Likelihood Ratio Tests and Inequality Constraints

By
Charles J. Geyer

Technical Report No. 610
School of Statistics
University of Minnesota
December, 18, 1995

SUMMARY

In likelihood ratio tests involving inequality-constrained hypotheses, the Neyman-Pearson test based on the least favourable parameter value in a compound null hypothesis can be extremely conservative. The ordinary parametric bootstrap is generally inconsistent and usually too liberal. Two methods of correcting the inconsistency of the parametric bootstrap are proposed: shrinking the constraint set toward the maximum likelihood estimate and superefficient estimation of the active set of constraints. Optimal shrinkage adjustment can be determined using bootstrap calibration. These methods are compared with the double bootstrap, the subsampling bootstrap, Bayes factors, and Bayesian P-values. The Bayesian methods are also too liberal if diffuse priors are used.

Keywords: BAYES FACTOR; BAYESIAN P-VALUE; DOUBLE BOOTSTRAP; HYPOTHESIS TEST; SUBSAMPLING BOOTSTRAP

1. INTRODUCTION

Consider the general hypothesis testing problem: given a statistical model $\{ P_\theta : \theta \in \Theta \}$ and arbitrary subsets Θ_0 and Θ_1 of Θ, conduct a test with null and alternative hypotheses

$$\begin{aligned} &H_0 : \theta \in \Theta_0 \\ &H_1 : \theta \in \Theta_1 \setminus \Theta_0. \end{aligned}$$

We are particularly interested in likelihood ratio tests involving multiple inequality constraints. The reason the alternative hypothesis is not denoted Θ_1 is that we prefer to use this notation for the constraint set in the maximization, which is usually the union of the null and alternative. Our examples are simple, multivariate analogues of one-sided tests, but we develop theory that handles the most general case for which asymptotics exist.

The standard procedure for such problems (Perlman, 1969; Robertson and Wegman, 1978; Warrack and Robertson, 1984; Wolak 1987; and Robertson, *et al.*, 1988, pp. 254 ff.) follows the Neyman-Pearson theory. If a test has nested critical regions and hence can be defined by a test statistic t, then the corresponding P-value, the infimum of significance levels at which the null hypothesis can be rejected, is

$$\sup_{\theta \in \Theta_0} P_\theta\Bigl(t(X) \geq t(x)\Bigr). \tag{1}$$

When the supremum in (1) is achieved at a point θ_{LF}, the *least favourable* parameter value, (1) is the same as

$$P_{\theta_{\mathrm{LF}}}\Bigl(t(X) \geq t(x)\Bigr). \tag{2}$$

It has recently been recognized that (2) may be overly conservative. Berger and Boos (1994) propose the P-value

$$\sup_{\theta \in C_\epsilon} P_\theta\Bigl(t(X) \geq t(x)\Bigr) + \epsilon. \tag{3}$$

where C_ϵ is an exact $1 - \epsilon$ confidence region for θ. This produces an exact test, which may be less conservative than (2). However, this procedure seems difficult to implement in complex problems and will not be considered further.

Asymptotic tests are different. In effect, they use the P-value

$$P_{\theta_0}\Bigl(t(X) \geq t(x)\Bigr), \tag{4}$$

where θ_0 is the true parameter value. There is no supremum over the null hypothesis as in (1), nor is (4) in general a limit of such suprema. When θ_{LF}

is far from θ_0, (4) seems preferable to (2), a notion implicit in the theory of asymptotic tests.

In inequality-constrained inference (4) usually depends on the unknown true parameter value θ_0 and hence cannot be calculated. One can estimate (4) using the parametric bootstrap, giving the P-value

$$P_{\hat{\theta}_0(x)}\Big(t(X) \geq t(x)\Big). \tag{5}$$

Unfortunately, the lack of continuity inherent in inequality constraints usually makes the bootstrap inconsistent. The double bootstrap can provide evidence that the single bootstrap works despite inconsistency (Geyer, 1991; Shaw and Geyer, submitted), but the double bootstrap is itself inconsistent, so its evidence is suspect.

Another approach is to fix the inconsistency in the bootstrap, replacing (5) with a consistent estimate of (4). The subsampling bootstrap (Politis and Romano, 1994) is a general method applicable to all problems of inconsistency of the bootstrap. Though it does apply to inequality constrained problems, better methods more specific for constraints can be found. Two such methods are proposed: shrinking the constraint set toward the maximum likelihood estimate (MLE) and superefficient estimation of the active set of constraints.

What about Bayesian inference? Doesn't Bayesian inference using Markov chain Monte Carlo make these problems easy, with no need for bootstraps, double bootstraps, adjustments to bootstraps, and appeals to asymptotics? In general it does not, because in inequality-constrained problems Bayes factors and Bayesian P-values (Rubin, 1984; Meng, 1994) depend strongly on the prior. In contrast to the simple problems studied by Berger and Sellke (1987) and Casella and Berger (1987), there is in general no ordering of Bayesian and frequentist inferences. A Bayes factor may be many orders of magnitude above or below the P-value depending on the prior, and diffuse priors produce extremely liberal inference, implying strong evidence against the null hypothesis where a frequentist sees no evidence.

2. MULTIVARIATE ONE-SIDED TESTS

Consider observing a d-dimensional normal random vector x with unknown mean θ and known nondegenerate covariance matrix Σ. If K is a closed convex cone in $\mathbb{R}^d$ (meaning $s\theta + t\phi \in K$ whenever $\theta, \phi \in K$ and $s, t \geq 0$) we call the test with

$$\begin{aligned} &H_0 : \theta = 0 \\ &H_1 : \theta \in K,\ \theta \neq 0 \end{aligned} \tag{6}$$

a *type I* multivariate one-sided test and the test with

$$\begin{aligned} H_0 &: \theta \in K \\ H_1 &: \theta \notin K \end{aligned} \tag{7}$$

a *type II* multivariate one-sided test. The terminology 'multivariate analogue of the one-sided test' was used by Kudô (1963) to describe is called 'type I' here.

There is a curious duality between the two types of test. Define the Mahalanobis inner product $\langle x, y \rangle = x^T \Sigma^{-1} y$ and norm $\|x\|^2 = \langle x, x \rangle$. The *polar* of a convex cone K is (Rockafellar, 1970, p. 121)

$$K^\circ = \{\, x : \langle x, y \rangle \leq 0, \ y \in K \,\} .$$

Let $y = P_C(x)$ denote the projection of the point x on the closed convex set C (meaning y is the unique closest point to x in C). Then Moreau's theorem (Rockafellar, 1970, Theorem 31.5) and the Pythagorean theorem imply

$$\|P_K(x)\|^2 = \|x - P_{K^\circ}(x)\|^2.$$

The left hand side is the likelihood ratio test statistic for the type I test (6) and the right hand side is the likelihood ratio test statistic for the type II test (7) with K replaced by K°. The two tests have the same test statistic, and in Neyman-Pearson theory the tests are the same, because the origin is the least favourable parameter value. But in general the tests are not the same because of the compound null hypothesis in the Type II case.

This duality is trivial in the univariate case. The test of $H_0 : \theta = 0$ versus $H_1 : \theta > 0$ is a type I test. Its dual type II test has $H_0 : \theta \leq 0$ versus $H_1 : \theta > 0$. Few statisticians bother to distinguish them, which is perhaps a reason why type II tests have received little attention.

The duality is not so obvious even in the simplest multivariate case when Σ is a constant times the identity, K is the positive orthant, and K° the negative orthant. Despite the duality, the tests are very different in interpretation. When H_0 is rejected, the conclusion for both types can be stated 'at least one θ_i is strictly greater than zero', but the meaning is very different, because the type I test assumes all of the θ_i are nonnegative and the type II test does not.

Type II tests will be used as an example of general inequality constrained inference. Despite their simplicity, the exhibit the features that make general inequality-constrained inference difficult. General methods that work well for these problems should also work well in general.

3. BOOTSTRAP ADJUSTMENT

This section gives an informal description of the asymptotics of maximum likelihood and likelihood ratio tests and of bootstrap adjustment. A more formal treatment is given in Section 5.

3.1. Constrained Maximum Likelihood Estimates

This section deals with the asymptotics and bootstrap adjustment of an MLE $\hat{\theta}_n$ constrained to lie in a closed subset C of $\mathbb{R}^d$. Tests, which involve MLEs for the null and alternative, involve only simple additions to the this basic theory.

