to \infty$$

if and only if $$H_n \to H \text{ weakly}$$

and $$\int |x|\, dH_n(x) \to \int |x|\, dH(x).$$

https://doi.org/10.1017/S0305004100072121 Published online by Cambridge University Press

The details of d_1 can be found in [1]. Now take H as the distribution of $\tau(X_1, G)$ with $X_1, X_2, \ldots$, independently distributed from G. Let $\hat{H}_n$ be the conditional distribution of $\tau(X_i^*, G)$ with X_i^*s independently distributed from $\hat{G}_n$. Clearly $H, \hat{H}_n \in \Gamma_1$ a.s. We can deduce from the Strong Law of Large Numbers (SLLN) that

$$d_1(\hat{H}_n^{\langle n\rangle}, H^{\langle n\rangle}) \leqslant d_1(\hat{H}_n, H) \to 0 \quad \text{a.s.}$$

On the other hand, SLLN and a slightly modified version of the Dominated Convergence Theorem imply the result

$$d_1(H^{\langle n\rangle}, \mathbf{E}_G \tau(X, G)) \to 0.$$

Thus

$$d_1(\hat{H}_n^{\langle n\rangle}, \mathbf{E}_G \tau(X, G)) \leqslant d_1(\hat{H}_n^{\langle n\rangle}, H^{\langle n\rangle}) + d_1(H^{\langle n\rangle}, \mathbf{E}_G \tau(X, G)) \to 0 \quad \text{a.s.}$$

and

$$\hat{H}_n^{\langle n\rangle} \to \mathbf{E}_G \tau(X, G)$$

weakly a.s. So result (i) follows.

To prove (ii), we let $\hat{J}_n$ be the conditional distribution of $\tau(X^*, \hat{G}_n)$ with $X^* \sim \hat{G}_n$. It is clear that $\hat{J}_n \in \Gamma_1$ a.s. Note that SLLN and the assumption

$$\frac{1}{n} \sum_{i=1}^{n} |\tau(X_i, \hat{G}_n) - \tau(X_i, G)| \to 0 \quad \text{a.s.}$$

together imply

$$\int |x| \, d\hat{J}_n(x) \to \mathbf{E}_G |\tau(X, G)| \quad \text{a.s.} \tag{†}$$

Also, Chebychev's inequality shows that

$$\tau(X^*, \hat{G}_n) - \tau(X^*, G) \,|\, X_1, \ldots, X_n \xrightarrow{\text{p}} 0 \quad \text{a.s.}$$

But we know that

$$\tau(X^*, G) \,|\, X_1, \ldots, X_n \xrightarrow{\mathscr{D}} \tau(X, G) \quad \text{a.s.}$$

using results from the proof of (i). Therefore, by Slutsky's Theorem,

$$\tau(X^*, \hat{G}_n) \xrightarrow{\mathscr{D}} \tau(X, G)$$

conditional on $X_1, \ldots, X_n$ a.s. Hence

$$\hat{J}_n \to H \text{ weakly} \quad \text{a.s.} \tag{‡}$$

From (†) and (‡), we have

$$d_1(\hat{J}_n, H) \to 0 \quad \text{a.s.}$$

Therefore $d_1(\hat{J}_n^{\langle n\rangle}, \mathbf{E}_G \tau(X, G)) \leqslant d_1(\hat{J}_n, H) + d_1(H^{\langle n\rangle}, \mathbf{E}_G \tau(X, G)) \to 0$ a.s.

So result (ii) follows.

2·4. *Asymptotic properties of ϵ-hybrid estimators.* In this section we shall develop the asymptotic properties of $\hat{\epsilon}_n$ and $T_{n,\hat{\epsilon}_n}$ as well as the theoretical optimal estimator $T_{n,\epsilon_n(F)}$. For simplicity, we write $\epsilon_n = \epsilon_n(F)$, $\hat{\epsilon}_n = \epsilon_n(\hat{F}_n)$, $\nu_n = \alpha(\hat{F}_n) - \alpha(F)$, $\omega_n = \alpha(F_{\hat{\theta}_n}) - \alpha(F)$, $\mu_n^* = \alpha(\hat{F}_n^*) - \alpha(\hat{F}_n)$, $\eta_n^* = \alpha(F_{\hat{\theta}_n^*}) - \alpha(F_{\hat{\theta}_n})$, and $\gamma = \alpha(F_{\theta_0}) - \alpha(F)$.

2·4·1. *Asymptotic joint distribution of pure bootstrap estimates.* First we study the

https://doi.org/10.1017/S0305004100072121 Published online by Cambridge University Press

joint distribution of the non-parametric bootstrap estimate $\alpha(\hat{F}_n)$ and the parametric bootstrap estimate $\alpha(F_{\hat{\theta}_n})$. By (A 1) we have

$$\sqrt{n}\begin{pmatrix}\nu_n\\ \omega_n\end{pmatrix} = \sqrt{n}\begin{pmatrix}\frac{1}{n}\sum_{i=1}^{n}\psi_F(X_i)\\ \beta(\hat{\boldsymbol{\theta}}_n)-\beta(\boldsymbol{\theta}_0)\end{pmatrix} + \sqrt{n}\begin{pmatrix}\mathrm{Rem}\,(\hat{F}_n,F)\\ \alpha(F_{\boldsymbol{\theta}_0})-\alpha(F)\end{pmatrix},$$

which represents the error of the pure bootstrap estimates for $\alpha(F)$. Here β and $\boldsymbol{\theta}_0$ are defined as in (A 3) and (A 6). We shall first prove the following theorem.

THEOREM 1. *The joint distribution of*

$$n^{-\frac{1}{2}}\sum_{i=1}^{n}\psi_F(X_i) \quad \textit{and} \quad n^{\frac{1}{2}}(\beta(\hat{\boldsymbol{\theta}}_n)-\beta(\boldsymbol{\theta}_0))$$

is asymptotically bivariate normal with zero mean and variance covariance matrix

$$\Sigma_F = \begin{pmatrix}\sigma_F^2 & \boldsymbol{\rho}_F^{\mathrm{T}} J_F(\boldsymbol{\theta}_0)^{-1}\nabla\beta|_{\boldsymbol{\theta}_0}\\ \nabla\beta|_{\boldsymbol{\theta}_0}^{\mathrm{T}} J_F(\boldsymbol{\theta}_0)^{-1}\boldsymbol{\rho}_F & \nabla\beta|_{\boldsymbol{\theta}_0}^{\mathrm{T}} J_F(\boldsymbol{\theta}_0)^{-1}K_F(\boldsymbol{\theta}_0)J_F(\boldsymbol{\theta}_0)^{-1}\nabla\beta|_{\boldsymbol{\theta}_0}\end{pmatrix},$$

where

$$\boldsymbol{\rho}_F = \mathbf{E}_F[\psi_F(X)\nabla\log f(X;\boldsymbol{\theta})|_{\boldsymbol{\theta}_0}],$$

$$\sigma_F^2 = \mathbf{E}_F[\psi_F^2(X)],$$

$$J_F(\boldsymbol{\theta}_0) = -\mathbf{E}_F\mathscr{H}\log f(X;\boldsymbol{\theta})|_{\boldsymbol{\theta}_0},$$

and

$$K_F(\boldsymbol{\theta}_0) = \mathrm{Var}_F[\nabla\log f(X;\boldsymbol{\theta})|_{\boldsymbol{\theta}_0}].$$

Proof of Theorem 1. By assumption (A 4), there exists a neighbourhood $N(\boldsymbol{\theta}_0)$ of $\boldsymbol{\theta}_0$ such that for all $\boldsymbol{\lambda},\boldsymbol{\theta}\in N(\boldsymbol{\theta}_0)$, we have

$$\mathbf{A}_n(\mathscr{X},\boldsymbol{\lambda}) = \mathbf{A}_n(\mathscr{X},\boldsymbol{\theta}) + B_n(\mathscr{X},\boldsymbol{\theta})(\boldsymbol{\lambda}-\boldsymbol{\theta}) + \tfrac{1}{2}C_n(\mathscr{X})W_n(\mathscr{X};\boldsymbol{\lambda},\boldsymbol{\theta})(\boldsymbol{\lambda}-\boldsymbol{\theta}) \tag{1}$$

where

$$\mathbf{A}_n(\mathscr{X},\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\nabla\log f(X_i;\boldsymbol{\theta}),$$

$$B_n(\mathscr{X},\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\mathscr{H}\log f(X_i;\boldsymbol{\theta}),$$

$$C_n(\mathscr{X}) = \frac{1}{n}\sum_{i=1}^{n}H(X_i) \quad \text{as defined in (A 4)},$$

and $W_n(\mathscr{X};\boldsymbol{\lambda},\boldsymbol{\theta})$ is a $p\times p$ matrix satisfying

$$\{W_n(\mathscr{X};\boldsymbol{\lambda},\boldsymbol{\theta})\}_{jl} = \sum_{k=1}^{p}(\lambda_k-\theta_k)\xi_{jkl}(\mathscr{X};\boldsymbol{\lambda},\boldsymbol{\theta})$$

for some $|\xi_{jkl}(\mathscr{X};\boldsymbol{\lambda},\boldsymbol{\theta})|<1$, and $j,k,l=1,2,\ldots,p$. Take $\boldsymbol{\theta}=\boldsymbol{\theta}_0$. Then SLLN and assumptions (A 3), (A 4) and (A 5) imply

$$\mathbf{A}_n(\mathscr{X},\boldsymbol{\theta}_0) = \frac{1}{n}\sum_{i=1}^{n}\nabla\log f(X_i;\boldsymbol{\theta})|_{\boldsymbol{\theta}_0} \xrightarrow{\text{a.s.}} \mathbf{E}_F\nabla\log f(X;\boldsymbol{\theta})|_{\boldsymbol{\theta}_0} = \mathbf{0},$$

$$B_n(\mathscr{X},\boldsymbol{\theta}_0) = \frac{1}{n}\sum_{i=1}^{n}\mathscr{H}\log f(X_i;\boldsymbol{\theta})|_{\boldsymbol{\theta}_0} \xrightarrow{\text{a.s.}} \mathbf{E}_F\mathscr{H}\log f(X;\boldsymbol{\theta})|_{\boldsymbol{\theta}_0} = -J_F(\boldsymbol{\theta}_0),$$

https://doi.org/10.1017/S0305004100072121 Published online by Cambridge University Press

and $$C_n(\mathscr{X}) = \frac{1}{n}\sum_{i=1}^{n} H(X_i) \xrightarrow{\text{a.s.}} \mathbf{E}_F H(X).$$

Thus, with probability 1 and for large enough n, there exists a sequence $\hat{\boldsymbol{\theta}}_n$ converging to $\boldsymbol{\theta}_0$ such that $\mathbf{A}_n(\mathscr{X}, \hat{\boldsymbol{\theta}}_n) = \mathbf{0}$. See [8, 4·2·2, pp. 144–9]. Then we have

$$\begin{aligned}\|W_n(\mathscr{X}; \hat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}_0)\| &\leqslant \|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\| \left[\sum_{j=1}^{p}\sum_{k=1}^{p}\sum_{l=1}^{p} \xi_{jkl}(\mathscr{X}; \hat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}_0)^2\right]^{\frac{1}{2}} \\ &\leqslant p^{\frac{3}{2}} \|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\| \\ &\to 0 \quad \text{a.s.} \quad \text{as } n \to \infty.\end{aligned}$$

Therefore, for sufficiently large n, we have from (1) that

$$\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0 = Q_n(\mathscr{X}; \hat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}_0)\, \mathbf{A}_n(\mathscr{X}, \boldsymbol{\theta}_0) \quad \text{a.s.} \tag{2}$$

where $Q_n(\mathscr{X}; \hat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}_0) = -(B_n(\mathscr{X}, \boldsymbol{\theta}_0) + \frac{1}{2}C_n(\mathscr{X})\, W_n(\mathscr{X}; \hat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}_0))^{-1} \to J_F(\boldsymbol{\theta}_0)^{-1}$ a.s.

since $J_F(\boldsymbol{\theta}_0)$ is positive definite by (A 5). Consider the Taylor's expansion for $\beta(\boldsymbol{\theta})$, which is valid by assumption (A 6),

$$\beta(\hat{\boldsymbol{\theta}}_n) - \beta(\boldsymbol{\theta}_0) = \nabla\beta|_{\boldsymbol{\theta}_0}^{\mathrm{T}}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) + \tfrac{1}{2}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)^{\mathrm{T}}\, \mathscr{H}\beta|_{\boldsymbol{\theta}_1}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$$

for some $\boldsymbol{\theta}_1$ with $\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_0\| < \|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\|$. Without ambiguity we hereby drop the arguments of $Q_n(\mathscr{X}; \hat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}_0)$ and $\mathbf{A}_n(\mathscr{X}, \boldsymbol{\theta}_0)$ for simplicity. So

$$\beta(\hat{\boldsymbol{\theta}}_n) - \beta(\boldsymbol{\theta}_0) = \mathbf{S}_n^{\mathrm{T}}\, \mathbf{A}_n \quad \text{a.s.}$$

where

$$\mathbf{S}_n^{\mathrm{T}} = \mathbf{S}_n(\mathscr{X}; \hat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}_0)^{\mathrm{T}} = \nabla\beta|_{\boldsymbol{\theta}_0}^{\mathrm{T}}\, Q_n + \tfrac{1}{2}\mathbf{A}_n^{\mathrm{T}}\, Q_n^{\mathrm{T}}\, \mathscr{H}\beta|_{\boldsymbol{\theta}_1}\, Q_n \to \nabla\beta|_{\boldsymbol{\theta}_0}^{\mathrm{T}}\, J_F(\boldsymbol{\theta}_0)^{-1} \quad \text{a.s.}$$

since $\|\mathscr{H}\beta|_{\boldsymbol{\theta}_1}\|$ is bounded in $N(\boldsymbol{\theta}_0)$ by (A 6).

