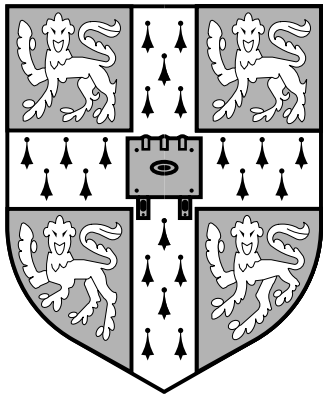
Some mathematical and theoretical aspects of the bootstrap

Richard John Samworth

St John's College and Statistical Laboratory

University of Cambridge



A dissertation submitted for the

degree of Doctor of Philosophy

Some mathematical and theoretical aspects of the bootstrap

Richard John Samworth

Abstract

This thesis contributes to the mathematical and theoretical understanding of the bootstrap, introduced into Statistics by Efron in 1979. It consists of four main chapters, the first of which examines cases of bootstrap inconsistency, with particular reference to the Hodges and Stein estimators. We evaluate a proposal of Beran for the diagnosis of inconsistency and investigate modifications, such as the m out of n bootstrap, which restore consistency.

The bootstrap distribution of the Stein estimator is of interest as a method of constructing confidence sets for the mean of a spherically symmetric distribution. We study these confidence sets in Chapter 2, where we also propose an analytic formulation, and give mathematical results as well as numerical simulations to suggest their improvements over the usual confidence set.

Recently, the bootstrap has been applied to classification problems in an attempt to reduce the error rate of a classifier, using a technique known as *bagging*. In Chapter 3 we study the bagged nearest-neighbour classifier, and advocate bagging with smaller resample sizes than the actual sample sizes.

The final chapter considers the question: 'What is the probability that the bootstrap performs badly?'. We formulate this problem in terms of the Mallows distance between the bootstrap distribution of a sample mean and its true sampling distribution, whose properties are found to depend on the tail of the underlying population.

Contents

Preface

The May 2003 issue of *Statistical Science* is a commemoration of the Silver Anniversary of the bootstrap, and is a fascinating reminder of the impact the bootstrap has had on statistical theory and practice. Born in May 1978, I am one of the first generation of 'bootstrap babies', brought up in the age of the personal computer, where numerical approximations of almost unlimited precision can be found to previously intractable problems.

Despite our newfound capabilities, however, it is important not to lose sight of the fact that practical implementation of the bootstrap is almost always a two-stage approximation procedure. The first step is to approximate the distribution of the statistic of interest, $\hat{\theta}$, under the unknown distribution of the data, $\mathcal{P}$, by the bootstrap distribution of $\hat{\theta}$; that is, the distribution of $\hat{\theta}$ under some data-driven estimate, $\hat{\mathcal{P}}$, of $\mathcal{P}$. The second stage is to simulate samples from $\hat{\mathcal{P}}$, computing $\hat{\theta}$ on each sample. Since the power of the modern computer generally enables sufficiently many samples to be drawn to ensure that the simulation error in the second step is small, the issue that lies at the heart of the bootstrap is the validity of the first approximation.

This thesis examines this bootstrap approximation in various contexts, and from different perspectives. Chapter 1 focuses on non-regular circumstances in which the standard bootstrap procedure is known to be inconsistent. This work is closely con-

nected with two other important statistical discoveries of the last century, namely the phenomena of superefficiency and Stein estimation. The surprisingly good performance of the standard bootstrap compared with its competitors leads us to question the prominence attached to bootstrap consistency. An abbreviated version of this chapter is to appear in the December 2003 issue of *Biometrika*.

A natural motivation for studying the bootstrap distribution of the Stein estimator is as a method for constructing improved confidence sets for the mean of a multivariate normal distribution, or more generally, a spherically symmetric distribution. This has applications to the problem of finding a confidence set for the regression coefficients in the linear model, and is the subject of Chapter 2. We present and discuss two different confidence set constructions, one via an analytic method and the other via the bootstrap.

Chapter 3 concerns an emerging and exciting area of applications of the bootstrap, to prediction and classification problems. Our focus is on bagging, short for 'bootstrap aggregating', where predictors may be improved by averaging the results of predictions made based on resamples of the data. We consider applying bagging to the nearest-neighbour classifier, showing in particular that the resample sizes must be small in comparison with the actual sample sizes for bagging to result in an asymptotic improvement. The idea of using reduced resample sizes in conjunction with the bootstrap is another theme discussed in Chapter 1.

Finally, in Chapter 4 we examine a different aspect of bootstrap performance, namely its large deviations properties. This work was partly motivated by some of the simulations in Chapter 1, although our results in this chapter concern the sample mean and the Mallows metric. Broadly, we find encouraging large deviations behaviour of the bootstrap if the tail of the underlying distribution of the data is finite, but that this is not necessarily the case if the tail is heavy.

It is a great pleasure to have the opportunity to thank my supervisor, Alastair Young, for his support, encouragement and sound advice. I have also been extremely fortunate to have shared an office with Christina Goldschmidt and, latterly, Michail Loulakis. They, as well as other friends and colleagues, especially Olly Johnson, Lara Jamieson, Gareth Birdsall, Amanda Turner, John Harper, Ruth King, Damon Wischik, Edward Crane and Paul Russell, have been very generous with their time and helped to make the Statslab such a friendly environment.

I am grateful to the Engineering and Physical Sciences Research Council for their financial support, and to Peter Hall for arranging a Visiting Fellowship for me at the Australian National University in the early summer of 2003.

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the text.

Chapter 3 of this thesis is based on work carried out during my visit to the Australian National University, and, excluding the Appendix but including Theorem 3.4.1, this chapter is joint work with Peter Hall. A condensed version has been submitted to the Journal of the Royal Statistical Society, Series B. Certain results in Chapter 4, specifically Proposition 4.2.1, Lemmas 4.5.6, 4.5.7 and 4.5.8 and Theorem 4.5.11, are joint work with Edward Crane. I intend to submit both Chapters 2 and 4 for publication very shortly.

R. J. Samworth

Cambridge, 2003

Chapter 1

Bootstrap Diagnostics and Inconsistency

1.1 Introduction

Asymptotic analysis, usually as the sample size tends to infinity, has been an important tool for developing and understanding many statistical procedures. The bootstrap is no exception, and limit theorems have played a prominent role ever since Efron (1979) introduced the idea and began the process of establishing its validity.

The potential of Efron's idea was quickly seized upon, and Bickel and Freedman (1981) gave general conditions under which the bootstrap could be expected to be consistent, as well as studying many examples. Singh (1981) provided a more detailed asymptotic treatment of the standardised sample mean, which revealed the second-order accuracy of the bootstrap, and hence the possibility of improvement over traditional normal approximation. The success of his analysis set a trend among many authors to use the

powerful machinery of Edgeworth expansions in the study of the bootstrap, notably Hall (1992), who built on earlier work in a non-bootstrap context by Bhattacharya and Rao (1976). Beran (1982) argued that the bootstrap is asymptotically minimax. More recently, saddlepoint approximations have been applied to examine the relative error properties of the bootstrap (e.g. Jing, Feuerverger and Robinson, 1994).

It is consistency, though, which is seen as the sine qua non for the bootstrap. Many authors refer to bootstrap 'failure' in cases of inconsistency, and 'success' otherwise. This terminology may be inappropriate, however, for two reasons. Firstly, the sample size may not be large enough for the asymptotics to accurately reflect the finite-sample situation. More importantly, consistency is a fixed parameter property: there is generally no guarantee that any convergence is uniform over the parameter space.

An important contribution to the study of bootstrap consistency was made by Beran (1997), who considered locally asymptotically normal models, and characterised consistency in terms of an asymptotic independence property. This result is the basis for his graphical diagnostic, intended to give justification for the validity of the standard bootstrap approach, or to warn of its possible unreliability. This idea was followed and developed in a more practical setting by Canty, Davison, Hinkley and Ventura (2000). Beran also proved that asymptotic superefficiency is a sufficient condition for bootstrap inconsistency, and cited the Hodges and Stein estimators as examples of this phenomenon.

Several issues arise in the implementation of Beran's diagnostic; these are discussed in Section 1.2. We are led to formalise the procedure with reference to the examples above, in order to compare it with various alternatives. Our conclusion is that well-known existing procedures may be more suitable for diagnosing inconsistency in these instances.

It is natural next to consider the best course of action if faced with the possibility that the standard n out of n bootstrap may be inconsistent. It has been suggested that one should reduce the bootstrap resample size, an idea which dates back to Bretagnolle (1983). The use of this device has been shown to lead to consistent estimators in wide generality, but typically there is an asymptotic loss of efficiency in cases where the standard bootstrap is known to work successfully. Recent work, such as Bickel, Götze and van Zwet (1997) and Politis, Romano and Wolf (1999), has focused on remedying these losses. If entirely successful, this would negate the need for a diagnostic; but even then, further questions, especially the difficult choice of the reduced bootstrap resample size, remain. We examine both theoretically and empirically in the Hodges and Stein examples whether efficiency losses are manifested in finite samples, whether an optimal choice of resample size can be suggested and also investigate other alternatives which restore consistency. All proofs are given in Section 1.5.

1.2 Local asymptotic normality and the bootstrap

In this section, we describe the locally asymptotically normal (LAN) model, which was introduced into Statistics by Le Cam (1960) in his study of asymptotically similar tests. In addition, we introduce the bootstrap and outline the concepts necessary to understand the relevant version of Beran's key theorem (Theorem 1.2.6).

Suppose $X_1, \ldots, X_n$ are independent and identically distributed random vectors in $\mathbb{R}^m$, and write $\mathbb{P}_\theta$ for the distribution of $X = (X_1, \ldots, X_n)$. The parameter θ belongs to a parameter space Θ, which we assume is an open subset of $\mathbb{R}^k$. Suppose that the components of X have density p_θ with respect to Lebesgue measure on $\mathbb{R}^m$, and for $h \in \mathbb{R}^k$, let $L_n(h, \theta)$ denote the log-likelihood ratio of $\mathbb{P}_{\theta + n^{-1/2}h}$ with respect to $\mathbb{P}_\theta$.

Thus,

$$L_n(h,\theta) = \log\Bigg(\prod_{i=1}^{n} \frac{p_{\theta+n^{-1/2}h}(X_i)}{p_\theta(X_i)} \Bigg).$$

Definition 1.2.1. *The model $\{\mathbb{P}_\theta : \theta \in \Theta\}$ is LAN at θ_0 if there exist a random vector $Y_n(\theta_0, X) \in \mathbb{R}^k$ and a non-singular $k \times k$ matrix $I(\theta_0)$ such that under $\mathbb{P}_{\theta_0}$ we have both $Y_n(\theta_0, X) \xrightarrow{d} N_k\big(0, I(\theta_0)\big)$, and*

$$L_n(h_n, \theta_0) = h^T Y_n(\theta_0, X) - \tfrac{1}{2} h^T I(\theta_0) h + o_p(1)$$

as $n \to \infty$, for every $h \in \mathbb{R}^k$ and every sequence (h_n) in $\mathbb{R}^k$ converging to h.

Local asymptotic normality acquires its name from the fact that the log-likelihood ratio in LAN models is asymptotically the same as that of $N\big(h, I^{-1}(\theta_0)\big)$ with respect to $N\big(0, I^{-1}(\theta_0)\big)$. Thus an LAN model $\{\mathbb{P}_{\theta_0+n^{-1/2}h} : h \in \mathbb{R}^k\}$ and the model $\big\{N\big(h, I^{-1}(\theta_0)\big) : h \in \mathbb{R}^k\big\}$ are similar in their statistical properties. Note here that the original model $\{\mathbb{P}_\theta : \theta \in \Theta\}$ has been reparametrised in terms of a local parameter $h = n^{1/2}(\theta - \theta_0)$.

A Taylor expansion argument shows that in our case of independent and identically distributed random variables, the LAN property is satisfied under mild regularity conditions on the log-likelihood $\ell_x(\theta) = \log p_\theta(x)$ (van der Vaart, 1998, pp. 93–95). The sequence $Y_n(\theta, x)$ and Fisher information matrix $I(\theta)$ are then related to the score function, $\nabla_\theta \ell_x(\theta)$, through

$$Y_n(\theta, X) = \frac{1}{n^{1/2}} \sum_{i=1}^{n} \nabla_\theta \ell_{X_i}(\theta) \quad \text{and} \quad I(\theta) = \mathbb{E}_\theta\big(\nabla_\theta \ell_X(\theta)\, \nabla_\theta \ell_X(\theta)^T\big).$$

Let $T_n = T_n(X)$ be an estimator of θ. Of interest is the *root* $n^{1/2}(T_n - \theta)$, and we denote its sampling distribution under $\mathbb{P}_\theta$ by $H_n(\theta)$. Statistical considerations such as the construction of confidence sets for θ motivate the study of such roots. If $\hat{\theta}_n = \hat{\theta}_n(X)$ is another estimator of θ, then the (parametric) bootstrap distribution

estimator of $H_n(\theta)$ is $H_n(\hat{\theta}_n)$. As defined, the bootstrap distribution is a random probability measure, although we usually study it as a conditional distribution, for given X.

Definition 1.2.2. *Suppose d is a metric on the space of probability measures on $\mathbb{R}^k$. We say that $H_n(\hat{\theta}_n)$ is d-consistent at θ_0 if for all $\epsilon > 0$,*

$$\mathbb{P}_{\theta_0}\{d\big(H_n(\hat{\theta}_n), H_n(\theta_0)\big) > \epsilon\} \to 0 \tag{1.1}$$

as $n \to \infty$.

We shall be primarily interested in the topology of weak convergence. If (1.1) holds for a metric which metrises weak convergence, we will simply say $H_n(\hat{\theta}_n)$ is consistent at θ_0. If, in addition, there exists a limit distribution $H(\theta_0)$ such that $H_n(\theta_0)$ converges in distribution to $H(\theta_0)$, we write $H_n(\hat{\theta}_n) \xrightarrow{d} H(\theta_0)$ in $\mathbb{P}_{\theta_0}$-probability as $n \to \infty$.

Often, consistency is proved by verifying the conditions of the following proposition, which is a version of Theorem 1 of Beran (1984).

Proposition 1.2.3. *Let $\theta_0 \in \Theta$, and suppose that the following conditions hold:*

(i) there exists a limit distribution $H(\theta_0)$ such that $H_n(\theta_n) \xrightarrow{d} H(\theta_0)$ as $n \to \infty$ for every sequence (θ_n) in Θ converging to θ_0;

(ii) there exists a sequence of estimators $(\hat{\theta}_n)$ such that $\hat{\theta}_n \to \theta_0$ in $\mathbb{P}_{\theta_0}$-probability as $n \to \infty$.

Then $H_n(\hat{\theta}_n) \xrightarrow{d} H(\theta_0)$ in $\mathbb{P}_{\theta_0}$-probability as $n \to \infty$.

Beran (1997) shows the importance of *local asymptotic equivariance* in determining bootstrap consistency:

Definition 1.2.4. *Suppose that* $H_n(\theta_0) \xrightarrow{d} H(\theta_0)$ *as* $n \to \infty$. *The sequence of estimators* (T_n) *of* θ *is locally asymptotically equivariant at* θ_0 *if for every* $h \in \mathbb{R}^k$ *and every sequence* (h_n) *in* $\mathbb{R}^k$ *converging to* h, *we have*

$$H_n(\theta_0 + n^{-1/2}h_n) \xrightarrow{d} H(\theta_0)$$

as $n \to \infty$.

Local asymptotic equivariance is a slightly stronger property than that of *regularity* (Hájek, 1970), which only requires the above convergence to hold with $h_n = h$ for all n.

Before we can state the main theorem, we need to define one final property of estimators, typically satisfied by maximum likelihood estimators in exponential families and, more generally, by one-step maximum likelihood estimators (van der Vaart, 1998, pp. 71–75) in LAN models.

Definition 1.2.5. *A sequence of estimators* $(T_{n,E})$ *is asymptotically efficient at* θ_0 *if, under* $\mathbb{P}_{\theta_0}$,

$$T_{n,E} = \theta_0 + n^{-1/2} I^{-1}(\theta_0) Y_n(\theta_0, X) + o_p(n^{-1/2})$$

as $n \to \infty$.

We suppose the existence of such a sequence of estimators, and write $K_n(\theta)$ for the joint distribution of $\big(n^{1/2}(T_n - T_{n,E}), Y_n(\theta, X)\big)$ under $\mathbb{P}_\theta$.

Theorem 1.2.6 (Beran). *Suppose that the model* $\{\mathbb{P}_\theta : \theta \in \Theta\}$ *is LAN at* θ_0 *and that* $H_n(\theta_0) \xrightarrow{d} H(\theta_0)$ *as* $n \to \infty$. *Suppose that the estimator* $\hat{\theta}_n$ *used to construct the bootstrap distribution satisfies the condition that* $n^{1/2}(\hat{\theta}_n - \theta_0)$ *converges in distribution, under* $\mathbb{P}_{\theta_0}$, *to a limit distribution which has full support in* $\mathbb{R}^k$. *Then the following are equivalent:*

(a) $H_n(\hat{\theta}_n) \xrightarrow{d} H(\theta_0)$ *in* $\mathbb{P}_{\theta_0}$*-probability as* $n \to \infty$*;*

(b) $K_n(\hat{\theta}_n) \xrightarrow{d} D(\theta_0) \times N\big(0, I^{-1}(\theta_0)\big)$ *in* $\mathbb{P}_{\theta_0}$*-probability as* $n \to \infty$*, for some distribution* $D(\theta_0)$ *such that* $H(\theta_0)$ *can be written as the convolution of* $D(\theta_0)$ *and* $N\big(0, I^{-1}(\theta_0)\big)$*;*

(c) the sequence of estimators (T_n) *are locally asymptotically equivariant at* θ_0 *with limit distribution* $H(\theta_0)$*.*

Thus, in LAN models, part (c) of the theorem gives a means of verifying the bootstrap consistency in part (a). Beran's diagnostic is based on the asymptotic independence in part (b) of the theorem:

(1) Given $X_1, \ldots, X_n$, compute $\hat{\theta}_n$, and then generate B independent bootstrap samples $X_i^* = \{X_{1,i}^*, \ldots, X_{n,i}^*\}$ for $i = 1, \ldots, B$ from $\mathbb{P}_{\hat{\theta}_n}$.

(2) Compute $T_{n,i}^* = T_n(X_i^*)$, and $T_{n,E,i}^* = T_{n,E}(X_i^*)$ for $i = 1, \ldots, B$.

(3) Compute $a_i^* = n^{1/2}(T_{n,i}^* - T_{n,E,i}^*)$ and $d_i^* = Y_n(\hat{\theta}_n, X_i^*)$ for $i = 1, \ldots, B$.

(4) Choose real-valued, continuous functions f and g on $\mathbb{R}^k$ and plot the pairs $\big\{\big(f(a_i^*), g(d_i^*)\big) : i = 1, \ldots, B\big\}$ to assess whether the approximate independence breaks down. If so, mistrust the bootstrap distribution from this data set.

In this author's experience, this procedure can be ambiguous. On what basis do we decide what does and what does not look independent? How large does n need to be before we should expect to see independence at points of local asymptotic equivariance? How should we choose the scalar summaries f and g in the multi-dimensional case? Figure 1.1 shows the result of applying the algorithm above to the Stein estimator (defined in Section 1.3.2) with the functions f and g both chosen to

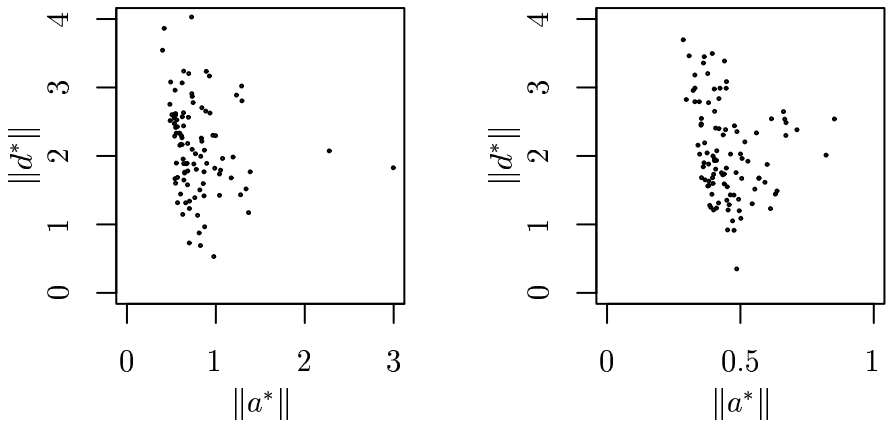


Figure 1.1: Beran's diagnostic applied to the Stein estimator. On the left-hand plot, $\theta = (0, 0, 0, 0, 0)$; on the right, $\theta = (-0.1, 0.1, 0, 0, 0)$. The choice of θ on the right ensures that the likelihood ratio test of size 0.05 of $H_0 : \theta_1 = \ldots = \theta_k$ versus $H_1 : H_0$ is not true, rejects H_0 with probability approximately 0.95 (see Section 1.3.2). Parameter values: $n = 1,000$, $B = 100$, $k = 5$.

be the Euclidean norm on $\mathbb{R}^k$. According to Theorem 1.2.6, we would like to be able to diagnose dependence on the left and independence on the right.

1.3 The Beran diagnostic and alternatives

1.3.1 The Hodges estimator

Let $X_1, \ldots, X_n$ be independent and identically distributed random variables, each distributed according to $N(\theta, 1)$, and let $\bar{X}_n = n^{-1} \sum_{i=1}^n X_i$. The Hodges estimator is defined by

$$T_{n,H}(\bar{X}_n) = \begin{cases} b\bar{X}_n & \text{if } |\bar{X}_n| \leq n^{-1/4} \\ \bar{X}_n & \text{otherwise,} \end{cases}$$

where $b \in (0,1)$. It is possibly the simplest example of an asymptotically superefficient estimator. The risk of the Hodges estimator, given by $\mathbb{E}_\theta\left(n^{1/2}(T_{n,H}-\theta)\right)$, converges to b^2 when $\theta = 0$ and to 1 otherwise (Lehmann, 1998, p. 442); the Cramér-Rao lower bound is 1. Though it is possible to modify the definition of the Hodges estimator to extend the set of asymptotic superefficiency to an arbitrary closed countable set (Le Cam, 1953), this is of limited practical benefit. The reason is that in one dimension, it is a feature of asymptotically superefficient estimators that at fixed n they should behave poorly in terms of risk near a point of asymptotic superefficiency (Le Cam, 1953; Hájek, 1972). Nevertheless, similar superefficient truncation estimators have been studied, for instance, in wavelet regression, where estimates of wavelet coefficients are discarded if smaller in modulus than some threshold value. Further details can be found in Canty, Davison, Hinkley and Ventura (2000).

We are interested in estimating the distribution $H_n(\theta)$ of $n^{1/2}(T_{n,H}-\theta)$, and consider the bootstrap approximation $H_n(\bar{X}_n)$. We will see in Section 1.4.1 that $H_n(\bar{X}_n)$ is consistent if and only if $\theta \neq 0$. We may take $T_{n,E} = \bar{X}_n$, so part (b) of Theorem 1.2.6 states that if $\theta \neq 0$, then $a^* = n^{1/2}(T^*_{n,H} - \bar{X}^*_n)$ and $d^* = n^{1/2}(\bar{X}^*_n - \bar{X}_n)$ are asymptotically independent in $\mathbb{P}_\theta$-probability, with marginal distributions a point mass at the origin and $N(0,1)$, respectively. Here, conditional on $X_1, \ldots, X_n$, we have that $X_1^*, \ldots, X_n^*$ are independent and identically distributed $N(\bar{X}_n, 1)$ random variables, $\bar{X}^*_n = n^{-1}\sum_{i=1}^n X_i^*$ and $T^*_{n,H} = T_{n,H}(\bar{X}^*_n)$.

In the remainder of this subsection, we assess the formal properties of the procedure described in the last paragraph of Section 1.2 when applied to this example. Expressed in the language of hypothesis testing, the clear implication of Beran's diagnostic is that we should take as our null hypothesis that the standard bootstrap works – in other words $\theta \neq 0$. This runs counter to the general philosophy of hypothesis tests, in which H_0 is the conservative hypothesis, to be rejected only if there is evidence

against it. More conventional, then, would be the null hypothesis that the standard bootstrap is inconsistent, i.e. $\theta = 0$. For this reason, we choose to swap over the null and alternative hypotheses.

With the new $\theta = 0$ null hypothesis, the theory of classical tests in exponential families gives that a uniformly most powerful unbiased (UMPU) test of size α is to reject H_0 if $n^{1/2}|\bar{X}_n| > \Phi^{-1}(1-\alpha/2)$, where Φ is the standard normal distribution function. Formalising Beran's method in this context requires the choice of a test statistic. Beran notes (albeit in the nonparametric bootstrap setting) that the sample correlation (or equivalently $T = \sum_{i=1}^{B} a_i^* d_i^*$) is unreliable due to the presence of outliers. For, either $a_i^* = 0$ or the points (a_i^*, d_i^*) lie on a line with gradient $-1/(1-b)$. Instead, he argues that a large proportion of points with $a_i^* = 0$ is evidence of independence, implicitly suggesting that we should take

$$T = \frac{1}{B}\sum_{i=1}^{B} \mathbb{1}_{\{a_i^*=0\}}$$

as our test statistic. We can compute the critical value, c, for the test as follows:

(1) Choose a test size, $\alpha \in (0,1)$, and an integer R such that $(R+1)(1-\alpha)$ is also an integer.

(2) For each $j = 1, \ldots, R$, repeat steps (3) to (5).

(3) Generate $\bar{Y}_{n,j} \sim N(0, 1/n)$.

(4) Generate $\bar{Y}^*_{n,i} \sim N(\bar{Y}_{n,j}, 1/n)$ independently for $i = 1, \ldots, B$.

(5) Compute $a_i^* = n^{1/2}\big(T_{n,H}(\bar{Y}^*_{n,i}) - \bar{Y}^*_{n,i}\big)$ for each $i = 1, \ldots, B$, and then evaluate $T_j^* = B^{-1}\sum_{i=1}^{B} \mathbb{1}_{\{a_i^*=0\}}$.

(6) Let $c = T^*_{((R+1)(1-\alpha))}$, i.e. the $((R+1)(1-\alpha))$th order statistic of $T_1^*, \ldots, T_R^*$.

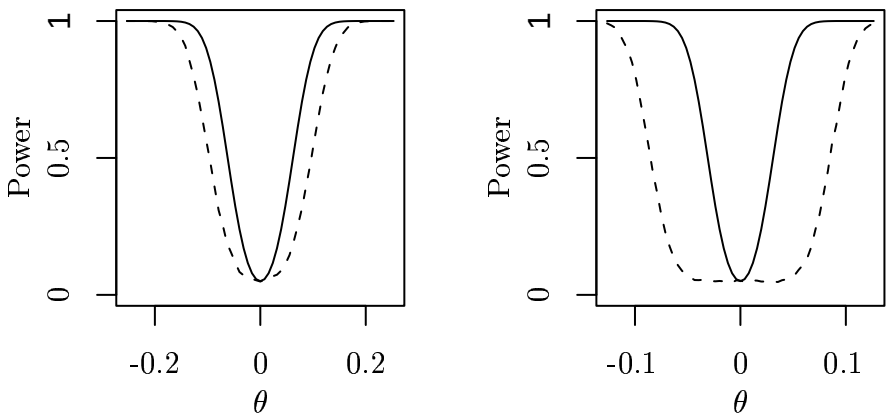


Figure 1.2: A comparison of the power functions of the UMPU test (solid) and the one derived from Beran's diagnostic (dashed). Parameter values: $\alpha = 0.05$, $b = 0.5$, $B = 100$, $R = 999$, $n = 1,000$ (left), $n = 4,000$ (right).

Our test function is

$$\phi(T) = \begin{cases} 1 & \text{if } T > c \\ \gamma & \text{if } T = c \\ 0 & \text{if } T < c \end{cases},$$

where $\gamma \in [0, 1]$ is chosen so that

$$\frac{1}{R}\sum_{j=1}^{R} \mathbb{1}_{\{T_j^* > c\}} + \frac{\gamma}{R}\sum_{j=1}^{R} \mathbb{1}_{\{T_j^* = c\}} = \alpha.$$

Figure 1.2 shows the power functions of the UMPU test and the one derived from Beran's diagnostic. In the latter case, 10,000 Monte-Carlo replications of each test were performed, ensuring a simulation standard error of no more than 0.005 at each point.

We find that Beran's test performs acceptably for small n, but very poorly as n increases. To explain this behaviour, note that if $\mathbb{P}_*$ denotes the conditional probability

of $X^* = (X_1^*, \ldots, X_n^*)$, given $X_1, \ldots, X_n$, then

$$\begin{aligned}
&\mathbb{P}_*(a^* = 0) \\
&\quad = \mathbb{P}_*(T_{n,H}^* = \bar{X}_n^*) = \mathbb{P}_*(|\bar{X}_n^*| > n^{-1/4}) \\
&\quad = \mathbb{P}_*\bigl(n^{1/2}(\bar{X}_n^* - \bar{X}_n) < -n^{1/4} - n^{1/2}\bar{X}_n\bigr) + \mathbb{P}_*\bigl(n^{1/2}(\bar{X}_n^* - \bar{X}_n) > n^{1/4} - n^{1/2}\bar{X}_n\bigr) \\
&\quad = \Phi(-n^{1/4} - n^{1/2}\bar{X}_n) + 1 - \Phi(n^{1/4} - n^{1/2}\bar{X}_n),
\end{aligned}$$

It follows that conditional on $X_1, \ldots, X_n$, we have

$$BT \sim \mathrm{Bin}\bigl(B, \Phi(-n^{1/4} - n^{1/2}\bar{X}_n) + 1 - \Phi(n^{1/4} - n^{1/2}\bar{X}_n)\bigr).$$

Writing $n^{1/2}\bar{X}_n = n^{1/2}\theta + Z$, where $Z \sim N(0,1)$, we see that the (unconditional) power function for the Beran test varies over scale $n^{-1/4}$, in the sense that its value at $\pm n^{-1/4}$ converges to a constant. On the other hand, the UMPU test has power function

$$\begin{aligned}
w(\theta) &= \mathbb{P}_\theta\bigl(n^{1/2}|\bar{X}_n| > \Phi^{-1}(1-\alpha/2)\bigr) \\
&= \Phi\bigl(\Phi^{-1}(\alpha/2) - n^{1/2}\theta\bigr) + 1 - \Phi\bigl(\Phi^{-1}(1-\alpha/2) - n^{1/2}\theta\bigr),
\end{aligned}$$

and so varies over scale $n^{-1/2}$.

1.3.2 The Stein estimator

Now suppose that $X_1, \ldots, X_n$ are independent and identically distributed random vectors in $\mathbb{R}^k$, where $k \geq 4$, and let $X = (X_1, \ldots, X_n)$. Each component of X has a k-variate normal distribution $N_k(\theta, I)$, with mean vector $\theta \in \mathbb{R}^k$ and identity covariance matrix. Write $\bar{X}_n = n^{-1}\sum_{i=1}^n X_i$, define $\mu : \mathbb{R}^k \to \mathbb{R}$ by $\mu(x) = k^{-1}\sum_{i=1}^k x_i$, and let e denote a k-vector of ones. The Stein estimator is defined by

$$T_{n,S}(\bar{X}_n) = \mu(\bar{X}_n)e + \left(1 - \frac{k-3}{n\|\bar{X}_n - \mu(\bar{X}_n)e\|^2}\right)\bigl(\bar{X}_n - \mu(\bar{X}_n)e\bigr).$$

Thus each component of $\bar{X}_n$ is 'shrunk' towards the mean of the nk observations. By contrast with the one-dimensional setting of Section 1.3.1, when $k \geq 4$, the Stein estimator is superefficient for every value of n; its risk less than the Cramér-Rao lower bound, namely k, for all $\theta \in \mathbb{R}^k$. The asymptotic risk is 3 when the components of θ are all equal, and k otherwise (Brandwein and Strawderman, 1990). That the set of points of asymptotic superefficiency is of Lebesgue measure zero does not detract from its practical importance due to its good finite-sample properties. In fact, the behaviour of the Stein estimator in this regular parametric setting is symptomatic of that of superefficient shrinkage estimators employed in more general problems such as kernel density estimation and nonparametric regression. There, the complexity of the parameter space allows far more severe forms of superefficiency (Brown, Low and Zhao, 1997).

We consider estimating the sampling distribution, $H_n(\theta)$, of $n^{1/2}(T_{n,S} - \theta)$, by the bootstrap approximation, $H_n(\bar{X}_n)$. We will see in Section 1.4.2 that $H_n(\bar{X}_n)$ is consistent if and only if the components of θ are not all equal. As in the Hodges example, we may take $T_{n,E} = \bar{X}_n$, so part (b) of Theorem 1.2.6 states that if the components of θ are not all equal then $a^* = n^{1/2}(T^*_{n,S} - \bar{X}^*_n)$ and $d^* = n^{1/2}(\bar{X}^*_n - \bar{X}_n)$ are asymptotically independent in $\mathbb{P}_\theta$-probability with marginal distributions a point mass at the origin and $N_k(0, I)$ respectively. Again, conditional on $X_1, \ldots, X_n$, we have that $X^*_1, \ldots, X^*_n$ are independent and identically distributed $N_k(\bar{X}_n, I)$ random vectors, $\bar{X}^*_n = n^{-1} \sum_{i=1}^n X^*_i$ and $T^*_{n,S} = T_{n,S}(\bar{X}^*_n)$. Note that in applying the diagnostic algorithm to this example, we are forced to choose scalar summaries of the data (cf. Figure 1.1).

As argued in the Hodges example, for the purposes of formal inference we should really be testing $H_0 : \theta_1 = \ldots = \theta_k$ against $H_1 : H_0$ is not true. Considered as a classical hypothesis testing problem, this is very similar to a situation in which one

would use a one-way analysis of variance (ANOVA) test, except that the covariance matrix of each component of X is I rather than $\sigma^2 I$, for some unknown scalar factor σ^2. With $\sigma^2 I$ covariance matrix, the ANOVA test is uniformly most powerful amongst the class of all tests invariant under location, scale and orthogonal transformations, and uniformly most powerful among those tests whose power functions depend on θ only through $\|\theta - \mu(\theta)e\|^2/\sigma^2$ (Lehmann, 1986, Chapters 6 and 7). However, in our situation the test is not invariant under scale transformations, so justification in terms of optimality criteria is lacking. It nevertheless remains a possibility to be considered. In such a test, we would reject H_0 if

$$F = \frac{\frac{1}{k-1} n\|\bar{X}_n - \mu(\bar{X}_n)e\|^2}{\frac{1}{k(n-1)} \sum_{i=1}^n n\|X_i - \bar{X}_n\|^2} > F^{(k-1,k(n-1))}(\alpha),$$

where $F^{(k-1,k(n-1))}(\alpha)$ is the upper α-point of the $F^{(k-1,k(n-1))}$ distribution. Note that the distribution of F under $\mathbb{P}_\theta$ is the same as that of

$$\frac{Y_1/(k-1)}{Y_2/\big(k(n-1)\big)},$$

where Y_1 has a non-central chi-squared distribution with $k-1$ degrees of freedom and non-centrality parameter $\lambda = n\|\theta - \mu(\theta)e\|^2$, and is independent of $Y_2 \sim \chi^2_{k(n-1)}$.

A natural alternative to the ANOVA test is a generalised likelihood ratio test. The maximum likelihood estimator of θ is $\mu(\bar{X}_n)e$ under the null hypothesis and $\bar{X}_n$ under the alternative hypothesis. Thus the generalised likelihood ratio is given by

$$\begin{aligned} L_X(H_0, H_1) &= \frac{\sup_{\theta \in \mathbb{R}^k} \exp\big(-\frac{1}{2}\sum_{i=1}^n \|X_i - \theta\|^2\big)}{\sup_{\theta_1 = \ldots = \theta_k} \exp\big(-\frac{1}{2}\sum_{i=1}^n \|X_i - \theta\|^2\big)} \\ &= \frac{\exp\big(-\frac{1}{2}\sum_{i=1}^n \|X_i - \bar{X}_n\|^2\big)}{\exp\big(-\frac{1}{2}\sum_{i=1}^n \|X_i - \mu(\bar{X}_n)e\|^2\big)} \\ &= \exp\big(n\|\bar{X}_n - \mu(\bar{X}_n)e\|^2/2\big), \end{aligned}$$

so we would reject H_0 if $n\|\bar{X}_n - \mu(\bar{X}_n)e\|^2 > \chi^2_{k-1}(\alpha)$, where $\chi^2_{k-1}(\alpha)$ is the upper α-point of the χ^2_{k-1} distribution. Justification for using this test can be expressed

in terms of the *shortcoming* of the test and its *Bahadur deficiency.* If we write $\Theta_1 = \mathbb{R}^k \setminus \{\theta : \theta_1 = \ldots = \theta_k\}$, the shortcoming of a test is defined, for each $\theta \in \Theta_1$, as the difference in power between the test in question and the most powerful test of the same size. Theorem 3.6.1 of Kallenberg (1978) states that the shortcoming of the likelihood ratio test based on sample of size n tends to zero, uniformly for $\theta \in \Theta_1$, as $n \to \infty$.

To describe Bahadur deficiency, let $N(\alpha, \beta, \theta)$ denote the number of observations needed for the likelihood ratio test of size α to achieve power β at $\theta \in \Theta_1$ and let $N^+(\alpha, \beta, \theta)$ denote the minimum of $N(\alpha, \beta, \theta)$ of over all size α tests. Then Corollary 5.3.6 of Kallenberg (1978) gives that, for each $\beta \in (0, 1)$ and $\theta \in \Theta_1$, there exists $A = A(\beta, \theta)$ such that

$$\limsup_{\alpha \to 0} \frac{N(\alpha, \beta, \theta) - N^+(\alpha, \beta, \theta)}{\log N^+(\alpha, \beta, \theta)} \leq A.$$

In this sense, the likelihood ratio test is Bahadur deficient of order $O\big(\log N^+(\alpha, \beta, \theta)\big)$ as $\alpha \to 0$.

We implement Beran's ideas as follows: given $X_1, \ldots, X_n$, construct the statistic $T = \sum_{i=1}^{B} \|a_i^*\| \|d_i^*\|$ after following steps (1)–(3) of the algorithm given at the end of Section 1.2. This can be compared with independently generated values of $T_1^*, \ldots, T_R^*$, where each T_j^*, for $j = 1, \ldots, R$, is the value of T when the original data are drawn from $N_k(0, I)$. Under the alternative hypothesis, we expect the value of the test statistic to be reduced. There is no need to consider randomised tests in this case. The proposition below validates the plotting of the power function of this test as a function of $\lambda = n\|\theta - \mu(\theta)e\|^2$.

Proposition 1.3.1. *When X has distribution $\mathbb{P}_\theta$ and X^* has distribution $\mathbb{P}_*$, the sampling distribution of T depends on θ only through $\lambda = n\|\theta - \mu(\theta)e\|^2$.*

Figure 1.3 suggests that the power function of the Beran test is uniformly smaller

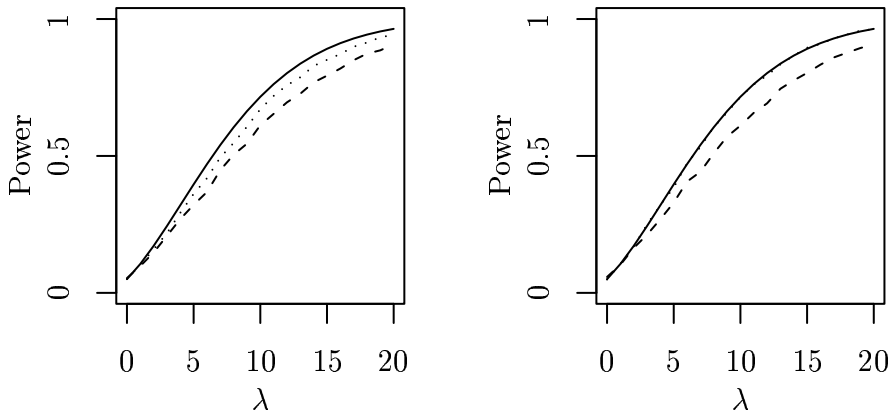


Figure 1.3: The power functions of the likelihood ratio test (solid), the ANOVA test (dotted), and Beran's method test (dashed) of $H_0 : \lambda = 0$ and $H_1 : \lambda > 0$. Parameter values: $n = 10$ (left), $n = 1000$ (right), $\alpha = 0.05$, $k = 5$, $B = 100$, $R = 999$; 10,000 Monte-Carlo repetitions at each value of λ. The dots are almost indistinguishable from the solid line on the right-hand plot.

than both the generalised likelihood ratio test and the ANOVA test for both small and large n. The ANOVA test is a little worse than the generalised likelihood ratio test for small n and as good for large n. This is unsurprising as the ANOVA test is analogous to using a t-test for a normal mean when the population standard deviation is known, while the likelihood ratio test is akin to the more standard z-test.

1.4 Restoring consistency to the bootstrap

1.4.1 The Hodges estimator

It was mentioned in Section 1.3.1 that when estimating the distribution $H_n(\theta)$ of $n^{1/2}(T_{n,H} - \theta)$, the parametric bootstrap distribution $H_n(\bar{X}_n)$ is consistent when $\theta \neq 0$

but inconsistent when $\theta = 0$. The bootstrap fails despite the fact that $H_n(\theta)$ converges pointwise for all $\theta \in \mathbb{R}$, with a limiting distribution $H(\theta)$ which is $N(0,1)$ when $\theta \neq 0$ and $N(0, b^2)$ when $\theta = 0$. To explain this behaviour, note that provided $b \neq 0$, we can compute the cumulative distribution function, $H_n(x, \theta)$ corresponding to the distribution $H_n(\theta)$ as follows:

$$
\begin{aligned}
&H_n(x,\theta) \\
&\quad = \mathbb{P}_\theta\big(n^{1/2}(T_{n,H} - \theta) \le x\big) = \mathbb{P}_\theta(T_{n,H} \le n^{-1/2}x + \theta) \\
&\quad = \mathbb{P}_\theta(\bar{X}_n \le n^{-1/2}x + \theta, |\bar{X}_n| > n^{-1/4}) + \mathbb{P}_\theta(b\bar{X}_n \le n^{-1/2}x + \theta, |\bar{X}_n| \le n^{-1/4}) \\
&\quad = \mathbb{P}_\theta\big\{\bar{X}_n \le \big(-n^{-1/4} \wedge (n^{-1/2}x + \theta)\big)\big\} \\
&\qquad + \mathbb{P}_\theta\big\{-n^{-1/4} < \bar{X}_n < \big(n^{-1/4} \wedge b^{-1}(n^{-1/2}x + \theta)\big)\big\} \\
&\qquad + \mathbb{P}_\theta(n^{-1/4} \le \bar{X}_n \le n^{-1/2}x + \theta) \\
&\quad = \mathbb{P}_\theta\big(n^{1/2}(\bar{X}_n - \theta) \le (-n^{1/4} - n^{1/2}\theta) \wedge x\big) \\
&\qquad + \mathbb{P}_\theta\big\{-n^{1/4} - n^{1/2}\theta < n^{1/2}(\bar{X}_n - \theta) < (n^{1/4} - n^{1/2}\theta) \wedge b^{-1}\big(x + (1-b)\theta n^{1/2}\big)\big\} \\
&\qquad + \mathbb{P}_\theta\big(n^{1/4} - n^{1/2}\theta \le n^{1/2}(\bar{X}_n - \theta) \le x\big).
\end{aligned}
$$

Thus

$$
H_n(x,\theta) = \begin{cases}
\Phi(x) & \text{if } x < -n^{1/4} - n^{1/2}\theta \\
\Phi(-n^{1/4} - n^{1/2}\theta) & \text{if } -n^{1/4} - n^{1/2}\theta \le x < -bn^{1/4} - n^{1/2}\theta \\
\Phi\big\{b^{-1}\big(x + (1-b)\theta n^{1/2}\big)\big\} & \text{if } -bn^{1/4} - n^{1/2}\theta \le x < bn^{1/4} - n^{1/2}\theta \\
\Phi(n^{1/4} - n^{1/2}\theta) & \text{if } bn^{1/4} - n^{1/2}\theta \le x < n^{1/4} - n^{1/2}\theta \\
\Phi(x) & \text{if } x \ge n^{1/4} - n^{1/2}\theta.
\end{cases} \tag{1.2}
$$

Under $\mathbb{P}_{\theta_0}$, we have that $n^{1/2}(\bar{X}_n - \theta_0)$ has a standard normal distribution for every n (so in particular the limit distribution has full support). It follows from Theorem 1.2.6 that $H_n(\bar{X}_n)$ will be a consistent estimator of $H_n(\theta_0)$ if and only if the sequence $(T_{n,H})$ is locally asymptotically equivariant at θ_0. The proof of the following proposition is similar to an argument in Putter and van Zwet (1996), and is given in Section 1.5:

Proposition 1.4.1. *The sequence $(T_{n,H})$ is locally asymptotically equivariant at θ_0 if and only if $\theta_0 \neq 0$.*

Remark: When $\theta_0 = 0$, Theorem 2.3 of Beran (1997) shows that $H_n(\bar{X}_n)$ converges in distribution, as a random element of the space of probability measures on the real line metrised by weak convergence, to the random probability measure $N\big((b-1)Z, b^2\big)$, where $Z \sim N(0,1)$.

The inconsistency of the standard n out of n bootstrap at the origin leads us to consider an m out of n parametric bootstrap, $H_m(\bar{X}_n)$, where $m \to \infty$ as $n \to \infty$, but $m = o(n)$. The rationale is as follows: since $H_n(\theta) \xrightarrow{d} H(\theta)$ as $n \to \infty$ for all $\theta \in \mathbb{R}$, consistent estimation of $H(\theta)$ and $H_n(\theta)$, or indeed $H_m(\theta)$, amount to the same thing. Thus the m out of n bootstrap may be thought of as an attempt to estimate $H_m(\theta)$ with the advantage that the parameter of the resampling distribution, $\bar{X}_n$, is likely to be closer to the true parameter θ than is $\bar{X}_m$. Indeed, as a consequence of Corollary 2.1(b) of Beran (1997), $H_m(\bar{X}_n)$ is consistent for all $\theta \in \mathbb{R}$ provided m tends to infinity slowly enough that $m = o(n)$.

