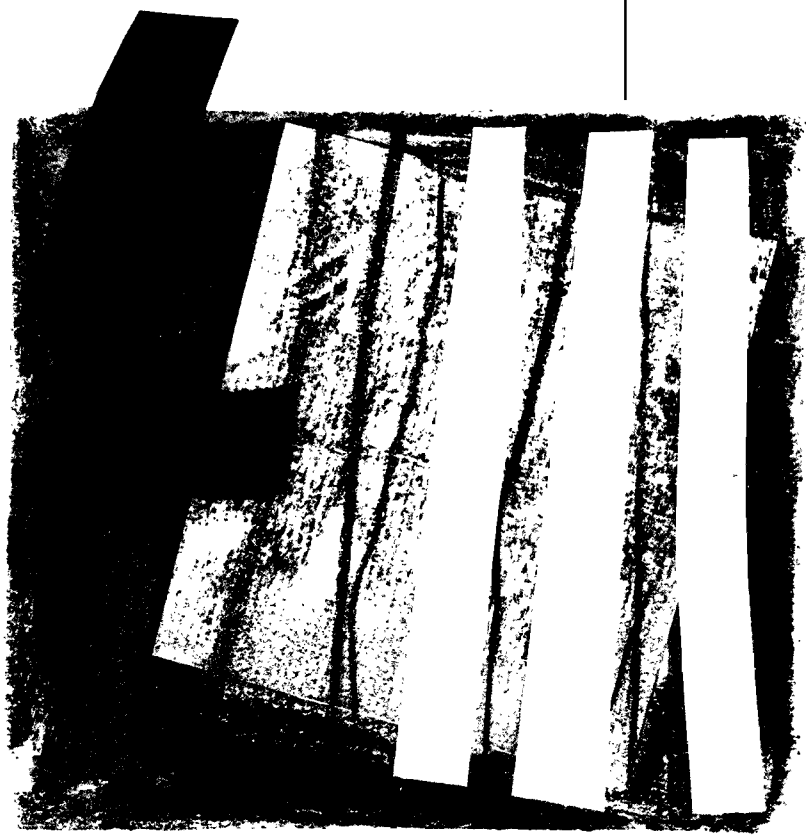
SUBSAMPLING INFERENCE IN CUBE ROOT ASYMPTOTICS WITH AN APPLICATION TO MANSKI'S MAXIMUM SCORE ESTIMATOR

Miguel A. Delgado,
Juan M. Rodríguez-Poo
Michael Wolf

00-78



WORKING PAPERS

Universidad Carlos III de Madrid

Working Paper 00-78
Statistics and Econometrics Series 37
November 2000

Departamento de Estadística y Econometría
Universidad Carlos III de Madrid
Calle Madrid, 126
28903 Getafe (Spain)
Fax (34) 91 624-98-49

SUBSAMPLING INFERENCE IN CUBE ROOT ASYMPTOTICS WITH AN APPLICATION TO MANSKI'S MAXIMUM SCORE ESTIMATOR

Miguel A. Delgado, Juan M. Rodríguez-Poo and Michael Wolf *

Abstract

Kim and Pollard (1990) showed that a general class of M-estimators converge at rate $n^{1/3}$ rather than at the standard rate $n^{1/2}$. Many times, this situation arises when the objective function is non-smooth. The limiting distribution is the (almost surely unique) random vector that maximizes a certain Gaussian process and is difficult to analyze analytically. In this paper, we propose the use of the subsampling method for inferential purposes. The general method is then applied to Manski' s maximum score estimator and its small sample performance is highlighted via a simulation study.

Keywords: Cube root asymptotics; Maximum score estimator; Subsampling.

*Delgado, Departamento de Estadística y Econometría, Universidad Carlos III de Madrid, C/ Madrid 126 28903 Getafe, Madrid, Spain; Rodríguez-Poo, Departamento de Economía Universidad de Cantabria, 39005 Santander, Spain; Wolf, Departamento de Estadística y Econometría, Universidad Carlos III de Madrid, e-mail: mwolf@est-econ.uc3m.es, Phone: 34-91-6249893. The first and the third author acknowledge financial support from the Dirección General de Enseñanza Superior Research Project PB98-0025 while the second author acknowledges financial support from the Dirección General de Enseñanza Superior Research Project PB96-1469-C05-03.

