Ann. Inst. Statist. Math.
Vol. 42, No. 4, 709-735 (1990)

BOOTSTRAP CHOICE OF TUNING PARAMETERS

CHRISTIAN LÉGER[1] AND JOSEPH P. ROMANO[2]

[1]*Département d'informatique et Recherche Opérationnelle, Université de Montréal, CP 6128, Succursale A, Montréal, Québec, Canada H3C 3J7*
[2]*Department of Statistics, Stanford University, Stanford, CA 94305, U.S.A.*

(Received June 5, 1989; revised December 11, 1989)

Abstract. Consider the problem of estimating $\theta = \theta(P)$ based on data x_n from an unknown distribution P. Given a family of estimators $T_{n,\beta}$ of $\theta(P)$, the goal is to choose β among $\beta \in I$ so that the resulting estimator is as good as possible. Typically, β can be regarded as a tuning or smoothing parameter, and proper choice of β is essential for good performance of $T_{n,\beta}$. In this paper, we discuss the theory of β being chosen by the bootstrap. Specifically, the bootstrap estimate of β, $\hat{\beta}_n$, is chosen to minimize an empirical bootstrap estimate of risk. A general theory is presented to establish the consistency and weak convergence properties of these estimators. Confidence intervals for $\theta(P)$ based on $T_{n,\hat{\beta}_n}$ are also asymptotically valid. Several applications of the theory are presented, including optimal choice of trimming proportion, bandwidth selection in density estimation and optimal combinations of estimates.

Key words and phrases: Bandwidth selection, bootstrap, confidence limits, density estimation, risk function.

1. Introduction

The bootstrap, first introduced by Efron (1979), is a general powerful technique predominantly used to estimate the sampling distribution of a statistic or an approximate pivot in order to construct confidence regions and hypothesis tests. Little theoretical attention has been given to other potential uses of the bootstrap, though Beran (1986) explored the bootstrap in the context of estimating the power of a test, and he mentions the possibility of choosing among different test statistics by taking the one with the largest estimated power. A relatively unexplored area of great potential use of the bootstrap is the estimation of risk functions of various estimators, with the goal of choosing the estimator with the best risk properties. Hall and Martin (1988) consider using the bootstrap to determine a shrinkage parameter in estimates of location based on an L^1 loss function. In this paper, we develop a general framework in which to establish

fundamental consistency and weak convergence results of estimators obtained by minimizing an empirical bootstrap estimate of risk over a given class of estimators.

To develop the problem, given data x_n from an unknown distribution P on a sample space S, the problem is to estimate and construct a confidence region for some unknown quantity $\theta(P)$. Although not necessary, assume now that $\theta(\cdot)$ is real-valued. The distribution P is unknown and is assumed to belong to a family $\boldsymbol{P}$ of distributions. Let $\{T_{n,\beta}: \beta \in I\}$ be a class of estimators based on x_n indexed by β in I. The fundamental problem addressed in this paper is how to choose β so that the resulting estimator is best among the given class. In general, β may be thought of as a tuning parameter, smoothing parameter, complexity parameter, etc. Given a loss function l, and some sequence of norming constants τ_n, the risk of using $T_{n,\beta}$ as an estimator of $\theta(P)$ is

$$R_n(\beta, P) = \boldsymbol{E}_P\{l[\tau_n | T_{n,\beta} - \theta(P)|]\}\,. \tag{1.1}$$

Let $\hat{Q}_n$ be an estimate of P. The bootstrap estimate of risk is then $R_n(\beta, \hat{Q}_n)$. Not worrying about problems of existence and uniqueness in the present section, define $\hat{\beta}_n$ to be the value of β minimizing $R_n(\beta, \hat{Q}_n)$. The resulting estimator is $T_{n,\hat{\beta}_n}$.

Note that the normalizing sequence τ_n in (1.1) is chosen so that $\tau_n[T_{n,\beta} - \theta(P)]$ has a nondegenerate limiting distribution. As expected, this is needed to obtain a useful (nondegenerate) asymptotic theory. However, in practice, one typically does not need to know the sequence τ_n. For example, if l is squared error, the value of β minimizing (1.1) is, in fact, independent of the choice of τ_n because τ_n is just a scaling factor. In smooth problems, $\tau_n = n^{1/2}$ works for any loss function.

The bootstrap method described here may be viewed as a competitor to cross-validation. Both are computer-intensive and applicable to complex problems. An introductory account of cross validation is given in Stone (1974). Some special examples that we develop are now described.

Example 1. *Mean versus median.* Suppose x_n consists of n independent identically distributed observations from a symmetric distribution P on the real line. A theoretically interesting question, though perhaps not practically important, is whether or not to use the sample mean or the sample median as an estimate of location. Thus, in this case, $T_{n,1}$ is the sample mean, $T_{n,2}$ is the sample median, and $I = \{1, 2\}$. As simple as this example may seem, we discuss it because it is a striking case where the method of cross validation fails miserably. Indeed, as argued in Stone (1977), if the actual population is normally distributed, the method of cross validation chooses the median over the mean with probability approximately 0.5008, for large n. The criterion is based on a squared error loss

function. If the criterion is changed to absolute error loss function, this large sample probability is about 0.5673. Moreover, a procedure that randomly selects (independent of x_n) the median with probability 0.5673 and the mean with probability 0.4327 has efficiency relative to the sample mean of 0.794, which is superior to the efficiency, 0.711, of the cross-validatory estimator. In contrast, regardless of the loss function, the bootstrap method with $\hat{Q}_n$ equal to the empirical distribution function of the data selects the sample median with probability approaching zero as n gets large. In fact, even for sample sizes as small as 10, the bootstrap method has an efficiency, with respect to squared error loss, of 0.95 (see Table 1).

A more interesting and difficult problem is the following.

Example 2. *Choosing a trimming proportion.* As in Example 1, suppose x_n consists of a sample of size n from a symmetric distribution P on the real line. Let $\theta(P)$ be the median of P. Let $T_{n,\beta}$ be a trimmed mean estimator of $\theta(P)$ with trimming parameter β. Specifically, if F is a distribution function on the line, and $F^{-1}(x) = \inf\{y: F(y) \geq x\}$, consider the functional

$$T_\beta(F) = (1 - 2\beta)^{-1} \int_\beta^{1-\beta} F^{-1}(t)\, dt .$$

When $\beta = 1/2$, $T_\beta(F)$ is defined to be $F^{-1}(1/2)$. Then, the β-trimmed mean is defined to be $T_{n,\beta} = T_\beta(\hat{F}_n)$, where $\hat{F}_n$ is the empirical distribution of the sample. This example is studied in Léger (1988), where the bootstrap approach is shown to be successful. Not surprisingly, the method of cross validation fails for this problem (see Pruitt (1988)). One might conjecture, however, that cross-validation might behave well if β is restricted to

Table 1. Combining the mean and median 10,000 simulations, 100 bootstrap replications.

Situation 1: Normal data

	$n=10$	$n=20$	$n=40$
Standardized MSE of $\bar{X}_n$	1.00	1.00	1.00
Standardized MSE of Med	1.384	1.479	1.517
Standardized MSE of $\hat{T}_{BOOT}$	1.048	1.079	1.088

Situation 2: Double exponential data

	$n=10$	$n=20$	$n=40$
Standardized MSE of $\bar{X}_n$	2.006	2.009	2.001
Standardized MSE of Med	1.466	1.335	1.242
Standardized MSE of $\hat{T}_{BOOT}$	1.745	1.567	1.432

$I = [\varepsilon, 1/2 - \varepsilon]$ for some $\varepsilon > 0$, since jackknife type estimates of variance of the sample median are known to be inconsistent. Even with such a restriction, Pruitt (1988) argues that cross-validation misbehaves.

In Example 2, we restrict attention to the class of trimmed mean estimators for reasons of simplicity, robustness and mathematical tractability. The cost is a possible loss of efficiency. However, it is known that a good choice of trimming proportion leads to an estimator which has good efficiency properties, even if it is not fully adaptive. For further discussion of this compromise approach, see Cox and Hinkley ((1974), Section 9.4).

Example 2 is more theoretically challenging than Example 1 because of the ambitious goal of choosing the best estimator among an infinite class of estimators. Another such example is choosing the smoothing or bandwidth parameter in density estimation. A distinct difference between density estimation problems and the two previous examples is that estimators are not smooth functionals of the distribution and do not converge at the usual $n^{-1/2}$ rate. Nevertheless, the methods described in Section 2 are applicable.

Example 3. *Bandwidth selection in density estimation.* Consider the problem of estimating an unknown density function f on the real line based on a sample $x_n = (X_1, \ldots, X_n)$ of size n from f. Specifically, consider estimating f at some fixed point t. A kernel density estimate of $f(t)$ with bandwidth parameter β is given by

$$\hat{g}_{n,\beta}(t) = \frac{1}{n\beta} \sum_{i=1}^{n} K\left(\frac{t - X_i}{\beta}\right), \tag{1.2}$$

for some (fixed) choice of kernel K. The question is how to choose β well to estimate $f(t)$. Alternatively, one may wish to estimate $f(t)$ for all t and construct a confidence band for $f(\cdot)$. In any case, care must be given so that $\hat{Q}_n$ is chosen properly. As will be seen, with an appropriately smooth choice of $\hat{Q}_n$, the bootstrap approach will yield an "optimal" choice of bandwidth parameter and also allow the construction of confidence regions.