A constraint set C is *Chernoff regular* at a point $\theta \in C$ if the limit

$$T_C(\theta) = \lim_{\tau \downarrow 0} \frac{C - \theta}{\tau}, \tag{8}$$

exists in the sense of Painlevé-Kuratowski set convergence (Chernoff, 1954; Geyer, 1994), in which case $T_C(\theta)$ is called the *tangent cone* to C at θ. $T_C(\theta)$ is *nontrivial* if θ is not an isolated point of C, which implies that $T_C(\theta)$ is not the zero cone $\{0\}$.

Let V denote the expected Fisher information. Define a random function

$$q(\delta) = \delta' Z - \tfrac{1}{2} \delta' V \delta$$

where Z is an $N(0, V)$ random vector, and let $\hat{\delta}(Z)$ denote the maximizer of q over $T_C(\theta_0)$. Then under fairly week regularity conditions (Geyer, 1994) $\sqrt{n}(\hat{\theta}_n - \theta_0)$ converges in law to $\hat{\delta}(Z)$.

To understand what goes wrong with the bootstrap, we need to examine a bit of the proof. If $\hat{\theta}_n$ is constrained to lie in C, then $\sqrt{n}(\hat{\theta}_n - \theta_0)$ is constrained to lie in $\sqrt{n}(C - \theta_0)$. (8) implies

$$\sqrt{n}(C - \theta_0) \to T_C(\theta_0), \qquad \text{as } n \to \infty,$$

which is why the tangent cone appears in the asymptotics. In the parametric bootstrap we use $\hat{\theta}_n$ as if it were the true parameter value and estimate the tangent cone by $\sqrt{n}(C - \hat{\theta}_n)$. But this does not converge to the tangent cone because it differs from the correct formula $\sqrt{n}(C - \theta_0)$ by the term $\sqrt{n}(\hat{\theta}_n - \theta_0)$, which does not converge to zero.

3.2. Adjusting the Bootstrap

Politis and Romano (1994) propose as a general solution to problems of inconsistency of the nonparametric bootstrap that it should use a bootstrap sample size m less than the actual sample size n such that $m \to \infty$ but

$m/n \to 0$ as $n \to \infty$ and that the bootstrap should resample the data without rather than with replacement. This procedure also corrects the parametric bootstrap in inequality-constrained problems (where the issue of sampling with or without replacement does not arise), because $\sqrt{n}(C - \theta_0)$ is replaced by $\sqrt{m}(C - \hat{\theta}_n)$ which does converge in probability to the tangent cone because $\sqrt{m}(\hat{\theta}_n - \theta_0)$ does converge in probability to zero. This subsampling bootstrap, however, is inappropriate for inequality-constrained inference because it attacks the problem in the wrong place. It adjusts the likelihood, which is generally consistent, rather than the inequality constraints, which are the source of the problem. For that reason we propose two new adjusted bootstraps.

The first adjustment method, suggested to me by Professor R. T. Rockafellar, is to shrink the constraint set toward the maximum likelihood estimate. Here we obtain a consistent estimator of the tangent cone by introducing a *shrinkage factor* sequence satisfying $\lambda_n \to 0$ and $\sqrt{n}\lambda_n \to \infty$. We impose the constraint

$$\sqrt{n}(\theta_n^* - \hat{\theta}_n) \in \lambda_n \sqrt{n}(C - \hat{\theta}_n), \tag{9}$$

where θ_n^* is the bootstrap estimator. The right hand side is a consistent estimator of the tangent cone, because $\lambda_n \sqrt{n}(\hat{\theta}_n - \theta_0)$ converges in probability to zero.

Formula (9) constrains θ_n^* to lie in

$$C_{\lambda_n} = \lambda_n C + (1 - \lambda_n)\hat{\theta}_n. \tag{10}$$

The *shrinkage adjusted bootstrap* samples from the distribution indexed by $\hat{\theta}_n$ and finds θ_n^* by maximizing the bootstrap likelihood over (10). Then $\sqrt{n}(\theta_n^* - \hat{\theta}_n)$ has the same asymptotic distribution as $\sqrt{n}(\hat{\theta}_n - \theta_0)$.

The second adjustment method involves the specific form of the constraint set C. Suppose it is defined by a finite set of equality and inequality constraints

$$C = \{\, \theta \in \Theta : g_i(\theta) = 0,\ i \in E \text{ and } g_i(\theta) \le 0,\ i \in I \,\}. \tag{11}$$

Let

$$A = \{\, i \in I : g_i(\theta_0) = 0 \,\}$$

indicate the true *active set* of constraints, the inequality constraints satisfied with equality at the true parameter value. Another explanation of the inconsistency in the bootstrap is that it never gets the active set right, even asymptotically. But A can be estimated superefficiently using estimators much like the original examples suggested by Hodges (Le Cam, 1953, p. 280 or Lehmann, 1983, p. 405). Define

$$\tilde{A}_n = \left\{ i \in I : |g_i(\hat{\theta}_n)| \le B n^{-\alpha} \right\} \tag{12}$$

for any constant $B > 0$ and any constant $\alpha \in (0, 1/2)$. If the constraints are differentiable at θ_0, then $\tilde{A}_n$ consistently estimates the true active set A. Let $\tilde{\theta}_n$ maximize the likelihood over the constraint set

$$\tilde{C}_n = \left\{ \theta \in \Theta : g_i(\theta) = 0,\ i \in E \cup \tilde{A}_n \text{ and } g_i(\theta) \le 0,\ i \in I \setminus \tilde{A}_n \right\},$$

which imposes inequality constraints in the estimated active set with equality. The *adjusted active set bootstrap* samples from the distribution indexed by $\tilde{\theta}_n$ (not $\hat{\theta}_n$) and finds θ_n^* by maximizing the bootstrap likelihood over C. Under certain regularity conditions (Section 5.2) on the constraint set, $\sqrt{n}(\theta_n^* - \tilde{\theta}_n)$ has the same asymptotic distribution as $\sqrt{n}(\hat{\theta}_n - \theta_0)$.

All three adjustment procedures suffer from arbitrariness. In reality there is just one sample and its size n does not 'go to infinity'. Hence mere consistency justifies any subsampling size $m < n$ for the subsampling bootstrap, any shrinkage factor $\lambda_n \in (0, 1)$ for the shrinkage bootstrap, or any adjusted active set $\tilde{A}_n$ between the full set I and the active set at the maximum likelihood estimate

$$\hat{A}_n = \left\{ i \in I : g_i(\hat{\theta}_n) = 0 \right\} \tag{13}$$

for the adjusted active set bootstrap. But these procedures are useful despite their arbitrariness. Overlapping and nonoverlapping batch means (Meketon and Schmeiser, 1984), special cases of the subsampling bootstrap, have long been used in time series, and arbitrariness of the batch size has been accepted as an inherent property of the method.

Moreover, a simple example shows that no data-dependent adjustment can obtain consistency. Consider the type II multivariate test (7) with K the positive orthant. If the true parameter value is $\theta_0 = 0$, then the sufficient statistic $\sqrt{n}\bar{x}_n$ has the same distribution for all n, and the scaled constraint set $\sqrt{n}(C - \theta_0) = K$ is also the same for all n. Thus no procedure can be consistent unless it is either exact for all n or depends on n in a way that does not involve data.

3.3. Likelihood Ratio Tests

In likelihood ratio tests there are two constraint sets Θ_0 and Θ_1 in $\mathbb{R}^d$ such that the maximum likelihood estimates $\hat{\theta}_0$ and $\hat{\theta}_1$ are obtained by maximizing over the closures of Θ_0 and Θ_1 respectively. The test statistic is

$$t(x) = l_x\Big(\hat{\theta}_1(x)\Big) - l_x\Big(\hat{\theta}_0(x)\Big), \tag{14}$$

l_x being the log likelihood. Θ_0 is the null hypothesis of the test. Θ_1 is either the alternative or the union of the null and the alternative depending how one prefers to define the likelihood ratio test statistic.

The asymptotics of the likelihood ratio test described in Chernoff (1954) and Geyer (1994) hold when the closures of Θ_0 and Θ_1 both contain the true parameter value θ_0 and both have nontrivial tangent cones at θ_0.