In Figure 1.4, we present a comparison of the errors in the bootstrap approximations $H_m(\bar{X}_n)$ as estimators of $H_n(\theta)$ for $m = n^{1/2}$, $m = n^{3/4}$ and $m = n$. These values of m are understood to be rounded to the nearest integer. We compare $H_m(\bar{X}_n)$ and $H_n(\theta)$ using the supremum metric, d, on the corresponding distribution functions, $H_m(x, \bar{X}_n)$ and $H_n(x, \theta)$, so that

$$d\big(H_m(\bar{X}_n), H_n(\theta)\big) = \sup_{x \in \mathbb{R}} \big| H_m(x, \bar{X}_n) - H_n(x, \theta) \big|. \tag{1.3}$$

This distance metrises convergence in distribution, by Pólya's theorem (van der Vaart, 1998, p. 12), and has the advantage of being considerably easier to compute in practice than other equivalent distances, such as the Lévy metric (Billingsley, 1995, p. 198).

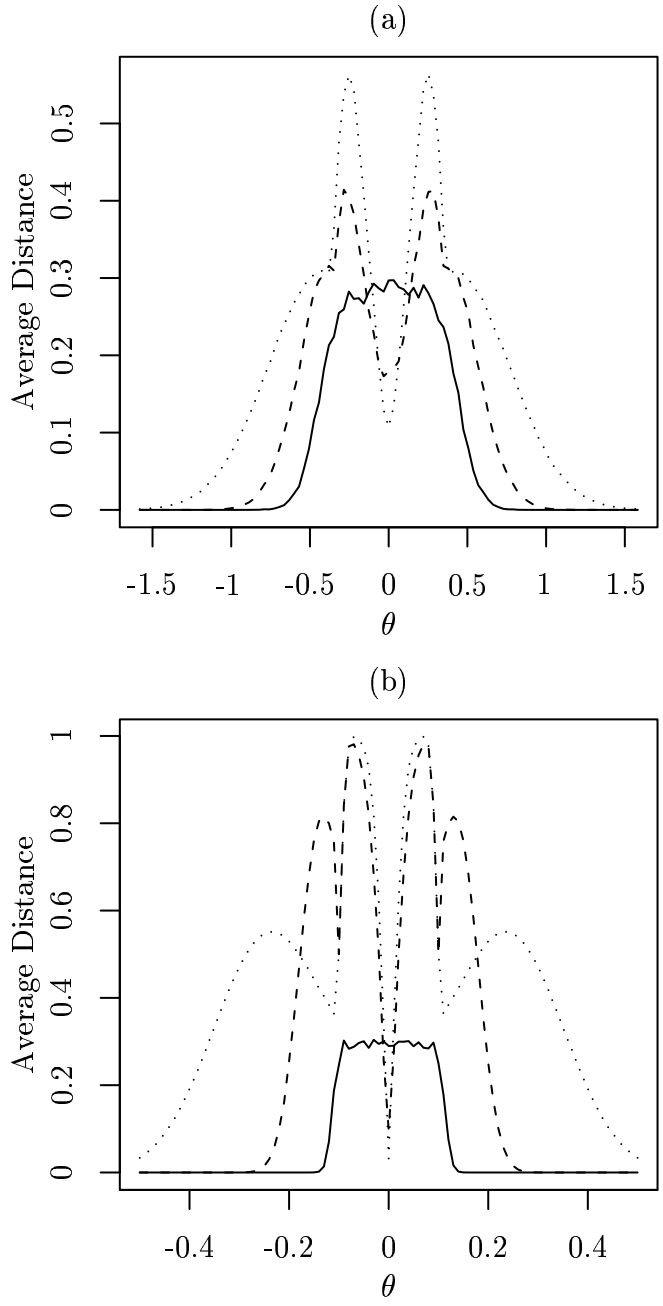


Figure 1.4: The distances $d\big(H_m(\bar{X}_n), H_n(\theta)\big)$, averaged over 1000 realisations of $\bar{X}_n$, with $m = n^{1/2}$ (dotted), $m = n^{3/4}$ (dashed), $m = n$ (solid). Parameter values: $b = 0.5$, (a) $n = 100$, (b) $n = 10,000$.

It is particularly interesting to note that, although smaller choices of m do improve the bootstrap performance in a very small neighbourhood of $\theta = 0$, the improvements come at the expense of considerably worse performance outside this neighbourhood. Treated as a problem in decision theory, the minimax rule appears to be to choose $m = n$, and this would agree with the Bayes rule unless most of the mass of the prior were concentrated in a very small neighbourhood of $\theta = 0$.

We give here a heuristic explanation for the results observed. Write

$$m^{1/2}\bar{X}_n = m^{1/2}\theta + m^{1/2}n^{-1/2}Z,$$

where $Z \sim N(0,1)$. From (1.2), we see that the magnitude of the error in the bootstrap approximation depends on the absolute value of the difference between $n^{1/2}\theta$ and $m^{1/2}\theta + m^{1/2}n^{-1/2}Z$. If $|\theta| \ll n^{-1/2}$, then the random term in the error, $m^{1/2}n^{-1/2}Z$, dominates. The variance of this term increases as m increases relative to n, although it always has zero expectation. However, for larger values of $|\theta|$ the difference between the two non-random terms, $m^{1/2}\theta$ and $n^{1/2}\theta$, is crucial. This is large in absolute value for small m relative to n, and decreases to zero as m increases to n.

We now investigate whether or not it is possible to retain the desirable characteristics of both methods by means of an empirical, data-driven choice of m. That is, if we let $m = f_n(|\bar{X}_n|)$, where $f_n : [0, \infty) \to \{1, \ldots, n\}$ is some suitably chosen non-decreasing function, can we achieve improved performance in a neighbourhood of $\theta = 0$ without loss elsewhere in the parameter space?

A simple class of possible choices of m is given by

$$m = \begin{cases} An^{\alpha} & \text{if } |\bar{X}_n| \leq Cn^{-\beta} \\ n & \text{if } |\bar{X}_n| > Cn^{-\beta}, \end{cases} \tag{1.4}$$

where $A, C > 0$, $\alpha \in (0,1)$ and $\beta \in (0, 1/2)$. Let $\mathcal{M}$ denote this class.

Proposition 1.4.2. *For any $m \in \mathcal{M}$ and any $\theta \in \mathbb{R}$, we have that $H_m(\bar{X}_n)$ is a consistent estimator of $H_n(\theta)$.*

Numerical studies suggest, however, that while improved performance in a small neighbourhood of $\theta = 0$ can be achieved, again this comes at the expense of worse performance outside this neighbourhood. Although the 'bad' neighbourhoods vanish in the limit as n tends to infinity, which ensures consistency, they remain a problem in finite samples. The problem occurs in the region, in this case where $|\theta| \approx Cn^{-\beta}$, in which the event $\{|\bar{X}_n| \leq Cn^{-\beta}\}$ has moderate probability. Considered as an attempt to estimate the optimal value $m_{\text{opt}} = m_{\text{opt}}(\theta)$, the rule in (1.4) is analogous to using the Hodges estimator as an estimator of θ, and suffers the same drawbacks. Of course, other more complicated empirical choices of m are possible, but the scope for improvement over the naive n out of n bootstrap appears small.

A further suggestion for restoring consistency, proposed by Putter and van Zwet (1996), involves a refined choice of parameter estimate in the bootstrap approximation: we replace $H_n(\bar{X}_n)$ by $H_n(\hat{\theta}_n)$ where $\hat{\theta}_n$ is chosen so that

(i) $\mathbb{P}_{\theta=0}(\hat{\theta}_n = 0) \to 1$ as $n \to \infty$;

(ii) $\mathbb{P}_{\theta\neq 0}(\hat{\theta}_n \neq 0) \to 1$ as $n \to \infty$.

The consistency of $H_n(\hat{\theta}_n)$ then follows from Corollary 1.1 of Putter and van Zwet (1996). The authors themselves suggest an estimator from the following class:

$$\hat{\theta}_n = \begin{cases} 0 & \text{if } |\bar{X}_n| \leq Cn^{-\beta} \\ \bar{X}_n & \text{if } |\bar{X}_n| > Cn^{-\beta}, \end{cases}$$

where $C > 0$ and $\beta \in (0, 1/2)$. Note that, when $C = 1$ and $\beta = 1/4$, this is the Hodges estimator with $b = 0$. Once again, however, the improvements in the immediate

vicinity of $\theta = 0$ are offset by severe losses elsewhere in the parameter space. For large n, comparing the expression for $H_n(x, \theta)$ in (1.2) with the corresponding expression for $H_n(x, \hat{\theta}_n)$, we see that, when $\theta \in (n^{-1/2}, Cn^{-\beta})$, it is likely that $n^{1/2}\hat{\theta}_n$ will be zero, whereas $n^{1/2}\theta$ may be large. Thus $H_n(x, \hat{\theta}_n)$ will be a poor estimator of $H_n(x, \theta)$ in this region of the parameter space.

1.4.2 The Stein estimator

Corollary 2.1(b) of Beran (1997) also applies to the Stein estimator, and gives that $H_m(\bar{X}_n)$ is consistent for $H_n(\theta)$ for all $\theta \in \mathbb{R}^k$, provided that $m = o(n)$, as before. By Pólya's theorem, we may still compare bootstrap approximations to $H_n(\theta)$ using the supremum distance on the corresponding distribution functions, and we continue to denote this metric by d. As explicit distribution functions are not available in this instance, comparisons must be based on the respective empirical distribution functions. The algorithm is as follows:

(i) Choose $B \in \mathbb{N}$ and repeat steps (ii) to (iv) for $i = 1, \ldots, B$.

(ii) Generate independent $\bar{X}_{n,1}, \ldots, \bar{X}_{n,R} \sim N_k(\theta, n^{-1}I)$ to compute $\hat{H}_{n,R}(\theta)$, the empirical distribution of $n^{1/2}\big(T_{n,S}(\bar{X}_{n,1}) - \theta\big), \ldots, n^{1/2}\big(T_{n,S}(\bar{X}_{n,R}) - \theta\big)$.

(iii) Generate independent $\bar{X}^*_{m,1}, \ldots, \bar{X}^*_{m,R}$, where, conditional on $\bar{X}_{n,j}$, we have $\bar{X}^*_{m,j} \sim N_k(\bar{X}_{n,j}, m^{-1}I)$ for $j = 1, \ldots, R$. Compute $\hat{H}_{m,R}(\bar{X}_n)$, the empirical distribution of $m^{1/2}\big(T_{m,S}(\bar{X}^*_{m,1}) - \bar{X}_{n,1}\big), \ldots, m^{1/2}\big(T_{m,S}(\bar{X}^*_{m,R}) - \bar{X}_{n,R}\big)$.

(iv) Compute

$$d_i = d\big(\hat{H}_{m,R}(\bar{X}_n), \hat{H}_{n,R}(\theta)\big).$$

(v) Compute $\bar{d} = B^{-1}\sum_{i=1}^{B} d_i$.

In Figure 1.5, we plot $\bar{d}$ as a function of $\lambda = n\|\theta - \mu(\theta)e\|^2$, for $m = n^{1/2}$, $m = n^{3/4}$ and $m = n$. Numerical studies show no qualitative change for different θ-directions.

As in the Hodges example, we find that improvements in a small neighbourhood of $\lambda = 0$ are possible, but that there is still a price to be paid in terms of poor performance for larger values of λ. A minimax approach to selecting m would suggest choosing $m = o(n)$ (perhaps $m = n^{3/4}$), whereas adopting a Bayesian decision principle would lead to the choice $m = n$ unless most of the mass of the prior distribution were concentrated on a small neighbourhood of $\lambda = 0$. The n out of n bootstrap performs better relative to the alternatives as n increases. Incidentally, when two samples of size $R = 500$ were drawn independently from $N_k(0, 1)$, the average over $B = 200$ realisations of the supremum distance between the empirical distribution functions was 0.059. Figure 1.5 therefore suggests that $H_n(\bar{X}_n)$ is a very good approximation to $H_n(\theta)$ for $\lambda \geq 10$.

To explain these observations, let Z, Z' denote independent standard k-variate normal random variables, and let $T^*_{m,S} = T_{m,S}(\bar{X}^*_n)$. Now, under $\mathbb{P}_\theta$,

$$\begin{aligned} n^{1/2}(T_{n,S} - \theta) &= n^{1/2}(\bar{X}_n - \theta) - \frac{(k-3)\, n^{1/2}\big(\bar{X}_n - \mu(\bar{X}_n)e\big)}{\|n^{1/2}(\bar{X}_n - \mu(\bar{X}_n)e)\|^2} \\ &\sim Z - \frac{(k-3)\big\{Z - \mu(Z)e + n^{1/2}\big(\theta - \mu(\theta)e\big)\big\}}{\|Z - \mu(Z)e + n^{1/2}\big(\theta - \mu(\theta)e\big)\|^2} \end{aligned} \tag{1.5}$$

and, under $\mathbb{P}_*$,

$$\begin{aligned} &m^{1/2}(T^*_{m,S} - \bar{X}_n) \\ &\quad = m^{1/2}(\bar{X}^*_m - \bar{X}_n) - \frac{(k-3)\, m^{1/2}\big(\bar{X}^*_m - \mu(\bar{X}^*_m)e\big)}{\|m^{1/2}(\bar{X}^*_m - \mu(\bar{X}^*_m)e)\|^2} \\ &\quad \sim Z' - \frac{(k-3)\big\{Z' - \mu(Z')e + m^{1/2}\big(Z - \mu(Z)e\big)/n^{1/2} + m^{1/2}\big(\theta - \mu(\theta)e\big)\big\}}{\big\|Z' - \mu(Z')e + m^{1/2}\big(Z - \mu(Z)e\big)/n^{1/2} + m^{1/2}\big(\theta - \mu(\theta)e\big)\big\|^2}. \end{aligned} \tag{1.6}$$

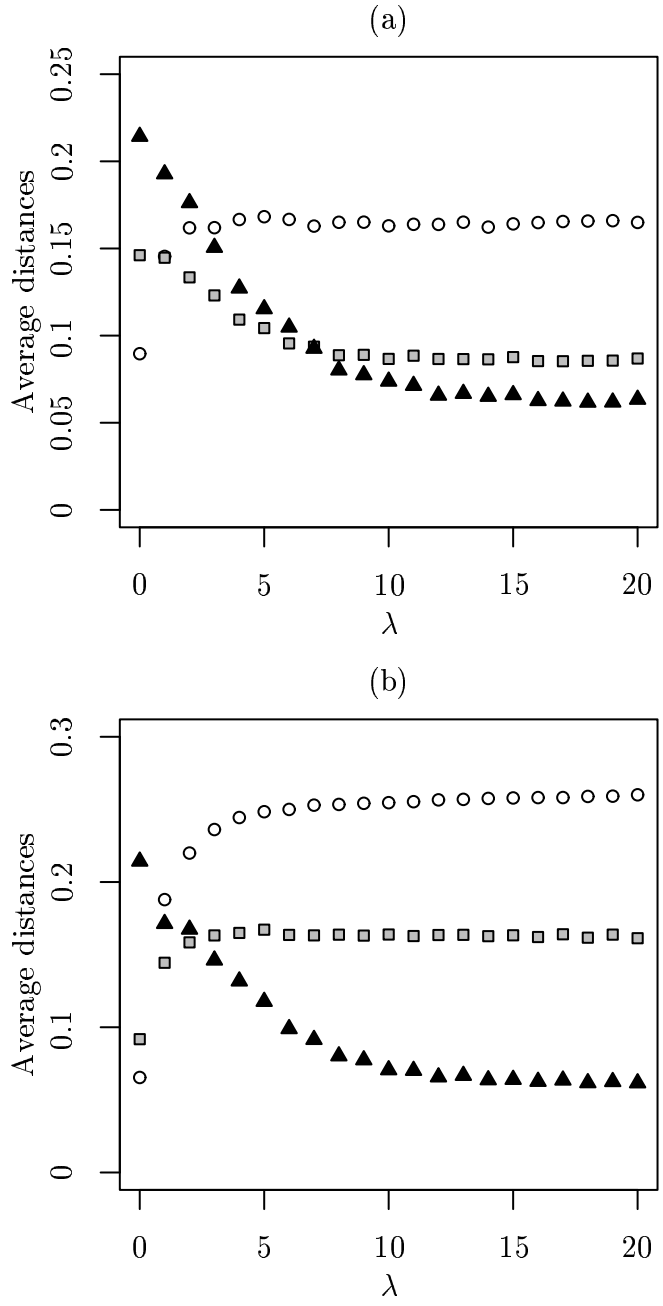


Figure 1.5: The average distances $d\big(\hat{H}_{m,R}(\bar{X}_n), \hat{H}_{n,R}(\theta)\big)$, with $m = n^{1/2}$ (circles), $m = n^{3/4}$ (grey squares), $m = n$ (black triangles). Parameter values: $R = 500$, $B = 200$, $k = 5$, $n^{1/2}\theta = (\lambda/2)^{1/2}(-1, 1, 0, 0, 0)$, (a) $n = 100$, (b) $n = 10,000$.

Comparing (1.5) and (1.6), we see that $H_m(\bar{X}_n)$ will be a good approximation to $H_n(\theta)$ when $m^{1/2}\big(Z - \mu(Z)e\big)/n^{1/2} + m^{1/2}\big(\theta - \mu(\theta)e\big)$ is close to $n^{1/2}\big(\theta - \mu(\theta)e\big)$. When λ is small, and in particular when $\lambda = 0$, the main error is likely to come from the random term, $m^{1/2}\big(Z - \mu(Z)e\big)/n^{1/2}$. Since we can choose m to tend to infinity as slowly as we like, we can make this error as small as we like, in probability. However, when λ is large, it is the difference between the two non-random terms, $m^{1/2}\big(\theta - \mu(\theta)e\big)$ and $n^{1/2}\big(\theta - \mu(\theta)e\big)$, which dominates. This decreases to zero as m increases towards n.

Note from (1.5) and (1.6) that when the components of θ are not all equal, $H_n(\theta)$ converges weakly to $N_k(0, I)$ and $H_n(\bar{X}_n)$ converges weakly in $\mathbb{P}_\theta$-probability to $N_k(0, I)$ also. This explains the fact that the bootstrap distribution is consistent in this instance. However, when the components of θ are equal, $H_n(\theta)$ converges to the the probability measure $\pi(0)$, where for any $h \in \mathbb{R}^k$, we define $\pi(h)$ to be the distribution of

$$Z - \frac{(k-3)\big(Z - \mu(Z)e + h - \mu(h)e\big)}{\|Z - \mu(Z)e + h - \mu(h)e\|^2},$$

with $Z \sim N_k(0, I)$. On the other hand, Theorem 2.3 of Beran (1997) shows that $H_n(\bar{X}_n)$ converges weakly, as a random element of the space of probability distributions on $\mathbb{R}^k$ metrised by weak convergence, to the random probability measure $\pi(Z')$, where $Z' \sim N_k(0, I)$ and is independent of Z.

Analogues of the empirical rules for choosing m and the Putter and van Zwet method of restoring consistency also exist for this problem. For instance, the latter may be implemented with

$$\hat{\theta}_n = \begin{cases} \mu(\bar{X}_n)e & \text{if } \|\bar{X}_n - \mu(\bar{X}_n)e\| \le Cn^{-\beta} \\ \bar{X}_n & \text{if } \|\bar{X}_n - \mu(\bar{X}_n)e\| > Cn^{-\beta}, \end{cases}$$

where $C > 0$ and $\beta \in (0, 1/2)$, in which case the resulting bootstrap approximation $H_n(\hat{\theta}_n)$ is a consistent estimator of $H_n(\theta)$ for all $\theta \in \mathbb{R}^k$, again by Corollary 1.1 of

		λ					
		0	1	2	5	10	20
$n = 100$	$H_n(\bar{X}_n)$	0.214	0.193	0.176	0.115	0.074	0.063
	$H_n(T_{n,S})$	0.141	0.160	0.153	0.108	0.071	0.061
$n = 10000$	$H_n(\bar{X}_n)$	0.214	0.171	0.167	0.118	0.071	0.062
	$H_n(T_{n,S})$	0.141	0.164	0.157	0.109	0.071	0.062

Table 1.1: The distances $d\big(H_n(\bar{X}_n), H_n(\theta)\big)$ and $d\big(H_n(T_{n,S}), H_n(\theta)\big)$. Parameter values: $R = 500$, $B = 200$, $k = 5$.

Putter and van Zwet (1996). Although numerical studies suggest it is possible to achieve minor improvements for a fixed n with a suitable choice of C, any choice of C will eventually be poor for sufficiently large n, because $n\|\bar{X}_n - \mu(\bar{X}_n)e\|^2$ has a non-central chi-squared distribution with $(k-1)$ degrees of freedom and non-centrality parameter λ, so the event $\{\|\bar{X}_n - \mu(\bar{X}_n)e\| \leq Cn^{-\beta}\}$ has moderate probability when $\lambda \approx C^2 n^{1-2\beta}$. Thus the event $\{\hat{\theta}_n = \mu(\bar{X}_n)e\}$ is eventually probable, even for large λ, and $H_n(\hat{\theta}_n)$ will then perform poorly. Similar remarks apply to empirical choices of m in the m out of n bootstrap.

In fact, it is another inconsistent alternative bootstrap distribution, $H_n(T_{n,S})$, which seems to come closest to improving the poor performance of $H_n(\bar{X}_n)$ near $\lambda = 0$ while retaining the good performance elsewhere in the parameter space (cf. Table 1.1). Applying Theorem 2.3 of Beran (1997) again, the random limiting distribution of $H_n(T_{n,S})$ when the components of θ are all equal is $\pi(V)$, where $V \sim \pi(0)$. Since we can construct V by shrinking $Z \sim N_k(0, I)$ towards $\mu(Z)e$, we expect that $\pi(V)$ will be closer to $\pi(0) = \pi\big(\mu(Z)e\big)$ than is $\pi(Z)$. This argument breaks down if $\|Z - \mu(Z)e\|$ is so small that the shrinkage factor is negative and large in modulus. However, this is a rare event, which has overall little effect.

1.5 Appendix

Proof of **Proposition 1.3.1**.

Recall that $T = \sum_{i=1}^{B} \|a_i^*\|\|d_i^*\|$ is a sum of B independent and identically distributed random variables, so it suffices to show the result for $\|a^*\|\|d^*\|$. Observe that

$$\begin{aligned} \|a^*\|\|d^*\| &= \frac{(k-3)\|n^{1/2}(\bar{X}_n^* - \bar{X}_n)\|}{\|n^{1/2}(\bar{X}_n^* - \mu(\bar{X}_n^*)e)\|} \\ &\sim \frac{(k-3)\|Z'\|}{\|Z' - \mu(Z')e + Z - \mu(Z)e + n^{1/2}(\theta - \mu(\theta)e)\|}, \end{aligned} \tag{1.7}$$

where Z, Z' are independent standard normal random variables on $\mathbb{R}^k$. The idea of the proof is to find the set of transformations of $\theta \in \mathbb{R}^k$ which preserve $\|\theta - \mu(\theta)e\|$, and show that the distribution of the random variable above is invariant under such transformations.

For $d \geq 0$, we seek to characterise the set $B_d = \{\theta \in \mathbb{R}^k : \|\theta - \mu(\theta)e\| = d\}$. Geometrically, we can consider $\theta - \mu(\theta)e$ as the orthogonal projection of θ onto the $(k-1)$-dimensional subspace $S = \{x \in \mathbb{R}^k : \langle x, e\rangle = 0\}$. (Here, and throughout, $\langle \cdot, \cdot \rangle$ denotes the Euclidean inner product.) Since $\theta \in S$ is in B_d if and only if $\|\theta\| = d$, it follows that B_d is a hyper-cylinder in $\mathbb{R}^k$, with axis along e (cf. Figure 1.6). Thus if $\theta, \theta' \in B_d$, we can write

$$\theta' = P(\theta - \mu(\theta)e) + \mu(\theta')e,$$

where P is a $k \times k$ orthogonal matrix mapping S into itself.

Note that if e is an eigenvector of P with eigenvalue 1, and $\theta \in S$, then

$$\langle P\theta, e\rangle = \langle \theta, P^T e\rangle = \langle \theta, e\rangle = 0,$$

so P maps S into itself. Now suppose $\theta, \theta' \in B_d \cap S$. We show that there exists an orthogonal matrix with eigenvalue 1 and corresponding eigenvector e which maps θ to θ'.

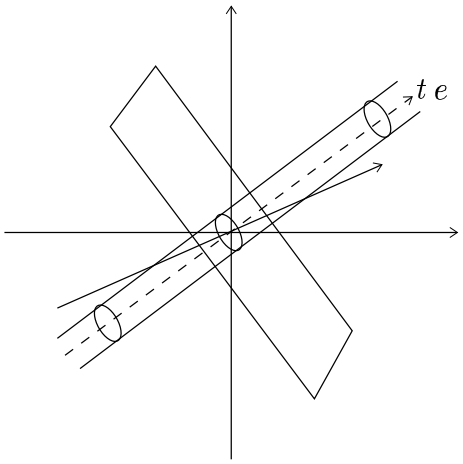


Figure 1.6: Diagram showing the set S and the hyper-cylinder B_d, which has axis along e.

Choose an orthogonal change of basis matrix A such that $Ae/k^{1/2} = (0, 0, \ldots, 0, 1)^T$. Then

$$\langle A\theta, Ae\rangle = \langle \theta, A^T Ae\rangle = \langle \theta, e\rangle = 0,$$

and similarly $\langle A\theta', Ae\rangle = 0$, so we can find a $(k-1) \times (k-1)$ orthogonal matrix B such that

$$\begin{pmatrix} & & & 0 \\ & B & & \vdots \\ & & & 0 \\ 0 & \cdots & 0 & 1 \end{pmatrix} A\theta = A\theta'.$$

Hence, if C denotes the $k \times k$ matrix obtained by extending B as above, then $A^T CA$ is orthogonal and $A^T CA\theta = \theta'$. Moreover, e is an eigenvector of $A^T CA$ with eigenvalue 1.

We see from (1.7) that adding $t\,e$ to θ, for some $t \in \mathbb{R}$, does not change the distribution of $\|a^*\|\|d^*\|$. Thus it suffices to show that, for $\theta \in B_d \cap S$, the distribution of $\|a^*\|\|d^*\|$ is the same when X has distribution $\mathbb{P}_\theta$ as when X has distribution $\mathbb{P}_{P\theta}$, provided

that P is orthogonal and $Pe = e$. Noting that $\mu(\theta) = 0$ for $\theta \in S$, we have

$$\frac{\|Z'\|}{\|Z' - \mu(Z')e + Z - \mu(Z)e + n^{1/2}P\theta\|} = \frac{\|P^T Z'\|}{\|P^T\big(Z' - \mu(Z')e + Z - \mu(Z)e\big) + n^{1/2}\theta\|}.$$

Now $Z' - \mu(Z')e \sim N_k(0, \Sigma)$, where $\Sigma = I - ee^T/k$, and it therefore follows that $P^T(Z' - \mu(Z')e) \sim N_k(0, P^T\Sigma P)$. But

$$P^T\Sigma P = P^T\Big(I - \frac{1}{k}ee^T\Big)P = P^TP - \frac{1}{k}(P^Te)(P^Te)^T = I - \frac{1}{k}ee^T.$$

Similarly, $P^T(Z - \mu(Z)e) \sim N_k(0, \Sigma)$, and the result follows. □

Proof of **Proposition 1.4.1**.

Suppose $\theta_0 \neq 0$, and let (θ_n) be *any* sequence converging to θ_0. We assume that $\theta_0 > 0$, as the other case is very similar. From (1.2) we see that $H_n(x, \theta_0) \to \Phi(x)$ as $n \to \infty$ for all $x \in \mathbb{R}$, so the result will follow if we show that $H_n(x, \theta_n) \to \Phi(x)$ as $n \to \infty$ for all $x \in \mathbb{R}$.

Given $\epsilon > 0$ with $\epsilon < \theta_0$, there exists $n_0 \in \mathbb{N}$ such that $|\theta_n - \theta_0| < \epsilon$ for all $n \geq n_0$. Moreover, there exists $n_1 \in \mathbb{N}$ such that

$$\Phi\big(n^{1/4} - n^{1/2}(\theta_0 - \epsilon)\big) \leq \epsilon/2$$

for all $n \geq n_1$. Observe from (1.2) that for $n \geq n_0$, $H_n(x, \theta_n)$ and $\Phi(x)$ agree on the interval $[n^{1/4} - n^{1/2}(\theta_0 - \epsilon), \infty)$. Thus, for $n \geq \max(n_0, n_1)$,

$$\begin{aligned} |H_n(x, \theta_n) - \Phi(x)| &\leq \sup_{x \leq n^{1/4} - n^{1/2}(\theta_0 - \epsilon)} |H_n(x, \theta_n) - \Phi(x)| \\ &\leq 2\Phi\big(n^{1/4} - n^{1/2}(\theta_0 - \epsilon)\big) \\ &\leq \epsilon. \end{aligned}$$

Conversely, if $\theta_0 = 0$, then $H_n(x, \theta_0) \to \Phi(b^{-1}x)$ as $n \to \infty$ for all $x \in \mathbb{R}$. Suppose that (θ_n) is a sequence such that for some non-zero $h \in \mathbb{R}$ and some sequence (h_n) converging to h, we can write $\theta_n = n^{-1/2}h_n$. Again from (1.2), we see that $H_n(x, \theta_n)$

and $\Phi\{b^{-1}(x+(1-b)\theta_n n^{1/2})\}$ agree on the interval $(-bn^{1/4}-h_n, bn^{1/4}-h_n)$. Since both are distribution functions, it follows that given $\delta>0$, there exists $n_0\in\mathbb{N}$ such that

$$\sup_{x\in\mathbb{R}}\left|H_n(x,\theta_n)-\Phi\{b^{-1}(x+(1-b)\theta_n n^{1/2})\}\right|\le\delta$$

for all $n\ge n_0$. Moreover, since $\Phi(\cdot)$ is uniformly continuous, there exists $n_1\in\mathbb{N}$ such that

$$\sup_{x\in\mathbb{R}}\left|\Phi\{b^{-1}(x+(1-b)\theta_n n^{1/2})\}-\Phi\{b^{-1}(x+(1-b)h)\}\right|\le\delta$$

for $n\ge n_1$. But then, for all $n\ge\max(n_0,n_1)$,

$$\begin{aligned}\sup_{x\in\mathbb{R}}|H_n(x,\theta_n)-\Phi(b^{-1}x)| &\ge \sup_{x\in\mathbb{R}}\left|\Phi\{b^{-1}(x+(1-b)h)\}-\Phi(b^{-1}x)\right|-2\delta\\ &= \left|\Phi\Big(\frac{(1-b)h}{2b}\Big)-\Phi\Big(\frac{-(1-b)h}{2b}\Big)\right|-2\delta,\end{aligned}$$

since the supremum is attained at $x=-(1-b)h/2$. Since $\delta>0$ was arbitrary, we see that the sequence $(T_{n,H})$ is not locally asymptotically equivariant at $\theta_0=0$. □

Proof of **Proposition 1.4.2**.

Recall the definition of the metric d in (1.3). We deal separately with the cases $\theta=0$ and $\theta\neq 0$. Let $m\in\mathcal{M}$, and $m^-=An^\alpha$. Given $\epsilon>0$, we have

$$\begin{aligned}\mathbb{P}_{\theta=0}\{d(H_m(\bar X_n),H_n(\theta))>\epsilon\} &= \mathbb{P}_{\theta=0}\{d(H_m(\bar X_n),H_n(\theta))>\epsilon,|\bar X_n|\le Cn^{-\beta}\}\\ &\quad+\mathbb{P}_{\theta=0}\{d(H_m(\bar X_n),H_n(\theta))>\epsilon,|\bar X_n|>Cn^{-\beta}\}\\ &\le \mathbb{P}_{\theta=0}\{d(H_{m^-}(\bar X_n),H_n(\theta))>\epsilon\}+2\Phi(-Cn^{1/2-\beta})\\ &\to 0\end{aligned}$$

as $n \to \infty$, by Corollary 2.1(b) of Beran (1997). On the other hand,

$$\begin{aligned}
\mathbb{P}_{\theta\neq 0}\big\{d\big(H_m(\bar{X}_n), H_n(\theta)\big) > \epsilon\big\} &= \mathbb{P}_{\theta\neq 0}\big\{d\big(H_m(\bar{X}_n), H_n(\theta)\big) > \epsilon, |\bar{X}_n| \le Cn^{-\beta}\big\} \\
&\quad + \mathbb{P}_{\theta\neq 0}\big\{d\big(H_m(\bar{X}_n), H_n(\theta)\big) > \epsilon, |\bar{X}_n| > Cn^{-\beta}\big\} \\
&\le \Phi(Cn^{1/2-\beta} - n^{1/2}\theta) - \Phi(-Cn^{1/2-\beta} - n^{1/2}\theta) \\
&\quad + \mathbb{P}_{\theta\neq 0}\big\{d\big(H_n(\bar{X}_n), H_n(\theta)\big) > \epsilon\big\} \\
&\to 0
\end{aligned}$$

as $n \to \infty$, by Theorem 1.2.6. □

Chapter 2

Small confidence sets for the mean of a spherically symmetric distribution

2.1 Introduction

Suppose that X has a k-dimensional spherically symmetric distribution about θ, with density $f(\|x-\theta\|^2)$. The usual $(1-\alpha)$-level confidence set for θ is

$$C^0(X) = \{\theta \in \mathbb{R}^k : \|X-\theta\|^2 \leq c^2\},$$

where c^2 satisfies

$$\int_{\mathbb{R}^k} f(\|x\|^2)\mathbb{1}_{\{\|x\|^2 \leq c^2\}}\, dx = 1-\alpha.$$

This chapter is concerned with the construction of improved confidence sets for θ when $k \geq 3$. Specifically, we consider sets of the form

$$\{\theta \in \mathbb{R}^k : \|T_S^+(X)-\theta\|^2 \leq v^2(\|X\|)\},$$

where

$$T_S^+(X) = \left(1 - \frac{a}{\|X\|^2}\right)_+ X$$

is a positive-part Stein estimator, $h_+ = \max(h, 0)$ and $a > 0$. Note that, in contrast to Chapter 1, and to simplify the calculations, we consider the version of the Stein estimator which shrinks the observations towards the origin. We investigate two methods of construction of the radius function $v(\cdot)$, both involving direct approximation of the upper α-point of the sampling distribution of $\|T_S^+(X) - \theta\|^2$. The first is an analytic procedure, giving an explicit expression for $v(\cdot)$ which is never larger than c and which can be considerably smaller. Despite this, subject to minor conditions on the underlying density, we are able to show that the resulting confidence set dominates $C^0(X)$ in terms of coverage probability, provided $\|\theta\|$ is either less than a given bound, or sufficiently large. Simulations suggest that dominance may well be attained for all values of $\|\theta\|$, at least for moderate or large k. An alternative to the analytic procedure is to apply the parametric bootstrap. Here, even greater improvement in volume over the original confidence set is possible, without the coverage probability dropping below the nominal level, but at the expense of a less explicit radius function.

Structurally, the confidence sets are of the same form as those of Casella and Hwang (1983), who consider only the multivariate normal case and who obtain their radius by modifying the solution to an empirical Bayes problem. However, the sets constructed in this paper, as well as having a more natural motivation, compare favourably in the region of the parameter space which is of most interest when applying the positive-part Stein estimator (c.f. the discussion in Section 2.6).

The importance of finding good confidence sets for the mean of a spherically symmetric distribution derives from their applications. For $n \geq k$, consider the linear model

$$\underset{n\times 1}{X} = \underset{n\times k}{A}\ \underset{k\times 1}{\theta} + \sigma\ \underset{n\times 1}{\epsilon}, \tag{2.1}$$

where the design matrix A is assumed to be of full rank k, and where the error vector ϵ has a density which is spherically symmetric about the origin. Of course, this model includes the standard linear model with normally distributed errors as an important special case. Hwang and Chen (1986) show how the problem of finding a confidence set for θ in the model (2.1) can be reduced to the simpler form studied in this chapter, provided that the error variance σ^2 is known.

Zellner (1976) cites several authors who have considered the linear model with spherically symmetric errors as a model for practical situations, and proposes other scenarios himself. Properties of the usual least squares estimator, $\hat{\theta} = (A^T A)^{-1} A^T X$, in the model (2.1), have been studied by Thomas (1970), Zellner (1976) and Box (1953), amongst others.

Interest in the problem of point estimation of θ when X has a multivariate normal distribution was sparked by the celebrated discovery of Stein (1956), who proved the existence of estimators which strictly dominate X with respect to the squared error loss function when $k \geq 3$. Brandwein and Strawderman (1978) and Brandwein (1979) extended these results to cover spherically symmetric distributions. It is now known that the Stein phenomenon applies to a very wide class of distributions and loss functions – see, for example, Brandwein and Strawderman (1990) or Evans and Stark (1996). By contrast, progress on the confidence set problem has been much slower, to the extent that results for confidence sets which strictly dominate the obvious confidence set in terms of volume are still restricted to the multivariate normal distribution. As several authors testify, this is not to do with the lesser importance of the confidence set problem, but rather because of its technical difficulty.

A loss function is rarely stated explicitly in the confidence set problem, though Casella and Hwang (1983) and Beran (1995) are exceptions in this regard. Instead, different confidence procedures tend to be compared according to four criteria: shape, cov-

erage probability, volume and conditional properties. It is the need to ensure good performance in all of these respects that makes the problem so demanding.

It is difficult to make concrete statements regarding the shape of a reasonable confidence set. At first sight, it is hard to look beyond a sphere when dealing with a spherically symmetric distribution. However, Berger (1980) gives a heuristic argument suggesting that this choice may not be so clear-cut. In fact, Faith (1976), Shinozaki (1989) and Tseng and Brown (1997) have all also proposed non-spherical confidence regions. There is a consensus that an acceptable confidence set should be at least connected, though this still seems to be quite a weak requirement. We suspect that most practitioners would be reluctant to use a confidence set unless its geometry were fairly well understood.

Fortunately, coverage probability and volume can be treated in a more satisfactory way, and they are of course intimately linked. According to Joshi (1969), a confidence set $C(X)$ strictly dominates $C^0(X)$ if

(i) $\mathbb{P}_\theta(C(X) \ni \theta) \geq \mathbb{P}_\theta(C^0(X) \ni \theta)$ for all $\theta \in \mathbb{R}^k$;

(ii) $\mathrm{Vol}\big(C(x)\big) \leq \mathrm{Vol}\big(C^0(x)\big)$ for all $x \in \mathbb{R}^k$,

with strict inequality either in (i) for some θ, or in (ii) for all x in some set with positive Lebesgue measure. Joshi also pointed out that two confidence sets should be considered equivalent if their symmetric difference has zero volume. Of course, the practitioner is more interested in a reduction in volume, provided that the coverage probability does not drop below the nominal level, than in increased coverage probability at a fixed volume.

Appreciation of the importance of the conditional properties of confidence sets began with Fisher (1956, 1959). Rules for satisfactory conditional performance were

formalised by Buehler (1959) and Robinson (1979a,b) in terms of a betting game between two players. Casella and Hwang (1986) were the first to consider the conditional properties of confidence sets for the mean of a multivariate normal distribution. Robinson (1979b), Lu and Berger (1989), Robert and Casella (1994) and Wang (2000) all take another approach, and discuss how to improve the reported confidence statements for the usual confidence set $C^0(X)$.

Tseng and Brown (1997) give an excellent review of the earlier literature on the multivariate normal confidence set problem, which, in addition to those references already given, includes Stein (1962, 1981) and Hwang and Casella (1982, 1984). Tseng and Brown themselves propose somewhat egg-shaped sets which have exact coverage probability and they also find sufficient conditions under which their sets uniformly dominate $C^0(X)$ in terms of volume. Unfortunately, as the authors themselves admit, these sufficient conditions do not appear to be entirely satisfactory, and it seems difficult to choose an optimal set from within the class they study. In addition, the shape of the sets can be quite complicated, although certain results concerning the geometry are obtained.

Previous work on the spherically symmetric case, such as Ki and Tsui (1985), Hwang and Chen (1986) and Robert and Casella (1990), has focused on proving that confidence sets of the same radius as $C^0(X)$ have higher coverage probability when recentred at a positive-part Stein estimator. In this chapter, we recognise that once a spherical confidence set centred at the positive-part Stein estimator has been decided upon, the ideal, exact, $(1-\alpha)$-level confidence set would be

$$\{\theta \in \mathbb{R}^k : \|T_S^+(X) - \theta\|^2 \leq w_\alpha(\theta)\},$$

where $w_\alpha(\theta)$ is the upper α-point of the sampling distribution of $\|T_S^+(X) - \theta\|^2$. Of course, this is not a feasible confidence set as the radius depends on the unknown θ. The approach taken in Section 2.3 is a direct, analytic estimation of $w_\alpha(\theta)$. Specif-

ically, we compute the first two non-zero terms in the Taylor series of $w_\alpha(\theta)$ about the origin, allowing us to write

$$w_\alpha(\theta) = w_\alpha(0) + \frac{1}{2} w''_\alpha(0)\|\theta\|^2 + o(\|\theta\|^2)$$

as $\|\theta\| \to 0$. Ignoring the $o(\|\theta\|^2)$ term, we estimate $\|\theta\|^2$ by an estimator $\|\hat{\theta}\|^2$ and obtain the confidence set

$$C(X) = \Big\{\theta \in \mathbb{R}^k : \|T_S^+(X) - \theta\|^2 \leq \min\Big(w_\alpha(0) + \frac{1}{2} w''_\alpha(0)\|\hat{\theta}\|^2, c^2\Big)\Big\}.$$

We are motivated by the knowledge that $T_S^+(X)$ performs best as an estimator of θ when $\|\theta\|$ is small, which suggests that this is the region of the parameter space where one would expect a spherical confidence set centred at the positive-part Stein estimator with radius $v(\|X\|) = c$ to show greatest improvement, in terms of coverage probability, over $C^0(X)$. Simulations in Hwang and Casella (1982) support this intuition. More importantly, this suggests that it is for small values of $\|X\|$ that we can hope to see the greatest reduction in volume while maintaining at least the same coverage probability as $C^0(X)$. The radius function $v(r)$ we propose attains the value $c - a/c$ at $r = 0$, which is rather smaller than the suggestion in Casella and Hwang (1983).

In Section 2.4, we make use of the simple analytic form of the radius function to prove some results about the properties of the confidence set. A particularly interesting feature of the work from a theoretical point of view is that the radius of the analytic confidence set depends on the density f only through c^2, and a quantity $f'(c^2)/f(c^2)$, called the Relative Increasing Rate (RIR) of f at c^2. Both Hwang and Chen (1986) and Robert and Casella (1990) have noted the importance of this latter quantity in establishing dominance of their recentred sets over $C^0(X)$. Simulations of the coverage probabilities are provided for three spherically symmetric densities: the k-variate normal, the multivariate t, and the double exponential. These latter distributions

were studied in Hwang and Chen (1986). The multivariate t distribution with N degrees of freedom has density

$$f(\|x\|^2) \propto \left(1 + \frac{\|x\|^2}{N}\right)^{-(N+k)/2}.$$

Relative to the normal model, it gives more flexibility to the practitioner, through the choice of the number of degrees of freedom, but is a close approximation to normality when the number of degrees of freedom is large (c.f. Zellner, 1976). The double exponential distribution with parameter d has density

$$f(\|x\|^2) \propto e^{-d\|x\|},$$

and the parameter choice $d = (k+1)^{1/2}$ ensures each component of X has unit variance.

As was suggested by the work in Chapter 1, another appealing approach to the problem for the modern statistician involves a parametric bootstrap procedure. In the related problem where we have independent random vectors $X_1, \ldots, X_n$, each having the same spherically symmetric distribution as X, we encounter a similar inconsistency problem to that in Section 1.4.2. Nevertheless, as was discussed in Chapter 1, inconsistency does not preclude the bootstrap performing successfully at finite sample sizes. We investigate the parametric bootstrap confidence set in Section 2.5, and present various comments and generalisations in Section 2.6. Most of the proofs and some ancillary results are deferred to Section 2.7.

2.2 Preliminaries

Suppose X is a random variable taking values in a sample space $\mathcal{X}$ which is endowed with a σ-algebra. Suppose also that θ is a parameter of the distribution of X, taking

values in a parameter space Θ which is also equipped with a σ-algebra. A confidence procedure C is a measurable subset of $\mathcal{X} \times \Theta$, and has two associated cross-sections. A confidence set $C(X)$ for θ is the X-section:

$$C(X) = \{\theta \in \Theta : (X, \theta) \in C\}.$$

The abuse of notation in which θ is used both for the parameter of interest and the index in the confidence set is standard and causes no confusion. The θ-section is

$$C(\theta) = \{x \in \mathcal{X} : (x, \theta) \in C\}.$$

As we will see, the θ-section is important for calculations involving conditional and unconditional coverage probabilities of confidence sets. Moreover, for $\theta_0 \in \Theta$, the set $C(\theta_0)$ can be seen as an acceptance region for a hypothesis test of $H_0 : \theta = \theta_0$. Birnbaum (1955) treats hypothesis testing as a problem in decision theory with $0-1$ loss, and shows that, under certain conditions on the spherically symmetric density f, a test of $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$ is admissible if and only if its acceptance region is convex (provided we regard tests whose acceptance regions differ only on a set of measure zero as equivalent). As our primary interest is finding confidence sets rather than in hypothesis testing, we shall follow Casella and Hwang (1983) and merely require that $C(\theta)$ should be connected for each $\theta \in \mathbb{R}^k$.

2.3 Constructing the analytic confidence set

We say that X has a k-variate spherically symmetric distribution about θ if $X - \theta$ has the same distribution as $P(X - \theta)$ for all $k \times k$ orthogonal matrices P. We assume that $k \geq 3$ and that the distribution, $\mathbb{P}_\theta$, of X has a density with respect to Lebesgue measure on $\mathbb{R}^k$, whose value at a point $x \in \mathbb{R}^k$ may therefore be written as

$f(\|x-\theta\|^2)$. We begin with a useful result concerning all estimators of θ of the form $\gamma(\|X\|)X$, where $\gamma : [0,\infty) \to \mathbb{R}$ is a measurable function:

Proposition 2.3.1. *For $\alpha \in (0,1)$, the upper α-point of the sampling distribution of $\|\gamma(\|X\|)X-\theta\|^2$ depends on θ only through $\|\theta\|$.*

Proof. The upper α-point, w, satisfies

$$\int_{\mathbb{R}^k} f(\|x-\theta\|^2)\mathbb{1}_{\{\|\gamma(\|x\|)x-\theta\|^2\le w\}}\,dx = 1-\alpha.$$

If P is a $k\times k$ orthogonal matrix, then

$$\int_{\mathbb{R}^k} f(\|x-P\theta\|^2)\mathbb{1}_{\{\|\gamma(\|x\|)x-P\theta\|^2\le w\}}\,dx = \int_{\mathbb{R}^k} f(\|P^Tx-\theta\|^2)\mathbb{1}_{\{\|\gamma(\|x\|)P^Tx-\theta\|^2\le w\}}\,dx,$$

from which the result follows, on substituting $y = P^Tx$. □

The positive-part Stein estimator

$$T_S^+(X) = \left(1-\frac{a}{\|X\|^2}\right)_+ X \tag{2.2}$$

is of the form $\gamma(\|X\|)X$, and we let $w_\alpha(\|\theta\|)$ denote the upper α-point of the sampling distribution of $\|T_S^+(X)-\theta\|^2$. The theorem below is the main theorem of this section, and is proved in Section 2.7.