In Section 2, a general methodology is described to analyze such problems. Several applications of the theory are discussed in Section 3. To summarize the main results of the paper, it is shown under conditions spelled out in Section 2 that the bootstrap choice of β, $\hat{\beta}_n$, converges to the optimal β in I. Typically, the optimal β, β_P, depends on the unknown probability P generating the data. Thus, the bootstrap approach "adapts" itself to the data to construct an (asymptotically) optimal estimator among

the given class. Moreover, the order of the difference between $T_{n,\hat{\beta}_n}$ and T_{n,β_P} is easily obtained. In addition, bootstrap confidence limits for $\theta(P)$ based on the bootstrap estimator $T_{n,\hat{\beta}_n}$ are asymptotically valid. The main results can be deduced from a single uniform weak convergence assumption. This assumption, with appropriate modifications when necessary, is verified in the examples. A modest simulation was done relating to Example 1. More extensive simulation results have already been reported (see Léger (1988) in the context of Example 2 and Faraway and Jhun (1988) in the context of Example 3). The numerical results are all extremely encouraging. The goal here, however, is not to establish the direct applicability of the bootstrap method in a specific problem; rather, it is to provide a general framework and theory for a wide class of problems.

Note that the main conclusions in this paper are developed for general loss functions, though the abstract formulation in Section 2 applies to general bounded, continuous loss functions. See Technical Remark 8 in Section 2 to see what additional assumptions are needed to include unbounded loss functions; typically, this just involves extra moment conditions. In any case, the theory and proofs do not rely on the particular form of the loss function in order to establish the claimed asymptotic properties. This stands in contrast to cross validation, where squared error loss is predominantly used to ease technical manipulations. More important, consistency and optimality properties for cross validation actually depend on the loss function (see Bowman *et al.* (1984)). Finally, Pruitt (1988) suggest that the success of cross validation hinges on the problem being hard enough so that "best" estimates converge at rates slower than $n^{-1/2}$, such as density estimation. In contrast, the bootstrap approach appears to be a powerful, successful approach in both regular and "hard" problems.

2. General formulation and analysis

Let $x_n = (X_1,\ldots, X_n)$ be a sample of size n from an unknown distribution P on some arbitrary sample space S. The distribution P is assumed to belong to a family $\boldsymbol{P}$. The model $\boldsymbol{P}$ may be "parametric" or "nonparametric". The framework presented in this section applies outside the i.i.d. case, but for simplicity we focus on this case for now. The problem is to estimate and construct a confidence region for some parameter $\theta(P)$. The range $\{\theta(P)\colon P \in \boldsymbol{P}\}$ will be denoted Θ. Usually, $\theta(\cdot)$ is real-valued, but a more abstract treatment is possible and necessary for some applications. For now, however, assume $\theta(\cdot)$ is real-valued.

Attention is focused on some class of estimators for $\theta(P)$, denoted by $T_{n,\beta} \equiv T_{n,\beta}(x_n)$, where β ranges over some index set I. The first question is to choose β, say $\hat{\beta}$, so that the resulting estimator $T_{n,\hat{\beta}}$ is best in some sense from the class of estimators $\{T_{n,\beta}\colon \beta \in I\}$. Typically, the law of $\tau_n[T_{n,\beta} - \theta(P)]$ converges weakly under P to a normal distribution with mean 0 and

variance $\sigma^2(\beta, P)$, for some choice of scale constants τ_n. In smooth problems, $\tau_n = n^{1/2}$. Since P is unknown, $\sigma^2(\beta, P)$ is usually unknown. Usually, $\sigma^2(\beta, P)$ possesses a unique minimum in β, so that the question is (asymptotically) meaningful.

Let $\hat{Q}_n$ be an estimate of P based on x_n. A bootstrap estimate of $\sigma^2(\beta, P)$ is then $\sigma^2(\beta, \hat{Q}_n)$. One possible approach to choosing β (and hence $T_{n,\beta}$) is to choose a value of β minimizing $\sigma^2(\beta, \hat{Q}_n)$. This is the approach taken by Jaeckel (1971) in the context of choosing a trimming proportion. Three main drawbacks are apparent with this approach if one desires a more general abstract theory. First, the analytical form of $\sigma^2(\beta, P)$ is often unknown. Second, since the choice of β is based on an asymptotic expression, $\sigma^2(\beta, P)$, for the "finite sample efficiency" of $T_{n,\beta}$, presumably a better (more ambitious) approach would be to estimate some finite sample characteristic of the distribution of $T_{n,\beta}$. Third, even for fixed β, it need not be the case that $\sigma^2(\beta, \hat{Q}_n) \to \sigma^2(\beta, P)$ in probability. Usually, fairly strong (or at least additional) assumptions that we will make are needed to obtain convergence of second moments. The approach taken below bypasses these difficulties. Later, we will explain the additional assumptions needed in Jaeckel's approach.

We now begin the general formulation. Slight variations are sometimes needed in application to examples. The goal of this section is not to provide a theorem which easily covers all applications. Rather, it is to present a fairly abstract formulation showing the structure and common features present in many typical examples.

The index set I is usually a subset of Euclidean space. In general, assume I is a metric space (or possibly a semi-metric space) with metric d_I. Introduce a loss function l, which is just an increasing map from the nonnegative real numbers to the nonnegative real numbers. Unless explicitly stated otherwise, we will assume l is bounded and uniformly continuous. The risk of using $T_{n,\beta}$ as an estimate of $\theta(P)$ is then given by (1.1). More generally, one need not assume the loss of estimating $\theta(P)$ by $T_{n,\beta}$ is a function of $|T_{n,\beta} - \theta(P)|$, though we restrict attention to this case here. In fact, the theory developed in this section may easily be generalized to that l can depend on (β, θ), as long as it is assumed the family of functions $\{l_{\beta,\theta}\}$ is uniformly bounded and equicontinuous (see Pollard (1984), p. 74).

Let $\hat{Q}_n$ be an estimate of P. A bootstrap estimate of $R_n(\beta, P)$ is then $R_n(\beta, \hat{Q}_n)$. Of course, the bootstrap estimate depends on $\hat{Q}_n$, and may be good or bad depending on such a choice. A bootstrap choice of β, say $\hat{\beta}_n$, is the value of β minimizing $R_n(\beta, \hat{Q}_n)$. In cases of nonuniqueness or nonexistence of a minimizing β, let $\hat{\beta}_n$ be any random variable that satisfies

$$R_n(\hat{\beta}_n, \hat{Q}_n) < \inf_{\beta \in I} R_n(\beta, \hat{Q}_n) + \varepsilon_n ,$$

where ε_n is any sequence of positive numbers tending to zero.

In general, the calculation of $\hat{\beta}_n$ involves resampling or simulation. Specifically, for $i = 1,\ldots, B$, let $y_{n,i}$ be a sample of size n from $\hat{Q}_n$. Then, a stochastic approximation to $R_n(\beta, \hat{Q}_n)$ based on B replicated data sets from $\hat{Q}_n$ is given by

$$\hat{R}_{n,B}(\beta, \hat{Q}_n) = \sum_{i=1}^{B} l[\tau_n | T_{n,\beta}(y_{n,i}) - \theta(\hat{Q}_n)|]/B .$$

A bootstrap choice of β would then involve minimizing this approximation over β. In some applications, the set I of possible values of β may be quite large so that this approach may not be computationally feasible. Instead, one may choose β values over some subset I_n of I, where I_n becomes "dense" in I. This subset may be determined by approximating I by some fixed set of a finite number of β values, so that these points form a "grid" or some ε net. Alternatively, the β values may be chosen at random by some probability measure on I. This may be viewed as a stochastic search procedure, as discussed by Beran and Millar (1987). In general, the theory presented in this paper applies even when such approximations are invoked.

We now begin the theoretical development. For fixed β, let $J_n(\beta, P)$ be the law of $\tau_n[T_{n,\beta} - \theta(P)]$ under P. The following assumption is weak, and is easy to check in applications. Assume that $J_n(\beta, P)$ converges weakly to the law, $J(\beta, P)$, of a random variable $Z(\beta, P)$. It then follows that

$$R_n(\beta, P) \to E[Z(\beta, P)] \equiv R(\beta, P) .$$

To ensure that the problem of choosing β is (asymptotically) well-defined, assume there exists a unique β_P in I so that, for any $\delta > 0$,

$$\inf_{\{\beta: d_I(\beta, \beta_P) > \delta\}} R(\beta, P) > R(\beta_P, P) . \tag{2.1}$$

This may be weakened somewhat, but for ease of exposition, we will assume (2.1).

To study the uniform behavior of $R_n(\beta, P)$, and subsequent processes indexed by β, we introduce the following terminology. Let $L_\infty(I)$ be the metric space of real-valued bounded functions on I, equipped with the metric induced by the uniform norm $|\cdot|_I$ defined by: if z_1 and z_2 are elements in $L_\infty(I)$, then

$$|z_1 - z_2| = \sup_{\beta \in I} |z_1(\beta) - z_2(\beta)| .$$

Endow $L_\infty(I)$ with the σ-field generated by the open balls. Let

$$Z_n(\beta, x_n, P) = \tau_n[T_{n,\beta} - \theta(P)]\,. \tag{2.2}$$

Regard $Z_n(\cdot, x_n, P)$ as a random element of $L_\infty(I)$. Issues of measurability will essentially be ignored because, in all the examples considered, the processes will be universally separable (see Pollard (1984), p. 38). In fact, for statistical purposes, there seems no loss in assuming I is countable so that questions of measurability are easily handled.