1 Introduction

Kim and Pollard (1990) studied a class of M-estimators defined by maximization of processes

$$P_n g_\theta = \frac{1}{n}\sum_{i=1}^{n} g_\theta(X_i),$$

where $\{X_i\}$ is an independent and identically distributed (i.i.d.) sequence with distribution P, $\{g_\theta : \theta \in \Theta\}$ is a class of functions indexed by a subset Θ of $\mathbb{R}^d$, and P_n is the empirical measure. When g_θ is sufficiently smooth, the resulting estimator $\hat{\theta}_n$ will typically converge at the standard rate $n^{1/2}$ to a limiting Gaussian distribution with mean zero; a basic example is the sample mean with $g_\theta(X_i) = -(X_i - \theta)^2$. However, for non-smooth g_θ these standard asymptotics tend to break down and the rate of convergence often slows to $n^{1/3}$; this is sometimes called the *sharp-edge effect*. A simple example dating back to Chernoff (1964) is the univariate modal interval estimator $\hat{\theta}_n$ that maximizes $P_n[\theta - a, \theta + a]$, the empirical measure of an interval of fixed width. In this instance, $g_\theta(X_i) = 1\{\theta - a \leq X_i \leq \theta + a\}$ with $1\{\cdot\}$ being the indicator function. Another example, and the one that motivated this paper, is the maximum score estimator proposed by Manski (1975, 1989); see Section 3 for details.

Considering the case of non-smooth g_θ, Kim and Pollard (1990) derived a general theorem for a sequence of estimators $\{\hat{\theta}_n\}$ under the following conditions

- $P_n g_{\hat{\theta}_n} \geq \sup_{\theta\in\Theta} P_n g_\theta - o_P(n^{-2/3})$;
- $\hat{\theta}_n$ converges in probability to the unique $\theta(P)$ that maximizes Pg_θ;
- $\theta(P)$ is an interior point of Θ
- Pg_θ is twice differentiable with second derivative matrix $-V$ at $\theta(P)$;
- $H(s,t) = \lim_{\beta\to\infty} \beta P g_{\{\theta(P)+s/\beta\}} g_{\{\theta(P)+t/\beta\}}$ exists for each s, t in $\mathbb{R}^d$;
- a set of further regularity conditions.

Now, let $Z(t)$ be a Gaussian process with continuous sample paths, expected value $-0.5t'Vt$ and covariance kernel H. Theorem 1.1 of Kim and Pollard (1990) states that if V is positive definite and if Z has nondegenerate increments, then $n^{1/3}(\hat{\theta}_n - \theta(P))$ converges in distribution to the (almost surely unique) random vector that maximizes Z. Denote this vector by M.

It is typically impossible to analyze the distribution of M analytically in order to base inferences for $\theta(P)$ on it. Also, given the non-standard asymptotics, it is doubtful that the standard bootstrap would work. In this paper, we will show how the subsampling method can be used to obtain asymptotically valid inferences.

The remainder of the paper is organized as follows. Section 2 describes the general methods. As an example, we consider inference based on Manski's maximum score estimator in Section 3. Finally, Section 4 concludes. The appendix contains some details concerning the computations as well as the simulation results.

2 Subsampling Inference in Cube Root Asymptotics

The subsampling method is designed to provide valid inferences under very weak assumptions. The original paper by Politis and Romano (1994) describes the construction of confidence regions for general parameters, while the methodology of subsampling hypothesis tests was introduced in Politis, Romano, and Wolf (1999; Section 2.6).

2.1 Confidence Regions

Suppose we are interested in a confidence region for $\theta(P)$. Define the sampling distribution

$$J_n(A,P) = Prob_P\{n^{1/3}(\hat{\theta}_n - \theta(P)) \in A\},$$

for any Borel set $A \in \mathbb{R}^d$. The goal is to consistently estimate the limiting value of $J_n(A,P)$, which shall be denoted by $J(A,P)$; hence, $J(P)$ is the law of M, the random vector that maximizes Z. Moving on to a normed statistic will then, with the help of the continuous mapping theorem, allow us to construct a confidence region with asymptotically correct coverage probability. To describe the method, let $Y_1,\ldots,Y_{N_n}$ be equal to the $N_n = \binom{n}{b}$ subsets of $\{X_1,\ldots,X_n\}$ of size $b < n$ ordered in any fashion. Let $\hat{\theta}_{n,b,i}$ be equal to the statistic $\hat{\theta}_b$ evaluated at the data set Y_i. The sampling distribution of $J_n(A,P)$ is then approximated by the subsampling distribution

$$\hat{J}_{n,b}(A) = N_n^{-1}\sum_{i=1}^{N_n} 1\{b^{1/3}(\hat{\theta}_{n,b,i} - \hat{\theta}_n) \in A\}.$$

Theorem 2.1 *Assume $\{\hat{\theta}_n\}$ is a sequence of estimators satisfying the conditions of Theorem 1.1 of Kim and Pollard (1990). Consider the Gaussian process with $Z(t)$ with continuous sample paths, expected value $-0.5t'Vt$ and covariance kernel H. Assume that V is positive definite and that Z has nondegenerate increments. Finally, assume $b/n \to 0$ and $b \to \infty$ as $n \to \infty$. Then,*