We know by the Central Limit Theorem (CLT) that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \psi_F(X_i) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} [\psi_F(X_i) - \mathbf{E}_F(\psi_F(X))] \xrightarrow{\mathscr{D}} N(0, \sigma_F^2)$$

and $$\sqrt{n}\, \mathbf{A}_n(\mathscr{X}, \boldsymbol{\theta}_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \{\nabla \log f(X_i; \boldsymbol{\theta})|_{\boldsymbol{\theta}_0} - \mathbf{E}_F\, \nabla \log f(X; \boldsymbol{\theta})|_{\boldsymbol{\theta}_0}\}$$

$$\xrightarrow{\mathscr{D}} N_p(\mathbf{0}, K_F(\boldsymbol{\theta}_0))$$

where $K_F(\boldsymbol{\theta}_0) = \mathrm{Var}_F[\nabla \log f(X; \boldsymbol{\theta})|_{\boldsymbol{\theta}_0}]$ exists by (A 5). Therefore

$$\begin{aligned}\sqrt{n}\begin{pmatrix} \frac{1}{n}\sum_{i=1}^{n} \psi_F(X_i) \\ \beta(\hat{\boldsymbol{\theta}}_n) - \beta(\boldsymbol{\theta}_0)\end{pmatrix} &= \begin{pmatrix} 1 & \mathbf{0}^{\mathrm{T}} \\ 0 & \mathbf{S}_n^{\mathrm{T}}\end{pmatrix} \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \begin{pmatrix} \psi_F(X_i) \\ \nabla \log f(X_i; \boldsymbol{\theta})|_{\boldsymbol{\theta}_0}\end{pmatrix} \\ &\xrightarrow{\mathscr{D}} N_2\left(\mathbf{0}, \begin{pmatrix} 1 & \mathbf{0}^{\mathrm{T}} \\ 0 & \mathbf{S}_n^{\mathrm{T}}\end{pmatrix} \mathrm{Var}_F\left[\begin{pmatrix} \psi_F(X) \\ \nabla \log f(X; \boldsymbol{\theta})|_{\boldsymbol{\theta}_0}\end{pmatrix}\right] \begin{pmatrix} 1 & 0 \\ \mathbf{0} & \mathbf{S}_n\end{pmatrix}\right) \\ &= N_2(\mathbf{0}, \Sigma_F),\end{aligned} \tag{3}$$

where Σ_F is defined in the statement of the theorem. ∎

https://doi.org/10.1017/S0305004100072121 Published online by Cambridge University Press

Applying Theorem 1 to the bivariate error of the pure bootstrap estimates, we conclude

$$\sqrt{n}\begin{pmatrix}\nu_n\\ \omega_n\end{pmatrix} = \sqrt{n}\begin{pmatrix}\frac{1}{n}\sum_{i=1}^{n}\psi_F(X_i)\\ \beta(\hat{\boldsymbol{\theta}}_n)-\beta(\boldsymbol{\theta}_0)\end{pmatrix}+\sqrt{n}\begin{pmatrix}\mathrm{Rem}\,(\hat{F}_n,F)\\ \alpha(F_{\theta_0})-\alpha(F)\end{pmatrix}$$

with

$$\sqrt{n}\begin{pmatrix}\frac{1}{n}\sum_{i=1}^{n}\psi_F(X_i)\\ \beta(\hat{\boldsymbol{\theta}}_n)-\beta(\boldsymbol{\theta}_0)\end{pmatrix} \xrightarrow{\mathscr{D}} N_2(\mathbf{0},\Sigma_F). \tag{4}$$

The exact asymptotic behaviour of the bivariate error depends on correctness of our proposed parametric model as well as the remainder function $\mathrm{Rem}\,(\hat{F}_n,F)$ (which depends on the functional α).

2·4·2. *Asymptotic joint bootstrap distribution of pure bootstrap estimates.* Next we consider bootstrap resampled version of the bivariate error of the pure bootstrap estimates, namely

$$\sqrt{n}\begin{pmatrix}\alpha(\hat{F}_n^*)-\alpha(\hat{F}_n)\\ \alpha(F_{\hat{\theta}_n^*})-\alpha(\hat{F}_n)\end{pmatrix} = \sqrt{n}\begin{pmatrix}\mu_n^*\\ \eta_n^*-\nu_n+\omega_n\end{pmatrix}.$$

Note that

$$\sqrt{n}\begin{pmatrix}\alpha(\hat{F}_n^*)-\alpha(\hat{F}_n)\\ \alpha(F_{\hat{\theta}_n^*})-\alpha(\hat{F}_n)\end{pmatrix} = \sqrt{n}\begin{pmatrix}\frac{1}{n}\sum_{i=1}^{n}\psi_{\hat{F}_n}(X_i^*)\\ \beta(\hat{\boldsymbol{\theta}}_n^*)-\beta(\hat{\boldsymbol{\theta}}_n)\end{pmatrix}+\sqrt{n}\begin{pmatrix}\mathrm{Rem}\,(\hat{F}_n^*,\hat{F}_n)\\ \alpha(F_{\hat{\theta}_n^*})-\alpha(\hat{F}_n)\end{pmatrix}.$$

A similar argument to that used in Section 2·4·1 yields the bootstrap version of Theorem 1, given as follows.

THEOREM 2. *The joint bootstrap distribution of*

$$n^{-\frac{1}{2}}\sum_{i=1}^{n}\psi_{\hat{F}_n}(X_i^*) \quad \textit{and} \quad n^{\frac{1}{2}}(\beta(\hat{\boldsymbol{\theta}}_n^*)-\beta(\hat{\boldsymbol{\theta}}_n))$$

conditional on $\mathscr{X}$ converges weakly to a bivariate normal distribution with zero mean and variance covariance matrix Σ_F given in Theorem 1, along almost all sample sequences $X_1, X_2, \ldots$.

Proof of Theorem 2. The proof follows closely that of Theorem 1. Take $\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_n$ in (1) which is valid because $\hat{\boldsymbol{\theta}}_n \in N(\boldsymbol{\theta}_0)$ a.s. for sufficiently large n. Also, replace $\mathscr{X}$ by $\mathscr{X}^*$. Thus

$$\mathbf{A}_n(\mathscr{X}^*,\boldsymbol{\lambda}) = \mathbf{A}_n(\mathscr{X}^*,\hat{\boldsymbol{\theta}}_n)+B_n(\mathscr{X}^*,\hat{\boldsymbol{\theta}}_n)\,(\boldsymbol{\lambda}-\hat{\boldsymbol{\theta}}_n)+\tfrac{1}{2}C_n(\mathscr{X}^*)\,W_n(\mathscr{X}^*;\boldsymbol{\lambda},\hat{\boldsymbol{\theta}}_n)\,(\boldsymbol{\lambda}-\hat{\boldsymbol{\theta}}_n)$$

for all $\boldsymbol{\lambda}\in N(\boldsymbol{\theta}_0)$. Since $\hat{\boldsymbol{\theta}}_n$ can be seen as a functional of $\hat{F}_n$, applying Lemma B(ii) together with (A 5) and (A 8), we get

$$\mathbf{A}_n(\mathscr{X}^*,\hat{\boldsymbol{\theta}}_n)\,|\,X_1,\ldots,X_n \xrightarrow{\mathrm{p}} \mathbf{E}_F\,\nabla \log f(X;\boldsymbol{\theta})|_{\boldsymbol{\theta}_0} = \mathbf{0} \quad \text{a.s.}$$

and

$$B_n(\mathscr{X}^*,\hat{\boldsymbol{\theta}}_n)\,|\,X_1,\ldots,X_n \xrightarrow{\mathrm{p}} \mathbf{E}_F\,\mathscr{H} \log f(X;\boldsymbol{\theta})|_{\boldsymbol{\theta}_0} = -J_F(\boldsymbol{\theta}_0) \quad \text{a.s.}$$

Moreover, using Lemma B(i) with (A 4), we have

$$C_n(\mathscr{X}^*)\,|\,X_1,\ldots,X_n \xrightarrow{\mathrm{p}} \mathbf{E}_F H(X) \quad \text{a.s.}$$

https://doi.org/10.1017/S0305004100072121 Published online by Cambridge University Press

All these convergences together with the fact that $W_n(\mathscr{X}^*;\boldsymbol{\lambda},\hat{\boldsymbol{\theta}}_n) = O(\|\boldsymbol{\lambda}-\hat{\boldsymbol{\theta}}_n\|)$ a.s. imply that there exists a sequence $\hat{\boldsymbol{\theta}}_n^*$ such that

$$\hat{\boldsymbol{\theta}}_n^*-\hat{\boldsymbol{\theta}}_n \,|\, X_1,\ldots,X_n \xrightarrow{\mathrm{p}} \mathbf{0} \quad \text{a.s.}$$

and
$$\mathbf{P}[\mathbf{A}_n(\mathscr{X}^*,\hat{\boldsymbol{\theta}}_n^*)=\mathbf{0}\,|\,X_1,\ldots,X_n]\to 1 \quad \text{a.s.}$$

or in abbreviation,

$$\mathbf{A}_n(\mathscr{X}^*,\hat{\boldsymbol{\theta}}_n^*)=\mathbf{0} \quad \text{with conditional probability (w.c.p.)} \to 1 \quad \text{a.s.}$$

Hence
$$W_n(\mathscr{X}^*;\hat{\boldsymbol{\theta}}_n^*,\hat{\boldsymbol{\theta}}_n)\,|\,X_1,\ldots,X_n \xrightarrow{\mathrm{p}} 0 \quad \text{a.s.}$$

Thus
$$\hat{\boldsymbol{\theta}}_n^*-\hat{\boldsymbol{\theta}}_n = Q_n(\mathscr{X}^*;\hat{\boldsymbol{\theta}}_n^*,\hat{\boldsymbol{\theta}}_n)\,\mathbf{A}_n(\mathscr{X}^*,\hat{\boldsymbol{\theta}}_n) \quad \text{w.c.p.}\to 1 \quad \text{a.s.}$$

where
$$Q_n(\mathscr{X}^*;\hat{\boldsymbol{\theta}}_n^*,\hat{\boldsymbol{\theta}}_n)\,|\,X_1,\ldots,X_n \xrightarrow{\mathrm{p}} J_F(\boldsymbol{\theta}_0)^{-1} \quad \text{a.s.}$$

Therefore
$$\beta(\hat{\boldsymbol{\theta}}_n^*)-\beta(\hat{\boldsymbol{\theta}}_n) = \mathbf{S}_n^{*\mathrm{T}}\,\mathbf{A}_n(\mathscr{X}^*,\hat{\boldsymbol{\theta}}_n) \quad \text{w.c.p.}\to 1 \quad \text{a.s.}$$

where

$$\mathbf{S}_n^{*\mathrm{T}} = \nabla\beta|_{\hat{\boldsymbol{\theta}}_n}^{\mathrm{T}}\,Q_n(\mathscr{X}^*;\hat{\boldsymbol{\theta}}_n^*,\hat{\boldsymbol{\theta}}_n)+\tfrac{1}{2}\mathbf{A}_n^{\mathrm{T}}(\mathscr{X}^*,\hat{\boldsymbol{\theta}}_n)\,Q_n^{\mathrm{T}}(\mathscr{X}^*;\hat{\boldsymbol{\theta}}_n^*,\hat{\boldsymbol{\theta}}_n)\,\mathscr{H}\beta|_{\boldsymbol{\theta}_1^*}\,Q_n(\mathscr{X}^*;\hat{\boldsymbol{\theta}}_n^*,\hat{\boldsymbol{\theta}}_n)$$

with
$$\|\boldsymbol{\theta}_1^*-\hat{\boldsymbol{\theta}}_n\| < \|\hat{\boldsymbol{\theta}}_n^*-\hat{\boldsymbol{\theta}}_n\| \quad \text{w.c.p.}\to 1 \quad \text{a.s.}$$

