A Fast Resample Method
for Parametric and Semiparametric Models

Timothy B. Armstrong\textsuperscript{a} \textsuperscript{1} and Marinho Bertanha\textsuperscript{b} and Han Hong\textsuperscript{c} \textsuperscript{2}

\textsuperscript{a,b,c} Department of Economics, Stanford University, CA, USA.

June 2012

Abstract

We propose a fast resample method that can be used to provide valid inference in nonlinear parametric and semiparametric models. This method does not require recomputation of the second stage estimator during each resample iteration but still provides valid inference under very weak assumptions for a large class of nonlinear models. These models can be highly nonlinear in the parameters that need to be estimated and can also be semiparametric through dependence on a first stage nonparametric functional estimation procedure. The fast resample method directly exploits the score function representations computed on each bootstrap sample, thereby reducing computational time considerably. This method is used to approximate the limit distribution of parametric and semiparametric estimators, possibly simulation based, that admit an asymptotic linear representation. It can also be used for bias reduction and variance estimation. Monte Carlo experiments demonstrate the desirable performance and vast improvement in the numerical speed of the fast bootstrap method.

\textbf{Key words:} Score function, Bootstrap, Subsampling, Nonlinear models.

\textbf{JEL Classification:} C12, C15, C22, C52.

\textsuperscript{1}Corresponding author. Email: tim.b.armstrong@gmail.com.

\textsuperscript{2}We thank Xiaohong Chen particularly for helpful comments, and acknowledge generous support by the National Science Foundation (SES 1024504). Armstrong acknowledges generous support from a fellowship from the endowment in memory of B.F. Haley and E.S. Shaw through the Stanford Institute for Economic Policy Research.
1 Introduction

Bootstrap, jackknife, subsampling and other resampling methods have become in recent years a burgeoning area in both theoretical and applied statistics, and are clearly beginning to impact developments in econometric methodology as well as various applied scientific fields. Their primary asset is to provide powerful statistical tools which are easy to implement and are valid where other more standard tools either fail or are difficult to implement. The main idea of resampling is to evaluate a statistic of interest at simulated data sets that are resampled from the data, and to use these statistics computed from the simulated data sets to build up an estimated sampling distribution. Resampling methods are attractive because they do not require analytic derivation of the limiting distribution and a consistent estimator for it. When the statistic of interest has a pivotal asymptotic distribution, bootstrap also provides a computer-automatic mechanism for improving the precision of asymptotic inference through Edgeworth expansions.

In many nonlinear econometric models, estimation of the parameter is computer-intensive and hence requires a considerable amount of time. This is due to either the complex nonlinearities of the model, or the need to rely on simulations when the model is too difficult to estimate in a direct way, such as in latent factor models, stochastic volatility models and diffusion processes, or in many nonlinear models based on micro-level data levels that require nested computation of a dynamic optimization problem or nested search of equilibrium conditions within each evaluation of the model parameters. The objective of the paper is to offer a fast procedure that delivers valid inference for parameters of interest while maintaining the attractiveness of the resampling methods. The idea is to avoid reestimating the parameter on each simulated data but rather use the asymptotic linear representation of the estimator and evaluate that on each simulated data set.

In addition to nonlinearity and the difficulty of numerical optimization, many semi-parametric estimators in economics also depend on a first step nonparametric estimator of an infinite dimensional function. While an extensive theory is available for demonstrating parametric rate of convergence to a limiting normal distribution (see for example the general results of Chen, Linton, and Van Keilegom (2003), the time series generalization by
Chen, Hahn, and Liao (2011), and the earlier contributions by Newey (1994) and Andrews (1994) in more specific models), practical inference for these models remains difficult, especially when the first stage nonparametric estimator is not orthogonal to the moment conditions used in the second stage to construct the second stage estimator.

For these models, there are several conventional approaches to compute the correct standard errors to take into account the statistical uncertainty introduced by the first stage nonparametric estimation. The first one is to derive the asymptotic distribution of the estimator analytically, and replace the asymptotic variance with a consistent estimate based on the sample data. This is in principle possible by following the pathwise derivative calculation in Newey (1994) and making use of Chen, Linton, and Van Keilegom (2003) to allow for general nonsmooth moment conditions. A second approach is resampling. In particular, Chen, Linton, and Van Keilegom (2003) provide high level conditions for the asymptotic normality of the second stage parametric estimator even for models when the criterion function is not smooth, and for the validity of the bootstrap procedure. Either bootstrap or subsampling will provide a valid inference procedure. A third approach is to make use of an insight in Newey (1994), who shows that the asymptotic variance of the second stage estimators does not depend on how the first stage nonparametric estimation method is implemented. The second stage estimator will have the same asymptotic variance regardless of whether the first stage is estimated using a kernel smoother or a sieve parametric approximation.

If the implementation of the first stage estimator can be modified to be a sieve parametric approximation method, then the (overidentifying) moment conditions that are used in obtaining the estimator can potentially be modified to a set of exactly identifying moment conditions, for both the first stage sieve parameters and the second stage structural parameters of the model. According to Newey (1994), if this is possible, the approximate variance of the second stage estimator can be read off from the lower diagonal of the variance-covariance matrix of the entire generalized method of moment estimator that includes both the first stage and second stage estimators. Computing the overall variance-covariance matrix is straightforward using the conventional sandwich formula for GMM
estimators. In particular, a recent paper by Ackerberg, Chen, and Hahn (2011) provides a formal justification of this procedure.

Unfortunately each of these approaches has its own disadvantages. The pathwise derivative calculation in Newey (1994) is often tedious and prone to errors in the analytic computation. The resulting asymptotic variance estimate can also be complex and is sensitive to coding errors. Resampling methods require recomputing the estimators repeatedly over many bootstrap iterations. Given the nonlinear nature of the method of moment estimators, this might not be computationally feasible. Replacing the first stage kernel smoother with a sieve parametric approach also seems at odds with the implementation of the semiparametric estimator, and might also lead to different point estimates for the second second structural parameters. In addition, implementing a first stage sieve parametric approach appears to be more difficult than implementing the kernel smoother.

The computation intensity of bootstrap is well recognized in nonlinear models, e.g. in Gonçalves and White (2004). For single stage parametric estimators, score function based resampling approaches have been considered by by Davidson and MacKinnon (1999), Andrews (2002) and Kline and Santos (2010). The present paper considers a semiparametric two stage setting in which a score function resampling approach provides a solution to the difficulties of computing the second stage extreme estimator numerically and calculating the asymptotic standard errors analytically.

The paper is organized as follows. In Section 2 we first outline our framework in the context of two stage semiparametric estimators in which the first stage is possibly non-parametric and the second stage is parametric. Then we describe how fast resampling works and how it can be used in conducting inference on the estimator. We allow for a broad category of direct estimators and simulation based indirect estimators admitting an asymptotically linear representation. The estimator can be either one stage or multi-stage, and the first stage can be parametric or nonparametric. Section 3 provides the formal results of the consistency of the fast resampling procedure, both bootstrap and subsampling cases. A brief discussion follows on how to extend the fast resampling procedure to multi-step estimation problems, and some primitive conditions that imply
stochastic equicontinuity when data is not iid. In section 4 we demonstrate through a Monte Carlo experiment that fast bootstrap achieves desirable performance and improves computational speed when compared with standard bootstrap. The proofs are found in the Appendix.

2 Framework and fast resampling methods

Consider an estimator \( \hat{\theta} \) of a parameter \( \theta_0 \in \Theta \in \mathbb{R}^d \) formed from a sample \( X = (X_1, \ldots, X_n) \) (iid or dependent data). The estimator \( \hat{\theta} \) can potentially depend on an initial estimate \( \hat{h} \) of a nuisance parameter \( h_0 \in H \) which can be either finite dimensional or infinite dimensional. Often \( \hat{\theta} \) is obtained by equating to approximately zero a set of moment conditions of dimension \( k \)

\[
\hat{g}_n (\theta, \hat{h}) \equiv \frac{1}{n} \sum_{i=1}^{n} g(X_i, \theta, \hat{h})
\]

such that

\[
g(\theta, h_0) \equiv E g(X_i, \theta, h_0) = 0 \quad \text{if and only if} \quad \theta = \theta_0.
\]

In the rest of the paper we will focus on this GMM setup. It can be modified with minor changes to the M-estimator framework. When the number of moment conditions 'k' in \( g(X_i, \theta, h) \) is greater than the number of parameters 'd', the GMM estimator \( \hat{\theta} \) is often defined as the minimizer of a quadratic objective function

\[
\hat{\theta} = \arg \min_{\theta \in \Theta} g_n \left( \theta, \hat{h} \right)' W g_n \left( \theta, \hat{h} \right) = \arg \min_{\theta \in \Theta} \| g_n \left( \theta, \hat{h} \right) \|.
\]

where \( W \) is a \( k \times k \) positive definite weighting matrix. Restricting attention to GMM estimators does not involve a loss of generality, since most two stage estimators can be written in this form. Alternatively, the GMM estimator can also be expressed in a Z-estimator form:

\[
\hat{A}_n \left( \hat{\theta} \right) \hat{g}_n \left( \hat{\theta}, \hat{h} \right) = o_p \left( \frac{1}{\sqrt{n}} \right),
\]
where $\hat{A}_n(\theta)$ is a matrix of linear combinations that can depend on the data and the parameter. For example, in (1), $\hat{A}_n(\theta) = \frac{\partial}{\partial \theta} \hat{g}_n(\theta, \hat{h})' \hat{W}$.