(i) $\hat{J}_{n,b}(A) \to J(A,P)$ *in probability, for each Borel set A.*

(ii) $\rho_d(\hat{J}_{n,b}, J(P)) \to 0$ *in probability, for every metric ρ_d that metrizes weak convergence on $\mathbb{R}^d$.*

(iii) For a norm $||\cdot||$ on $\mathbb{R}^d$, define a univariate 'normed' distribution $\hat{J}_{n,b,||\cdot||}$ in the following way:

$$\hat{J}_{n,b,||\cdot||}(x) = N_n^{-1}\sum_{i=1}^{N_n} 1\{b^{1/3}||\hat{\theta}_{n,b,i} - \hat{\theta}_n|| \le x\}.$$

For $\alpha \in (0,1)$, let

$$c_{n,b,||\cdot||}(1-\alpha) = \inf\{x : \hat{J}_{n,b,||\cdot||}(x) \ge 1-\alpha\}.$$

Then,

$$Prob_P\{n^{1/3}||\hat{\theta}_n - \theta(P)|| \leq c_{n,b,||\cdot||}(1-\alpha)\} \to 1-\alpha \text{ as } n \to \infty.$$

Thus, the asymptotic coverage probability under P of the confidence region $\{\theta : n^{1/3}||\theta - \hat{\theta}_n|| \leq c_{n,b,||\cdot||}(1-\alpha)\}$ is the nominal level $1-\alpha$.

Proof: Theorem 1.1 of Kim and Pollard (1990) implies that $J_n(P)$ converges in distribution to M, the (almost surely unique) random vector that maximizes Z. The remainder of the proof now follows easily from Theorem 3.3.1 of Politis, Romano, and Wolf (1999). Note that this theorem deals with α-mixing, dependent data, but i.i.d. variables can be considered a special case of them. ■

Remark 2.1 For all but very small sample sizes, the number N_n of subsets of size b will be too large in order to calculate the exact subsampling distribution $\hat{J}_{n,b}(\cdot)$. In that case one can select a random subset, with or without replacement, $\{I_1, \ldots, I_B\}$ of $\{1, \ldots, N_n\}$ and base the subsampling distribution on the statistics $\hat{\theta}_{n,b,I_1}, \ldots, \hat{\theta}_{n,b,I_B}$ only by considering

$$\tilde{J}_{n,b}(x) = B^{-1} \sum_{i=1}^{B} 1\{b^{1/3}(\hat{\theta}_{n,b,I_i} - \hat{\theta}_n) \in A\}.$$

As long as $B \to \infty$ as $n \to \infty$, the asymptotic validity of the subsampling confidence regions is not affected; this can be seen by arguments very similar to that of Corollary 2.4.1 of Politis, Romano, and Wolf (1999).

Confidence regions are, maybe, most attractive in situations where the parameter of interest is univariate and the region simplifies to an interval, which can be written down easily. An alternative inference method that is equally attractive in multi- and univariate situations are hypothesis tests. Also, they allow us to directly focus on smooth functions of the parameter of interest.

2.2 Hypothesis Tests

We wish to test the null hypothesis H_0: $f(\theta(P)) = 0$ against the alternative hypothesis H_1: $f(\theta(P)) \neq 0$. Here, $f(\cdot)$ is a smooth function from $\mathbb{R}^d$ to $\mathbb{R}^k$ and it is assumed that the Jacobian matrix of $f(\cdot)$ evaluated at $\theta(P)$ is of full rank. The basis for our test will be a statistic $t_n = t_n(X_1, \ldots, X_n)$ that converges to zero in probability under the null and that converges to a positive constant in probability under the alternative. An obvious choice is $t_n = ||f(\hat{\theta}_n)||$, where $||\cdot||$ is any norm on $\mathbb{R}^k$. The final test statistic then becomes $T_n = n^{1/3}t_n$. Define

$$Q_n(x, P) = Prob_P\{T_n \leq x\}.$$

To describe the test construction, let $Y_1, \ldots, Y_{N_n}$ be equal to the $N_n = \binom{n}{b}$ subsets of $\{X_1, \ldots, X_n\}$ of size $b < n$ ordered in any fashion. Let $t_{n,b,i}$ be equal to the statistic t_b

evaluated at the data set Y_i. The sampling distribution of T_n is then approximated by the subsampling distribution

$$\hat{Q}_{n,b}(x) = N_n^{-1} \sum_{i=1}^{N_n} 1\{b^{1/3} t_{n,b,i} \leq x\}.$$

Using this estimated sampling distribution, the critical value for the test is obtained as the $1-\alpha$ quantile of $\hat{Q}_{n,b}$; specifically, define

$$\hat{q}_{n,b}(1-\alpha) = \inf\{x : \hat{Q}_{n,b}(x) \geq \alpha\}.$$

Finally, the nominal level α test rejects H_0 if and only if $T_n > \hat{q}_{n,b}(1-\alpha)$.