Note that $\hat{\boldsymbol{\theta}}_n\to\boldsymbol{\theta}_0$ a.s. and $\nabla\beta(\boldsymbol{\theta})$ is continuous in $\boldsymbol{\theta}$. Thus,

$$\nabla\beta|_{\hat{\boldsymbol{\theta}}_n}^{\mathrm{T}} \to \nabla\beta|_{\boldsymbol{\theta}_0}^{\mathrm{T}} \quad \text{a.s.}$$

and
$$\mathbf{S}_n^{*\mathrm{T}}\,|\,X_1,\ldots,X_n \xrightarrow{\mathrm{p}} \nabla\beta|_{\boldsymbol{\theta}_0}^{\mathrm{T}}\,J_F(\boldsymbol{\theta}_0)^{-1} \quad \text{a.s.}$$

According to Lemma B(ii) and assumptions (A 2) and (A 8), we have also

$$\frac{1}{n}\sum_{i=1}^{n}\psi_{\hat{F}_n}(X_i^*)\,|\,X_1,\ldots,X_n \xrightarrow{\mathrm{p}} \mathbf{E}_F\,\psi_F(X) \quad \text{a.s.}$$

Note that
$$\mathbf{E}_F[\psi_F^2(X)+\|\nabla\log f(X;\boldsymbol{\theta})|_{\boldsymbol{\theta}_0}\|^2] < \infty$$

by (A 2) and (A 5), and

$$\frac{1}{n}\sum_{i=1}^{n}\left\|\begin{pmatrix}\psi_{\hat{F}_n}(X_i)\\ \nabla\log f(X_i;\boldsymbol{\theta})|_{\hat{\boldsymbol{\theta}}_n}\end{pmatrix}-\begin{pmatrix}\psi_F(X_i)\\ \nabla\log f(X_i;\boldsymbol{\theta})|_{\boldsymbol{\theta}_0}\end{pmatrix}\right\|^2 \to 0 \quad \text{a.s.}$$

by (A 8). Hence Lemma A(ii) implies

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\begin{pmatrix}\psi_{\hat{F}_n}(X_i^*)\\ \nabla\log f(X_i^*;\boldsymbol{\theta})|_{\hat{\boldsymbol{\theta}}_n}\end{pmatrix}|X_1,\ldots,X_n \xrightarrow{\mathscr{D}} N_{1+p}\left(\mathbf{0},\mathrm{Var}_F\begin{bmatrix}\psi_F(X)\\ \nabla\log f(X;\boldsymbol{\theta})|_{\boldsymbol{\theta}_0}\end{bmatrix}\right) \quad \text{a.s.}$$

Therefore

$$\sqrt{n}\begin{pmatrix}\frac{1}{n}\sum_{i=1}^{n}\psi_{\hat{F}_n}(X_i^*)\\ \beta(\hat{\boldsymbol{\theta}}_n^*)-\beta(\hat{\boldsymbol{\theta}}_n)\end{pmatrix}|X_1,\ldots,X_n = \begin{pmatrix}1 & \mathbf{0}^{\mathrm{T}}\\ 0 & \mathbf{S}_n^{*\mathrm{T}}\end{pmatrix}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\begin{pmatrix}\psi_{\hat{F}_n}(X_i^*)\\ \nabla\log f(X_i^*;\boldsymbol{\theta})|_{\hat{\boldsymbol{\theta}}_n}\end{pmatrix}|X_1,\ldots,X_n$$

$$\xrightarrow{\mathscr{D}} N_2(\mathbf{0},\Sigma_F) \quad \text{a.s.}$$

This completes the proof. ∎

https://doi.org/10.1017/S0305004100072121 Published online by Cambridge University Press

Thus, by Theorem 2, we have a result analogous to (4):

$$\sqrt{n}\begin{pmatrix}\mu_n^* \\ \eta_n^* - \nu_n + \omega_n\end{pmatrix} = \sqrt{n}\begin{pmatrix}\frac{1}{n}\sum_{i=1}^{n}\psi_{\hat{F}_n}(X_i^*) \\ \beta(\hat{\boldsymbol{\theta}}_n^*) - \beta(\hat{\boldsymbol{\theta}}_n)\end{pmatrix} + \sqrt{n}\begin{pmatrix}\mathrm{Rem}\,(\hat{F}_n^*, \hat{F}_n) \\ \alpha(F_{\hat{\boldsymbol{\theta}}_n}) - \alpha(\hat{F}_n)\end{pmatrix}$$

with
$$\sqrt{n}\begin{pmatrix}\frac{1}{n}\sum_{i=1}^{n}\psi_{\hat{F}_n}(X_i^*) \\ \beta(\hat{\boldsymbol{\theta}}_n^*) - \beta(\hat{\boldsymbol{\theta}}_n)\end{pmatrix} | X_1, \ldots, X_n \xrightarrow{\mathscr{D}} N_2(\mathbf{0}, \Sigma_F) \quad \text{a.s.} \tag{5}$$

2·4·3. *Asymptotic distribution of optimal ϵ-hybrid estimators.* For convenience we hereby let

$$\begin{pmatrix}U \\ V\end{pmatrix} \sim N(\mathbf{0}, \Sigma_F)$$

be a bivariate random variable having the same distribution as the limit obtained in (5). Define a function $\zeta: \mathbf{R} \to [0, 1]$ as

$$\zeta(x) = \begin{cases} 0 & \text{if } x < 0, \\ x & \text{if } 0 \leqslant x \leqslant 1, \\ 1 & \text{if } x > 1. \end{cases}$$

Note that ζ is a bounded continuous function on $\mathbf{R}$. Define also

$$\Upsilon_F \equiv \frac{\mathbf{E}[V(V-U)]}{\mathbf{E}[(V-U)^2]}.$$

The next theorem summarizes the main results of this paper.

THEOREM 3. *Under assumptions* (A 1) *to* (A 9), *the theoretical and empirical optimal tuning parameters and ϵ-hybrid estimators satisfy the following*:

(i) *Suppose $\gamma = 0$, that is, $\alpha(F_{\boldsymbol{\theta}_0}) = \alpha(F)$, which covers the case where the postulated parametric model is correct. Then we have, as $n \to \infty$,*

$$\epsilon_n \longrightarrow \zeta(\Upsilon_F) \quad \textit{and} \quad \hat{\epsilon}_n \xrightarrow{\mathscr{D}} \zeta\left(\frac{Z + \Upsilon_F}{Z+1}\right),$$

where Z is a random variable with a χ_1^2 distribution. Also, the corresponding ϵ-hybrid estimators have the following asymptotic distributions,

$$\sqrt{n}\,[T_{n,\epsilon_n} - \alpha(F)] \xrightarrow{\mathscr{D}} V + (U - V)\,\zeta(\Upsilon_F)$$

and
$$\sqrt{n}\,[T_{n,\hat{\epsilon}_n} - \alpha(F)] \xrightarrow{\mathscr{D}} V + (U-V)\,\zeta\left(\frac{(V-U)^2 + \mathbf{E}[V(V-U)]}{(V-U)^2 + \mathbf{E}[(V-U)^2]}\right),$$

as $n \to \infty$.

(ii) *Suppose $\gamma \neq 0$, that is, $\alpha(F_{\boldsymbol{\theta}_0}) \neq \alpha(F)$. Then we have, as $n \to \infty$,*

$$\epsilon_n = 1 + o(n^{-\frac{1}{2}}) \quad \textit{and} \quad \hat{\epsilon}_n = 1 + o_{\mathrm{p}}(n^{-\frac{1}{2}}).$$

The corresponding ϵ-hybrid estimators satisfy

$$\sqrt{n}\,[T_{n,\epsilon_n} - \alpha(F)] \xrightarrow{\mathscr{D}} U \quad \textit{and} \quad \sqrt{n}\,[T_{n,\hat{\epsilon}_n} - \alpha(F)] \xrightarrow{\mathscr{D}} U,$$

as $n \to \infty$.

https://doi.org/10.1017/S0305004100072121 Published online by Cambridge University Press

For $\gamma = 0$, Theorem 3 (i) shows that our asymptotically optimal ϵ is $\zeta(\Upsilon_F)$. Both the true and empirical optimal ϵ-hybrid estimates have a rate of convergence of order $O(n^{-\frac{1}{2}})$. While the true optimal T_{n,ϵ_n} is asymptotically normal, its empirical version $T_{n,\hat{\epsilon}_n}$ has an asymptotic distribution very close to normal with a slightly greater dispersion. On the other hand, if $\gamma \neq 0$, the theoretical optimal tuning parameter $\epsilon_n \to 1$ as $n \to \infty$ and our asymptotically optimal choice is therefore NB. This sounds reasonable as we have postulated a wrong model for the underlying distribution and a non-parametric estimate is therefore intuitively superior. Our empirical optimal $\hat{\epsilon}_n$ is consistent with the theoretical optimal ϵ_n as n tends to infinity. The theoretical and empirical estimates, T_{n,ϵ_n} and $T_{n,\hat{\epsilon}_n}$, are consistent with each other asymptotically and have the same asymptotic distribution $N(0, \sigma_F^2)$ as does NB.

Proof of Theorem 3. Suppose first that $\gamma = 0$. From (4) and (A 7), we have by Slutsky's Theorem,

$$\sqrt{n}\begin{pmatrix}\nu_n\\ \omega_n\end{pmatrix} \xrightarrow{\mathcal{D}} \begin{pmatrix}U\\ V\end{pmatrix}. \tag{6}$$

Consider $\quad \mathrm{MSE}_F(T_{n,\epsilon}) = \epsilon^2 \mathbf{E}_F[\nu_n^2] + (1-\epsilon)^2 \mathbf{E}_F[\omega_n^2] + 2\epsilon(1-\epsilon)\,\mathbf{E}_F[\nu_n \omega_n].$

This is a simple quadratic polynomial in ϵ and is minimized over $\epsilon \in [0,1]$ at

$$\epsilon = \epsilon_n = \epsilon_n(F) = \zeta\left(\frac{\mathbf{E}_F[\omega_n(\omega_n - \nu_n)]}{\mathbf{E}_F[(\omega_n - \nu_n)^2]}\right),$$

provided the denominator is non-zero, or equivalently, PB and NB coincide with probability less than 1 under F. We can assume so because the case where PB and NB are almost everywhere identical is not of interest. The result $\epsilon_n \to \zeta(\Upsilon_F)$ follows by using (6) and (A 9). Similarly, we obtain an expression for the bootstrap estimate of ϵ_n,

$$\hat{\epsilon}_n = \zeta\left(\frac{\mathbf{E}_{\hat{F}_n}[(\eta_n^* + \omega_n - \nu_n)(\eta_n^* - \mu_n^* + \omega_n - \nu_n)]}{\mathbf{E}_{\hat{F}_n}[(\eta_n^* - \mu_n^* + \omega_n - \nu_n)^2]}\right). \tag{7}$$

Substituting ϵ_n and $\hat{\epsilon}_n$ for ϵ respectively in $T_{n,\epsilon}$, we have

$$\sqrt{n}\,[T_{n,\epsilon_n} - \alpha(F)] = \sqrt{n}\,[\epsilon_n(\nu_n - \omega_n) + \omega_n] \tag{8}$$

and

$$\sqrt{n}\,[T_{n,\hat{\epsilon}_n} - \alpha(F)] = \sqrt{n}\,[\hat{\epsilon}_n(\nu_n - \omega_n) + \omega_n]. \tag{9}$$

Note that all the quantities (7), (8) and (9) are continuous functions of $\sqrt{n}\,(\nu_n, \omega_n)$, $n\mathbf{E}_{\hat{F}_n}[\mu_n^{*2}]$, $n\mathbf{E}_{\hat{F}_n}[\eta_n^{*2}]$, $n\mathbf{E}_{\hat{F}_n}[\mu_n^* \eta_n^*]$, $\sqrt{n}\,\mathbf{E}_{\hat{F}_n}[\mu_n^*]$, and $\sqrt{n}\,\mathbf{E}_{\hat{F}_n}[\eta_n^*]$ with

$$\sqrt{n}\begin{pmatrix}\nu_n\\ \omega_n\end{pmatrix} \xrightarrow{\mathcal{D}} \begin{pmatrix}U\\ V\end{pmatrix}$$

and

$$\begin{pmatrix} n\mathbf{E}_{\hat{F}_n}[\mu_n^{*2}] \\ n\mathbf{E}_{\hat{F}_n}[\eta_n^{*2}] \\ n\mathbf{E}_{\hat{F}_n}[\mu_n^* \eta_n^*] \\ \sqrt{n}\,\mathbf{E}_{\hat{F}_n}[\mu_n^*] \\ \sqrt{n}\,\mathbf{E}_{\hat{F}_n}[\eta_n^*] \end{pmatrix} \longrightarrow \begin{pmatrix} \mathbf{E}[U^2] \\ \mathbf{E}[V^2] \\ \mathbf{E}[UV] \\ 0 \\ 0 \end{pmatrix} \text{ a.s.} \tag{10}$$

https://doi.org/10.1017/S0305004100072121 Published online by Cambridge University Press

The almost sure convergence (10) follows from (A 9) and noting that (A 7) and (5) together imply

$$\sqrt{n}\begin{pmatrix}\mu_n^* \\ \eta_n^*\end{pmatrix} | X_1, \ldots, X_n \xrightarrow{\mathscr{D}} \begin{pmatrix}U \\ V\end{pmatrix} \quad \text{a.s.}$$

by Slutsky's Theorem.