Under mild conditions, $\hat{\theta}$ usually has the following influence function representation that depends on $\hat{h}$:

$$\sqrt{n}(\hat{\theta} - \theta_0) = - (\Gamma_1' W T_1)^{-1} \Gamma_1' W \sqrt{n} g_n \left( \theta_0, \hat{h} \right) + o_p(1) \quad (3)$$

Here, $\Gamma_1 = \frac{\partial}{\partial \theta} E g(X, \theta, h_0) \bigg|_{\theta = \theta_0}$. Hence asymptotic normality of $\sqrt{n}(\hat{\theta} - \theta_0)$ depends on the validity of the following condition:

$$\sqrt{n} g_n \left( \theta_0, \hat{h} \right) \overset{d}{\rightarrow} N(0, V).$$

Under suitable conditions, one can separate the dependence of $g(\cdot)$ on the initial estimate $\hat{h}$ and obtain an asymptotic linear influence function representation of the following form:

$$\sqrt{n} g_n \left( \theta_0, \hat{h} \right) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[ g(X_i, \theta_0, h_0) + \psi(X_i) \right] + o_p(1) \quad (4)$$

for some function $\psi(X_i)$ that represents the impact of replacing $h_0$ with $\hat{h}$ on the second stage estimator. Assuming that the sample mean in equation 4 has a finite second moment, the normalized estimator $\sqrt{n} \left( \hat{\theta} - \theta_0 \right)$ will converge to a normal distribution with variance matrix $(\Gamma_1' W T_1)^{-1} \Gamma_1' W V W T_1 (\Gamma_1' W T_1)^{-1}$ where

$$V = E \left[ g(X, \theta_0, h_0) + \psi(X) \right] \left[ g(X, \theta_0, h_0) + \psi(X) \right]' .$$

One approach to inference is to estimate this asymptotic distribution using $\hat{W}$ and estimates $\hat{\Gamma}$ and $\hat{V}$ of $\Gamma$ and $V$. Although it is usually possible to obtain consistent estimates of these matrices, in particular $\hat{\Gamma}$ and $\hat{W}$, estimating $V$ can be burdensome and often times requires substantial and difficult analytic calculations.

One alternative is to bootstrap. The multinomial nonparametric bootstrap works well in cross section i.i.d data sets. When the data generating processes are different, other resampling methods can be used instead. For example, with stationary time series observations, a block bootstrap method can be used. It is convenient to now define two
ways to bootstrap data: one for the case the data is iid and another one for when the data comes from a stationary dependent process.

**Definition 1** (Multinomial Bootstrap). In case the observed sample \( X = (X_1, \ldots, X_n) \) is drawn iid from some distribution \( F \), a multinomial bootstrap procedure generates a new sample conditional on the observed one by drawing \( X^* = (X^*_1, \ldots, X^*_n) \) iid from the empirical distribution \( F_n(x) = \frac{1}{n} \sum_{i=1}^n I[X_i \leq x] \).

**Definition 2** (Moving Block Bootstrap). In case the observed sample \( X = (X_1, \ldots, X_n) \) is drawn from a stationary dependent process, a moving block bootstrap procedure (MBB) is defined as follows. Let \( X_{n+i} = X_i \) and the \( t \)-th block of the data with \( b < n \) elements to be \( B_{t,b} = \{X_t, \ldots, X_{t+b-1}\} \) for \( t = 1, \ldots, n \). Let \( k = \lfloor \frac{n}{b} \rfloor \), and \( I_1, \ldots, I_k \) be a sequence of iid Uniform\( \{1, \ldots, n\} \). The MBB sample \( X^* \) will be all the \( X_i \)'s in \( \{B_{I_1,b}, B_{I_2,b}, \ldots, B_{I_k,b}\} \). That is, \( X^* = (X^*_1, \ldots, X^*_l) \), where \( X^*_1 = X_{I_1}; \ldots; X^*_b = X_{I_1+b-1}; \ldots; X^*_b = X_{I_2}; \ldots; X^*_l = X_{I_k+b-1} \) and \( l = kb \leq n \). In other words, the \( k \) random numbers \( I_j \) choose \( k \) out of \( n \) blocks of size \( b \).

Chen, Linton, and Van Keilegom (2003) provide conditions under which the distribution of \( \hat{\theta}^* \) estimated using the multinomial bootstrap procedure provides a consistent estimate of the asymptotic variance of \( \hat{\theta} \) in the iid case, which are generalized to allow for dependent data in Chen, Hahn, and Liao (2011). In other words, after generating \( X^* \) from \( X \) according to definition 1 above, \( \hat{\theta}^* \) is obtained by minimizing the expression

\[
\left[ \sum_{i=1}^n g(X^*_i, \hat{\theta}, \hat{h}^*) - g(X_i, \hat{\theta}, \hat{h}) \right] \left[ \sum_{i=1}^n g(X^*_i, \hat{\theta}, \hat{h}^*) - g(X_i, \hat{\theta}, \hat{h}) \right]^T \quad (5)
\]

over \( \theta \), where \( \hat{h}^* \) is an estimate of \( h_0 \) formed using the bootstrapped sample \( X^* \). The bootstrap estimate of the distribution of \( \sqrt{n}(\hat{\theta} - \theta_0) \) is the distribution of \( \sqrt{n}(\hat{\theta}^* - \hat{\theta}) \) conditional on the data \( X \).

Drawing repeated simulations from the bootstrap distribution requires repeatedly solving a minimization problem. This may be difficult computationally. We propose a bootstrap procedure that avoids this minimization problem, but also does not require computing an estimate of \( V \). Rather than minimizing the expression in (5), we base our bootstrap procedure on the influence function representation in (3).
Using either multinomial or MBB bootstrap draws $X^*$, and an estimate $\hat{h}^*$ of $h_0$ based on these draws, our bootstrap procedure estimates the distribution of $\sqrt{n}(\hat{\theta} - \theta_0)$ using the distribution of

$$
\hat{\eta}^* = -\left(\hat{\Gamma}_1^* \hat{W} \hat{\Gamma}_1^*\right)^{-1} \hat{\Gamma}_1^* \hat{W} \sqrt{n} \left( g_n^* \left( \hat{\theta}, \hat{h}^* \right) - g_n \left( \hat{\theta}, \hat{h} \right) \right)
$$

(6)

can be replaced by

$$
\bar{\eta}^* = -\left(\hat{\Gamma}_1^* \hat{W} \hat{\Gamma}_1^*\right)^{-1} \hat{\Gamma}_1^* \hat{W} \sqrt{n} \left( g_n^* \left( \hat{\theta}, \hat{h}^* \right) + g_n \left( \hat{\theta}, \hat{h}^* \right) - 2 g_n \left( \hat{\theta}, \hat{h} \right) \right)
$$

(8)


Both (6) and (8) will be shown in the next subsection to be consistent bootstrap methods under certain conditions.

When the size of the data set is sufficiently large, the subsampling method of Politis, Romano, and Wolf (1999) can also be used in place of the bootstrap. Just like for the bootstrap, the algorithm for subsampling depends on whether the data is i.i.d or serially dependent.
Definition 3 (Stationary Subsample). In case the observed sample $X = (X_1, \ldots, X_n)$ is drawn from a stationary dependent process, subsamples are chosen to be blocks of size $b$ of consecutive observations, the first one being $\{X_1, \ldots, X_b\}$, and the last one $\{X_{n-b+1}, \ldots, X_n\}$. This gives $q = n - b + 1$ blocks. For each block $t = 1, \ldots, q$, let $\hat{h}_t$ be the first stage estimate computed from this subsample. Define

$$g_t(\theta, h) = \frac{1}{b} \sum_{l=t}^{t+b-1} g(X_l, \theta, h)$$

$$\bar{\eta}_t = - \left( \Gamma' \tilde{W} \Gamma \right)^{-1} \Gamma' \tilde{W} \sqrt{b} \left( g_t(\hat{\theta}, \tilde{h}_t) - g_n(\hat{\theta}, \hat{h}) \right)$$

$$L_n^D(x) = \frac{1}{q} \sum_{t=1}^{q} I[\bar{\eta}_t \leq x]$$

Under suitable regularity conditions, the empirical distribution of $\bar{\eta}_t, t = 1, \ldots, q$, defined above to be $L_n^D(x)$, will provide a consistent estimate of the asymptotic distribution of $\sqrt{n} \left( \hat{\theta} - \theta_0 \right)$. This will be formally demonstrated by showing that $L_n^D(x)$ converges in probability to the asymptotic distribution $J(x)$ of $\sqrt{n} \left( \hat{\theta} - \theta_0 \right)$.

