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,

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

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 $Y_n(\theta_0, X) \xrightarrow{d} N_k\bigl(0, I(\theta_0)\bigr)$, 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 $N\bigl(h, I^{-1}(\theta_0)\bigr)$ with respect to $N\bigl(0, I^{-1}(\theta_0)\bigr)$. Thus an LAN model $\{\mathbb{P}_{\theta_0+n^{-1/2}h} : h \in \mathbb{R}^k\}$ and the model $\bigl\{N\bigl(h, I^{-1}(\theta_0)\bigr) : h \in \mathbb{R}^k\bigr\}$ 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

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

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 $\epsilon > 0$,*

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

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

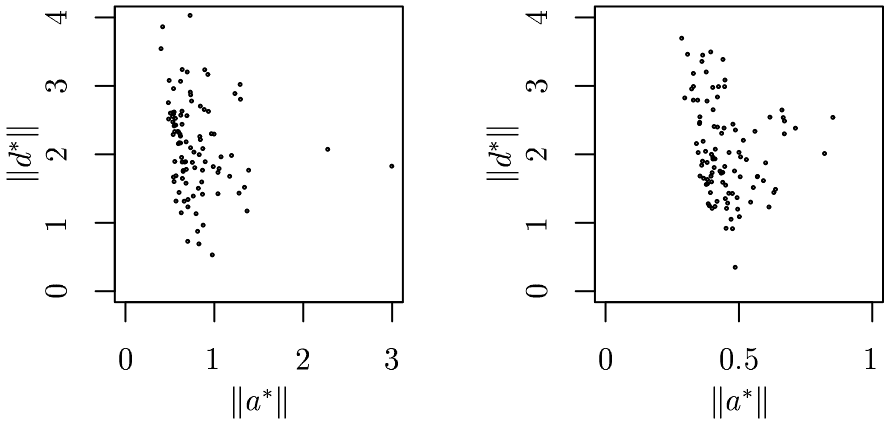

Figure 1.1: Beran's diagnostic applied to the Stein estimator. On the left-hand plot, $\theta = (0, 0, 0, 0, 0)$; on the right, $\theta = (-0.1, 0.1, 0, 0, 0)$. The choice of θ on the right ensures that the likelihood ratio test of size 0.05 of $H_0 : \theta_1 = \ldots = \theta_k$ versus $H_1 : H_0$ is not true, rejects H_0 with probability approximately 0.95 (see Section 1.3.2). Parameter values: $n = 1000$, $B = 100$, $k = 5$.

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 non-parametric 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 $T_1^*, \ldots, T_R^*$.

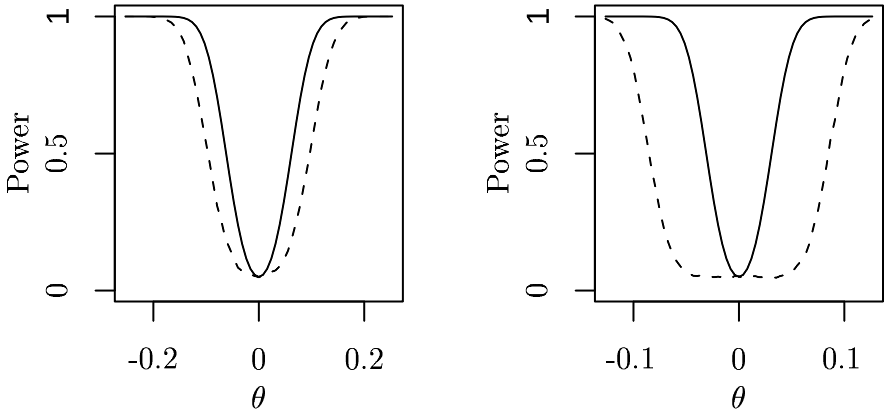

Figure 1.2: A comparison of the power functions of the UMPU test (solid) and the one derived from Beran's diagnostic (dashed). Parameter values: $\alpha = 0.05$, $b = 0.5$, $B = 100$, $R = 999$, $n = 1000$ (left), $n = 4000$ (right).

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

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

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

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

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 non-parametric 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 (c.f. 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

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

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

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

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 $O\bigl(\log N^+(\alpha, \beta, \theta)\bigr)$ 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