The following theorem gives the consistency of this procedure both under the null and under the alternative hypothesis.

Theorem 2.2 *Assume $\{\hat{\theta}_n\}$ is a sequence of estimators satisfying the conditions of Theorem 1.1 of Kim and Pollard (1990). Consider the Gaussian process with $Z(t)$ with continuous sample paths, expected value $-0.5t'Vt$ and covariance kernel H. Assume that V is positive definite and that Z has nondegenerate increments. Finally, assume $b/n \to 0$ and $b \to \infty$ as $n \to \infty$. Then,*

(i) If $f(\theta(P)) = 0$, $Prob_P\{T_n > \hat{q}_{n,b}(1-\alpha)\} \to \alpha$ as $n \to \infty$.

(ii) If $f(\theta(P)) \neq 0$, $Prob_P\{T_n > \hat{q}_{n,b}(1-\alpha)\} \to 1$ as $n \to \infty$.

Proof: Theorem 1.1 of Kim and Pollard (1990) implies that under the null hypothesis t_n converges to zero in probability while under the alternative hypothesis it converges to a positive constant in probability. The proof now follows from Theorem 2.6.1 of Politis, Romano, and Wolf (1999) if we can show that T_n has a nondegenerate, continuous limiting distribution under the null. To see why this is true, denote by M the (almost surely unique) random vector that maximizes Z. Also, let J be the $d \times k$ Jacobian matrix of $f(\cdot)$ evaluated at $\theta(P)$, assumed to be of full rank. By the delta method, $G_n(P)$ then converges in distribution to $||J'M||$. ■

Remark 2.2 For all but very small sample sizes, the number N_n of subsets of size b will be too large in order to calculate the exact subsampling distribution $\hat{G}_{n,b}(\cdot)$. In that case one can select a random subset, with or without replacement, $\{I_1, \ldots, I_B\}$ of $\{1, \ldots, N_n\}$ and base the subsampling distribution on the statistics $t_{n,b,I_1}, \ldots, t_{n,b,I_B}$ only by considering

$$\tilde{G}_{n,b}(x) = B^{-1} \sum_{i=1}^{B} 1\{b^{1/3} t_{n,b,I_i} \leq x\}.$$

As long as $B \to \infty$ as $n \to \infty$, the asymptotic validity of the subsampling hypothesis test is not affected; this can be seen by arguments very similar to that of Corollary 2.4.1 of Politis, Romano, and Wolf (1999).

2.3 Other Rates of Convergence

Generalizing the results of Kim and Pollard (1990), van der Vaart and Wellner (1996; Theorem 3.2.10) presented a theorem for M-estimators with a general rate of convergence r_n. The subsampling methods presented in the two previous subsections can be easily adapted to other rates of convergences by replacing $n^{1/3}$ and $b^{1/3}$ by r_n and r_b, respectively, everywhere. Note that the non-standard examples of van der Vaart and Wellner all exhibit cube root asymptotics, though.

2.4 Choice of b

The main practical problem in applying the subsampling method is the choice of the subsample (or block) size b; the problem is analogous to the choice of the bandwidth in smoothing problems. Unfortunately, the asymptotic requirements $b \to \infty$ and $b/n \to \infty$ as $n \to \infty$ give little guidance when faced with a finite sample. Instead, we propose a calibration algorithm to estimate a 'good' block size in practice. The idea will be illustrated in the context of hypothesis tests but a similar algorithm works for the construction of confidence regions.

Let us assume that the null hypothesis is true. In finite samples, a subsampling hypothesis test will typically not exhibit level exactly equal to α; moreover, the actual rejection probability generally depends on the block size b. Indeed, one can think of the actual level λ of a subsampling test as a function of the block size b, conditional on the underlying probability mechanism P and the nominal level α. The idea is now to adjust the 'input' b in order to obtain the actual level close to the nominal one. Hence, one can consider the block size calibration function $h : b \to \lambda$. If $h(\cdot)$ were known, one could construct an 'optimal' test by finding $\tilde{b}$ that minimizes $|h(b) - \alpha|$ and use $\tilde{b}$ as the block size; note that $|h(b) - \alpha| = 0$ may not always have a solution.