When the data is i.i.d, blocks do not have to be limited to consecutive ones and can be drawn from any $b$ out of the total of $n$ observations. The subsample algorithm is modified for i.i.d. data as follows.

Definition 4 (IID Subsample). In case the observed sample $X = (X_1, \ldots, X_n)$ is iid, the set $S$ of all subsamples of size $b$ contains all $(\binom{n}{b})$ combinations of observations in $X$. For each subsample $S \in S$, let $\tilde{h}_S$ be the first stage parametric estimate computed from this subsample. Define

$$g_S(\theta, h) = \frac{1}{b} \sum_{i \in S} g(X_i, \theta, h)$$

$$\bar{\eta}_S = - \left( \Gamma' \tilde{W} \Gamma \right)^{-1} \Gamma' \tilde{W} \sqrt{b} g_S(\hat{\theta}, \tilde{h}_S)$$

$$L_n^I(x) = \frac{1}{\binom{n}{b}} \sum_{S \in S} I[\bar{\eta}_S \leq x]$$
Again, the consistency of this subsampling procedure is demonstrated by showing that $L_n^I(x)$ converges in probability to $J(x)$, the limiting distribution of $\sqrt{n}(\hat{\theta} - \theta_0)$.

Each of the resampling methods described above requires a consistent estimate $\hat{\Gamma}_1$ of $\Gamma_1$. For example, when the moment function is differentiable with respect to $\theta$, one can use

$$\hat{\Gamma}_1 = \frac{\partial}{\partial \theta} \hat{g}_n(\hat{\theta}, \hat{h}).$$

When $\hat{g}_n(\hat{\theta}, \hat{h})$ is not differentiable in $\theta$, numerical derivative can be used to compute $\hat{\Gamma}_1$ and is consistent under weak conditions on the step size parameters (e.g. Hong, Mahajan, and Nekipelov (2010)). Alternatively, the posterior variance from running $\hat{g}_n(\theta, \hat{h})$ through a monte carlo markov chain will provide a consistent estimate of $(\Gamma_1'W\Gamma_1)^{-1}$.

When the moment conditions are exactly identifying as in M-estimator settings, $W$ is irrelevant and this posterior variance will estimate $\Gamma_1^{-1}$ consistently. When the moment conditions are overidentifying, it is possible to vary the choice of $W$ to generate different sequences of markov chains. The posterior variances of the markov chains can generate a sufficient number of equations to recover the elements of $\Gamma_1$.

In each of the methods discussed above, one can obviously replace $\hat{\Gamma}_1$ and $\hat{W}$ by their bootstrap analogs $\hat{\Gamma}_1^*$ and $\hat{W}^*$. The results that will be stated below will remain the same for these alternative estimates of $\Gamma_1$ and $W$. It appears a priori though that it should be more precise to use $\hat{\Gamma}_1$ and $\hat{W}$.

Generalized empirical likelihood methods are also becoming increasingly popular for estimating GMM models. These methods look for a saddle point in the space of $\theta, \lambda$ of a function $\frac{1}{n} \sum_{i=1}^{n} \rho \left( \lambda' g \left( X_i, \theta, \hat{h} \right) \right)$ where the scalar function $\rho(\cdot)$ is convex, finite and three times differentiable. For each $\theta$, computing $\hat{\lambda}(\theta)$ is often not difficult, but optimizing over $\theta$ can be intensive. For each model, the first order conditions of the concentrated generalized empirical likelihood function $\frac{1}{n} \sum_{i=1}^{n} \rho \left( \hat{\lambda}(\theta)' g \left( X_i, \theta, \hat{h} \right) \right)$ can be treated as a set of exactly identifying moment conditions for $\theta$, and the fast resampling methods described above can be applied to this new set of moment conditions.
3 Consistency of the fast resample algorithm

In this section, we show that our modified resampling procedure is consistent under similar conditions to those used to show consistency of the procedure based on the centered bootstrap objective function (5) in Chen, Linton, and Van Keilegom (2003), following a similar proof strategy. We first present a consistency theorem for the bootstrap procedure and then a subsampling consistency theorem.

3.1 Assumptions and Asymptotic distribution

In order to show consistency of the proposed procedures, we first state the assumptions for asymptotic normality, which were addressed by Chen, Linton, and Van Keilegom (2003) for this general GMM framework. This will help us build our results in the next subsections.

In what follows, whenever we write a parameter space with a δ-subscript, we’re dealing with an open δ-ball around the true parameter value; e.g. \( \Theta_\delta = \{ \theta \in \Theta : \| \theta - \theta_0 \| \leq \delta \} \) and \( \mathcal{H}_\delta = \{ h \in \mathcal{H} : \| h - h_0 \| \leq \delta \} \). Wherever unambiguous, we abuse the norm notation and use \( \| \cdot \| \) to denote norms on the Euclidean and \( \mathcal{H} \) spaces, besides the already defined quadratic form GMM objective function in (1). It is also helpful to define a pathwise derivative of \( g(\theta, h) \) wrt \( h \). We say \( g(\theta, h) \) is differentiable at \( h \in \mathcal{H} \) in the direction \( [\hat{h} - h] \in \mathcal{H} \) if the limit \( \Gamma_2(\theta, h)[\hat{h} - h] = \lim_{\tau \to 0} \frac{[g(\theta, h + \tau(\hat{h} - h)) - g(\theta, h)]}{\tau} \) exists.

**Assumption 1.** The estimator \( \hat{\theta} \) is a function of the observed data \( X \) such that \( \hat{\theta} \xrightarrow{p} \theta_0 \) and

\[
\| g_n(\hat{\theta}, \hat{h}) \| = \inf_{\theta \in \Theta_\delta} \| g_n(\theta, \hat{h}) \| + o_p(n^{-1/2})
\]

**Assumption 2.**

(i) For some \( \delta > 0 \), \( \Gamma_1(\theta, h_0) = \frac{\partial}{\partial h} g(\theta, h_0) \) exists for \( \forall \theta \in \Theta_\delta \) and is continuous at \( \theta = \theta_0 \)

(ii) \( \Gamma_1 = \Gamma_1(\theta_0, h_0) \) has full column rank.

**Assumption 3.** For some \( \delta > 0 \), and for \( \forall \theta \in \Theta_\delta \), the pathwise derivative \( \Gamma_2(\theta, h_0)[h - h_0] \) exists in all directions \( [h - h_0] \in \mathcal{H} \). Also, for all positive sequences \( \delta_n = o(1) \), and \( \forall (\theta, h) \in \Theta_{\delta_n} \times \mathcal{H}_{\delta_n} \),
Assumption 4. \( P\{\hat{h} \in \mathcal{H}\} \to 1 \) and \( \|\hat{h} - h_0\| = o_p(n^{-1/4}) \).

Assumption 5. For all positive sequences \( \delta_n = o(1) \),
\[
\sup_{\theta \in \Theta_{n}, \ h \in \mathcal{H}} \delta_n \| g_n(\theta, h) - g(\theta, h) - g_n(\theta_0, h_0) \| = o_p(n^{-1/2})
\]

Assumption 6. For a finite matrix \( V \),
\[
(i) \ \sqrt{n} \{ g_n(\theta_0, h_0) + \Gamma_2(\theta_0, h_0)[\hat{h} - h_0] \} \overset{d}{\to} N(0; V)
\]
\[
(ii) \ \Gamma_2(\theta_0, h_0)[\hat{h} - h_0] = O_p(n^{-1/2})
\]

The following theorem is proven in Chen, Linton, and Van Keilegom (2003).

Theorem 1. Under assumptions 1 to 6,
\[
\sqrt{n}(\hat{\theta} - \theta_0) \overset{d}{\to} N(0, (\Gamma'_1 W \Gamma_1)^{-1} \Gamma'_1 W V W \Gamma_1 (\Gamma'_1 W \Gamma_1)^{-1}).
\]

3.2 Consistency of fast bootstrap

In this section, we use \( \overset{d}{\to} \) to denote conditional weak convergence in probability, that is, conditional on the observed sample \( X \), the bootstrapped statistic converges in distribution.

Definition 5. Let \( T^*_n \) be some bootstrapped statistic assuming values in \( \mathbb{R}^d \), for some finite \( d \). We say that \( T^*_n \overset{p}{\to} X \iff d^*_BL (T^*_n, X) \overset{p}{\to} 0 \) where
\[
d^*_BL (T^*_n, X) = \sup_{H \in BL} |E[H(T^*_n)|X] - EH(X)|
\]
\[
BL = \left\{ H : \mathbb{R}^d \to \mathbb{R} \ st \ ||H||_\infty \leq 1, ||H||_{Lip} \leq 1 \right\}
\]

In order to show consistency of the fast bootstrap procedure, we need to make the following additional assumptions.
Assumption 7. For some $\delta > 0$, and for $\forall (\theta, h) \in \Theta_{\delta} \times H_{\delta}$, the pathwise derivative $\Gamma_2(\theta, h)[h' - h]$ exists in all directions $[h' - h] \in H$. Also, for all positive sequences $\delta_n = o(1)$, and $\forall(\theta, h) \in \Theta_{\delta_n} \times H_{\delta_n}$ and $h' \in H_{\delta_n}$,
\[
\exists c \geq 0 : \|g(\theta, h) - g(\theta, h') - \Gamma_2(\theta, h)[h - h']\| \leq c\|h - h'\|^2
\]

Assumption 8. $\hat{\Gamma}_1$ and $\hat{W}$ are functions of the observed data $X$, and $\hat{h}^*$, of the bootstrapped data $X^*$, such that: $\hat{\Gamma}_1 \overset{p}{\rightarrow} \Gamma_1$, $\hat{W} \overset{p}{\rightarrow} W$, and $n^{1/4}\|\hat{h}^* - h_0\| \overset{p}{\rightarrow} 0$.