Theorem 2.3.2. *Suppose that $w_\alpha(0) > 0$, and that the spherically symmetric density f is twice continuously differentiable. Then*

$$w_\alpha(\|\theta\|) = w_\alpha(0) + \frac{1}{2}w_\alpha''(0)\|\theta\|^2 + o(\|\theta\|^2)$$

as $\|\theta\|\to 0$, where $w_\alpha(0) = (c-a/c)^2$ and

$$\frac{1}{2}w_\alpha''(0) = \left\{\frac{1}{k}\Big(1-\frac{a}{c^2}\Big)\Big(\frac{a(k-1)}{c^2+a} - \frac{2ac^2}{(c^2+a)^2} - \frac{2a^2}{c^2+a}\frac{f'(c^2)}{f(c^2)}\Big) + \frac{a(k-1)}{c^2k}\right\}.$$

The condition that $w_\alpha(0) > 0$ is equivalent to requiring that $\alpha < \mathbb{P}_0(\|X\|^2 > a)$, which in turn is equivalent to $c^2 > a$; this will rarely be restrictive in practice. For instance, when f is the standard k-variate normal density, James and Stein (1961) showed that the ordinary Stein estimator

$$T_S(X) = \Big(1 - \frac{a}{\|X\|^2}\Big)X$$

strictly dominates X in the point estimation problem with squared error loss for $a \in \big(0, 2(k-2)\big)$, and that $a = k-2$ is the optimal choice. In this case, the confidence set

$$\{\theta \in \mathbb{R}^k : \|X - \theta\|^2 \leq k - 2\}$$

has only about 50% coverage probability.

Having computed $w_\alpha(0)$ and $w''_\alpha(0)$, a natural, theoretical confidence set for θ, of nominal $(1-\alpha)$-level coverage, is

$$C(X) = \Big\{\theta \in \mathbb{R}^k : \|T_S^+(X) - \theta\|^2 \leq \min\Big(w_\alpha(0) + \frac{1}{2}w''_\alpha(0)\|\theta\|^2, c^2\Big)\Big\}.$$

Of course, this confidence set cannot be used in practice, as the radius depends on the unknown $\|\theta\|^2$. However, we can estimate $\|\theta\|^2$ from the data. There are many possible ways of doing this: a bootstrap approach would suggest that the square of the population mean should be estimated by the square of the sample mean, namely $\|X\|^2$; the minimum variance unbiased estimate is $\|X\|^2 - k$, but we do not wish to allow our estimate to be negative, so $(\|X\|^2 - k)_+$ is another possibility; $\|T_S^+(X)\|^2$ is a third option.

Simulations of the coverage probabilities of these confidence sets with $a = k-2$ were performed. These confirmed that estimating $\|\theta\|^2$ by $(\|X\|^2 - k)_+$ or $\|T_S^+(X)\|^2$ fails to yield a $(1-\alpha)$-level confidence set. The principal reason for this is that in these cases, the radius of the confidence set takes the constant value $c - a/c$ for $\|X\|^2 \leq a$.

Thus the coverage probability drops by at least $F_{k,(c-a/c)^2}(a)$ at $\|\theta\| = c - a/c$, where $F_{k,\lambda}(\cdot)$ is the distribution function of a non-central chi-squared random variable with k degrees of freedom and non-centrality parameter λ.

On the other hand, estimating $\|\theta\|^2$ by $\|X\|^2$ is more successful. This has the additional advantage of simplicity, and the resulting confidence set is

$$C(X) = \Big\{\theta \in \mathbb{R}^k : \|T_S^+(X) - \theta\|^2 \leq \min\Big(w_\alpha(0) + \frac{1}{2}w''_\alpha(0)\|X\|^2, c^2\Big)\Big\}. \qquad (2.3)$$

As noted in Section 2.1, an extremely interesting feature of this confidence set is that it depends on the density f only through c^2 and the Relative Increasing Rate of f at c^2. Typically, c^2 will be large enough to ensure that the RIR at c^2 is negative, with very negative values indicating that the distribution has light tails. For the three distributions mentioned in Section 2.1, namely the standard multivariate normal, the multivariate t with N degrees of freedom, and the double exponential with parameter d, the relative increasing rates at c^2 are

$$-\frac{1}{2}, \quad -\frac{N+k}{2(N+c^2)} \quad \text{and} \quad -\frac{d}{2c},$$

respectively.

Since $C^0(X)$ is minimax (Stein, 1962), a necessary condition for the confidence set $C(X)$ in (2.3) to dominate $C^0(X)$ in coverage probability is that $w''_\alpha(0) > 0$. Perhaps surprisingly in view of the results of Hwang and Chen (1986) and Robert and Casella (1990), this condition corresponds to the RIR at c^2 being *less* than some positive bound depending on a, c^2 and k. One of the themes of these previous works is that a confidence set of the same radius as $C^0(X)$ has uniformly higher coverage probability when recentred at a positive-part Stein estimator, provided that the RIR at c^2 is *greater* than some negative bound. As mentioned in the previous paragraph, however, this positive bound will almost certainly be unrestrictive in practice.

The choice of a is more delicate in the spherically symmetric case than the multivariate normal. For general spherically symmetric distributions, Brandwein (1979) showed that the ordinary Stein estimator $T_S(X)$ dominates X with respect to squared error loss for $k \geq 4$ and

$$a \in \left(0, \frac{2(k-2)}{k\mathbb{E}_0(\|X\|^{-2})}\right],$$

provided $\mathbb{E}_0(\|X\|^{-2})$ is finite. For unimodal spherically symmetric distributions, Brandwein and Strawderman (1978) proved that the range can be improved to

$$a \in \left(0, \frac{2k}{(k+2)\mathbb{E}_0(\|X\|^{-2})}\right],$$

for $k \geq 4$, and $a \in (0, 0.375]$ for $k = 3$. For $k \geq 4$, this upper bound is

$$\frac{2(k-2)k}{k+2} \quad \text{and} \quad \frac{2(k-1)(k-2)k}{(k+1)(k+2)}$$

for the multivariate t distribution with N degrees of freedom, and the double exponential distribution with parameter $d = (k+1)^{1/2}$, respectively. These upper bounds are close to $2(k-2)$ for moderate and large k, but no optimal estimator is given.

In their study of confidence sets in spherically symmetric distributions, Hwang and Chen (1986) find values of a_0 such that sets of form

$$\{\theta \in \mathbb{R}^k : \|T_S^+(X) - \theta\|^2 \leq c^2\} \tag{2.4}$$

have uniformly higher coverage probability than $C^0(X)$, for all $a \in (0, a_0]$. These values arise from showing the derivative of the coverage probability with respect to a is non-negative for $a \in (0, a_0]$, and are therefore relatively weak. In the special cases of the multivariate t and double exponential distributions, Hwang and Chen (1986) succeed in giving somewhat improved bounds on a for domination, although these are still rather smaller than $k-2$. The same authors also demonstrate that

$$a \leq -(k-2)\frac{f(c^2)}{f'(c^2)}$$

is a necessary condition for domination in terms of coverage probability, provided $f'(c^2) > 0$. Note that this bound is equal to $2(k-2)$ in the multivariate normal case, and will typically be even larger for multivariate t distributions and the double exponential distribution with parameter $d = (k+1)^{1/2}$. Numerical studies and a heuristic argument in Hwang and Chen (1986) suggest that this necessary condition is close to being sufficient.

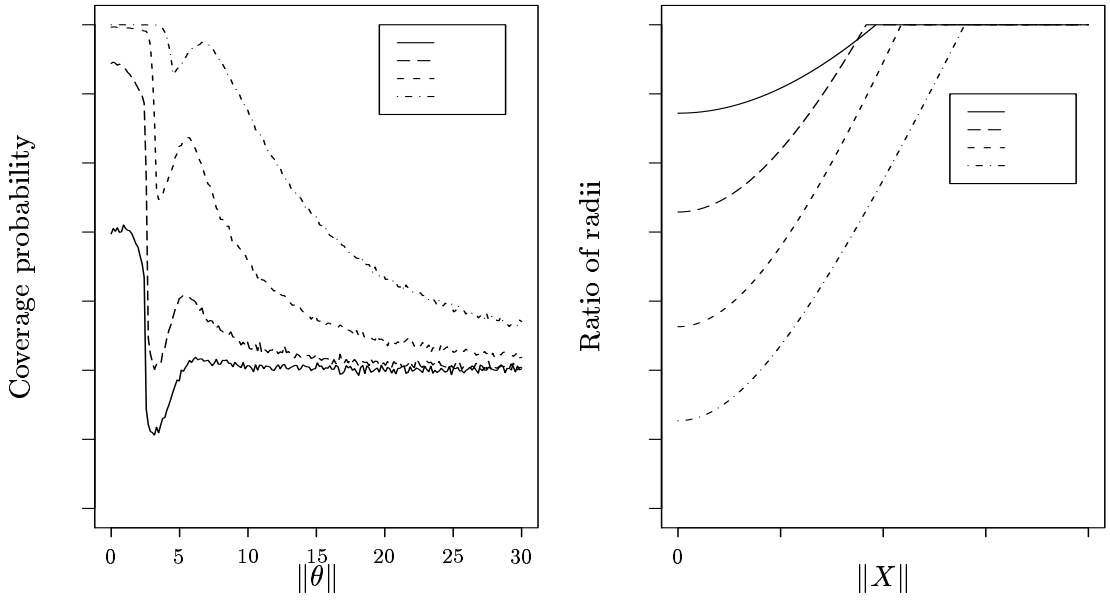
An immediate corollary of Theorem 5.1 of Hwang and Chen (1986) is that, under their mild conditions on f,

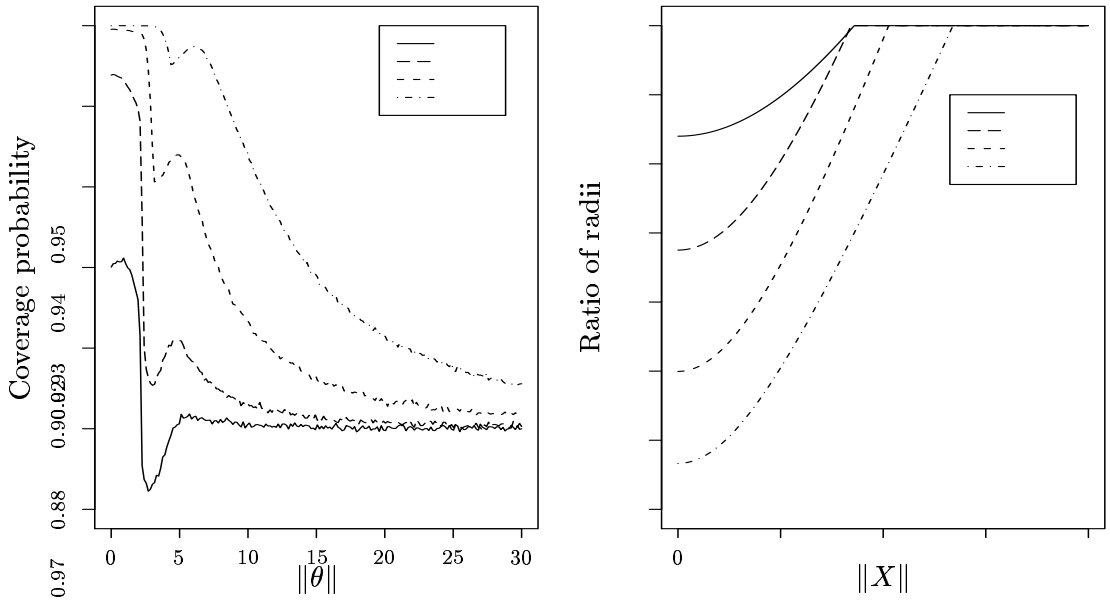
$$a = -\frac{(k-2)f(c^2)}{2f'(c^2)} \tag{2.5}$$

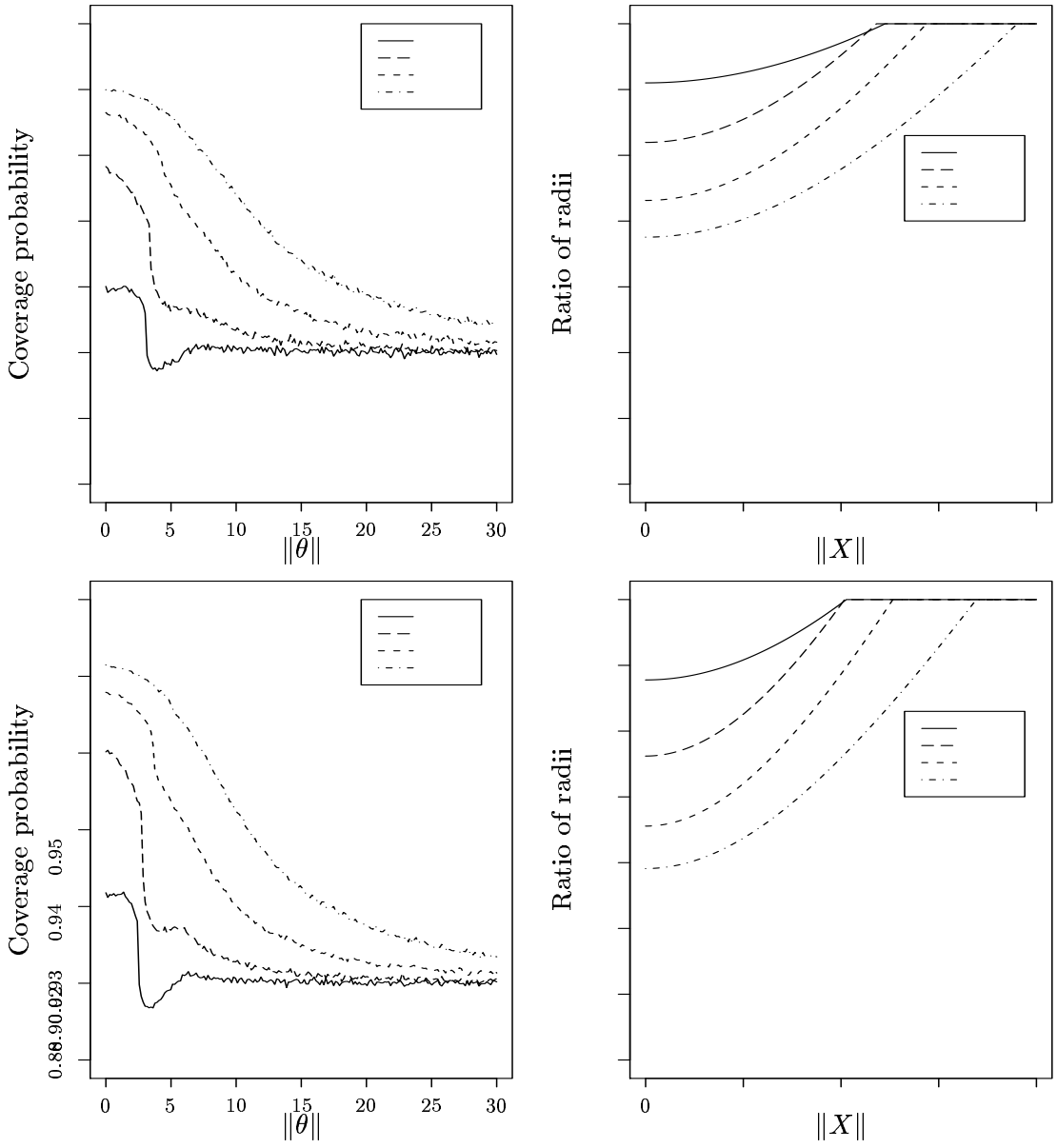
gives the greatest coverage probability for large $\|\theta\|$ for confidence sets of the form (2.4). Moreover, their theorem also shows that this coverage probability is greater than $1-\alpha$ for sufficiently large $\|\theta\|$. Since (2.4) is the same as our analytic confidence set $C(X)$ in (2.3) for sufficiently large $\|X\|$ provided that $w''_\alpha(0) > 0$, it follows that $C(X)$ strictly dominates $C^0(X)$ in coverage probability for sufficiently large $\|\theta\|$.

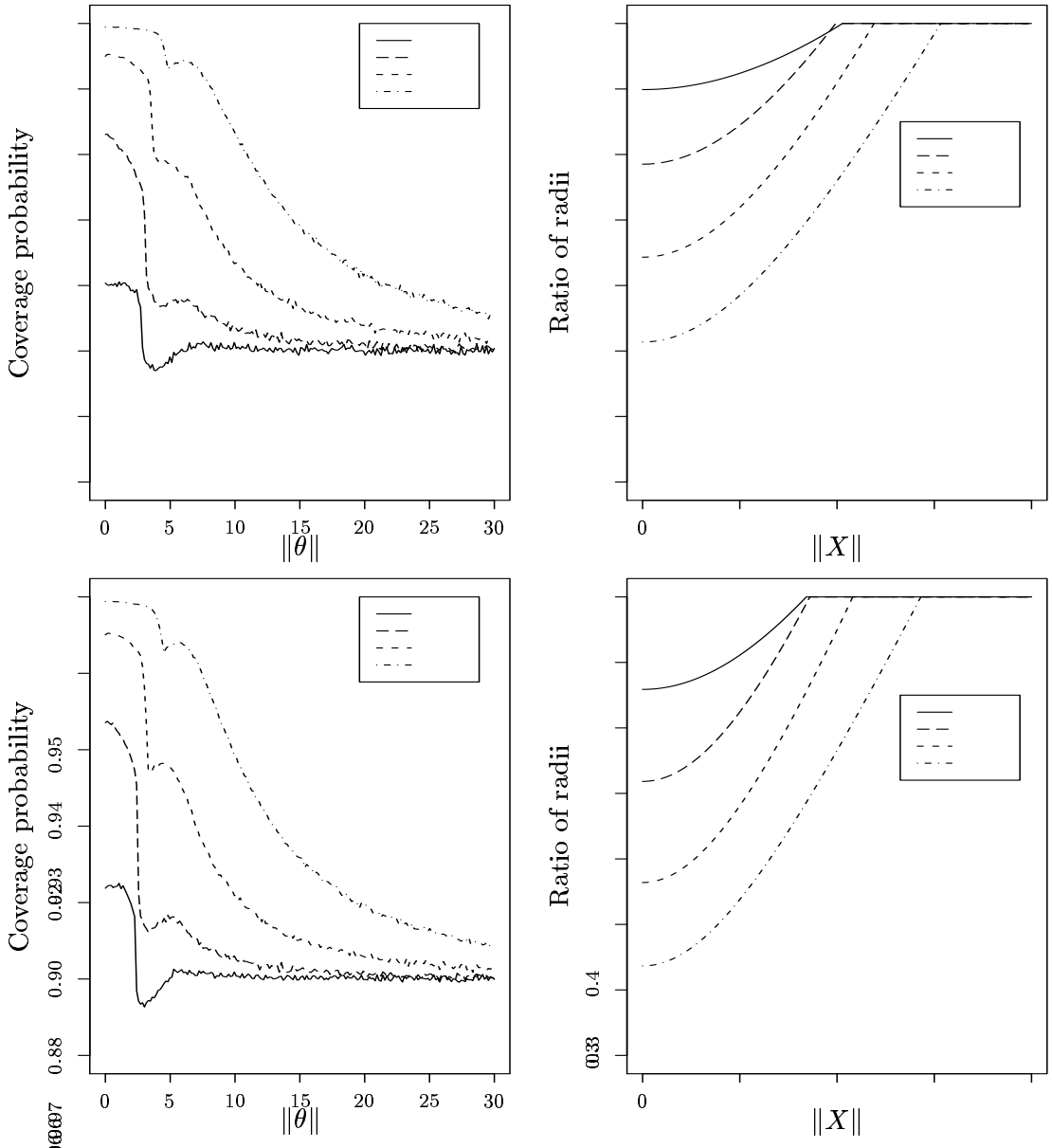
As the value of a in (2.5) above is close to $k-2$ in the cases considered here, we take $a = k-2$ in our numerical studies of the confidence set (2.3), which are presented in Figures 2.1, 2.2 and 2.3. This has the additional advantage of simplifying comparison between different distributions. In each figure, 400,000 Monte-Carlo repetitions were used to approximate the coverage probability at each value of θ, giving a simulation error standard deviation of about 0.0005 at each point.

It appears that the confidence set (2.3) dominates $C^0(X)$ in terms of coverage probability for all of the distributions considered, apart from possibly in a narrow middle range of values of $\|\theta\|$ for small values of k. These exceptions are similar to those found in Casella and Hwang (1983) and the problems are sufficiently small that they can be ignored in most practical contexts. In view of the point estimation results


Coverage probability
0
5
10
15
20
25
30
$\|\theta\|$
Ratio of radii
0
$\|X\|$


Coverage probability
0.88
0.90
0.94
0.95
0
5
10
15
20
25
30
$\|\theta\|$
Ratio of radii
0
$\|X\|$


Coverage probability
0.88
0.90
0.92
0.94
0.95
0
5
10
15
20
25
30
$\|\theta\|$
Ratio of radii
0
$\|X\|$
Coverage probability
0.94
0.96
0.97
0.98
0.99
1
0
5
10
15
20
25
30
$\|\theta\|$
Ratio of radii
0
$\|X\|$


Coverage probability
0
5
10
15
20
25
30
$\|\theta\|$
Ratio of radii
0
$\|X\|$
Coverage probability
0.88
0.90
0.94
0.95
0.98
0.99
1
0
5
10
15
20
25
30
$\|\theta\|$
Ratio of radii
0.4
0
$\|X\|$

$\|X\|$	0	1	2	3	4	5	6	8	10	20
$k = 5$	0.64	0.68	0.79	0.98	1.13	1.07	1.05	1.03	1.02	1.00
$k = 10$	0.43	0.48	0.62	0.85	1.18	1.21	1.13	1.07	1.04	1.01

Table 2.1: Ratio of the radii of the confidence set (2.3) to the corresponding ones of Casella and Hwang (1983) in the k-variate normal case and with $\alpha = 0.05$.

$\|\theta\|$	0	1	2	3	4	5	6
$k = 5$	0.99	0.97	0.87	0.58	0.24	0.05	0.01
$k = 10$	0.97	0.95	0.83	0.58	0.26	0.06	0.01

Table 2.2: Estimates of the probability that the radius of the confidence set (2.3) is less than c in the k-variate normal case and with $\alpha = 0.05$.

of Brandwein and Strawderman (1978) mentioned above, the choice of a is almost certainly too large for these small values of k, although we do not pursue this matter further here.

On the other hand, for all values of k, the radii show a great improvement over those of $C^0(X)$, especially for small $\|X\|$, as one would expect from their construction. In fact, in the k-variate normal case, the radii tend to be considerably smaller than those of Casella and Hwang (1983), with which they are directly comparable, for small values of $\|X\|$, at the expense of being slightly larger for larger values of $\|X\|$ (c.f. Table 2.1). The ratio of the radii of (2.3) and $C^0(X)$ in the best case, that is when $\|X\| = 0$, is $1 - a/c^2$. Thus, for fixed a, the maximum improvement in volume is greater for distributions with lighter tails. Table 2.2 gives estimates of the probability that the radius of the confidence set (2.3) is less than c, which, of course, is a decreasing function of $\|\theta\|$.

2.4 Properties of the analytic confidence set

We have already discussed the need for the RIR of f at c^2 to be negative in order for the confidence sets $C(X)$ in (2.3) to have adequate coverage probability. In this section we will see that the sets have several desirable properties provided also that the RIR at c^2 is not *too* negative. These results are more in line with the works of Hwang and Chen (1986) and Robert and Casella (1990), where dominance occurs provided the tails of the distribution are heavy enough. Throughout this section, we assume $k \geq 3$, and that f is twice continuously differentiable. Proofs are deferred to Section 2.7. We start with some elementary bounds, which give simple yet general conditions under which Proposition 2.4.2 and Lemma 2.4.3 hold.

Lemma 2.4.1. *(i) Let $a \in (0, k-1]$ and $\alpha \in \big(0\,,\, \mathbb{P}_0(\|X\|^2 > a)\big)$, and suppose that $f'(c^2)/f(c^2) \geq -1/2$. Then*

$$\frac{1}{2} w''_\alpha(0) \leq \frac{k-1}{k}.$$

(ii) Let $a > 0$ and $\alpha \in \big(0, \mathbb{P}_0(\|X\|^2 > a)\big)$, and suppose that $f'(c^2)/f(c^2) \leq 0$. Then

$$\frac{1}{2} w''_\alpha(0) \geq \frac{a(k-1)}{c^2 k}.$$

The next three results concern the θ-section associated with the confidence set (2.3), given by

$$C(\theta) = \Big\{ x \in \mathbb{R}^k : \|T_S^+(x) - \theta\|^2 \leq \min\Big(w_\alpha(0) + \frac{1}{2} w''_\alpha(0) \|x\|^2, c^2 \Big) \Big\}.$$

The first is an extension of Theorem A1 of Casella and Hwang (1983).

Proposition 2.4.2. *Let $a > 0$, and suppose that $0 < w''_\alpha(0)/2 < 1$. Then $C(\theta)$ is connected, for all $\theta \in \mathbb{R}^k$.*

In fact, $C(\theta)$ can possess a stronger property when $\|\theta\|$ lies in a range which is of particular importance to us (c.f. Section 2.6).

Lemma 2.4.3. *Let $a > 0$, and suppose that $0 < w''_\alpha(0)/2 < 1$. For $\|\theta\| \leq c - a/c$, if $x \in C(\theta)$, then so is $t\,x$, for all $t \in [0,1]$.*

In order to present the main theorem of this section, we let

$$C^0(\theta) = \{x \in \mathbb{R}^k : \|x - \theta\|^2 \leq c^2\}$$

denote the θ-section corresponding to the usual confidence set $C^0(X)$.

Theorem 2.4.4. *Let $a \in (0, k-1]$, $\alpha \in \big(0\,,\, \mathbb{P}_0(\|X\|^2 > a)\big)$ and also suppose that $-1/2 \leq f'(c^2)/f(c^2) \leq 0$. If*

$$\|\theta\|^2 \leq \min\left(w_\alpha(0), \frac{c^2 - a}{2w''_\alpha(0)c^4}\Big\{2w''_\alpha(0)c^4 - (c^2 - a)a\Big\}\right),$$

then $C^0(\theta) \subseteq C(\theta)$.

The upper bound on range of values of $\|\theta\|$ for which the conclusion of the theorem holds is the best possible, and the theorem is clearly non-vacuous since it holds for $\|\theta\| = 0$. In fact, the upper bound corresponds to a point just before the sharp drop in coverage probability seen in Figures 2.1, 2.2 and 2.3. For instance, when $k = 5$, $\alpha = 0.05$ and f is the k-variate normal density, we have $C^0(\theta) \subseteq C(\theta)$ for $\|\theta\| \leq 2.7$.

An obvious corollary of this theorem is that $C(X)$ dominates $C^0(X)$ in terms of coverage probability for the particular range of values of $\|\theta\|$ above. Moreover, it also has implications for the conditional properties of $C(X)$, which we now describe.

When making an assertion of the form

$$\mathbb{P}_\theta(C(X) \ni \theta) = 1 - \alpha,$$

the statistician is averaging (integrating) over the sample space. However, the confidence set must be specified on the basis of observing $X = x$, say. The statistician should, therefore, question whether the probability assertion is still valid in the light of the data. For instance, such considerations provide a strong criticism of confidence sets centred at the ordinary, as opposed to positive-part, Stein estimator. For, if $\|x\|$ were very small, the confidence set would presumably be well away from the origin, and the statistician would be unable to justify the probability statement in the light of the data, whatever the true value of θ. Put another way, a hypothetical opponent of the statistician could specify a very small sphere A centred at the origin, staking an amount α to win $1 - \alpha$ that $C(x)$ does not contain θ if $x \in A$, and not making a bet otherwise. Under infinitely many hypothetical repetitions of the experiment with a referee who knows the true value of θ, the opponent would win almost surely.

More formally, Buehler (1959) and Robinson (1979a) introduced various criteria for judging the conditional performance of a confidence set. In our situation, if A is a subset of $\mathbb{R}^k$ of positive Lebesgue measure, Robinson calls A a negatively biased relevant subset for $C(X)$ if there exists $\epsilon > 0$ such that

$$\mathbb{P}_\theta(C(X) \ni \theta \,|\, X \in A) \leq 1 - \alpha - \epsilon$$

for all $\theta \in \mathbb{R}^k$, and advocates that one should not use a confidence set if there exists a negatively biased relevant subset.

Another simple corollary of Theorem 2.4.4 is that for any subset A of $\mathbb{R}^k$ of positive Lebesgue measure, we have

$$\mathbb{P}_\theta(C(X) \ni \theta \,|\, X \in A) \geq \mathbb{P}_\theta(C^0(X) \ni \theta \,|\, X \in A) \tag{2.6}$$

for $\|\theta\|$ in the given range. Casella and Hwang (1986) show that for any $u > 0$, there exists $\delta = \delta(u) > 0$ such that

$$\mathbb{P}_\theta(C^0(X) \ni \theta \,|\, \|X\|^2 \leq u) > 1 - \alpha$$

for all $\|\theta\|^2 < \delta$, and a very similar argument shows that, for any $\xi \in \mathbb{R}^k$ and $u > 0$, there exists $\delta = \delta(u) > 0$ such that

$$\mathbb{P}_\theta(C^0(X) \ni \theta | \, \|X - \xi\|^2 \leq u) > 1 - \alpha \tag{2.7}$$

for all $\|\theta - \xi\|^2 < \delta$. Combining (2.6) and (2.7), we see that there are no negatively biased relevant spheres centred at ξ for $C(X)$, provided

$$\|\xi\|^2 < \min\left(w_\alpha(0), \frac{c^2 - a}{2w''_\alpha(0)c^4}\left\{2w''_\alpha(0)c^4 - (c^2 - a)a\right\}\right).$$

2.5 The bootstrap confidence set

Here we investigate another way of approximating the ideal confidence set

$$\{\theta \in \mathbb{R}^k : \|T_S^+(X) - \theta\|^2 \leq w_\alpha(\|\theta\|)\}. \tag{2.8}$$

In a parametric bootstrap procedure, we estimate θ by $\hat{\theta}$, say, and approximate (2.8) by

$$\{\theta \in \mathbb{R}^k : \|T_S^+(X) - \theta\|^2 \leq w^*_\alpha(\|\hat{\theta}\|)\}, \tag{2.9}$$

where $w^*_\alpha(\|\hat{\theta}\|) = \inf\{x \in \mathbb{R} : \mathbb{P}_*(\|T_S^+(X^*) - \hat{\theta}\|^2 \leq x) \geq 1 - \alpha\}$. Here, the conditional density of X^* given X is $f(x - \hat{\theta})$, and $\mathbb{P}_*$ denotes the probability under this conditional distribution. In practice, $w^*_\alpha(\|\hat{\theta}\|)$ is still unavailable explicitly, but we can approximate it to any required degree of accuracy (in probability) by Monte-Carlo simulation. The following algorithm, which first approximates the radius of the bootstrap confidence set at a fixed number of equally spaced points, and then uses linear interpolation to find the radius for the observed value of $\|X\|$, greatly improves the computational efficiency:

(i) Choose $r_{1,\max} \in (0, \infty)$ and $M_1 \in \mathbb{N}$, set $r_{\max} = (r_{1,\max}, 0, \ldots, 0)$ and then set $r_j = j r_{\max}/M_1$ for $j = 0, 1, \ldots, M_1$.

(ii) Choose a large integer B_1 such that $(B_1+1)(1-\alpha)$ is an integer.

(iii) For each $j = 0, 1, \ldots, M_1$ repeat steps (iv) to (vi).

(iv) Generate independent and identically distributed random vectors $X_1^*, \ldots, X_{B_1}^*$ with density $f(\|x - r_j\|^2)$.

(v) Compute $U_i = \|T_S^+(X_i^*) - r_j\|^2$ for $i = 1, \ldots, B_1$.

(vi) Estimate $w_\alpha(\|r_j\|)$ by $w_\alpha^*(\|r_j\|) = U_{((B_1+1)(1-\alpha))}$, i.e. the $((B_1+1)(1-\alpha))$th order statistic of $U_1, \ldots, U_{B_1}$.

(vii) Choose $\theta_{1,\max} \in (0, \infty)$ and $M_2 \in \mathbb{N}$, set $\theta_{\max} = (\theta_{1,\max}, 0, \ldots, 0)$ and then set $\theta_j = j\theta_{\max}/M_2$ for $j = 0, 1, \ldots, M_2$.

(viii) Choose a large integer B_2 and an estimator $\|\hat{\theta}\| = \|\hat{\theta}(X)\|$ of $\|\theta\|$.

(ix) For each $j = 0, 1, \ldots, M_2$ repeat steps (x) to (xiii).

(x) Generate independent and identically distributed random vectors $X_1, \ldots, X_{B_2}$ with density $f(\|x - \theta_j\|^2)$.

(xi) Compute $V_i = \|T_S^+(X_i) - \theta_j\|^2$ and $\|\hat{\theta}_i\| = \|\hat{\theta}(X_i)\|$ for $i = 1, \ldots, B_2$.

(xii) For each $i = 1, \ldots, B_2$, find an integer s such that $r_s \leq \|\hat{\theta}_i\| < r_{s+1}$ and approximate $w_\alpha(\|\theta_j\|)$ by $w_\alpha^*(\|\hat{\theta}_i\|)$, which is obtained by linear interpolation between $w_\alpha^*(r_s)$ and $w_\alpha^*(r_{s+1})$.

(xiii) Approximate $\mathbb{P}_{\theta_j}\left(\|T_S^+(X) - \theta_j\|^2 \leq w_\alpha^*(\|\theta_j\|)\right)$ by $B_2^{-1} \sum_{i=1}^{B_2} \mathbb{1}_{\{V_i \leq w_\alpha^*(\|\hat{\theta}_i\|)\}}$.

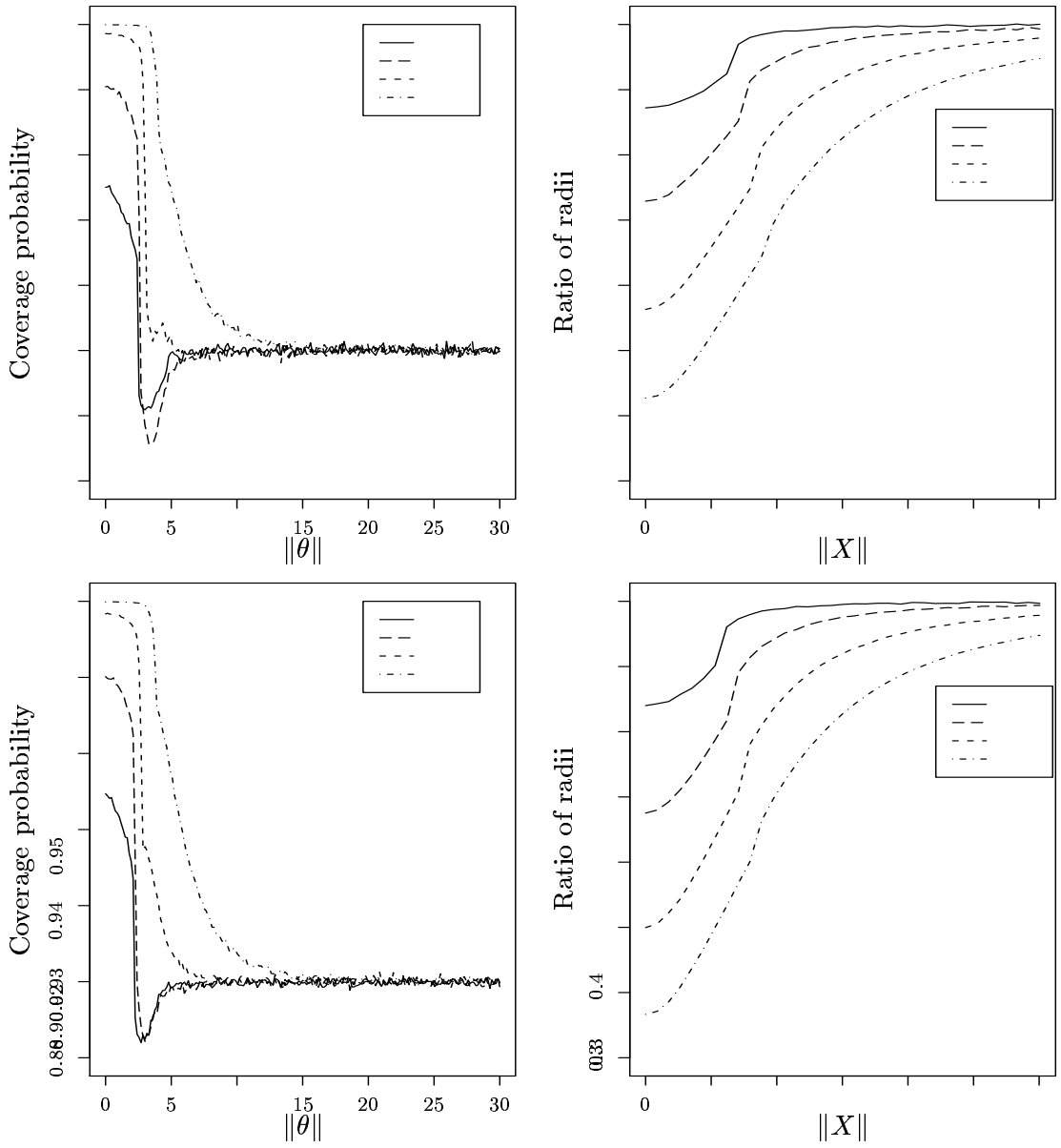
It is possible to generate random vectors from many spherically symmetric distributions as

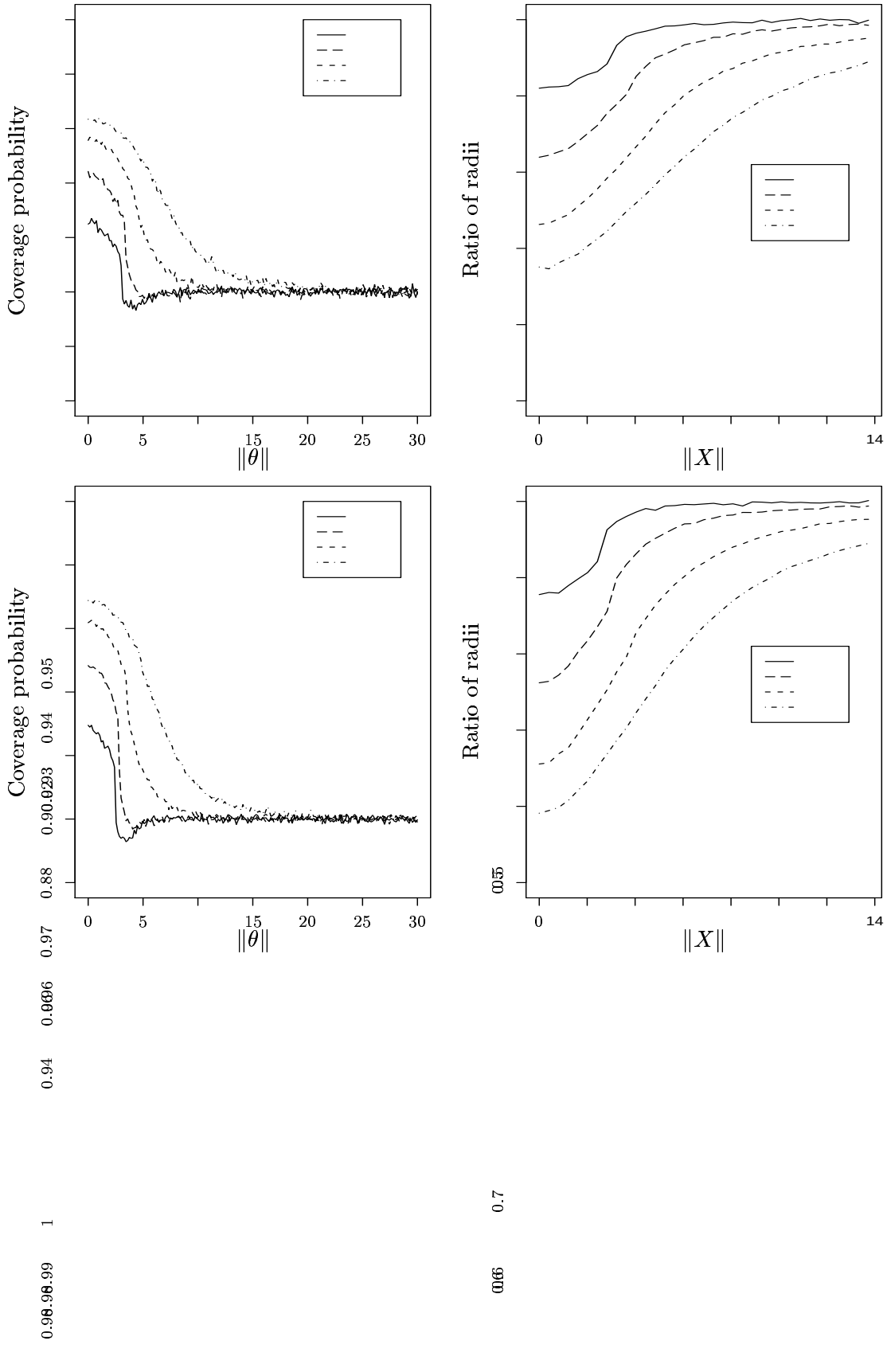
$$X = RU + \theta,$$

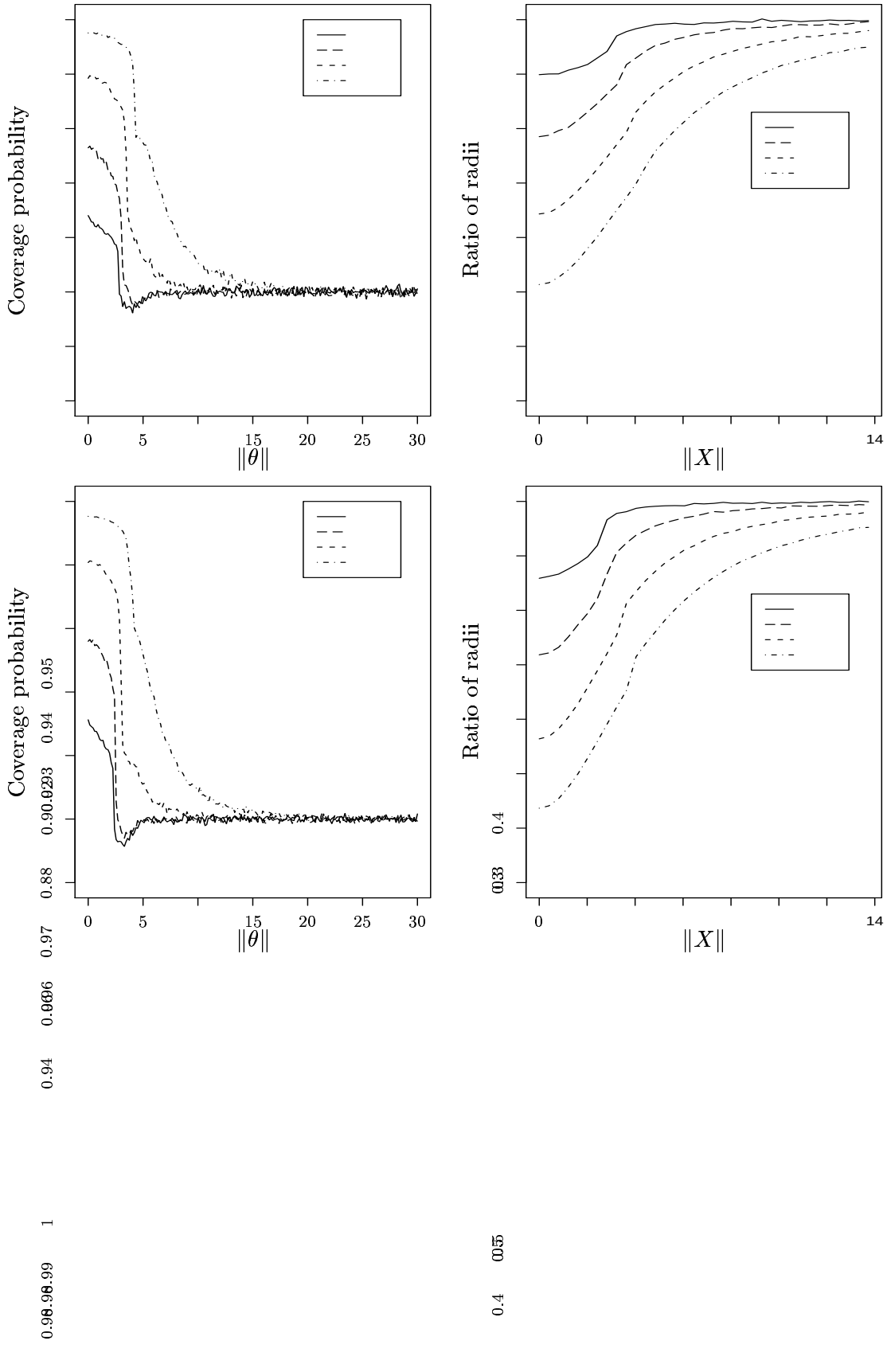
where R has the same density as $\|X\|$, and U is independent of R and has a uniform distribution on the unit sphere $S = \{x \in \mathbb{R}^k : \|x\| = 1\}$. It follows that R has density proportional to $r^{k-1}f(r^2)$ (Fang, Kotz and Ng, 1989, p. 35), while U has the same distribution as $Y/\|Y\|$, where $Y \sim N_k(0, I)$. For the double exponential distribution with parameter d, we have $R \sim \Gamma(k, d)$. We can simulate random vectors from a multivariate t distribution with N degrees of freedom as follows: generate $Z \sim n/\chi^2_n$, and, conditional on Z, generate $X \sim N_k(0, ZI)$ (Zellner, 1976).

The results of simulating the coverage probabilities of the bootstrap confidence sets are given in Figures 2.4, 2.5 and 2.6, for the same distributions that were considered in Figures 2.1, 2.2 and 2.3, and using $\|\hat{\theta}\| = \|X\|$.