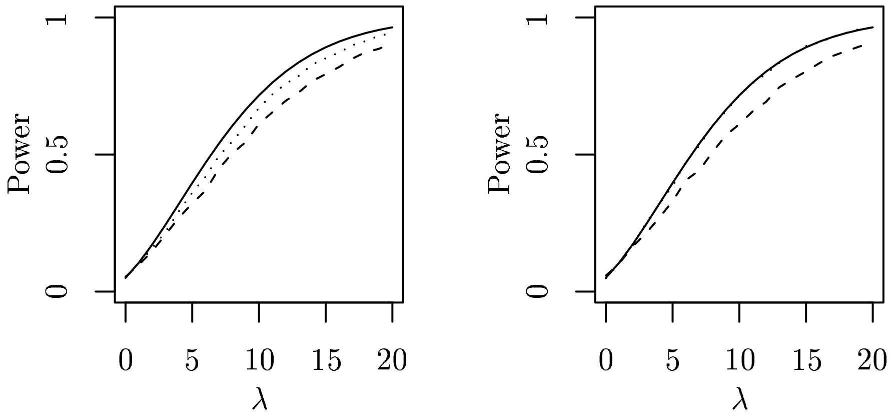

Figure 1.3: The power functions of the likelihood ratio test (solid), the ANOVA test (dotted), and Beran's method test (dashed) of $H_0 : \lambda = 0$ and $H_1 : \lambda > 0$. Parameter values: $n = 10$ (left), $n = 1000$ (right), $\alpha = 0.05$, $k = 5$, $B = 100$, $R = 999$; 10,000 Monte-Carlo repetitions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

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

Thus

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

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 Polya'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).

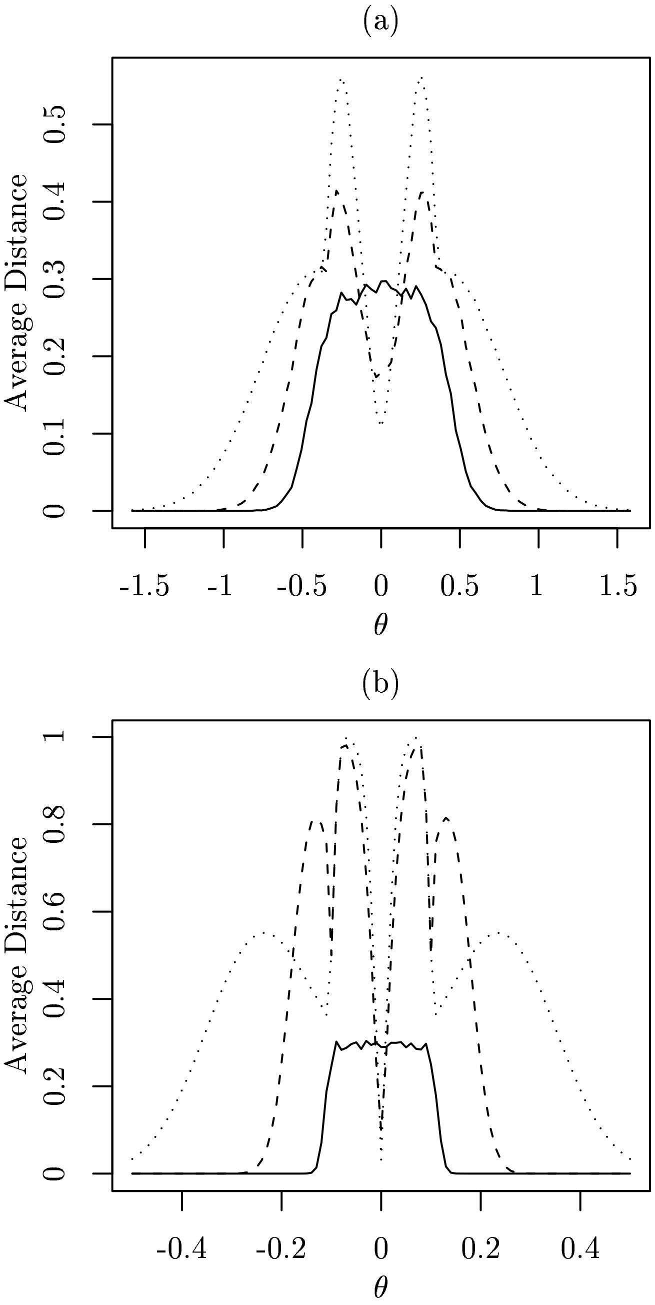

Figure 1.4: The distances $d\big(H_m(\bar{X}_n), H_n(\theta)\big)$, averaged over 1000 realisations of $\bar{X}_n$, with $m = n^{1/2}$ (dotted), $m = n^{3/4}$ (dashed), $m = n$ (solid). Parameter values: $b = 0.5$, (a) $n = 100$, (b) $n = 10,000$.

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 in 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) $\mathbb{P}_{\theta=0}(\hat{\theta}_n = 0) \rightarrow 1$ as $n \rightarrow \infty$

(ii) $\mathbb{P}_{\theta\neq0}(\hat{\theta}_n \neq 0) \rightarrow 1$ as $n \rightarrow \infty$.