The nontriviality assumption plays the same role as the assumption of *nested hypotheses* in the unconstrained case. The classical distinction between nested and nonnested hypotheses breaks down when there are inequality constraints. In general, as with type II multivariate one-sided tests, the null hypothesis is not contained in the closure of the alternative, and the hypotheses are not 'nested' in the usual sense. But if θ_0 is on the boundary between the hypotheses, so that both tangent cones are nontrivial, the likelihood ratio test statistic does behave as in the classical 'nested' case, converging to a nontrivial random variable rather than to infinity.

Thus the proper distinction is between nontrivial and trivial tangent cones rather than nested or nonnested hypotheses. The regular, nontrivial case occurring when θ_0 lies in the intersection of the closures of Θ_0 and Θ_1. Presumably there do exist problems in which one should apply the constrained analogue of Cox's test for nonnested hypotheses (Cox, 1961; Kent, 1986), but such tests are not considered here.

The procedures of adjusting the bootstrap to make it consistent are much the same when there are two constraint sets Θ_0 and Θ_1 as when there is just one. The shrinkage bootstrap uses the constraint sets

$$\Theta_{i,\lambda} = \lambda\Theta_i + (1-\lambda)\hat{\theta}, \qquad i = 0, 1. \tag{15}$$

An exception can be made when Θ_1 involves no inequality constraints so the unadjusted bootstrap consistently estimates the asymptotic distribution of $\hat{\theta}_1$. Then Θ_1 may be used instead of the shrunken version.

The adjusted active set bootstrap is more complicated and perhaps not applicable in general. It is applicable to specific problems in which the null hypothesis simply imposes additional constraints, keeping those for the alternative hypothesis, that is

$$\Theta_k = \{\, \theta \in \Theta : g_i(\theta) = 0,\, i \in E_k \text{ and } g_i(\theta) \le 0,\, i \in I_k \,\}, \qquad k = 0, 1$$

with $E_1 \subseteq E_0$ and $I_1 \subseteq I_0$. Define

$$\tilde{A}_n = \Big\{\, i \in I_0 : |g_i(\hat{\theta}_n)| \le Bn^{-\alpha} \,\Big\} \tag{16}$$

for any constant B and any constant $\alpha \in (0, 1/2)$. Then $\tilde{A}_n$ consistently estimates the true active set A_0 for the null hypothesis and $\tilde{A}_n \cap I_1$ consistently estimates the true active set A_1 for the alternative hypothesis, and the adjusted bootstrap sampling from the distribution indexed by $\tilde{\theta}_n$ that maximizes the likelihood over

$$\tilde{C}_n = \left\{ \theta \in \Theta_0 : g_i(\theta) = 0,\ i \in E_0 \cup \tilde{A}_n \text{ and } g_i(\theta) \leq 0,\ i \in I_0 \setminus \tilde{A}_n \right\}$$

consistently estimates the sampling distribution of the test statistic.

4. A CONCRETE EXAMPLE

For a simple example, we examine a type II multivariate one-sided test (7) with K the positive orthant. Then $\hat{\theta}_0$ is found by setting the negative components of x to zero. That is, $P_K(x) = y$ where $y_i = x_i$ if $x_i \geq 0$ and $y_i = 0$ otherwise. This procedure of setting the negative components to zero only works because of the assumed independence of the components of X and the orthogonality of the constraints. In general it is necessary to look at the signs of Lagrange multipliers to determine which inequality constraints hold with equality. By focussing on this simple example we avoid computational complexities that obscure the inferential issues. General inequality constrained inference does not present any essential difficulties (Geyer, 1991; Fletcher, 1987; Gill, Murray, and Wright, 1981).

We take $d = 20$, and the observation x to be the vector

$$\begin{array}{rrrrr} -0.2360 & 1.3422 & -0.0380 & 1.9903 & 1.7873 \\ 1.2240 & 1.6539 & 0.9006 & -0.8150 & 0.7314 \\ -1.4778 & -1.8332 & -1.6111 & -1.2611 & 1.9145 \\ -0.8351 & -0.5660 & -0.5925 & -2.3522 & 1.9421 \end{array} \tag{17}$$

which was chosen so that the bootstrap P-value would be less than .05 but the Neyman-Pearson P-value would be greater than .05 and about half the components would have each sign. Except for these considerations, this particular vector of observations was chosen haphazardly. It was not chosen with knowledge of how the Bayesian analysis would turn out.

4.1. The Neyman-Pearson Test

It is obvious that in our example the point θ_{LF} for which we obtain the worst case P-value is the origin (Robertson and Wegman, 1978), where the sampling distribution of the test statistic can be calculated analytically. The test statistic is

$$t(x) = \|x - \hat{\theta}_0\|^2 = \sum_{i=1}^{d} x_i^2 1\{x_i < 0\}$$

Since the X_i are independent, $P(X_i < 0) = 1/2$, and X_i^2 is chi-squared distributed with one degree of freedom, the distribution of $t(X)$ is a binomial mixture of chi-squares

$$P(t(X) \geq c) = \sum_{k=0}^{d} 2^{-d} \binom{d}{k} P(\chi_k^2 \geq c),$$

where χ_k^2 denotes a chi-squared random variable with k degrees of freedom (χ_0^2 being concentrated at zero). For our example $t(x) = 17.35355$, giving $P = 0.0846$. By conventional standards, the P-value is not impressive.

The point $\theta_{\mathrm{LF}} = 0$ that gives the worst-case P-value, is a very unlikely parameter value. A test of the point null hypothesis $\theta = 0$ against the unrestricted alternative has a test statistic $t(x) = \sum_i x_i^2 = 39.32743$, giving $P = 0.0061$. So there is a serious question whether the Neyman-Pearson P-value .085 has much relevance.

4.2. The Ordinary Parametric Bootstrap Test

The ordinary, unadjusted parametric bootstrap uses the same test test statistic $t(x) = 17.35355$ but compares it to the simulation distribution obtained by simulating data from the normal distribution centered at $\hat{\theta}_0(x)$, which is (17) with the negative components set to zero. Using 10 million bootstrap iterations gave $P = 0.0111$. This is now 'statistically significant' by conventional criteria and is less than one seventh the Neyman-Pearson, worst-case value.

A scientist wishing to reject the null hypothesis, would be much happier with the ordinary bootstrap test. There is, however, the problem that the bootstrap, being inconsistent, lacks justification.

4.3. The Shrinkage Adjusted Bootstrap

For $0 < \lambda < 1$, the constraint set for the shrinkage bootstrap is

$$C_\lambda = \left\{ \theta : \theta \geq (1 - \lambda)\hat{\theta}_0(x) \right\}$$

We are to simulate from $\hat{\theta}_0(x)$ and use the constraint set C_λ as the null hypothesis in the bootstrap simulations. In our special case, because of the translation invariance of the model, this is the same as simulating from $\lambda\hat{\theta}_0(x)$ and using the original constraint set C as the null hypothesis.

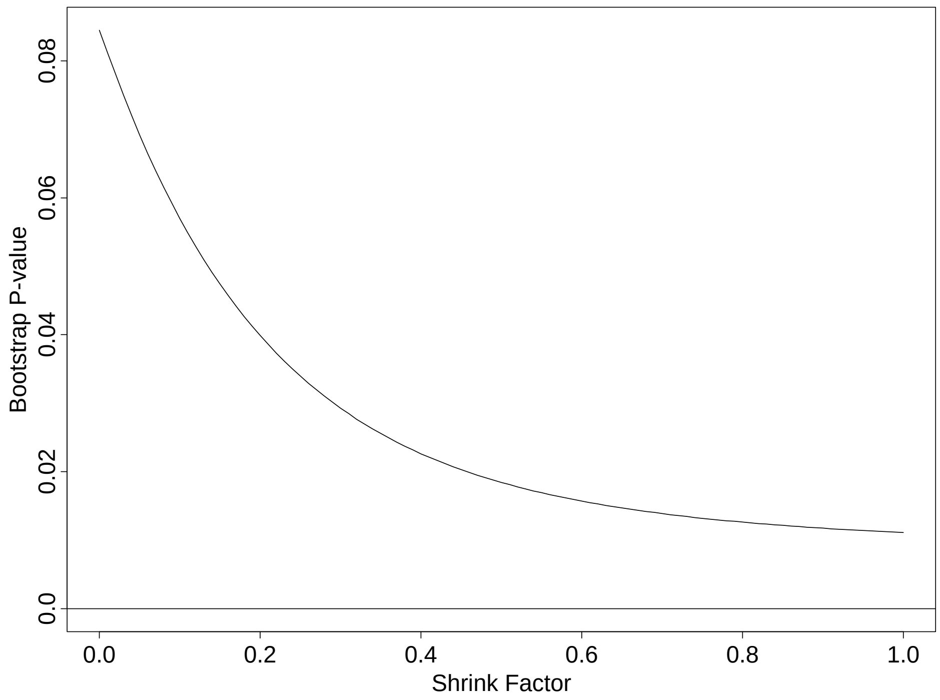

Figure 1: Adjusted Bootstrap P-value Curve. The P-value was calculated at 21 equally spaced λ values. The smooth curve is the interpolating cubic spline.