Let $J_n(P)$ be the distribution, in $L_\infty(I)$, of $Z_n(\cdot, x_n, P)$ under P. We wish to study the asymptotic behavior of $J_n(P)$. In general, we say a sequence of probability measures ν_n on $L_\infty(I)$ converges weakly to a probability measure ν if all bounded, continuous, measurable real-valued functions f defined on $L_\infty(I)$ satisfy $\int f d\nu_n \to \int f d\nu$.

THEOREM 2.1. *Let C_P be a set of sequences $\{P_n; n \geq 1\}$ of probability measures containing the sequence $\{P, P, \ldots\}$. Suppose that, for every sequence $\{P_n\} \in C_P$, $J_n(P_n)$ converges weakly to a common limit law $J(P)$, where $J(P)$ is the law of a process on $L_\infty(I)$ whose paths are continuous and lie in a separable subset of $L_\infty(I)$. Also, assume* (2.1) *and that l is bounded and continuous.*

(i) *Then,*

$$\sup_{\beta \in I} |R_n(\beta, P_n) - R(\beta, P)| \to 0\,. \tag{2.3}$$

(ii) *If $x_n = (X_1, \ldots, X_n)$ is a sample from P, let $\hat{Q}_n$ be an estimate of P based on x_n. Assume that $\hat{Q}_n$ falls in C_P with probability one under P. Then,*

$$\sup_{\beta \in I} |R_n(\beta, \hat{Q}_n) - R(\beta, P)| \to 0 \tag{2.4}$$

almost surely. Define $\hat{\beta}_n$ to be any random variable satisfying

$$R_n(\hat{\beta}_n, \hat{Q}_n) < \inf_{\beta \in I} R_n(\beta, \hat{Q}_n) + \varepsilon_n\,, \tag{2.5}$$

where ε_n is any sequence of positive numbers tending to 0. *Then,*

$$d_I(\hat{\beta}_n, \beta_P) \to 0 \tag{2.6}$$

almost surely.

(iii) *If $\hat{\gamma}_n$ is any sequence satisfying $d_I(\hat{\gamma}_n, \beta_P) \to 0$ in probability under P, then the law of $\tau_n[T_{n,\hat{\gamma}_n} - \theta(P)]$ tends weakly to $J(\beta_P, P)$. In particular, set $\hat{T}_n \equiv T_{n,\hat{\beta}_n}$, and let $L_n(P)$ be the law of $\tau_n[\hat{T}_n - \theta(P)]$ under P. Then, $L_n(P)$ tends weakly to $J(\beta_P, P)$.*

(iv) *Finally*,

$$\tau_n[\hat{T}_n - T_{n,\beta_P}] \to 0$$

in probability.

PROOF. (i) Let $Z_n(\cdot)$ and $Z(\cdot)$ be processes on $L_\infty(I)$ with distributions $J_n(P_n)$ and $J(P)$. By Dudley's version of Skorohod's almost sure representation theorem (see Pollard (1984), Section IV.3), we might as well assume these processes are defined on a common probability space and

$$\sup_{\beta \in I} |Z_n(\beta) - Z(\beta)| \to 0$$

almost surely. Then,

$$\sup_{\beta \in I} |R_n(\beta, P_n) - R(\beta, P)| \le E\left[\sup_{\beta \in I} |l(|Z_n(\beta)|) - l(|Z(\beta)|)| \right].$$

By the assumptions on l, the expression inside the expectation tends to 0 a.s., so we can apply dominated convergence. If you are worried about the measurability of this expression, argue as on Pollard ((1984), p. 74).

(ii) By (2.3), (2.4) trivially follows. Now, (2.6) follows by (2.1) and (2.4).

(iii) Let $Z_n(\cdot)$ and $Z(\cdot)$ be processes with laws $J_n(P)$ and $J(P)$. The assumptions imply $(Z_n(\cdot), \hat{\gamma}_n)$ converges weakly under P, on the product space of $L_\infty(I) \times I$, to the law of $(Z(\cdot), \gamma_P)$. By Skorohod's almost sure representation theorem, there exists Z_n^*, γ_n^* and Z^*, all defined on a common probability space, such that (Z_n^*, γ_n^*) has the same distribution as $(Z_n, \hat{\gamma}_n)$, Z^* has the same distribution as Z,

$$\sup_{\beta \in I} |Z_n^*(\beta) - Z^*(\beta)| \to 0$$

almost surely and $\gamma_n^* \to \gamma_P$ almost surely. Since $Z^*(\cdot)$ has continuous paths, it follows that $Z_n^*(\gamma_n^*) \to Z^*(\gamma_P)$ almost surely. Hence, the law of $Z_n^*(\gamma_n^*)$, which is equal to the law of $\tau_n[T_{n,\hat{\gamma}_n} - \theta(P)]$, converges weakly to the law of $Z^*(\gamma_P)$, which is equal to $J(\beta_P, P)$.

(iv) The proof of (iii) shows $Z_n^*(\gamma_n^*) - Z_n^*(\gamma_P) \to 0$ almost surely. The result follows.

The above theorem is analogous to Theorem 1 of Beran (1984), where a general result on the asymptotic consistency of the bootstrap method is given in the context of estimating a sampling distribution. Beran introduces C_P as a device indicating the assumption of smoothness of $J_n(\beta, P)$ in P

(for fixed β), so that the weak convergence of $J_n(\beta, P)$ to $J(\beta, P)$ is (locally) uniform in P in some sense. Here, we have the added complexity that we are studying the behavior of an entire class of estimators, so that it becomes fruitful to study $J_n(P)$ as a distribution on $L_\infty(I)$.

To summarize Theorem 2.1, (2.3) and (2.4) combine to yield the result that the bootstrap estimate of risk, $R_n(\beta, \hat{Q}_n)$ is (asymptotically) uniformly close in β to the actual risk $R_n(\beta, P)$ of the estimator $T_{n,\beta}$. By (2.6), the resulting bootstrap choice $\hat{\beta}_n$ of β asymptotically tends to the "optimal" value β_P. Finally, the resulting estimator $T_{n,\hat{\beta}_n}$ is asymptotically equivalent to the "best" choice T_{n,β_P}, as they have the same asymptotic distributions. In fact, the difference $T_{n,\hat{\beta}_n} - T_{n,\beta_P}$ is $o(\tau_n^{-1})$ in probability. Note that the proof shows other possible equivalent statements of Theorem 2.1 are possible. For example, let $\beta_{P,n}$ be the value of β minimizing the finite sample risk function $R_n(\beta, P)$ (or at least minimizing it within ε_n analogous to (2.5)). Then, the estimator $T_{n,\hat{\beta}_n}$ is equal to $T_{n,\beta_{P,n}}$ to $o(\tau_n^{-1})$ in probability.

In nice "smooth" problems, the asymptotic distribution $J(\beta, P)$ of $n^{1/2}[T_{n,\beta} - \theta(P)]$ is normal with mean 0 and variance $\sigma^2(\beta, P)$. In this case, regardless of the actual loss function, β_P in (2.1) is simply the value of β minimizing $\sigma^2(\beta, P)$. Hence, for any choice of loss function l used in the construction of the bootstrap estimate $\hat{\beta}_n$, we have, under the conditions of Theorem 2.1, $\hat{\beta}_n \to \beta_P$ a.s. Thus, the loss function may be viewed merely as a means to a construction of a "good" estimate, and the actual choice may not be crucial.

Based on the data-based bootstrap estimator $T_{n,\hat{\beta}_n}$, it may be desired to form a confidence region for $\theta(P)$. This involves estimating the distribution, $L_n(P)$, of $\tau_n[T_{n,\hat{\beta}_n} - \theta(P)]$. The bootstrap solution is to estimate $L_n(P)$ by $L_n(\hat{Q}_n)$. In general, the calculation of $L_n(\hat{Q}_n)$ would involve a double bootstrap, because the calculation of the estimator $T_{n,\hat{\beta}_n}$ itself involves a bootstrap loop. To describe the consistency of this bootstrap approximation, we need some terminology. Let $L_n(x, P)$ and $J(x, \beta, P)$ be the cumulative distribution functions corresponding to the laws $L_n(P)$ and $J(\beta, P)$. Also, let

$$L_n^{-1}(\alpha, P) = \inf_x \{x: L_n(x, P) \geq \alpha\},$$

and similarly define $J^{-1}(\alpha, \beta, P)$. Theorem 2.2 below establishes the consistency of bootstrap confidence limits for $\theta(P)$ based on the data-based bootstrap choice of β.

THEOREM 2.2. *Assume the conditions of Theorem* 2.1. *Also, assume* $J(x, \beta_P, P)$ *is continuous and strictly increasing on its support as a function of* x.

(i) *Then,*

$$\sup_x |L_n(x, P) - L_n(x, \hat{Q}_n)| \to 0$$

almost surely.

(ii)

$$P\{\tau_n[\hat{T}_n - \theta(P)] \geq L_n^{-1}(\alpha, \hat{Q}_n)\} \to 1 - \alpha .$$

Hence, the nominal $1 - \alpha$ *one-sided confidence interval*

$$(-\infty, \hat{T}_n - \tau_n^{-1} L_n^{-1}(\alpha, \hat{Q}_n)]$$

has an asymptotic coverage probability equal to $1 - \alpha$.

PROOF. (i) Let $\{P_n\}$ be any sequence in C_P. By an analogous argument to Theorem 2.1(ii), $\hat{\gamma}_n$ tends in probability under P_n to γ_P. Now, argue as in the proof of Theorem 2.1(iii), letting $Z_n(\cdot)$ be a process with law $J_n(P_n)$ instead of $J_n(P)$. The same argument shows $L_n(P_n)$ tends weakly to $J(\beta_P, P)$. Since $\hat{Q}_n$ falls in C_P with probability one (under P) and the limit distribution $J(\cdot, \beta_P, P)$ is continuous, (i) follows.