Assumption 9. For the same $V$ as in assumption (6),
\[
\sqrt{n}\{g_n^*(\theta, \hat{h}) - g_n(\hat{\theta}, \hat{h}) + \Gamma_2(\hat{\theta}, \hat{h})[\hat{h}^* - \hat{h}]\} \overset{p}{\rightarrow} N(0; V)
\]

Assumption 10. For all positive sequences $\delta_n = o(1)$,
\[
\sup_{\theta \in \Theta_{\delta_n}, h \in H_{\delta_n}} n^{1/2}\|g_n^*(\theta, h) - g_n^*(\theta_0, h_0) + g_n(\theta_0, h_0) - g_n(\theta, h)\| \overset{p}{\rightarrow} 0
\]

Theorem 2. Under assumptions 1 to 9
\[
\hat{\eta}^* \overset{p}{\rightarrow} N\left(0, (\Gamma_1 W^{T} \Gamma_1)^{-1} \Gamma_1 W V W^{T} \Gamma_1 (\Gamma_1 W^{T} \Gamma_1)^{-1}\right).
\]

Under assumptions 1 to 10
\[
\hat{\eta}^* \overset{p}{\rightarrow} N\left(0, (\Gamma_1 W^{T} \Gamma_1)^{-1} \Gamma_1 W V W^{T} \Gamma_1 (\Gamma_1 W^{T} \Gamma_1)^{-1}\right).
\]

where $\hat{\eta}^*$ and $\bar{\eta}^*$ are given by equations (6) and (8), respectively.

Proof. In the appendix. \qed

The same idea can be used to construct an estimate of the optimal weighting matrix in a GMM procedure for a $\chi^2$ overidentification test. The optimal weighting matrix, which is the inverse of the variance of the moment condition $\sqrt{n}g_n\left(\theta_0, \hat{h}\right)$, can be estimated from the bootstrap sample of $\sqrt{n}\left(g_n^*(\hat{\theta}, \hat{h}^*) - g_n(\hat{\theta}, \hat{h})\right)$. 13
3.3 Consistency of fast subsampling

This section shows that the two step fast subsampling algorithm is consistent under very weak conditions similar to those in Politis, Romano, and Wolf (1999). Because the two step structure involves a nonlinear moment condition for each subsample where the complete sample parameter estimates enter nonlinearly, which differs from the original setup in Politis, Romano, and Wolf (1999) that depends only on the parameter estimates, we do require an additional empirical process stochastic equicontinuity assumption.

Most of assumptions in the bootstrap section apply to the subsample case with the corresponding interpretations. As in the bootstrap case, we require assumption 5 to hold for stationary serially dependent data under a mixing condition. In the i.i.d case assumption 5 will be applicable for each possibly noncontiguous subsampling block. For mixing data this assumption will be applied to each contiguous sample block.

**Assumption 11.** One of the following two conditions holds: (1) The data is iid and the subsample distribution is constructed using all \(b\) out of \(n\) blocks like in definition 4; (2) The data is stationary and strong mixing, such that the mixing coefficients satisfy \(\alpha_X(m) \rightarrow 0\) as \(m \rightarrow \infty\), and the subsample distribution is constructed using contiguous blocks like in definition 3.

**Theorem 3.** Under assumptions 1 to 6 and 11, and definitions (3) and (4), if \(b \rightarrow \infty\) and \(b/n \rightarrow 0\), then at every continuity point \(x\) of \(J(x)\), both \(L_{nD}^D(x) \xrightarrow{p} J(x)\) and \(L_{nI}^I(x) \xrightarrow{p} J(x)\).

3.4 Multi-stage estimation methods

The fast bootstrap method for two step estimators can be easily generalized to multi-stage estimators, at least when all the stages are parametric except the first step estimator. This only requires using the bootstrap distribution of estimator from each previous stage to form the distribution of the score functions of the estimator for each step, without the need to recompute each of the \(m\)th stage estimator during the bootstrap procedure except for the initial stage if there exists one.
To be more specific, consider a $M$ step estimator where the parameter estimated at stage $m$ is denoted $\theta_m, m = 1, \ldots, M$. The initial stage $\theta_1$ can also depend on an estimated nonparametric function $\theta_0 \equiv h$. There need no be an initial nonparametric stage either, in which case the estimator is entirely parametric. Suppose each of the $m = 1, \ldots, M$ stages is estimated by a GMM objective function (or an estimation method that can be equivalently formulated as a GMM method such as a maximum likelihood estimator) as previously, except that each stage of GMM method is indexed by $m$:

$$
\hat{\theta}^m = \arg \min_{\theta_m \in \Theta} \tilde{g}^m_\theta \left( \theta_m, \hat{\theta}_{m-1} \right) \tilde{W}^m \hat{g}^m_\theta \left( \theta_m, \hat{\theta}_{m-1} \right).
$$

(9)

For $m = 1$, follow section 3.2 to estimate the distribution of $\sqrt{n} \left( \hat{\theta} - \theta_1 \right)$ using the bootstrap distribution of 

$$
\hat{\eta}^*_1 = - \left( \hat{\Gamma}^1 \hat{W}_1 \hat{\Gamma}^1 \right)^{-1} \hat{\Gamma}^1 \hat{W}_1 \sqrt{n} \left( g_{n,1}^* \left( \hat{\theta}_1, \hat{h}^* \right) - g_n^1 \left( \hat{\theta}_1, \hat{h} \right) \right)
$$

(10)

Then for each of the subsequent stages $m = 2, \ldots, M$, recursively define

$$
\hat{\eta}^*_m = - \left( \hat{\Gamma}^m \hat{W}_m \hat{\Gamma}^m \right)^{-1} \hat{\Gamma}^m \hat{W}_m \sqrt{n} \left( g_{n,m}^* \left( \hat{\theta}_m, \hat{\theta}_{m-1} + \frac{1}{\sqrt{n}} \hat{\eta}^*_m \right) - g_{n}^m \left( \hat{\theta}_m, \hat{\theta}_{m-1} \right) \right)
$$

(11)

The bootstrap distribution of $\hat{\eta}^*_m$ naturally estimates the distribution of $\sqrt{n} \left( \hat{\theta}_m - \theta_0 \right)$. The validity of this procedure follows immediately from a straightforward extension of the section 3.2. Intuitively, its validity depends crucially on the fact that

$$
\sqrt{n} \left( \hat{\theta}_{m-1}^* - \hat{\theta}_{m-1} \right) = \hat{\eta}_m^* + o_p^*(1).
$$

where $o_p^*(1)$ denotes a random variable that converges in probability to zero in the weak conditional convergence sense. Similarly, the fast subsampling method can also be extended straightforwardly to multi-stage estimation procedures. Consider the steps in section 3.3 as giving the procedure for stage 1: $\hat{\theta}_1$. Then similarly define recursively for $m = 2, \ldots, M$:

$$
\hat{\eta}^m_\delta = - \left( \hat{\Gamma}^m_1 \hat{W}_m \hat{\Gamma}^m_1 \right)^{-1} \hat{\Gamma}^m_1 \hat{W}_m \sqrt{b} \hat{g}^m_\delta \left( \hat{\theta}_m, \hat{\theta}_{m-1} + \frac{1}{\sqrt{b}} \hat{\eta}^m_\delta \right),
$$

and use the subsample distribution of $\hat{\eta}^m_\delta$ to estimate the distribution of $\sqrt{n} \left( \hat{\theta}_m - \theta_m \right)$.
4 Monte Carlo simulations

4.1 Setup

To study the finite sample performance of the score function resampling method we conduct a small monte carlo simulation of a two step model. In the simulation model the first stage and second stage parameters are denoted $\alpha$ and $\beta$ respectively. The dependent variables $s_t$ and $y_t$ are related to the independent variables $x_t$ and $z_t$ through the following relations.

\[
\begin{align*}
  s_t &= I \{ z_t' \alpha_0 + u_t > 0 \} \\
  y_t &= x_t' \beta_0 + h ( z_t' \alpha_0 ) + \epsilon_t
\end{align*}
\]