Hence, using Slutsky's Theorem again on (7), (8) and (9) and putting

$$Z = \frac{(V-U)^2}{\mathrm{E}[(V-U)^2]} \sim \chi_1^2,$$

the required convergence results for $\hat{\epsilon}_n$ and the optimal ϵ-hybrid estimators follow immediately.

Next we look at the case where $\gamma \neq 0$. Application of Slutsky's Theorem to (4) and (A 7) now gives us

$$\sqrt{n}\begin{pmatrix}\nu_n \\ \omega_n - \gamma\end{pmatrix} \xrightarrow{\mathscr{D}} \begin{pmatrix}U \\ V\end{pmatrix}. \tag{11}$$

By (A 9) we have $n\nu_n^2$ and $n(\omega_n-\gamma)^2$ to be uniformly integrable. Therefore ω_n^2 is also uniformly integrable. Thus

$$\mathbf{E}_F[(\omega_n-\nu_n)^2] = \gamma^2 + o(1)$$

and

$$\sqrt{n}\,\mathbf{E}_F[\nu_n(\omega_n-\nu_n)] = \mathbf{E}_F[\sqrt{n}\,\nu_n\,\omega_n] - \mathbf{E}_F[\sqrt{n}\,\nu_n^2] = o(1)$$

since $\omega_n \xrightarrow{\mathrm{p}} \gamma$ and $\sqrt{n}\,\nu_n \xrightarrow{\mathscr{D}} U$. Therefore we have

$$\epsilon_n = \zeta\left(1 + \frac{\mathbf{E}_F[\nu_n(\omega_n-\nu_n)]}{\mathbf{E}_F[(\omega_n-\nu_n)^2]}\right) = 1 + o(n^{-\frac{1}{2}}).$$

The almost sure convergence (10) still holds as the value of γ does not affect the asymptotic behaviour of the random quantities in question. Thus we can deduce

$$\begin{aligned}
\mathbf{E}_{\hat{F}_n}[\mu_n^*(\eta_n^*-\mu_n^*)] &= n^{-1}\mathbf{E}[U(V-U)] + o(n^{-1}) \quad \text{a.s.},\\
\mathbf{E}_{\hat{F}_n}[(\eta_n^*-\mu_n^*)^2] &= n^{-1}\mathbf{E}[(V-U)^2] + o(n^{-1}) \quad \text{a.s.},\\
\mathbf{E}_{\hat{F}_n}[\eta_n^*] &= o(n^{-\frac{1}{2}}) \quad \text{a.s.},\\
\mathbf{E}_{\hat{F}_n}[\mu_n^*] &= o(n^{-\frac{1}{2}}) \quad \text{a.s.}
\end{aligned}$$

and hence

$$\hat{\epsilon}_n = 1 + o_{\mathrm{p}}(n^{-\frac{1}{2}}),$$

using expression (7) again. Lastly we consider the actual hybrid estimates T_{n,ϵ_n} and $T_{n,\hat{\epsilon}_n}$. Clearly,

$$T_{n,\epsilon_n} - \alpha(F) = \epsilon_n\,\nu_n + (1-\epsilon_n)\,\omega_n = \nu_n + o_{\mathrm{p}}(n^{-\frac{1}{2}}),$$

and

$$T_{n,\hat{\epsilon}_n} - \alpha(F) = \nu_n + o_{\mathrm{p}}(n^{-\frac{1}{2}}).$$

This completes the whole proof. ∎

Now we look at a particular case where $\Theta \subseteq \mathbf{R}$ with $F = F_{\theta_0}$, so that our postulated parametric model is correct. We want to find out what kind of conditions guarantee the convergence of the optimal ϵ_n to 0, or in other words, PB being our asymptotically optimal choice. First we need a lemma as stated below, the proof of which follows that of Theorem 8·1·3 in [**4**].

https://doi.org/10.1017/S0305004100072121 Published online by Cambridge University Press

Let $\mathscr{C}(\bar{\mathbf{R}})$ be the space of continuous real-valued functions on $\mathbf{R}$ with finite limits at $\pm\infty$. Equip $\mathscr{C}(\bar{\mathbf{R}})$ with the sup-norm $\|\cdot\|_\infty$.

LEMMA C. *Suppose the mapping $\theta \mapsto F_\theta$ from Θ to $\mathscr{C}(\bar{\mathbf{R}})$ is continuous. Suppose also that α is a statistical functional defined on $\mathscr{C}(\bar{\mathbf{R}})$. Assume α is Hadamard differentiable at F_{θ_0} for some $\theta_0 \in \Theta$, such that*

(i) *F_{θ_0} is strictly increasing, and*

(ii) *as $\delta \to 0$,*

$$\frac{f_{\theta_0+\delta} - f_{\theta_0}}{\delta f_{\theta_0}} \to \frac{\partial}{\partial\theta} \log f_\theta \Big|_{\theta_0} \quad \text{in } \mathscr{L}^2(F_{\theta_0}),$$

where $f_\theta = F'_\theta$ exists and $\mathscr{L}^2(F_{\theta_0})$ is the usual integrated squared norm with respect to the measure F_{θ_0} on a suitable space of real functions. Then we have

$$\alpha(F_{\theta_0+\delta}) - \alpha(F_{\theta_0}) - \int \psi_{F_{\theta_0}}(x)\, dF_{\theta_0+\delta}(x) = o(\delta).$$

Proof of Lemma C. Since α is Hadamard differentiable at F_{θ_0}, its derivative $\alpha'_{F_{\theta_0}} : \mathscr{C}(\bar{\mathbf{R}}) \to \mathbf{R}$ exists, where

$$\alpha'_{F_{\theta_0}}(G - F_{\theta_0}) = \int \psi_{F_{\theta_0}}(x)\, dG(x) \quad \forall G \in \mathscr{C}(\bar{\mathbf{R}}).$$

Define a function $g : \mathbf{R} \to \mathbf{R}$ as

$$g(x) = \int_{-\infty}^{x} \frac{\partial}{\partial\theta} \log f_\theta(t) \Big|_{\theta_0} dF_{\theta_0}(t).$$

By the regularity properties of $\{F_\theta\}$, g is well-defined and $g \in \mathscr{C}(\bar{\mathbf{R}})$. Now define

$$H_\delta = \begin{cases} \dfrac{F_{\theta_0+\delta} - F_{\theta_0}}{\delta} & \text{if } \delta \neq 0, \\ g & \text{if } \delta = 0. \end{cases}$$

Consider the function space

$$K = \{H_\delta : \delta \in [-1, 1]\}$$

and the mapping $\phi : [-1, 1] \to \mathscr{C}(\bar{\mathbf{R}})$ such that $\phi(\delta) = H_\delta$. We claim that ϕ is continuous on $[-1, 1]$. For $\delta \neq 0$, H_δ is obviously continuous in δ since $\theta \mapsto F_\theta$ is assumed to be continuous. So ϕ is continuous at all $\delta \neq 0$. Moreover, using the Cauchy–Schwarz inequality and assumption (ii), we have

$$\begin{aligned} \|H_\delta - g\|_\infty &= \sup_{x \in \mathbf{R}} \left| \frac{F_{\theta_0+\delta}(x) - F_{\theta_0}(x)}{\delta} - \int_{-\infty}^{x} \frac{\partial}{\partial\theta} \log f_\theta(t) \Big|_{\theta_0} dF_{\theta_0}(t) \right| \\ &\leqslant \sup_{x \in \mathbf{R}} \left[\int \mathbf{1}\{t \leqslant x\}\, dF_{\theta_0}(t) \right]^{\frac{1}{2}} \left\| \frac{f_{\theta_0+\delta} - f_{\theta_0}}{\delta f_{\theta_0}} - \frac{\partial}{\partial\theta} \log f_\theta \Big|_{\theta_0} \right\|_{\mathscr{L}^2(F_{\theta_0})} \to 0 \end{aligned}$$

as $\delta \to 0$. Therefore ϕ is continuous at $\delta = 0$ as well and our claim is valid. Thus $K = \phi[-1, 1]$ is compact. By Hadamard differentiability of α at F_{θ_0},

$$\frac{\alpha(F_{\theta_0} + \delta H) - \alpha(F_{\theta_0}) - \alpha'_{F_{\theta_0}}(\delta H)}{\delta} \to 0 \quad \text{as } \delta \to 0$$

https://doi.org/10.1017/S0305004100072121 Published online by Cambridge University Press

uniformly for all $H \in K$. Therefore

$$\frac{\alpha(F_{\theta_0+\delta})-\alpha(F_{\theta_0})-\alpha'_{F_{\theta_0}}(F_{\theta_0+\delta}-F_{\theta_0})}{\delta} = \frac{\alpha(F_{\theta_0}+\delta H_\delta)-\alpha(F_{\theta_0})-\alpha'_{F_{\theta_0}}(\delta H_\delta)}{\delta} \to 0 \quad \text{as } \delta \to 0.$$

Hence the result follows. ∎

The following theorem characterizes a common situation in which PB is the asymptotically optimal choice.

THEOREM 4. *Suppose* $\Theta \subseteq \mathbf{R}$ *and* $F = F_{\theta_0}$. *Assume conditions prescribed in Lemma C. Then we have* $\Upsilon_{F_{\theta_0}} = 0$. *Hence*

$$\epsilon_n \to 0 \quad \textit{and} \quad \hat{\epsilon}_n \xrightarrow{\mathscr{D}} \frac{Z}{Z+1}$$

with $$Z \sim \chi^2_1.$$

Proof of Theorem 4. Using notation given in Theorem 1, we have clearly

$$J_{F_\theta}(\theta) = K_{F_\theta}(\theta),$$

and $$\rho_{F_{\theta_0}} = \int \psi_{F_{\theta_0}}(x) \frac{\partial}{\partial \theta} \log f(x;\theta)\Big|_{\theta_0} f(x;\theta_0)\,dx = \frac{\partial}{\partial \theta}\int \psi_{F_{\theta_0}}(x)\,dF_\theta(x)\Big|_{\theta_0}.$$

Suppose all the conditions prescribed in Lemma C hold. Then we have

$$\begin{aligned}\mathbf{E}[V(V-U)] &= \frac{\partial \alpha(F_\theta)}{\partial \theta}\Big|_{\theta_0} J_{F_{\theta_0}}(\theta_0)^{-1}\left[\frac{\partial \alpha(F_\theta)}{\partial \theta}\Big|_{\theta_0} - \rho_{F_{\theta_0}}\right] \\ &= \frac{\partial \alpha(F_\theta)}{\partial \theta}\Big|_{\theta_0} J_{F_{\theta_0}}(\theta_0)^{-1}\left\{\lim_{\delta \to 0}\frac{\alpha(F_{\theta_0+\delta})-\alpha(F_{\theta_0})-\alpha'_{F_{\theta_0}}(F_{\theta_0+\delta}-F_{\theta_0})}{\delta}\right\} = 0\end{aligned}$$

according to Lemma C. Therefore $\Upsilon_{F_{\theta_0}} = 0$. The convergence of the optimal tuning parameters follows from Theorem 3(i). ∎

Theorem 4 is a fairly general result since the conditions in Lemma C are usually fulfilled by reasonable classes of distributions. Clearly, the theoretical optimal T_{n,ϵ_n} is asymptotic to PB under the conditions of Theorem 4. However, the empirical estimate $T_{n,\hat{\epsilon}_n}$ still has a non-trivial limit distribution.

The case in which $\alpha(F) = \alpha(F_{\theta_0})$ and $F \neq F_{\theta_0}$ may give rise to quite different results. We shall come back to this in Section 3·2 where an example will be given to illustrate the non-trivial asymptotic behaviour in this case.