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 Polya'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 $n^{1/2}\bigl(T_{n,S}(\bar{X}_{n,1}) - \theta\bigr), \ldots, n^{1/2}\bigl(T_{n,S}(\bar{X}_{n,R}) - \theta\bigr)$

(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 $m^{1/2}\bigl(T_{m,S}(\bar{X}^*_{m,1}) - \bar{X}_{n,1}\bigr), \ldots, m^{1/2}\bigl(T_{m,S}(\bar{X}^*_{m,R}) - \bar{X}_{n,R}\bigr)$

(iv) Compute

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

(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 $\lambda = 0$ as 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 $\mathbb{P}_*$,

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

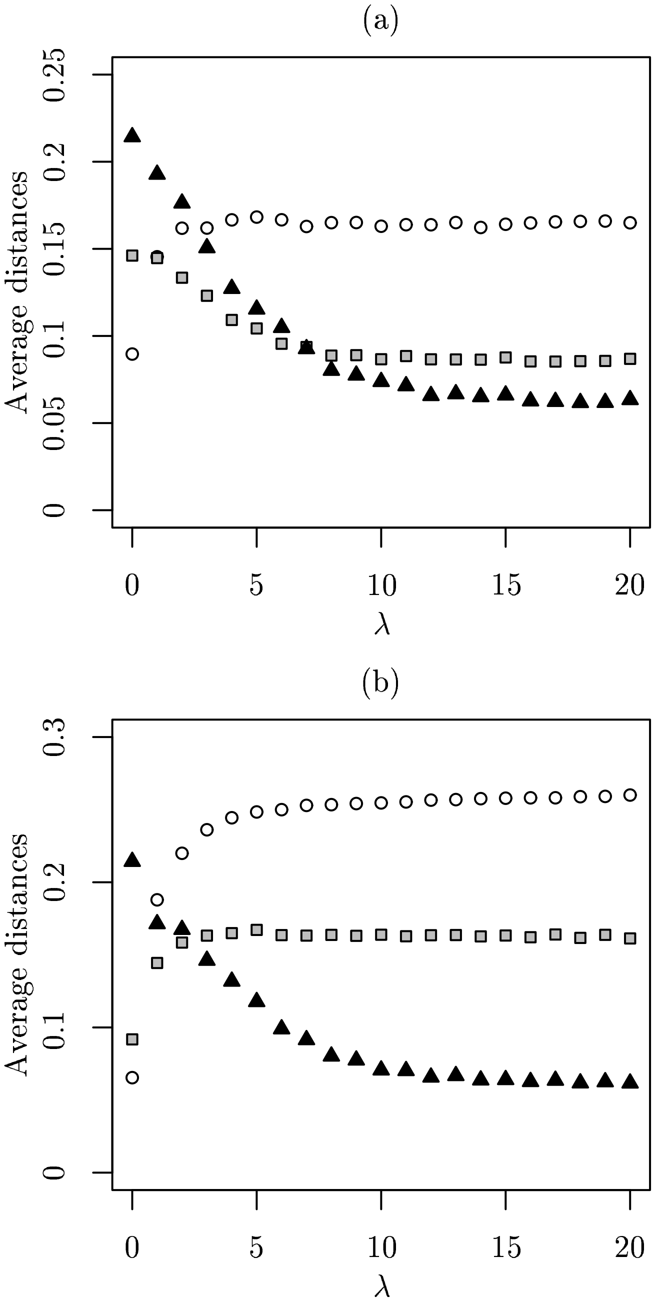

Figure 1.5: The average distances $d\big(\hat{H}_{m,R}(\bar{X}_n), \hat{H}_{n,R}(\theta)\big)$, with $m = n^{1/2}$ (circles), $m = n^{3/4}$ (grey squares), $m = n$ (black triangles). Parameter values: $R = 500$, $B = 200$, $k = 5$, $n^{1/2}\theta = (\lambda/2)^{1/2}(-1, 1, 0, 0, 0)$, (a) $n = 100$, (b)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

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 (c.f. 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 (c.f.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'.

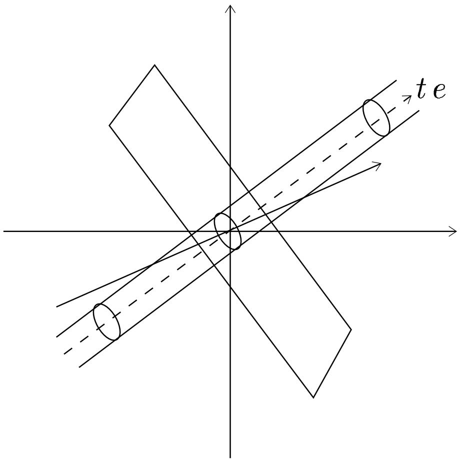

Figure 1.6: Diagram showing the set S and the hyper-cylinder B_d, which has axis along e.

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

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

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 $\Phi\bigl\{b^{-1}\bigl(x+(1-b)\theta_n n^{1/2}\bigr)\bigr\}$ 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

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

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

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

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

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

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

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

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

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

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