(ii) By (i) and the assumptions on the limit law $J(\cdot, \beta_P, P)$, $L_n^{-1}(\alpha, \hat{Q}_n)$ tends to $J^{-1}(\alpha, \beta_P, P)$ in probability under P. Also, the law under P of $\tau_n[\hat{T}_n - \theta(P)]$ tends weakly to $J(\beta_P, P)$ by Theorem 2.1(iii). Combining these facts yields the result.

We conclude this section with some remarks on the above theorems. The reader may wish to skip to Section 3 at a first reading.

Technical Remarks.

1. In smooth problems, it is easy to see why the assumption in Theorem 2.1 of $J_n(P_n)$ converging weakly to $J(P)$ should be true. Indeed, the analysis of $J_n(P_n)$ can be deduced from smoothness of the estimators $T_{n,\beta}$. To see why, consider the case where $T_{n,\beta}$ is a functional $T_\beta(\cdot)$ on $\boldsymbol{P}$, and $T_{n,\beta}(x_n) = T_\beta(\hat{P}_n)$, where $\hat{P}_n$ is the empirical measure based on a sample of size n. Often, $T_\beta(\cdot)$ is differentiable in the sense that it satisfies an approximation like

$$T_\beta(P_n) = T_\beta(P) + \int f_{\beta,P} d(P_n - P) + R_{n,\beta}(P_n, P) ,$$

where the remainder term $R_{n,\beta}(P_n, P)$ tends to 0 as P_n tends to P in an appropriate sense. Suppose, further, that the remainder term $R_{n,\beta}$ is small, uniformly in β. Then, the analysis of $J_n(P_n)$ can be deduced from the term $\int f_{\beta,P_n} d(\hat{P}_n - P_n)$. From the linear structure, $J_n(P_n)$ should behave like the distribution of a mean 0 Gaussian process $Z_n(\cdot)$ on $L_\infty(I)$ with covariance

$$\text{Cov}\,[Z_n(\beta_1), Z_n(\beta_2)] = \int f_{\beta_1, P_n} f_{\beta_2, P_n} dP_n\,.$$

With entropy and conditions of smoothness in P on the $f_{\beta, P}$, this argument can be formalized by appealing to the many recent results on central limit theorems for empirical processes indexed by classes of functions (see, e.g. Pollard (1984) and Sheehy and Wellner (1988)).

2. Applying a central limit theorem as explained in the above remark typically allows one to conclude that the limiting distribution $J(P)$ is the law of a process on $L_\infty(I)$ whose paths are uniformly continuous with respect to some metric d_I. If the metric d_I also makes I totally bounded, then the paths of this limiting process lie in a separable subset of $L_\infty(I)$.

3. The assumption that $J(x, \beta_P, P)$ is continuous and strictly increasing on its support is usually easy to verify; direct considerations often yield that $J(x, \beta_P, P)$ is a Gaussian distribution function.

4. The weak convergence of $J_n(P_n)$ to $J(P)$ may be too strong. However, we can sometimes argue from direct considerations that the bootstrap choice, $\hat{\beta}_n$, of β lies in some subset I_0 (usually compact) of I with probability tending to one. Then, it may be possible to apply the above theorems with I replaced by I_0 to obtain the same conclusions.

5. The hypotheses and conclusions of Theorems 2.1 and 2.2 may vary to obtain almost sure convergence results or convergence in probability results. We do not dwell on this technical distinction and the possible variations of the main theme of fundamental convergence results. However, in some applications, we may be forced to settle for a result like (2.6) with almost sure convergence replaced by convergence in probability.

6. An alternative route to proving (iii) and (iv) of Theorem 2.1 is the following. By some method, argue that (2.6) holds. Then, show that the sequence of processes $Z_n(\beta, x_n, P_n)$ defined by (2.2), where x_n is a sample from P_n, is stochastically equicontinuous whenever $\{P_n\} \in C_P$. For the definition of stochastic equicontinuity, see Pollard ((1984), p. 139). This is slightly weaker than the hypothesis of $J_n(P_n)$ converging weakly to $J(P)$. However, the stronger hypothesis tidily yields the main results quite readily.

7. A more general result is possible by letting the estimate $\hat{Q}_n$ of P depend on β. That is, let $\hat{\beta}_n$ be the bootstrap choice of β obtained by minimizing $R_n(\beta, \hat{Q}_{n,\beta})$, where $\hat{Q}_{n,\beta}$ is an estimate of P which may depend on β. The reason one might wish to allow this possibility is that $R_n(\beta, \hat{Q}_n)$ may be a good estimate of the risk of $T_{n,\beta}$ for some β, but it may not be for all β. To illustrate this point, consider Example 1 introduced in Section 1, with l being squared error loss. If $\hat{Q}_n$ is the empirical distribution of the data, the bootstrap estimate of variance of the sample mean is just $\hat{\sigma}^2/n$, where $\hat{\sigma}^2$ is the usual bootstrap estimate of the population variance. This estimate is quite good; on the other hand, the bootstrap estimate of the

variance of the sample median need not be so good. In general, its convergence rate is much slower than if we replace $\hat{Q}_n$ by a smoother estimate of P (see Hall *et al.* (1989)).

8. It may be desirable and convenient to consider unbounded loss functions. However, additional assumptions are generally needed to establish the results given in Theorems 2.1 and 2.2. Recall Z_n defined in (2.2). One essentially needs to show that the collection of random variables $l|Z_n(\beta, x_n, P_n)|$ as β and n varies is uniformly integrable with $\{P_n\}$ in $\boldsymbol{C}_P$. Consider Example 1 introduced in Section 1 with squared error loss. The bootstrap estimate of risk for the sample median is not consistent without an additional moment assumption which is not needed to show weak convergence of the bootstrap sampling distribution of the sample median. See Ghosh *et al.* (1984). Unfortunately, verifying the uniform integrability assumption can be difficult. However, in some examples where estimators are sums of independent random variables, the special properties of mean squared error can be exploited for a direct calculation (see Example 3 in Section 3).

9. An alternative approach for a bootstrap choice of β which does not make explicit use of a loss function and does not need additional assumptions to the ones given in Theorems 2.1 and 2.2 is the following. For each fixed β, let $\hat{C}_{n,\beta}(1-2\alpha)$ be a nominal $1-2\alpha$ bootstrap confidence interval for $\theta(P)$ based on the estimator $T_{n,\beta}$, so that

$$\hat{C}_{n,\beta}(1-2\alpha) = [\hat{T}_{n,\beta} - \tau_n^{-1}J_n^{-1}(1-\alpha, \beta, \hat{Q}_n), \hat{T}_{n,\beta} - \tau_n^{-1}J_n^{-1}(\alpha, \beta, \hat{Q}_n)] .$$

Let $\hat{\beta}_n$ be the value of β such that the length of $\hat{C}_{n,\beta}$ is minimized. That is, $\hat{\beta}_n$ is the value of β such that the nominal $1-2\alpha$ confidence interval based on $T_{n,\beta}$ has the shortest length. Then, the conclusions of Theorems 2.1 and 2.2 hold for such a choice of $\hat{\beta}_n$. In particular, suppose $J(\beta, P)$ is asymptotically normal with mean 0 and variance $\sigma^2(\beta, P)$. Then, $\hat{\beta}_n$ converges a.s. to β_P minimizing $\sigma^2(\beta, P)$. This approach is used in Léger (1988).

3. Applications and examples

Example 2 (*continued*). Consider the situation of Example 2 introduced in Section 1. The goal is to choose a trimming proportion β among β in $I = [\varepsilon, 1/2 - \varepsilon]$, for some $\varepsilon > 0$. We now give a sketch of why the results of the theorems in Section 2 hold. An alternative direct approach to this example is well-studied in Léger (1988), and makes direct use of Technical Remarks 1 and 6 of Section 2.

Let $q_n(\cdot)$ be the quantile process defined by

$$q_n(t) = n^{1/2}[\hat{F}_n^{-1}(t) - F^{-1}(t)] .$$

To proceed in applying Theorems 2.1 and 2.2, we must study the process

$$Z_n(\beta) = n^{1/2}[T_\beta(\hat{F}_n) - T_\beta(F)] .$$

Then,

$$Z_n(\beta) = (1 - 2\beta)^{-1} \int_\beta^{1-\beta} q_n(t)\, dt .$$

Regard $Z_n(\cdot) = T(q_n(\cdot))$ as a map taking q_n in $L_\infty([\varepsilon, 1 - \varepsilon])$ to Z_n in $L_\infty(I)$. Since this map is linear and even uniformly continuous, the weak convergence properties of $Z_n(\cdot)$ can be derived from known properties of the quantile process. Moreover, bootstrap versions of these processes behave (asymptotically) as the original processes, by virtue of the validity of bootstrapping the quantile process (see Bickel and Freedman (1981), for example). In summary, the bootstrap allows one to choose an (asymptotically) optimal trimming proportion in I without assumptions on P, other than those used for establishing the bootstrap consistency of the quantile process. As in Section 2, we have assumed that l is bounded (see Léger (1988) for the case of squared error loss). Extensive simulation results in Léger (1988) are extremely encouraging for this example.