In the first equation, $u_t \sim N(0, 1)$ and are independent of $z_t$, which allows for $\hat{\alpha}$ to be obtained from a standard probit regression. In the second equation, $\epsilon_t$ can be correlated with both $x_t$ and $z_t$. But we assume that there exist a set of instrumental variables $w_t$ that are median independent of $\epsilon_t$, in the sense that median \( \{ \epsilon_t | w_t \} = 0 \). More specifically, we simulate the data set using $\alpha_0 = (1, 1)'$, $\beta_0 = 3$, $h ( x ) = \sin ( x )$, and the following data generating mechanism:

\[
\begin{bmatrix}
  z_{1,t} \\
  z_{2,t} \\
  u_t \\
  x_t \\
  \epsilon_t \\
  w_t
\end{bmatrix}
\sim iid N
\begin{bmatrix}
  0 \\
  0.1 \\
  0 \\
  0.1 \\
  0.5 \\
  0
\end{bmatrix}
\begin{bmatrix}
  8 & 0.1 & 0 & 0.1 & 0.3 & 0.5 \\
  0.1 & 9 & 0 & 0.2 & 0.1 & 0 \\
  0 & 0 & 1 & 0 & 0.5 & 0 \\
  0.1 & 0.2 & 0 & 10 & 0.9 & 1 \\
  0.3 & 0.1 & 0.5 & 0.9 & 1 & 0 \\
  0.5 & 0 & 0 & 1 & 0 & 10
\end{bmatrix}.
\]

Given the $\hat{\alpha}$ estimated by a probit regression in the first stage, we use a set of median based moment conditions

\[ g ( \nu_t, \beta, \alpha ) = w_t \left[ 0.5 - I \left( y_t \leq x_t' \beta + h ( z_t' \alpha ) \right) \right] \text{ where } \nu_t = ( s_t, z_t, y_t, x_t, w_t ). \]
to obtain an estimate $\hat{\beta}$ of $\beta_0$. In particular,

$$\hat{\beta} = \arg \max_{\beta \in \mathcal{R}} Q_n(\beta, \hat{\alpha}) \equiv g_n(\beta, \hat{\alpha})^2,$$

where

$$g_n(\beta, \hat{\alpha}) = \frac{1}{n} \sum_{i=1}^{n} g(\nu_t, \beta, \hat{\alpha}).$$

We consider sample sizes in the range of 100, 200, 500, 1000, 2000, 5000 and 1000. For each sample size we draw 10000 samples of simulations. The number bootstrap samples used in each simulation draw is 1000.

For each simulation, we generate $\nu_t$ for the largest sample size (10000). Then, for each sample size inside each simulation, we apply the following procedure for the subsample of corresponding size of generated sample:

1. Estimate $\alpha_0$ consistently by PROBIT using $s_t$ and $z'_t$ in a first stage.

2. Estimate $\beta_0$ and $\Gamma_1$ consistently in a second stage, where

$$\Gamma_1 = \frac{\partial}{\partial \beta} E[g(\nu_t, \beta, \alpha)] \bigg|_{(\beta_0, \alpha_0)} = E \left[w_t f_{\varepsilon|x,w}(0; x_t, w_t) x_t \right].$$

We obtain $\hat{\beta}$ by grid-search minimization of the $Q_n(\beta)$ objective function. The algorithm is initialized with two inputs:

(i) an initial guess for $\hat{\beta}$ that comes from running a simple quantile regression of $y_t - h(z_t'\hat{\alpha})$ on $x_t$ using Koenker’s algorithm;

(ii) an interval that contains $\hat{\beta}$. We use $[-50; 50]$.

Next, we compute residuals $\hat{\varepsilon}_t$ and estimate $\Gamma_1$ by $\hat{\Gamma}_1 = \frac{1}{nh} \sum_{t=1}^{n} x_t k \left( \frac{\hat{\varepsilon}_t}{h} \right) w_t$. We use an 4th order kernel with $h = n^{-0.001}$.

3. Bootstrap the statistic below 1000 times:

$$\hat{\eta}^* = -\hat{\Gamma}_1^{-1} \sqrt{n} \left[ g_n^*(\hat{\beta}, \hat{\alpha}) - g_n(\hat{\beta}, \hat{\alpha}) \right]$$
to approximate the distribution of
\[ \eta = -\Gamma_1^{-1} \sqrt{n} [g_n (\beta_0, \hat{\alpha}) - g (\beta_0, \alpha_0)] \]
where \( g (\beta, \alpha) = E [g (\nu_t, \beta, \alpha)] \) and \( g_n^* (\beta, \alpha) = \frac{1}{n} \sum_{i=1}^{n} g (\nu_t^*, \beta, \alpha) \).

4) compute quantiles \((c_{1-\alpha/2}, c_1 - \alpha/2)\) from bootstrapped distribution of \(\hat{\eta}^*\), and calculate confidence intervals for \(\beta_0\):
\[ [\hat{\beta} - c_{1-\alpha/2} n^{-\frac{1}{2}}; \hat{\beta} - c_2 n^{-\frac{1}{2}}] \]

5) So far we have 1000 C.I. for each sample size. Calculate the proportion of these that contain the true \(\beta_0\).

6) We also compute the classic inference confidence interval derived from the asymptotic distribution of \(\sqrt{n} (\hat{\beta} - \beta_0)\).

We know that \(\sqrt{n} (\hat{\beta} - \beta_0) = -\Gamma_1^{-1} \sqrt{n} [g_n (\beta_0, \hat{\alpha})] \xrightarrow{d} N (0; \Gamma_1^{-2} \Omega)\)
where
\[ \Omega = VAR \left\{ g (m_t, \beta_0, \alpha_0) + \left[ \frac{\partial}{\partial \alpha} g (\beta_0, \alpha) \right] I (\alpha_0)^{-1} \left[ \frac{\partial \ln f (s | z_t; \alpha)}{\partial \alpha} \right] \right\} I (\alpha_0) = -E \left[ \frac{\partial^2 \ln f (s | z_t; \alpha)}{\partial \alpha \partial \alpha'} \right] |_{\alpha_0} \], the information matrix of a PROBIT model.

The estimation of \(\Omega\) combines the following parts:
- \(a_t = g (m_t, \hat{\beta}, \hat{\alpha})\) for \(g (m_t, \beta_0, \alpha_0)\);.
- \(b = \frac{1}{n} \sum_{t=1}^{n} u_t k \left( \frac{h'}{h'} \right) h' (z_t' \hat{\alpha}) z_t'\) for \(\frac{\partial}{\partial \alpha} g (\beta_0, \alpha) |_{\alpha_0}\);.
- \(c = \frac{1}{n} \sum_{t=1}^{n} \hat{\lambda}_t (\hat{z}_t + z_t' \hat{\alpha}) z_t z_t'\) for \(I (\alpha_0)\), where \(\hat{\lambda}_t = q_t \phi (q_t z_t' \hat{\alpha}) / \Phi (q_t z_t' \hat{\alpha})\), \(q_t = 2s_t - 1\), \(\phi \) & \(\Phi\) are the pdf and cdf of a standard normal.
- \(d_t = \hat{\lambda}_t z_t\) for \(\frac{\partial \ln f (s | z_t; \alpha)}{\partial \alpha} \right|_{\alpha_0}\)
- \(\hat{\Omega} = \frac{1}{n} \sum_{t=1}^{n} \{a_t + bcd_t\}^2\)

The asymptotic \((1 - \alpha)\)-level confidence interval for \(\beta_0\) is obtained in the usual way:
\[ [\hat{\beta} - e_{1-\alpha/2} \left( \frac{\hat{\Omega}}{n \Gamma_1^2} \right)^{0.5}; \hat{\beta} + e_{1-\alpha/2} \left( \frac{\hat{\Omega}}{n \Gamma_1^2} \right)^{0.5}] \), where the values of \(e\) is the standard Normal quantile: \(\Phi (e_{1-\alpha/2}) = 1 - \frac{\alpha}{2} \).
4.2 Results

The results from the Monte Carlo simulations are reported in tables 1 to 8 below. Table 1 tabulates the mean, standard deviation and the mean square errors of the sample parameter estimates across the 1000 simulation draws. As the sample increases, both the standard deviations and the mean square errors decrease monotonically to zero, assuring that the sample parameter estimates are consistent. Table 2 reports similar statistics for the properly normalized and centered estimates that admit a normal limiting distribution. As predicted by the asymptotic distribution theory, the standard deviations and mean square errors stabilize for large sample sizes, although $\beta$ appears to be more difficult to estimate than $\alpha_1$ and $\alpha_2$.

Tables 3 and 4 report the performance of estimates formed by averaging over the bootstrap simulations, respectively for the fast bootstrap method and for the conventional bootstrap method. These estimates theoretically should have larger biases than the original sample estimates.

Tables 5 and 6 report the performance of biased corrected estimates using respectively the fast bootstrap method and the conventional bootstrap method. These biased corrected estimates are defined by two times the initial sample estimates minus the averaged bootstrap estimates. We can see that for the first two parameters $\alpha_1$ and $\alpha_2$, the bias corrected estimates reported in table 5 have visibly smaller bias than the estimates reported in table 3. However, the benefit of bias reduction is not significant for $\beta$. On the other hand, bias correction tends to increase variation, while taking the average over the bootstrap sample reduces variation. The net impact of bias reduction on the mean squared error of the estimates are very small. It seems clear however, that the fast bootstrap bias reduced estimates outperform the conventional bootstrap bias reduced estimates.