It is a problem that we have no idea what λ to use. In this example, though not in general, $\lambda = 0$ gives the Neyman-Pearson P-value, which we suspect of being too conservative, and $\lambda = 1$ gives the ordinary bootstrap P-value, which we suspect of being too liberal, but we have no idea where λ should be and asymptotics provide no guide. It is not the case that the smaller λ the better, since the asymptotics says that λ should be large compared to $n^{-1/2}$. Without getting into massive simulation (that will come later), it seems the best thing to do is to plot the adjusted bootstrap P-value as a function of the shrinkage λ (Figure 1).

This curve should constitute a satisfactory statistical inference in many cases. Though it does not give a single number as the inference, this will not be important if the inference is not critical. The lack of precision might even be considered good, since it prevents anyone from getting overly excited about the magic .05 significance level. In this example, the curve stays below .05 down to $\lambda = .135$, a considerable amount of shrinkage.

4.4. *The Ordinary Double Bootstrap*

A very simple way of thinking of the double bootstrap (Beran, 1988) of a significance test is that it just bootstraps a bootstrap estimate. The ordinary bootstrap P-value

$$p(x) = P_{\hat{\theta}_0(x)}\Big(t(X) \ge t(x)\Big)$$

is not the right thing because $\hat{\theta}_0(x)$ is not θ_0. One way to think of $p(x)$ is that it is just another test statistic—we reject the null when $p(x)$ is small—having an unknown sampling distribution. If we thought bootstrapping was appropriate when the test statistic was $t(x)$, then we should still think it is a good idea. The way to deal with not knowing the sampling distribution of $p(x)$ is to bootstrap.

Thus we simulate new data x^* from $P_{\hat{\theta}_0(x)}$, and for each simulated x^* we calculate $p(x^*)$. By averaging over many simulations we estimate the double bootstrap P-value

$$P_{\hat{\theta}_0(x)}\Big(p(X) \le p(x)\Big) \tag{18}$$

by the fraction of times that $p(x^*)$ is less than or equal to $p(x)$. Since $p(x^*)$ is itself calculated by simulation, we have a loop within a loop. For each x^* we calculate $\hat{\theta}_0(x^*)$ and simulate new data x^{**} from $P_{\hat{\theta}_0(x^*)}$, estimating $p(x^*)$ by the fraction of times that $t(x^{**})$ is greater than or equal to $t(x^*)$.

Not only does the double bootstrap provide what is hoped to be a better P-value, it also provides a simulation study of the single bootstrap (Geyer, 1991) by simulating the sampling distribution of $p(x^*)$. If the ordinary single bootstrap were doing the right thing and needed no correction, $p(X)$ would have a Uniform$(0,1)$ distribution, at least in the lower tail. If we make a quantile-quantile plot of the bootstrap samples $p(x^*)$, we can see how close its distribution is to the uniform. In our example, this plot (Figure 2) clearly shows the single bootstrap does not work. A 5000 by 5000 double bootstrap (5000 iterations of both inner and outer loops) estimates (18) to be $P = 0.0208$. The double bootstrap indicates that we need to increase the ordinary bootstrap P-value from .011 to .021.

There are two problems with the double bootstrap. First, like the single bootstrap, it is not consistent, though there is hope that the double bootstrap is better than the single bootstrap in the sense that the test statistic $p(X)$ is closer to pivotal than $t(X)$ (Beran, 1988). Second, the diagnostic nature of the iterated bootstrap is unsatisfactory. If the quantile-quantile plot from the double bootstrap is satisfactory, it shows that the single bootstrap is working, and the effort spent on the double bootstrap was unnecessary, except for the reassurance it provides. If the quantile-quantile plot is unsatisfactory, as it

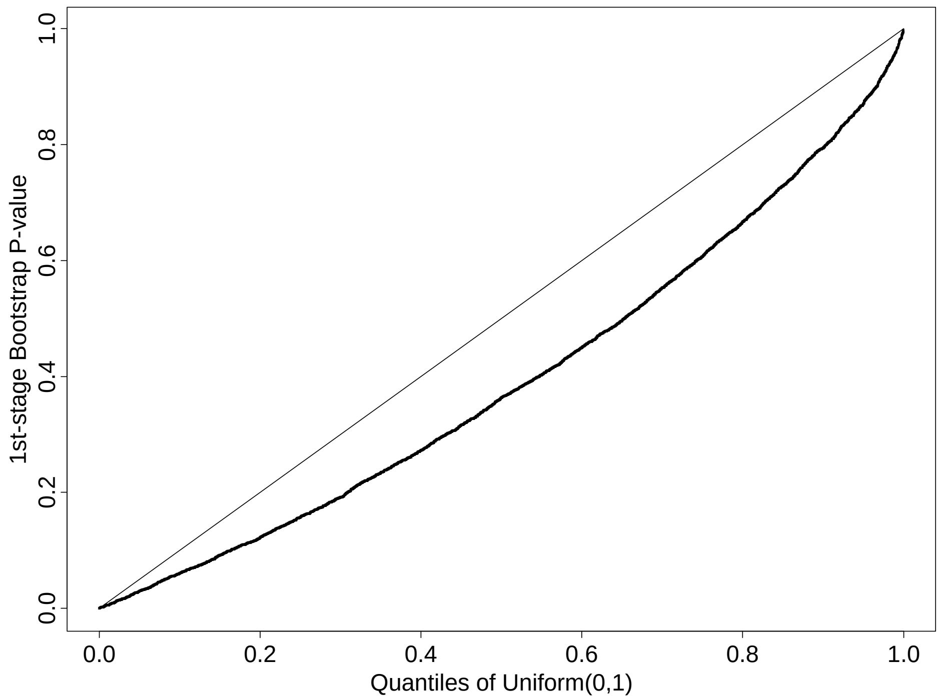

Figure 2: Quantile-Quantile Plot for Double Bootstrap

is in our example, this shows that the double bootstrap is necessary, but it does not show that it is sufficient. Only a triple bootstrap could do that.

4.5. Calibrating the Shrinkage Bootstrap

In inequality-constrained inference, a better method of bootstrap iteration is to use the bootstrap to calibrate the shrinkage factor. We do this by bootstrapping the shrinkage bootstrap and choosing the shrinkage factor λ that gives the sampling distribution of the bootstrap P-value $p_\lambda(X)$ that is most nearly uniformly distributed, reporting $p_\lambda(x)$, where λ is the chosen shrinkage factor and x the observed data.

In this double bootstrap we are trying to calculate

$$P_{\hat{\theta}_0(x)}\Big(p_\lambda(X) \le p_\lambda(x)\Big), \tag{19}$$

which we estimate by the fraction of times that $p_\lambda(x^*)$ is less than or equal to $p_\lambda(x)$, where $x^* \sim P_{\hat{\theta}_0(x)}$. We estimate $p_\lambda(x^*)$ by simulating data $x^{**} \sim P_{\hat{\theta}_0(x^*)}$ and calculating the fraction of times that $t_\lambda(x^{**})$ is greater than or equal to