Example 4. *Linear combinations of estimates.* Given two estimators $U_n = U_n(x_n)$ and $V_n = V_n(x_n)$ of some parameter θ, the problem is to combine them to produce a more efficient estimator. Let

$$T_{n,\beta} = \beta U_n + (1 - \beta) V_n , \tag{3.1}$$

so that the problem is to choose β optimally among β in the real line, or possibly some restricted subset of the real line. Let $\hat{\beta}$ be the bootstrap estimate as defined by (2.5). Let $D_n(P^{(n)})$ be the distribution of $\tau_n[U_n - \theta, V_n - \theta]$ under $P^{(n)}$ for some sequence τ_n tending to infinity. Note that we are not necessarily assuming that the data x_n is made up of n independent and identically distributed components; however, when x_n is a sample of size n from a fixed distribution P, then $P^{(n)}$ is just the product law P^n. The parameter θ is still regarded as some functional of $P^{(n)}$. We will assume Assumption A, given below, to be verified in two subexamples. The assumption essentially amounts to being able to bootstrap the joint distribution of (U_n, V_n), and hence we can draw on well-known theory for its verification. The true data distribution is denoted $P_0^{(n)}$, and the true value of θ is $\theta_0 = \theta(P_0^{(n)})$.

ASSUMPTION A. Let $\mathcal{C}_0$ be a set of sequences of distributions $\{P^{(n)}\}$ of x_n containing the sequence $P_0^{(n)}$ and satisfying $D_n(P^{(n)})$ converges weakly

to a continuous limit distribution D_0, whenever $\{P^{(n)}\} \in \mathbf{C}_0$. If (U_0, V_0) has distribution D_0, then

$$E[l|\beta U_0 + (1-\beta)V_0|] \tag{3.2}$$

has a unique minimum, β_0, in β; that is, assume (2.1).

The first result below applies to a bootstrap choice of β in an index set I which is a bounded subset of the real line. The choice $I = [0, 1]$ corresponds to choosing some convex combination of U_n and V_n.

THEOREM 3.1. *(Bounded I) Assume Assumption* A *with the optimal β_0 belonging to a bounded set I. Let $\hat{Q}_n$ be an estimate of $P_0^{(n)}$ based on x_n, such that $\{\hat{Q}_n\}$ falls in $\mathbf{C}_0$ with probability one. Then, the assumptions of Theorem* 2.1 *and Theorem* 2.2 *are satisfied. Consequently, the bootstrap estimate $\hat{\beta}_n$ (in I) satisfies $\hat{\beta}_n \to \beta_0$ almost surely. Moreover, bootstrap confidence intervals for θ based on $T_{n,\hat{\beta}_n}$ are asymptotically valid in the sense of Theorem* 2.2.

Typically, the distribution D_0 of (U_0, V_0) is asymptotically bivariate Gaussian with mean 0 and covariance matrix $\Sigma = (\sigma_{i,j})$. In this case, regardless of the loss function (as long as it is assumed to be increasing), the value of β minimizing (3.2) is

$$\beta_0 = \frac{\sigma_{2,2} - \sigma_{1,2}}{\sigma_{1,1} + \sigma_{2,2} - 2\sigma_{1,2}} \tag{3.3}$$

and is well-defined so that (3.2) holds.

PROOF OF THEOREM 3.1. The proof of Theorem 3.1 is immediate from the fact that, whenever (X_n, Y_n) converges in distribution to (X, Y) in $\mathbf{R}^2$, then the distribution of $\beta X_n + (1-\beta)Y_n$, regarded as a random element of $L_\infty(I)$, converges weakly to the law of the process $\beta X + (1-\beta)Y$. Indeed, the mapping g from $\mathbf{R}^2$ to $L_\infty[0, 1]$ taking (x, y) into the function $g(x, y) = \beta x + (1-\beta)y$ is uniformly continuous:

$$\begin{aligned} \|g(x, y) - g(x_n, y_n)\| &= \sup_\beta |\beta(x_n - x) + (1-\beta)(y_n - y)| \\ &\le M \max [|x_n - x|, |y_n - y|] , \end{aligned}$$

where $M = \sup \{\beta: \beta : \in I\}$. Hence, the weak convergence of the processes under study follows from the continuous mapping theorem.

In general, the optimal β_0 need not be known to belong to some

bounded set I. For example, in the asymptotically bivariate Gaussian case, the solution β_0 given by (3.3) will not necessarily lie in a bounded set, and it may be desirable to choose a linear combination of U_n and V_n by allowing β to be any real number. The next result covers this case. Note, however, that the main weak convergence assumption of Theorem 2.1 and Theorem 2.2 is not implied by these assumptions as it does not hold. Nevertheless, the main conclusions of the theorems all follow. See Technical Remarks 4 and 6 of the previous section to appreciate why.

THEOREM 3.2. *(Unbounded I) Let $\mathbf{C}_0$ be a set of sequences of distributions $\{P^{(n)}\}$ of x_n satisfying Assumption* A. *Also, assume* (3.2). *Let $\hat{Q}_n$ be an estimate of $P_0^{(n)}$ based on x_n, such that $\{\hat{Q}_n\}$ falls in $\mathbf{C}_0$ with probability one. Then, the conclusions of Theorem* 2.1 *and Theorem* 2.2 *all hold. In particular, the bootstrap estimate $\hat{\beta}_n$ (in I) satisfies $\hat{\beta}_n \to \beta_0$ almost surely. Moreover, bootstrap confidence intervals for θ based on $T_{n,\hat{\beta}_n}$ are asymptotically valid in the sense of Theorem* 2.2.

PROOF OF THEOREM 3.2. To prove the theorem, the following small result is needed. The family of functions indexed by β in $\mathbf{R}$, mapping (u, v) in $\mathbf{R}^2$ to $l\{\beta u + (1-\beta)v\}$ in $\mathbf{R}$, is equicontinuous; that is, for all (u, v) and $\varepsilon > 0$, there exists a $\delta > 0$, depending possibly on (u, v) but not on β, such that

$$|l(u, v) - l(x, y)| < \varepsilon$$

whenever $\rho((u, v), (x, y)) < \delta$. Here, ρ is the usual Euclidean metric. This result is left to the reader to prove; note the importance of assuming l bounded, continuous and monotone on the positive half of the real line. The result would be false if $l(t) = |t|$, for example. Now, it follows from this result (see Pollard (1984), p. 74, equation 20) that, for any sequence $\{P^{(n)}\}$ in $\mathbf{C}_0$, the risk function

$$R_n(\beta, P^{(n)}) = E_{P^{(n)}}[l\{\tau_n|\beta U_n + (1-\beta)V_n - \theta(P^{(n)})|\}]$$

tends to (3.2) uniformly in β. Hence, the bootstrap estimate of risk, $R_n(\beta, \hat{Q}_n)$, tends to (3.2) uniformly in β with probability one. Thus, by the uniqueness assumption of β_0 minimizing (3.2), the bootstrap estimate $\hat{\beta}_n$ tends to the (asymptotically) optimal value β_0. Note that we are not claiming that the distribution of the process $\tau_n[U_n + (1-\beta)V_n - \theta(P^{(n)})]$, regarded as a random element of $L_\infty(\mathbf{R})$, converges to a weak limit. Nevertheless, (2.3) and (2.4) hold as well. To see why, given that $\hat{\beta}_n \to \beta_0$ almost surely, one can restrict attention to β values in some compact subset I_0 and apply Theorem 3.1 to obtain all of the conclusions of Theorems 2.1 and 2.2.

Example 4(1). *Combining the sample mean and sample median.* Let x_n be a sample of size n from a symmetric law on the line having a distribution function F. The problem is to estimate $\theta(F) = F^{-1}(1/2)$. Let $U_n(x_n)$ be the sample mean and let $V_n(x_n)$ be the sample median. The following proposition establishes the validity of the hypotheses of the two previous theorems.

PROPOSITION 3.1. *Fix F, a symmetric distribution on the line with unique median $\theta(F)$ and having a finite (nonzero)variance $\sigma^2(F)$. Let $\mathbf{C}_F$ be the set of sequences of distributions $\{F_n\}$ satisfying F_n converges weakly to F, $\sigma^2(F_n) \to \sigma^2(F)$ and*

$$\lim_{n\to\infty} n^{1/2}\left[F_n(\theta(F_n) + n^{-1/2}x) - \frac{1}{2}\right] = xf(\theta(F)) \tag{3.4}$$

for every real x. Let $\{F_n\}$ be any sequence in $\mathbf{C}_F$ and let x_n^* be a sample of size n from F_n. Then, the joint distribution of

$$n^{1/2}[U_n(x_n^*) - \theta(F_n), V_n(x_n^*) - \theta(F_n)]$$

converges weakly to a bivariate normal distribution with mean 0 and covariance matrix $\Sigma = (\sigma_{i,j})$, where $\sigma_{1,1} = \sigma^2(F)$, $\sigma_{2,2} = 1/[4f^2(\theta(F))]$, and $\sigma_{1,2}$ is the covariance between X and $-1\ (X \le \theta(F))/f(\theta(F))$ when X has distribution F.

The proof of Proposition 3.1 is similar to the proofs of convergence of the marginal distributions, for which the reader is referred to Bickel and Freedman (1981), Beran (1984) and Sheehy and Wellner (1988). The joint asymptotic normality is easily obtained, for example, by applying an appropriate linear representation of the sample median (see Serfling (1980), Theorem 2.5).

If $\hat{F}_n$ is the empirical distribution of the data, it remains to show that $\{\hat{F}_n\}$ falls in $\mathbf{C}_F$ with probability one. The only difficulty is showing that (3.4) holds almost surely when $F_n = \hat{F}_n$. Beran (1984) shows that this convergence holds in probability, but Sheehy and Wellner (1988) show this convergence to hold almost surely. Hence, the conclusions of Theorems 3.1 and 3.2 hold.