Table 7 reports the empirical coverage probabilities across simulations of the confidence intervals for $\beta_0$ at three confidence levels, generated according to the fast bootstrap, the conventional bootstrap and the analytic calculation of the asymptotic distribution. Both the fast bootstrap confidence interval and the analytic asymptotic confidence interval outperform the conventional bootstrap method in terms of providing more precise
coverage of the true parameter $\beta_0$. This is an interesting result, particularly given that the conventional bootstrap method is one that is most likely used by empirical researchers.

Finally, table 8 shows the average time in minutes required to run a simulation of this Monte Carlo exercise, for each sample size and both resampling methods. We can see that, even for this simple example, the conventional bootstrap method can be quite time consuming. The fast-bootstrap was about 10 to 20 times faster depending on the sample size.

[Table 1 about here.]

[Table 2 about here.]

[Table 3 about here.]

[Table 4 about here.]
5 Conclusion

We have proposed a fast resampling method. The method directly exploits estimating (score) functions computed on each resampled draw and avoids recomputing the estimators for each of them. Fast resampling is easy to perform, and achieves satisfactory performance while improving considerably numerical speed. These advantages should be of interest for applied researchers using nonlinear and dynamic models to conduct effective inference.

Our analysis also suggests that while analytical or numerical variance formulas, resampling and MCMC can each be used to obtain valid asymptotic inference, using them in combination instead of in isolation can offer more powerful tools for computing standard errors and constructing confidence intervals and test statistics.

References


6 Appendix

6.1 Fast Bootstrap - Proof of Theorem (2)

Proof. First of all, two results will be important for this proof. The first one says that $X_n \xrightarrow{p} X$ if and only if for every subsequence $\{n_k\}$ there is a further subsequence $\{n_{k_l}\}$ such that $X_{n_{k_l}} \xrightarrow{a.s.} X$. For the second result, note that $X_n^{*} \xrightarrow{p} X$ is the same as $d^*_{BL}(X_n^{*}, X) \xrightarrow{p} 0$ by definition (5). By the first result, for every subsequence $\{n_k\}$ there is a further subsequence $\{n_{k_l}\}$ such that $d^*_{BL}(X_{n_{k_l}}^{*}, X) \xrightarrow{a.s.} 0$. In other words, there exists a set $\Omega : P(\Omega) = 1$ for which $\sup_{H \in BL} |E[H(X^*_n) | X|(\omega) - EH(X)| \to 0 \forall \omega \in \Omega$. Given that convergence in $d_{BL}$ is equivalent to convergence in distribution, and that once we fix $\omega$ the conditioning sample $X$ is fixed, we have that this is the same as $X_{n_{k_l}}^{*}(\omega) \xrightarrow{d} X$, where $X_{n_{k_l}}^{*}(\omega)$ is the random variable generated by bootstrapping $X(\omega)$. In summary, our second result tells us that $X_n^{*} \xrightarrow{p} X$ if and only if for every subsequence $\{n_k\}$ there is a further subsequence $\{n_{k_l}\}$ and a set $\Omega : P(\Omega) = 1$ such that $\forall \omega \in \Omega, X_{n_{k_l}}^{*}(\omega) \xrightarrow{d} X$.

We will proof the result for $\hat{\eta}^*$ first. Consider an arbitrary subsequence $\{n_k\}$ of $\hat{\eta}^*$. We know by the first result that there is a further subsequence $\{n_{k_l}\}$ and a set $\Omega$ of probability one such that for an arbitrary $\omega \in \Omega$ everything in the assumptions that was convergence in probability now becomes convergence of constants; and, by the second result, every bootstrap conditional distribution convergence now becomes convergence in distribution. That said, our goal is to show

$$\hat{\eta}_{n_{k_l}}^*(\omega) \xrightarrow{d} N(0, (\Gamma'_1 W \Gamma_1)^{-1} \Gamma'_1 W V \Gamma_1 (\Gamma'_1 W \Gamma_1)^{-1})$$

In what follows, we’ll do everything along this sub-subsequence $\{n_{k_l}\}$ and for a fixed $\omega \in \Omega$, but we will suppress the subscripts for notational ease where unnecessary. Starting from

$$\hat{\eta}^* = - \left( \hat{\Gamma}'_1 \hat{W} \hat{\Gamma}_1 \right)^{-1} \hat{\Gamma}'_1 \hat{W} \sqrt{n} \left[ g'_n \left( \hat{\theta}, \hat{h}^* \right) - g_n \left( \hat{\theta}, \hat{h} \right) \right]$$

we note that $- \left( \hat{\Gamma}'_1 \hat{W} \hat{\Gamma}_1 \right)^{-1} \hat{\Gamma}'_1 \hat{W} \to - (\Gamma'_1 W \Gamma_1)^{-1} \Gamma'_1 W$, so that we only have to worry about convergence in distribution of the $\sqrt{n} \left[ g'_n \left( \hat{\theta}, \hat{h}^* \right) - g_n \left( \hat{\theta}, \hat{h} \right) \right]$ part.

Rewriting it,
\[
\sqrt{n} \left[ g_n^* \left( \hat{\theta}, \hat{h}^* \right) - g_n \left( \hat{\theta}, \hat{h} \right) \right] = \\
\sqrt{n} \left\{ g_n^* \left( \hat{\theta}, \hat{h}^* \right) - g_n \left( \hat{\theta}, \hat{h} \right) \right\} + \Gamma_2 \left( \hat{\theta}, \hat{h} \right) \left[ \hat{h}^* - \hat{h} \right] \] (I)
\[
+ \sqrt{n} \left\{ g \left( \hat{\theta}, \hat{h}^* \right) - g \left( \hat{\theta}, \hat{h} \right) \right\} - \Gamma_2 \left( \hat{\theta}, \hat{h} \right) \left[ \hat{h}^* - \hat{h} \right] \] (II)
\[
+ \sqrt{n} \left\{ g_n^* \left( \hat{\theta}_0, h_0 \right) + g_n \left( \hat{\theta}_0, h_0 \right) - g_n \left( \hat{\theta}_0, \hat{h}^* \right) \right\} \] (III)
\[
- \sqrt{n} \left\{ g_n^* \left( \hat{\theta}, \hat{h} \right) - g_n^* \left( \hat{\theta}_0, h_0 \right) + g_n \left( \hat{\theta}_0, h_0 \right) - g_n \left( \hat{\theta}, \hat{h} \right) \right\} \] (IV)
\[
+ \sqrt{n} \left\{ g_n \left( \hat{\theta}, \hat{h}^* \right) - g_n \left( \hat{\theta}_0, h_0 \right) - g \left( \hat{\theta}, \hat{h}^* \right) \right\} \] (V)
\[
- \sqrt{n} \left\{ g_n \left( \hat{\theta}, \hat{h} \right) - g_n \left( \hat{\theta}_0, h_0 \right) - g \left( \hat{\theta}, \hat{h} \right) \right\} \] (VI)
\]

For Part (I), we know by A10 that \((I) \xrightarrow{p} N(0; V)\).

For the rest, given that \(\| \hat{\theta} - \theta_0 \| \to 0, n^{1/4} \| \hat{h} - h_0 \| \to 0 \) and \(n^{1/4} \| \hat{h}^* - h_0 \| \xrightarrow{P} 0\), we can find a positive sequence \(\delta_{\omega k_i} \to 0\) such that the event
\[
\{ \| \hat{\theta}_{nk_i} (\omega) - \theta_0 \| \leq \delta_{\omega k_i}, n^{1/4} \| \hat{h}_{nk_i} (\omega) - h_0 \| \leq \delta_{\omega k_i}, n^{1/4} \| \hat{h}^*_{nk_i} (\omega) - h_0 \| \leq \delta_{\omega k_i} \}
\]
happens with probability approaching one, where the only source of randomness is the last term. Similarly to the argument used in the proof of Theorem (2), we can do everything that follows as if \(\hat{\theta}_{nk_i} (\omega) \in \Theta_{\delta_{\omega k_i}}, \) and \(\hat{h}_{nk_i} (\omega), \hat{h}^*_{nk_i} (\omega) \in H_{\delta_{\omega k_i}}\).

For part (II), we use A7 to get
\[
(II) \leq c \cdot n^{1/2} \| \hat{h}^* - \hat{h} \|^2
\]
From A8 we know that \(n^{1/4} \| \hat{h}^* - h_0 \| \xrightarrow{P} 0\). Combining this with the triangular inequality and A4 gives \(n^{1/4} \| \hat{h}^* - \hat{h} \| \xrightarrow{P} 0\). Applying this result above with the continuous mapping theorem gives \((II) \xrightarrow{P} 0\).