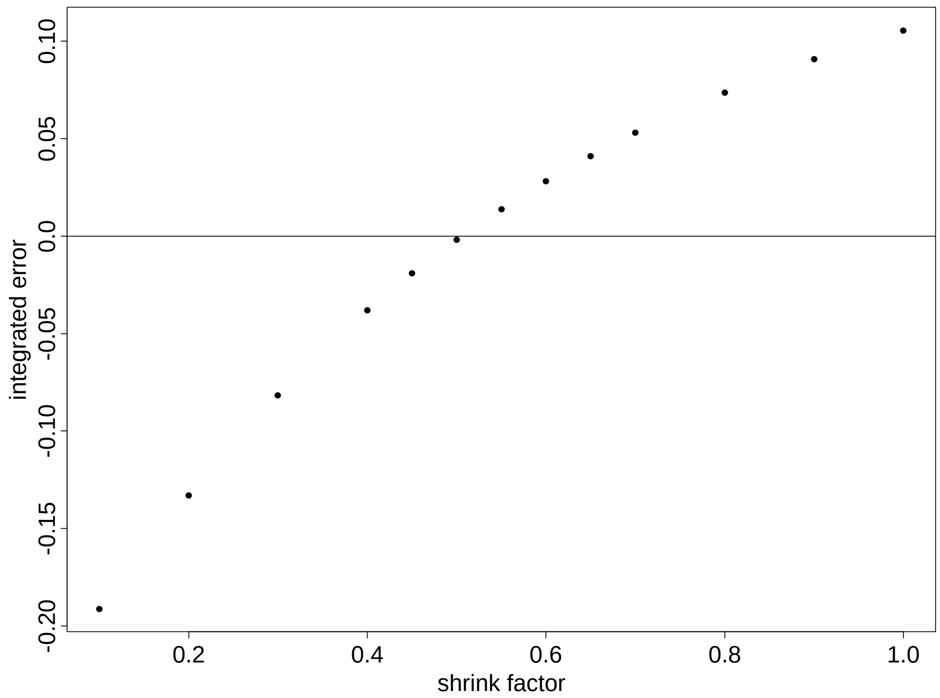

Figure 3: Plot Used to Estimate Optimal Shrinkage. Plot of the area between the line $y = x$ and the empirical distribution of $p_\lambda(x^*)$ for various λ. Each point involves a 5000 by 5000 double bootstrap.

$t(x^*)$, where $t_\lambda(x)$ indicates the likelihood ratio test statistic calculated using the shrunken constraint sets (15).

It is not clear how best to determine the optimal shrinkage factor, but the following procedure seems reasonable. For each λ on a grid of values between zero and one, bootstrap the shrinkage bootstrap and make a quantile-quantile plot like Figure 2. Integrate the difference between the theoretical and empirical distribution functions, that is, calculate the mean difference between the abscissa and the ordinate in the quantile-quantile plot. (For Figure 2 this is a large positive number, because the empirical distribution function bows below the straight line.) Figure 3 shows such a plot.

Interpolation indicates that the curve crosses zero at $\lambda = 0.5052$, and that is the λ we use for our final inference. Using 10 million bootstrap iterations gave $P = 0.01835$, roughly the same as that given by the ordinary double bootstrap. Figure 4 shows the performance of the shrinkage bootstrap at the nearby grid point $\lambda = .5$.

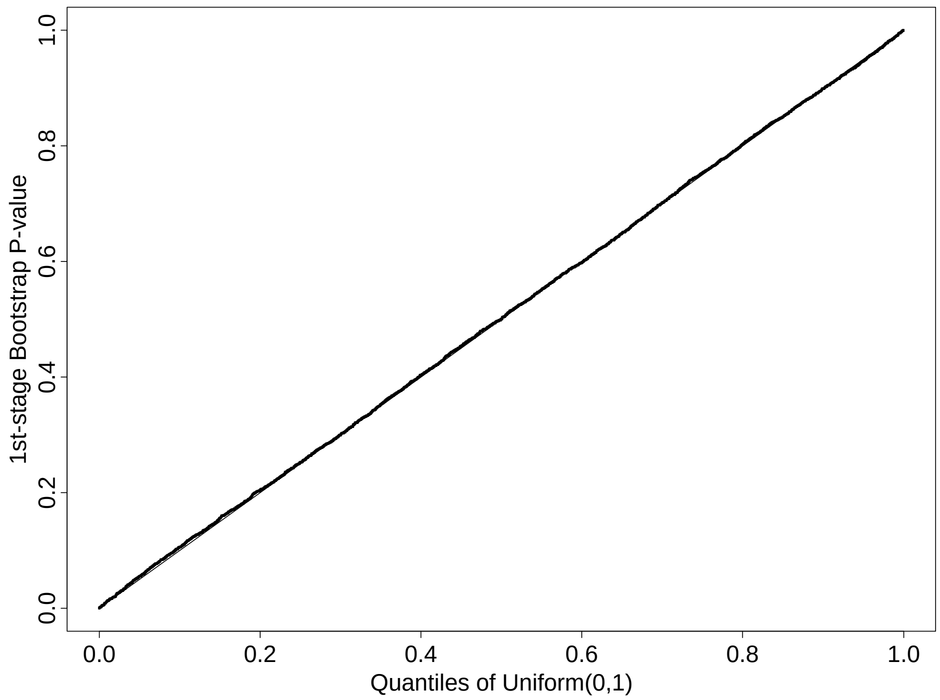

Figure 4: Quantile-Quantile Plot for Double Bootstrap with Shrinkage Factor $\lambda = .5$. This is a 5000 by 5000 double bootstrap.

From the figure it seems that this adjusted bootstrap works perfectly, and an appeal to asymptotics to justify the bootstrap is not really necessary. There is still a nagging bit of residual doubt. Our P-value (19) is still not the right thing because $\hat{\theta}_0(x)$ is not θ_0. Moreover, by adopting a data-dependent scheme, we have lost consistency (Section 3.2).

It should also be conceded that there is no guarantee that it will be so easy to chose the optimal λ in other problems. Here we used the integrated signed difference because the quantile-quantile plots were bowed down for $.5 < \lambda < 1$ and bowed up for $0 < \lambda < .5$. If the behaviour of the quantile-quantile plots were more complicated, a different criterion would have to be used. No theory says there always exists a λ that does a good job.

Despite these disclaimers, the shrinkage bootstrap with optimal shrinkage estimated by bootstrap calibration seems to be all one could ask of a frequentist inference in a complex problem. The calculations for Figure 3 ran overnight, about twelve hours on a fast workstation (20 million floating point operations per second) and would have taken several days on older equip-

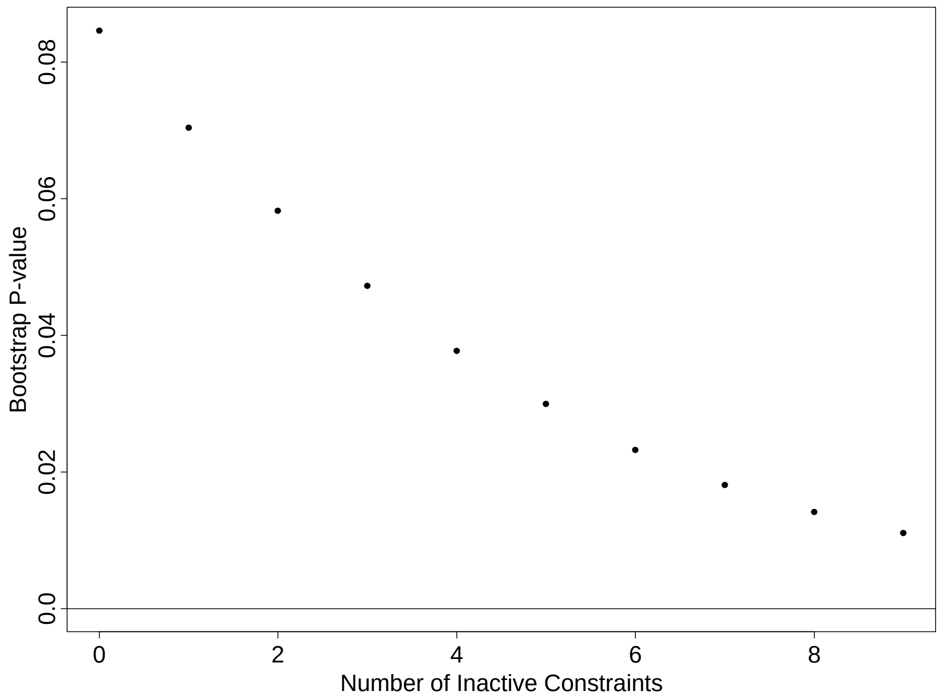

Figure 5: Adjusted Active Set Bootstrap P-values.

ment. Figure 1, which took less than an hour and could have done with less precision in only a few minutes, would suffice for many analyses. Nevertheless, if one insists on a single number to be the best possible P-value, then the $P = .01835$ produced in this section is that number.