In this example, it may be desirable to symmetrize the empirical distribution and resample from a symmetric distribution. In addition, one might wish to smooth the empirical distribution. The same conclusions could easily be obtained in an analogous manner as long as the estimating sequence of distributions falls in $\mathbf{C}_F$ with probability one. For example, one might symmetrize the empirical distribution about the sample median.

Then, the same results hold under no additional assumptions.

Example 4(2). *Combining independent estimates.* Let y_n be a sample of size n from a distribution P_y and z_m be a sample of size $m = m(n)$ from a distribution P_z. Let $\theta = \theta(P)$ be the functional of interest. It is assumed $\theta(P_y) = \theta(P_z)$, though it need not be true that $P_y = P_z$. Let U_n be an estimate of θ based on y_n and let V_n be an estimate of θ based on $z_{m(n)}$. The problem is to combine the estimates by a suitable choice of β in (3.1). Under Assumption A, and the further assumption that the limiting distribution D_0 is bivariate Gaussian with mean 0, the resulting (asymptotically) best choice of β is, by (3.3), given by $\beta_0 = \sigma_{2,2}/(\sigma_{1,1} + \sigma_{2,2})$, where $\sigma_{2,2}$ is the asymptotic variance of V_n and $\sigma_{1,1}$ is the asymptotic variance of U_n. Thus, the bootstrap method of choosing β is equivalent, to first order, to choosing β by the usual weights determined by the inverse proportional to the variance and is dependent on the choice of the loss function l. For squared error loss function and unbiasedness of the estimators U_n and V_n, the bootstrap choice of β is exactly equal to $\hat{\beta} = \hat{\sigma}_{2,2}/(\hat{\sigma}_{1,1} + \hat{\sigma}_{2,2})$, where $\hat{\sigma}_{2,2}$ and $\hat{\sigma}_{1,1}$ are bootstrap estimates of the variances of V_n and U_n, respectively. In general, $T_{n,\hat{\beta}}$ is (asymptotically) as good as T_{n,β_0} and the conclusions of Theorems 3.1 and 3.2 hold.

Example 4(3). *Shrinkage estimators.* Let $U_n(x_n)$ be an estimate of $\theta(P)$. The goal is to choose β to minimize

$$E_P[l\{\tau_n|(1-\beta)U_n + \beta - \theta(P)|\}] .$$

The case $l(|t|) = |t|$ corresponds to L^1-shrinkage, as considered by Hall and Martin (1988). This is a special case in our context of combining estimates because we can take $V_n = 1$ (or $V_n = \tau_n^{-1}$, which is convenient for asymptotic purposes in verifying Assumption A). The above theorems apply immediately to this situation for bounded loss functions. Typically, $\tau_n = n^{1/2}$, as is the case under the assumptions of Hall and Martin, and the optimal value of β (as is its bootstrap estimate) is of order n^{-1} and so does not play a role in first order asymptotics.

We conclude this subsection by pointing out that, in general, one might wish to consider linear combinations of estimators of the form

$$T_{n,\beta} = \sum_{i=1}^{d} \beta_i U_{n,i}$$

by an appropriate choice of $\beta = (\beta_1,\ldots,\beta_d)$. Analogous theoretical results are obtained in the same manner. In particular, Assumption A is modified so that $D_n(P^{(n)})$ refers to the distribution of

$$\tau_n[U_{n,1} - \theta, \ldots, U_{n,d} - \theta]$$

under $P^{(n)}$. As an example, Cox and Hinkley ((1974), p. 347) consider estimating location by a linear combination of order statistics. The methods discussed here allow one to choose the appropriate combination in an optimal way.

Example 5. *Unbiased risk reduction.* Let x_n be a sample of size n from a distribution P and consider the estimation of some functional $\theta(P)$ by some estimator $U_n(x_n)$. Because the estimator may be biased, for example, one might wish to consider alternative estimators of the form

$$T_{n,\beta} = \beta_1 + \beta_2 U_n(x_n) ,$$

with the idea that some linear transformation of U_n is a better estimator than U_n. For example, suppose one always chooses $\beta_2 = 1$ but considers choices for β_1 other than zero. If the criterion is squared error loss and if

$$E[U_n(x_n)] = \theta(P) + b_n(P) ,$$

where $b_n(P)$ is the bias of $U_n(x_n)$ under P, then the "best" choice of β_1 is $-b_n(P)$. In general, the optimal choices of β_1 and β_2 will depend on both P and the loss function, and the bootstrap offers an approach for correcting for bias or, more generally, reducing risk. As before, the bootstrap choice of $\beta = (\beta_1, \beta_2)$ minimizes the empirical risk function. Thus, for squared error loss, the bootstrap choice of β_1 is $-b_n(\hat{Q}_n)$, where $\hat{Q}_n$ is some estimate of P.

The mathematical development of the bootstrap choice of β is, of course, similar to Example 4, as it is really a special case. However, due to the importance of bias reduction, we prefer to distinguish its special features.

Example 5(1). *Unbiased risk estimation.* Suppose that P has a density $f(x - \theta)$ for some location parameter θ and that f is known. Consider the estimation of θ based on a single observation X. Any location equivariant estimator of θ takes the form $X + \beta_1$. In this case, the optimal choice of β_1 does not depend on θ but rather on the choice of the loss function l. Hence, the parametric bootstrap approach also exactly yields the best choice of β_1, as long as one estimates P by some distribution in the parametric family. The resulting estimator is the minimum risk equivariant (see Lehmann (1983), Chapter 3). For squared-error loss, the estimator is unbiased for θ. In general, the estimator is risk-unbiased. The point of this example was to see the connection with risk-unbiasedness and to see how the choice of an estimator is influenced by the loss function. It is also

reassuring to know the bootstrap approach reproduces the exact solution when it exists.

Example 5(2). *Estimation of variance in a linear model.* Suppose x_n consists of n observations X_i, where $X_i = \mu + \varepsilon_i$ and the ε_i are independent and identically distributed with mean 0 and variance σ^2. The problem is to estimate $\theta = \sigma^2$. Let $U_n = \Sigma(X_i - \bar{X}_n)^2/(n-1)$, where $\bar{X}_n = \Sigma X_i/n$.

First, consider the case where the ε_i are assumed to be normally distributed. Also, take $\beta_1 = 0$, so that the problem is then reduced to the proper choice of β_2. The mean squared error of $\beta_2 U_n$ is easily seen to be

$$\frac{2\beta_2^2\sigma^4}{n-1} + (\beta_2 - 1)^2\sigma^4, \tag{3.5}$$

and so the optimal choice of β_2 for squared error loss is $(n-1)/(n+1)$. Again, the parametric bootstrap approach yields the same exact answer. Note that the new estimator $(n-1)U_n/(n+1)$ is now biased but has a smaller risk function than the unbiased estimator U_n.

Now, suppose the ε_i are only assumed to have mean 0 and variance σ^2. Then, the mean-squared error of $\beta_2 U_n$ is given by (3.5) plus the additional term $\beta_2^2\kappa_4/n$, where κ_4 is the fourth cumulant of ε_i. Alternatively, $\kappa_4 = \mu_4 - 3\mu_2^2$, where μ_j is the j-th moment of ε_i. Then, the optimal choice of β_2 becomes

$$\beta_{2,0} = \beta_{2,0}(n) = \frac{1}{1 + \dfrac{2}{n-1} + \dfrac{\kappa}{n}}, \tag{3.6}$$

where $\kappa = \kappa_4/\sigma^4$ is the kurtosis of the ε_i. Since κ is not assumed known, the bootstrap solution amounts to replacing κ by a sample estimate $\hat{\kappa}$ in (3.6). Thus, the bootstrap approach is not exact for finite samples in this case. Note that the optimal solution (3.6) depends on n and differs from 1 by order n^{-1}, regardless of σ^2 and κ. This is typical for bias reduction; that is, the removal of bias does not typically enter into first order asymptotic properties. This example can be easily generalized to the case where the mean of X_i is a linear function of some covariates.

An example where the removal of bias does enter into first order asymptotic properties is the following.

Example 5(3). *Uniform scale family.* Consider the estimation of θ based on a sample $x_n = (X_1, \ldots, X_n)$ of size n from a uniform distribution on $[0, \theta]$. Let $U_n(x_n) = \max(X_1, \ldots, X_n)$ be the maximum likelihood estimate of

θ. The mean squared error of $\beta_2 U_n$ is easily calculated to be

$$\frac{\beta_2^2 n\theta^2}{(n+2)(n+1)^2} + \theta^2\left(\frac{n\beta_2}{n+1} - 1\right)^2 .$$

This is minimized when β_2 is $1 + n^{-1} + o(n^{-1})$. Since it is independent of θ, the bootstrap also yields the optimal value of β_2. In this example, U_n converges to θ at rate n, and the removal of the bias is reflected in first order asymptotic properties. Indeed, $n[U_n - \theta]$ converges weakly to $-\theta X$, where X has the exponential distribution with mean 1. However, $n[((n+1)/n)U_n - \theta]$ converges weakly to $-\theta X + \theta$, and so has an asymptotic bias equal to 0.