In principle, we could simulate $h(\cdot)$ if P were known by generating data of size n according to P and constructing subsampling hypothesis tests for $\theta(P)$ for a number of different block sizes b. This process is then repeated many times and for a given b one estimates $h(b)$ as the fraction of tests that reject the null. The method we propose is identical except that P is replaced by an estimate $\hat{P}_n$ that is consistent for P at least under the null. The choice of $\hat{P}_n$ should be made on a case-by-case analysis. To reflect the null hypothesis as much as possible, it should ideally satisfy $f(\theta(\hat{P}_n)) = 0$ but this will not always be possible (for example, see Subsection 3.1). However, a sensible choice that is always available is the empirical distribution of the observed data $\{X_1, \ldots, X_n\}$; in this instance one can take $\theta(\hat{P}_n) = \hat{\theta}_n$ and for large n it will be assured that $f(\theta(\hat{P}_n)) \approx 0$ in case the null hypothesis is true.

Algorithm 2.1 (Choice of the Block Size)

1. Fix a selection of reasonable block sizes b between limits b_{low} and b_{up}.
2. Generate K pseudo sequences $X^*_{k,1}, \ldots, X^*_{k,n}$, $k = 1, \ldots, K$ i.i.d. according to $\hat{P}_n$. For each sequence, $k = 1, \ldots, K$, and for each b, construct a subsampling hypothesis test $\mathrm{HT}_{k,b}$ for $H_0 : f(\theta(P)) = f(\theta(\hat{P}_n))$. Let $\mathrm{HT}_{k,b} = 1$ if H_0 is rejected and 0 otherwise.
3. Compute $\hat{h}(b) = K^{-1} \sum_{k=1}^{K} \mathrm{HT}_{k,b}$.
4. Find the value $\tilde{b}$ that minimizes $|\hat{h}(b) - \alpha|$.

Remark 2.3 Algorithm 2.1 is by an order of magnitude more expensive than the computation of the final subsampling hypothesis test once the block size has been determined. While it is advisable to choose the selection of candidate block sizes in Step 2 as fine as possible (ideally, include every integer between b_{low} and b_{up}), this may computationally not be feasible, especially in simulation studies. In those instances, a coarser grid should be employed.

Remark 2.4 The idea of Algorithm 2.1 is to find a good block size in case the null hypothesis is true, that is, to bring the actual level close to the nominal level. No attempt is made to maximize power, since this would require to act as if the alternative hypothesis were true. Since we do not know in practice which hypothesis is true, we stick to the common principle of "honoring the null".

Remark 2.5 A similar algorithm can be employed for the construction of confidence regions by focusing on the confidence level of the region (rather than the significance level of the test). The details are straightforward.

3 The Maximum Score Estimator

There are quite a few examples of M-estimators displaying cube root examples, e.g., the modal interval, the shorth estimator (Kim and Pollard, 1990), monotone densities, and current status (van der Waart and Wellner, 1996; Subsection 3.2.3). The example that motivated this paper is Manski's maximum score estimator in binary response models.

3.1 Background and General Discussion

Consider the binary response model of the form

$$y = \begin{cases} 1 & \text{if } z'\beta(P) + u \geq 0, \\ 0 & \text{otherwise,} \end{cases}$$

where y is a scalar dependent variable, z is a d-dimensional vector of explanatory variables, u is an unobserved error variable, and $\beta(P)$ is a vector of regression parameters. In addition, introduce the notation $F(t|z) = P(u \leq t|Z = z)$.

If the distribution of u conditional on z is known to belong to a parametric family, $\beta(P)$ can be estimated by maximum likelihood, among other techniques; for example, see McFadden (1974) or Amemiya (1985, Chapter 9). One disadvantage of this method, in contrast to ordinary linear least squares regression, is that if the distributional form of $u|z$ is misspecified, the estimation of $\beta(P)$ will always be inconsistent. Another disadvantage is that the estimation is not robust against heteroskedasticity. These facts have lead to a search for estimation methods that do not require specifying the distribution of $u|z$; see Cosslett (1983, 1987), Han (1987), Ichimura (1993), Klein and Spady (1993), Powell et al. (1986) and Stoker (1986), among others. However, all of those estimators are somewhat restrictive in that they require that z and u be independent, or that u be limited to heteroskedasticity of certain limited form, or that the distribution of z be known up to a finite-dimensional parameter; see the discussion in Horowitz (1992). The most general estimator by far is the maximum score estimator of Manski (1975, 1985). Its main assumption is that the median of $u|z$ is equal to zero; and it allows the dispersion of u to depend on z in a much more general way than can be accommodated by any of the previously mentioned papers.