The coverage probabilities and radii exhibit many of the same features as those of the analytic confidence set (2.3) for small $\|\theta\|$ and $\|X\|$ respectively. However, we find that it is possible to achieve an even smaller radius for larger $\|X\|$ by bootstrapping, while retaining coverage probability at the nominal level. Of course, it is much harder to prove any results concerning the properties of the bootstrap confidence set, such as those presented in Section 2.4 for the analytic confidence set, as the radius is given in a less explicit form. Nevertheless, Beran (1995) has studied the large k asymptotics of similar bootstrap confidence sets centred at the positive-part Stein estimator in the multivariate normal case, using a different approach involving a geometrical risk criterion as well as coverage probability. Beran obtains the radii for his confidence sets in a different way, however, and his simulation results suggest greater undercoverage problems, which persist for larger values of k.


Coverage probability
0.88
0.90
0.92
0.93
0.94
0.95
0
5
15
20
25
30
$\|\theta\|$
Ratio of radii
0.3
0.4
0
$\|X\|$


Coverage probability
0.88
0.90
0.92
0.94
0.95
0.97
0.99
1
0
5
15
20
25
30
$\|\theta\|$
Ratio of radii
0.55
0.6
0.7
0
14
$\|X\|$


Coverage probability
0.88
0.90
0.92
0.94
0.95
0.97
0.99
1
0
5
15
20
25
30
$\|\theta\|$
Ratio of radii
0.4
0.5
0
14
$\|X\|$

2.5.1 The unknown scale factor case

Recall the linear model (2.1) introduced in Section 2.1:

$$\underset{n\times 1}{X} = \underset{n\times k}{A}\ \underset{k\times 1}{\theta} + \sigma\ \underset{n\times 1}{\epsilon}$$

Throughout this chapter, we have assumed that the scale factor, σ, is known, so that without loss of generality we were able to take $\sigma = 1$. In practice, however, σ^2 is usually unknown, but can be estimated from the data. The canonical model is where $Z = (X^T, Y^T)^T$ has a $(k+\nu)$-dimensional spherically symmetric density with location parameter $\theta' = (\theta^T, 0^T)^T$ and covariance matrix $\sigma^2 I_{k+\nu}$. Here, $I_{k+\nu}$ denotes the $(k+\nu)\times(k+\nu)$ identity matrix, and, despite the increased dimension, we write the density of Z as $f_{\sigma^2}(\|z-\theta'\|)^2$. The appropriate version of the positive-part Stein estimator in this set-up is

$$T_S^+(X,Y) = \left(1 - \frac{a\|Y\|^2/\nu}{\|X\|^2}\right)_+ X.$$

In the multivariate normal case,

$$\frac{\|Y\|^2}{\nu} \sim \frac{\sigma^2}{\nu}\chi^2_\nu,$$

and is independent of X; James and Stein (1961) showed that $a = \nu(k-2)/(\nu+2)$ is the optimal choice with respect to quadratic loss for a point estimate of θ. Analytic theory for confidence sets is very difficult when σ^2 is unknown (though we expect a good approximation to the known-σ^2 case in the limit as $\nu \to \infty$). Bootstrapping, however, remains a viable possibility, and in Table 2.3 we present some coverage probabilities of the confidence set

$$\left\{\theta \in \mathbb{R}^k : \|T_S^+(X,Y) - \theta\|^2 \le w_\alpha^*(\|X\|, \|Y\|)\right\}, \tag{2.10}$$

where

		$\|\theta\|/\sigma$								
		0	1	2	3	4	8	12	16	20
	100	0.996	0.996	0.994	0.974	0.944	0.938	0.936	0.937	0.937
ν	1000	0.998	0.998	0.997	0.974	0.953	0.949	0.948	0.950	0.948
	10000	0.999	0.998	0.997	0.975	0.953	0.950	0.950	0.950	0.950

Table 2.3: Coverage probabilities of the confidence set (2.10) in the k-variate normal case. Parameter values: $\alpha = 0.05$, $a = \nu(k-2)/(\nu+2)$, $k = 10$.

$$w_\alpha^*(\|X\|, \|Y\|) = \inf\{x \in \mathbb{R} : \mathbb{P}_*(\|T_S^+(X^*, Y^*) - X\|^2 \leq x) \geq 1 - \alpha\}.$$

Here, the conditional density of (X^*, Y^*) given (X, Y) is $f_{\|Y\|^2/\nu}(\|z - X'\|)$, where $X' = (X^T, 0^T)^T$, and $\mathbb{P}_*$ denotes the corresponding probability measure. The usual confidence set in this situation is

$$\left\{\theta \in \mathbb{R}^k : \|X - \theta\|^2 \leq \frac{k}{\nu}\|Y\|^2 F_\alpha(k, \nu)\right\}, \tag{2.11}$$

where $F_\alpha(k, \nu)$ is the upper α-point of an F-distribution with k and ν degrees of freedom. This is an exact $(1-\alpha)$-level confidence set, since it follows from Theorem 11 of Kelker (1970) that

$$\frac{\|X - \theta\|^2/k}{\|Y\|^2/\nu}$$

has an F-distribution with k and ν degrees of freedom, regardless of the spherically symmetric distribution. Table 2.4 gives the ratios of the radii of (2.10) to the corresponding radii of (2.11). We find that it is possible to achieve similar gains in volume to those observed in Section 2.5 for the known covariance matrix case, but that ν needs to be very large before undercoverage ceases to be a problem.

		$\|X\|$							
		0	1	2	3	4	6	10	20
$\frac{\|Y\|^2}{\nu\sigma^2}$	0.6	0.57	0.61	0.69	0.82	0.87	0.93	0.96	0.97
	0.8	0.57	0.60	0.67	0.77	0.85	0.91	0.95	0.97
	1	0.57	0.60	0.66	0.72	0.83	0.90	0.94	0.97
	1.2	0.57	0.59	0.65	0.70	0.80	0.88	0.94	0.97
	1.4	0.57	0.59	0.63	0.69	0.77	0.87	0.94	0.96

Table 2.4: The ratio of the radii of the confidence set (2.10) to the corresponding radii of (2.11) in the k-variate normal case. Parameter values: $\alpha = 0.05$, $k = 10$, $a = \nu(k-2)/(\nu+2)$, $\nu = 100$.

2.6 Comments and generalisations

We have seen that the confidence sets (2.3), (2.9) and (2.10) successfully harness the power of the positive-part Stein estimator to produce confidence sets which can be much smaller than the usual set $C^0(X)$, while still maintaining adequate coverage probability. Unfortunately, we were unable to provide a proof that the confidence set (2.3) strictly dominates $C^0(X)$ in terms of coverage probability for sufficiently large k.

The assumption of a spherically symmetric density may be generalised as follows. If $Y = \mu + B^T X$, where $\mu \in \mathbb{R}^m$, B is a $k \times m$ matrix with $B^T B = \Sigma$ having rank k and X has a k-dimensional spherically symmetric density f about the origin, we say Y has an m-dimensional elliptically symmetric distribution with parameters μ and Σ. If $m = k$, it follows from Theorem 2.16 of Fang, Kotz and Ng (1989) that $\Sigma^{-1/2}Y$ has spherically symmetric density f about $\theta = \Sigma^{-1/2}\mu$, so here the problem reduces to the spherically symmetric case provided Σ is known. In particular, if

$Y \sim N_k(\mu, \sigma^2\Sigma)$, where σ is an unknown scale factor and Σ is a $k \times k$ known, positive definite matrix, then the transformed vector $X = \Sigma^{-1/2}Y$ satisfies $X \sim N_k(\theta, \sigma^2 I_k)$, where $\theta = \Sigma^{-1/2}\mu$. As an application, consider a one-way analysis of variance with k cells and n_i observations in cell i, for $i = 1, \ldots, k$. If Y denotes the vector of cell means, then Σ is a diagonal matrix with diagonal entries $n_1^{-1}, \ldots, n_k^{-1}$.

One extension which is especially important for our work is the choice of origin for the Stein estimator. Both (2.3) and (2.9) only represent a significant improvement over $C^0(X)$ if θ is reasonably near the origin. If a prior estimate of θ, say θ_0, is available, then one should redefine the positive-part Stein estimator as

$$T_S^+(X) = \theta_0 + \left(1 - \frac{a}{\|X - \theta_0\|^2}\right)_+ (X - \theta_0),$$

and replace the radius function $v^2(\|X\|)$ by $v^2(\|X - \theta_0\|)$. The region of greatest improvement is then near θ_0, so our confidence sets will perform particularly well if the prior guess is nearly correct.

2.7 Appendix

In this section we give the proofs omitted in the main text. To aid exposition, we return to the more general estimators of θ of the form $\gamma(\|X\|)X$, which were introduced in Section 2.3. Throughout, we assume γ satisfies the following conditions:

(i) $\gamma(r) \geq 0$ for all $r \in [0, \infty)$;

(ii) γ is non-decreasing, and there exists $r_0 \in [0, \infty)$ such that $\gamma(r)$ is strictly increasing for $r \geq r_0$;

(iii) $\gamma(r)$ is twice continuously differentiable for $r > r_0$.

Observe that the positive-part Stein estimator defined in (2.2) satisfies these conditions with $r_0 = a^{1/2}$, while the ordinary Stein estimator does not. We continue to write $w_\alpha(\|\theta\|)$ for the upper α-point of the sampling distribution of $\|\gamma(\|X\|)X - \theta\|^2$. The next two lemmas are needed to compute the first two non-zero terms in the Taylor series of $w_\alpha(0)$ in Lemma 2.7.3.

Suppose $\{F(x,\lambda) : \lambda \in \Lambda \subseteq \mathbb{R}\}$ is a non-empty family of distribution functions on the real line. We think of $F(x,\lambda)$ as $\mathbb{P}_\theta(\|T_S^+(X) - \theta\|^2 \leq x)$ when $\|\theta\| = \lambda$. We say that $F(x,\lambda)$ is strictly increasing in x at (x_0,λ) if $F(x,\lambda) < F(x_0,\lambda)$ for all $x < x_0$ and $F(x,\lambda) > F(x_0,\lambda)$ for all $x > x_0$.

Lemma 2.7.1. *Let $\alpha \in (0,1)$ and $w_\alpha(\lambda) = \inf\{x \in \mathbb{R} : F(x,\lambda) \geq 1-\alpha\}$. Fix $\lambda_0 \in \Lambda$ and let $x_0 = w_\alpha(\lambda_0)$.*

(i) Suppose that $F(x,\lambda)$ is continuous in λ at (x,λ_0) for all x in some neighbourhood of x_0, and that $F(x,\lambda_0)$ is strictly increasing in x at (x_0,λ_0). Then $w_\alpha(\lambda)$ is continuous in λ at λ_0.

(ii) Suppose that both of the partial derivatives $\partial F/\partial x$ and $\partial F/\partial\lambda$ exist in some neighbourhood of (x_0,λ_0) and are (jointly) continuous at (x_0,λ_0), and that $\frac{\partial F}{\partial x}(x_0,\lambda_0) > 0$. Then $w_\alpha(\lambda)$ is differentiable with respect to λ at λ_0, and this derivative is given by

$$w'_\alpha(\lambda_0) = \frac{-\frac{\partial F}{\partial \lambda}(x_0,\lambda_0)}{\frac{\partial F}{\partial x}(x_0,\lambda_0)}.$$

(iii) Suppose that, for some $n \geq 2$, all of the nth order partial derivatives of F exist in some neighbourhood of (x_0,λ_0) and are continuous at (x_0,λ_0), and that $\frac{\partial F}{\partial x}(x_0,\lambda_0) > 0$. Then $w_\alpha(\lambda)$ is n times differentiable with respect to λ at λ_0.

Proof. Choose $\delta > 0$ such that $F(x,\lambda)$ is continuous in λ at (x,λ_0), for all $x \in \mathbb{R}$ such that $|x - x_0| \leq \delta$. Let $\epsilon_1 = F(x_0+\delta,\lambda_0) - (1-\alpha)$, let $\epsilon_2 = (1-\alpha) - F(x_0-\delta,\lambda_0)$,

and set $\epsilon = \min(\epsilon_1, \epsilon_2)$, so that $\epsilon > 0$.

Now choose $h = h(\epsilon) > 0$ such that both

$$|F(x_0 + \delta, \lambda) - F(x_0 + \delta, \lambda_0)| \leq \epsilon/2 \quad \text{and} \quad |F(x_0 - \delta, \lambda) - F(x_0 - \delta, \lambda_0)| \leq \epsilon/2$$

for all $\lambda \in \Lambda$ such that $|\lambda - \lambda_0| \leq h$. For such λ, we have

$$\begin{aligned} F(x_0 - \delta, \lambda) &\leq F(x_0 - \delta, \lambda_0) + |F(x_0 - \delta, \lambda) - F(x_0 - \delta, \lambda_0)| \\ &\leq 1 - \alpha - \epsilon_2 + \epsilon/2 \\ &\leq 1 - \alpha - \epsilon/2, \end{aligned}$$

and similarly $F(x_0 + \delta, \lambda) \geq 1 - \alpha + \epsilon/2$. Hence $w_\alpha(\lambda_0) - \delta \leq w_\alpha(\lambda) \leq w_\alpha(\lambda_0) + \delta$ for all $\lambda \in \Lambda$ such that $|\lambda - \lambda_0| \leq h$, which proves (i).

To prove (ii), we first show that there exists a constant $C > 0$ such that

$$\left| \frac{w_\alpha(\lambda) - w_\alpha(\lambda_0)}{\lambda - \lambda_0} \right| \leq C$$

for all $\lambda \in \Lambda$ with $|\lambda - \lambda_0|$ sufficiently small. We choose

$$C = \frac{\left|\frac{\partial F}{\partial \lambda}(x_0, \lambda_0)\right|}{\frac{\partial F}{\partial x}(x_0, \lambda_0)} + 1.$$

Given $\epsilon > 0$, choose $h = h(\epsilon) > 0$ such that both

$$\left| \frac{\partial F}{\partial \lambda}(x, \lambda) - \frac{\partial F}{\partial \lambda}(x_0, \lambda_0) \right| \leq \frac{\epsilon}{2}$$

for all $x \in [x_0, x_0 + Ch]$ and $|\lambda - \lambda_0| \leq h$, and

$$\left| \frac{\partial F}{\partial x}(x, \lambda_0) - \frac{\partial F}{\partial x}(x_0, \lambda_0) \right| \leq \frac{\epsilon}{2C}$$

for all $x \in [x_0, x_0 + Ch]$. Then for all $\lambda \in \Lambda$ with $|\lambda - \lambda_0| \leq h$, we have

$$\begin{aligned} &F(x_0 + C|\lambda - \lambda_0|, \lambda) \\ &\qquad \geq F(x_0 + C|\lambda - \lambda_0|, \lambda_0) - |\lambda - \lambda_0| \Big(\Big| \frac{\partial F}{\partial \lambda}(x_0, \lambda_0) \Big| + \frac{\epsilon}{2} \Big) \\ &\qquad \geq 1 - \alpha + C|\lambda - \lambda_0| \Big(\frac{\partial F}{\partial x}(x_0, \lambda_0) - \frac{\epsilon}{2C} \Big) - |\lambda - \lambda_0| \Big(\Big| \frac{\partial F}{\partial \lambda}(x_0, \lambda_0) \Big| + \frac{\epsilon}{2} \Big) \\ &\qquad \geq 1 - \alpha - \epsilon|\lambda - \lambda_0|. \end{aligned}$$

But $\epsilon > 0$ was arbitrary, so $w_\alpha(\lambda) \leq w_\alpha(\lambda_0) + C|\lambda - \lambda_0|$ for all $\lambda \in \Lambda$ with $|\lambda - \lambda_0| \leq h$. Similarly, by reducing $h > 0$ if necessary, we may assume $w_\alpha(\lambda) \geq w_\alpha(\lambda_0) - C|\lambda - \lambda_0|$ for all $\lambda \in \Lambda$ with $|\lambda - \lambda_0| \leq h$.

Since $\frac{\partial F}{\partial x}(x_0, \lambda_0) > 0$ and $\partial F/\partial x$ and $\partial F/\partial \lambda$ are continuous at (x_0, λ_0), for $|\lambda - \lambda_0|$ sufficiently small there is a unique solution in x to the equation $F(x, \lambda) = 1 - \alpha$. This solution satisfies

$$1 - \alpha = F(x_0, \lambda_0) + \frac{\partial F}{\partial x}(x_0, \lambda_0)(x - x_0) + \frac{\partial F}{\partial \lambda}(x_0, \lambda_0)(\lambda - \lambda_0) + o(|x - x_0| + |\lambda - \lambda_0|)$$

as $x \to x_0$ and $\lambda \to \lambda_0$. Thus, given $\epsilon > 0$, there exists $h = h(\epsilon) > 0$ such that for $\lambda \in \Lambda$ with $|\lambda - \lambda_0| \leq h$, we have

$$\left| \frac{x - x_0}{\lambda - \lambda_0} + \frac{\frac{\partial F}{\partial \lambda}(x_0, \lambda_0)}{\frac{\partial F}{\partial x}(x_0, \lambda_0)} \right| \leq \frac{\epsilon(|x - x_0| + |\lambda - \lambda_0|)}{\frac{\partial F}{\partial x}(x_0, \lambda_0)(\lambda - \lambda_0)} \leq \frac{(C + 1)\epsilon}{\frac{\partial F}{\partial x}(x_0, \lambda_0)},$$

which proves (ii).

Part (iii) follows immediately from (ii) by induction and using the chain and quotient rules. □

We now specialise to the case where

$$w_\alpha(\lambda) \equiv w_\alpha(\|\theta\|) = \inf\{x \in \mathbb{R} : \mathbb{P}_\theta(\|\gamma(\|X\|)X - \lambda\|^2 \leq x) \geq 1 - \alpha\}.$$

Lemma 2.7.2. *Suppose that the spherically symmetric density, f, is twice continuously differentiable. If $w_\alpha(0) > 0$, then $w_\alpha(\|\theta\|)$ is twice differentiable with respect to $\|\theta\|$ in a neighbourhood of the origin.*

Proof. The proof involves checking the conditions of Lemma 2.7.1. We will simply show that

$$\frac{\partial}{\partial x}\mathbb{P}_\theta(\|\gamma(\|X\|)X - \theta\|^2 \leq x) \Big|_{(w_\alpha(0),0)} > 0,$$

from which it will be seen that the other conditions follow similarly. For $\|\theta\|^2 < x$, we can write

$$\mathbb{P}_\theta(\|\gamma(\|X\|)X - \theta\|^2 \le x) = \int_{\mathbb{R}^k} f(\|t-\theta\|^2)\mathbb{1}_{\{\|\gamma(\|t\|)t-\theta\|^2 \le x\}}\, dt.$$

The proof of Proposition 2.3.1 shows that we may assume $\theta = (\|\theta\|, 0, \ldots, 0)$. Following Hwang and Casella (1982), we note that the integrand in the expression above depends on t only through $r = \|t\|$ and the angle β between t and θ, defined for $t \neq 0$ and $\|\theta\| > 0$ by

$$\cos\beta = \frac{\theta^T t}{\|\theta\|\|t\|} = \frac{t_1}{\|t\|},$$

where $t = (t_1, \ldots, t_k)$. The angle β may be defined arbitrarily for $t = 0$ or $\|\theta\| = 0$. We therefore transform to these spherical coordinates, with, say $y_3 = t_3, \ldots, y_k = t_k$. The Jacobian determinant of the inverse transformation is $|J| = (r^2 \sin\beta)/t_2$, and we may integrate out $y_3, \ldots, y_k$ to obtain, for $\|\theta\|^2 < x$,

$$\mathbb{P}_\theta(\|\gamma(\|X\|)X-\theta\|^2 \le x) = K\int_0^\pi \int_0^{r_+(\beta,\|\theta\|)} r^{k-1}\sin^{k-2}\beta\, f(r^2-2r\|\theta\|\cos\beta+\|\theta\|^2)\, dr\, d\beta, \tag{2.12}$$

where

$$K = 2\pi\prod_{p=1}^{k-3}\Big(\int_0^\pi \sin^p t\, dt\Big) = \frac{2\pi^{(k-1)/2}}{\Gamma\big((k-1)/2\big)},$$

and $r_+ = r_+(\beta, \|\theta\|)$ is the unique positive solution to

$$\gamma^2(r)r^2 - 2\|\theta\|\cos\beta\gamma(r)r + \|\theta\|^2 - x = 0.$$

The uniqueness of the solution of the above equation is guaranteed by the condition that $\|\theta\|^2 < x$, which ensures that the other root of the quadratic in $\gamma(r)r$ is negative. Thus

$$\gamma(r_+)r_+ = \|\theta\|\cos\beta + \big(x - \|\theta\|^2\sin^2\beta\big)^{1/2} > 0,$$

so in particular $r_+ > a^{1/2}$. To simplify notation, we write (2.12) as

$$\mathbb{P}_\theta(\|\gamma(\|X\|)X - \theta\|^2 \le x) = K\int_0^\pi \int_0^{r_+(\beta,\|\theta\|)} h(r,\beta,\|\theta\|)\, dr\, d\beta.$$

Since the integrand in the above expression is continuous, and since r_+ is differentiable with respect to x, we find for $(x, \|\theta\|)$ in a neighbourhood of $\big(w_\alpha(0), 0\big)$ that

$$\frac{\partial}{\partial x}\mathbb{P}_\theta(\|\gamma(\|X\|)X-\theta\|^2 \le x) = K\int_0^\pi \frac{h(r_+,\beta,\|\theta\|)}{\gamma(r_+)+\gamma'(r_+)r_+}\,\frac{1}{2\big(x-\|\theta\|^2\sin^2\beta\big)^{1/2}}\,d\beta. \tag{2.13}$$

Evaluating (2.13) at $(x, \|\theta\|) = \big(w_\alpha(0), 0\big)$, and observing that $r_+(\beta, 0)$ does not depend on β, gives

$$\begin{aligned}\frac{\partial}{\partial x}\mathbb{P}_\theta(\|\gamma(\|X\|)X-\theta\|^2 \le x)\,\Big|_{(w_\alpha(0),0)} &= \frac{Kr_+^{k-1}f(r_+^2)}{2\big(\gamma(r_+)+\gamma'(r_+)r_+\big)w_\alpha^{1/2}(0)}\int_0^\pi \sin^{k-2}\beta\,d\beta \\ &> 0.\end{aligned}$$

□

We are now in a position to compute the required terms in the Taylor series of $w_\alpha(0)$.

Lemma 2.7.3. *Suppose that $w_\alpha(0) > 0$ and that f is twice continuously differentiable. Then $w'_\alpha(0) = 0$, and*

$$\begin{aligned}\frac{k}{2}w''_\alpha(0) = \frac{f'(r_+^2)}{f(r_+^2)}w_\alpha^{1/2}(0)\Big\{-2\big(\gamma(r_+)+\gamma'(r_+)r_+\big)r_+ + 4r_+ - \frac{2r_+}{\gamma(r_+)+\gamma'(r_+)r_+}\Big\} \\ - \frac{(k-1)w_\alpha^{1/2}(0)}{r_+\big(\gamma(r_+)+\gamma'(r_+)r_+\big)} + \frac{w_\alpha^{1/2}(0)\big(2\gamma'(r_+)+\gamma''(r_+)r_+\big)}{\big(\gamma(r_+)+\gamma'(r_+)r_+\big)^2} + k - 1,\end{aligned} \tag{2.14}$$

where r_+ is the unique positive solution to $\gamma(r)r = w_\alpha^{1/2}(0)$.

Proof. Transforming to spherical polar coordinates as in the proof of Lemma 2.7.2 gives, for $\|\theta\| < w_\alpha^{1/2}(\|\theta\|)$,

$$K\int_0^\pi\int_0^{r_+(\beta,\|\theta\|)} h(r,\beta,\|\theta\|)\,dr d\beta = 1-\alpha, \tag{2.15}$$

where $r_+ = r_+(\beta, \|\theta\|)$ satisfies

$$\gamma(r_+)r_+ = \|\theta\| \cos\beta + \big(w_\alpha(\|\theta\|) - \|\theta\|^2 \sin^2\beta\big)^{1/2},$$

and $h(r, \beta, \|\theta\|)$ is the integrand in (2.12). By Lemma 2.7.2, for sufficiently small $\|\theta\|$, we may differentiate (2.15) with respect to $\|\theta\|$ to obtain

$$-2\int_0^\pi \int_0^{r_+(\beta,\|\theta\|)} (r\cos\beta - \|\theta\|) r^{k-1} \sin^{k-2}\beta\, f'(r^2 - 2\|\theta\| r\cos\beta + \|\theta\|^2)\, dr d\beta \\ + \int_0^\pi h(r_+, \beta, \|\theta\|) \frac{\partial r_+}{\partial \|\theta\|}\, d\beta = 0, \tag{2.16}$$

where

$$\big(\gamma(r_+) + \gamma'(r_+)r_+\big)\frac{\partial r_+}{\partial\|\theta\|} = \cos\beta + \frac{w'_\alpha(\|\theta\|) - 2\|\theta\|\sin^2\beta}{2\big(w_\alpha(\|\theta\|) - \|\theta\|^2\sin^2\beta\big)^{1/2}}. \tag{2.17}$$

Evaluating (2.16) at $\|\theta\| = 0$ gives

$$-2\int_0^\pi \int_0^{r_+} r^k \cos\beta \sin^{k-2}\beta\, f'(r^2)\, dr d\beta \\ + \frac{r_+^{k-1} f(r_+^2)}{\gamma(r_+) + \gamma'(r_+)r_+} \int_0^\pi \sin^{k-2}\beta \left\{\cos\beta + \frac{w'_\alpha(0)}{2w_\alpha^{1/2}(0)}\right\} d\beta = 0,$$

where we have used the fact that $r_+ = r_+(\beta, 0)$, which is the positive solution to $\gamma(r)r = w_\alpha^{1/2}(0)$, does not depend on β. But $\sin^{k-2}\beta$ is symmetric about $\beta = \pi/2$, while $\cos\beta$ is anti-symmetric about $\beta = \pi/2$, so $w'_\alpha(0) = 0$.

Differentiating again, we find

$$2\int_0^\pi \int_0^{r_+(\beta,\|\theta\|)} r^{k-1}\sin^{k-2}\beta\, f'(r^2 - 2\|\theta\| r\cos\beta + \|\theta\|^2)\, dr d\beta \\ + 4\int_0^\pi \int_0^{r_+(\beta,\|\theta\|)} (r\cos\beta - \|\theta\|)^2 r^{k-1}\sin^{k-2}\beta\, f''(r^2 - 2\|\theta\| r\cos\beta + \|\theta\|^2)\, dr d\beta \\ - 4\int_0^\pi (r_+\cos\beta - \|\theta\|) r_+^{k-1}\sin^{k-2}\beta \frac{\partial r_+}{\partial\|\theta\|} f'(r^2 - 2\|\theta\| r\cos\beta + \|\theta\|^2)\, d\beta \\ + \int_0^\pi \left\{\frac{\partial^2 r_+}{\partial\|\theta\|^2} h(r_+, \beta, \|\theta\|) + \frac{\partial h}{\partial r_+}\left(\frac{\partial r_+}{\partial\|\theta\|}\right)^2\right\} d\beta = 0. \tag{2.18}$$

We obtain $\partial^2 r_+/\partial\|\theta\|^2$ by differentiating (2.17):

$$\big(2\gamma'(r_+)+\gamma''(r_+)r_+\big)\left(\frac{\partial r_+}{\partial\|\theta\|}\right)^2+\big(\gamma(r_+)+\gamma'(r_+)r_+\big)\frac{\partial^2 r_+}{\partial\|\theta\|^2}$$
$$=\frac{2\big(w''_\alpha(\|\theta\|)-2\sin^2\beta\big)\big(w_\alpha(\|\theta\|)-\|\theta\|^2\sin^2\beta\big)-\big(w'_\alpha(\|\theta\|)-2\|\theta\|\sin^2\beta\big)^2}{4\big(w_\alpha(\|\theta\|)-\|\theta\|^2\sin^2\beta\big)^{3/2}},$$

and also note that

$$\frac{\partial h}{\partial r_+}=(k-1)r_+^{k-2}\sin^{k-2}\beta\, f(r_+^2-2\|\theta\|r_+\cos\beta+\|\theta\|^2)$$
$$+2(r_+-\|\theta\|\cos\beta)r_+^{k-1}\sin^{k-2}\beta\, f'(r_+^2-2\|\theta\|r_+\cos\beta+\|\theta\|^2).$$

We are now in a position to evaluate (2.18) at $\|\theta\|=0$ to obtain

$$\begin{aligned}0&=2\int_0^\pi\int_0^{r_+}\big\{2f''(r^2)r^2\cos^2\beta+f'(r^2)\big\}r^{k-1}\sin^{k-2}\beta\,dr\,d\beta\\
&\quad-\frac{4r_+^k f'(r_+^2)}{\gamma(r_+)+\gamma'(r_+)r_+}\int_0^\pi\cos^2\beta\sin^{k-2}\beta\,d\beta\\
&\quad+\frac{r_+^{k-1}f(r_+^2)}{\gamma(r_+)+\gamma'(r_+)r_+}\int_0^\pi\sin^{k-2}\beta\Big\{\frac{w''_\alpha(0)-2\sin^2\beta}{2w_\alpha^{1/2}(0)}-\frac{\big(2\gamma'(r_+)+\gamma''(r_+)r_+\big)\cos^2\beta}{\big(\gamma(r_+)+\gamma'(r_+)r_+\big)^2}\Big\}\,d\beta\\
&\quad+\frac{r_+^{k-2}}{\big(\gamma(r_+)+\gamma'(r_+)r_+\big)^2}\big\{2r_+^2f'(r_+^2)+(k-1)f(r_+^2)\big\}\int_0^\pi\cos^2\beta\sin^{k-2}\beta\,d\beta\\
&\equiv J_1+J_2+J_3+J_4,\end{aligned}$$

say. To evaluate J_1, we convert temporarily back to Cartesian coordinates. Recalling that in the spherical transformation, $t_1=r\cos\beta$, we have

$$\begin{aligned}J_1&=\frac{2}{K}\int_{\mathbb{R}^k}\big\{2f''(\|t\|^2)t_1^2+f'(\|t\|^2)\big\}\,dt=\frac{2}{K}\int_{\mathbb{R}^k}\Big\{2f''(\|t\|^2)\frac{\|t\|^2}{k}+f'(\|t\|^2)\Big\}\,dt\\
&=2\int_0^\pi\int_0^{r_+}\Big\{2f''(r^2)\frac{r^2}{k}+f'(r^2)\Big\}r^{k-1}\sin^{k-2}\beta\,dr\,d\beta=\frac{2}{k}f'(r_+^2)r_+^kI_{k-2},\end{aligned}$$

where $I_k=\int_0^\pi\sin^k\beta\,d\beta$, and the last equality follows on integrating the first term by

parts. Moreover,

$$J_2 = -\frac{4r_+^k f'(r_+^2)}{\gamma(r_+) + \gamma'(r_+)r_+}(I_{k-2} - I_k)$$

$$J_3 = \frac{r_+^{k-1} f(r_+^2)}{\gamma(r_+) + \gamma'(r_+)r_+}\left\{ \frac{w_\alpha''(0) I_{k-2}}{2w_\alpha^{1/2}(0)} - \frac{I_k}{w_\alpha^{1/2}(0)} - \frac{\big(2\gamma'(r_+) + \gamma''(r_+)r_+\big)(I_{k-2} - I_k)}{\big(\gamma(r_+) + \gamma'(r_+)r_+\big)^2} \right\}$$

$$J_4 = \frac{r_+^{k-2}}{\big(\gamma(r_+) + \gamma'(r_+)r_+\big)^2}\big(2r_+^2 f'(r_+^2) + (k-1)f(r_+^2)\big)(I_{k-2} - I_k).$$

Putting this all together, and noting that $I_k/I_{k-2} = (k-1)/k$, gives the required expression after some simplification. □

The expression in Lemma 2.7.3 is still rather complicated. As we saw in the statement of Theorem 2.3.2, however, the terms simplify considerably when we restrict attention to the positive-part Stein estimator.

Proof of **Theorem 2.3.2**.
For $r > a^{1/2}$, we have $\gamma(r) + \gamma'(r)r = (1 + a/r^2)$ and $2\gamma'(r) + \gamma''(r)r = -2a/r^3$. Moreover, if Z has density f, then

$$1 - \alpha = \mathbb{P}(\|Z\|^2 \leq c^2) = \mathbb{P}(\gamma^2(\|Z\|)\|Z\|^2 \leq \gamma^2(c)c^2),$$

so $w_\alpha(0) = (c - a/c)^2$. Lastly, $r_+ = r_+(\beta, 0)$ satisfies $r_+\gamma(r_+) = w_\alpha^{1/2}(0)$, so that $r_+ = c > a^{1/2}$. Substituting these expressions into (2.14) gives the result. □

2.7.1 Proofs of the properties of the analytic confidence set

Proof of **Lemma 2.4.1**.
To prove (i), recall from Theorem 2.3.2 that

$$\frac{1}{2}w_\alpha''(0) = \frac{1}{k}\Big(1 - \frac{a}{c^2}\Big)\Big(\frac{a(k-1)}{c^2 + a} - \frac{2ac^2}{(c^2+a)^2} - \frac{2a^2}{c^2+a}\frac{f'(c^2)}{f(c^2)}\Big) + \frac{a(k-1)}{c^2 k} \qquad (2.19)$$

and that the condition on α means that $c^2 > a$. Hence, for $a \in (0, k-1]$,

$$\frac{1}{2}w''_\alpha(0) \leq \frac{1}{k}\Big(1 - \frac{a}{c^2}\Big)\Big(\frac{k-1+a}{2}\Big) + \frac{a(k-1)}{c^2 k} \leq \frac{k-1}{k}.$$

To prove (ii), we note that

$$\frac{1}{2}w''_\alpha(0) \geq \frac{1}{k}\Big(1 - \frac{a}{c^2}\Big)\Big(\frac{a(k-3)c^2}{(c^2+a)^2}\Big) + \frac{a(k-1)}{c^2 k} \geq \frac{a(k-1)}{c^2 k}.$$

□

The next two results are slightly more general than the versions stated in Section 2.4. They specialise to cover Proposition 2.4.2 and Lemma 2.4.3 respectively, on replacing $\gamma(\|x\|)x$ by $T_S^+(x)$, and $v^2(r)$ by its expression in (2.3).

Proposition 2.7.4. *Let $\theta \in \mathbb{R}^k$, $a > 0$ and $\gamma(r) = (1 - a/r^2)_+$. Suppose that $v : [0, \infty) \to [0, \infty)$ is a non-decreasing function with the property that there exists $r^{**} \in (a^{1/2}, \infty]$ such that $v^2(r)$ is twice differentiable for $r \in (a^{1/2}, r^{**})$ with*

$$\frac{d^2 v^2}{dr^2} < 2 + \frac{6a^2}{r^4} + \frac{4a\|\theta\|}{r^3}$$

for all $r \in (a^{1/2}, r^{**})$, and $v^2(r)$ is constant for $r \geq r^{**}$. Then the set

$$C(\theta) = \{x \in \mathbb{R}^k : \|\gamma(\|x\|)x - \theta\|^2 \leq v^2(\|x\|)\}$$

is connected. In particular, if $d^2v^2/dr^2 < 2$ for all $r \in (a^{1/2}, r^{**})$, then $C(\theta)$ is connected for every $\theta \in \mathbb{R}^k$.

Proof. By Theorem 3.1 of Casella and Hwang (1983), the set $C(\theta)$ is connected if and only if the set

$$S(\|\theta\|) = \big\{r \in [0, \infty) : \big(\gamma(r)r - \|\theta\|\big)^2 \leq v^2(r)\big\}$$

is an interval. For $\|\theta\| > 0$, let $r^*(\|\theta\|)$ denote the positive root of $\gamma(r)r = \|\theta\|$ and define $r^*(0) = a^{1/2}$. Let

$$S_1 = S_1(\|\theta\|) = \big\{r \in [0,\infty) : \big(\gamma(r)r - \|\theta\|\big)^2 \le v^2(r),\ r \le r^*(\|\theta\|)\big\}$$
$$S_2 = S_2(\|\theta\|) = \big\{r \in [0,\infty) : \big(\gamma(r)r - \|\theta\|\big)^2 \le v^2(r),\ r \ge r^*(\|\theta\|)\big\}.$$

Since $r^*(\|\theta\|) \in S_1 \cap S_2$, it follows that $S(\|\theta\|) = S_1 \cup S_2$ is an interval if and only if both S_1 and S_2 are intervals.

For $r \le r^*(\|\theta\|)$, the function $f(r) = (\gamma(r)r - \|\theta\|)^2 - v^2(r)$ is decreasing in r, which proves that S_1 is an interval. If $r \in \big(r^*(\|\theta\|), r^{**}\big)$, we have

$$f'(r) = 2\big(\gamma(r) + \gamma'(r)r\big)\big(\gamma(r)r - \|\theta\|\big) - \frac{dv^2(r)}{dr},$$

so that

$$\begin{aligned} f''(r) &= 2\big(\gamma(r) + \gamma'(r)r\big)^2 + 2\big(2\gamma'(r) + \gamma''(r)r\big)\big(\gamma(r)r - \|\theta\|\big) - \frac{d^2v^2(r)}{dr^2} \\ &= 2\Big(1 + \frac{a}{r^2}\Big)^2 - \frac{4a}{r^3}\Big(r - \frac{a}{r} - \|\theta\|\Big) - \frac{d^2v^2(r)}{dr^2} \\ &= 2 + \frac{6a^2}{r^4} + \frac{4a\|\theta\|}{r^3} - \frac{d^2v^2(r)}{dr^2} \\ &> 0. \end{aligned}$$

Since $f\big(r^*(\|\theta\|)\big) \le 0$, there can therefore be at most one root to the equation $f(r) = 0$ for $r \in \big(r^*(\|\theta\|), r^{**}\big)$. For $r \ge r^{**}$, we have that $f(r)$ is increasing, and this proves that S_2 is an interval, as required. □

Lemma 2.7.5. *Assume the hypotheses of Proposition 2.7.4, including the requirement that $d^2v^2/dr^2 < 2$ for all $r \in (a^{1/2}, r^{**})$, and also assume that $dv^2/dr < 2r$ for all $r \in (a^{1/2}, r^{**})$. For $\|\theta\| \le v(0)$, if $x \in C(\theta)$, then so is $t\,x$, for all $t \in [0,1]$.*

Proof. Observe that $x \in C(\theta)$ if and only if $f(r) \leq 0$, where

$$\begin{aligned} f(r) &= \gamma^2(r)r^2 - 2\|\theta\|\gamma(r)r\cos\beta + \|\theta\|^2 - v^2(r) \\ &= (\gamma(r)r - \|\theta\|\cos\beta)^2 + \|\theta\|^2\sin^2\beta - v^2(r), \end{aligned}$$

where $r = \|x\|$, and where, for $r > 0$ and $\|\theta\| > 0$, we have $\cos\beta = x^T\theta/(\|x\|\|\theta\|)$. The angle β may be defined arbitrarily for $\|x\| = 0$ or $\|\theta\| = 0$. Since $f(0) \leq 0$ by hypothesis, it suffices to show that f has at most one non-negative root for each $\beta \in [0, \pi]$. Let $r^*(\|\theta\|)$ denote the positive root of $\gamma(r)r = \|\theta\|\cos\beta$ if it exists, and otherwise set $r^*(\|\theta\|) = a^{1/2}$. Then $f(r)$ is decreasing for $r \leq r^*(\|\theta\|)$ and if $r \in \big(r^*(\|\theta\|), r^{**}\big)$, we have

$$\begin{aligned} f''(r) &= 2 - \frac{d^2v^2(r)}{dr^2} + \frac{6a^2}{r^4} + \frac{4a\|\theta\|\cos\beta}{r^3} \\ &> \frac{6a^2}{r^4} + \frac{4a\|\theta\|\cos\beta}{r^3}. \end{aligned}$$

Thus f is strictly convex if $\beta \in [0, \pi/2]$, or if $\|\theta\| = 0$, or if $\beta \in (\pi/2, \pi]$, $\|\theta\| > 0$ and $r \leq -3a/(2\|\theta\|\cos\beta)$. In the remaining case where $\beta \in (\pi/2, \pi]$, $\|\theta\| > 0$ and $r > -3a/(2\|\theta\|\cos\beta)$, we have

$$\begin{aligned} f'(r) &= 2r - \frac{dv^2(r)}{dr} - \frac{2a^2}{r^3} - 2\Big(1 + \frac{a}{r^2}\Big)\|\theta\|\cos\beta \\ &> -\frac{2a^2}{r^3} - \frac{2a\|\theta\|\cos\beta}{r^2}. \end{aligned}$$

But this final expression is positive for $r > -a/(\|\theta\|\cos\beta)$. Finally, f is increasing for $r > r^{**}$, and the result follows. □

Proof of **Theorem 2.4.4**.

In view of Lemma 2.4.3, it suffices to show that the boundary $\partial C^0(\theta)$ of $C^0(\theta)$ lies inside $C(\theta)$. Suppose $x \in \partial C^0(\theta)$, and define r, β as in the proof of Lemma 2.7.5. Then

$$r = \|\theta\|\cos\beta + \big(c^2 - \|\theta\|^2\sin^2\beta\big)^{1/2}.$$

In their Theorem 2.1, Hwang and Casella (1982) prove that $C^0(\theta)$ is contained in the set

$$\{x \in \mathbb{R}^k : \|T_S^+(x) - \theta\|^2 \leq c^2\},$$

for $\|\theta\| \leq c$, and the result is trivial if either $\|\theta\| = 0$ or $r \leq a^{1/2}$. Therefore, for $0 < \|\theta\| \leq w_\alpha(0)$ and $r > a^{1/2}$, we let

$$\begin{aligned} f(\beta, \|\theta\|) &= \gamma^2(r)r^2 - 2\|\theta\|\gamma(r)r\cos\beta + \|\theta\|^2 - w_\alpha(0) - \frac{1}{2}w''_\alpha(0)r^2 \\ &= -\frac{a^2}{c^2} + \frac{a^2}{r^2} + \frac{2a\|\theta\|\cos\beta}{r} - \frac{1}{2}w''_\alpha(0)r^2 \\ &= -\frac{a^2}{c^2} + \frac{a^2}{r^2} + \frac{a(r^2 + \|\theta\|^2 - c^2)}{r^2} - \frac{1}{2}w''_\alpha(0)r^2, \end{aligned}$$

where $r = \|\theta\|\cos\beta + \left(c^2 - \|\theta\|^2\sin^2\beta\right)^{1/2}$, so that it is enough to show $f(\beta, \|\theta\|) \leq 0$ for all $\beta \in [0, \pi]$ and $\|\theta\|$ in the given range.

Since $\partial r/\partial\beta = -\|\theta\| r \sin\beta/(c^2 - \|\theta\|^2\sin^2\beta)^{1/2}$, we find

$$\frac{\partial f}{\partial\beta} = \frac{-2\|\theta\|\sin\beta}{\left(c^2 - \|\theta\|^2\sin^2\beta\right)^{1/2}}\left\{\frac{a(c^2 - \|\theta\|^2 - a)}{r^2} - \frac{1}{2}w''_\alpha(0)r^2\right\},$$

from which we deduce that f has turning points at $\beta = 0, \pi$ and possibly at β^*, where

$$\left\{\|\theta\|\cos\beta^* + \left(c^2 - \|\theta\|^2\sin^2\beta^*\right)^{1/2}\right\}^4 = \frac{2a(c^2 - \|\theta\|^2 - a)}{w''_\alpha(0)}.$$

Since r is a decreasing function of $\beta \in [0, \pi]$, a solution to this last equation exists if and only if

$$(c - \|\theta\|)^4 \leq \frac{2a(c^2 - \|\theta\|^2 - a)}{w''_\alpha(0)} \leq (c + \|\theta\|)^4.$$

Observe first that

$$f(\pi, \|\theta\|) = \left(c - \frac{a}{c - \|\theta\|}\right)^2 - \left(c - \frac{a}{c}\right)^2 - \frac{1}{2}w''_\alpha(0)(c - \|\theta\|)^2 \leq 0.$$

Next,

$$\begin{aligned} f(0,\|\theta\|) &\le -\frac{a^2}{c^2} + \frac{a^2}{(c+\|\theta\|)^2} + \frac{2a\|\theta\|}{c+\|\theta\|} - \frac{a(k-1)}{c^2k}(c+\|\theta\|)^2 \\ &\le \frac{1}{c^2(c+\|\theta\|)^2}\left\{\|\theta\|\Big(2ac^3 - \frac{4a(k-1)c^3}{k}\Big) + \|\theta\|^2\Big(2ac^2 - \frac{6ac^2(k-1)}{k}\Big)\right\} \\ &\le 0. \end{aligned}$$

Finally,

$$\begin{aligned} f(\beta^*,\|\theta\|) &= \frac{1}{r^2}\left\{a\Big(1-\frac{a}{c^2}\Big)\Big(\frac{2a(c^2-\|\theta\|^2-a)}{w''_\alpha(0)}\Big)^{1/2} - 2a(c^2-\|\theta\|^2-a)\right\} \\ &= \frac{a(c^2-\|\theta\|^2-a)^{1/2}}{r^2}\left\{\Big(1-\frac{a}{c^2}\Big)\Big(\frac{2a}{w''_\alpha(0)}\Big)^{1/2} - 2(c^2-\|\theta\|^2-a)^{1/2}\right\}. \end{aligned}$$

Thus we find $f(\beta^*,\|\theta\|)$ is non-positive for

$$\|\theta\|^2 \le c^2 - a - \Big(1-\frac{a}{c^2}\Big)^2\frac{a}{2w''_\alpha(0)} = \frac{c^2-a}{2w''_\alpha(0)c^4}\Big\{2w''_\alpha(0)c^4 - (c^2-a)a\Big\}.$$

□

Chapter 3

The bagged nearest-neighbour classifier

3.1 Introduction

Suppose we observe data pairs $\mathcal{L} = \{(X_i, Y_i) : i = 1, \ldots, n\}$, in which $X_1, \ldots, X_n$ are the independent variables, or inputs, and $Y_1, \ldots, Y_n$ are the dependent variables, or responses. In the language of machine learning, $\mathcal{L}$ is often called a training, or learning, set. On the basis of $\mathcal{L}$, a *predictor* assigns the response of an arbitrary input x. When the response is a class variable, i.e. each response belongs to finite set of unordered elements, a predictor is referred to as a *classifier*.