2·4·4. *Asymptotic MSE of optimal ε-hybrid estimators.*

Case 1. Suppose that $\gamma = 0$. According to assumption (A 9) we deduce that both $n[T_{n,\epsilon_n}-\alpha(F)]^2$ and $n[T_{n,\hat{\epsilon}_n}-\alpha(F)]^2$ are uniformly integrable. Therefore,

$$n\,\mathrm{MSE}_F(T_{n,\epsilon_n}) \to \mathbf{E}[\{V+(U-V)\,\zeta(\Upsilon_F)\}^2]$$

and similarly,

$$n\,\mathrm{MSE}_F(T_{n,\hat{\epsilon}_n}) \to \mathbf{E}\left[\left\{V+(U-V)\,\zeta\left(\frac{(V-U)^2+\mathbf{E}[V(V-U)]}{(V-U)^2+\mathbf{E}[(V-U)^2]}\right)\right\}^2\right].$$

https://doi.org/10.1017/S0305004100072121 Published online by Cambridge University Press

Now we consider three separate cases.

(i) $\Upsilon_F \geqslant 1$: Here $\zeta(\Upsilon_F) = 1$. Both the asymptotic errors satisfy

$$\left.\begin{aligned}\sqrt{n}\,[T_{n,\epsilon_n} - \alpha(F)]\\ \sqrt{n}\,[T_{n,\hat{\epsilon}_n} - \alpha(F)]\end{aligned}\right\} \xrightarrow{\mathscr{D}} U$$

and

$$\left.\begin{aligned}n\,\mathrm{MSE}_F(T_{n,\epsilon_n})\\ n\,\mathrm{MSE}_F(T_{n,\hat{\epsilon}_n})\end{aligned}\right\} \to \mathbf{E}[U^2].$$

Our asymptotically optimal choice is NB and the empirical version $T_{n,\hat{\epsilon}_n}$ is consistent with the true optimal estimate T_{n,ϵ_n} asymptotically.

(ii) $0 < \Upsilon_F < 1$: Then $\zeta(\Upsilon_F) = \Upsilon_F$. Here T_{n,ϵ_n} is asymptotically normal with variance strictly less than those of U and V since Υ_F is the unique and non-trivial $\epsilon \in (0,1)$ which minimizes the asymptotic variance of convex combinations of PB and NB. The asymptotic distribution and MSE of $T_{n,\hat{\epsilon}_n}$ are no less complicated than our previous general results. Note that $\mathrm{MSE}_F(T_{n,\hat{\epsilon}_n})$ is asymptotically slightly bigger than $\mathrm{MSE}_F(T_{n,\epsilon_n})$ owing to the more dispersed asymptotic distribution of $T_{n,\hat{\epsilon}_n}$, although they both have the same rate of convergence.

(iii) $\Upsilon_F \leqslant 0$: Then $\zeta(\Upsilon_F) = 0$. So now

$$\sqrt{n}\,[T_{n,\epsilon_n} - \alpha(F)] \xrightarrow{\mathscr{D}} V$$

and

$$n\,\mathrm{MSE}_F(T_{n,\epsilon_n}) \to \mathbf{E}[V^2].$$

PB is the asymptotically optimal choice. Again there is no simpler corresponding expression for $T_{n,\hat{\epsilon}_n}$.

Case 2. Suppose now $\gamma \neq 0$. Note that both the theoretical and empirical optimal ϵ-hybrid estimators have the same asymptotic distribution $N(0, \sigma_F^2)$ from Theorem 3(ii). We should expect their MSEs to be asymptotic to σ_F^2/n as well. This is certainly the case for T_{n,ϵ_n} since with probability 1, T_{n,ϵ_n} satisfies

$$n[T_{n,\epsilon_n} - \alpha(F)]^2 \leqslant 2n\nu_n^2 + 2[\sqrt{n}(1-\epsilon_n)]^2\,\omega_n^2,$$

where $\sqrt{n}\,(1-\epsilon_n) \to 0$. Therefore it is uniformly integrable and hence

$$n\,\mathrm{MSE}_F(T_{n,\epsilon_n}) \to \mathbf{E}[U^2] = \sigma_F^2.$$

Under the further condition that $n(1-\hat{\epsilon}_n)^2$ be uniformly integrable, we have also

$$n[T_{n,\hat{\epsilon}_n} - \alpha(F)]^2 \leqslant 2n\nu_n^2 + 4n(\omega_n - \gamma)^2 + 4n(1-\hat{\epsilon}_n)^2\,\gamma^2$$

to be uniformly integrable, so that

$$n\,\mathrm{MSE}_F(T_{n,\hat{\epsilon}_n}) \to \sigma_F^2.$$

Note that the condition that $n(1-\hat{\epsilon}_n)^2$ be uniformly integrable is usually satisfied under stricter moment conditions on the underlying distribution F.

2.5. *Summary and comments.* We have shown that PB is the asymptotically optimal estimate under some mild conditions if our proposed parametric model is correct, as stated in Theorem 4. However the empirical optimal $\hat{\epsilon}_n$ has a non-degenerate limit distribution corresponding to the random variable $Z/(Z+1)$, where $Z \sim \chi_1^2$. The density function of $Z/(Z+1)$ is given by the formula

$$f_{Z/(Z+1)}(x) = \{2\pi t(1-t)^3\}^{-\frac{1}{2}} \exp\left\{-\frac{t}{2(1-t)}\right\}.$$

https://doi.org/10.1017/S0305004100072121 Published online by Cambridge University Press

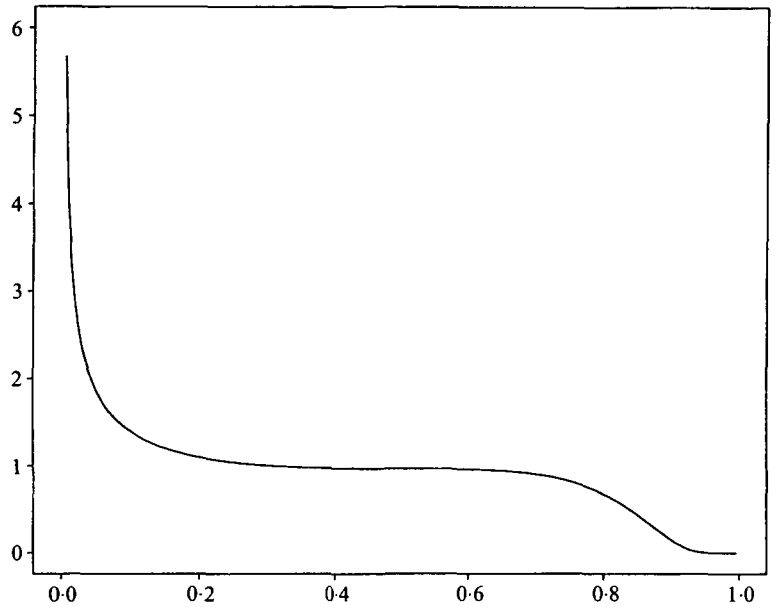


Fig. 1. Density function of $(Z/Z+1)$ where $Z \sim \chi^2_1$.

Figure 1 shows that this function is highly concentrated around zero, and is more or less evenly distributed over the rest of the interval [0, 1] with moderate density. This means we are quite likely to end up with a non-trivial convex combination of PB and NB as our hybrid estimate, where PB plays a much more substantial role in the combination. Nevertheless, $\mathrm{MSE}_F(T_{n,\hat{\epsilon}_n})$ proves to be only slightly bigger than $\mathrm{MSE}_F(T_{n,\epsilon_n})$ and both have the same convergence rate of order $O(n^{-1})$.

On the other hand, if the proposed parametric model is wrong and $\alpha(F_{\theta_0}) \neq \alpha(F)$, then we have $\hat{\epsilon}_n \xrightarrow{\mathrm{p}} 1$ and our empirical estimate $T_{n,\hat{\epsilon}_n}$ is exactly consistent with the true optimal T_{n,ϵ_n} which is asymptotic to NB. Similarly their MSEs are asymptotically identical.

One nice property of $T_{n,\hat{\epsilon}_n}$ is that whether or not our suggested parametric model is correct, it can adjust itself automatically to do the right thing. It should be borne in mind that the pure and hybrid bootstrap estimates considered so far have the same convergence rate of order $O(n^{-\frac{1}{2}})$. (In the case of PB, we need the assumption that our postulated model is correct.) This is due to the fact that these estimates are all based on the empirical distribution $\hat{F}_n$, which is known to be a $\sqrt{n}$-consistent estimate of F under certain regularity conditions. What the hybrid estimate does is to reduce the first order MSE size without changing the convergence rate.

There exist particular cases when $\alpha(F) = \alpha(F_{\theta_0})$ where T_{n,ϵ_n} is asymptotic to neither PB nor NB but a non-trivial convex combination of them. This means $T_{n,\hat{\epsilon}_n}$ can sometimes be not just a trade-off but a strictly better estimate than the pure bootstraps. Section 3·2 gives a particular example to illustrate this point.

Finally, a remark should be made on the computational requirements of our empirical optimal ϵ-hybrid estimator $T_{n,\hat{\epsilon}_n}$. Recall that the empirical optimal tuning parameter is given by

$$\hat{\epsilon}_n = \zeta\left(\frac{\mathbf{E}_{\hat{F}_n}[(\alpha(F_{\hat{\theta}_n^*}) - \alpha(\hat{F}_n))\,(\alpha(F_{\hat{\theta}_n^*}) - \alpha(\hat{F}_n^*))]}{\mathbf{E}_{\hat{F}_n}[(\alpha(F_{\hat{\theta}_n^*}) - \alpha(\hat{F}_n^*))^2]}\right).$$

https://doi.org/10.1017/S0305004100072121 Published online by Cambridge University Press

In general the functional α cannot be expressed as a directly computable formula. Bootstrap resampling is usually required to approximate the bootstrap estimates involved in the above formula for $\hat{\epsilon}_n$. The presence of quantities like $\alpha(\hat{F}_n^*)$ and $\alpha(F_{\hat{\theta}_n^*})$ generally calls for a double bootstrap procedure which is very time-consuming. There exist other non-parametric methods of estimating MSEs without the need for bootstrap resampling. For example, jackknife estimates of standard error and bias can be used to derive an estimate of $\mathrm{MSE}_F(\cdot)$. Alternatively, the delta method, similar to the jackknife, is also applicable here. See [3] for details of these estimation methods. It is shown that these methods can be regarded in some sense as approximations to the bootstrap method. We may thus expect some loss of accuracy from saving the requirement of a second level of bootstrapping.

3. *Examples of ϵ-hybrid estimator*

Two separate examples are considered here to verify the asymptotic theory of $T_{n,\epsilon}$, as developed in Section 2·4. Both examples focus on the estimation of variance with the negative exponential family indexed by the scale parameter θ as our proposed underlying parametric model. The second example illustrates the possibility of $T_{n,\hat{\epsilon}_n}$ being superior to either PB or NB.

Note that in the case where $\alpha(F)$ is the variance of F, there exist closed form expressions for $\alpha(F)$ so that no double bootstrap procedure is required. Moreover, the MSE of the corresponding parametric and non-parametric bootstrap estimates of $\alpha(F)$ can be expressed explicitly in terms of population moments, which allows us to compute the MSE exactly. The same is also true for the empirical optimal tuning parameter $\hat{\epsilon}_n$ which admits an explicit formula in terms of sample moments.

3·1. *Example* 1: *True distribution is Gamma.* Let $X_1, X_2, \ldots, X_n$ be a random sample drawn from F, taken to be a $\Gamma(\alpha, \theta)$ distribution, whose probability density function is given by

$$\frac{\theta^\alpha x^{\alpha-1} e^{-\theta x}}{\Gamma(\alpha)} \quad \text{for } x > 0,$$

where α and θ are the shape and scale parameters respectively. We are interested in estimating the functional $\alpha(F) \equiv \mathrm{Var}_F(X)$ with $X \sim F$. The parametric model contemplated here is the family of exponential distributions indexed by the scale parameter θ, that is,

$$\mathscr{F} = \{F_\theta : F_\theta(x) = \mathbf{1}_{\{x \geqslant 0\}}(1 - e^{-\theta x}), \quad \theta > 0\}.$$

The MLE of θ assuming $\mathscr{F}$ to be correct is easily found to be $\hat{\theta}_n \equiv 1/\bar{X}$. The two bootstrap estimates are then:

$$\alpha(F_{\hat{\theta}_n}) = \bar{X}^2, \tag{PB}$$

$$\alpha(\hat{F}_n) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2. \tag{NB}$$

Figure 2 shows the ratio of the respective mean squared errors (MSEs) under a Gamma distribution with varying α and $\theta = 1$ for different sample sizes n. It is found that PB outperforms NB if the true distribution is close enough to our proposed model, which is the case $\alpha = 1$.

https://doi.org/10.1017/S0305004100072121 Published online by Cambridge University Press

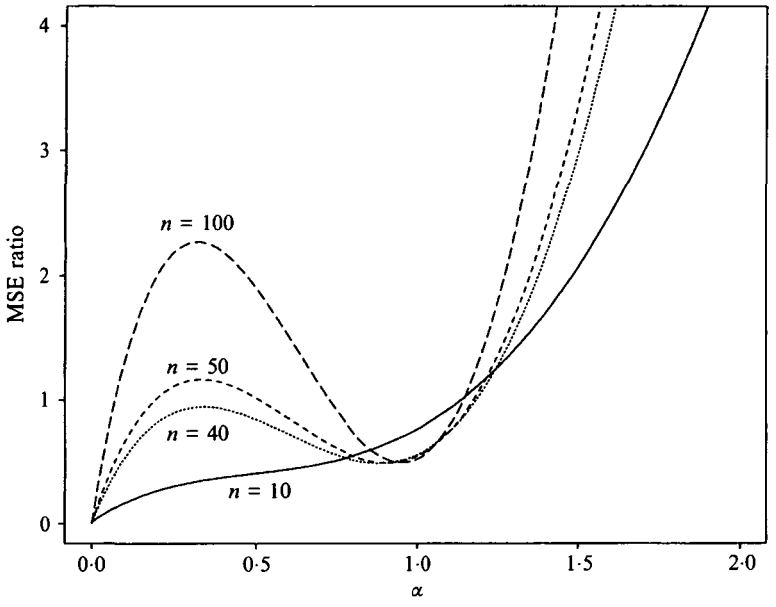


Fig. 2. Example 1. – Ratio of MSE(PB) to MSE(NB) under $\Gamma(\alpha, 1)$ for different sample sizes n.