4.6. The Adjusted Active Set Bootstrap

Like the shrinkage bootstrap, the adjusted active set bootstrap has a problem of arbitrariness. The estimated active set $\tilde{A}$ can be any set between the maximum likelihood active set $\hat{A}$ (11 constraints) and the full set I (20 constraints). There are 9 constraints that may be in or out of $\tilde{A}$ and hence $2^9 = 512$ choices for $\tilde{A}$. Since we cannot examine them all, it seems reasonable to add additional constraints one at a time in order of the size of the residuals. The first estimator examined is $\hat{\theta}$, corresponding to $\tilde{A} = \hat{A}$. The second sets the smallest positive component (0.7314) of $\hat{\theta}$ to zero. The third sets the next smallest, and so forth. The last sets all components to zero, corresponding to $\tilde{A} = I$. Figure 5 shows the resulting bootstrap P-values. Imposing with equality the two constraints with smallest residuals at the maximum likelihood estimate gives $P = .0181$, nearly the same as the P-values obtained from the double bootstrap and the shrinkage bootstrap with shrinkage chosen by bootstrap calibration.

It seems difficult to calibrate this procedure via bootstrapping the adjusted bootstrap because of the discreteness of $\tilde{A}$. To calibrate active set adjustment, we would need a rule for adjusting the active set that was completely specified and could be applied by the computer to the bootstrap iterations. It may be just coincidence, but the two additional constraints imposed to get the 'right' $P = .0181$ are the two constraints with residuals less than one, which is their standard error. Perhaps that is a reasonable rule. In our example, because of the independence of the components of the data vector and the orthogonality of the constraints, the $g_i(\hat{\theta})$ are independent. In general, they will not be. Perhaps the residuals should be examined sequentially. Impose the constraint with the smallest residual, find the corresponding $\tilde{\theta}$, look at the new residuals $g_i(\tilde{\theta})$, and so forth.

We shall not attempt to decide what is best here. Bootstrap calibration will be more complicated here than with shrinkage adjustment and perhaps should not be attempted. Active set adjustment seems more appropriate when one wants a good, simple, easily understood analysis. Its rule is 'bootstrap using a point $\tilde{\theta}$ that has a few additional constraints imposed with equality' because the maximum likelihood estimate typically does not have enough constraints satisfied with equality. The main point is that one should either impose additional constraints or shrink.

4.7. Bayes Factors

The Bayesian procedures we shall apply to our example will all be based on an improper, uniform prior. Under this prior, the posterior probability of the first orthant is

$$\prod_{i=1}^{d} \Phi(x_i) = 3.385 \times 10^{-11},$$

where Φ denotes the standard normal cumulative distribution function. This of course bears no relation to a frequentist P-value and only indicates that the first orthant is very small.

Strictly speaking a Bayes factor is undefined for improper priors, but by symmetry we may take the prior probability of the first orthant to be $2^{-20} = 9.537 \times 10^{-7}$. The Bayes factor, the ratio of posterior to prior odds is then 3.55×10^{-5}. Using a very diffuse proper prior that is symmetric under rotation about the origin would produce much the same results. Our best estimate of the frequentist P-value is over 500 times this Bayes factor.

4.8. Bayesian P-values

The Bayesian P-value for this problem, more precisely the posterior predictive P-value (Meng, 1994) is the posterior expectation of the frequentist P-value considered a function of θ

$$\int P_\theta\bigl(t(X) \geq t(x)\bigr)\pi(\theta|x)\,d\theta, \tag{20}$$

where $\pi(\theta|x)$ is the posterior density. This cannot be calculated exactly, but is easily done by a simulation computationally resembling the double bootstrap. One simulates θ^* values from the posterior, for each such θ^* simulates data $x^* \sim P_{\theta^*}$, and estimates $p(\theta^*) = P_{\theta^*}\bigl(t(X) \geq t(x)\bigr)$ by the fraction of $t(x^*)$ that are greater than or equal to $t(x)$. Then (20) is estimated by the average of the $p(\theta^*)$.

For our example, using the improper, uniform prior, the Bayesian P-value is $P = 0.00124$. Our best estimate of the frequentist P-value is 15 times larger. This hybrid of Bayesian and frequentist inference is not so wildly liberal as the Bayes factor but is still far too liberal.

The reason for the marked differences between the Bayesian and frequentist inferences here are, of course, attributable to the use of diffuse priors which put most of the prior probability far away from the boundary of the null hypothesis. It is an interesting open question how one should formulate sensible priors for this problem.

5. THEORY

The asymptotics of inequality-constrained maximum likelihood estimates and likelihood ratio tests are described by Chernoff (1954), Le Cam (1970), Self and Liang (1987), and Geyer (1994). This section develops the corresponding theory for our two types of bootstrap adjustment. We shall not attempt to develop the weakest possible stochastic regularity conditions. In particular we assume the likelihood is well defined in a neighbourhood of the true parameter value and continuous. These conditions can presumably be relaxed to permit the likelihood to be defined only on the constraint set and be only upper semicontinuous, obtaining a theory with the flavour of Geyer (1994), but that is a subject for further research.

Let $P_{\theta,n}$ denote the probability distribution of the data for parameter θ and sample size n, and let

$$l_n(\phi, \theta) = \log \frac{dP_{\phi,n}}{dP_{\theta,n}}$$

denote the Radon-Nikodym derivative with respect to $P_{\theta,n}$ of the part of $P_{\phi,n}$ that is absolutely continuous with respect to $P_{\theta,n}$. Then we assume

(a) $P_{\phi,n}$ is defined for all ϕ in some neighbourhood of the true parameter value θ_0.

(b) The log likelihood function $\phi \mapsto l_n(\phi, \theta)$ is continuous for each θ in a neighbourhood of the true parameter value θ_0 and for all data values.

A likelihood problem satisfies the LAN (local asymptotic normality) conditions at a point θ for a rate sequence $\delta_n \to 0$ (usually $\delta_n = n^{-1/2}$) if (Le Cam and Yang, 1990, pp. 54 ff.) if (a) holds and also

(c) The sequences $P_{\theta,n}$ and $P_{\theta+\delta_n\phi_n}$ are contiguous for any bounded sequence ϕ_n.

(d) There exist random vectors S_n and nonrandom, positive definite matrices $V_n \to V$ such that for any bounded sequence ϕ_n

$$l_n(\theta + \delta_n\phi_n, \theta) - \left[\phi_n' S_n - \tfrac{1}{2}\phi_n' V_n \phi_n\right]$$

converges in probability to zero under $P_{\theta,n}$.

We shall say the problems satisfies the ULAN (uniformly LAN) conditions if (d) is replaced by

(e) There exist random vectors S_n and nonrandom, positive definite matrices $V_n \to V$ such that for any $R > 0$

$$\sup_{\|\phi\| \le R} l_n(\theta + \delta_n\phi, \theta) - \left[\phi' S_n - \tfrac{1}{2}\phi' V_n \phi\right]$$

converges in probability to zero under $P_{\theta,n}$.

It is well known (Le Cam and Yang, 1990, p. 59) that the LAN conditions are not sufficient for asymptotics of maximum likelihood. Some uniformity of convergence is necessary. The ULAN condition is stronger than necessary but allows a simple treatment of the asymptotics of the bootstrap.

The continuity assumption (b) makes the log likelihood function a random element of the space $C(\mathbb{R}^d)$ of all continuous functions on $\mathbb{R}^d$ with the topology of uniform convergence on compact sets, which is a complete separable metric space.

Lemma 1. If the continuity and ULAN conditions hold at θ_0, then for any nonrandom bounded sequence ψ_n the random function h_n defined by

$$h_n(\phi) = l_n(\theta_0 + \delta_n \psi_n + \delta_n \phi, \theta_0 + \delta_n \psi_n). \tag{21}$$

converges in law in $C(\mathbb{R}^d)$ under $P_{\theta_0 + \delta_n \psi_n, n}$ to the random function q defined by

$$q(\phi) = \phi' Z - \tfrac{1}{2} \phi' V \phi,$$

where Z is an $N(0, V)$ random vector.