Classification problems have a long history, dating at least from the famous Iris data example of Fisher (1936). A medical application is described in Breiman, Friedman, Olshen and Stone (1984) as follows:

> At the University of California, San Diego Medical Center, when a heart

> attack patient is admitted, 19 variables are measured during the first 24 hours. These include blood pressure, age, and 17 other ordered and binary variables summarizing the medical symptoms considered as important indicators of the patient's condition.
>
> The goal of a recent medical study ... was the development of a method to identify high risk patients (those who will not survive at least 30 days) on the basis of the initial 24-hour data.

Hand (1981) gives details of classification problems in archaeology, agriculture, speech recognition and cardiac wave analysis, amongst many others. Apart from the breadth of application, however, it is important to note that in many cases the data to be classified may take values in high-dimensional spaces, or even, as in the last two examples above, infinite-dimensional function spaces. In such settings, techniques assuming the existence of densities for the underlying populations may not be viable.

Nearest-neighbour methods are one of the oldest approaches to classification, beginning with the work of Fix and Hodges (1951). Nevertheless, they are constantly being adapted to new settings – see, for example, Kuncheva and Bezdek (1998) and Mollineda, Ferri and Vidal (2000). A major attraction of nearest-neighbour classifiers is their simplicity. For implementation they require only a measure of distance on the sample space, along with training samples; hence their popularity as a starting-point for refinement and improvement.

Bagging is a means of improving the performance of a classifier (or, more generally, a predictor) by combining the results of many empirically simulated predictions. Breiman (1996) introduced the technique, and also coined the term, which is derived from **b**ootstrap **agg**regat**ing**. Bagging classifiers can sometimes, although not always, reduce the error rate (Bay, 1999). An overview of bagging and related techniques,

such as boosting, can be found in the recent book by Hastie, Tibshirani and Friedman (2001).

In this chapter we show that, in circumstances where the relative densities of two populations can be meaningfully defined, a bagged nearest-neighbour classifier can converge, as the training-sample sizes increase, to the optimal Bayes classifier. However, we obtain this limit if and only if the simulated training samples are of asymptotically negligible size relative to the respective actual training samples, and if the simulated training-sample sizes diverge together with the actual training-sample sizes. If, as is commonly the case in practice, the simulated training-sample sizes are the same as those of the actual training samples, then the nearest-neighbour classifier does not converge to the Bayes classifier. These results apply to both with- and without-replacement bagging, which are the two approaches most commonly used in practice.

The extent of the improvements can be determined by probability calculations, which we discuss theoretically and illustrate numerically. It is shown, for example, that in the case of with-replacement bagging, the resample size should be at most 69% of the original sample size if the bagged nearest-neighbour classifier is to asymptotically improve on the performance of its unbagged counterpart. The ceiling is reduced to 50% in the case of without-replacement bagging.

These results are of interest for two reasons. Firstly, because the majority of bagging experiments employ relatively large resample sizes; much of the evidence against the performance of bagged nearest-neighbour classifiers (e.g. Breiman, 1996; Bay, 1999) is for full-size resamples. Secondly, because in most statistical work, the performance of classifiers is discussed in settings where population densities are well-defined and estimable. However, to construct the Bayes classifier only the relative densities are needed, weighted by the prior probabilities of each class. Indeed, the optimal Bayes

classifier depends only on whether the weighted density ratio is greater than 1 or less than 1, rather than its exact value. We show that our result about convergence of the nearest-neighbour classifier to the Bayes rule can be set up in this context, making it relevant to a relatively general class of problems.

Friedman and Hall (2000) and Buja and Stuetzle (2000a,b) have carried out investigations into the theoretical properties of bagging in contexts other than classification problems. In Friedman and Hall (2000), smooth estimators are decomposed into linear and higher order parts and it is argued that bagging reduces the variance of the nonlinear component, without affecting the linear part. Buja and Stuetzle (2000a) concentrate on U-statistics, and show that bagging has a second-order effect on variance, squared bias and mean squared error; Buja and Stuetzle (2000b) examine statistical functionals and give the von Mises expansions of the bagged functionals.

Bühlmann and Yu (2002) do discuss the performance of bagging in prediction problems, including bagging with smaller resample sizes than the original sample sizes, but their focus is on decision trees (Hastie, Tibshirani and Friedman, 2001, pp. 266–279) rather than nearest-neighbour methods. Their theoretical results show that bagging smoothes such hard decisions, reducing the variance and mean squared error, and they provide simulations to suggest it can also reduce the error rate, or risk, in classification.

The literature on nonparametric methods of classification is also extensive, including techniques which converge to the Bayes rule. An example in the univariate case is the approach of Stoller (1954), which is based on the empirical distribution function. In multivariate settings, methods founded on nonparametric density estimation are sometimes used; these are discussed by Hand (1981). Marron (1983) finds such a classification rule whose risk converges to the Bayes risk at the optimal rate (which depends on the dimension of the inputs and the smoothness of the underlying den-

sities). In the univariate case, Cover (1968) showed that the risk of the nearest-neighbour classifier converges to its limit at rate m^{-2}, where m is the sample size, under smoothness conditions. The limit here generally exceeds the risk of the Bayes classifier. Fukunaga and Hummels (1987) gave a heuristic argument to suggest a rate of $m^{-2/d}$ in d-variate settings, and this was later proved, again under smoothness assumptions, by Psaltis, Snapp and Venkatesh (1994).

The rest of this chapter is organised as follows. In Section 3.2 we introduce the Bayes, nearest-neighbour and bagged nearest-neighbour classifiers, define the risk of a classifier and give an informal derivation of the large-sample risk of the nearest-neighbour classifier using a marked Poisson process argument. This argument is developed in Section 3.3 to study the asymptotic risk of the bagged nearest-neighbour classifier. In these two sections we work with Euclidean data, and assume the existence of population densities, which aids intuition. Section 3.4 shows how to extend the ideas of Sections 3.2 and 3.3 to cover cases where only relative densities exist. This section also contains a statement of the main theorem, Theorem 3.4.1, concerning the convergence of the bagged nearest-neighbour classifier to the Bayes classifier, under these relaxed assumptions and in the case where the simulated training-sample sizes are asymptotically negligible in comparison with the actual training-sample sizes. The problem of choosing the simulated training-sample size is studied in Section 3.5 using the technique of cross-validation. Finally, the Appendix in Section 3.6 is devoted to formal justification of the heuristic methods of Sections 3.2 and 3.3, and the proof of Theorem 3.4.1.

3.2 Definitions of classifiers, and basic properties

3.2.1 The nearest-neighbour, bagged nearest-neighbour and Bayes classifiers

Assume there are two populations, Π_X and Π_Y, from which we have random samples $\mathcal{X}$ and $\mathcal{Y}$. Suppose $\mathcal{X}$ is of size m and $\mathcal{Y}$ is of size n. The nearest-neighbour classifier, based on $\mathcal{X}$ and $\mathcal{Y}$, assigns a new data value, z, to Π_X or Π_Y according as z is nearest to an element of $\mathcal{X}$ or $\mathcal{Y}$, respectively. Now draw resamples $\mathcal{X}^*$ and $\mathcal{Y}^*$, of sizes $m_1 \le m$ and $n_1 \le n$, by resampling randomly, with or without replacement, from $\mathcal{X}$ and $\mathcal{Y}$ respectively. The bagged version of the nearest-neighbour classifier allocates z to Π_X if the nearest-neighbour classifier, based on independent realisations of $\mathcal{X}^*$ and $\mathcal{Y}^*$, assigns it more often to Π_X than to Π_Y. For the sake of definiteness we shall always treat the version of the bagged nearest-neighbour classifier which uses an infinite number of simulations in the 'majority-vote' step.

Breiman (1996) provides the following simple explanation as to why bagging the nearest-neighbour classifier with full-size resamples does not change its performance. The probability that $\mathcal{X}^* \cup \mathcal{Y}^*$ contains the nearest-neighbour to an arbitrary z is at least $1 - 1/e \approx 0.632 > 1/2$. It follows that, with probability one, the bagged nearest-neighbour classifier agrees with its unbagged counterpart. By reducing the resample sizes, however, we can ensure the probability that the nearest-neighbour belongs to $\mathcal{X}^* \cup \mathcal{Y}^*$ falls below $1/2$. It is in these circumstances that we may expect bagging to improve the nearest-neighbour classifier, through its ability to 'explore' more of the data in the vicinity of z.

Of course, nearest-neighbour classification requires a measure of distance, so we would generally assume that the data in $\mathcal{X}$ and $\mathcal{Y}$ take values in a space on which a met-

ric, or norm, has been defined. We will exploit this generality in Section 3.4 and Theorem 3.4.1, but for the remainder of this section we work with data in $\mathbb{R}^k$, and assume that the populations Π_X and Π_Y have densities f and g respectively, which are continuous almost everywhere.

If the prior probabilities of the populations Π_X and Π_Y are p and $1-p$ respectively, where $p \in (0,1)$, then the 'ideal' Bayes classifier $\mathcal{C}_{\text{Bayes}}$ assigns z to Π_X or Π_Y according as $p\,f(z) - (1-p)\,g(z)$ is positive or negative. Equivalently, z is assigned to Π_X if the probability

$$q(z) = \frac{p\,f(z)}{p\,f(z) + (1-p)\,g(z)} \tag{3.1}$$

exceeds $1/2$, and to Π_Y if $q(z) < 1/2$. We may choose to classify z arbitrarily if $p\,f(z) = (1-p)\,g(z)$.

3.2.2 Error rates of Bayes and nearest-neighbour classifiers

We write, for example, $\{\mathcal{C}(z) = X\}$ for the event that a general classifier $\mathcal{C}$ assigns the data value z to the population Π_X. The average error rate, or risk, of the classifier $\mathcal{C}$ is defined by

$$\text{Risk}(\mathcal{C}) = p \int_{\mathbb{R}^k} \mathbb{P}\big(\mathcal{C}(z) = Y\big)\, f(z)\, dz + (1-p) \int_{\mathbb{R}^k} \mathbb{P}\big(\mathcal{C}(z) = X\big)\, g(z)\, dz.$$

In the language of decision theory, $\text{Risk}(\mathcal{C})$ is the Bayes risk of the decision rule $\mathcal{C}$ with respect to 0-1 loss and the prior which places probabilities p and $1-p$ on Π_X and Π_Y respectively. The Bayes classifier defined in Section 3.2.1 above derives its name from the fact that it is the Bayes decision rule for this problem. It is only an ideal classifier, because in practice f and g are unknown.

The risk of the Bayes classifier is

$$\begin{aligned}\mathrm{Risk}(\mathcal{C}_{\mathrm{Bayes}}) &= p \int_{\mathbb{R}^k} \mathbb{1}_{\{pf(z)<(1-p)g(z)\}}\, f(z)\, dz + (1-p) \int_{\mathbb{R}^k} \mathbb{1}_{\{pf(z)>(1-p)g(z)\}}\, g(z)\, dz \\ &= \int_{\mathbb{R}^k} \min\{p\, f(z), (1-p)\, g(z)\}\, dz. \end{aligned} \tag{3.2}$$

A heuristic derivation of the large-sample limit of risk of the nearest-neighbour classifier $\mathcal{C}_{\mathrm{NN}}$ can be deduced by a point-process approximation, as follows. Provided $m/(m+n) \to p$ as $m \to \infty$, the distribution of data in the neighbourhood of z converges to that of a marked Poisson process, $\mathcal{P}$, in which each point has one of two marks, chosen independently of the marks for all other points (see e.g. Kingman, 1993, Chapter 5). The probability that the mark at a given point of $\mathcal{P}$ equals X is the same as the proportion of points from Π_X, relative to points from either Π_X or Π_Y, that occur in an infinitesimal neighbourhood of z. That is, it equals $q(z)$, defined at (3.1), and likewise the probability that the mark at the given point of $\mathcal{P}$ is Y, equals $1 - q(z)$. It follows that the point in $\mathcal{P}$ that is nearest to z will be of type X with probability $q(z)$. Hence, as $m \to \infty$,

$$\begin{aligned}\mathrm{Risk}(\mathcal{C}_{\mathrm{NN}}) &\to p \int_{\mathbb{R}^k} \{1 - q(z)\}\, f(z)\, dz + (1-p) \int_{\mathbb{R}^k} q(z)\, g(z)\, dz \\ &= 2\, p\, (1-p) \int_{\mathbb{R}^k} \frac{f(z)\, g(z)}{p\, f(z) + (1-p)\, g(z)}\, dz. \end{aligned} \tag{3.3}$$

Observe that if $p\, f(z) \geq (1-p)\, g(z)$, then

$$\frac{2\, p\, (1-p)\, f(z)\, g(z)}{p\, f(z) + (1-p)\, g(z)} \geq \frac{p\, (1-p)\, f(z)\, g(z) + (1-p)^2\, g(z)^2}{p\, f(z) + (1-p)\, g(z)} = (1-p)\, g(z),$$

so that

$$\lim_{m \to \infty} \mathrm{Risk}(\mathcal{C}_{\mathrm{NN}}) \geq \mathrm{Risk}(\mathcal{C}_{\mathrm{Bayes}}).$$

The inequality is strict except in the pathological case where $p = 1/2$ and $f = g$ almost everywhere.

The expression (3.3) for the asymptotic risk of the nearest-neighbour classifier has been known since the work of Cover and Hart (1967), who derived it using a different method. We include the derivation above because the marked Poisson process argument will allow us to deduce the asymptotic risk of the bagged nearest-neighbour classifier in the next section. The argument above is formalised in the Section 3.6.1.

3.3 The bagged nearest-neighbour classifier

In this section we again work informally, postponing rigorous argument to Section 3.6.2. Recall that bagging involves drawing resamples $\mathcal{X}^*$ uniformly at random from $\mathcal{X}$, and $\mathcal{Y}^*$ uniformly at random from $\mathcal{Y}$. The resample sizes are m_1 and n_1, respectively. Both with- and without-replacement resampling are possible, so that in the with-replacement resampling case, $\mathcal{X}^*$ and $\mathcal{Y}^*$ may contain repeats. Recall the marked Poisson process, $\mathcal{P}$, introduced in Section 3.2.2. It represents a large-sample approximation to the data in the neighbourhood of z. The mark of each point in $\mathcal{P}$ is either X or Y, and the nearest-neighbour classifier allocates z to the population indicated by the mark of the point in $\mathcal{P}$ that is nearest to z.

The bagged nearest-neighbour classifier $\mathcal{C}_{\text{Bagg}}$ applies the majority-vote rule to the outcomes of the nearest-neighbour classifier based on independent resamples $\mathcal{P}^*$ drawn from $\mathcal{P}$. In the case of with-replacement resampling, $\mathcal{P}^*$ will typically involve repeated data, but the number of repeats of any given data value is not used by the bagged classifier. Therefore we may disregard repeats, and view $\mathcal{P}^*$ as simply a randomly chosen subset of $\mathcal{P}$.

Hence we may consider $\mathcal{P}^*$ as having been obtained from $\mathcal{P}$ by the point-process operation of 'thinning'. That is, to produce $\mathcal{P}^*$, each point in $\mathcal{P}$ is 'killed' with a

certain probability ρ and 'kept' with probability $1-\rho$, where $0 \le \rho \le 1$, independently of all other points. This is true for either with-replacement or without-replacement resampling.

The value of ρ depends on the large-sample limit of the 'sampling ratios' m_1/m and n_1/n, and on whether the resampling is done with or without replacement. In order to make this statement more precise, we now assume $m/(m+n) \to p \in (0,1)$ as $m \to \infty$, and study the case where m_1/m and n_1/n both converge to ℓ as $m \to \infty$. Non-degenerate results are obtained when the limit $\ell \in [0,1]$ for with-replacement resampling, and $\ell \in [0,1)$ for without-replacement resampling. We also suppose that either with-replacement bagging, or without-replacement bagging, is used throughout; we do not, for example, resample with-replacement for one type of data and without-replacement for the other.

With these assumptions, the thinning probability ρ is identical for type X and type Y data. For with-replacement resampling, $\rho = e^{-\ell}$, reflecting the fact that the limiting probability that a resample of size m_1 excludes d specified data points in $\mathcal{X}$ is

$$\lim_{m\to\infty} \Bigl(1 - \frac{d}{m}\Bigr)^{m_1} = \rho^d .$$

If resampling is without replacement then we take $\rho = 1 - \ell$. To appreciate why, note that the limiting probability that a resample of size m_1, drawn from $\mathcal{X}$ without replacement, excludes d specified data points in $\mathcal{X}$, is

$$\lim_{m\to\infty} \Bigl(1-\frac{d}{m}\Bigr) \Bigl(1-\frac{d}{m-1}\Bigr) \cdots \Bigl(1-\frac{d}{m-m_1+1}\Bigr)$$
$$= \lim_{m\to\infty} \Bigl(1-\frac{m_1}{m}\Bigr) \Bigl(1-\frac{m_1}{m-1}\Bigr) \cdots \Bigl(1-\frac{m_1}{m-d+1}\Bigr) = \rho^d .$$

Let $T_j = T_j(z)$ denote the type, either X or Y, of the point in $\mathcal{P}$ that is jth nearest to z. In either the with- or without-replacement case, if $\rho < 1$, it follows from the

definition of the thinned point process $\mathcal{P}^*$ that

$$\begin{aligned}\pi(\mathcal{P}) &\equiv \mathbb{P}\big(\text{the point in } \mathcal{P}^* \text{ that is nearest to } z, \text{ has mark } X \,\big|\, \mathcal{P}\big) \\ &= \sum_{j=1}^{\infty} \rho^{j-1}\,(1-\rho)\,\mathbb{1}_{\{T_j = X\}}.\end{aligned}$$

The marks of the points of $\mathcal{P}$ are independent and identically distributed as X or Y, the former having the probability $q(z)$ defined in (3.1). Thus the random variables $I_j = \mathbb{1}_{\{T_j = X\}}$ are independent and identically distributed, taking the value 1 with probability $q(z)$ and 0 with probability $1 - q(z)$.

Suppose we generate $\mathcal{P}^*$ a total of B independent times, on each occasion starting from the same $\mathcal{P}$. If $\pi(\mathcal{P}) \neq 1/2$, then by the weak law of large numbers, the conditional probability, given $\mathcal{P}$, that for the majority of resampled point processes $\mathcal{P}^*$ the nearest point in $\mathcal{P}^*$ to z has mark X, converges to $\mathbb{1}_{\{\pi(\mathcal{P}) > 1/2\}}$ as $B \to \infty$. Therefore, since $\mathbb{P}\big(\pi(\mathcal{P}) = 1/2\big) = 0$, and since we always treat the infinite-simulation case of bagging, we have that as $m \to \infty$,

$$\begin{aligned}\mathbb{P}\big(\mathcal{C}_{\text{Bagg}}(z) = X\big) &\to \mathbb{P}\bigg(\sum_{j=1}^{\infty} \rho^{j-1}\,(1-\rho)\,I_j > \frac{1}{2}\bigg) \\ &\equiv P\big(\rho, q(z)\big)\,,\end{aligned} \tag{3.4}$$

say. It also follows that

$$\text{Risk}\big(\mathcal{C}_{\text{Bagg}}\big) \to p \int_{\mathbb{R}^k} \big\{1 - P\big(\rho, q(z)\big)\big\}\, f(z)\, dz + (1-p) \int_{\mathbb{R}^k} P\big(\rho, q(z)\big)\, g(z)\, dz, \tag{3.5}$$

as $m \to \infty$.

The value of $P\big(\rho, q(z)\big)$ when $\rho = 1$, corresponding to $\ell = 0$ under either resampling scheme, is defined by taking the limit as $\rho \nearrow 1$ in (3.4). We have

$$\mathbb{E}\bigg(\sum_{j=1}^{\infty} \rho^{j-1}\,(1-\rho)\,I_j\bigg) = q(z)$$

and

$$\begin{aligned}\mathrm{Var}\Big(\sum_{j=1}^{\infty} \rho^{j-1}\,(1-\rho)\,I_j\Big) &= q(z)\big(1-q(z)\big)(1-\rho)^2 \sum_{j=1}^{\infty} \rho^{2(j-1)} \\ &= q(z)\big(1-q(z)\big)\Big(\frac{1-\rho}{1+\rho}\Big) \\ &\to 0\end{aligned}$$

as $\rho \nearrow 1$. Hence, by Chebychev's inequality,

$$P\big(\rho, q(z)\big) \to \mathbb{1}_{\{q(z)>1/2\}}$$

as $\rho \nearrow 1$, provided $q(z) \neq 1/2$, so we take $P\big(1, q(z)\big) = \mathbb{1}_{\{q(z)>1/2\}}$ if $q(z) \neq 1/2$.

In order for the bagged nearest-neighbour classifier to converge to the Bayes classifier, it is necessary and sufficient that the probability at (3.4) should equal 1 if $q(z) > 1/2$, and equal 0 if $q(z) < 1/2$; see Section 3.2.1. From (3.4) and the argument in the previous paragraph, we see that this property holds if and only if $\rho = 1$; that is, if and only if m_1/m and n_1/n both converge to 0. Provided this constraint holds, the risk of the bagged nearest-neighbour classifier converges to the risk of the Bayes classifier, defined at (3.2). This is also the limit, as $\rho \nearrow 1$, of the risk at (3.5) above.

A number of properties can be deduced from (3.4). For example, if $\rho \in [0, 1/2]$ then

$$P\big(\rho, q(z)\big) = P(I_1 = 1) = q(z).$$

It follows that the asymptotic risk of the bagged nearest-neighbour classifier is the same as that for the regular nearest-neighbour classifier if $\rho \in [0, 1/2]$, and is generally reduced if $\rho > 1/2$. Since, in the cases of with- and without-replacement bagging, the respective values of ρ are the limits of $e^{-m_1/m}$ and $1 - (m_1/m)$, bagging the nearest-neighbour classifier asymptotically improves performance if $m_1 < m \log 2 \approx 0.69\, m$ in the with-replacement case, and if $m_1 < m/2$ in the without-replacement case, but not otherwise. Therefore, reducing the sampling ratio, m_1/m, does not immediately

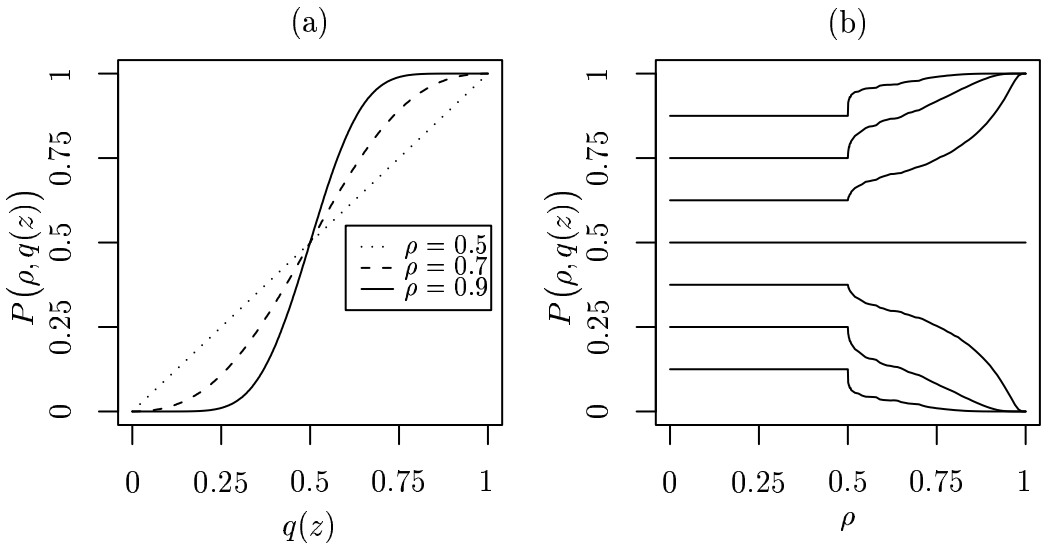


Figure 3.1: Panel (a) plots $P\big(\rho, q(z)\big)$ against $q(z)$ for fixed ρ; in (b), $P\big(\rho, q(z)\big)$ is plotted against ρ, with $q(z)$ ranging from 1/8 (bottom) to 7/8 (top) in steps of 1/8.

lead to a reduction in risk; it has to be reduced below the threshold, 0.69 or 0.5, in the cases of with- or without-replacement bagging, respectively.

3.3.1 Numerical studies

Figure 3.1(a) shows $P\big(\rho, q(z)\big)$ as a function of $q(z)$, for $\rho = 0.5$, 0.7 and 0.9. For $\rho = 0.5$, this corresponds to the probability that the nearest-neighbour classifier assigns z to Π_X, namely $q(z)$. The other two curves were obtained by simulation, and show the convergence of the large-sample approximation of the bagged nearest-neighbour classifier to the Bayes classifier defined in the last paragraph of Section 3.2.1. Figure 3.1(b) gives the complementary plot of $P\big(\rho, q(z)\big)$ as a function of ρ, for $q(z) = \frac{1}{8}\,(\frac{1}{8})\,\frac{7}{8}$. Contrary, perhaps, to first appearances, $P\big(\rho, q(z)\big)$ is

continuous at $\rho = 1/2$ for each value of $q(z)$. To see this, define

$$p_i(r) = \mathbb{P}\Big(\sum_{j=i}^{\infty} \rho^{j-1}\,(1-\rho)\,I_j > r\Big),$$

where, as above, $I_1, I_2, \ldots$ are independent and identically distributed Bernoulli$\big(q(z)\big)$ random variables. Then, for $\rho \in [1/2, 1)$, we have

$$\begin{aligned} p_1(1/2) &= q(z)\,p_2(\rho - 1/2) + \big(1 - q(z)\big)\,p_2(1/2) \\ &= q(z)\,p_1\Big(\frac{\rho - 1/2}{\rho}\Big) + \big(1 - q(z)\big)\,p_1\Big(\frac{1}{2\rho}\Big) \\ &\to q(z) \end{aligned}$$

as $\rho \searrow 1/2$.

To demonstrate the asymptotic improvement in risk of the bagged nearest-neighbour classifier over its nearest-neighbour counterpart, we study two examples. In the first, we choose $f(z) = \phi(z)$, the standard normal density, and set $g(z) = \phi(z - \mu)$, for $\mu \in [0, 3]$. In Figure 3.2(a) we plot the asymptotic risk of the bagged nearest-neighbour classifier, given by (3.5), as a function of μ, for $\rho = 0.5$, 0.7, 0.9 and 1. Recall that the cases $\rho = 0.5$ and $\rho = 1$ correspond to the asymptotic risks of the nearest-neighbour classifier, given at (3.3), and the Bayes classifier, given at (3.2), respectively. The asymptotic risks in these cases are found by numerical integration of (3.3) and (3.2) respectively, while simulation can be used in the other two cases. In this problem, the graphs of the functions $y = f(z)$ and $y = g(z)$ cross at only one point, so we expect classification to be relatively straightforward, provided μ is not too small. In the second of these two examples we choose f to have the mixture density

$$f(z) = \frac{1}{5}\sum_{i=0}^{4} \phi(z - 4\,i)\,, \tag{3.6}$$

and set $g(z) = f(z-\mu)$ for $\mu \in [0, 2]$. For these versions of f and g, Figure 3.2(b) again plots the asymptotic risk of the bagged nearest-neighbour classifier as a function of μ,

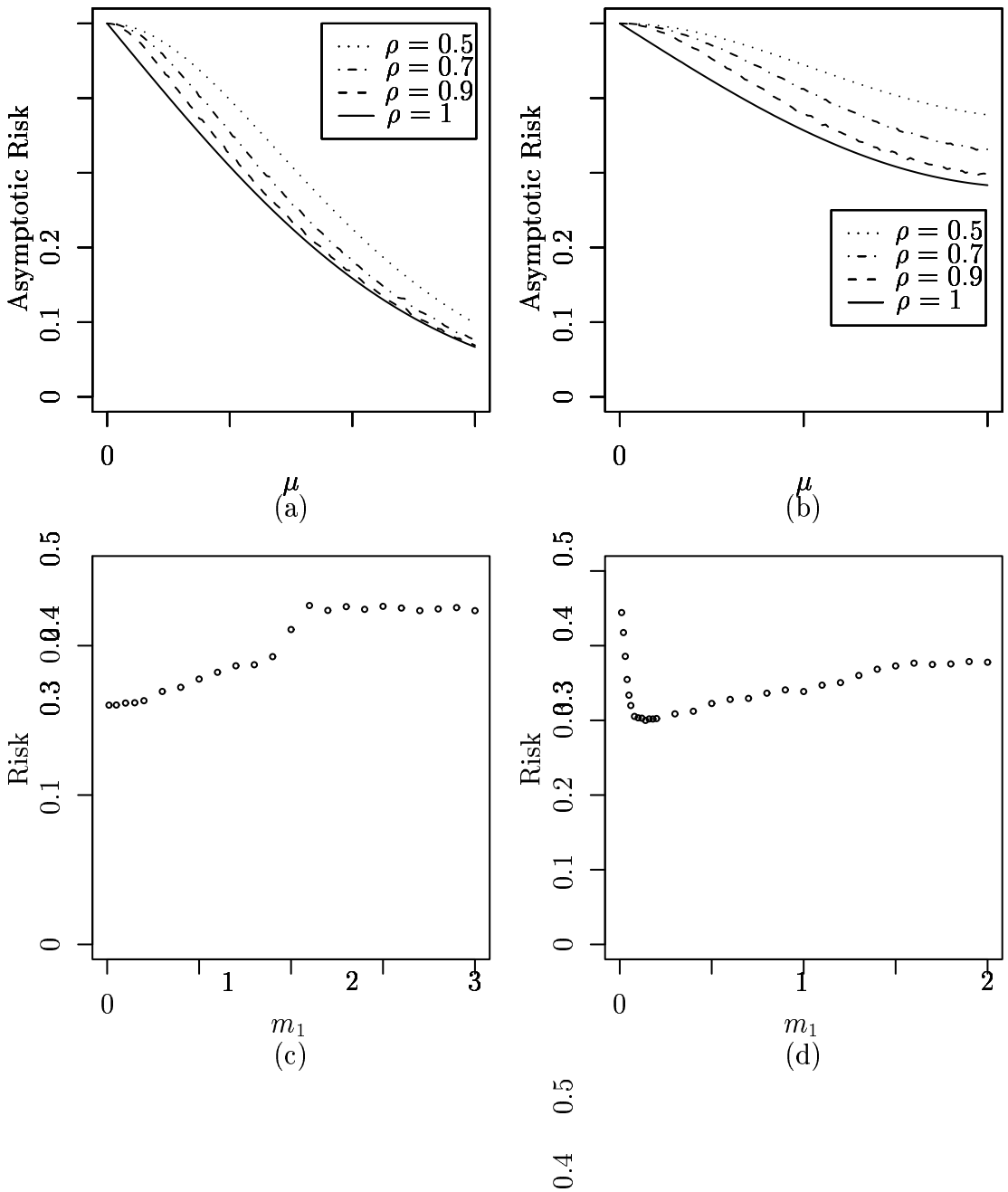


Figure 3.2: Panels (a) and (b) show the asymptotic risk of the bagged nearest-

for the same values of ρ. Here the densities cross at nine points, and classification is therefore much more difficult. In both examples, we observe that the bagged nearest-neighbour classifier can provide considerable asymptotic improvement in risk over the nearest-neighbour classifier.

To ascertain the extent of the possible improvement in practice, we studied both examples above with $\mu = 2$ and simulated training samples of size $m = n = 200$. The risks of the bagged nearest-neighbour classifier as functions of m_1 are given in panels (c) and (d) of Figure 3.2. Resampling was done without-replacement in Figure 3.2(c), and with-replacement in Figure 3.2(d), in order to show that the risk becomes flat at different points, specifically at $m_1 \approx 0.5\,m$ and $m_1 \approx 0.7\,m$ in panels (c) and (d) respectively. Behaviour for relatively small m_1 is virtually identical for either resampling scheme. Two further points are worthy of note: (i) in both examples, the optimal choice of m_1 is much smaller than m; and (ii) while the optimal choice appears to be $m_1 = 1$ in the first example, this is a poor choice in the more complicated second example. There, $m_1 = 14$ is optimal.

3.4 Relative densities

One of the major attractions of classifiers based on nearest-neighbour methods is that they only require a notion of distance on the sample space. This opens up the possibility of studying the bagged nearest-neighbour classifier in general circumstances, extending beyond Euclidean-data models. Up to this point, we have assumed the existence of densities f and g for the populations Π_X and Π_Y. However, standard definitions of f and g, requiring differentiable distribution functions, are not meaningful in some cases, for example when the generic elements X and Y of $\mathcal{X}$ and $\mathcal{Y}$ respectively are random functions. Significant progress can nevertheless be made un-

der less restrictive assumptions, involving little more than the existence of relative densities rather than actual densities, and continuity, as a function of ball radius, of the probability that X or Y lies in a ball.

To appreciate how this is done, suppose X and Y take values in a common sample space, $\mathcal{B}$, which we take to be a separable Banach space equipped with a norm, $\|\cdot\|$. We make the following continuity assumption on balls:

(**A1**): For all $z \in \mathcal{B}$, and for $Z = X$ and $Z = Y$, the function $\pi_Z(\delta|z) = \mathbb{P}(\|Z-z\| \le \delta)$ is continuous in $\delta \in [0, \infty)$, with $\pi_Z(0|z) = 0$.

We define relative density in terms of ratios of probabilities that X and Y lie in balls. Specifically, given $\eta > 0$ let $\mathcal{S}_X(\eta)$ denote the set of $z \in \mathcal{B}$ such that, for all $\delta \in (0, \eta]$,

$$\pi_Y(\delta|z) > 0 \quad \text{and} \quad \frac{p\,\pi_X(\delta|z)}{(1-p)\,\pi_Y(\delta|z)} \ge 1 + \eta. \tag{3.7}$$

Similarly, define $\mathcal{S}_Y(\eta)$ to be the set of $z \in \mathcal{B}$ such that, for all $\delta \in (0, \eta]$,

$$\pi_X(\delta|z) > 0 \quad \text{and} \quad \frac{(1-p)\,\pi_Y(\delta|z)}{p\,\pi_X(\delta|z)} \ge 1 + \eta. \tag{3.8}$$

Thus, if $z \in \mathcal{S}_X(\eta)$ then we can fairly say that the distribution of X has greater density than that of Y in the neighbourhood of z, weighted by the prior probabilities p and $1-p$, without having to specify what we mean by the densities of the distributions of X or Y.

It is possible to construct realistic, non-Euclidean examples where (**A1**) holds and the definitions of the relative densities in (3.7) and (3.8) are meaningful. When $\mathcal{B}$ denotes a function space, $\|\cdot\|$ might be an L_r norm for some $r \ge 1$. However, it is easier to construct examples in the case of a component-wise supremum norm, as follows. Assume X and Y may be represented as

$$X = \sum_{j=1}^{\infty} \alpha_j\, U_j\, \psi_j \quad \text{and} \quad Y = \sum_{j=1}^{\infty} \alpha_j\, V_j\, \psi_j,$$

where $\alpha_1, \alpha_2, \ldots$ is a sequence of nonnegative constants satisfying $\sum_{j=1}^{\infty} j\,\alpha_j < \infty$, $U, U_1, U_2, \ldots$ and $V, V_1, V_2, \ldots$ are sequences of independent and identically distributed random variables with finite variance, and $\psi_1, \psi_2, \ldots$ is a sequence of bounded, orthogonal functions. In this model, the distance between the distributions of X and Y expresses the distance between the distributions of U and V. Given functions

$$u = \sum_{j=1}^{\infty} \alpha_j\, u_j\, \psi_j \quad \text{and} \quad v = \sum_{j=1}^{\infty} \alpha_j\, v_j\, \psi_j$$

in the common sample space of X and Y, define

$$\|u - v\| = \max_{j \in \mathbb{N}} \alpha_j\, |u_j - v_j|.$$

Then, for example,

$$\mathbb{P}(\|X - u\| \le \delta) = \prod_{j=1}^{\infty} \mathbb{P}\Big(|U_j - u_j| \le \frac{\delta}{\alpha_j} \Big).$$

Using this formula and its analogue for $\mathbb{P}(\|Y - u\| \le \delta)$, we may determine whether $u \in \mathcal{S}_X(0)$ or $u \in \mathcal{S}_Y(0)$, where for both $Z = X$ and $Z = Y$, we define

$$\mathcal{S}_Z(0) = \bigcup_{\eta > 0} \mathcal{S}_Z(\eta).$$

We conclude this section by illustrating the theory that can be developed when the distributions of X and Y are smooth in the sense of (**A1**), and relative densities are defined using (3.7) and (3.8). Theorem 3.4.1 below asserts that the bagged nearest-neighbour classifier, applied to $z \in \mathcal{B}$, converges to the generalised Bayes classifier, which assigns z to Π_X or Π_Y according as $z \in S_X(0)$ or $z \in S_Y(0)$ respectively.

We define

$$\mathcal{S}_Z(\eta, \epsilon) = \{z \in \mathcal{S}_Z(\eta) : \mathbb{P}(\|Z - z\| \le \eta) > \epsilon\}.$$

Restricting attention to $z \in \mathcal{S}_Z(\eta, \epsilon)$ for some $\epsilon > 0$, rather than just to $z \in \mathcal{S}_Z(\eta)$, ensures that the density of the distribution of Z is not too small in the neighbourhood of z. Our assumptions on the resample sizes are:

(**A2**): The resample sizes m_1 and n_1 satisfy

$$\min(m_1, n_1) \to \infty, \quad \max\Big(\frac{m_1}{m}, \frac{n_1}{n}\Big) \to 0 \quad \text{and} \quad \frac{m_1}{n_1} \to \frac{p}{1-p}$$

as $m \to \infty$.

Before we can state the main theorem, we need one final technical condition on the asymptotic independence of the two types of mark on the points. Order the distances of the training-sample points from z as $W_1(z) < \ldots < W_{m+n}(z)$, so that $W_j = \|Z_j - z\|$, for some $Z_j = Z_j(z) \in \mathcal{Z} = \mathcal{X} \cup \mathcal{Y}$, and for $j = 1, \ldots, m+n$.

(**A3**): For each $\eta, \epsilon > 0$, we have

$$\sup_{z \in \mathcal{S}_X(\eta,\epsilon)} \max_{1 \le i < j \le m+n} \mathrm{Cov}(\mathbb{1}_{\{Z_i \in \mathcal{X}\}}, \mathbb{1}_{\{Z_j \in \mathcal{X}\}}) \to 0$$

$$\sup_{z \in \mathcal{S}_Y(\eta,\epsilon)} \max_{1 \le i < j \le m+n} \mathrm{Cov}(\mathbb{1}_{\{Z_i \in \mathcal{Y}\}}, \mathbb{1}_{\{Z_j \in \mathcal{Y}\}}) \to 0$$

as $m \to \infty$.

This condition may be loosely expressed as 'knowing the mark on one point in the combined sample gives negligible information about the mark on any other point in the limit as the sample sizes tend to infinity'. It is highly plausible, and may be redundant, but we were unable to prove the theorem below without it.

Theorem 3.4.1. *Assume* (**A1**), (**A2**) *and* (**A3**). *Then, for each* $\eta, \epsilon > 0$,

$$\inf_{z \in \mathcal{S}_X(\eta,\epsilon)} \mathbb{P}\big(\mathcal{C}_{\mathrm{Bagg}}(z) = X\big) \to 1 \quad \textit{and} \quad \inf_{z \in \mathcal{S}_Y(\eta,\epsilon)} \mathbb{P}\big(\mathcal{C}_{\mathrm{Bagg}}(z) = Y\big) \to 1$$

as $m \to \infty$.

Corollary 3.4.2. *In addition to* (**A1**), (**A2**) *and* (**A3**), *assume that*

$$\mathbb{P}\big(X \in \mathcal{S}_X(0) \cup \mathcal{S}_Y(0)\big) = 1 \quad \textit{and} \quad \mathbb{P}\big(Y \in \mathcal{S}_X(0) \cup \mathcal{S}_Y(0)\big) = 1.$$

Then

$$\mathrm{Risk}(\mathcal{C}_{\mathrm{Bagg}}) \to p\,\mathbb{P}\big(X \in \mathcal{S}_Y(0)\big) + (1-p)\,\mathbb{P}\big(Y \in \mathcal{S}_X(0)\big)$$

as $m \to \infty$.

3.5 Choice of sampling ratio by cross-validation

Let $\mathcal{X}_i = \mathcal{X} \setminus \{X_i\}$ and $\mathcal{Y}_i = \mathcal{Y} \setminus \{Y_i\}$ denote the two samples after the ith data value has been dropped, where $1 \le i \le m$ or $1 \le i \le n$ in the respective cases. Write $\mathcal{C}_{\text{Bagg},-X_i}$ and $\mathcal{C}_{\text{Bagg},-Y_i}$ for the bagged nearest neighbour classifiers based on the sample pairs $(\mathcal{X}_i, \mathcal{Y})$ and $(\mathcal{X}, \mathcal{Y}_i)$, respectively, rather than on $(\mathcal{X}, \mathcal{Y})$. The classifier $\mathcal{C}_{\text{Bagg},-X_i}$ is constructed by sampling m_1 data from $\mathcal{X}_i$ and n_1 data from $\mathcal{Y}$, using either with- or without-replacement sampling; and analogously for $\mathcal{C}_{\text{Bagg},-Y_i}$. To simplify optimisation we shall put $r = m_1/m$ and take $n_1 = \lfloor rn \rfloor$, the integer part of rn, so that optimisation is over only a single parameter. Both with- and without-replacement resampling could be used, and cross-validation methods employed to minimise risk over both approaches as well as over r. However, for simplicity we shall assume that just one of the two types of resampling is employed, and that optimisation over r is attempted for just that type.

A cross-validation based estimator of risk is

$$\widehat{\text{Risk}}(r) = \frac{p}{m} \sum_{i=1}^{m} \mathbb{1}_{\{\mathcal{C}_{\text{Bagg},-X_i}(X_i)=Y\}} + \frac{1-p}{n} \sum_{i=1}^{n} \mathbb{1}_{\{\mathcal{C}_{\text{Bagg},-Y_i}(Y_i)=X\}}.$$

We suggest that r be chosen so as to minimise $\widehat{\text{Risk}}(r)$.

3.5.1 Numerical properties

Panels (a) and (b) of Figure 3.3 show plots of $\widehat{\text{Risk}}(r)$ against $r = m_1/m$ for two typical datasets, with the same distribution pairs as were used in panels (a) and (b), respectively, of Figure 3.2. The sample size is $m = n = 200$, as in the case of panels (c) and (d) of Figure 3.2. Panels (c) and (d) give the frequencies with which different values of m_1 are selected by cross-validation. The results reflect very closely the fact,

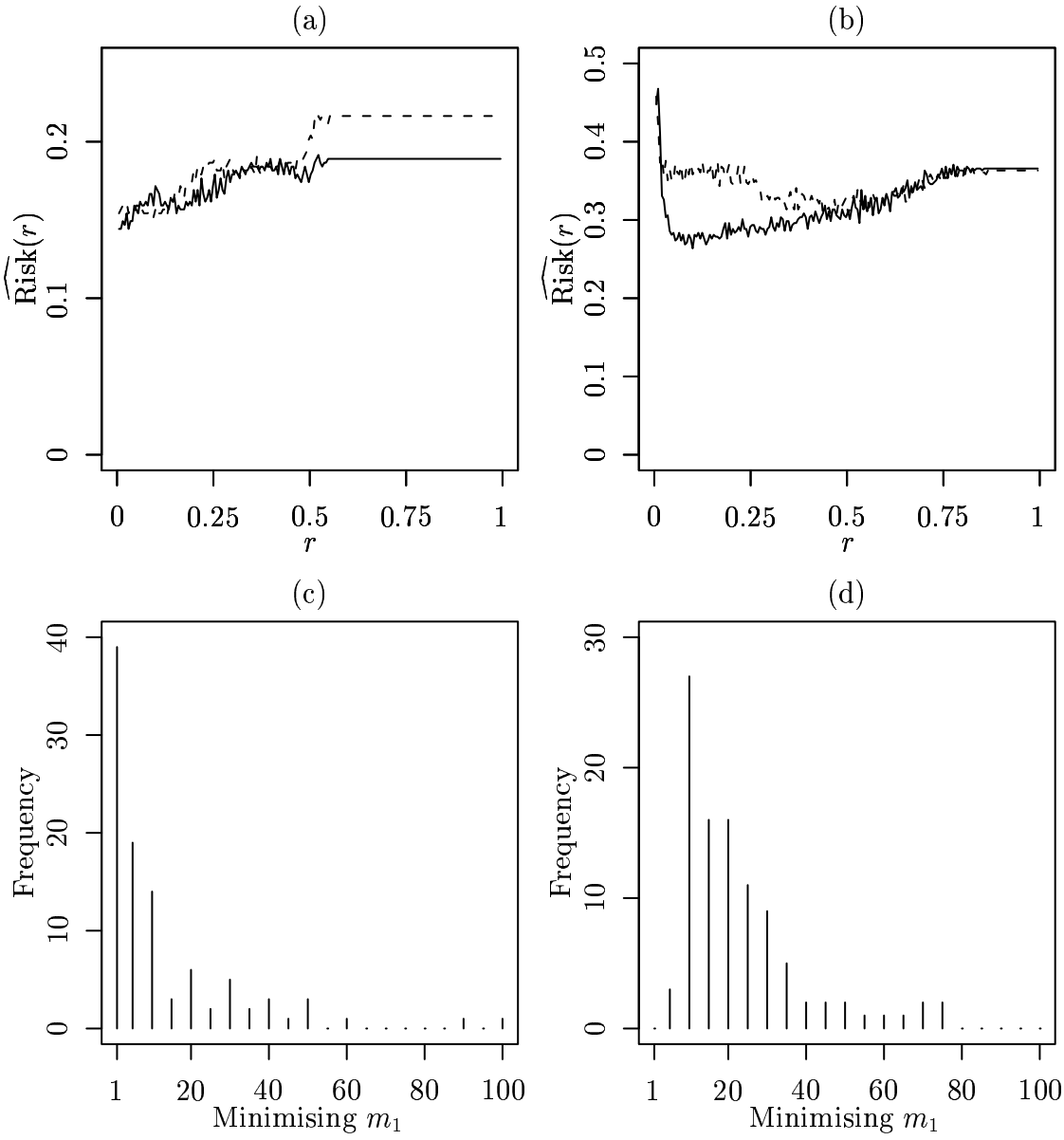


Figure 3.3: Sample sizes are $m = n = 200$, and the densities are as in the respective panels in Figure 1.2, but with $\mu = 2$. Panels (a) and (b) plot the cross-validation criterion for two typical samples. Panels (c) and (d) give the frequencies, from 100 simulations, with which cross-validation selects different values of m_1. Resampling is without-replacement in panels (a) and (c), and with-replacement in (b) and (d).