Using A10 for parts (III) and (IV) and A5 for parts (V) and (VI), we conclude that they all converge in probability to zero. Therefore, putting all parts together and using Slutsky, \(\hat{\eta}_{nk_i} (\omega) \xrightarrow{d} N(0, (\Gamma_1^T W T_1)^{-1} \Gamma_1^T W V W T_1 (\Gamma_1^T W T_1)^{-1})\). Given that the initial subsequence \(\{n_k\}\) and fixed \(\omega\) were arbitrary,
\[
\hat{\eta}^* \xrightarrow{P} N(0, (\Gamma_1^T W T_1)^{-1} \Gamma_1^T W V W T_1 (\Gamma_1^T W T_1)^{-1})
\]
A similar argument applies to \(\bar{\eta}^*\).
\[
\bar{\eta}^* = - \left( \Gamma_1^T W T_1 \right)^{-1} \Gamma_1^T W \sqrt{n} \left[ g_n^* \left( \hat{\theta}, \hat{h} \right) + g_n \left( \hat{\theta}, \hat{h}^* \right) - 2 g_n \left( \hat{\theta}, \hat{h} \right) \right]
\]
Like before we have to focus on the \( \sqrt{n} \left( g_n^* \left( \hat{\theta}, \hat{h} \right) + g_n \left( \hat{\theta}, \hat{h}^* \right) - 2g_n \left( \hat{\theta}, \hat{h} \right) \right) \) part.

Rewrite is as:
\[
\sqrt{n} \left( g_n^* \left( \hat{\theta}, \hat{h} \right) + g_n \left( \hat{\theta}, \hat{h}^* \right) - 2g_n \left( \hat{\theta}, \hat{h} \right) \right) =
\sqrt{n} \left\{ g_n \left( \hat{\theta}, \hat{h} \right) - 2g_n \left( \hat{\theta}, \hat{h} \right) \right\} + \sum_{q \geq 1} \left( g \left( \hat{\theta}, \hat{h}^* \right) - g \left( \hat{\theta}, \hat{h} \right) \right) - \sum_{q \geq 1} \left( g \left( \hat{\theta}, \hat{h}^* \right) - g \left( \hat{\theta}, \hat{h} \right) \right) \right\}
\]
\[
= \left\{ g \left( \hat{\theta}, \hat{h}^* \right) - g \left( \hat{\theta}, \hat{h} \right) \right\} - \left( g \left( \hat{\theta}, \hat{h} \right) - g \left( \hat{\theta}, \hat{h} \right) \right)
\]
\[
+ \sup_{d(h,\theta_0) \leq \delta} \left| \Gamma \right| g \left( \hat{\theta}, \hat{h} \right) - g \left( \hat{\theta}, \hat{h} \right) \right)
\]

Apart from parts (III) and (IV), these are the same parts as in the proof of \( \hat{\gamma}^* \) above.

The same logic applies except that we don’t need to use A10. Hence,
\[
\hat{\gamma}^* \xrightarrow{p} N(0, (\Gamma_1 W \Gamma_1)^{-1} \Gamma_1 W V W \Gamma_1 (\Gamma_1 W \Gamma_1)^{-1}).
\]

\[\square\]

6.2 Fast Subsampling - Proof of Theorem (3)

**Proof.** First note that under assumptions 1 to 6, as shown in the proof of theorem (1), equation (3) holds, and it holds that \( \hat{\Gamma} \xrightarrow{p} \Gamma, \hat{W} \xrightarrow{p} W, \sqrt{n} \hat{g}_n \left( \theta, \hat{h} \right) \xrightarrow{d} N(0, V) \). Under these assumptions it also holds that \( \sup_{d(h,\theta_0) \leq \delta} \left| g \left( \theta, h \right) - g \left( \theta, \theta_0 \right) \right| = O(\delta) \).

Define
\[
L_n \left( x, \theta, M \right) = \frac{1}{q} \sum_{S} f \left[ M \sqrt{b} g_S (\theta, \hat{h}_S) \leq x \right].
\]

For i.i.d data, \( q = n - b + 1 \), and \( S \) ranges over the contiguous blocks \( \{1, \ldots, b\}, \{2, \ldots, b+1\}, \ldots, \{n - q + 1, n\} \). For i.i.d data, \( q = \binom{n}{b} \), \( S \subseteq \{1, \ldots, n\}, |S| = b \). The inequality in \( L_n \left( x, \theta, M \right) \) refers to the product order on \( \mathbb{R}^d \). We want to show that
\[
L_n \left( x, \hat{\theta}, \hat{M} \right) \xrightarrow{p} J(x)
\]
for all continuity points \( x \) of \( J(x) \) where \( \hat{M} = - \left( \hat{\Gamma}_1 \hat{W} \hat{\Gamma}_1 \right)^{-1} \hat{\Gamma}_1 \hat{W} \). The standard argument will show that \( L_n \left( x, \theta_0, M_0 \right) \xrightarrow{p} J(x) \) where \( M_0 = - \left( \Gamma_1 W \Gamma_1 \right)^{-1} \Gamma_1 W \), so it is enough to show that \( \| L_n \left( x, \hat{\theta}, \hat{M} \right) - L_n \left( x, \theta_0, M_0 \right) \| \xrightarrow{p} 0 \).
Let $\delta_n = K/\sqrt{n}$ and let $k_n$ be a sequence converging to zero such that $\|\hat{M} - M_0\| \leq k_n$ with probability approaching 1. With probability that can be made arbitrarily close to one by making $K$ and $n$ large, the quantity of interest is bounded by

$$\sup_{\|\theta - \theta_0\| < \delta_n, \|M - M_0\| < k_n} \|L_n(x, \theta, M) - L_n(x, \theta_0, M_0)\|,$$

so it suffices to show that this converges in probability to zero for every $K$. If $x$ is a continuity point of the limiting distribution $J$, then, by consistency of $L_n(x, \theta_0, M_0)$, $L_n(x + \eta \mu, \theta_0, M_0) - L_n(x, \theta_0, M_0)$ will converge in probability to something that can be made arbitrarily small by making $\eta$ small (here, $\eta$ is a vector of ones). Similarly, $L_n(x, \theta_0, M_0) - L_n(x - \eta \mu, \theta_0, M_0)$ converges in probability to something that can be made small by making $\eta$ small. Thus, it suffices to show that, for any $\eta > 0$, we will have

$$L_n(x - \eta \mu, \theta_0, M_0) - \eta \leq L_n(x, \theta, M) \leq L_n(x + \eta \mu, \theta_0, M_0) + \eta$$

for all $\theta$ and $M$ with $\|\theta - \theta_0\| < \delta_n$ and $\|M - M_0\| < k_n$ with probability approaching 1. This will follow if

$$\left\{ \sup_{\|\theta - \theta_0\| < \delta_n, \|M - M_0\| < k_n} L_n(x, \theta, M) - L_n(x + \eta \mu, \theta_0, M_0) \right\} \lor 0$$

and

$$\left\{ \sup_{\|\theta - \theta_0\| < \delta_n, \|M - M_0\| < k_n} L_n(x + \eta \mu, \theta_0, M_0) - L_n(x, \theta, M) \right\} \lor 0$$

converge in probability to zero. The quantities in the two preceding displays are bounded by

$$\sup_{\|\theta - \theta_0\| < \delta_n, \|M - M_0\| < k_n} \frac{1}{q} \sum_S I \left[ \sqrt{b} \|G_S(\theta, \hat{h}_S) - M_0g_S(\theta_0, \hat{h}_S)\| \geq \eta \right]$$

$$\leq \frac{1}{q} \sum_S I \left[ \sup_{\|\theta - \theta_0\| < \delta_n, \|M - M_0\| < k_n} \sqrt{b} \|G_S(\theta, \hat{h}_S) - M_0g_S(\theta_0, \hat{h}_S)\| \geq \eta \right]$$

$$\leq \frac{1}{q} \sum_S I \left[ \sup_{\|\theta - \theta_0\| < \delta_n, \|M - M_0\| < k_n} \|G_S(\theta, \hat{h}_S) - M_0g_S(\theta_0, \hat{h}_S)\| \geq \eta/2 \right]$$

$$+ \frac{1}{q} \sum_S I \left[ \sup_{\|\theta - \theta_0\| < \delta_n, \|M - M_0\| < k_n} \sqrt{b} \|G(\theta, \hat{h}_S) - M_0g(\theta_0, \hat{h}_S)\| \geq \eta/2 \right]$$

26
where \( G_S(\theta, h) = \sqrt{b}[g_S(\theta, h) - g(\theta, h)] \).