Proof. Except in events whose probability goes to zero as $n \to \infty$

$$l_n(\theta_0 + \delta_n \psi_n + \delta_n \phi, \theta_0 + \delta_n \psi_n) = l_n(\theta_0 + \delta_n \psi_n + \delta_n \phi, \theta_0) - l_n(\theta_0 + \delta_n \psi_n, \theta_0)$$

(Le Cam and Yang, 1990, p. 56). So by (e) h_n is equal to the random function w_n defined by

$$w_n(\phi) = \phi'(S_n - V_n \psi_n) - \tfrac{1}{2} \phi' V_n \phi \tag{22}$$

except for terms that are $o_p(1)$ uniformly on compact sets. That is, $h_n = w_n + o_p(1)$ considered as random elements of $C(\mathbb{R}^d)$ under $P_{\theta_0, n}$ and also under $P_{\theta_0 + \delta_n \psi_n, n}$, since, by contiguity, any sequence of random variables that converges in probability to zero under $P_{\theta_0, n}$ also converges in probability to zero under $P_{\theta_0 + \delta_n \psi_n, n}$.

Now suppose $\psi_n \to \psi$ so that $S_n - V_n \psi_n$ converges in law to $N(-V\psi, V)$. Then by contiguity (Le Cam and Yang, 1990, pp. 24 and 81) $S_n - V_n \psi_n$ converges to $N(0, V)$ under $P_{\theta_0 + \delta_n \psi_n, n}$. Since for any subsequence there is a subsubsequence such that ψ_n converges and for each such subsubsequence the limit is the same, $S_n - V_n \psi_n$ converges to $N(0, V)$ under $P_{\theta_0 + \delta_n \psi_n, n}$ regardless of whether ψ_n converges.

Now by the continuous mapping theorem w_n converges in law to q, and hence h_n also converges in law to q (under $P_{\theta_0 + \delta_n \psi_n, n}$). □

In addition to the assumptions already made, we must assume that the maximum likelihood estimate is δ_n-consistent. When $\delta_n = n^{-1/2}$ this is implied by consistency and the other regularity conditions (Geyer, 1994). We also need to assume that the bootstrap maximum likelihood estimator θ_n^* is δ_n-consistent, i. e. that $\theta_n^* - \hat{\theta}_n = O_p(\delta_n)$. This can always be forced, because $\hat{\theta}_n$ is known and can be used in calculating θ_n^*.

5.1. *The Shrinkage Adjusted Bootstrap*

Theorem 1. Under the conditions assumed above (ULAN, continuity, consistency of maximum likelihood, and δ_n-consistency of both $\hat{\theta}_n$ and θ_n^*), the shrinkage adjusted bootstrap with shrinkage factor λ_n satisfying $\lambda_n \to 0$ and $\delta_n^{-1}\lambda_n \to \infty$ is consistent.

Proof. Consider a nonrandom sequence θ_n such that $\psi_n = \delta_n^{-1}(\theta_n - \theta_0)$ is bounded. Then by the Lemma h_n converges to q regardless of the sequence ψ_n. Also $\lambda_n \delta_n^{-1}(C - \theta_n) \to T_C(\theta_0)$. Let θ_n^* be the bootstrap maximum likelihood estimate simulating from $P_{\theta_n,n}$ and maximizing over $\lambda_n C + (1-\lambda_n)\theta_n$. Then invoking the Skorohod and Prohorov theorems as in Theorem 4.4 of Geyer (1994) and using the fact that the sum of uniformly converging and epiconverging functions epiconverges (Attouch, 1984, Theorem 2.15) proves that the asymptotic distribution of θ_n^* does not depend on the sequence θ_n and is asymptotic distribution of the MLE, the distribution of the maximizer of q over $T_C(\theta_0)$. Now we apply the Skorohod theorem to the MLE, getting an almost surely convergent sequence $\hat{\theta}_n$ having the same properties assumed for θ_n. In the Skorohod representation the law of θ_n^* converges almost surely to the correct distribution, and this implies convergence in probability of the distribution of θ_n^* to correct asymptotic distribution. □

The same argument applied simultaneously to the constrained estimates for null and alternative hypotheses shows the shrinkage bootstrap consistently calculates the sampling distribution of the likelihood ratio test statistic.

5.2. *The Adjusted Active Set Bootstrap*

The argument here is more complicated. We need a *constraint qualification* assumption, either of the equivalent hypotheses of the following theorem.

Theorem 2. (Rockafellar and Wets, forthcoming). Suppose the constraint functions g_i are continuously differentiable, and suppose that there does not exist a multiplier vector μ indexed by $E \cup I$, such that (1) $\mu_i \geq 0$ and $\mu_i g_i(\theta_0) = 0$ for all $i \in I$, (2) $\mu_i \neq 0$ for some $i \in E \cup I$, and (3) $\sum_i \mu_i \nabla g_i(\theta_0) = 0$. An equivalent condition is that the gradients $\nabla g_i(\theta_0)$, $i \in E$ are linearly independent and the set

$$W = \left\{ v \in \mathbb{R}^d : \langle \nabla g_i(\theta_0), v \rangle = 0,\ i \in E \text{ and } \langle \nabla g_i(\theta_0), v \rangle < 0,\ i \in A \right\}$$

is nonempty. Then the closure of W

$$\left\{ v \in \mathbb{R}^d : \langle \nabla g_i(\theta_0), v \rangle = 0,\ i \in E \text{ and } \langle \nabla g_i(\theta_0), v \rangle \leq 0,\ i \in A \right\} \tag{23}$$

is the tangent cone, C is Clarke regular at θ_0, and the constraint qualification conditions hold at every point in some neighbourhood of θ_0.

A much simpler constraint qualification condition, which implies those of the theorem, is that the $\nabla g_i(\theta_0)$, $i \in E \cup A$ are linearly independent.

Theorem 3. Under the conditions assumed above (ULAN, continuity, consistency of maximum likelihood, δ_n-consistency of both $\hat{\theta}_n$ θ_n^*, and constraint qualification), the adjusted active set bootstrap is consistent.

Proof. Consider a nonrandom sequence θ_n such that $\psi_n = \delta_n^{-1}(\theta_n - \theta_0)$ is bounded so that h_n converges to q regardless of the sequence ψ_n. Also assume that θ_n eventually satisfies the correct active set, i. e., $g_i(\theta_n) = 0$, for all $i \in A$. Then Clarke regularity implies

$$T_C(\theta_0) \subseteq \liminf_n \delta_n^{-1}(C - \theta_n).$$

Next we establish

$$\limsup_n \delta_n^{-1}(C - \theta_n) \subseteq T_C(\theta_0). \tag{24}$$

A vector v lies in the left hand side of (24) if and only if there is a subsequence n_k and a sequence ϕ_k in C satisfying

$$\delta_{n_k}^{-1}(\phi_k - \theta_{n_k}) \to v,$$

which implies that $\delta_{n_k}^{-1}(\phi_k - \theta_0)$ is also bounded. By differentiability of the constraints

$$\delta_{n_k}^{-1}\Big(g_i(\phi_k) - g_i(\theta_{n_k})\Big) \to \langle \nabla g_i(\theta_0), v \rangle.$$

For $i \in I$, the left hand side is identically zero, and for $i \in A$, $g_i(\phi_k) \leq 0$ and eventually $g_i(\theta_{n_k}) = 0$. Hence v is in the tangent cone (23). This proves

$$\delta_n^{-1}(C - \theta_n) \to T_C(\theta_0).$$

Now the proof continues as in the proof of Theorem 1. Two invocations of Skorohod and one of Prohorov show the bootstrap is consistent. □

5.3. A Counterexample

That some regularity conditions like constraint qualification are necessary is shown by the following counterexample. The model is Normal(θ, I) with

$\theta = (u, v) \in \mathbb{R}^2$. There are two inequality constraints. The first is $v \geq 0$. The second is $g_2(\theta) \geq 0$, where

$$g_2(\theta) = \begin{cases} u^2 \sin(\pi \log_4 u)^2 - v, & u > 0 \\ -v, & u = 0 \\ -u^2 - v, & u < 0 \end{cases}$$

The true parameter value is $(0, 0)$.

This example satisfies the regularity conditions for $\hat{\theta}_n$ to have asymptotics, the tangent cone being $\{ (u, v) : u \geq 0, v = 0 \}$, and the asymptotic distribution of the maximum likelihood estimate being the projection of a Normal(θ, I) random vector on the tangent cone giving a random vector $(U, 0)$, where $U = \max(0, Z)$, Z standard normal.