3·1·1. *Theoretical asymptotic results.* First we apply the asymptotic theory developed in Section 2·4 to this particular example. Note that for any distributions F and G,

$$\alpha(G) = \alpha(F) + \int \psi_F(x)\, dG(x) + \text{Rem}\,(G, F),$$

where $$\psi_F(x) = x^2 - \mathbf{E}_F[X^2] - 2\mathbf{E}_F[X]\,(x - \mathbf{E}_F[X])$$

and $$\text{Rem}\,(G, F) = -(\mathbf{E}_G[X] - \mathbf{E}_F[X])^2.$$

Suppose the true underlying distribution F is $\Gamma(\alpha_1, \theta_1)$. The postulated parametric family is negative exponential so that $f(x;\theta) = \theta e^{-\theta x}$ for $x > 0$. Therefore, we have $\alpha(F) = \alpha_1/\theta_1^2$ and $\beta(\theta) \equiv \alpha(F_\theta) = \theta^{-2}$. The MLE of θ is obtained as $\hat{\theta}_n = 1/\bar{X}$. The nearest θ describing F is taken to be the solution θ_0 for θ to the equation

$$\mathbf{E}_F\left[\frac{\partial}{\partial\theta}\log f(X;\theta)\right] = 0.$$

The solution is given by $\theta_0 = \theta_1/\alpha_1$. It is easy to check that assumptions (A 1)–(A 9) are satisfied here. After a few simple calculations we obtain

$$\begin{pmatrix} U \\ V \end{pmatrix} \sim N(\mathbf{0}, \Sigma_F),$$

such that $$\Sigma_F = \frac{2\alpha_1}{\theta_1^4}\begin{pmatrix} \alpha_1 + 3 & 2\alpha_1 \\ 2\alpha_1 & 2\alpha_1^2 \end{pmatrix}.$$

The two cases $\gamma = 0$ and $\gamma \neq 0$ correspond to $\alpha_1 = 1$ and $\alpha_1 \neq 1$ respectively, since

$$\gamma \equiv \alpha(F_{\theta_0}) - \alpha(F) = \frac{\alpha_1(\alpha_1 - 1)}{\theta_1^2}.$$

https://doi.org/10.1017/S0305004100072121 Published online by Cambridge University Press

Case 1. Suppose first that $\alpha_1 = 1$. The optimal ϵ_n is given, for large n, by

$$\zeta\left(\frac{6+5n}{2(n^2-n+6)}\right) = \frac{5}{2n} + O(n^{-2}).$$

We have $$\Upsilon_F = \frac{\mathbf{E}[V(V-U)]}{\mathbf{E}[(V-U)^2]} = \frac{4\alpha_1^2(\alpha_1-1)}{4\alpha_1^3-6\alpha_1^2+6\alpha_1} = 0.$$

Therefore we deduce $$\epsilon_n \longrightarrow 0 \quad \text{and} \quad \hat{\epsilon}_n \xrightarrow{\mathscr{D}} \frac{Z}{Z+1},$$

where $Z \sim \chi_1^2$. Moreover, the asymptotic behaviour of T_{n,ϵ_n} and $T_{n,\hat{\epsilon}_n}$ is given by

$$\sqrt{n}\,[T_{n,\epsilon_n} - \alpha(F)] \xrightarrow{\mathscr{D}} V \sim N(0, 4\theta_1^{-4}),$$

and $$\sqrt{n}\,[T_{n,\hat{\epsilon}_n} - \alpha(F)] \xrightarrow{\mathscr{D}} V + \frac{(U-V)^3}{(U-V)^2+4\theta_1^{-4}},$$

where $$\begin{pmatrix} U \\ V \end{pmatrix} \sim N\left(\mathbf{0}, 4\theta_1^{-4}\begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}\right).$$

Case 2. Suppose now that $\alpha_1 \neq 1$. Here ϵ_n satisfies

$$\epsilon_n = 1 + \frac{1}{n}\left[\frac{15\alpha_1^2-17\alpha_1+6}{\alpha_1(\alpha_1-1)^2}\right] + O(n^{-2})$$

for sufficiently large n. Therefore we have

$$\sqrt{n}\,[T_{n,\epsilon_n} - \alpha(F)] \xrightarrow{\mathscr{D}} U \sim N\left(0, \frac{2\alpha_1(\alpha_1+3)}{\theta_1^4}\right).$$

A similar argument and the fact that

$$\mathbf{E}_{\hat{F}_n} \operatorname{Rem}(\hat{F}_n^*, \hat{F}_n) = \frac{1}{n}\left[\frac{1}{n}\sum_{i=1}^n X_i^2 - \left(\frac{1}{n}\sum_{i=1}^n X_i\right)^2\right] = O(n^{-1}) \quad \text{a.s.}$$

imply $$\hat{\epsilon}_n = 1 + O_p(n^{-1})$$

and $$\sqrt{n}\,[T_{n,\hat{\epsilon}_n} - \alpha(F)] \xrightarrow{\mathscr{D}} N(0, 2\alpha_1(\alpha_1+3)\,\theta_1^{-4}).$$

3·1·2. *Simulation results.* Here we take $\theta_1 = 1$ so that our parameter of interest has the value α_1. Two hundred data samples of size n were drawn from $\Gamma(\alpha_1, 1)$, for $n = 10, 10^2, 10^3, 10^4$ and $\alpha_1 = 0{\cdot}5, 1, 1{\cdot}5, 4$. For each n and α_1 the frequency distributions of $\sqrt{n}\,[T_{n,\epsilon_n} - \alpha_1]$ and $\sqrt{n}\,[T_{n,\hat{\epsilon}_n} - \alpha_1]$ based on the 200 samples were formed into a series of histograms. They are compared with the densities of the theoretical limit distributions. Theory shows that

(i) for $\alpha_1 = 1$,

$$\sqrt{n}\,[T_{n,\epsilon_n} - 1] \xrightarrow{\mathscr{D}} N(0, 4)$$

and $$\sqrt{n}\,[T_{n,\hat{\epsilon}_n} - 1] \xrightarrow{\mathscr{D}} V + \frac{(U-V)^3}{(U-V)^2+4},$$

https://doi.org/10.1017/S0305004100072121 Published online by Cambridge University Press

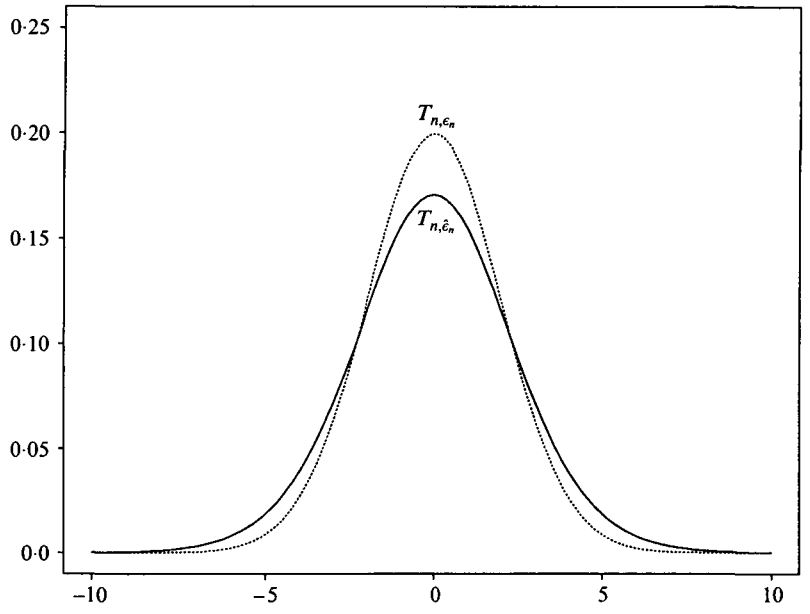


Fig. 3. Example 1. – Comparison of asymptotic distributions of $\sqrt{n}[T_{n,\hat{\epsilon}_n}-1]$ and $\sqrt{n}[T_{n,\epsilon_n}-1]$ under $\Gamma(1,1)$.

where $$\begin{pmatrix} U \\ V \end{pmatrix} \sim N\left(\mathbf{0}, \begin{pmatrix} 8 & 4 \\ 4 & 4 \end{pmatrix}\right).$$

Note that $$\mathrm{Var}\left[V+\frac{(U-V)^3}{(U-V)^2+4}\right] \approx 5{\cdot}868,$$

which is bigger than the asymptotic variance 4 of the normalized optimal T_{n,ϵ_n}.

Figure 3 compares the densities of the above two limit distributions.

(ii) for $\alpha_1 \neq 1$,

$$\left.\begin{matrix} \sqrt{n}[T_{n,\epsilon_n}-\alpha_1] \\ \sqrt{n}[T_{n,\hat{\epsilon}_n}-\alpha_1] \end{matrix}\right\} \xrightarrow{\mathscr{D}} N(0, 2\alpha_1(\alpha_1+3)).$$

Frequency histograms of $\sqrt{n}[T_{n,\hat{\epsilon}_n}-\alpha_1]$ corresponding to $n = 10^4$ are shown in Figure 4, together with their respective asymptotic density curves. Results for $\sqrt{n}[T_{n,\epsilon_n}-\alpha_1]$ are fairly similar and so they are not presented here. Observe that the empirical histograms shown in Figure 4 follow closely their limiting normal or almost normal density curves. Table 1 lists the sample variances of $\sqrt{n}[T_{n,\epsilon_n}-\alpha_1]$ and $\sqrt{n}[T_{n,\hat{\epsilon}_n}-\alpha_1]$ together with their asymptotic variances. Note that the figures for the sample variances listed in Table 1 were obtained by averaging over two hundred random samples drawn from $\Gamma(\alpha_1, 1)$. It is found that the experimental results agree remarkably with the theoretical asymptotic results.