To obtain a general result on the bootstrap estimation of bias, we will need the following assumption which is quite similar to Assumption A of Example 4. Here, we consider the i.i.d. case, with the extension to other situations left to the reader. Let $D_n(P)$ be the law of $\tau_n[U_n(x_n) - \theta(P)]$ based on a sample x_n of size n from P, for some sequence τ_n tending to ∞. For simplicity, we focus on the additive bias adjustment by always setting $\beta_2 = 1$. Similar results could be obtained by considering a multiplicative adjustment or a combination of the two. For convenience, we modify the notation so the problem is to choose β among the class of estimators

$$T_{n,\beta}(x_n) = \beta\tau_n^{-1} + U_n(x_n) .$$

Multiplying by τ_n^{-1} does not change the resulting choice of estimators; it merely changes the name of the index β.

ASSUMPTION B. Let $\boldsymbol{C}_P$ be a set of sequences of distributions $\{P_n\}$ satisfying $D_n(P_n)$ converges weakly to a continuous limit distribution D_P whenever $\{P_n\} \in \boldsymbol{C}_P$, and $\theta(P_n) \to \theta(P)$. If U has distribution D_P, then $E_P[l|\beta + U|]$ has a unique minimum $\beta(P)$ in β.

THEOREM 3.3. *Let x_n be a sample of size n from P. Assume Assumption* B. *Let $\hat{Q}_n$ be an estimate of P based on x_n such that $\{\hat{Q}_n\}$ falls in $\boldsymbol{C}_P$ with probability one. Then, the bootstrap estimate $\hat{\beta}_n$ minimizing $E_{\hat{Q}_n} l\{\tau_n|\beta\tau_n^{-1} + U_n(x_n^*) - \theta(\hat{Q}_n)|\}$ converges to $\beta(P)$ almost surely, for any bounded, continuous loss function l. Moreover, bootstrap confidence intervals for $\theta(P)$ based on $T_{n,\hat{\beta}_n}$ are asymptotically valid.*

The proof is omitted, as it is completely analogous to the proof of Theorem 3.2. Note that when D_P in Assumption B is normal with mean μ_P and variance σ_P^2, then the optimal (asymptotic) value of β is $-\mu_P$. More-

over, μ_P and hence the optimal $\beta(P)$ are typically zero. In this sense, the problem is typically distinct from that of Example 4 in that the asymptotic solution is often degenerate (because the optimal value of 0 is known and does not depend on P). In Example 5(3), however, the limiting distribution of U_n was not normal and did not have mean 0, and the bias reduction was nontrivial even at the first order level. Also, the choice of the loss function was important.

Thus, the optimal value of β is typically 0 in asymptotically normal situations. Although the study of second order properties are beyond the scope of this paper, the bootstrap estimator $T_{n,\hat{\beta}_n}$ is generally better than $T_{n,0}$. To heuristically appreciate why, U_n typically has an expected value equal to

$$\theta(P) + \frac{a(P)}{n} + O(n^{-2}) . \tag{3.7}$$

Then, for squared error loss, the estimator $\beta + U_n$ has the same variance as U_n and an expected value equal to (3.7) plus β. The bootstrap choice of β amounts to choosing β to be $a(\hat{Q}_n)/n$, for some estimate $\hat{Q}_n$ of P. Since $a(\hat{Q}_n)$ also typically has an expectation equal to $a(P) + O(n^{-1})$, it should follow that $U_n - a(\hat{Q}_n)/n$ has an expectation equal to $\theta(P)$ plus a term of order n^{-2}. This result is typically true in smooth parametric and smooth nonparametric problems. For more details of bootstrap bias reduction, see Efron (1979). Also, the recent work of Hall and Martin (1988) considers the iterative bootstrap reduction of bias. This works by iterating the above method. That is, given the new bias is the reduced estimator $U_n^1 = U_n + \hat{\beta}_n$, consider choosing β among the new class of estimators $U_n^1 + \beta$. Applying the bootstrap procedure yields $U_n^1 + \hat{\beta}_n^1$. Repeat the procedure for this new estimator, and so on. See Hall and Martin (1988) for a most interesting discussion of iterative bootstrap methods.

Example 1. *Mean versus median* (*continued*). Before discussing the details of Example 1, consider the more general situation where the index set I is a finite set, say $I = \{1, 2, \ldots, d\}$. In this case, the main weak convergence hypothesis in Theorem 2.1 is reduced to studying the weak convergence of Z_n given by (2.2) as a random variable on $\boldsymbol{R}^d$. Actually, based on the fact that real numbers $y_{n,j}$ converge to y_j, for $j = 1, \ldots, d$ implies that $y_{n,j}$ converges to y_j uniformly in j, one can deduce the conclusions of Theorem 2.1 under a weaker assumption. That is, one only needs to study the asymptotic behavior of $Z_n(\beta, x_n, P_n)$ for a fixed β. Specifically, the assumption that $J_n(P_n)$ converges weakly to $J(P)$ is replaced by that $J_n(\beta, P_n)$ converges weakly to $J(\beta, P)$ for each β. The details are left to the reader. The reason that such an assumption is nice is

that its verification can be deduced from known properties about the bootstrap distribution of $T_{n,\beta}$ for a fixed β. The behavior of the bootstrap distribution of various types of estimators has been well-studied in Bickel and Freedman (1981) and Beran (1984), for example.

The problem of choosing β among a finite set I has several potential applications to model selection and regression problems. For example, consider fitting a polynomial of degree less than or equal to d to data (x_i, y_i) or consider the problem of selecting which of a finite number of variables should be included in a regression equation. While certain asymptotic results may be deduced for these problems by the methods in this paper, we defer them to subsequent work. In such problems, it seems necessary to study asymptotic properties where d tends to infinity with n if one is to believe the appropriateness and validity of bootstrap approximations in finite samples.

In the mean versus the median example, the verification of the main weak convergence assumption can be deduced from Proposition 3.1. In the case of squared error loss (or any loss function $l(t)$ bounded above by t^2), one also needs to know that bootstrap estimates of variance of the sample mean and sample median are consistent. Under the hypothesis of a finite variance of the underlying population (as already assumed in Proposition 3.1), this is easily seen to be the case (see Ghosh *et al.* (1984)).

It is perhaps worthwhile to note the following. The bootstrap approach asymptotically picks the best estimator $T_{n,i}$ with probability approaching one. On the other hand, if $T_{n,1}$ is the sample mean and $T_{n,2}$ is the sample median and the underlying distribution is normal with mean 0, it is not the case that the probability that $|T_{n,1}| < |T_{n,2}|$ occurs approaches one; this follows from Proposition 3.1.

To gain some insight into how well the bootstrap works for small sample sizes, some simulation results are presented in Table 1. In particular, the loss function is squared error loss and $\hat{Q}_n$ is the empirical distribution of the data X_i. This is especially convenient because the bootstrap estimate of the mean squared error of the sample mean can be calculated without simulation and is equal to

$$\hat{\sigma}_n^2 = \sum_{i=1}^{n} (X_i - \bar{X}_n)^2 / n \,.$$

On the other hand, the bootstrap estimate of the mean squared error of the sample median is calculated by simulation of 100 bootstrap samples. The resulting estimator, denoted $\hat{T}_{BOOT}$ in Table 1, picks the estimator with the smallest bootstrap estimate of the mean squared error. In Table 1, the mean squared errors of the sample mean and sample median are also reported for comparison. Actually, all mean squared errors are multiplied by the sample size n for easier comparison over values of n. For example,

consider situation 1 where the underlying population is normal with variance one. The sample mean has a standardized mean squared error (MSE) of 1.00. The sample median, in the case $n = 20$, has a standardized MSE of 1.479; this number is obtained based on 10,000 simulations, using the variance reduction technique of Johnstone and Velleman (1985). The bootstrap estimator $\hat{T}_{BOOT}$, based on 10,000 simulations, has a mean squared error of 1.079, and suffers only a small loss in efficiency relative to the efficient sample mean. The results are slightly less favorable in situation 2 where the underlying population is double exponential, but are still encouraging. Perhaps the results would improve with a better estimate of the bootstrap variance of the sample median (see Hall *et al.* (1989)).

Example 3. *Bandwidth selection in density estimation* (*continued*). Recall the setting of Example 3 introduced in Section 1. Assume the kernel K in (1.2) is bounded, has mean 0, has an integrable j-th derivative $K^{(j)}$ for $j = 0, 1, 2$ satisfying $K^{(j)}(x) \to 0$ as $|x| \to \infty$. Also, assume K^3 is integrable and set

$$C_1 = \int x^2 K(x)\,dx$$

and

$$C_2 = \int K^2(x)\,dx\,.$$

We will also assume the unknown density f is bounded and twice differentiable with $f^{(2)}$ uniformly continuous. These assumptions can be weakened somewhat, particularly if one is not interested in estimating f everywhere, but we do not dwell on the best technical assumptions needed here.

First, consider the problem of estimating $\theta = f(t)$ at some fixed t. Recall the following facts, as developed in Parzen (1962). If h_n is a fixed bandwidth sequence and if $nh_n^2 \to \infty$ and $h_n \to 0$, then $\hat{g}_{n,h_n}(t)$ is asymptotically normal. Moreover,

$$h_n^{-2}[E_f \hat{g}_{n,h_n}(t) - f(t)] \to C_1 f^{(2)}(t)/2$$

and

$$nh_n \operatorname{Var}_f[\hat{g}_{n,h_n}(t)] \to f(t)\,C_2\,.$$

Assuming $f(t)$ and $f^{(2)}(t)$ are nonzero, the asymptotically optimal choice of h_n minimizing the mean squared error then satisfies

$$n^{1/5} h_n \to [f(t)\,C_2]^{1/5}[C_1 f^{(2)}(t)]^{-2/5}\,. \tag{3.8}$$

Hence, it is convenient for asymptotic purposes to reparametrize the problem by setting $\hat{f}_{n,\beta}(t) = \hat{g}_{n,n^{-1/5}\beta}(t)$, and the problem is to choose β so that $\hat{f}_{n,\beta}(t)$ best estimates $f(t)$. Thus, the asymptotic best choice of β, β_f, depends on f and, for squared error loss, is given by the right side of (3.8).