The estimated active set eventually imposes both inequality constraints, and $\tilde{\theta}$ is the maximizer of the likelihood over the set

$$\tilde{C} = \{(0, 0)\} \cup \left\{ (4^{-k}, 0) : k \in \mathbb{N} \right\}, \tag{25}$$

which is the set of θ such that $g_1(\theta) = 0$ and $g_2(\theta) = 0$. This set is not Chernoff regular, and $\sqrt{n}(\tilde{\theta}_n - \theta_0)$ does not converge in law to any distribution (Geyer, 1994), though it is bounded in probability. The adjusted active set bootstrap fails to be consistent because $\sqrt{n}\tilde{\theta}_n$ is the projection of a standard normal random vector on the set $\sqrt{n}\tilde{C}$, and does not converge in probability to zero. Thus $\sqrt{n}(C - \tilde{\theta}_n)$ contains subsequences that converge to vectors $(u, 0)$ with $u < 0$, and such vectors are not in the tangent cone.

5.4. Variations on Adjustment

In some problems it may be necessary to use non-isotropic shrinkage or active set adjustment. This is obvious for active set adjustment, because multiplying the constraints by different constant factors does not change the constraint set C but does change the estimated active set given by (12) or (16). When the $g_i(\hat{\theta}_n)$ have very different asymptotic variances, it would seem sensible take this into account in constructing the estimated active set.

The case for non-isotropic shrinkage is less clear, but it is clear that one could use shrinkage of the form

$$C_\Lambda = \Lambda C + (I - \Lambda)\hat{\theta}$$

instead of (10), where Λ is a diagonal matrix with diagonal entries in $(0, 1)$ with $\Lambda_n \to 0$ and $\sqrt{n}\,\mathrm{diag}(\Lambda_n) \to \infty$, and this would not change the consistency proof. This would seem sensible when the original parameterization of the problem is ill-conditioned so that different components of $\hat{\theta}$ have very different asymptotic variances.

6. DISCUSSION

Geyer (1991) claimed that the double bootstrap provided a 'sound recipe' for data analysis in inequality constrained maximum likelihood problems. This was always a bit suspect in that neither the ordinary bootstrap nor the ordinary double bootstrap are consistent, although the ordinary double bootstrap seems to do a reasonable job in the problem studied in Geyer (1991) and in the example in Section 4. The double bootstrap is also a great deal of work, especially in problems (Shaw and Geyer, submitted) where maximum likelihood estimates are difficult to calculate, although the bootstrap recycle algorithm of Newton and Geyer (1994) can save some of the work. There is thus a need both for simple procedures not involving the double bootstrap and for a consistent double bootstrap. The adjusted bootstraps in Sections 4.3, 4.5, and 4.6 meet these needs.

There is also a more general point to be made about inequality constraints. They make a difference in every aspect of statistical inference. The problem studied here tells us something about tests of significance and competing Bayesian procedures, about the parametric bootstrap, and about the double bootstrap and bootstrap adjustment and calibration. Many procedures thought to be well understood suddenly become problematical when general inequality constraints are introduced. We have seen this is the case with the standard Neyman-Pearson tests with P-values given by (1), with the ordinary bootstrap and double bootstrap, and with Bayes factors and Bayesian P-values.

The argument given here that the generality claimed for the Neyman-Pearson theory is spurious and only makes sense for the special cases in which uniformly most something or other tests exist is not new. Fisher (1973, Section 4.5) made exactly the same point: the Neyman-Pearson theory has trouble with compound null hypotheses. Fisher's example was like ours in that it involved inequality constraints, although it was not convincing because, while Fisher claimed the Neyman-Pearson theory could not handle his example, he had no good analysis of his own to offer. Fisher's example now seems ill chosen because it does not distinguish amongst the methods discussed here, the Neyman-Pearson test based on the least favourable parameter value and the ordinary parametric bootstrap giving much the same answers. Fisher's basic point, however, does seem correct.

As for the comparison of 'frequentist' tests and their Bayesian competitors, it is not clear what one is to make of the example studied here. It does at least show that the phenomena studied by Berger and Sellke (1987) and Casella and Berger (1987) are not general and that the comparison of Bayesian and frequentist 'tests' does not go all one way. The example also contradicts the conventional wisdom that Bayesian inference under inequality constraints is unproblematical.

References

Attouch, H. (1984) *Variational Convergence of Functions and Operators.* Boston: Pitman.

Beran, R. (1988) Prepivoting test statistics: A bootstrap view of asymptotic refinements. *J. Am. Statist. Ass.*, **83**, 687–697.

Berger, J. O., and Sellke, T. (1987) Testing a point null hypothesis: The irreconcilability of P values and evidence (with discussion). *J. Am. Statist. Ass.*, **82**, 112–139.

Berger, R. L., and Boos, D. D. (1994) P values maximized over a confidence set for the nuisance parameter. *J. Am. Statist. Ass.*, **89**, 1012–1016.

Casella, G., and Berger, R. L. (1987) Reconciling Bayesian and frequentist evidence in the one-sided testing problem (with discussion). *J. Am. Statist. Ass.*, **82**, 106–139.

Chernoff, H. (1954) On the distribution of the likelihood ratio. *Ann. Math. Statist.*, **25**, 573–578.

Cox, D. R. (1961) Tests of separate families of hypotheses. *Proc. 4th Berkeley Symp.*, **1**, 105–123.

Fisher, R. A. (1973) *Statistical methods and scientific inference* (3rd ed., rev.) New York: Hafner.

Fletcher, R. (1987) *Practical Methods of Optimization* (2nd ed.) New York: John Wiley.

Geyer, C. J. (1991) Constrained maximum likelihood exemplified by isotonic convex logistic regression. *J. Am. Statist. Ass.*, **86**, 717–724.

Geyer, C. J. (1994) On the asymptotics of constrained M-estimation. *Ann. Statist.*, **22**, 1993–2010.

Gill, P. E., Murray, W., and Wright, M. E. (1981), *Practical Optimization.* New York: Academic Press.

Kent, J. T. (1986) The underlying structure of nonnested hypothesis tests. *Biometrika*, **73**, 333–343.

Kudô, A. (1963) A multivariate analogue of the one-sided test. *Biometrika*, **50**, 403–418.

Le Cam, L. (1953) On some asymptotic properties of maximum likelihood estimates and related Bayes estimates. *Univ. of Calif. Publ. in Statist.*, **1**, 277–330.

Le Cam, L. (1970) On the assumptions used to prove asymptotic normality of maximum likelihood estimates. *Ann. Math. Statist.*, **41**, 802–828.

Le Cam, L. and Yang, G. L. (1990) *Asymptotics in Statistics: Some Basic Concepts.* New York: Springer-Verlag.

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

Meng, X.-L. (1994) Posterior predictive p-values. *Ann. Statist.* **22**, 1142–1160.

Meketon, M. S. and Schmeiser B. W. (1984) Overlapping batch means: something for nothing? In S. Sheppard, U. Pooch, and D. Pegden (eds.) *Proceedings of the 1984 Winter Simulation Conference*, 227–230.

Newton, M. A. and Geyer, C. J. (1994) Bootstrap recycling: A Monte Carlo alternative to the nested bootstrap. *J. Am. Statist. Ass.*, **89**, 905–912.

Perlman, M. D. (1969) One-sided testing problems in multivariate analysis. *Ann. Math. Statist.*, **40**, 549–567.

Politis, D. N. and Romano, J. P. (1994) Large sample confidence regions based on subsamples under minimal assumptions. *Ann. Statist.* **22**, 2031–2050.

Robertson, T. and Wegman, E. J. (1978) Likelihood ratio tests for order restrictions in exponential families. *Ann. Statist.*, **6**, 485–505.

Robertson, T., Wright, F. T., and Dykstra, R. L. (1988) *Order Restricted Statistical Inference.* New York: John Wiley.

Rockafellar, R. T. (1970) *Convex Analysis.* Princeton University Press.

Rockafellar, R. T. and Wets, R. J. B. (forthcoming) *Variational Analysis.* New York: Springer-Verlag.

Rubin, D. B. (1984) Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Ann. Statist.*, **12**, 1151-1172.

Self, S. G., and Liang, K.-Y. (1987) Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *J. Am. Statist. Ass.*, **82**, 605–610.

Shaw, F. H., and Geyer, C. J. (submitted) Estimation and testing in constrained covariance component models.

Warrack, G. and Robertson, T. (1984) A likelihood ratio test regarding two nested but oblique order-restricted hypotheses. *J. Am. Statist. Ass.*, **79**, 881–886.

Wolak, F. A. (1987) An exact test for multiple inequality and equality constraints in the linear regression model. *J. Am. Statist. Ass.*, **82**, 782–793.