In order to verify our theoretical results further, frequency distributions of $\hat{\epsilon}_n$ for $\alpha_1 = 1$ and $n(1-\hat{\epsilon}_n)$ for $\alpha_1 = 4$ were also constructed. Figure 5 shows the situations for the largest attempted sample size $n = 10^4$. Note the close resemblance of the case $\alpha_1 = 1$ and the density of $Z/(Z+1)$ with $Z \sim \chi^2_1$ as shown in Figure 1. Also, in the case where $\alpha_1 = 4$ we find that $n(1-\hat{\epsilon}_n)$ seems to be bounded in a very small interval. These observations are all consistent with our asymptotic theory.

https://doi.org/10.1017/S0305004100072121 Published online by Cambridge University Press

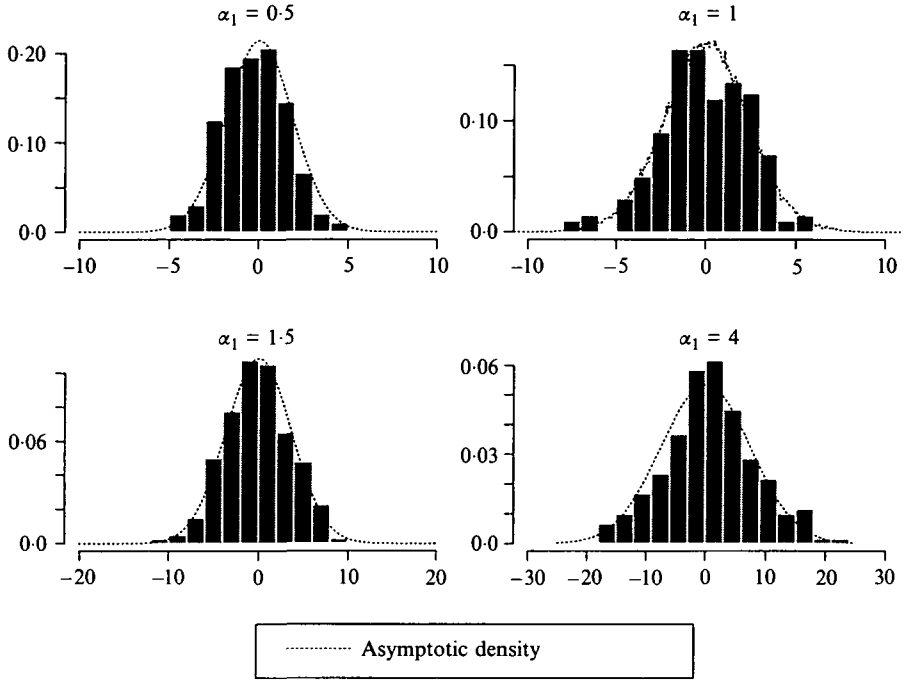


Fig. 4. Example 1. – Frequency histograms of $\sqrt{n}[T_{n,\hat{\epsilon}_n} - \alpha_1]$ under $\Gamma(\alpha_1, 1)$ for $n = 10000$ and $\alpha_1 = 0{\cdot}5, 1, 1{\cdot}5, 4$. The theoretical asymptotic densities are also plotted for comparison. In the case of $\alpha_1 = 1$, the asymptotic density is obtained by simulation of 10^5 data points.

Table 1. *Sampling results from Example* 1

α_1	n	Sample of variance $\sqrt{n}[T_{n,\hat{\epsilon}_n} - \alpha(F)]$	Sample variance of $\sqrt{n}[T_{n,\epsilon_n} - \alpha(F)]$
1	10	6·244	5·227
	100	5·554	3·657
	1000	5·076	3·823
	10000	6·054	4·392
Asymptotic variance:		**5·868**	**4·000**
0·5	10	1·931	0·804
	100	2·410	2·329
	1000	3·114	3·156
	10000	3·172	3·181
Asymptotic variance:		**3·500**	**3·500**
1·5	10	12·79	16·61
	100	13·66	13·57
	1000	14·69	13·61
	10000	13·11	13·04
Asymptotic variance:		**13·50**	**13·50**
4	10	37·63	35·80
	100	60·19	59·16
	1000	58·01	57·86
	10000	56·45	56·43
Asymptotic variance:		**56·00**	**56·00**

https://doi.org/10.1017/S0305004100072121 Published online by Cambridge University Press

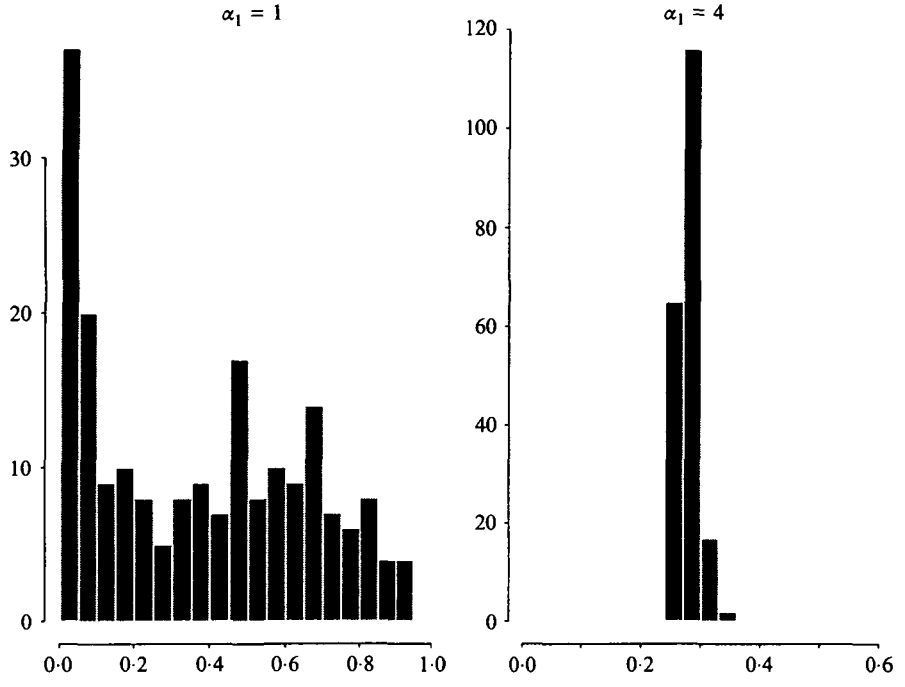


Fig. 5. Example 2. – Frequency histograms of $\hat{\epsilon}_n$ under $\Gamma(1, 1)$ and $n(1-\hat{\epsilon}_n)$ under $\Gamma(4, 1)$ for $n = 10000$.

Lastly, the performances of T_{n,ϵ_n} and $T_{n,\hat{\epsilon}_n}$ are compared with the pure PB and NB by means of their MSEs, for different sample sizes n over a range of α_1. The MSEs of T_{n,ϵ_n} and $T_{n,\hat{\epsilon}_n}$ were approximated by averaging over two hundred random samples, whereas the MSEs of PB and NB were computed exactly from their closed form expressions. Figure 6 demonstrates two typical cases where $n = 10$ and $n = 200$. It is found that the theoretically optimal T_{n,ϵ_n} performs pretty well uniformly over $\alpha_1 \in [0, 2]$. The empirical $T_{n,\hat{\epsilon}_n}$ is excellent for values of α_1 where NB is favoured. It is a bit less than optimal when PB is favoured, where its asymptotic variance is slightly bigger than the optimal one. Generally speaking, $T_{n,\hat{\epsilon}_n}$ performs reasonably well over a wide class of underlying distributions uniformly.

3·2. *Example* 2: *True distribution is Normal.* This example illustrates a possible case where $F \neq F_{\theta_0}$ and yet $\alpha(F) = \alpha(F_{\theta_0})$.

Suppose the true underlying distribution F is normal with mean μ and variance μ^2. The contemplated parametric model is still negative exponential indexed by the scale parameter θ. Our purpose is to estimate $\alpha(F) = \mathrm{Var}_F[X] = \mu^2$. So we have the same PB and NB as in Example 1.

3·2·1. *Theoretical asymptotic results.* As before, we perform the steps necessary for deriving the asymptotic properties for this example. The influence function $\psi_G(\cdot)$ corresponding to $\alpha(\cdot)$ is the same as in the previous example. Here the nearest θ describing F is easily found to be $\theta_0 = \mu^{-1}$, giving $\alpha(F_{\theta_0}) = \alpha(F) = \mu^2$. A little algebra shows that

$$\Sigma_F = \begin{pmatrix} 2\mu^4 & 0 \\ 0 & 4\mu^4 \end{pmatrix}.$$

https://doi.org/10.1017/S0305004100072121 Published online by Cambridge University Press

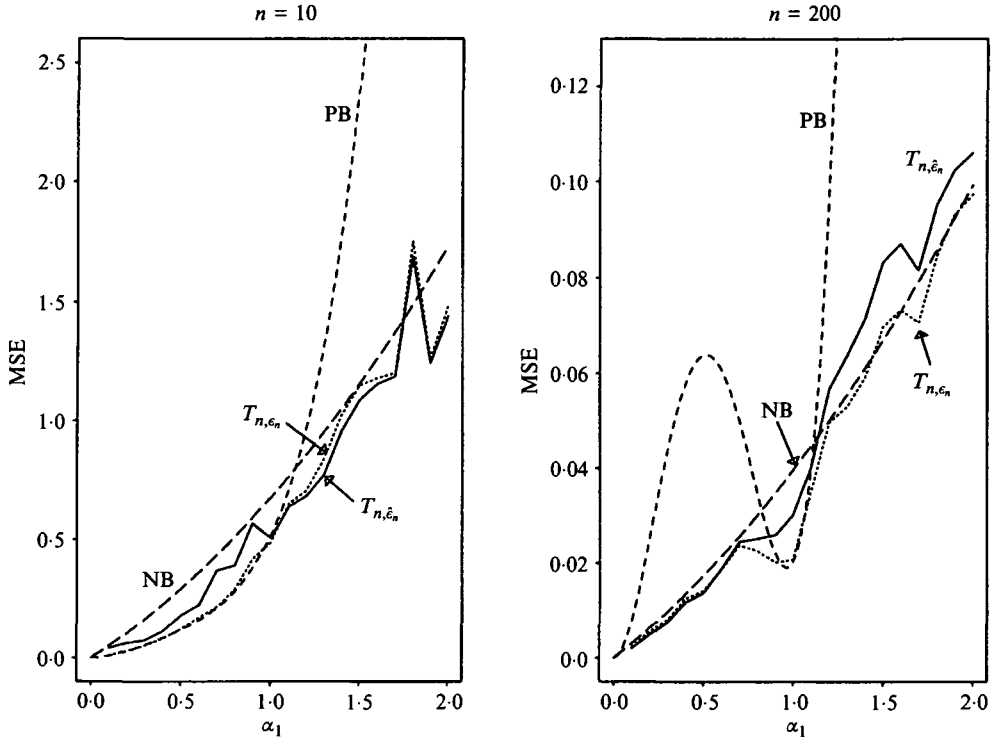


Fig. 6. Example 1. – Comparison of MSEs of $T_{n,\hat{\epsilon}_n}$ and T_{n,ϵ_n} with those of PB and NB under $\Gamma(\alpha_1, 1)$ for $n = 10$ and 200.

Suppose now $$\begin{pmatrix} U \\ V \end{pmatrix} \sim N(\mathbf{0}, \Sigma_F).$$

Then we have $$\epsilon_n \to \zeta\left(\frac{\mathbf{E}[V(V-U)]}{\mathbf{E}[(V-U)^2]}\right) = \frac{2}{3},$$

and $$\sqrt{n}\,[T_{n,\epsilon_n} - \mu^2] \xrightarrow{\mathscr{D}} \tfrac{1}{3}V + \tfrac{2}{3}U \sim N\left(0, \frac{4\mu^4}{3}\right).$$

Thus we have a non-trivial asymptotically optimal combination of PB and NB. Note also that

$$n\,\mathrm{MSE}_F(\mathrm{NB}) \to 2\mu^4 \quad \text{and} \quad n\,\mathrm{MSE}_F(\mathrm{PB}) \to 4\mu^4,$$

whereas $$n\,\mathrm{MSE}_F(T_{n,\epsilon_n}) \to \frac{4\mu^4}{3}.$$

This shows that the optimal hybrid estimator T_{n,ϵ_n} actually outperforms both PB and NB asymptotically in terms of MSEs. As for the asymptotics of our empirical estimate $T_{n,\hat{\epsilon}_n}$, we have

$$\hat{\epsilon}_n \xrightarrow{\mathscr{D}} \frac{Z + 2/3}{Z + 1}$$

with $Z \sim \chi^2_1$, and

$$\sqrt{n}\,[T_{n,\hat{\epsilon}_n} - \mu^2] \xrightarrow{\mathscr{D}} V + (U-V)\frac{(U-V)^2 + 4\mu^4}{(U-V)^2 + 6\mu^4}.$$

https://doi.org/10.1017/S0305004100072121 Published online by Cambridge University Press

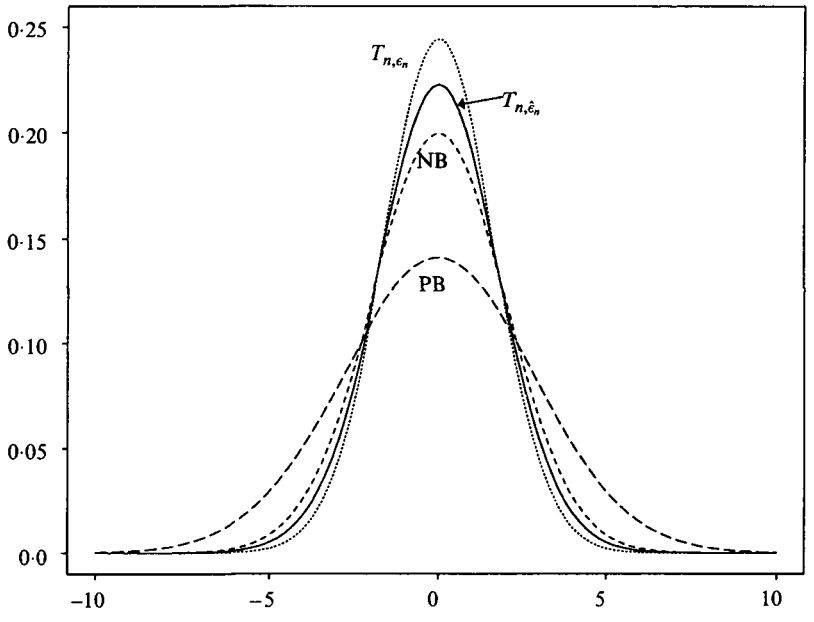


Fig. 7. Example 2. – Comparison of asymptotic distributions of $\sqrt{n}[T_{n,\hat{\epsilon}_n}-\sqrt{2}]$, $\sqrt{n}[T_{n,\epsilon_n}-\sqrt{2}]$, $\sqrt{n}[\mathrm{NB}-\sqrt{2}]$ and $\sqrt{n}[\mathrm{PB}-\sqrt{2}]$ under $N(2^{\frac{1}{4}}, 2^{\frac{1}{2}})$.