Among many equivalent definitions, the maximum score estimator can be defined as the value $\hat{\beta}_n$ that maximizes the 'score' function

$$P_n g_\beta = \frac{1}{n} \sum_{i=1}^{n} [1\{y_i = 1, z_i'\beta \geq 0\} + 1\{y_i = 0, z_i'\beta < 0\}].$$

Obviously, this fits in the general framework of Section 1 letting $X_i = (z_i', y_i)'$ and $\theta = \beta$; as always, it is maintained that the $\{X_i\}$ are i.i.d. according to some unknown distribution P. Because $\hat{\beta}_n$ is only determined up to scalar multiples, it is usually assumed that it is standardized to have unit length; however, other standardizations are also possible (e.g., Horowitz; 1992). Similarly, the regression equation can be rescaled to ensure that $\beta(P)$ is of unit length as well. The parameter space thus can be identified with the unit sphere in $\mathbb{R}^d$.

Despite the attractive properties of the maximum score estimator, to the best of our knowledge, is has to date not been used for inferential purposes. For many years, its distributional properties remained unknown; for example, see the discussion in Amemiya (1985, pages 345-346). This gap was filled by Kim and Pollard (1990) but the limiting distribution turned out to be untractable and unamiable to the bootstrap.

One solution to this problem was proposed in Horowitz (1992). In this paper, he modified Manski's maximum score estimator by smoothing the score function so that it becomes continuous and differentiable. This smoothed estimator is consistent and, after centering and suitable normalization, is asymptotically normally distributed. Therefore, under the assumptions established in the paper, statistical inference is possible with this estimator if the estimation sample is large enough to make the asymptotic normal approximation accurate. In

addition, the standard asymptotics of this estimator permit an application of the bootstrap which significantly improves finite sample performance (see Horowitz, 1996). The problem with Horowitz's estimator is that smoothness of the score function induces a bias which needs to be eliminated by imposing further smoothness restrictions on $F(t|z)$. Mainly, he needs to assume (for an alternative set of assumptions, see Pollard, 1993) that $F(t|x)$ has uniformly bounded derivative $\dot{F}(t|x)$ with respect to t that satisfies $|\dot{F}(t_1|x) - \dot{F}(t_2|x)| \leq C_1|t_1 - t_2|^\alpha$ for some $0 < \alpha \leq 1$ and some constant C_1. On the other hand, Manski's estimator only requires the condition $|2F(t|z) - 1| \geq C_2|t|$ for $|z| < \delta$ and $|t| < \delta$. The same can be stated for the marginal density of Z, and therefore the smoothed maximum score estimator needs somewhat stronger assumptions than the original maximum score estimator. According to these findings, it is clear that in some situations Manski's (1975, 1985) estimator is preferred to the smoothed one proposed by Horowitz (1992). Furthermore, the simulation studies in Horowitz (1992) indicate that even for samples of size $n = 1,000$ the normal approximation is rather inaccurate (although the bootstrap greatly improves finite sample performance). The application of the subsampling method as outlined in Section 2 allows for the construction of hypothesis tests for $\beta(P)$ or smooth functions of $\beta(P)$ based on the original maximum score estimator. For the exact regularity conditions ensuring cube root asymptotics of the estimator, the reader is referred to Example 6.4 of Kim and Pollard (1990).

3.2 Simulation Study

The goal of this subsection is to shed some light on the small sample performance of the subsampling hypothesis test applied to the maximum score estimator. The simulation design is very similar to that of Horowitz (1992) but we use the more common standardization of β having unit length. The dimension of the predictor variable is $d = 2$ and the regression parameter is $\beta(P) = (\sqrt{1/2}, \sqrt{1/2})'$. The predictor variables z_1 and z_2 have a joint standard normal distribution and we consider three different distributions for the error u:

Distribution L:	$\sqrt{2}\,u \sim$ logistic with median 0 and variance 1;
Distribution $T3$:	$\sqrt{2}\,u \sim$ Student's t distribution with 3 degrees of freedom, normalized to have variance 1;
Distribution H:	$\sqrt{2}\,u = 0.25(1 + 2w^2 + w^4)v$, where $w = z_1 + z_2$ and $v \sim$ logistic with median 0 and variance 1.

The factor $\sqrt{2}$ is due to the rescaling of the Horowitz (1992) design so that the usual unit-length standardization for β is achieved.

The null hypothesis is H_0: $\beta_2(P) = 1/\sqrt{2}$; we consider the two nominal levels $\alpha = 0.05$ and $\alpha = 0.1$; the sample sizes included are $n = 100$ and $n = 200$; all simulations are based on $1,000$ replications per scenario; for further details concerning the computations, see Appendix A.

A few words concerning the automatic choice of the block size are in order. As stated earlier, in our Algorithm 2.1 one would ideally want to sample from a distribution $\hat{P}_n$ that imposes the null hypothesis. In our application, this would require a consistent estimator of the distribution of the error u. Unfortunately, in binary response regressions—in contrast to ordinary least square regressions—such an estimator is not available. We therefore take $\hat{P}_n$ to be the empirical distribution of the observed data $\{X_1, \ldots, X_n\}$ and $\beta_2(\hat{P}_n) = \hat{\beta}_{2,n}$.