Classifier	*Risk*	
	Example 1	Example 2
$\mathcal{C}_{\text{Bayes}}$	0.158	0.283
$\mathcal{C}_{\text{Bagg}}$ with optimal m_1	0.160	0.300
$\mathcal{C}_{\text{Bagg}}$ with CV	0.163	0.311
$\mathcal{C}_{\text{NN}}$	0.224	0.379

Table 3.1: Risks of the four classifiers. Sample sizes are $m = n = 200$. Resampling is without-replacement in Example 1, and with-replacement in Example 2.

indicated in Figure 3.2, that $m_1 = 1$ and $m_1 = 14$ are optimal in the respective cases of panels (a) and (b). Analogously to the case of panels (c) and (d) of Figure 3.2, resampling is without-replacement for panels (a) and (c) and with-replacement for panels (b) and (d) of Figure 3.3.

Not only does cross-validation (CV) lead to an appropriate choice of m_1, it also results in only a small increase in risk over that of the bagged nearest-neighbour classifier with the optimal choice of m_1. Table 3.1 compares these risks with those of the Bayes and nearest-neighbour classifiers in the context of the two numerical examples treated in panels (c) and (d) of Figure 3.2.

3.6 Appendix

3.6.1 Asymptotics of the nearest-neighbour classifier

In this section we formalise the remarks made in Section 3.2.2 leading to the derivation of the large-sample risk of the nearest-neighbour classifier. First we establish

some notation and discuss the convergence of the distribution of the data in the neighbourhood of $z \in \mathbb{R}^k$ to that of a marked Poisson process.

Let $\mathcal{X} = \{X_1, \ldots, X_m\}$ and $\mathcal{Y} = \{Y_1, \ldots, Y_n\}$ be independent random samples in $\mathbb{R}^k$, with densities f and g which are continuous almost everywhere. Suppose that $m/(m+n) \to p \in (0,1)$ as $m \to \infty$.

Let $z^1, \ldots, z^d$ be distinct continuity points of f and g with $f(z^i) > 0$ and $g(z^i) > 0$, and let $\lambda^1, \ldots, \lambda^d$ be arbitrary, positive real numbers. For $i = 1, \ldots, d$, choose a decreasing sequence of closed balls $(B_m^i)_{m \geq 1}$ centred at z^i such that, for sufficiently large m,

$$\mathbb{P}\big(X \in B_m^i\big) = \frac{\lambda^i}{m} \equiv p_m^i,$$

say, and write $q_m^i = \mathbb{P}\big(Y \in B_m^i\big)$. For Borel sets $B \in \mathbb{R}^k$, we write

$$N_m^X(B) = \sum_{j=1}^{m} \mathbb{1}_{\{X_j \in B\}} \quad \text{and} \quad N_m^Y(B) = \sum_{j=1}^{n} \mathbb{1}_{\{Y_j \in B\}}$$

for the number of elements in B from $\mathcal{X}$ and $\mathcal{Y}$ respectively.

We shall need to make extensive use of the following elementary lemma:

Lemma 3.6.1. *Suppose $x > 0$ and $x_m \to x$ as $m \to \infty$. Then*

$$\Big(1 - \frac{x_m}{m}\Big)^m \to e^{-x}$$

as $m \to \infty$, uniformly for $x \in (0, K]$, for each $K > 0$.

Proof. The argument in Burkill (1962), p. 179, shows that for $x \in (0, K]$ and sufficiently large m,

$$\Big|\Big(1 - \frac{x}{m}\Big)^m - e^{-x}\Big| \leq \frac{x^2\, e^{-x}}{m - x^2} \leq \frac{K^2}{m - K^2}.$$

Moreover, if $\epsilon \in (0,1)$, then

$$\Big(1 + \frac{\epsilon}{m}\Big)^m - 1 = \sum_{t=1}^{m} \binom{m}{t} \Big(\frac{\epsilon}{m}\Big)^t \leq \sum_{t=1}^{\infty} \frac{\epsilon^t}{2^{t-1}} = \frac{\epsilon}{1 - \epsilon/2} \leq 2\epsilon.$$

Thus, we can choose m_0 large enough such that, for all $m \geq m_0$,

$$\frac{K^2}{m - K^2} \leq \epsilon, \quad |x_m - x| \leq \epsilon \quad \text{and} \quad \Big(1 + \frac{\epsilon}{m}\Big)^m \leq 1 + 2\epsilon.$$

Then, for all $m \geq m_0$ and $x \in (0, K]$,

$$\begin{aligned}\Big|\Big(1 - \frac{x_m}{m}\Big)^m - e^{-x}\Big| &\leq \Big|\Big(1 - \frac{x + x_m - x}{m}\Big)^m - \Big(1 - \frac{x}{m}\Big)^m\Big| + \Big|\Big(1 - \frac{x}{m}\Big)^m - e^{-x}\Big| \\ &\leq 2\epsilon + \epsilon = 3\epsilon.\end{aligned}$$

□

Suppose $r_1, \ldots, r_d \in \mathbb{N}$ satisfy $r_1 + \ldots + r_d \leq m$ and let $m_0 = m - r_1 - \ldots - r_d$. Then for sufficiently large m (in particular, large enough such that $B_m^1, \ldots, B_m^d$ are disjoint), we have

$$\begin{aligned}\mathbb{P}\big(N_m^X(B_m^1) = r_1, \ldots, N_m^X(B_m^d) = r_d\big) &= \frac{m!}{m_0!\, r_1! \ldots r_d!}(p_m^1)^{r_1} \ldots (p_m^d)^{r_d}\Big(1 - \sum_{i=1}^d p_m^i\Big)^{m_0} \\ &\to \exp\Big(-\sum_{i=1}^d \lambda^i\Big)\frac{(\lambda^1)^{r_1}}{r_1!} \ldots \frac{(\lambda^d)^{r_d}}{r_d!}\end{aligned}$$

as $m \to \infty$, uniformly for $\lambda^1, \ldots, \lambda^d \in (0, K]$, for each $K > 0$. Thus, in this sense, the number of elements from $\mathcal{X}$ in each infinitesimal neighbourhood has an asymptotic Poisson distribution, and the numbers in disjoint neighbourhoods are asymptotically independent. Before we can deduce the corresponding result for the $\mathcal{Y}$-sample, we need the following lemma:

Lemma 3.6.2. *For each $i = 1, \ldots, d$,*

$$\frac{q_m^i}{p_m^i} \to \frac{g(z^i)}{f(z^i)}$$

as $m \to \infty$, uniformly for $\lambda^i \in (0, K]$, for each $K > 0$, and uniformly for continuity points z^i of f and g such that $f(z^i) \in [1/C, C]$ and $g(z^i) \in [1/C, C]$, for each $C \geq 1$.

Proof. Fix $i \in \{1, \ldots, d\}$, $C \geq 1$ and $K > 0$. Choose a decreasing sequence of closed balls $(B_m)_{m \geq 1}$ centred at z^i such that, for sufficiently large m,

$$\mathbb{P}\big(X \in B_m\big) = \frac{K}{m}.$$

Given $\epsilon > 0$ small enough such that $f(z^i) > 2\epsilon$, choose m_0 such that, for all $m \geq m_0$,

$$\sup_{z \in B_m} |f(z) - f(z^i)| \leq \epsilon \quad \text{and} \quad \sup_{z \in B_m} |g(z) - g(z^i)| \leq \epsilon.$$

Note that $B_m^i \subseteq B_m$ for each $m \in \mathbb{N}$ and $\lambda^i \in (0, K]$. Now, for all $m \geq m_0$,

$$\begin{aligned} &\mathrm{Vol}(B_m^i)\big(f(z^i) - \epsilon\big) \leq p_m^i \leq \mathrm{Vol}(B_m^i)\big(f(z^i) + \epsilon\big) \\ \text{and} \quad &\mathrm{Vol}(B_m^i)\big(g(z^i) - \epsilon\big) \leq q_m^i \leq \mathrm{Vol}(B_m^i)\big(g(z^i) + \epsilon\big). \end{aligned}$$

Hence, for all $m \geq m_0$,

$$\frac{g(z^i) - \epsilon}{f(z^i) + \epsilon} \leq \frac{q_m^i}{p_m^i} \leq \frac{g(z^i) + \epsilon}{f(z^i) - \epsilon},$$

so that

$$\left| \frac{q_m^i}{p_m^i} - \frac{g(z^i)}{f(z^i)} \right| \leq \frac{2\epsilon\big(f(z^i) + g(z^i)\big)}{f(z^i)^2} \leq 4\epsilon C^3.$$

□

Now suppose $s_1, \ldots, s_d \in \mathbb{N}$ satisfy $s_1 + \ldots + s_d \leq n$, and let $n_0 = n - s_1 - \ldots - s_d$. For sufficiently large m, we have

$$\begin{aligned} \mathbb{P}\big(N_m^Y(B_m^1) = s_1, \ldots, N_m^Y(B_m^d) = s_d\big) &= \frac{n!}{n_0!\, s_1! \ldots s_d!} (q_m^1)^{s_1} \ldots (q_m^d)^{s_d} \Big(1 - \sum_{i=1}^d q_m^i\Big)^{n_0} \\ &\to \exp\Big(-\frac{(1-p)}{p} \sum_{i=1}^d \frac{g(z^i)}{f(z^i)} \lambda^i\Big) \frac{\Big(\frac{(1-p)g(z^1)\lambda^1}{pf(z^1)}\Big)^{s_1}}{s_1!} \ldots \frac{\Big(\frac{(1-p)g(z^d)\lambda^d}{pf(z^d)}\Big)^{s_d}}{s_d!} \end{aligned}$$

as $m \to \infty$, uniformly for $\lambda^1, \ldots, \lambda^d \in (0, K]$, for each $K > 0$, and uniformly for continuity points $z^1, \ldots, z^d$ of f and g such that $f(z^i) \in [1/C, C]$ and $g(z^i) \in [1/C, C]$, for each $C \geq 1$.

We now discuss the asymptotic probability of each of the two types of mark, for which it suffices to consider neighbourhoods of a single point. Let $z \in \mathbb{R}^k$ be a continuity point of f and g, set $U_j = \|X_j - z\|$ for $j = 1, \ldots, m$ and $V_j = \|Y_j - z\|$ for $j = 1, \ldots, n$. As prior to Theorem 3.4.1, we order $U_1, \ldots, U_m$ and $V_1, \ldots, V_n$ as $W_1(z) < \ldots < W_{m+n}(z)$, so that for $j = 1, \ldots, m+n$, we may write $W_j = \|Z_j - z\|$, say, for some $Z_j = Z_j(z) \in \mathcal{Z} = \mathcal{X} \cup \mathcal{Y}$. Fix $0 < \lambda^1 < \ldots < \lambda^d$. Analogously to the above situation of d distinct points, for $i = 1, \ldots, d$, choose a decreasing sequence of closed balls $(\tilde{B}^i_m)_{m \geq 1}$ centred at z and of radius δ^i_m such that, for sufficiently large m,

$$\mathbb{P}\big(X \in \tilde{B}^i_m\big) = \frac{\lambda^i}{m} \equiv \tilde{p}^i_m,$$

say. Write $\tilde{q}^i_m = \mathbb{P}\big(Y \in \tilde{B}^i_m\big)$.

Denote the common density of $U_1, \ldots, U_m$ by h_U and the common density of $V_1, \ldots, V_n$ by h_V. We will need the following result, the proof of which is almost identical to that of Lemma 3.6.2 and is omitted.

Lemma 3.6.3. *For each $i = 1, \ldots, d$,*

$$\frac{h_V(\delta^i_m)}{h_U(\delta^i_m)} \to \frac{g(z)}{f(z)}$$

as $m \to \infty$, uniformly for $\lambda^i \in (0, K]$, for each $K > 0$, and uniformly for continuity points z of f and g such that $f(z) \in [1/C, C]$ and $g(z) \in [1/C, C]$, for each $C \geq 1$.

Recall from Section 3.2.1 that we define

$$q(z) = \frac{p\, f(z)}{p\, f(z) + (1-p)\, g(z)}.$$

We study the asymptotic probability of the points having each type of mark by first conditioning on their distances from z:

Lemma 3.6.4. *For each $d \in \mathbb{N}$ and $(\epsilon_1, \ldots, \epsilon_d) \in \{0,1\}^d$, we have*

$$\mathbb{P}\big(\mathbb{1}_{\{Z_1 \in \mathcal{X}\}} = \epsilon_1, \ldots, \mathbb{1}_{\{Z_d \in \mathcal{X}\}} = \epsilon_d | W_1 = \delta^1_m, \ldots, W_d = \delta^d_m\big) \to \prod_{i=1}^{d} \big\{ q(z)^{\epsilon_i} \big(1 - q(z)\big)^{1-\epsilon_i} \big\}$$

as $m \to \infty$*, uniformly for* $\lambda^d \in (0, K], \lambda^{d-1} \in (0, \lambda^d), \ldots, \lambda^1 \in (0, \lambda^2)$*, for each* $K > 0$*, and uniformly for continuity points* z *of* f *and* g *such that* $f(z) \in [1/C, C]$ *and* $g(z) \in [1/C, C]$*, for each* $C \geq 1$*.*

Proof. In this proof, all sums and products without specified ranges are assumed to run from $i = 1, \ldots, d$, and $\sum_{\boldsymbol{\eta}}$ is shorthand for $\sum_{\eta_1=0}^{1} \ldots \sum_{\eta_d=0}^{1}$. For sufficiently large m, we have

$$\begin{aligned}
&\mathbb{P}\big(\mathbb{1}_{\{Z_1 \in \mathcal{X}\}} = \epsilon_1, \ldots, \mathbb{1}_{\{Z_d \in \mathcal{X}\}} = \epsilon_d \;\big|\; W_1 = \delta_m^1, \ldots, W_d = \delta_m^d\big) \\
&= \frac{\frac{m!}{(m-\sum \epsilon_i)!}\{\prod h_U(\delta_m^i)^{\epsilon_i}\}(1-\tilde{p}_m^d)^{m-\sum \epsilon_i}\frac{n!}{(n-d+\sum \epsilon_i)!}\{\prod h_V(\delta_m^i)^{1-\epsilon_i}\}(1-\tilde{q}_m^d)^{n-d+\sum \epsilon_i}}{\sum_{\boldsymbol{\eta}}\frac{m!}{(m-\sum \eta_i)!}\{\prod h_U(\delta_m^i)^{\eta_i}\}(1-\tilde{p}_m^d)^{m-\sum \eta_i}\frac{n!}{(n-d+\sum \eta_i)!}\{\prod h_V(\delta_m^i)^{1-\eta_i}\}(1-\tilde{q}_m^d)^{n-d+\sum \eta_i}} \\
&\to \frac{1}{\sum_{\boldsymbol{\eta}}\big\{\frac{(1-p)g(z)}{pf(z)}\big\}^{\sum(\epsilon_i-\eta_i)}}
\end{aligned}$$

as $m \to \infty$, uniformly for $\lambda^d \in (0, K], \lambda^{d-1} \in (0, \lambda^d), \ldots, \lambda^1 \in (0, \lambda^2)$, for each $K > 0$, and uniformly for continuity points z of f and g such that $f(z) \in [1/C, C]$ and $g(z) \in [1/C, C]$, for each $C \geq 1$. But

$$\begin{aligned}
\frac{1}{\sum_{\boldsymbol{\eta}}\big\{\frac{(1-p)g(z)}{pf(z)}\big\}^{\sum(\epsilon_i-\eta_i)}} &= \frac{\{pf(z)\}^{\sum \epsilon_i}}{\{(1-p)g(z)\}^{\sum \epsilon_i}\sum_{j=0}^{d}\binom{d}{j}\big\{\frac{pf(z)}{(1-p)g(z)}\big\}^j} \\
&= \frac{\{pf(z)\}^{\sum \epsilon_i}}{\{(1-p)g(z)\}^{\sum \epsilon_i}\big\{1+\frac{pf(z)}{(1-p)g(z)}\big\}^d} \\
&= \prod\big\{q(z)^{\epsilon_i}\big(1-q(z)\big)^{1-\epsilon_i}\big\},
\end{aligned}$$

as required. □

The following theorem shows that the marks on the points are asymptotically independent, each point having mark X with probability $q(z)$ and mark Y with probability $1 - q(z)$.

Theorem 3.6.5. *For each $d \in \mathbb{N}$ and $(\epsilon_1, \ldots, \epsilon_d) \in \{0,1\}^d$, we have*

$$\mathbb{P}(\mathbb{1}_{\{Z_1 \in \mathcal{X}\}} = \epsilon_1, \ldots, \mathbb{1}_{\{Z_d \in \mathcal{X}\}} = \epsilon_d) \to \prod_{i=1}^{d} \big\{ q(z)^{\epsilon_i} \big(1 - q(z)\big)^{1-\epsilon_i} \big\}$$

as $m \to \infty$, uniformly for continuity points z of f and g such that $f(z) \in [1/C, C]$ and $g(z) \in [1/C, C]$, for each $C \geq 1$.

Proof. Fix $\epsilon > 0$ and $K > 0$. Define a decreasing sequence of closed balls centred at z and of radius K_m such that, for sufficiently large m,

$$\mathbb{P}(\|X - z\| \leq K_m) = \frac{K}{m}.$$

Observe that, for sufficiently large m,

$$\mathbb{P}(W_d \geq K_m) \leq \mathbb{P}(U_d \geq K_m) = \sum_{j=0}^{d-1} \binom{m}{j} \Big(\frac{K}{m}\Big)^j \Big(1 - \frac{K}{m}\Big)^{m-j} \leq 2 \sum_{j=0}^{d-1} \frac{K^j}{j!} e^{-K}.$$

But $\sum_{j=0}^{d-1} \frac{K^j}{j!} e^{-K} \to 0$ as $K \to \infty$, so we can choose our value of K above such that

$$\mathbb{P}(W_d \leq K_m) \geq 1 - \epsilon$$

for all $m \geq m_0$, say. By Lemma 3.6.4, there exists m_1 such that, for all $m \geq m_1$, all $0 < \lambda^1 < \ldots < \lambda^d \leq K$ and all continuity points z of f and g such that $f(z) \in [1/C, C]$ and $g(z) \in [1/C, C]$,

$$\Big| \mathbb{P}\big(\mathbb{1}_{\{Z_1 \in \mathcal{X}\}} = \epsilon_1, \ldots, \mathbb{1}_{\{Z_d \in \mathcal{X}\}} = \epsilon_d \;\big|\; W_1 = \delta_m^1, \ldots, W_d = \delta_m^d\big) - \prod_{i=1}^{d} \big\{ q(z)^{\epsilon_i} \big(1 - q(z)\big)^{1-\epsilon_i} \big\} \Big| \leq \epsilon.$$

For $i = 1, \ldots, d$, let $H_i(\delta) = \mathbb{P}(W_i \leq \delta)$. Then, for all $m \geq \max(m_0, m_1)$ and all continuity points z of f and g such that $f(z) \in [1/C, C]$ and $g(z) \in [1/C, C]$,

$$\begin{aligned} \Big| \mathbb{P}(\mathbb{1}_{\{Z_1 \in \mathcal{X}\}} = \epsilon_1, \ldots, \mathbb{1}_{\{Z_d \in \mathcal{X}\}} = \epsilon_d) - \prod_{i=1}^{d} \big\{ q(z)^{\epsilon_i} \big(1 - q(z)\big)^{1-\epsilon_i} \big\} \Big| \\ \leq \epsilon + \int_0^{K_m} \int_0^{\delta_m^d} \ldots \int_0^{\delta_m^2} \epsilon \; dH_1(\delta) \ldots dH_d(\delta) \leq 2\epsilon. \end{aligned}$$

□

Setting $d = 1$ in the Theorem 3.6.5, the following corollary is immediate:

Corollary 3.6.6. *As $m \to \infty$,*

$$\mathbb{P}\big(\mathcal{C}_{\mathrm{NN}}(z) = X\big) \to q(z),$$

uniformly for continuity points z of f and g with $f(z) \in [1/C, C]$ and $g(z) \in [1/C, C]$, for each $C \geq 1$.

Finally, we can establish the asymptotic risk of the nearest-neighbour classifier, as given in (3.3).

Theorem 3.6.7. *As $m \to \infty$,*

$$\mathrm{Risk}(\mathcal{C}_{\mathrm{NN}}) \to 2\,p\,(1-p) \int_{\mathbb{R}^k} \frac{f(z)\,g(z)}{p\,f(z) + (1-p)\,g(z)}\,dz.$$

Proof. By the monotone convergence theorem,

$$\int_{\mathbb{R}^k} f(z)\,\mathbb{1}_{\{f(z)\in[1/C,C]\}}\,dz \to 1$$

as $C \to \infty$, and the same result holds if f is replaced by g. Thus, given $\epsilon > 0$, there exists $C \geq 1$ such that

$$\int_{\mathbb{R}^k} f(z)\,\mathbb{1}_{\{f(z)\in[1/C,C]\}}\,dz \geq 1-\epsilon \quad \text{and} \quad \int_{\mathbb{R}^k} g(z)\,\mathbb{1}_{\{g(z)\in[1/C,C]\}}\,dz \geq 1-\epsilon.$$

By Corollary 3.6.6, we can choose m_0 large enough such that

$$\big|\mathbb{P}\big(\mathcal{C}_{\mathrm{NN}}(z) = X\big) - q(z)\big| \leq \epsilon,$$

for all $m \geq m_0$ and all continuity points z of f and g such that $f(z) \in [1/C, C]$ and $g(z) \in [1/C, C]$. Since the set of points where either f or g is not continuous has Lebesgue measure zero, we have

$$p \int_{\mathbb{R}^k} \big|\mathbb{P}\big(\mathcal{C}_{\mathrm{NN}}(z) = Y\big) - \big(1 - q(z)\big)\big|\, f(z)\,\mathbb{1}_{\{f(z)\in[1/C,C]\}}\,dz$$
$$+ (1-p) \int_{\mathbb{R}^k} \big|\mathbb{P}\big(\mathcal{C}_{\mathrm{NN}}(z) = X\big) - q(z)\big|\, g(z)\,\mathbb{1}_{\{g(z)\in[1/C,C]\}}\,dz \leq \epsilon,$$

for all $m \geq m_0$. Recalling that

$$p \int_{\mathbb{R}^k} \{1-q(z)\} f(z)\, dz + (1-p) \int_{\mathbb{R}^k} q(z) g(z)\, dz = 2p(1-p) \int_{\mathbb{R}^k} \frac{f(z)\, g(z)}{pf(z) + (1-p)g(z)}\, dz,$$

it follows that for all $m \geq m_0$,

$$\left| \mathrm{Risk}(\mathcal{C}_{\mathrm{NN}}) - 2\, p\, (1-p) \int_{\mathbb{R}^k} \frac{f(z)\, g(z)}{p\, f(z) + (1-p)\, g(z)}\, dz \right| \leq 2\epsilon.$$

□

3.6.2 Asymptotics of the bagged nearest-neighbour classifier

The arguments of Section 3.6.1 were developed mainly in order to study the corresponding properties of the bagged nearest-neighbour classifier, which is the purpose of this section. It consists of formalising the remarks made in Section 3.3, the first step of which concerns the thinned marked Poisson process.

We retain the notation and assumptions of the previous section, and also denote by $\mathcal{X}^* = \{X_1^*, \ldots, X_{m_1}^*\}$ and $\mathcal{Y}^* = \{Y_1^*, \ldots, Y_{n_1}^*\}$ the resamples obtained by sampling uniformly at random from $\mathcal{X}$ and $\mathcal{Y}$ respectively, either with- or without-replacement. For Borel sets $B \in \mathbb{R}^k$, we write

$$N_m^{X^*}(B) = \sum_{j=1}^{m_1} \mathbb{1}_{\{X_j^* \in B\}} \quad \text{and} \quad N_m^{Y^*}(B) = \sum_{j=1}^{n_1} \mathbb{1}_{\{Y_j^* \in B\}}$$

for the number of elements in B from $\mathcal{X}^*$ and $\mathcal{Y}^*$ respectively.

Suppose $r_1^*, \ldots, r_d^* \in \mathbb{N}$ satisfy $r_1^* + \ldots + r_d^* \leq m_1$, and let $m_0^* = m_1 - r_1^* - \ldots - r_d^*$. Write

$$r = \sum_{i=1}^{d} r_i \quad \text{and} \quad r^* = \sum_{i=1}^{d} r_i^*.$$

We deal first with the case of with-replacement resampling:

Proposition 3.6.8. *Suppose $m_1/m \to \ell \in (0,1]$, and let $\rho = e^{-\ell}$. Then*

$$\mathbb{P}\big(N_m^{X^*}(B_m^1) = r_1^*, \ldots, N_m^{X^*}(B_m^d) = r_d^*\big) \to \prod_{i=1}^{d}\Big\{\exp\big(-(1-\rho)\lambda^i\big)\frac{\big((1-\rho)\lambda^i\big)^{r_i^*}}{r_i^*!}\Big\}$$

as $m \to \infty$, uniformly for $\lambda^1, \ldots, \lambda^d \in (0, K]$, for each $K > 0$.

Proof. In this proof, and that of Proposition 3.6.9 below, all sums and products without specified ranges are assumed to run from $i = 1, \ldots, d$, and $\sum_{\mathbf{r}\in\mathcal{R}}$ denotes the sum over all d-tuples $(r_1, \ldots, r_d)$ of integers r_i such that

$$r_i \in \{r_i^*, r_i^* + 1, \ldots, m\},$$

for each $i = 1, \ldots, d$, and $r = m - m_0$, where $m_0 \geq 0$. First note that, by the principle of inclusion-exclusion,

$$\begin{aligned}\mathbb{P}\big(N_m^{X^*}(B_m^1) = r_1^*, \ldots, N_m^{X^*}(B_m^d) = r_d^* \;\big|\; N_m^X(B_m^1) = r_1, \ldots, N_m^X(B_m^d) = r_d\big)& \\ = \Big\{\prod\binom{r_i}{r_i^*}\Big\}\Big\{\sum_{t=0}^{r^*}(-1)^t\binom{r^*}{t}\Big(1 - \frac{r - r^* + t}{m}\Big)^{m_1}\Big\}&.\end{aligned}$$

Thus,

$$\begin{aligned}&\mathbb{P}\big(N_m^{X^*}(B_m^1) = r_1^*, \ldots, N_m^{X^*}(B_m^d) = r_d^*\big) \\ &= \sum_{\mathbf{r}\in\mathcal{R}}\Big\{\prod\binom{r_i}{r_i^*}\frac{(p_m^i)^{r_i}}{r_i!}\Big\}\Big\{\sum_{t=0}^{r^*}(-1)^t\binom{r^*}{t}\Big(1 - \frac{r - r^* + t}{m}\Big)^{m_1}\Big\}\frac{m!}{m_0!}\Big(1 - \sum p_m^i\Big)^{m_0}.\end{aligned} \tag{3.9}$$

We wish to take the limit as $m \to \infty$ of each term in (3.9). This is justified by the dominated convergence theorem, since each term of the outer sum in (3.9) is no greater than

$$2^{r^*}\exp\Big(-\sum\lambda^i\Big)\prod\Big\{\binom{r_i}{r_i^*}\frac{(\lambda^i)^{r_i}}{r_i!}\Big\},$$

and

$$2^{r^*}\exp\Big(-\sum\lambda^i\Big)\prod\Big\{\sum_{r_i=r_i^*}^{\infty}\binom{r_i}{r_i^*}\frac{(\lambda^i)^{r_i}}{r_i!}\Big\} = 2^{r^*}\prod\Big\{\frac{(\lambda^i)^{r_i^*}}{r_i^*!}\Big\}.$$

Thus, uniformly for $\lambda^1, \ldots, \lambda^d \in (0, K]$, as $m \to \infty$,

$$\begin{aligned}
&\mathbb{P}\big(N_m^{X^*}(B_m^1) = r_1^*, \ldots, N_m^{X^*}(B_m^d) = r_d^*\big) \\
&\qquad \to \prod \Big\{ e^{-\lambda^i} \sum_{r_i = r_i^*}^{\infty} \binom{r_i}{r_i^*} \frac{(\lambda^i)^{r_i}}{r_i!} \rho^{r_i - r_i^*} \Big\} \sum_{t=0}^{r^*} (-1)^t \binom{r^*}{t} \rho^t \\
&\qquad = \prod \Big\{ e^{-\lambda^i} \sum_{r_i = r_i^*}^{\infty} \binom{r_i}{r_i^*} \frac{(\lambda^i)^{r_i}}{r_i!} \rho^{r_i - r_i^*} (1-\rho)^{r_i^*} \Big\} \\
&\qquad = \prod \Big\{ e^{-\lambda^i} \frac{\big((1-\rho)\lambda^i\big)^{r_i^*}}{r_i^*!} \sum_{r_i = r_i^*}^{\infty} \frac{(\lambda^i \rho)^{r_i - r_i^*}}{(r_i - r_i^*)!} \Big\} \\
&\qquad = \prod \Big\{ \exp\big(-(1-\rho)\lambda^i\big) \frac{\big((1-\rho)\lambda^i\big)^{r_i^*}}{r_i^*!} \Big\},
\end{aligned}$$

as required. □

Proposition 3.6.9 below gives the corresponding result for without-replacement resampling.

Proposition 3.6.9. *Suppose $m_1/m \to \ell \in (0,1)$, and let $\rho = 1 - \ell$. Then*

$$\mathbb{P}\big(N_m^{X^*}(B_m^1) = r_1^*, \ldots, N_m^{X^*}(B_m^d) = r_d^*\big) \to \prod_{i=1}^{d} \Big\{ \exp\big(-(1-\rho)\lambda^i\big) \frac{\big((1-\rho)\lambda^i\big)^{r_i^*}}{r_i^*!} \Big\}$$

as $m \to \infty$, uniformly for $\lambda^1, \ldots, \lambda^d \in (0, K]$, for each $K > 0$.

Proof. By a combinatorial argument,

$$\begin{aligned}
&\mathbb{P}\big(N_m^{X^*}(B_m^1) = r_1^*, \ldots, N_m^{X^*}(B_m^d) = r_d^* \;\big|\; N_m^X(B_m^1) = r_1, \ldots, N_m^X(B_m^d) = r_d\big) \\
&\qquad = \frac{m_1!}{r_1^*! \ldots r_d^*! \, m_0^*!} \prod \Big\{ \prod_{t=0}^{r_i^* - 1} (r_i - t) \Big\} \prod_{t=0}^{m_0^* - 1} (m_0 - t) \prod_{t=0}^{m_1 - 1} \frac{1}{m - t} \\
&\qquad = \prod \binom{r_i}{r_i^*} \prod_{t=0}^{r^* - 1} \Big(\frac{m_1 - t}{m - t} \Big) \prod_{t=0}^{m_0^* - 1} \Big(\frac{m - r - t}{m - r^* - t} \Big).
\end{aligned}$$

Thus,

$$\mathbb{P}\big(N_m^{X^*}(B_m^1) = r_1^*, \ldots, N_m^{X^*}(B_m^d) = r_d^*\big)$$
$$= \sum_{\mathbf{r}\in\mathcal{R}} \Big\{ \prod \binom{r_i}{r_i^*} \frac{(p_m^i)^{r_i}}{r_i!} \Big\} \prod_{t=0}^{r^*-1} \Big(\frac{m_1 - t}{m - t}\Big) \prod_{t=0}^{m_0^*-1} \Big(\frac{m - r - t}{m - r^* - t}\Big) \frac{m!}{m_0!} \Big(1 - \sum p_m^i\Big)^{m_0} \tag{3.10}$$

Now, in order to apply the dominated convergence theorem again, note that each term of the outer sum in (3.10) is no greater than

$$\exp\Big(-\sum \lambda^i\Big) \prod \Big\{ \binom{r_i}{r_i^*} \frac{(\lambda^i)^{r_i}}{r_i!} \Big\},$$

which can be summed as in the proof of Proposition 3.6.8. Continuing to mimic the proof of Proposition 3.6.8, it suffices to show that

$$\prod_{t=0}^{r^*-1} \Big(\frac{m_1 - t}{m - t}\Big) \prod_{t=0}^{m_0^*-1} \Big(\frac{m - r - t}{m - r^* - t}\Big) \to \prod \{\rho^{r_i - r_i^*} (1-\rho)^{r_i^*}\}$$

as $m \to \infty$. Now,

$$\prod_{t=0}^{r^*-1} \Big(\frac{m_1 - t}{m - t}\Big) \to \ell^{r^*}$$

as $m \to \infty$, so it remains to prove that

$$\prod_{t=0}^{m_0^*-1} \Big(\frac{m - r - t}{m - r^* - t}\Big) \to (1-\ell)^{r - r^*}$$

as $m \to \infty$. But,

$$\log\Big\{ \prod_{t=0}^{m_0^*-1} \Big(\frac{m - r - t}{m - r^* - t}\Big) \Big\} = \sum_{t=0}^{m_0^*-1} \log\Big(1 - \frac{r - r^*}{m - r^* - t}\Big)$$
$$= -\sum_{t=0}^{m_0^*-1} \sum_{s=1}^{\infty} \frac{\big(r - r^*\big)^s}{s\big(m - r^* - t\big)^s}$$
$$= -\sum_{s=1}^{\infty} \frac{\big(r - r^*\big)^s}{s} \sum_{t=0}^{m_0^*-1} \frac{1}{\big(m - r^* - t\big)^s}. \tag{3.11}$$

To show that the sum of the terms in (3.11) with $s \geq 2$ is negligible, observe that for sufficiently large m,

$$\begin{aligned}\sum_{s=2}^{\infty} \sum_{t=0}^{m_0^*-1} \frac{(r-r^*)^s}{s(m-r^*-t)^s} &\leq \sum_{s=2}^{\infty} \frac{(r-r^*)^s m_1}{(m-m_1)^s} \\ &\leq \frac{2\ell(r-r^*)}{1-\ell} \sum_{s=1}^{\infty} \Big(\frac{r-r^*}{m-m_1}\Big)^s \\ &\to 0\end{aligned}$$

as $m \to \infty$. To find the limit as $m \to \infty$ of the $s=1$ term in (3.11), we write

$$\sum_{t=1}^{a} \frac{1}{t} = \log a + \gamma + \epsilon(a),$$

where γ is Euler's constant, and $\epsilon(a) \to 0$ as $a \to \infty$. Thus

$$\begin{aligned}-(r-r^*) \sum_{t=0}^{m_0^*-1} \frac{1}{m-r^*-t} &= -(r-r^*)\{\log(m-r^*) - \log(m-m_1+1) \\ &\qquad + \epsilon(m-r^*) - \epsilon(m-m_1+1)\} \\ &\to (r-r^*)\log(1-\ell)\end{aligned}$$

as $m \to \infty$, as required. □

Recall that $m/(m+n) \to p \in (0,1)$ as $m \to \infty$. We state below the corresponding Poisson limits for the $\mathcal{Y}^*$-sample in both the with- and without-replacement cases. The proofs mirror those for the $\mathcal{X}^*$-sample, taken in conjunction with Lemma 3.6.2, and are omitted.

Corollary 3.6.10. *In the with-replacement case, suppose that $n_1/n \to \ell \in (0,1]$ as $m \to \infty$, and set $\rho = e^{-\ell}$. In the without-replacement case, suppose $n_1/n \to \ell \in (0,1)$ as $m \to \infty$, and set $\rho = 1-\ell$. Then*

$$\begin{aligned}&\mathbb{P}\big(N_m^{Y^*}(B_m^1) = s_1^*, \ldots, N_m^{Y^*}(B_m^d) = s_d^*\big) \to \\ &\qquad \prod_{i=1}^{d} \left\{ \exp\Big(-\frac{(1-\rho)(1-p)g(z^i)\lambda^i}{pf(z^i)}\Big) \frac{\big(\frac{(1-\rho)(1-p)g(z^i)\lambda^i}{pf(z^i)}\big)^{s_i^*}}{s_i^*!} \right\}\end{aligned}$$

as $m \to \infty$*, uniformly for* $\lambda^1, \ldots, \lambda^d \in (0, K]$*, for each* $K > 0$*, and uniformly for continuity points* $z^1, \ldots, z^d$ *of* f *and* g *such that* $f(z^i) \in [1/C, C]$ *and* $g(z^i) \in [1/C, C]$*, for each* $C \geq 1$*.*

We now wish to study the conditional probability, given the original samples, that the nearest element of either resample to z is from $\mathcal{X}$. Recalling the definition of $Z_1, \ldots, Z_{m+n}$ prior to Lemma 3.6.3, and with $\mathcal{Z}^* = \mathcal{X}^* \cup \mathcal{Y}^*$, and, for $j = 1, \ldots, m+n$, define

$$\rho_j = \mathbb{P}(Z_j \notin \mathcal{Z}^* | Z_1, \ldots Z_{j-1} \notin \mathcal{Z}^*; \mathcal{X}, \mathcal{Y}).$$

With this notation,

$$\begin{aligned} \mathbb{P}(\text{Nearest element of } \mathcal{Z}^* \text{ to } z \text{ has mark } X | \mathcal{X}, \mathcal{Y}) &= \sum_{j=1}^{m+n} \rho_1 \ldots \rho_{j-1}(1-\rho_j) \mathbb{1}_{\{Z_j \in \mathcal{X}\}} \\ &\equiv \pi(z|\mathcal{Z}), \end{aligned}$$

say. Our study of the asymptotic behaviour of this random variable begins with the following lemma:

Lemma 3.6.11. *In the with-replacement case, suppose both* m_1/m *and* n_1/n *converge to* $\ell \in (0, 1]$ *as* $m \to \infty$*, and set* $\rho = e^{-\ell}$*. In the without-replacement case, suppose both* m_1/m *and* n_1/n *converge to* $\ell \in (0, 1)$ *as* $m \to \infty$*, and set* $\rho = 1 - \ell$*. Let* $d \in \mathbb{N}$*, and let* $I_1, \ldots, I_d$ *be independent Bernoulli*$\big(q(z)\big)$ *random variables. Then*

$$\sum_{j=1}^{d} \rho_1 \ldots \rho_{j-1}(1-\rho_j) \mathbb{1}_{\{Z_j \in \mathcal{X}\}} \xrightarrow{d} \sum_{j=1}^{d} \rho^{j-1}(1-\rho) I_j$$

as $m \to \infty$ *uniformly for continuity points* z *of* f *and* g *such that* $f(z) \in [1/C, C]$ *and* $g(z) \in [1/C, C]$*, for each* $C \geq 1$*.*

Proof. For $j \in \{1, \ldots, m+n\}$, let $R_j = \big\{r \in \{0, 1, \ldots, j-1\} : r \leq m,\ j-1-r \leq n\big\}$.

In the with-replacement case, for $r \in R_j$, we have

$$\begin{aligned}\mathbb{P}\Big(Z_1,\ldots,Z_{j-1} \notin \mathcal{Z}^*, Z_j \in \mathcal{Z}^* \;\Big|\; \sum_{s=1}^{j-1} \mathbb{1}_{\{Z_s \in \mathcal{X}\}} = r \text{ and } Z_j \in \mathcal{X}\Big) \\ = \Big\{\Big(1-\frac{r}{m}\Big)^{m_1} - \Big(1-\frac{r+1}{m}\Big)^{m_1}\Big\}\Big(1-\frac{j-1-r}{n}\Big)^{n_1} \\ \to (\rho^r - \rho^{r+1})\rho^{j-1-r} = \rho^{j-1}(1-\rho)\end{aligned}$$

as $m \to \infty$. In the without-replacement case, for $r \in R_j$,

$$\begin{aligned}\mathbb{P}\Big(Z_1,\ldots,Z_{j-1} \notin \mathcal{Z}^*, Z_j \in \mathcal{Z}^* \;\Big|\; \sum_{s=1}^{j-1} \mathbb{1}_{\{Z_s \in \mathcal{X}\}} = r \text{ and } Z_j \in \mathcal{X}\Big) \\ = \Big(\frac{m_1}{m}\Big)\Big(1-\frac{r}{m-1}\Big)\Big(1-\frac{r}{m-2}\Big)\cdots\Big(1-\frac{r}{m-m_1+1}\Big) \\ \times \Big(1-\frac{j-1-r}{n}\Big)\Big(1-\frac{j-1-r}{n-1}\Big)\cdots\Big(1-\frac{j-1-r}{n-n_1+1}\Big) \\ \to \rho^{j-1}(1-\rho)\end{aligned}$$

as $m \to \infty$, by the same method as was used in the proof of Proposition 3.6.9. Note that the limit is identical for both types of resampling, and does not depend upon r. Hence, using Theorem 3.6.5 and Slutsky's theorem, we have

$$\begin{aligned}\rho_1\ldots\rho_{j-1}(1-\rho_j)\mathbb{1}_{\{Z_j\in\mathcal{X}\}} &= \mathbb{P}(Z_1,\ldots,Z_{j-1}\notin\mathcal{Z}^*, Z_j\in\mathcal{Z}^*|\mathcal{X},\mathcal{Y})\mathbb{1}_{\{Z_j\in\mathcal{X}\}} \\ &\xrightarrow{d} \rho^{j-1}(1-\rho)I_j\end{aligned}$$

as $m \to \infty$, uniformly for continuity points z of f and g such that $f(z) \in [1/C, C]$ and $g(z) \in [1/C, C]$, for each $C \geq 1$. In fact, though, Theorem 3.6.5 yields that

$$(\mathbb{1}_{\{Z_1\in\mathcal{X}\}},\ldots,\mathbb{1}_{\{Z_d\in\mathcal{X}\}}) \xrightarrow{d} (I_1,\ldots,I_d)$$

as $m \to \infty$, uniformly for continuity points z of f and g such that $f(z) \in [1/C, C]$ and $g(z) \in [1/C, C]$, for each $C \geq 1$. Applying the multidimensional version of Slutsky's theorem (Pollard, 2002, p. 175), we deduce the required result. □

Lemma 3.6.11 is not quite strong enough to deduce the asymptotic probability that the bagged nearest-neighbour classifier assigns z to Π_X. For this, we need a further piece of machinery:

Lemma 3.6.12. *Suppose* $\{A_{m,d} : m \in \mathbb{N}, d = 1, \ldots, d_m\}$ *is a triangular array of random variables satisfying the following three conditions:*

(i) $A_{m,d} \xrightarrow{d} A_d$, *say, as* $m \to \infty$, *for each* $d \in \mathbb{N}$;

(ii) $A_d \xrightarrow{d} A$, *say, as* $d \to \infty$;

(iii) given $\epsilon > 0$, *there exists* $d_0' \in \mathbb{N}$ *such that*

$$\mathbb{P}(|A_{m,d} - A_{m,d_0'}| \geq \epsilon) \leq \epsilon,$$

for all sufficiently large m *and* $d = d_0', \ldots, d_m$.

Then $A_{m,d_m} \xrightarrow{d} A$ *as* $m \to \infty$.

Proof. Suppose $\psi : \mathbb{R} \to \mathbb{R}$ is a uniformly continuous function, bounded by C_0. Given $\epsilon > 0$, choose $\delta > 0$ such that

$$|\psi(x) - \psi(y)| \leq \epsilon$$

for all $|x - y| \leq \delta$. Now, by condition (iii), choose $d_0' \in \mathbb{N}$ and $m_0' \in \mathbb{N}$ such that

$$\mathbb{P}(|A_{m,d} - A_{m,d_0'}| \geq \delta) \leq \epsilon$$

for all $m \geq m_0'$ and $d = d_0', \ldots, d_m$. Then, for such m and d,

$$\begin{aligned}\big|\mathbb{E}\big(\psi(A_{m,d})\big) - \mathbb{E}\big(\psi(A_{m,d_0'})\big)\big| &\leq \Big|\mathbb{E}\Big\{\big(\psi(A_{m,d}) - \psi(A_{m,d_0'})\big)\mathbb{1}_{\{|A_{m,d} - A_{m,d_0'}| \leq \delta\}}\Big\}\Big| \\ &\quad + \Big|\mathbb{E}\Big\{\big(\psi(A_{m,d}) - \psi(A_{m,d_0'})\big)\mathbb{1}_{\{|A_{m,d} - A_{m,d_0'}| > \delta\}}\Big\}\Big| \\ &\leq \epsilon + 2\,C_0\,\epsilon = (2\,C_0 + 1)\,\epsilon.\end{aligned}$$

Now, by condition (ii), choose $d_0 \geq d_0'$ such that $\big|\mathbb{E}\big(\psi(A_d)\big) - \mathbb{E}\big(\psi(A)\big)\big| \leq \epsilon$ for $d \geq d_0$, and by condition (i), choose $m_0 \geq m_0'$ such that $\big|\mathbb{E}\big(\psi(A_{m,d_0})\big) - \mathbb{E}\big(\psi(A_{d_0})\big)\big| \leq \epsilon$ for $m \geq m_0$. Then, for $m \in \mathbb{N}$ such that $m \geq m_0$ and $d_m \geq d_0$,

$$\begin{aligned}
&\big|\mathbb{E}\big(\psi(A_{m,d_m})\big) - \mathbb{E}\big(\psi(A)\big)\big| \\
&\qquad \leq \big|\mathbb{E}\big(\psi(A_{m,d_m})\big) - \mathbb{E}\big(\psi(A_{m,d_0'})\big)\big| + \big|\mathbb{E}\big(\psi(A_{m,d_0'})\big) - \mathbb{E}\big(\psi(A_{m,d_0})\big)\big| \\
&\qquad\qquad + \big|\mathbb{E}\big(\psi(A_{m,d_0})\big) - \mathbb{E}\big(\psi(A_{d_0})\big)\big| + \big|\mathbb{E}\big(\psi(A_{d_0})\big) - \mathbb{E}\big(\psi(A)\big)\big| \\
&\qquad \leq (2\,C_0 + 1)\,\epsilon + (2\,C_0 + 1)\,\epsilon + \epsilon + \epsilon \\
&\qquad = 4\,(C_0 + 1)\,\epsilon.
\end{aligned}$$

□

Theorem 3.6.13. *In the with-replacement case, suppose both m_1/m and n_1/n converge to $\ell \in (0,1]$ as $m \to \infty$, and set $\rho = e^{-\ell}$. In the without-replacement case, suppose both m_1/m and n_1/n converge to $\ell \in (0,1)$ as $m \to \infty$, and set $\rho = 1 - \ell$. Let $I_1, I_2, \ldots$ be independent Bernoulli$\big(q(z)\big)$ random variables. Then*

$$\mathbb{P}\big(\mathcal{C}_{\text{Bagg}}(z) = X\big) \to \mathbb{P}\bigg(\sum_{j=1}^{\infty} \rho^{j-1}(1-\rho) I_j > \frac{1}{2}\bigg)$$

as $m \to \infty$, uniformly for continuity points z of f and g such that $f(z) \in [1/C, C]$ and $g(z) \in [1/C, C]$, for each $C \geq 1$.