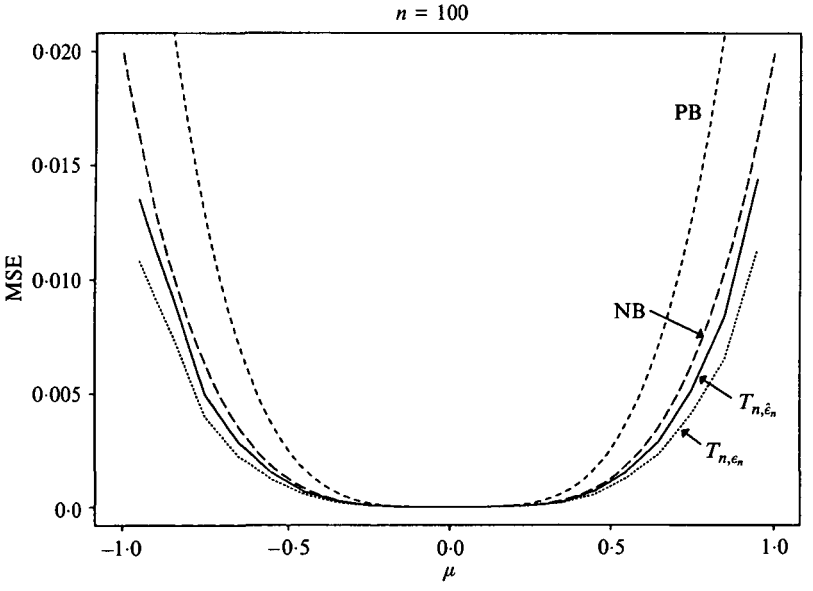


Fig. 8. Example 2. – Comparison of MSEs of $T_{n,\hat{\epsilon}_n}$ and T_{n,ϵ_n} with those of PB and NB under $N(\mu, \mu^2)$ for $n = 100$.

The variance of the above limit distribution is difficult to compute explicitly. However, we should expect it to lie strictly between $4\mu^4/3$ and $2\mu^4 = \min(2\mu^4, 4\mu^4)$. Figure 7 compares the densities of the asymptotic distributions of $T_{n,\hat{\epsilon}_n}$, T_{n,ϵ_n}, PB and NB, for the case where $\mu^4 = 2$. Clearly, both optimal ϵ-hybrid estimators are better than either PB or NB. This example convinces us that using a convex combination of PB and NB can be strictly beneficial in some circumstances.

https://doi.org/10.1017/S0305004100072121 Published online by Cambridge University Press

3·2·2. *Simulation results.* Five hundred samples of size n were drawn from $N(\mu, \mu^2)$, for different values of n and μ. The MSEs of T_{n,ϵ_n} and $T_{n,\hat{\epsilon}_n}$ were compared with those of NB and PB. Here MSE(PB) and MSE(NB) were exact, whereas MSE(T_{n,ϵ_n}) and MSE($T_{n,\hat{\epsilon}_n}$) were approximated by averaging over the 500 samples. The case for $n = 100$ is shown in Figure 8. We see the MSE of the hybrid estimator is strictly less than that of either PB or NB throughout the entire range of μ except at the point $\mu = 0$. Of course the theoretically optimal T_{n,ϵ_n} performs even better than the empirical $T_{n,\hat{\epsilon}_n}$, as can be seen from Figure 8.

4. *Further discussion and comments*

A few questions concerning our ϵ-hybrid approach will be raised and discussed briefly in this section.

4·1. *Restrictions on tuning parameter.* The pure and ϵ-hybrid bootstrap estimators discussed so far can be expressed in the form $\hat{\epsilon}_n \alpha_n(\hat{F}_n) + (1-\hat{\epsilon}_n)\alpha_n(F_{\hat{\theta}_n})$, where $\hat{\epsilon}_n$ is some random quantity bounded within the interval $[0, 1]$ based on the given sample data. In fact, the pure bootstrap estimates correspond to non-random values of $\hat{\epsilon}_n$. There is no particular reason why we must restrict our choice of $\hat{\epsilon}_n$ to the interval $[0, 1]$. If ϵ is allowed to assume values in the whole real line, the corresponding asymptotic results become as follows:

(i) If $\alpha(F) = \alpha(F_{\theta_0})$, then

$$\epsilon_n \to \Upsilon_F \quad \text{and} \quad \hat{\epsilon}_n \xrightarrow{\mathscr{D}} \frac{Z+\Upsilon_F}{Z+1}$$

with $Z \sim \chi^2_1$.

(ii) If $\alpha(F) \neq \alpha(F_{\theta_0})$, then

$$\sqrt{n}\,(\epsilon_n - 1) \longrightarrow 0 \quad \text{and} \quad \sqrt{n}\,(\hat{\epsilon}_n - 1) \xrightarrow{\mathrm{p}} 0.$$

The asymptotic distributions of T_{n,ϵ_n} and $T_{n,\hat{\epsilon}_n}$ follow in an obvious way. The only possible trouble occurs when deducing the large sample behaviour of their MSEs where the property of uniform integrability is made use of. However we note that both ϵ_n and $\hat{\epsilon}_n$ are still bounded asymptotically with probability one. Only when $\Upsilon_F \notin [0, 1]$ and $\alpha(F) = \alpha(F_{\theta_0})$ should ϵ_n and $\hat{\epsilon}_n$ converge to some limit beyond the range $[0, 1]$. The danger of an unbounded optimal ϵ is thus minimal.

4·2. *Dependence of functionals on sample size.* The asymptotic theory developed in Section 2·4 is valid under the assumption that $\alpha(F)$ is independent of n. What happens if the parameter of interest is of the form $\alpha_n(F)$? This is in fact a rather common situation. Take an example of the standard deviation of the variance-stabilized sample correlation coefficient of a bivariate sample, which is defined as

$$\left\{ \mathrm{Var}_F \left[\tanh^{-1} \left(\frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\left(\sum_{i=1}^{n} (X_i - \bar{X})^2\right)} \sqrt{\left(\sum_{i=1}^{n} (Y_i - \bar{Y})^2\right)}} \right) \right] \right\}^{\frac{1}{2}}, \tag{12}$$

where $\{(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)\}$ is a random sample drawn from a bivariate

https://doi.org/10.1017/S0305004100072121 Published online by Cambridge University Press

distribution F and $(\bar{X}, \bar{Y})$ is the usual sample mean. This quantity depends on the sample size n in a complicated and implicit manner. Whether the asymptotic theory still applies here remains an open question.

However we can often rewrite $\alpha_n(F)$ as an asymptotic expansion such as

$$\alpha_n(F) = \sum_{i=1}^{k} b_i(n)\,\alpha^{(i)}(F) + o(b_k(n)) \quad \text{as } n \to \infty,$$

where $\{\alpha^{(1)}(F), \alpha^{(2)}(F), \ldots\}$ is a sequence of sample-size free functionals of F and $b_i(n)$ is some distribution-free function of n such that

$$\frac{b_{i+1}(n)}{b_i(n)} = o(1) \quad \text{as } n \to \infty.$$

It is therefore possible to derive the asymptotic properties corresponding to the separate $\alpha^{(i)}$s. The overall asymptotic behaviour of any estimate of $\alpha_n(F)$ may be obtained by combining these individual results. In fact, sensible conclusions relevant to estimation of $\alpha_n(F)$ may often be expected to be obtained by study of estimation of the first order term $\alpha^{(1)}(F)$ alone.

Silverman and Young [**9**] show that it is indeed the case in the example where $\alpha_n(F)$ is given by (12). They simplify their study of estimation of (12) via an approximate formula given by [**6**, p. 251], namely

$$\alpha_n(F) = n^{-\frac{1}{2}}\alpha^{(1)}(F) + O(n^{-\frac{3}{2}}),$$

where the first order term $\alpha^{(1)}(F)$ admits an explicit expression in terms of moments of F.

4·3. *Other combinations of* NB *and* PB. Obviously $T_{n,\epsilon} = \epsilon\mathrm{NB} + (1-\epsilon)\,\mathrm{PB}$ is not the only way of mixing NB with PB. It has been chosen because of its relatively simple structure which allows for easy derivation of an optimal estimator as well as its asymptotic properties. Hjort [**5**] suggests a slightly different intermediate bootstrap estimate (IB), $\alpha(\epsilon\hat{F}_n + (1-\epsilon)F_{\hat{\theta}_n})$. Note that IB reduces to $T_{n,\epsilon}$ if α is a linear functional. How IB differs from $T_{n,\epsilon}$ theoretically remains to be studied. It seems inevitable that the optimal choice of ϵ in IB is much harder to determine because MSE_F (IB) is no longer a simple quadratic polynomial in ϵ, unless α is 'linear' enough. A direct approach to finding the empirical optimal ϵ for IB is to minimize the empirical version of MSE, given by

$$\mathrm{MSE}_{\hat{F}_n}(\alpha(\epsilon\hat{F}_n^* + (1-\epsilon)F_{\hat{\theta}_n^*})),$$

over ϵ in some predetermined interval. This process would typically call for a double bootstrap procedure, in which double bootstrap resamples are drawn from $\epsilon\hat{F}_n^* + (1-\epsilon)F_{\hat{\theta}_n^*}$ at some mesh of ϵ values and the empirical optimal ϵ chosen from this finite set of values, possibly by means of interpolation. The need for double bootstrap resampling from a variety of distributions of the form $\epsilon\hat{F}_n^* + (1-\epsilon)F_{\hat{\theta}_n^*}$ makes the estimate IB practically less attractive than $T_{n,\hat{\epsilon}_n}$ which requires double bootstrap resampling from $\hat{F}_n^*$ and $F_{\hat{\theta}_n^*}$ only, as has been pointed out in Section 2·5.

4·4. *Criterion of optimality.* Apart from the mean squared error, other criteria may sometimes be necessary, especially when the second moment of the bootstrap estimate fails to exist. Any measure of estimation error is suitable as a criterion

https://doi.org/10.1017/S0305004100072121 Published online by Cambridge University Press

provided it is well defined, non-parametric, easy to calculate and sensitive to the difference between NB and PB. It is hoped that the resulting hybrid estimate is rather stable despite the choice of criterion employed in its construction.

I should like to thank my research supervisor Dr Alastair Young for many helpful suggestions and discussions on the ideas and results presented in this paper, and Dr Pat Altham for her advice on many practical areas. This work was carried out whilst in receipt of a research scholarship from the Croucher Foundation.

REFERENCES

[1] P. J. Bickel and D. A. Freedman. Some asymptotic theory for the bootstrap. *Ann. Statist.* **9** (1981), 1196–1217.
[2] B. Efron. Bootstrap methods: another look at the jackknife. *Ann. Statist.* **7** (1979), 1–26.
[3] B. Efron. Jackknife-after-bootstrap standard errors and influence functions. (With discussion.) *J. Roy. Statist. Soc. Ser. B* **54** (1992), 83–127.
[4] L. T. Fernholz. *von Mises Calculus for Statistical Functionals* (Springer-Verlag, 1983).
[5] N. L. Hjort. Contribution to the discussion of David Hinkley's lectures on bootstrapping techniques. Written version presented at Nordic Conference in Mathematical Statistics. *Scand. J. Statist.* to appear.
[6] M. G. Kendall and A. Stuart. *The Advanced Theory of Statistics*, vol. 1, 4th edn. (Griffin, 1977).
[7] C. Léger and J. P. Romano. Bootstrap choice of tuning parameters. *Ann. Inst. Stat. Math.* **42** (1990), 709–735.
[8] R. J. Serfling. *Approximation Theorems of Mathematical Statistics* (Wiley, 1980).
[9] B. W. Silverman and G. A. Young. The bootstrap: to smooth or not to smooth? *Biometrika* **74** (1987), 469–479.

https://doi.org/10.1017/S0305004100072121 Published online by Cambridge University Press