The simulation results are presented in Table 1 in Appendix B. As expected, the optimal fixed block size generally depends on the error distribution. The automatic choice of block size is seen to work quite well in general. When comparing our results with those for the smooth maximum score estimator reported in Horowitz (1992), it turns out that they are much better when the inference for the smoothed estimator is based on the normal approximation and comparable when it is based on the bootstrap.

Remark 3.1 A formal proof that the (standard) bootstrap does not work for Manski's maximum score estimator is beyond the scope of this paper. However, we provide some simulation results concerning the performance of a bootstrap test for the null hypothesis H_0: $\beta_2(P) = 1/\sqrt{2}$; we consider the two nominal levels $\alpha = 0.05$ and $\alpha = 0.1$; the sample sizes included are $n = 100$, $n = 200$, and $n = 500$; all simulations are based on $1,000$ replications per scenario and the number of bootstrap samples is 500. The results are presented in Table 2. It is apparent that the level of the bootstrap test tends to zero as the sample size increases.

4 Conclusions

In this paper, we discussed inference for a class of estimators that converge at the non-standard rate of $n^{1/3}$. Such a general class was considered by Kim and Pollard (1990) where the estimators maximize a non-smooth objective function given i.i.d. data and a set of regularity conditions. The limiting distribution is the (almost surely unique) random vector that maximizes a certain Gaussian process and cannot be analyzed analytically. Instead, we showed how to construct confidence regions and hypothesis tests based on the subsampling method that have asymptotically correct confidence and significance level, respectively. The main drawback of our method is its computational burden due to the expensive algorithm to determine the block size to be used in practice; hence, it is a drawback that will diminish over time taking into account the development of fast computers.

As an application, we considered Manski's (1975, 1985) maximum score estimator in binary response regression models. It is the most general estimator available for such models but has so far not been used for inferential purposes because of its difficult limiting distribution. However, our general subsampling hypothesis tests were seen to apply and some simulation studies showed good finite sample performance.

References

Amemiya, T. (1985). *Advanced Econometrics.* Harvard University Press, Cambridge, Massachusetts.

Cosslett, S.R. (1983). Distribution-free maximum likelihood estimator of the binary choice model. *Econometrica* **51**, 765–782.

Cosslett, S.R. (1987). Efficiency bounds for distribution-free estimators of the binary choice and the censored regression models. *Econometrica* **55**, 559–585.

Han, A.K. (1987). Non-parametric analysis of a generalized regression model: the maximum rank correlation estimator. *Journal of Econometrics* **35**, 303–316.

Horowitz, J.L. (1996). Bootstrap critical values for tests based on the smoothed maximum score estimator. Technical Report. Department of Economics. University of Iowa.

Kim, J. and Pollard, D.B. (1990). Cube root asymptotics. *Annals of Statistics* **18**, 191–219.

Chernoff, H. (1964). Estimation of the mode. *Annals of the Institute of Mathematical Statistics* **16**, 31–41.

Horowitz, J.L. (1992). A smoothed maximum score estimator for the binary response model. *Econometrica* **60**, 505–531.

Ichimura, H. (1993). Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *Journal of Econometrics* **58**, 71–120.

Klein, R.L. and Spady, R.H. (1993). An efficient semiparametric estimator for discrete choice models. *Econometrica* **61**, 387–421.

Manski, C.F. (1975). Maximum score estimation of the stochastic utility model of choice. *Journal of Econometrics* **3**, 205–228.

Manski, C.F. (1985). Semiparametric analysis of discrete response: Asymptotic properties of the maximum score estimator. *Journal of Econometrics* **27**, 313–333.

McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. In *Frontiers in Econometrics*, ed. by P. Zarembka. Academic Press, New York, 105–142.

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

Politis, D.N., Romano, J.P., and Wolf, M. (1999). *Subsampling.* Springer, New York.

Pollard, D.B. (1993). The asymptotics of a binary choice model. Technical report, Department of Statistics, Yale University.

Powell, J.L., Stock, J.H., and Stoker, T.M. (1986). Semiparametric estimation of weighted averaged derivatives. Working paper No. 1793-86, Alfred P. Sloan School of Management, MIT, Cambridge, MA.

Stoker, T.M. (1986). Consistent estimation of scaled coefficients. *Econometrica* **54**, 1461–1481.

van der Vaart, A. and Wellner, J. (1996). *Weak Convergence and Empirical Processes.* Springer, New York.