Proof. Since

$$\mathbb{P}\big(\mathcal{C}_{\text{Bagg}}(z) = X\big) = \mathbb{P}\bigg(\sum_{j=1}^{m+n} \rho_1 \ldots \rho_{j-1}(1-\rho_j)\mathbb{1}_{\{Z_j \in \mathcal{X}\}} > \frac{1}{2}\bigg),$$

and $\mathbb{P}\big(\sum_{j=1}^{\infty} \rho^{j-1}(1-\rho) I_j = 1/2\big) = 0$, it suffices to show that

$$\sum_{j=1}^{m+n} \rho_1 \ldots \rho_{j-1}(1-\rho_j)\mathbb{1}_{\{Z_j \in \mathcal{X}\}} \xrightarrow{d} \sum_{j=1}^{\infty} \rho^{j-1}(1-\rho) I_j$$

as $m \to \infty$, uniformly for continuity points z of f and g such that $f(z) \in [1/C, C]$ and $g(z) \in [1/C, C]$, for each $C \geq 1$. We apply Lemma 3.6.12 with

$$A_{m,d} = \sum_{j=1}^{d} \rho_1 \dots \rho_{j-1}(1-\rho_j)\mathbb{1}_{\{Z_j \in \mathcal{X}\}}$$
$$A_d = \sum_{j=1}^{d} \rho^{j-1}(1-\rho)I_j$$
$$A = \sum_{j=1}^{\infty} \rho^{j-1}(1-\rho)I_j$$

and $d_m = m + n$. Condition (i) of Lemma 3.6.12 was proved in Lemma 3.6.11, (ii) is immediate, and it remains to prove (iii). Given $\epsilon \in (0, \ell)$, there exists $m_0 \in \mathbb{N}$ such that

$$\left|\frac{m_1}{m} - \ell\right| \leq \epsilon \quad \text{and} \quad \left|\frac{n_1}{n} - \ell\right| \leq \epsilon$$

for all $m \geq m_0$. It is this part of the proof which breaks down when $\ell = 0$, since in that case $\ell - \epsilon$ is not positive. Let $m \geq m_0$, $d_0' \in \mathbb{N}$ and $d \in \{d_0' + 1, \dots, m + n\}$. In the with-replacement resampling case,

$$\begin{aligned}
&\sum_{j=d_0'+1}^{d} \rho_1 \dots \rho_{j-1}(1-\rho_j)\mathbb{1}_{\{Z_j \in \mathcal{X}\}} \\
&\qquad \leq \sum_{j=d_0'+1}^{d} \max_{r \in R_j} \left[\left\{ \left(1 - \frac{r}{m}\right)^{m_1} - \left(1 - \frac{r+1}{m}\right)^{m_1} \right\} \left(1 - \frac{j-1-r}{n}\right)^{n_1} \right] \\
&\qquad \leq \sum_{j=d_0'+1}^{d} \max_{r \in R_j} \left\{ \left(1 - \frac{r}{m}\right)^{m_1} \left(1 - \frac{j-1-r}{n}\right)^{n_1} \right\} \\
&\qquad \leq \sum_{j=d_0'+1}^{d} \max_{r \in R_j} \left(e^{-r(\ell-\epsilon) - (j-1-r)(\ell-\epsilon)} \right) \\
&\qquad \leq \frac{e^{-d_0'(\ell-\epsilon)}}{1 - e^{-(\ell-\epsilon)}} \to 0
\end{aligned}$$

as $d'_0 \to \infty$. Similarly, in the without-replacement case,

$$\sum_{j=d'_0+1}^{d} \rho_1 \dots \rho_{j-1}(1-\rho_j)\mathbb{1}_{\{Z_j \in \mathcal{X}\}} \leq \sum_{j=d'_0+1}^{d} \binom{m_1}{m} \max_{r \in R_j} \left\{ \left(1 - \frac{r}{m}\right)^{m_1} \left(1 - \frac{j-1-r}{n}\right)^{n_1} \right\}$$
$$\leq \frac{e^{-d'_0(\ell-\epsilon)}}{1 - e^{-(\ell-\epsilon)}} \to 0$$

as $d'_0 \to \infty$. This proves a slightly stronger version of (iii), and the theorem therefore follows. □

The asymptotic risk of the bagged nearest-neighbour classifier is given below for the case where the sampling ratios both converge to the same, non-zero limit. It is derived from Theorem 3.6.13 in the same way as the asymptotic risk of the nearest-neighbour classifier was obtained in Theorem 3.6.7 from Theorem 3.6.6. The proof is therefore omitted.

Theorem 3.6.14. *In the with-replacement case, suppose both m_1/m and n_1/n converge to $\ell \in (0,1]$ as $m \to \infty$, and set $\rho = e^{-\ell}$. In the without-replacement case, suppose both m_1/m and n_1/n converge to $\ell \in (0,1)$ as $m \to \infty$, and set $\rho = 1 - \ell$. Let $I_1, I_2, \dots$ be independent Bernoulli$\big(q(z)\big)$ random variables, and define*

$$P\big(\rho, q(z)\big) = \mathbb{P}\Big(\sum_{j=1}^{\infty} \rho^{j-1}\,(1-\rho)\,I_j > \frac{1}{2}\Big).$$

Then

$$\mathrm{Risk}(\mathcal{C}_{\mathrm{Bagg}}) \to p \int_{\mathbb{R}^k} \big\{1 - P\big(\rho, q(z)\big)\big\}\, f(z)\, dz + (1-p) \int_{\mathbb{R}^k} P\big(\rho, q(z)\big)\, g(z)\, dz$$

as $m \to \infty$.

It only remains to deal with the $\ell = 0$ case. This was the content of Theorem 3.4.1.

Proof of **Theorem 3.4.1**.

Recall that

$$\mathbb{P}\big(\mathcal{C}_{\mathrm{Bagg}}(z) = X\big) = \mathbb{P}\big(\pi(z|\mathcal{Z}) > 1/2\big),$$

where

$$\pi(z|\mathcal{Z}) = \sum_{j=1}^{m+n} \rho_1 \dots \rho_{j-1}(1-\rho_j)\mathbb{1}_{\{Z_j \in \mathcal{X}\}},$$

and ρ_j and Z_j are defined prior to Lemma 3.6.11. Recall also from the proof of Lemma 3.6.11 that $R_j = \big\{r \in \{0,1,\dots,j-1\} : r \le m,\ j-1-r \le n\big\}$. We prove the first part of Theorem 3.4.1; the second part follows analogously. The first step is to prove that $\mathrm{Var}\big(\pi(z|\mathcal{Z})\big) \to 0$ as $m \to \infty$, uniformly for $z \in \mathcal{S}_X(\eta,\epsilon)$, for each $\eta, \epsilon > 0$.

Fix $\eta, \epsilon > 0$. Since a random variable bounded between 0 and a, for some $a > 0$, has variance no greater than a^2, we have in the with-replacement case,

$$\begin{aligned}
&\sum_{j=1}^{m+n} \mathrm{Var}\big(\rho_1 \dots \rho_{j-1}(1-\rho_j)\mathbb{1}_{\{Z_j \in \mathcal{X}\}}\big) \\
&\qquad \le \sum_{j=1}^{m+n} \max_{r \in R_j} \Big[\Big\{\Big(1-\frac{r}{m}\Big)^{m_1} - \Big(1-\frac{r+1}{m}\Big)^{m_1}\Big\}\Big(1-\frac{j-1-r}{n}\Big)^{n_1}\Big]^2 \\
&\qquad \le \sum_{j=1}^{m+n} \max_{r \in R_j} \Big[\exp\Big(-\frac{rm_1}{m}\Big)\exp\Big(-(j-1-r)\frac{n_1}{n}\Big)\Big\{1-\Big(\frac{m-(r+1)}{m-r}\Big)^{m_1}\Big\}\Big]^2 \\
&\qquad \le \Big\{1-\Big(1-\frac{1}{m}\Big)^{m_1}\Big\}^2 \sum_{j=1}^{m+n} \exp\Big(-2(j-1)\frac{n_1}{n} + 2(j-1)\Big|\frac{m_1}{m} - \frac{n_1}{n}\Big|\Big).
\end{aligned}$$

To deal with the term outside the sum, observe that, for $m \ge 2$,

$$-m_1 \log\Big(1-\frac{1}{m}\Big) = m_1 \sum_{t=1}^{\infty} \frac{1}{tm^t} \le \frac{m_1}{m}\Big(1 + \sum_{t=1}^{\infty}\frac{1}{m^t}\Big) = \frac{m_1}{m}\Big(1+\frac{1}{m-1}\Big).$$

Now, given $\delta \in (0,1)$, we can find $m_0 \in \mathbb{N}$ such that, for $m \ge m_0$,

$$\frac{1}{m-1} \le \delta, \quad \frac{n_1}{n} - \Big|\frac{m_1}{m} - \frac{n_1}{n}\Big| \ge (1-\delta)\frac{m_1}{m} \quad \text{and} \quad \frac{(1-e^{-(1+\delta)m_1/m})^2}{1-e^{-2(1-\delta)m_1/m}} \le \delta.$$

Then, for $m \geq m_0$,

$$\begin{aligned}
&\sum_{j=1}^{m+n} \mathrm{Var}\big(\rho_1 \ldots \rho_{j-1}(1-\rho_j)\mathbb{1}_{\{Z_j \in \mathcal{X}\}}\big) \\
&\qquad \leq \Big\{1 - \exp\Big(-(1+\delta)\frac{m_1}{m}\Big)\Big\}^2 \sum_{j=1}^{\infty} \exp\Big(-2(j-1)(1-\delta)\frac{m_1}{m}\Big) \\
&\qquad = \frac{\big(1 - e^{-(1+\delta)m_1/m}\big)^2}{1 - e^{-2(1-\delta)m_1/m}} \\
&\qquad \leq \delta.
\end{aligned}$$

Similarly, in the without-replacement case, given $\delta \in (0,1)$, we can find $m_0 \in \mathbb{N}$ such that, for $m \geq m_0$,

$$\frac{n_1}{n} - \Big|\frac{m_1}{m} - \frac{n_1}{n}\Big| \geq (1-\delta)\frac{m_1}{m} \quad \text{and} \quad \Big(\frac{m_1}{m}\Big)^2 \frac{1}{1 - e^{-2(1-\delta)m_1/m}} \leq \delta.$$

In this case,

$$\begin{aligned}
&\sum_{j=1}^{m+n} \mathrm{Var}\big(\rho_1 \ldots \rho_{j-1}(1-\rho_j)\mathbb{1}_{\{Z_j \in \mathcal{X}\}}\big) \\
&\qquad \leq \sum_{j=1}^{m+n} \Big(\frac{m_1}{m}\Big)^2 \max_{r \in R_j} \Big\{\Big(1 - \frac{r}{m}\Big)^{m_1} \Big(1 - \frac{j-1-r}{n}\Big)^{n_1}\Big\}^2 \\
&\qquad \leq \sum_{j=1}^{m+n} \Big(\frac{m_1}{m}\Big)^2 \exp\Big(-2(j-1)\frac{n_1}{n} + 2(j-1)\Big|\frac{m_1}{m} - \frac{n_1}{n}\Big|\Big) \\
&\qquad \leq \Big(\frac{m_1}{m}\Big)^2 \frac{1}{1 - e^{-2(1-\delta)m_1/m}} \\
&\qquad \leq \delta.
\end{aligned}$$

Therefore, to show that $\mathrm{Var}\big(\pi(z|\mathcal{Z})\big) \to 0$ as $m \to \infty$, uniformly for $z \in \mathcal{S}_X(\eta, \epsilon)$, it remains to prove that

$$\sum_{i \neq j} \mathrm{Cov}\big(\rho_1, \ldots, \rho_{i-1}(1-\rho_i)\mathbb{1}_{\{Z_i \in \mathcal{X}\}}\,,\, \rho_1, \ldots, \rho_{j-1}(1-\rho_j)\mathbb{1}_{\{Z_j \in \mathcal{X}\}}\big) \to 0$$

as $m \to \infty$, uniformly for $z \in \mathcal{S}_X(\eta, \epsilon)$. But, for each $z \in \mathcal{S}_X(\eta, \epsilon)$,

$$\sum_{i \neq j} \bigl| \mathrm{Cov}\bigl(\rho_1, \ldots, \rho_{i-1}(1-\rho_i)\mathbb{1}_{\{Z_i \in \mathcal{X}\}}\,,\, \rho_1, \ldots, \rho_{j-1}(1-\rho_j)\mathbb{1}_{\{Z_j \in \mathcal{X}\}}\bigr) \bigr|$$

$$\leq \sum_{i \neq j} \Bigl\{ \min_{r \in R_i} p_r^i \min_{s \in R_j} p_s^j \bigl| \mathrm{Cov}(\mathbb{1}_{\{Z_i \in \mathcal{X}\}}, \mathbb{1}_{\{Z_j \in \mathcal{X}\}}) \bigr| + \max_{r \in R_i} p_r^i \max_{s \in R_j} p_s^j - \min_{r \in R_i} p_r^i \min_{s \in R_j} p_s^j \Bigr\}$$

where, in the with-replacement case,

$$p_r^i = \Bigl\{ \Bigl(1 - \frac{r}{m}\Bigr)^{m_1} - \Bigl(1 - \frac{r+1}{m}\Bigr)^{m_1} \Bigr\} \Bigl(1 - \frac{i-1-r}{n}\Bigr)^{n_1}.$$

Now observe that, given $\delta \in (0,1)$, for sufficiently large m, we have

$$\sum_{i \neq j} \Bigl\{ \max_{r \in R_i} p_r^i \max_{s \in R_j} p_s^j - \min_{r \in R_i} p_r^i \min_{s \in R_j} p_s^j \Bigr\}$$

$$\leq \sum_{i \neq j} \Bigl[\Bigl\{ 1 - \Bigl(1 - \frac{1}{m}\Bigr) \Bigr\}^2 e^{-(i-1+j-1)(1-\delta)m_1/m}$$

$$- \min_{r \in R_i} \min_{s \in R_j} \bigl\{ e^{-rm_1/m} - e^{-(r+1)m_1/m} \bigr\} \bigl\{ e^{-sm_1/m} - e^{-(s+1)m_1/m} \bigr\} e^{-(i-1-r+j-1-s)(1+\delta)n_1/n} \Bigr]$$

$$\leq \sum_{i \neq j} \Bigl\{ (1 - e^{-(1+\delta)m_1/m})^2 e^{-(i-1+j-1)(1-\delta)m_1/m} - e^{-(i-1+j-1)(1+2\delta)m_1/m} (1 - e^{-m_1/m})^2 \Bigr\}$$

$$\leq \frac{\bigl\{1 - e^{-(1+\delta)m_1/m}\bigr\}^2}{\bigl\{1 - e^{-(1-\delta)m_1/m}\bigr\}^2} - \Bigl\{ \frac{(1-\delta)(1-e^{-m_1/m})^2}{(1 - e^{-(1+2\delta)m_1/m})^2} - \frac{(1-\delta)(1-e^{-m_1/m})^2}{1 - e^{-2(1+2\delta)m_1/m}} \Bigr\}.$$

But

$$\frac{\bigl\{1 - e^{-(1+\delta)x}\bigr\}^2}{\bigl\{1 - e^{-(1-\delta)x}\bigr\}^2} - \Bigl\{ \frac{(1-\delta)(1-e^{-x})^2}{(1 - e^{-(1+2\delta)x})^2} - \frac{(1-\delta)(1-e^{-x})^2}{1 - e^{-2(1+2\delta)x}} \Bigr\} \to \frac{(1+\delta)^2}{(1-\delta)^2} - \frac{(1-\delta)}{(1+2\delta)^2},$$

as $x \to 0$, and this limit may be made arbitrarily close to zero by choosing $\delta > 0$ sufficiently small. Moreover,

$$\sum_{i \neq j} \min_{r \in R_i} p_r^i \min_{s \in R_j} p_s^j \bigl| \mathrm{Cov}(\mathbb{1}_{\{Z_i \in \mathcal{X}\}}, \mathbb{1}_{\{Z_j \in \mathcal{X}\}}) \bigr| \to 0$$

as $m \to \infty$, uniformly for $z \in \mathcal{S}_X(\eta, \epsilon)$, by assumption (**A3**) and the fact that

$$\sum_{i \neq j} \min_{r \in R_i} p_r^i \min_{s \in R_j} p_s^j \leq 1.$$

The without-replacement case may be handled similarly. The next part of the proof consists of showing that

$$\liminf_{m\to\infty} \inf_{z\in\mathcal{S}_X(\eta,\epsilon)} \mathbb{E}\big(\pi(z|\mathcal{Z})\big) > \frac{1}{2}.$$

Let M and N be the number of distinct values in $\mathcal{X}^*$ and $\mathcal{Y}^*$, respectively. For example, in the case of without-replacement resampling, $M = m_1$ and $N = n_1$. Now, $\mathbb{E}\big(\pi(z|\mathcal{Z})\big) = \mathbb{E}\big(Q(z|M,N)\big)$, where

$$Q(z|M,N) = \mathbb{P}(\text{Nearest element of } \mathcal{Z}^* \text{ to } z \text{ is in } \mathcal{X}|M,N).$$

Writing

$$F_X(\delta|z) = 1 - \big(1-\pi_X(\delta|z)\big)^M \quad \text{and} \quad F_Y(\delta|z) = 1 - \big(1-\pi_Y(\delta|z)\big)^N$$

for the distribution functions of the nearest element of $\mathcal{X}^*$ to z and the nearest element of $\mathcal{Y}^*$ to z respectively, we have

$$\begin{aligned} Q(z|M,N) &= \int_0^\infty \big(1-F_Y(\delta|z)\big)\, dF_X(\delta|z) \\ &= M \int_0^\infty \big(1-\pi_Y(\delta|z)\big)^N \big(1-\pi_X(\delta|z)\big)^{M-1} d\pi_X(\delta|z) \\ &\equiv \frac{1}{2} + Q_1(z|M,N), \end{aligned}$$

where

$$Q_1(z|M,N) = M \int_0^\infty \big(1-\pi_X(\delta|z)\big)^{2M-1} \left\{ \frac{\big(1-\pi_Y(\delta|z)\big)^N}{\big(1-\pi_X(\delta|z)\big)^M} - 1 \right\} d\pi_X(\delta|z).$$

The distributions of M and N depend only on m, n, m_1 and n_1, not on z or on the distributions of X or Y. Moreover, we can write $M = (1-\Delta_1)m_1$ and $N = (1-\Delta_2)n_1$, where Δ_1 and Δ_2 are independent random variables satisfying $\mathbb{P}(0 \le \Delta_j \le 1) = 1$ and $\mathbb{P}(\Delta_j > \gamma) \to 0$, as $m \to \infty$, for each $\gamma > 0$ and for $j = 1, 2$. Observe that, for

any $\gamma > 0$,

$$\begin{aligned}
&\sup_{z \in \mathcal{S}_X(\eta,\epsilon)} \mathbb{E}\big\{ |Q_1|(1 - \mathbb{1}_{\{\Delta_1 \le \gamma,\, \Delta_2 \le \gamma\}})\big\} \\
&\qquad \le \sup_{z \in \mathcal{S}_X(\eta,\epsilon)} \mathbb{E}\Big\{ 2M \int_0^\infty \big(1 - \pi_X(\delta|z)\big)^{M-1} d\pi_X(\delta|z)\, (1 - \mathbb{1}_{\{\Delta_1 \le \gamma,\, \Delta_2 \le \gamma\}})\Big\} \\
&\qquad = \mathbb{E}\Big\{ 2M \int_0^1 y^{M-1}\, dy\, (1 - \mathbb{1}_{\{\Delta_1 \le \gamma,\, \Delta_2 \le \gamma\}})\Big\} \\
&\qquad = 2\mathbb{E}(1 - \mathbb{1}_{\{\Delta_1 \le \gamma,\, \Delta_2 \le \gamma\}}) \\
&\qquad \to 0 \qquad (3.12)
\end{aligned}$$

as $m \to \infty$. Moreover, if $\Delta_1 \le \gamma$ and $\Delta_2 \le \gamma$, then

$$\log\Big(\frac{\big(1 - \pi_Y(\delta|z)\big)^N}{\big(1 - \pi_X(\delta|z)\big)^M} \Big) \ge n_1 \log\big(1 - \pi_Y(\delta|z)\big) - (1-\gamma) m_1 \log\big(1 - \pi_X(\delta|z)\big).$$

Now, for any $\pi_Y \in [0, \xi]$, where $\xi < 1$, we have

$$\log(1 - \pi_Y) = -\pi_Y(1 + \theta_Y), \quad \text{where} \quad |\theta_Y| = \Big| \sum_{i=1}^\infty \frac{\pi_Y^i}{i+1} \Big| \le \frac{\pi_Y}{2(1-\pi_Y)} \le \frac{\xi}{2(1-\xi)}.$$

Thus, since $-\log(1-x) \ge x$ for all $x \ge 0$,

$$\begin{aligned}
\log\Big(\frac{\big(1 - \pi_Y(\delta|z)\big)^N}{\big(1 - \pi_X(\delta|z)\big)^M} \Big) &\ge -n_1 \pi_Y(\delta|z)(1 + \theta_Y) + (1-\gamma) m_1 \pi_X(\delta|z) \\
&= m_1 \pi_X(\delta|z) \Big\{ \frac{-n_1 \pi_Y(\delta|z)}{m_1 \pi_X(\delta|z)} (1 + \theta_Y) + 1 - \gamma \Big\}.
\end{aligned}$$

Since $m_1/n_1 \to p/(1-p)$ as $m \to \infty$, we can find $m_0 \in \mathbb{N}$ such that

$$\frac{m_1 \pi_X(\delta|z)}{n_1 \pi_Y(\delta|z)} \ge 1 + \frac{\eta}{2},$$

for all $\delta \in (0, \eta]$, all $z \in \mathcal{S}_X(\eta, \epsilon)$ and all $m \ge m_0$. Now choose $\xi \in (0,1)$ and $\gamma \in (0,1)$ small enough such that

$$\frac{-1}{1 + \eta/2} \Big(1 + \frac{\xi}{2(1-\xi)} \Big) + 1 - \gamma = c_1,$$

for some $c_1 > 0$. Then, for $z \in \mathcal{S}_X(\eta, \epsilon)$ and $\delta \in (0, \eta]$ satisfying $\pi_Y(\delta|z) \in [0, \xi]$, and $m \geq m_0$,

$$\begin{aligned}\frac{\big(1-\pi_Y(\delta|z)\big)^N}{\big(1-\pi_X(\delta|z)\big)^M} - 1 &= \exp\biggl\{\log\Bigl(\frac{\big(1-\pi_Y(\delta|z)\big)^N}{\big(1-\pi_X(\delta|z)\big)^M}\Bigr)\biggr\} - 1 \\ &\geq \exp\bigl\{m_1\pi_X(\delta|z)c_1\bigr\} - 1 \\ &\geq m_1\pi_X(\delta|z)c_1.\end{aligned}$$

Choose δ'_z such that $\pi_Y(\delta'_z|z) = \xi$, set $\delta_z = \min(\delta'_z, \eta)$ and set

$$\xi_0 = \min\Bigl(\frac{\xi(1-p)(1+\eta)}{p}, \epsilon\Bigr),$$

so that $\pi_X(\delta_z|z) \geq \xi_0$ for $z \in \mathcal{S}_X(\eta, \epsilon)$. Then, for $m \geq m_0$,

$$\begin{aligned}&\inf_{z\in\mathcal{S}_X(\eta,\epsilon)} \mathbb{E}\biggl[M\int_0^{\delta_z}\big(1-\pi_X(\delta|z)\big)^{2M-1}\biggl\{\frac{\big(1-\pi_Y(\delta|z)\big)^N}{\big(1-\pi_X(\delta|z)\big)^M}-1\biggr\}d\pi_X(\delta|z)\,\mathbb{1}_{\{\Delta_1\leq\gamma,\,\Delta_2\leq\gamma\}}\biggr] \\ &\geq m_1c_1\,\mathbb{E}\biggl\{M\int_0^{\xi_0}(1-y)^{2M-1}y\,dy\mathbb{1}_{\{\Delta_1\leq\gamma\}}\biggr\}\mathbb{P}(\Delta_2\leq\gamma) \\ &= \frac{m_1c_1}{2}\mathbb{E}\biggl\{\Bigl(-\xi_0(1-\xi_0)^{2M} - \frac{1}{2M+1}(1-\xi_0)^{2M+1} + \frac{1}{2M+1}\Bigr)\mathbb{1}_{\{\Delta_1\leq\gamma\}}\biggr\}\mathbb{P}(\Delta_2\leq\gamma) \\ &\geq \frac{c_1}{8} \qquad (3.13)\end{aligned}$$

for sufficiently large m. Moreover,

$$\begin{aligned}\sup_{z\in\mathcal{S}_X(\eta,\epsilon)}\mathbb{E}\biggl|M\int_{\delta_z}^{\infty}\big(1-\pi_X(\delta|z)\big)^{2M-1}\biggl\{\frac{\big(1-\pi_Y(\delta|z)\big)^N}{\big(1-\pi_X(\delta|z)\big)^M}-1\biggr\}d\pi_X(\delta|z)\,\mathbb{1}_{\{\Delta_1\leq\gamma,\,\Delta_2\leq\gamma\}}\biggr| \\ \leq \mathbb{E}\biggl\{2M\int_{\xi_0}^1(1-y)^{M-1}\,dy\,\mathbb{1}_{\{\Delta_1\leq\gamma\}}\biggr\} \\ = \mathbb{E}\bigl\{2(1-\xi_0)^M\mathbb{1}_{\{\Delta_1\leq\gamma\}}\bigr\} \\ \to 0 \qquad (3.14)\end{aligned}$$

as $m \to \infty$. Combining (3.12), (3.13) and (3.14) yields

$$\liminf_{m\to\infty}\inf_{z\in\mathcal{S}_X(\eta,\epsilon)}\mathbb{E}\bigl(\pi(z|\mathcal{Z})\bigr) = \frac{1}{2} + \liminf_{m\to\infty}\inf_{z\in\mathcal{S}_X(\eta,\epsilon)}\mathbb{E}\bigl(Q_1(z|M,N)\bigr) > \frac{1}{2}.$$

To complete the proof, given $\delta > 0$, choose $m^\dagger \in \mathbb{N}$ and $\alpha > 0$ such that

$$\inf_{z \in \mathcal{S}_X(\eta,\epsilon)} \mathbb{E}\big(\pi(z|\mathcal{Z})\big) - \frac{1}{2} > \alpha \quad \text{and} \quad \sup_{z \in \mathcal{S}_X(\eta,\epsilon)} \mathrm{Var}\big(\pi(z|\mathcal{Z})\big) \leq \alpha^2 \delta.$$

for $m \geq m^\dagger$. Then, by Chebychev's inequality, for $m \geq m^\dagger$,

$$\begin{aligned}
\inf_{z \in \mathcal{S}_X(\eta,\epsilon)} \mathbb{P}\big(\mathcal{C}_{\mathrm{Bagg}}(z) = X\big) &= \inf_{z \in \mathcal{S}_X(\eta,\epsilon)} \mathbb{P}\big(\pi(z|\mathcal{Z}) > 1/2\big) \\
&\geq 1 - \sup_{z \in \mathcal{S}_X(\eta,\epsilon)} \frac{\mathrm{Var}\big(\pi(z|\mathcal{Z})\big)}{\big\{\mathbb{E}\big(\pi(z|\mathcal{Z})\big) - 1/2\big\}^2} \\
&\geq 1 - \delta.
\end{aligned}$$

□

Proof of **Corollary 3.4.2**.

We may write

$$\begin{aligned}
&\mathrm{Risk}(\mathcal{C}_{\mathrm{Bagg}}) \\
&\quad = p\,\mathbb{P}\big(\mathcal{C}_{\mathrm{Bagg}}(X) = Y,\, X \in \mathcal{S}_X(0)\big) + p\,\mathbb{P}\big(\mathcal{C}_{\mathrm{Bagg}}(X) = Y,\, X \in \mathcal{S}_Y(0)\big) \\
&\qquad + (1-p)\mathbb{P}\big(\mathcal{C}_{\mathrm{Bagg}}(Y) = X,\, Y \in \mathcal{S}_X(0)\big) + (1-p)\mathbb{P}\big(\mathcal{C}_{\mathrm{Bagg}}(Y) = X,\, Y \in \mathcal{S}_Y(0)\big).
\end{aligned}$$

Given $\delta > 0$, choose $\eta_0 > 0$ such that

$$\mathbb{P}\big(X \in \mathcal{S}_X(0)\big) - \mathbb{P}\big(X \in \mathcal{S}_X(\eta_0)\big) \leq \delta \quad \text{and} \quad \mathbb{P}\big(Y \in \mathcal{S}_Y(0)\big) - \mathbb{P}\big(Y \in \mathcal{S}_Y(\eta_0)\big) \leq \delta.$$

Now choose $\epsilon_0 > 0$ such that

$$\mathbb{P}\big(X \in \mathcal{S}_X(\eta_0)\big) - \mathbb{P}\big(X \in \mathcal{S}_X(\eta_0, \epsilon_0)\big) \leq \delta \text{ and } \mathbb{P}\big(Y \in \mathcal{S}_Y(\eta_0)\big) - \mathbb{P}\big(Y \in \mathcal{S}_Y(\eta_0, \epsilon_0)\big) \leq \delta,$$

so that

$$\begin{aligned}
\mathrm{Risk}(\mathcal{C}_{\mathrm{Bagg}}) \leq\; & p\,\mathbb{P}\big(\mathcal{C}_{\mathrm{Bagg}}(X) = Y,\, X \in \mathcal{S}_X(\eta_0, \epsilon_0)\big) + p\,\mathbb{P}\big(X \in \mathcal{S}_Y(0)\big) \\
& + (1-p)\mathbb{P}\big(Y \in \mathcal{S}_X(0)\big) + (1-p)\mathbb{P}\big(\mathcal{C}_{\mathrm{Bagg}}(Y) = X,\, Y \in \mathcal{S}_Y(\eta_0, \epsilon_0)\big) + 2\delta.
\end{aligned}$$

By Theorem 3.4.1, we can find $m^\dagger \in \mathbb{N}$ large enough such that, for all $m \geq m^\dagger$,

$$\sup_{z \in \mathcal{S}_X(\eta_0, \epsilon_0)} \mathbb{P}\big(\mathcal{C}_{\text{Bagg}}(z) = Y\big) \leq \delta \quad \text{and} \quad \sup_{z \in \mathcal{S}_Y(\eta_0, \epsilon_0)} \mathbb{P}\big(\mathcal{C}_{\text{Bagg}}(z) = X\big) \leq \delta.$$

Thus

$$\text{Risk}(\mathcal{C}_{\text{Bagg}}) \leq p\,\mathbb{P}\big(X \in \mathcal{S}_Y(0)\big) + (1-p)\mathbb{P}\big(Y \in \mathcal{S}_X(0)\big) + 3\delta$$

for all $m \geq m^\dagger$.

□

Chapter 4

Some asymptotic results for the bootstrap distribution of the sample mean

4.1 Introduction

In Chapter 1, we argued that Edgeworth expansions and saddlepoint approximations have provided much of the theoretical underpinning for the bootstrap. They give a mathematical basis for assessing its properties and comparing its performance with other techniques.

Edgeworth expansions provide the order of magnitude (in probability) of the absolute error between a bootstrap distribution and the true distribution it estimates. Results are now known for many statistics of practical interest, such as smooth functions of a multidimensional sample mean (Hall, 1992) and M-estimators (Lahiri, 1992, 1994),

and are often cited in support of the bootstrap.

Saddlepoint approximations offer a different form of evidence, namely the order of the *relative* error in a bootstrap approximation. This information is particularly relevant when the magnitude of the feature of the underlying population of interest, such as a tail probability, may be very small. Statistics studied from this perspective include sample means (Jing, Feuerverger and Robinson, 1994) and smooth functions of sample means (Robinson and Skovgaard, 1998).

However, these are not the only aspects of bootstrap performance which merit consideration. In this chapter, we take a different approach, as a first attempt to answer the basic question, 'What is the probability that the bootstrap performs badly?'. A mathematical formulation of this problem involves appropriate choices both of a statistic and of a distance between distributions. We work with the univariate sample mean, and the Mallows distance (Mallows, 1972), whose properties were exploited effectively in a bootstrap context by Bickel and Freedman (1981). The main objective is to study the rate of decay of the probability that the distance between the true distribution of the normalised sample mean and its bootstrap approximation exceeds a given threshold.

In Sections 4.2 and 4.3, we review the Mallows distance and show how to reduce our bootstrap problem to one of studying the Mallows distance between a distribution and the empirical distribution of a sample. The main essence of the results in Sections 4.4 and 4.5 is that rate of decay of the probability of poor bootstrap performance depends on the tail of the underlying population. In Section 4.4, we give an explicit bound on the probability of the Mallows distance exceeding a threshold, and show that, under certain tail and smoothness conditions, this bound may decay exponentially in the sample size; that is, the bound is no more than $e^{-n\delta}$, for some $\delta > 0$ and sufficiently large sample sizes n. This may be interpreted as a mathematical statement that

in such cases the probability of poor bootstrap performance decays satisfactorily. However, by choosing a distribution with a sufficiently heavy tail, we can ensure the bound decays no faster than $\exp(-n^{\beta})$, for any given $\beta \in (0,1)$.

Section 4.5 provides further supporting evidence. For example, where the underlying population is of bounded support, it is shown that a large deviations upper bound exists on the probability of the Mallows distance exceeding a threshold. However, for distributions with heavy (polynomial) tails, the empirical distributions fail to satisfy a large deviations principle in the Mallows topology. This shows the delicacy of Sanov's theorem, which says that the empirical measures do satisfy a large deviations principle in the (coarser) weak topology. Results are not known for populations with infinite but light tails, such as the exponential and normal distributions, and these remain interesting topics for further research. The proofs omitted in the main text are given in Section 4.6.

4.2 The Mallows distance on the real line

Let $\mathcal{F}$ denote the set of all distribution functions on the real line and, for $r \geq 1$, let $\mathcal{F}_r = \{F \in \mathcal{F} : \int_{-\infty}^{\infty} |x|^r \, dF(x) < \infty\}$. For $F, G \in \mathcal{F}_r$, the Mallows metric $d_r(F,G)$ is defined by

$$d_r(F,G) = \inf_{\mathcal{T}_{X,Y}} \left\{ \mathbb{E}|X - Y|^r \right\}^{1/r},$$

where $\mathcal{T}_{X,Y}$ is the set of all joint distributions of pairs of random variables X and Y whose marginal distributions are F and G respectively. In a slight abuse of notation, we also write $d_r(X,Y)$ for $d_r(F,G)$, where this will cause no confusion. The following results about d_r are proved in Bickel and Freedman (1981) and Major (1978).

(a) If $(F_n) \in \mathcal{F}$ and $F \in \mathcal{F}$, then $d_r(F_n, F) \to 0$ as $n \to \infty$ if and only if, for every

bounded, continuous function $g : \mathbb{R} \to \mathbb{R}$, we have

$$\int_{-\infty}^{\infty} g(x)\, dF_n(x) \to \int_{-\infty}^{\infty} g(x)\, dF(x)$$

as $n \to \infty$, and also

$$\int_{-\infty}^{\infty} |x|^r\, dF_n(x) \to \int_{-\infty}^{\infty} |x|^r dF(x)$$

as $n \to \infty$. Thus, convergence in the Mallows metric d_r is equivalent to convergence in distribution together with convergence of the rth absolute moments.

(b) If $a \in \mathbb{R}$ and X, Y are random variables with finite rth absolute moments, then

$$d_r(aX, aY) = |a| d_r(X, Y).$$

(c) The infimum in the definition of the Mallows metric is attained by the following construction: let $U \sim U(0,1)$, and define $X = F^{-1}(U)$, $Y = G^{-1}(U)$. Here, F^{-1} and G^{-1} are the left-continuous versions of the respective inverse functions, so that, for example, $F^{-1}(p) = \inf\{x \in \mathbb{R} : F(x) \geq p\}$. Thus

$$d_r(F, G) = \left(\int_0^1 |F^{-1}(p) - G^{-1}(p)|^r\, dp \right)^{1/r}.$$

It is therefore more convenient in much of what follows to work with the set

$$\mathcal{G}_r = \left\{ G : (0,1) \to \mathbb{R} : G \text{ is left-continuous, increasing and } \int_0^1 |G(p)|^r dp < \infty \right\}$$

equipped with the L_r-norm restricted to this set:

$$\|G\|_r = \left(\int_0^1 |G(p)|^r\, dp \right)^{1/r}.$$

The map from $(\mathcal{F}_r, d_r)$ to $(\mathcal{G}_r, \|\cdot\|_r)$ which sends a distribution function to its left-continuous inverse is a distance-preserving bijection.

(d) Suppose X and Y have distributions in $\mathcal{F}_2$. Then

$$d_2(X, Y)^2 = d_2\big(X - \mathbb{E}(X)\,,\, Y - \mathbb{E}(Y)\big)^2 + \big(\mathbb{E}(X) - \mathbb{E}(Y)\big)^2.$$

(e) Suppose $X_1, \ldots, X_n$ are independent, $Y_1, \ldots, Y_n$ are independent, that all the distributions are in $\mathcal{F}_2$, and that $\mathbb{E}(X_i) = \mathbb{E}(Y_i)$ for $i = 1, \ldots, n$. Then

$$d_2\Big(\sum_{i=1}^n X_i\,,\,\sum_{i=1}^n Y_i\Big)^2 \le \sum_{i=1}^n d_2(X_i, Y_i)^2.$$

Equality is attained if $X_1, \ldots, X_n$ are independent $N(\mu, \sigma_X^2)$ random variables and $Y_1, \ldots, Y_n$ are independent $N(\mu, \sigma_Y^2)$ random variables, for some $\mu \in \mathbb{R}$ and $\sigma_X^2, \sigma_Y^2 \ge 0$.

We add two further properties in Proposition 4.2.1 below. The completeness of the Mallows metric on a separable metric space is already known (Dobrushin, 1970), but we can give a much simpler argument for the case of distributions on the real line.

Proposition 4.2.1. *The metric space $(\mathcal{F}_r, d_r)$ is separable and complete.*

Proof. To show separability, consider the set

$$\mathcal{H}_r = \{H \in \mathcal{G}_r : H(p) \in \mathbb{Q} \text{ for all } p \in \mathbb{Q} \cap (0,1)\}.$$

Note that any function in $\mathcal{G}_r$ is determined by its values at the rational points in $(0,1)$, and that $\mathcal{H}_r$ is countable. Moreover, given $\epsilon > 0$ and any $G \in \mathcal{G}_r$, we can choose values $H(p)$ for $p \in \mathbb{Q} \cap (0,1)$ such that

$$|H(p) - G(p)| \le \epsilon$$

and $H(p_1) \le H(p_2)$ whenever $p_1 \le p_2$. Extending H to a left-continuous function on $(0,1)$ (which is necessarily increasing), we have $|H(p) - G(p)| \le \epsilon$ for all $p \in (0,1)$, so

$$\|H - G\|_r \le \epsilon,$$

and moreover,

$$\|H\|_r \le \|G\|_r + \|H - G\|_r \le \|G\|_r + \epsilon,$$

so $H \in \mathcal{H}_r$. Hence $\mathcal{H}_r$ is dense in $\mathcal{G}_r$, for each r. Consequently, the distribution functions corresponding to the functions in $\mathcal{H}_r$ are dense in $\mathcal{F}_r$.

Now suppose (F_n) is a Cauchy sequence in $(\mathcal{F}_r, d_r)$. Let $U \sim U(0,1)$, and for each $n \in \mathbb{N}$, let $X_n = F_n^{-1}(U)$. Then X_n has distribution function F_n, and for each $m, n \in \mathbb{N}$, we have

$$d_r(F_m, F_n) = \left(\mathbb{E}|X_m - X_n|^r\right)^{1/r}.$$

Thus (X_n) is a Cauchy sequence in L_r. But L_r is complete (Billingsley, 1995, p. 243), so there exists a random variable $X \in L_r$ such that

$$\mathbb{E}|X_n - X|^r \to 0$$

as $n \to \infty$. Hence if F is the distribution function of X, then

$$d_r(F_n, F) \leq \left(\mathbb{E}|X_n - X|^r\right)^{1/r} \to 0$$

as $n \to \infty$. □

4.3 The Mallows distance and the bootstrap

Suppose $X_1, \ldots, X_n$ are independent random variables, each having distribution function F with mean μ and finite variance. Let $\bar{X}_n = n^{-1} \sum_{i=1}^n X_i$ denote the sample mean. If we are interested in making inference about μ, a natural root to consider is $n^{1/2}(\bar{X}_n - \mu)$, whose sampling distribution under F we denote by $H_n(F)$. Conditional on $X_1, \ldots, X_n$, let $X_1^*, \ldots, X_n^*$ be independent and identically distributed random variables drawn from the empirical distribution of the sample, whose distribution function, $\hat{F}_n$, is given by

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{X_i \leq x\}}$$

for $x \in \mathbb{R}$. The nonparametric bootstrap approximates the sampling distribution of $n^{1/2}(\bar{X}_n - \mu)$ by that of $n^{1/2}(\bar{X}_n^* - \bar{X}_n)$, where $\bar{X}_n^* = n^{-1}\sum_{i=1}^n X_i^*$. In other words, conditional on $X_1, \ldots, X_n$, we approximate $H_n(F)$ by $H_n(\hat{F}_n)$. The calculation below, which follows Shao and Tu (1995), shows how the properties of the Mallows distance d_2 outlined in Section 4.2 make it suitable for studying the performance of the bootstrap approximation in this context:

$$\begin{aligned} d_2\big(H_n(\hat{F}_n), H_n(F)\big) &= d_2\left(\frac{1}{n^{1/2}}\sum_{i=1}^n (X_i^* - \bar{X}_n)\,,\ \frac{1}{n^{1/2}}\sum_{i=1}^n (X_i - \mu)\right) \\ &\leq \frac{1}{n^{1/2}}\left(\sum_{i=1}^n d_2(X_i^* - \bar{X}_n\,,\ X_i - \mu)^2\right)^{1/2} \\ &= d_2(X_1^* - \bar{X}_n\,,\ X_1 - \mu) \\ &\leq d_2(X_1^*, X_1) \\ &= d_2(\hat{F}_n, F). \end{aligned}$$

Thus, in particular, the distance between the distribution of the root of interest, $H_n(F)$, and its bootstrap approximation, $H_n(\hat{F}_n)$, is stochastically dominated by the distance between the true and empirical distributions.

It follows immediately by property (a) in Section 4.2 and the strong law of large numbers that $d_2(\hat{F}_n, F) \to 0$ almost surely as $n \to \infty$. In straightforward cases, we can give a limiting distribution for $n^{1/2}d_2(\hat{F}_n, F)$, as in Theorem 4.3.2 below. Let $D = D[0,1]$ denote the space of left-continuous, real-valued functions on $[0,1]$ possessing right limits at each point. We may equip D with the uniform norm, so that for $x, y \in D$, we define

$$\|x - y\|_\infty = \sup_{p \in [0,1]} |x(p) - y(p)|.$$

A technical complication in Theorem 4.3.2 arises from the fact that the normed space $(D, \|\cdot\|_\infty)$ is non-separable, and the σ-algebra, $\mathcal{D}$, generated by the open balls is

strictly smaller than the Borel σ-algebra, $\mathcal{D}_{\text{Borel}}$, generated by the open sets. This creates measurability problems, as explained in Chibisov (1965), which lead us to work with the space $(D, \mathcal{D}, \|\cdot\|_\infty)$. A consequence of using the ball σ-algebra is that we must make a slight modification to the notion of weak convergence, in line with Billingsley (1999), p. 67:

Definition 4.3.1. *If $(Y_n)_{n\geq 0}$ is a sequence of random elements of $(D, \mathcal{D}, \|\cdot\|_\infty)$, we write $Y_n \xrightarrow{d^\circ} Y_0$ as $n \to \infty$ if*

$$\mathbb{E}\big(f(Y_n)\big) \to \mathbb{E}\big(f(Y_0)\big)$$

as $n \to \infty$, for all bounded, continuous functions $f : D \to \mathbb{R}$ which are $\mathcal{D}$-measurable.

Recall that a Brownian bridge $B = \big(B(p)\big)_{0\leq p\leq 1}$ is a zero mean Gaussian process with

$$\operatorname{Cov}\big(B(p), B(q)\big) = p(1-q)$$

for $p \leq q$. For $p \in (0,1)$, let $\xi_p = \inf\{x \in \mathbb{R} : F(x) \geq p\}$.

Theorem 4.3.2. *Suppose that the limits $\xi_0 = \lim_{p\searrow 0} \xi_p$ and $\xi_1 = \lim_{p\nearrow 1} \xi_p$ exist in $\mathbb{R}$, and that F has a density f such that $f(\xi_p)$ is positive and continuous for $p \in [0,1]$. Let $B = \big(B(p)\big)_{0\leq p\leq 1}$ denote a Brownian bridge. Then*

$$n^{1/2} d_2(\hat{F}_n, F) \xrightarrow{d} \left(\int_0^1 \frac{B^2(p)}{f^2(\xi_p)}\, dp\right)^{1/2}$$

as $n \to \infty$.

Proof. Theorem 1 on pp. 640–641 of Shorack and Wellner (1986), together with Corollary 1 on p. 48 of the same book, give that

$$f(\xi_p)\, n^{1/2}\big(\hat{F}_n^{-1}(p) - \xi_p\big) \xrightarrow{d^\circ} B(p)$$

on $(D, \mathcal{D}, \|\cdot\|_\infty)$, as $n \to \infty$. Now, with probability one, B belongs to the space $(C[0,1], \|\cdot\|_\infty)$ of continuous real-valued functions on $[0,1]$ equipped with the uniform norm, and moreover this space is separable. We can therefore apply the version of the continuous mapping theorem for $\xrightarrow{d^\circ}$ convergence (Billingsley, 1999, pp. 67–68) to a composition map $h(p) = h_2\big(h_1(p)\big)$ from $(D, \mathcal{D}, \|\cdot\|_\infty)$ to $\mathbb{R}$. The individual maps $h_1 : (D, \mathcal{D}, \|\cdot\|_\infty) \to (D, \mathcal{D}, \|\cdot\|_\infty)$ and $h_2 : (D, \mathcal{D}, \|\cdot\|_\infty) \to \mathbb{R}$ are defined by

$$h_1(G)(p) = \frac{G^2(p)}{f^2(\xi_p)} \quad \text{and} \quad h_2(G) = \left(\int_0^1 G(p)\,dp\right)^{1/2}.$$

Observe that the continuity of h_1 follows from the fact that $f(\xi_p)$ attains its (positive) infimum for some $p \in [0,1]$. We conclude that

$$n^{1/2} d_2(\hat{F}_n, F) = \left(\int_0^1 n\big(\hat{F}_n^{-1}(p) - \xi_p\big)^2\,dp\right)^{1/2} \xrightarrow{d^\circ} \left(\int_0^1 \frac{B^2(p)}{f^2(\xi_p)}\,dp\right)^{1/2}$$

as $n \to \infty$. The result follows on noting that any bounded, continuous function from $\mathbb{R}$ to $\mathbb{R}$ is (Borel) measurable. □

4.4 An exponential bound?

The inequality below, derived in Serfling (1980), pp. 75–76, from a lemma of Hoeffding (1963), is crucial for obtaining the main bound of this section in Theorem 4.4.3.