In order to apply the theorems of Section 2, first assume $I = [a, b]$, where $a > 0$ and $b < \infty$. The main weak convergence assumption of Theorems 2.1 and 2.2 is verified by the following.

PROPOSITION 3.2. *Let $\mathbf{C}_f$ be a set of sequences of distributions $\{F_n\}$ on the line with densities $\{f_n\}$ such that, for $j = 0, 1, 2$, $f_n^{(j)}$ converges uniformly to $f^{(j)}$. Let $X_{n,1}, \ldots, X_{n,n}$ be a sample of size n from f_n and set*

$$\hat{f}_{n,\beta}(t) = \beta^{-1} n^{-4/5} \sum_{i=1}^{n} K\left(\frac{t - X_{n,i}}{n^{-1/5}\beta} \right).$$

Let

$$Z_n(\beta) = Z_n(\beta, t) = n^{2/5}[\hat{f}_{n,\beta}(t) - f_n(t)], \tag{3.9}$$

so that the appropriate normalization is $\tau_n = n^{2/5}$. Then, $Z_n(\cdot)$, regarded as a random element of $L_\infty(I)$, converges weakly to a continuous Gaussian process Z with mean

$$E[Z(\beta)] = \beta^2 C_1 f^{(2)}(t)/2 \tag{3.10}$$

and covariance function

$$\operatorname{Cov}[Z(\beta_1), Z(\beta_2)] = f(t)(\beta_1\beta_2)^{-1/2} \int K(rz) K(r^{-1}z)\, dz, \tag{3.11}$$

where $r = (\beta_1/\beta_2)^{1/2}$.

The proof of Proposition 3.2 is relatively straightforward because $Z_n(\cdot)$ is a sum of independent identically distributed variables. It is similar to the proof of Theorem 2.1 of Romano (1988*a*) and Lemma 4.1 of Romano (1988*b*). The only difficulty is verifying tightness, but this can readily be obtained by application of Theorem 12.3 of Billingsley (1968).

Note that the optimal value of β actually depends on the loss function l, mainly due to the fact that the limiting process $Z(\cdot)$ does not have mean 0. In general, the optimal value of β is the value of β minimizing $E[l\{|Z(\beta)|\}]$, where $Z(\beta)$ is normal with mean given by (3.10) and variance given by (3.11) with $\beta = \beta_1 = \beta_2$.

In order to apply Proposition 3.2 to verify the assumptions of Theorems 2.1 and 2.2, we need to specify an appropriate resampling

distribution $\hat{Q}_n$. Given $x_n = (X_1, \ldots, X_n)$ from f, let $\hat{Q}_n$ be the distribution with density $\hat{g}_{n,h_n}(\cdot)$ given by (1.2). If $nh_n^5/\log(n) \to \infty$ and $h_n \to 0$, then $\hat{Q}_n$ falls in $\boldsymbol{C}_f$ with probability one (see Silverman (1978)). Furthermore, if $nh_n^5 \to c$ for some $c < \infty$, then the bootstrap will not work even though this is the optimal rate for estimating $f(t)$. The reason can be traced to the fact that $\hat{g}^{(2)}_{n,h_n}$ does not consistently estimate $f^{(2)}$ for such a sequence h_n. Hence, the bootstrap sampling distribution does a poor job of estimating the bias component (see Romano (1988*a*, 1988b) where this is observed in the context of modal estimation).

The convergence of Z_n to Z in Proposition 3.2 does not extend to the case $I = [0, \infty)$. While Theorems 2.1 and 2.2 provide results concerning optimal choices of β in $[a, b]$, we would like to extend these to β in $[0, \infty)$. By Technical Remark 4 of Section 2, it suffices to show that the bootstrap choice, $\hat{\beta}_n$, of β lies in some $[a, b]$, where $a > 0$ and $b < \infty$ with probability approaching one. To do this, consider the squared error loss. Then, $\hat{\beta}_n$ is obtained by minimizing $R_n(\beta, \hat{Q}_n)$. To see, for example, that the minimizing β is bounded away from infinity, consider the behavior of $R_n(\beta_n, \hat{Q}_n)$, where β_n is any sequence tending to ∞. A direct calculation shows that $R_n(\beta_n, \hat{Q}_n) \to \infty$ because a too large bandwidth β_n results in a large bias component for the risk function. Similarly, if $\beta_n \to 0$, the variance component gets large. For more details, see equation (4.15) of Parzen (1962), but generalized to the case that f varies with n. In any case, the minimizing β must be bounded away from 0 and ∞ and so that Proposition 3.2 is again applicable. For other loss functions, similar arguments work as well.

In the case of constructing a confidence band for $\theta = f(\cdot)$, θ is not real-valued. However, by considering $Z_n(\beta, t)$ defined by (3.9) as process on the product space of $[a, b]$ and the real line, bootstrap convergence results about the bootstrap choice of β can similarly be deduced from weak convergence results of the behavior of Z_n under sequences f_n. A start in this direction is given in Bickel and Rosenblatt (1973) who consider $Z_n(\beta, t)$ as a process in t under a fixed f, whereas in the case of fixed t above, we considered Z_n as a process in β. The technical considerations of treating Z_n as a process in both β and t under general sequences f_n will be treated elsewhere, as the calculations are too involved. In principle, however, the technical approach is straightforward since $Z_n(\beta, t)$ is still a sum of i.i.d. variables. Faraway and Jhun (1988) consider this problem of bootstrap bandwidth selection and constructing confidence bands for f. They conclude, based on fairly extensive simulation results, that the bootstrap outperforms cross-validation.

Bandwidth selection for other functionals of a density may also be considered by similar methods. For example, to select the bandwidth to estimate the mode of f, the main weak convergence hypothesis of Theorems 2.1 and 2.2 follows as in the proof of Theorem 2.1 of Romano (1988*a*).

REFERENCES

Beran, R. (1984). Bootstrap methods in statistics, *Jber. d. Dt. Math.-Verein*, **36**, 847–856.

Beran, R. (1986). Simulated power functions, *Ann. Statist.*, **14**, 151–173.

Beran, R. and Millar, W. (1987). Stochastic estimation and testing, *Ann. Statist.*, **15**, 1131–1154.

Bickel, P. J. and Freedman, D. A. (1981). Some asymptotic theory for the bootstrap, *Ann. Statist.*, **9**, 1196–1217.

Bickel, P. and Rosenblatt, M. (1973). On some global measures of the deviations of density function estimates, *Ann. Statist.*, **1**, 1071–1095.

Billingsley, P. (1968). *Convergence of Probability Measures*, Wiley, New York.

Bowman, A., Hall, P. and Titterington, D. (1984). Cross-validation in nonparametric estimation of probabilities and probability densities, *Biometrika*, **71**, 341–351.

Cox, D. and Hinkley, D. (1974). *Theoretical Statistics*, Chapman and Hall, London.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife, *Ann. Statist.*, **7**, 1–26.

Faraway, J. and Jhun, M. (1988). Bootstrap choice of bandwidth for density estimation, Tech. Report No. 157, Department of Statistics, University of Michigan.

Ghosh, M., Parr, W., Singh, K. and Babu, G. (1984). A note on bootstrapping the sample median, *Ann. Statist.*, **12**, 1130–1135.

Hall, P. and Martin, M. (1988). On bootstrap resampling and iteration, *Biometrika*, **75**, 661–671.

Hall, P., DiCiccio, T. and Romano, J. (1989). On smoothing and the bootstrap, *Ann. Statist.*, **17**, 692–704.

Jaeckel, L. (1971). Some flexible estimates of location, *Ann. Math. Statist.*, **43**, 1041–1067.

Johnstone, I. and Velleman, P. (1985). Efficient scores, variance decompositions, and Monte Carlo swindles, *J. Amer. Statist. Assoc.*, **80**, 851–862.

Léger, C. (1988). Use of the bootstrap in an adaptive statistical procedure, Tech. Report No. 296, Department of Statistics, Stanford University.

Lehmann, E. (1983). *Theory of Point Estimation*, Wiley, New York.

Parzen, E. (1962). On estimation of a probability density and mode, *Ann. Statist.*, **33**, 1065–1076.

Pollard, D. (1984). *Convergence of Stochastic Processes*, Springer, New York.

Pruitt, R. (1988). Cross-validation in the one sample location problem, Tech. Report No. 510, School of Statistics, University of Minnesota.

Romano, J. (1988*a*). On weak convergence and optimality of kernel density estimates of the mode, *Ann. Statist.*, **16**, 629–647.

Romano, J. (1988*b*). Bootstrapping the mode, *Ann. Inst. Statist. Math.*, **40**, 565–586.

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

Sheehy, A. and Wellner, J. (1988). Uniformity in P of some limit theorems for empirical measures and processes, Tech. Report 134, Revision 2, Department of Statistics, University of Washington.

Silverman, B. (1978). Weak and strong uniform consistency of the kernel estimate of a density and its derivatives, *Ann. Statist.*, **6**, 177–184.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions, *J. Roy. Statist. Soc. Ser. B*, **36**, 111–147.

Stone, M. (1977). Asymptotics for and against cross-validation, *Biometrika*, **64**, 29–35.