A Details concerning the Computations

A problem with our simulation study is the great computational expense. To start with, the score function $P_n g_\beta$ is nontrivial to maximize because it is non-differentiable and has many local maxima. At least for higher-dimensional β it would seem that the only way to go is a global optimization routine; for a general overview and many software pointers, see the web site `http://solon.cma.univie.ac.at/~neum/glopt.html`. We tried a couple of these routines but being in a low-dimensional problem ($d = 2$), it turned out to be more efficient to employ a grid search. To explain how, first introduce the alternative notation

$$S_n(\beta) = P_n g_\beta$$

and then define the univariate score function

$$S_{n,uni}(\beta_2) = \max\{S_n((\sqrt{1-\beta_2^2}, \beta_2)'), S_n((-\sqrt{1-\beta_2^2}, \beta_2)')\}.$$

We computed $S_{n,uni}(\beta_2)$ over the grid $\{0, 0.01, 0.02, \ldots, 0.99, 1\}$ to find the maximizing $\hat{\beta}_{2,n}$; note that $\hat{\beta}_{1,n}$ can be found easily from there but we did not need it.

The next problem lies in the fact that subsampling in conjunction with the block size Algorithm 2.1 is comparable to a double bootstrap and thus very expensive. To keep the computational burden manageable, we did several things. First, for the subsamples we only used the 'continuous' blocks of size b, that is, $Y_1 = \{X_1, \ldots, X_b\}$ to $Y_{n-b+1} = \{X_{n-b+1}, \ldots, X_n\}$. This actually corresponds to subsampling for time series (Politis, Romano, and Wolf; 1999, Chapter 3) but it also works for independent data. The advantage is that only a total of $n-b+1$ subsamples are used as opposed to B (typically taken to be equal to 1,000) subsamples for the stochastic approximation mentioned in Remark 2.2. To make sure that the simulations were not affected by this shortcut, we compared the two alternatives—time series way vs. stochastic approximation for i.i.d. data—in several scenarios with fixed block sizes and the results were identical up to simulation error.

Second, the number of 'reasonable' block sizes included in Algorithm 2.1 was limited to three and they were selected according to some prior simulations. Obviously, this cannot be done with a real data set and one has to use a finer grid; however, it becomes feasible then because only a single hypothesis test needs to be constructed instead of 1,000 tests for the simulation study.

Third, the number of of bootstrap samples in Algorithm 2.1 was taken to be $K = 200$, while for a real application we would recommend $K = 1,000$.

Still, reducing the computational burden in these ways, the simulations for the sample size $n = 200$ ran nearly a week for each scenario, using stand-alone C++ code on a supercomputer HP-Convex Exemplar SPP S2000. This explains why we did not include sample sizes of the order $n = 500$ or $n = 1,000$ in the study.

B Tables

Table 1: Estimated levels of nominal 5% and 10% subsampling hypothesis tests based on 1,000 replications. Columns 2 to 4 list the results for fixed block sizes while column 5 lists the results for the automatic choice of block size.

Error distribution = L, $n = 100$				
Target	$b = 10$	$b = 20$	$b = 30$	$\tilde{b}$
0.05	0.04	0.05	0.08	0.04
0.10	0.05	0.07	0.14	0.10
Error distribution = $T3$, $n = 100$				
Target	$b = 10$	$b = 20$	$b = 30$	$\tilde{b}$
0.05	0.03	0.05	0.08	0.03
0.10	0.04	0.07	0.13	0.06
Error distribution = H, $n = 100$				
Target	$b = 20$	$b = 30$	$b = 40$	$\tilde{b}$
0.05	0.03	0.04	0.09	0.06
0.10	0.04	0.07	0.13	0.10
Error distribution = L, $n = 200$				
Target	$b = 30$	$b = 45$	$b = 60$	$\tilde{b}$
0.05	0.03	0.05	0.11	0.06
0.10	0.06	0.09	0.16	0.10
Error distribution = $T3$, $n = 200$				
Target	$b = 30$	$b = 45$	$b = 60$	$\tilde{b}$
0.05	0.02	0.05	0.10	0.05
0.10	0.05	0.09	0.15	0.09
Error distribution = H, $n = 200$				
Target	$b = 50$	$b = 65$	$b = 80$	$\tilde{b}$
0.05	0.03	0.05	0.09	0.04
0.10	0.04	0.08	0.13	0.09

Table 2: Estimated levels of nominal 5% and 10% bootstrap hypothesis tests based on 1,000 replications.

Error distribution = L			
Target	$n = 100$	$n = 200$	$n = 500$
0.05	0.03	0.01	0.00
0.10	0.04	0.01	0.00
Error distribution = $T3$			
Target	$n = 100$	$n = 200$	$n = 500$
0.05	0.01	0.00	0.00
0.10	0.02	0.00	0.00
Error distribution = H			
Target	$n = 100$	$n = 200$	$n = 500$
0.05	0.03	0.01	0.00
0.10	0.04	0.01	0.00