Lemma 4.4.1. *Let $F \in \mathcal{F}$, and suppose $p \in (0,1)$ is such that there exists a unique $x \in \mathbb{R}$ such that $F(x_-) \le p \le F(x)$, where $F(x_-) = \lim_{y \nearrow x} F(y)$. Then, for any $\epsilon > 0$,*

$$\mathbb{P}(|\hat{F}_n^{-1}(p) - \xi_p| \ge \epsilon) \le 2e^{-2n\delta_\epsilon^2},$$

where $\delta_\epsilon = \min\{F(\xi_p + \epsilon) - p\,,\, p - F(\xi_p - \epsilon)\}$.

Let $B = \{p \in (0,1) :$ there exist $x_0 < x_1$ satisfying $F(x_0) = F(x_1) = p\}$. If $p \in B$, then F is constant in a right-neighbourhood of ξ_p, so B is countable.

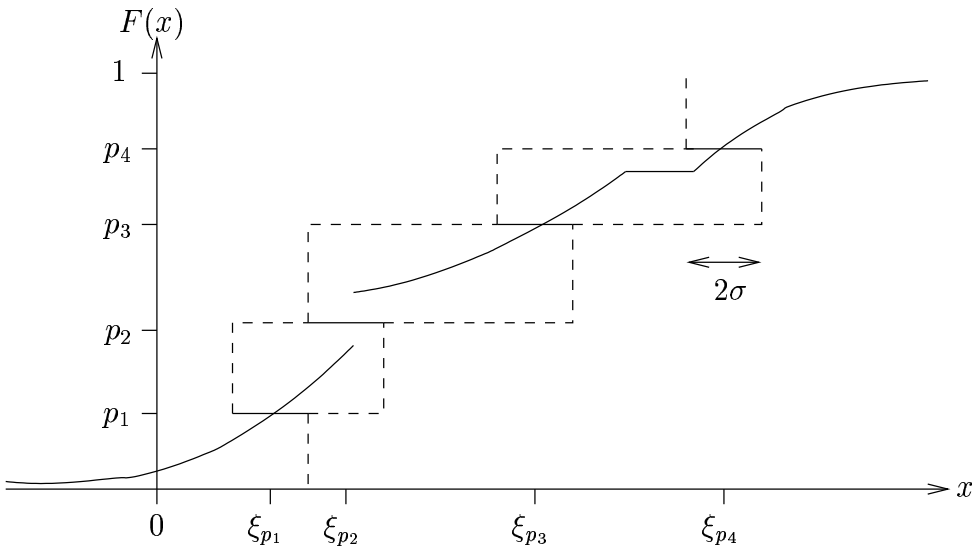


Figure 4.1: A plot of a distribution function F and the construction used in the proof of Theorem 4.4.2.

Theorem 4.4.2. *Let $F \in \mathcal{F}_2$, $\sigma > 0$ and $n \in \mathbb{N}$. Suppose $p_1, \ldots, p_n$ are such that $p_i \in \big((i-1)/n, i/n\big]$ and $p_i \notin B$ for $i = 1, \ldots, n$. Let*

$$\epsilon = \int_0^{p_1} (\xi_{p_1} + \sigma - \xi_p)^2 \, dp + \sum_{i=1}^{n-1} (\xi_{p_{i+1}} - \xi_{p_i} + 2\sigma)^2 (p_{i+1} - p_i) + \int_{p_n}^1 \big(\xi_p - (\xi_{p_n} - \sigma)\big)^2 \, dp.$$

Then

$$\mathbb{P}\big(d_2(\hat{F}_n, F)^2 > \epsilon\big) \leq 2 \sum_{i=1}^{n} e^{-2n\delta_\sigma(p_i)^2},$$

where $\delta_\sigma(p) = \min\{F(\xi_p + \sigma) - p \, , \, p - F(\xi_p - \sigma)\}$.

Proof. See Figure 4.1. We write

$$\begin{aligned} d_2(\hat{F}_n, F)^2 &= \int_0^1 \big(\hat{F}_n^{-1}(p) - \xi_p\big)^2 \, dp \\ &= \int_0^{p_1} \big(\hat{F}_n^{-1}(p) - \xi_p\big)^2 \, dp + \sum_{i=1}^{n-1} \int_{p_i}^{p_{i+1}} \big(\hat{F}_n^{-1}(p) - \xi_p\big)^2 \, dp + \int_{p_n}^1 \big(\hat{F}_n^{-1}(p) - \xi_p\big)^2 \, dp. \end{aligned}$$

Observe that $\hat{F}_n^{-1}(p)$ is constant for $p \in (0, p_1)$, so that if $\xi_{p_1} - \sigma \le \hat{F}_n^{-1}(p_1) \le \xi_{p_1} + \sigma$, then

$$\int_0^{p_1} \left(\hat{F}_n^{-1}(p) - \xi_p\right)^2 dp \le \int_0^{p_1} (\xi_{p_1} + \sigma - \xi_p)^2 \, dp.$$

A similar argument applies for the interval $(p_n, 1)$, and since $\hat{F}_n$ is a increasing function, it follows that for $i = 1, \ldots, n-1$,

$$\int_{p_i}^{p_{i+1}} \left(\hat{F}_n^{-1}(p) - \xi_p\right)^2 dp \le (\xi_{p_{i+1}} - \xi_{p_i} + 2\sigma)^2 (p_{i+1} - p_i),$$

whenever

$$\xi_{p_i} - \sigma \le \hat{F}_n^{-1}(p_i) \le \xi_{p_i} + \sigma \quad \text{and} \quad \xi_{p_{i+1}} - \sigma \le \hat{F}_n^{-1}(p_{i+1}) \le \xi_{p_{i+1}} + \sigma.$$

For $i = 1, \ldots, n$, let

$$B_i = \{\xi_{p_i} - \sigma \le \hat{F}_n^{-1}(p_i) \le \xi_{p_i} + \sigma\}.$$

Then by Lemma 4.4.1,

$$\mathbb{P}\big(d_2(\hat{F}_n, F)^2 > \epsilon\big) \le \mathbb{P}\Big(\bigcup_{i=1}^n B_i^c\Big) \le \sum_{i=1}^n \mathbb{P}(B_i^c) \le 2 \sum_{i=1}^n e^{-2n\delta_\sigma(p_i)^2},$$

where $\delta_\sigma(p)$ is as stated in the theorem. □

We are particularly interested in values of $p_1, \ldots, p_n$ satisfying

(a) $p_1 \in \big(1/(2n), 1/n\big]$, $p_n \in \big(1 - 1/n, 1 - 1/(2n)\big]$ and $p_i \in \big(i/n, (i+1)/n\big]$ for $i = 2, \ldots, n-1$;

(b) $p_i \notin B$ for $i = 1, \ldots, n$.

Theorem 4.4.3. *Given any $\epsilon > 0$, there exists $\sigma > 0$ such that for sufficiently large n, and all $p_1, \ldots, p_n$ satisfying conditions (a) and (b) above, we have*

$$\mathbb{P}\big(d_2(\hat{F}_n, F)^2 > \epsilon\big) \le 2 \sum_{i=1}^n e^{-2n\delta_\sigma(p_i)^2}. \tag{4.1}$$

Proof. The proof is a matter of showing that the positive ϵ in Theorem 4.4.2 can be made arbitrarily small by choosing $\sigma > 0$ suitably small and n sufficiently large. Observe that for $x > 0$,

$$x^2\big(1 - F(x)\big) = x^2 \int_x^\infty dF(y) \le \int_x^\infty y^2\, dF(y),$$

and since $F \in \mathcal{F}_2$, we may apply the dominated convergence theorem to conclude that $1 - F(x) = o(x^{-2})$ as $x \to \infty$, and similarly $F(x) = o(x^{-2})$ as $x \to -\infty$. Thus $(1-p)\xi_p^2 \to 0$ as $p \to 1$ and $p\xi_p^2 \to 0$ as $p \to 0$. Hence, given $\epsilon > 0$, we may choose n_0 large enough such that

$$\xi_{1/(2n)}^2 \le \frac{\epsilon n}{16} \quad \text{and} \quad \xi_{1-1/(2n)}^2 \le \frac{\epsilon n}{16}$$

as well as

$$\int_0^{1/n} (\xi_{1/n} - \xi_p)^2\, dp \le \frac{\epsilon}{8} \quad \text{and} \quad \int_{1-1/n}^1 (\xi_p - \xi_{1-1/n})^2\, dp \le \frac{\epsilon}{8}$$

for $n \ge n_0$. For such n, and for $p_1, \ldots, p_n$ satisfying conditions (a) and (b),

$$\begin{aligned}
\sum_{i=1}^{n-1} (\xi_{p_{i+1}} - \xi_{p_i})^2 (p_{i+1} - p_i) &\le \max_{1 \le i \le n-1} (p_{i+1} - p_i)(\xi_{p_n} - \xi_{p_1})^2 \\
&\le \frac{2}{n}\big(\xi_{1-1/(2n)} - \xi_{1/(2n)}\big)^2 \\
&\le \frac{4}{n}\big(\xi_{1-1/(2n)}^2 + \xi_{1/(2n)}^2\big) \\
&\le \frac{\epsilon}{2}.
\end{aligned}$$

Finally, choose $\sigma > 0$ small enough such that, for all $n \ge n_0$,

$$\int_0^{1/n} (\xi_{1/n} + \sigma - \xi_p)^2\, dp \le \frac{\epsilon}{6}, \quad \int_{1-1/n}^1 \big(\xi_p - (\xi_{1-1/n} - \sigma)\big)^2\, dp \le \frac{\epsilon}{6},$$

and

$$\sum_{i=1}^{n-1} (\xi_{p_{i+1}} - \xi_{p_i} + 2\sigma)^2 (p_{i+1} - p_i) \le \frac{2\epsilon}{3}.$$

Theorem 4.4.2 now completes the proof. □

Remark: Since $d_2(\hat{F}_n, F)$ is stochastically dominated by $d_2\big(H_n(\hat{F}_n), H_n(F)\big)$, the same bound (4.1) holds for $\mathbb{P}\big\{d_2\big(H_n(\hat{F}_n), H_n(F)\big)^2 > \epsilon\big\}$ under the conditions of the theorem.

Although the bound (4.1) appears at first sight to give a very satisfactory mathematical answer to the original question posed in the introduction concerning the probability of poor bootstrap performance for the sample mean, it is in fact not always the case that (4.1) is a genuine exponential bound in n. For instance, if for $x > 1$ and some $m > 3$,

$$F(x) = 1 - \frac{1}{x^{m-1}},$$

so that F has density $f(x) = (m-1)/x^m$ for $x > 1$, and $\xi_p = (1-p)^{-1/(m-1)}$, then

$$\sum_{i=1}^{n} e^{-2n\delta_\sigma(p_i)^2} \geq e^{-2n\sigma^2 f^2(\xi_{1-1/n})} = \exp\big(-2(m-1)^2\sigma^2 n^{(m-3)/(m-1)}\big).$$

Note that the power of n may be made arbitrarily close to zero by choosing m sufficiently close to 3. The problems here are caused by the heavy tails in the underlying distribution. The following result, however, gives simple conditions under which the bound (4.1) decays exponentially in n.

Corollary 4.4.4. *Suppose that the limits $\xi_0 = \lim_{p\searrow 0}\xi_p$ and $\xi_1 = \lim_{p\nearrow 1}\xi_p$ exist in $\mathbb{R}$, and that F has a density f such that $f(\xi_p)$ is positive and continuous for $p \in [0,1]$. Then given any $\epsilon > 0$, there exists $\delta = \delta(\epsilon) > 0$ such that, for all sufficiently large $n \in \mathbb{N}$,*

$$\mathbb{P}\big(d_2(\hat{F}_n, F)^2 > \epsilon\big) \leq e^{-n\delta}.$$

Proof. Let

$$\alpha = \inf_{p\in[0,1]} f(\xi_p),$$

so that $\alpha > 0$. Then by the mean value theorem, $\delta_\sigma(p) \geq \alpha\sigma$ for each $p \in [0,1]$. By Theorem 4.4.3 therefore, given $\epsilon > 0$, there exists $\sigma = \sigma(\epsilon) > 0$ such that, for

sufficiently large n,

$$\mathbb{P}\big(d_2(\hat{F}_n, F)^2 > \epsilon\big) \leq 2ne^{-2n\alpha^2\sigma^2}.$$

Hence the result holds for any $\delta \in (0, 2\alpha^2\sigma^2)$. □

4.5 A Large Deviations Principle?

In the light of Section 4.4, it is natural to ask whether the sequence of empirical distribution functions $(\hat{F}_n)$ satisfies a large deviations principle (LDP) in the topology generated by the Mallows metric. The answer is in general negative, though it is true under certain conditions which are described in this section. First, we recall some standard definitions and results on large deviations, which may be found in Dembo and Zeitouni (1995).

Definition 4.5.1. *Let $\mathcal{X}$ be a topological space. A function $I : \mathcal{X} \to [0, \infty]$ is called a rate function if it is lower semi-continuous; that is, if for each $\alpha \in [0, \infty)$, the level set $\{x \in \mathcal{X} : I(x) \leq \alpha\}$ is closed. A good rate function is a rate function for which the level sets are compact subsets of $\mathcal{X}$.*

Let $\bar{A}$ and A° denote the closure and interior, respectively, of a set A, and let $\mathcal{B}$ denote the Borel σ-algebra of $\mathcal{X}$. Write $M(\mathcal{X})$ for the space of probability measures on $\mathcal{X}$.

Definition 4.5.2. *A sequence of probability measures (μ_n) on $(\mathcal{X}, \mathcal{B})$ satisfies an LDP with rate function I if, for all $A \subseteq \mathcal{B}$,*

$$-\inf_{x \in A^\circ} I(x) \leq \liminf_{n\to\infty} \frac{1}{n} \log \mu_n(A) \leq \limsup_{n\to\infty} \frac{1}{n} \log \mu_n(A) \leq -\inf_{x \in \bar{A}} I(x).$$

Remark: If $\mathcal{X} = \mathbb{R}$, and (μ_n) satisfies an LDP, we may also say that (F_n) satisfies an LDP, where F_n is the distribution function corresponding to μ_n. Similarly, if (X_n)

is a sequence of random elements of $\mathcal{X}$ such that X_n is distributed according to μ_n, we may also say (X_n) satisfies an LDP in $\mathcal{X}$.

If μ and ν are probability measures, we write $\nu \ll \mu$ if ν is absolutely continuous with respect to μ. In this case, we also write $d\nu/d\mu$ for the Radon–Nikodým derivative of ν with respect to μ. If $\phi : \mathcal{X} \to \mathbb{R}$ is a bounded, continuous function, $x \in \mathbb{R}$ and $\delta > 0$, define an open set in $M(\mathcal{X})$ by

$$U_{\phi,x,\delta} = \Big\{\nu \in M(\mathcal{X}) : \Big|\int_{\mathcal{X}} \phi(y)\, d\nu(y) - x\Big| < \delta\Big\}.$$

The collection $\{U_{\phi,x,\delta}\}$ generates the weak topology, and the Borel σ-algebra in $M(\mathcal{X})$, equipped with the weak topology, is the σ-algebra generated by the open sets in the weak topology.

Theorem 4.5.3 (Sanov's Theorem). *Let $\mathcal{X}$ be a complete, separable metric space, and let μ be a probability measure on $\mathcal{X}$. If $X_1, \ldots, X_n$ are independent and identically distributed according to μ, and $\hat{\mu}_n$ denotes their empirical measure, then the sequence $(\hat{\mu}_n)$ satisfies an LDP in $M(\mathcal{X})$, equipped with the weak topology, with good rate function*

$$I(\nu) = \begin{cases} \int_{-\infty}^{\infty} \frac{d\nu}{d\mu}(x) \log\big(\frac{d\nu}{d\mu}(x)\big)\, d\mu(x) & \text{if } \nu \ll \mu \\ \infty & \text{otherwise.} \end{cases}$$

Proposition 4.5.4 (Contraction Principle). *Let $\mathcal{X}$ and $\mathcal{Y}$ be Hausdorff spaces, and let $f : \mathcal{X} \to \mathcal{Y}$ be a continuous function. If (X_n) satisfies an LDP in $\mathcal{X}$ with good rate function I, then $\big(f(X_n)\big)$ satisfies an LDP in $\mathcal{Y}$ with good rate function*

$$J(y) = \inf\{I(x) : f(x) = y\}.$$

Proposition 4.5.5 (Inverse Contraction Principle). *Let $\mathcal{X}$ and $\mathcal{Y}$ be Hausdorff spaces, and let $f : \mathcal{X} \to \mathcal{Y}$ be a continuous bijection. Suppose (X_n) is a sequence of random elements of $\mathcal{X}$ such that the following two conditions hold:*

(a) $\big(f(X_n)\big)$ *satisfies an LDP in* $\mathcal{Y}$ *with rate function* $J(y)$*;*

(b) for all $\alpha \in [0, \infty)$ *there exists a compact set* K_α *in* $\mathcal{X}$ *such that*

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}(X_n \notin K_\alpha) \leq -\alpha.$$

Then (X_n) *satisfies an LDP in* $\mathcal{X}$ *with good rate function* $I(x) = J\big(f(x)\big)$.

Remark: Condition (b) is usually known as exponential tightness. Since $(\mathcal{F}_2, d_2)$ is complete and separable, it follows from Lemma 2.6 of Lynch and Sethuraman (1987), that exponential tightness is a necessary condition for the sequence of empirical distribution functions $(\hat{F}_n)$ to satisfy an LDP with a good rate function.

In trying to strengthen the topology in which we hope an LDP will hold, we therefore first need to characterise the compact sets in $(\mathcal{F}_2, d_2)$. This task is complicated by the following lemma, whose proof is given in Section 4.6.

Lemma 4.5.6. *For each* $r \geq 1$*, each* $F \in \mathcal{F}_r$ *and every* $\epsilon > 0$*, the closed ball* $\bar{B}(F, \epsilon) = \{G \in \mathcal{F}_r : d_r(F, G) \leq \epsilon\}$ *is not compact.*

Nevertheless, the next two lemmas, whose proofs are also deferred to Section 4.6, provide enough compact sets to study exponential tightness in $(\mathcal{F}_2, d_2)$. In fact, we work with compact sets in $(\mathcal{G}_2, \|\cdot\|_2)$ for convenience. We let $\mathcal{H}$ denote the set of pairs (H_1, H_2) of functions $H_1 : (0,1) \to [0,\infty)$ and $H_2 : (0,1) \to [0,\infty)$ such that $H_1(\epsilon) \to 0$ as $\epsilon \to 0$, and $H_2(1-\epsilon) \to 0$ as $\epsilon \to 0$.

Lemma 4.5.7. *For* $(H_1, H_2) \in \mathcal{H}$*, let*

$$K_{H_1,H_2} = \Big\{ G \in \mathcal{G}_2 : \int_0^\epsilon G^2(p)\, dp \leq H_1^2(\epsilon) \text{ and } \int_{1-\epsilon}^1 G^2(p)\, dp \leq H_2^2(1-\epsilon)\ \forall \epsilon \in (0,1) \Big\}.$$

Then K_{H_1,H_2} *is compact in* $(\mathcal{G}_2, \|\cdot\|_2)$.

Lemma 4.5.8. *If K is a compact subset of $(\mathcal{G}_2, \|\cdot\|_2)$, then there exists a pair $(H_1, H_2) \in \mathcal{H}$ such that $K \subseteq K_{H_1,H_2}$.*

Theorem 4.5.9. *If F has bounded support, then the sequence $(\hat{F}_n)$ of empirical distribution functions is exponentially tight in $(\mathcal{F}_2, d_2)$.*

Proof. It suffices to find a pair $(H_1, H_2) \in \mathcal{H}$ such that $\mathbb{P}(\hat{F}_n^{-1} \notin K_{H_1,H_2}) = 0$ for all $n \in \mathbb{N}$. If F has bounded support, then $\xi_0 = \lim_{p \searrow 0} \xi_p$ and $\xi_1 = \lim_{p \nearrow 1} \xi_p$ exist in $\mathbb{R}$. For $\epsilon \in (0,1)$, let

$$H_1(\epsilon) = \epsilon^{1/2} \max(|\xi_0|, |\xi_1|) \quad \text{and} \quad H_2(\epsilon) = H_1(1-\epsilon).$$

Then $(H_1, H_2) \in \mathcal{H}$, and

$$\int_0^\epsilon \big(\hat{F}_n^{-1}(p)\big)^2 \, dp \le \epsilon \max(\xi_0^2, \xi_1^2) = H_1^2(\epsilon),$$

and similarly $\int_{1-\epsilon}^1 \big(\hat{F}_n^{-1}(p)\big)^2 \, dp \le H_2^2(1-\epsilon)$. Hence $\mathbb{P}(\hat{F}_n^{-1} \notin K_{H_1,H_2}) = 0$ for all $n \in \mathbb{N}$. □

Corollary 4.5.10. *If F has bounded support, then $d_2\big(H_n(\hat{F}_n), H_n(F)\big)$ satisfies the large deviations upper bound for semi-infinite intervals with a good rate function. In other words, there exists a good rate function I such that, for every $\epsilon > 0$,*

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}\big\{ d_2\big(H_n(\hat{F}_n), H_n(F)\big) \ge \epsilon \big\} \le -\inf_{x \ge \epsilon} I(x).$$

Proof. By Sanov's theorem, Theorem 4.5.9 and the Inverse Contraction Principle, the sequence $(\hat{F}_n)$ satisfies an LDP in $(\mathcal{F}_2, d_2)$ with good rate function

$$I_1(F') = \begin{cases} \int_{-\infty}^{\infty} \frac{d\mu_{F'}}{d\mu_F}(x) \log\Big(\frac{d\mu_{F'}}{d\mu_F}(x)\Big) \, dF(x) & \text{if } \mu_{F'} \ll \mu_F \\ \infty & \text{otherwise,} \end{cases}$$

where μ_F and $\mu_{F'}$ are the probability measures corresponding to the distribution functions F and F' respectively. Since the function $\psi : (\mathcal{F}_2, d_2) \to \mathbb{R}$ defined by

$\psi(F') = d_2(F', F)$ is continuous, the Contraction Principle implies that $d_2(\hat{F}_n, F)$ satisfies an LDP in $\mathbb{R}$ with good rate function

$$I(x) = \inf\{I_1(F') : F' \in \mathcal{F}_2,\ d_2(F', F) = x\}.$$

Since $d_2\big(H_n(\hat{F}_n), H_n(F)\big)$ is stochastically dominated by $d_2(\hat{F}_n, F)$, the result for $d_2\big(H_n(\hat{F}_n), H_n(F)\big)$ follows. □

We say that $F \in \mathcal{F}_2$ has a polynomial tail if there exists an $m > 2$ such that $x^m\big(1 - F(x)\big) \to \infty$ as $x \to \infty$, or $|x|^m F(x) \to \infty$ as $x \to -\infty$.

Theorem 4.5.11. *If $F \in \mathcal{F}_2$ has a polynomial tail, then the sequence $(\hat{F}_n)$ of empirical distribution functions is not exponentially tight in $(\mathcal{F}_2, d_2)$.*

Proof. It suffices to show that for any $(H_1, H_2) \in \mathcal{H}$, we have

$$\mathbb{P}(\hat{F}_n^{-1} \notin K_{H_1,H_2}) \geq e^{-n}$$

for sufficiently large $n \in \mathbb{N}$. Now, we may assume without loss of generality that $x^m\big(1 - F(x)\big) \to \infty$ as $x \to \infty$, for some $m > 2$. Let $X_{(n)} = \max_{1 \leq i \leq n} X_i$. Then

$$\mathbb{P}(\hat{F}_n^{-1} \notin K_{H_1,H_2}) \geq \mathbb{P}\big(X_{(n)} > n^{1/2} H_2(1 - 1/n)\big) = 1 - F\big(n^{1/2} H_2(1 - 1/n)\big)^n.$$

Choose $n_0 \in \mathbb{N}$ large enough such that $H_2(1 - 1/n) \leq 1$ and $1 - F(n^{1/2}) \geq n^{-m/2}$ for all $n \geq n_0$. Then, for $n \geq n_0$,

$$\mathbb{P}(\hat{F}_n^{-1} \notin K_{H_1,H_2}) \geq 1 - F(n^{1/2})^n \geq 1 - \left(1 - \frac{1}{n^{m/2}}\right)^n \geq 1 - \exp(-n^{-(m/2-1)}).$$

But, for sufficiently large n,

$$1 - \exp(-n^{-(m/2-1)}) \geq \frac{1}{2n^{m/2-1}} \geq e^{-n}.$$

□

Remark: In view of the remark following the statement of the Inverse Contraction principle (Proposition 4.5.5), Theorem 4.5.11 shows that the sequence $(\hat{F}_n)$ does not satisfy an LDP in $(\mathcal{F}_2, d_2)$.

4.6 Appendix

Proof of **Lemma 4.5.6**.

We prove that the closed ball is not sequentially compact. Fix $\epsilon > 0$, $F \in \mathcal{F}_r$, and, for $p \in (0,1)$, let $\xi_p = \inf\{x \in \mathbb{R} : F(x) \geq p\}$. Consider the sequence of distribution functions (F_n) given by

$$F_n(x) = \begin{cases} F(x) & \text{if } x < \xi_{1-1/n} \\ 1 - 1/n & \text{if } \xi_{1-1/n} \leq x < \xi_{1-1/n} + \epsilon n^{1/r} \\ F(x - \epsilon n^{1/r}) & \text{if } x \geq \xi_{1-1/n} + \epsilon n^{1/r}. \end{cases}$$

Thus

$$F_n^{-1}(p) = \begin{cases} \xi_p & \text{if } p \leq 1 - 1/n \\ \xi_p + \epsilon n^{1/r} & \text{if } p > 1 - 1/n. \end{cases}$$

It follows that

$$d_r(F_n, F) = \left(\int_0^1 |F_n^{-1}(p) - \xi_p|^r \, dp \right)^{1/r} = \left(\int_{1-1/n}^1 \epsilon^r n \, dp \right)^{1/r} = \epsilon.$$

On the other hand, we have $|F_n(x) - F(x)| \leq 1/n$ for all $x \in \mathbb{R}$ and $n \in \mathbb{N}$, so if a subsequence (F_{n_k}) satisfied $d_r(F_{n_k}, G) \to 0$ as $k \to \infty$, then we would have to have $G \equiv F$. Since $d_r(F_{n_k}, F) = \epsilon$ for all $k \in \mathbb{N}$, no convergent subsequence can exist. $\square$

Proof of **Lemma 4.5.7**.

Take a sequence $(G_n) \in K_{H_1,H_2}$. For each $m \in \mathbb{N}$, choose $\epsilon_m \in (0, 1/2)$ such that $H_1(\epsilon) \leq 1/m$ and $H_2(1-\epsilon) \leq 1/m$ for each $\epsilon \in (0, \epsilon_m]$. We claim that there exists

an infinite subset N_1 of $\mathbb{N}$ such that

$$\int_{\epsilon_1}^{1-\epsilon_1} \big(G_{n_1}(p) - G_{n_2}(p)\big)^2 \, dp \leq 1$$

for all $n_1, n_2 \in N_1$. To see why this is the case, observe first that $G(\epsilon_1) \geq -H_1(\epsilon_1)/\epsilon_1^{1/2}$ and $G(1-\epsilon_1) \leq H_2(1-\epsilon_1)/\epsilon_1^{1/2}$ for each $G \in K_{H_1,H_2}$. Now we may partition the interval $[\epsilon_1, 1-\epsilon_1]$ into d equally spaced divisions and note that given any $\delta > 0$, there exist $-H_1(\epsilon_1)/\epsilon_1^{1/2} \leq x_0 \leq x_1 \leq \ldots \leq x_d \leq H_2(1-\epsilon_1)/\epsilon_1^{1/2}$ and an infinite subset N_1 of $\mathbb{N}$ such that

$$\Big| G_n\Big(\epsilon_1 + \frac{i}{d}(1-2\epsilon_1)\Big) - x_i \Big| \leq \delta$$

for all $i = 0, 1, \ldots, d$ and $n \in N_1$. For $i = 0, 1, \ldots, d$, let $p_i = \epsilon_1 + i(1-2\epsilon_1)/d$. Then, for $n_1, n_2 \in N_1$,

$$\begin{aligned}
&\int_{\epsilon_1}^{1-\epsilon_1} \big(G_{n_1}(p) - G_{n_2}(p)\big)^2 \, dp \\
&\qquad = \sum_{i=1}^{d} \int_{p_{i-1}}^{p_i} \big(G_{n_1}(p) - G_{n_2}(p)\big)^2 \, dp \\
&\qquad \leq \frac{1}{d} \sum_{i=1}^{d} \big(x_i - x_{i-1} + 2\delta\big)^2 \\
&\qquad \leq \frac{1}{d} \Big\{ \Big(\frac{H_2(1-\epsilon_1)}{\epsilon_1^{1/2}} + \frac{H_1(\epsilon_1)}{\epsilon_1^{1/2}} \Big)^2 + 4\delta \Big(\frac{H_2(1-\epsilon_1)}{\epsilon_1^{1/2}} + \frac{H_1(\epsilon_1)}{\epsilon_1^{1/2}} \Big) + 4\delta^2 \Big\} \\
&\qquad \leq 1,
\end{aligned}$$

for sufficiently small $\delta > 0$, and sufficiently large $d \in \mathbb{N}$.

In a similar manner, we may find a sequence of infinite subsets (N_m) of $\mathbb{N}$ with $N_1 \supseteq N_2 \supseteq \ldots$, such that for each $m \in \mathbb{N}$,

$$\int_{\epsilon_m}^{1-\epsilon_m} \big(G_{n_1}(p) - G_{n_2}(p)\big)^2 \, dp \leq \frac{1}{m}$$

for all $n_1, n_2 \in N_m$. Now construct the diagonal subsequence (n_k), by taking n_k to be the kth smallest element of N_k, for each $k \in \mathbb{N}$. The sequence (G_{n_k}) is a subsequence

of the original sequence (G_n), and is a Cauchy sequence in $(\mathcal{G}_2, \|\cdot\|_2)$ because, for $k \leq l$,

$$\begin{aligned}
&\int_0^1 \big(G_{n_k}(p) - G_{n_l}(p)\big)^2 \, dp \\
&\qquad \leq 2\int_0^{\epsilon_k} \big(G_{n_k}^2(p) + G_{n_l}^2(p)\big)\, dp + \int_{\epsilon_k}^{1-\epsilon_k} \big(G_{n_k}(p) - G_{n_l}(p)\big)^2 \, dp \\
&\qquad\qquad + 2\int_{1-\epsilon_k}^1 \big(G_{n_k}^2(p) + G_{n_l}^2(p)\big)\, dp \\
&\qquad \leq 4\big(H_1^2(\epsilon_k) + H_2^2(1-\epsilon_k)\big) + \frac{1}{k} \\
&\qquad \to 0
\end{aligned}$$

as $k \to \infty$. But $(\mathcal{G}_2, \|\cdot\|_2)$ is complete, so (G_{n_k}) converges in $(\mathcal{G}_2, \|\cdot\|_2)$, so K_{H_1,H_2} is sequentially compact. □

Proof of **Lemma 4.5.8**.

Suppose the lemma is false. Then without loss of generality, we may assume there exist $\epsilon > 0$ and a sequence $(G_n) \in K$ such that

$$\int_0^{1/n} G_n^2(p)\, dp \geq \epsilon.$$

Since K is compact, there exist a strictly increasing sequence $(n_k) \in \mathbb{N}$, $G \in K$ and $k_0 \in \mathbb{N}$ such that

$$\int_0^1 \big(G_{n_k}(p) - G(p)\big)^2 \, dp \leq \frac{\epsilon}{4}$$

for all $k \geq k_0$. By restricting attention to a further subsequence if necessary, we may assume

$$\int_{1/n_{k+1}}^{1/n_k} G_{n_k}^2(p)\, dp \geq \frac{\epsilon}{2}$$

for each $k \in \mathbb{N}$. But then, by Minkowski's inequality,

$$\begin{aligned}
\left(\int_0^1 G^2(p)\,dp\right) &\geq \sum_{k=k_0}^{\infty} \int_{1/n_{k+1}}^{1/n_k} G^2(p)\,dp \\
&\geq \sum_{k=k_0}^{\infty} \left(\int_{1/n_{k+1}}^{1/n_k} G_{n_k}^2(p)\,dp - \int_{1/n_{k+1}}^{1/n_k} \big(G_{n_k}(p) - G(p)\big)^2\,dp\right)^2 \\
&\geq \sum_{k=k_0}^{\infty} \left(\frac{\epsilon}{2} - \frac{\epsilon}{4}\right)^2 \\
&= \infty,
\end{aligned}$$

which contradicts the fact that $G \in (\mathcal{G}_2, \|\cdot\|_2)$. □

Bibliography

Bay, S. D. (1999), *Nearest neighbor classification from multiple feature subsets*, Intelligent Data Analysis, **3**, 191–209.

Beran, R. J. (1982), *Estimated sampling distributions: the bootstrap and competitors*, Ann. Statist., **10**, 212–225.

Beran, R. J. (1984), *Bootstrap methods in statistics*, Jahresber. Deutsch. Math.-Verein., **86**, 212–225.

Beran, R. J. (1995), *Stein confidence sets and the bootstrap*, Statistica Sinica, **5**, 109–127.

Beran, R. J. (1997), *Diagnosing bootstrap success*, Ann. Inst. Statist. Math., **49**, 1–24.

Berger, J. (1980), *A robust generalized Bayes estimator and confidence region for a multivariate normal mean*, Ann. Statist., **8**, 716–761.

Bhattacharya, R. N. and Rao, R. R. (1976), *Normal Approximation and Asymptotic Expansions*, Wiley, New York.

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

Bickel, P. J., Götze, F. and van Zwet, W. R. (1997), *Resampling fewer than n observations: gains, losses and remedies for losses*, Statistica Sinica, **7**, 1–31.

Billingsley, P. (1995), *Probability and Measure*, Third ed., Wiley, New York.

Billingsley, P. (1999), *Convergence of Probability Measures*, Second ed., Wiley, New York.

Birnbaum, A. (1955), *Characterizations of complete classes of tests of some multiparametric hypotheses, with applications to likelihood ratio tests*, Ann. Math. Statist., **26**, 21–36.

Box, G. E. P. (1953), *Spherical distributions. (Abstract)*, Ann. Math. Statist., **24**, 687–688.

Brandwein, A. R. (1979), *Minimax estimation of the mean of spherically symmetric distributions under general quadratic loss*, J. Mult. Anal., **9**, 579–588.

Brandwein, A. R. and Strawderman, W. E. (1978), *Minimax estimation of location parameters for spherically symmetric unimodal distributions under quadratic loss*, Ann. Statist., **6**, 377–416.

Brandwein, A. R. and Strawderman, W. E. (1990), *Stein estimation: the spherically symmetric case*, Statist. Sci., **5**, 356–369.

Breiman, L. (1996), *Bagging predictors*, Machine Learning, **24**, 123–140.

Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. (1984), *Classification and Regression Trees*, Chapman and Hall, New York.

Bretagnolle, J. (1983), *Lois limites du bootstrap de certaines fonctionnelles*, Ann. Inst. Henri Poincaré, **19**, 281–296.

Brown, L. D., Low, M. G. and Zhao, L. H. (1997), *Superefficiency in nonparametric function estimation*, Ann. Statist., **25**, 2607–2625.

Buehler, R. J. (1959), *Some validity criteria for statistical inferences*, Ann. Math. Statist., **8**, 716–761.

Bühlmann, P. and Yu, B. (2002), *Analyzing bagging*, Ann. Statist., **30**, 927–961.

Buja, A. and Stuetzle, W. (2000a), *The effect of bagging on variance, bias and mean squared error*, Preprint.

Buja, A. and Stuetzle, W. (2000b), *Smoothing effects of bagging*, Preprint.

Burkill, J. C. (1962), *A First Course in Mathematical Analysis*, Cambridge University Press.

Canty, A. J., Davison, A. C., Hinkley, D. V. and Ventura, V. (2000), *Bootstrap Diagnostics*, Preprint.

Casella, G. and Hwang, J. T. (1983), *Empirical Bayes confidence sets for the mean of a multivariate normal distribution*, J. Amer. Statist. Assoc., **78**, 688–698.

Casella, G. and Hwang, J. T. (1986), *Confidence sets and the Stein effect*, Comm. Statist. Theory Methods, **15**, 2043–2063.

Chibisov, D. M. (1965), *An investigation of the asymptotic power of the tests of fit*, Theor. Prob. Appl., **10**, 421–437.

Cover, T. M. (1968), *Rates of convergence for nearest neighbor procedures*, Proceedings of the Hawaii International Conference on System Sciences, Eds. B. K. Kinariwala and F. F. Kuo, pp. 413–415, University of Hawaii Press, Honolulu.

Cover, T. M. and Hart, P. E. (1967), *Nearest neighbor pattern classification*, IEEE Trans. Inform. Theory, **13**, 21–27.

Dembo, A. and Zeitouni, O. (1995), *Large Deviations Techniques and Applications*, Second ed., Jones and Bartlett, 1995.

Dobrushin, R. L. (1970), *Presecribing a system of random variables by conditional distributions*, Theor. Prob. Appl., **15**, 458–486.

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

Evans, S. and Stark, P. (1996), *Shrinkage estimators, Skorokhod's problem and stochastic integration by parts*, Ann. Statist., **24**, 809–815.

Faith, R. E. (1976), *Minimax Bayes set and point estimators of a multivariate normal mean*, Technical Report 66, Univ. Michigan.

Fang, K. T., Kotz, S. and Ng, K. W. (1989), *Symmetric multivariate and related distributions*, Chapman and Hall, London.

Fisher, R. A. (1936), *The use of multiple measurements in taxonomic problems*, Ann. Eugenics, **7**, 179–188.

Fisher, R. A. (1956), *Statistical Methods and Scientific Inference*, Oliver and Boyd, Edinburgh.

Fisher, R. A. (1959), *Mathematical probability in the natural sciences*, Technometrics, **1**, 21–29.

Fix, E. and Hodges, J. (1951), *Discriminatory analysis, nonparametric discrimination: consistency properties.* Technical Report No. 4, Project No. 21–49–004, USAF School of Aviation Medicine, Randolph Field, Texas.

Friedman, J. H. (2000), *On bagging and nonlinear estimation*, Preprint.

Fukunaga, K. and Hummels, D. M. (1987), *Bias of nearest neighbor error estimates*, IEEE Trans. Patt. Anal. Mach. Intel., **9**, 103–112.

Hájek, J. (1970), *A characterization of limiting distributions of regular estimates*, Zeit. Wahrscheinlichkeitsth., **14**, 323–330.

Hájek, J. (1972), *Local asymptotic minimax and admissibility in estimation*, Proc. Sixth Berkeley Symp. on Math. Statist. Prob., **1**, 175–194, Univ. California Press, Berkeley.

Hall, P. G. (1992), *The Bootstrap and Edgeworth Expansion*, Springer-Verlag, New York.

Hand, D. J. (1981), *Discrimination and Classification*, Wiley, New York.

Hastie, T. J., Tibshirani, R. J. and Friedman, J. H. (2001), *The Elements of Statistical Learning: Data mining, Inference and Prediction*, Springer-Verlag, New York.

Hoeffding, W. (1963), *Probability inequalities for sums of bounded random variables*, J. Amer. Statist. Assoc., **58**, 13–30.

Hwang, J. T. and Casella, G. (1982), *Minimax confidence sets for the mean of a multivariate normal distribution*, Ann. Statist., **10**, 868–881.

Hwang, J. T. and Casella, G. (1984), *Improved set estimators for a multivariate normal mean*, Statist. Decisions (Suppl.), **1**, 3–16.

Hwang, J. T. and Chen, J. (1986), *Improved confidence sets for the coefficients of a linear model with spherically symmetric errors*, Ann. Statist., **14**, 444–460.

James, W. and Stein, C. M. (1961), *Estimation with quadratic loss*, Proc. Fourth Berkeley Symposium, **1**, 361–379, Univ. California Press, Berkeley.

Jing, B-Y, Feuerverger, A. and Robinson, J. (1994), *On the bootstrap saddlepoint approximations*, Biometrika, **81**, 211-215.

Joshi, V. M. (1969), *Admissibility of the usual confidence sets for the mean of a univariate or bivariate normal distribution*, Ann. Math. Statist., **38**, 1868–1875.

Kallenberg, W. C. M. (1978), *Asymptotic optimality of likelihood ratio tests in exponential families*, Mathematisch Centrum, Amsterdam.

Kelker, D. (1970), *Distribution theory of spherical distributions and a location-scale parameter generalisation*, Sankhya Ser. A, **32**, 419–430.

Ki, F. and Tsui, K-W (1985), *Improved confidence set estimators of a multivariate normal mean and generalizations*, Ann. Inst. Statist. Math., **37**, 487–498.

Kingman, J. F. C. (1993), *Poisson Processes*, Oxford University Press.

Kuncheva, L. I. and Bezdek, J. C. (1998), *Nearest prototype classification: clustering, genetic algorithms, or random search?*, IEEE Trans. Syst. Man Cybernet. Ser. C, **28**, 160–164.

Lahiri, S. N. (1992), *On bootstrapping M-estimators*, Sankhya Ser. A, **54**, 157–170.

Lahiri, S. N. (1994), *On two-term Edgeworth expansions and bootstrap approximations for studentized multivariate M-estimators*, Sankhya Ser. A, **56**, 201–226.

Le Cam, L. (1953), *On some asymptotic properties of maximum likelihood estimates and related Bayes estimates*, Univ. Calif. Pub. Statist., **1**, 277–330.

Le Cam, L. (1960), *Locally asymptotically normal families of distributions*, Univ. Calif. Pub. Statist., **45**, 169–180.

Lehmann, E. L. (1986), *Testing Statistical Hypotheses*, Second ed., Springer-Verlag, New-York.

Lehmann, E. L. (1998), *Theory of Point Estimation*, Second ed., Springer-Verlag, New-York.

Lu, K. L. and Berger, J. O. (1989), *Estimated confidence procedures for multivariate normal means*, J. Statist. Plann. Inference, **23**, 1–19.

Lynch, J. and Sethuraman, J. (1987), *Large deviations for processes with independent increments*, Ann. Probab., **15**, 610–627.

Major, P. (1978), *On the invariance principle for sums of independent, identically distributed random variables*, J. Mult. Anal., **8**, 487–501.

Mallows, C. L. (1972), *A note on asymptotic joint normality*, Ann. Math. Statist., **43**, 508–515.

Marron, J. S. (1983), *Optimal rates of convergence to Bayes risk in nonparametric discrimination*, Ann. Statist., **11**, 1142–1155.

Mollineda, R. A., Ferri, F. J. and Vidal, E. (2000), *Merge-based prototype selection for nearest-neighbor classification.* Proc. Fourth World Multiconf. on Systemics, Cybernetics and Informatics, **7**, Eds. Z. Huang, C. Sun, and I. McDonald, 640–645, International Institute of Informatics and Systemics, Orlando, Florida.

Politis, D. N., Romano, J. P. and Wolf, M. (1999), *Subsampling*, Springer-Verlag, New York.

Pollard, D. B. (2002), *A User's Guide to Measure Theoretic Probability*, Cambridge University Press.

Psaltis, D., Snapp, R. R. and Venkatesh, S. S. (1994), *On the finite sample performance of the nearest neighbor classifier*, IEEE Trans. Inform. Theory, **40**, 820–837.

Putter, H. and van Zwet, W. R. (1996), *Resampling: Consistency of substitution estimators*, Ann. Statist., **24**, 2297–2318.

Robert, C. and Casella, G. (1990), *Improved confidence sets for spherically symmetric distributions*, J. Mult. Anal., **32**, 84–94.

Robert, C. and Casella, G. (1994), *Improved confidence statements for the usual multivariate normal confidence set*, Statist. Decision Theory and Related Topics V, 351–368, Academic, New York.

Robinson, G. K. (1979a), *Conditional properties of statistical procedures*, Ann. Statist., **7**, 742–755.

Robinson, G. K. (1979b), *Conditional properties of statistical procedures for location and scale parameters*, Ann. Statist., **7**, 756–771.

Robinson, J. and Skovgaard, I. M. (1998), *Bounds for probabilities of small relative errors for empirical saddlepoint and bootstrap tail approximations*, Ann. Statist., **26**, 2369–2394.

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

Shao, J. and Tu, D. (1995), *The Jackknife and Bootstrap*, Springer-Verlag, New York.

Shinozaki, N. (1989), *Improved confidence sets for the mean of a multivariate normal distribution*, Ann. Inst. Statist. Math., **41**, 331–346.

Shorack, G. R. and Wellner, J. A. (1986), *Empirical Processes with Applications to Statistics*, Wiley, New York.

Singh, K. (1981), *On the asymptotic validity of Efron's bootstrap*, Ann. Statist., **9**, 1187–1195.

Stein, C. (1956), *Inadmissibility of the usual estimator for the mean of a multivariate normal distribution*, Proc. Third Berkeley Symposium, **1**, 197–206. Univ. of California Press.

Stein, C. (1962), *Confidence sets for the mean of a multivariate normal distribution (with discussion)*, J. Roy. Statist. Soc. Ser. B, **24**, 265–296.

Stein, C. (1981), *Estimation of the mean of a multivariate normal distribution*, Ann. Statist., **9**, 1135–1151.

Stoller, D. S. (1954), *Univariate two-population distribution-free discrimination*, J. Amer. Statist. Assoc., **49**, 770–777.

Thomas, D. H. (1970), *Some contributions to radial probability distributions*, Ph.D. dissertation, Wayne State University.

Tseng, Y-L. and Brown, L.D. (1997), *Good exact confidence sets for a multivariate normal mean*, Ann. Statist., **25**, 2228–2258.

van der Vaart, A. W. (1998), *Asymptotic Statistics*, Cambridge Univ. Press.

Wang, H. (2000), *Improved confidence estimators for the multivariate normal confidence set*, Statistica Sinica, **10**, 659–664.

Zellner, A. (1976), *Bayesian and non-Bayesian analysis of the regression model with multivariate student-t error terms*, J. Amer. Statist. Assoc., **71**, 400–405.