Each of these terms can be shown to converge in probability to zero by showing that the variation components and expectations converge to zero. Each of the two terms is a sample average of terms with same expectations. The expected value of the first term is

\[
P\left( \sup_{\|\theta-\theta_0\|<\delta_n, \|M-M_0\|<k_n} \|MG_S(\theta, \hat{h}_S) - M_0G_S(\theta_0, \hat{h}_S)\| \geq \eta/2 \right)
\]

which goes to zero as long as the limiting process \( MG(\theta, h) \) is stochastically equicontinuous in \((M, \theta, h)\) when this space has the product norm formed from the norms we have been using for \( M \) and \( \theta \). This can be shown to be valid because it can bounded by

\[
M_0 \sup_{\|\theta-\theta_0\|<\delta_n} \|M_0G_S(\theta, \hat{h}_S) - M_0G_S(\theta_0, \hat{h}_S)\| + k_n \sup_{\|\theta-\theta_0\|<\delta_n} G_S(\theta, \hat{h}_S),
\]

and because Assumptions 1 to 6 imply (as stated at the beginning of this proof) that the first term is \( o_p(1) \) and that

\[
\sup_{\|\theta-\theta_0\|<\delta_n} G_S(\theta, \hat{h}_S) = O_p(1).
\]

The variation component of the first term is given by

\[
\frac{1}{q} \sum_{S} I \left( \sup_{\|\theta-\theta_0\|<\delta_n, \|M-M_0\|<k_n} \|MG_S(\theta, \hat{h}_S) - M_0G_S(\theta_0, \hat{h}_S)\| \geq \eta/2 \right) - P\left( \sup_{\|\theta-\theta_0\|<\delta_n, \|M-M_0\|<k_n} \|MG_S(\theta, \hat{h}_S) - M_0G_S(\theta_0, \hat{h}_S)\| \geq \eta/2 \right).
\]

This variation component is \( o_p(1) \) in the i.i.d case by application of an Hoeffding inequality. It is also \( o_p(1) \) in the stationary mixing case because its variance goes to zero as long as \( b \to 0 \) (where \( q = O(n) \)) because of the mixing condition 11. Both are verified in Politis, Romano, and Wolf (1999), and therefore we do not repeat the arguments. The variation component of the second term goes to zero by essentially the same argument.
The expected value of the second term is

\[
P \left( \sup_{\|\theta - \theta_0\| < \delta_n, \|M - M_0\| < k_n} \sqrt{b} \|M g(\theta, \hat{h}_S) - M_0 g(\theta_0, \hat{h}_S)\| \geq \eta/2 \right)
\]

\[
= P \left( \sup_{\|\theta - \theta_0\| < \delta_n, \|M - M_0\| < k_n} \sqrt{b} \|M[g(\theta, \hat{h}_S) - g(\theta_0, \hat{h}_S)] + (M - M_0)g(\theta_0, \hat{h}_S)\| \geq \eta/2 \right)
\]

\[
\leq P \left( \sqrt{b} \|M\| \sup_{\|\theta - \theta_0\| < \delta_n} \|g(\theta, \hat{h}_S) - g(\theta_0, \hat{h}_S)\| + k_n \sqrt{b} \|g(\theta_0, \hat{h}_S)\| \geq \eta/2 \right).
\]

The first term in this event is of order \(O_P(\sqrt{b}\delta_n) = O_P(\sqrt{b}/\sqrt{n})\) by the implication of assumptions 1 to 6 that \(\sup_{d(h, \hat{h}_0) < \epsilon, d(\theta, \theta_0) \leq \delta} |g(\theta, h) - g(\theta_0, h)| = O(\delta)\), and the second term is of order \(\delta_n\) since \(g(\theta_0, \hat{h}_S)\) converges at a \(\sqrt{b}\) rate. Thus, the probability of this event goes to zero if \(b/n \to 0\).

\[\Box\]

### 6.3 Stochastic Equicontinuity

The goal of this section is to briefly discuss some primitive conditions behind stochastic equicontinuity (assumptions 5 and 10) which was assumed in order to obtain our results. Note that the general framework and assumptions allow for both a parametric and non-parametric first stage estimator, and for the special case without the first stage estimation.

In the iid case with parametric first stage, there are some known easy-to-verify primitive conditions on the class of moment condition functions that imply stochastic equicontinuity both for the empirical and bootstrapped moments. Examples of works that analyzed these conditions are: Pakes and Pollard (1989), Chen, Linton, and Van Keilegom (2003), and Kosorok (2008). Generalization to dependent data is developed in Chen, Hahn, and Liao (2011). Lemma 4.3 in Chen (2007) presented a sufficient condition for stochastic equicontinuity when data is beta-mixing. The high level conditions in Chen, Linton, and Van Keilegom (2003) allow for the second stage moment condition to depend on the entire unknown first stage nonparametric function in a general manner. Verification of this condition usually depends on a case by case basis.

When the data is generated by a stationary process that satisfy some mixing conditions, there are some empirical process results that rely on primitive conditions similar to those
mentioned above, and that are easier to verify when the first stage is parametric. We’ll state these conditions and results below.

Some empirical process notation will be necessary. The class of functions $\mathcal{F}$ is defined to be the class of all possible moment condition functions in our GMM framework. That is,

$$\mathcal{F} = \{ f : \mathcal{X} \to \mathbb{R}^k \text{ s.t. } \exists (\theta, h) \in \Theta \times \mathcal{H} \text{ with } f(X) \equiv g(X, \theta, h) \}$$  \hfill (12)$$

where $\mathcal{X}$ is the space where the random variables we sample from assume their values. Following Pollard (1984), we will define below a certain type of class of functions that will be relevant for us.

**Definition 6.** Let $\mathcal{D}$ be a class of subsets of some space $S$. The class $\mathcal{D}$ is called a VC Class if there exists constants $A$ and $v$ such that for every subset $S_0$ of $S$ with $N$ points, the number of sets in $\{ S_0 \cap D : D \in \mathcal{D} \}$ is less or equal than $A \cdot N^v$. That said, a class of functions $\mathcal{F}$ is called a VC-subgraph class if the graphs

\[ \{(x, t) : f(x) \leq t \leq 0 \text{ or } f(x) \geq t \geq 0 \} \]

of all functions $f$ in $\mathcal{F}$ form a VC class of sets.

In particular, we will need to assume the following condition:

**Assumption 12.** Assume $\mathcal{F}$ defined in (12) is a VC-subgraph class of measurable functions and there exists an envelope function $F > 0$ (i.e. $\forall f \in \mathcal{F} : |f(x)| \leq F(x)$), and $E_p F^p < \infty$ for $2 < p < \infty$.

If the $\mathcal{F}$ defined in (12) is a VC-subgraph class of functions, then we will be able to derive the stochastic continuity conditions under some assumptions on the form of the dependence of the data. In order to state this dependence assumption, we need the following definition.

**Definition 7.** Let $\{X_i\}_{i=-\infty}^{\infty}$ be a strictly stationary sequence of random variables, that is, for each $i_1, \ldots, i_n \in \mathbb{Z}$, $n \in \mathbb{N}$, the distribution of the vector $(X_{i_1+k}, \ldots, X_{i_n+k})$ does
not depend on \( k \in \mathbb{Z} \). Define the following sigma-algebras \( A_0 = \sigma (X_i : i \leq 0) \) and \( B_n = \sigma (X_i : i \geq n) \). The \( \beta \)-mixing coefficient of dependence is given:

\[
\beta_n = \beta (A_0, B_n) = \frac{1}{2} \sup \sum_{(i,j) \in I \times J} |P (A_i \cap B_j) - P (A_i) P (B_j)|
\]

where the sup is taken over all finite measurable partitions \( \{A_i\}_{i \in I} \) and \( \{B_i\}_{i \in J} \) of \( A_0 \) and \( B_n \), respectively.

Before we move on to the actual assumptions and results, we need to define the following real-valued mappings on \( F \). \( P (f) = E_P (f) \), where the expectation is taken with respect to the distribution \( P \) of the data; \( P_n (f) = \frac{1}{n} \sum_{i=1}^{n} f (X_i) \); and \( v_n (f) = n^{1/2} (P_n - P) (f) \). In terms of the GMM framework, if \( f (X) \equiv g(X, \theta, h) \), then \( P (f) = g(\theta, h) \), \( P_n (f) = g_n (\theta, h) \), and \( v_n (f) = n^{1/2} (g_n (\theta, h) - g(\theta, h)) \).

Lemma 4.2 in Chen (2007) provides a set of sufficient conditions for stochastic equicontinuity when the data is beta-mixing. We briefly summarize her conditions and refer the reader to details of the original Handbook of Econometrics Chapter in Chen (2007).

**Proposition 1.** (Lemma 4.2, Chen (2007)) If (1) \( \{X_i\}_{i=\infty}^{-\infty} \) is a strictly stationary sequence of random variables with \( \sum_{k=1}^{\infty} k^{\frac{r-2}{2}} \beta_k < \infty \) for some \( r > 2 \); (2) The moment conditions and the first step function only depend on finitely many \( X_i \); (3) Each of the moment conditions is a sum of a Hölder continuous component and a component that is locally uniformly \( L_r (P) \) continuous with respect to both the second stage parameter and the first stage nonparametric function; (4) The parameter space of the second stage parameter is compact, and the parameter space of the first stage nonparametric function has a finite covering entropy, then

\[
\sup_{f \in F_{\delta_n}} |v_n (f) - v_n (f_0)| \overset{P}{\rightarrow} 0
\]

where \( f_0 (X) = g(X, \theta_0, h_0) \), and \( F_{\delta_n} \) stands for an open \( \delta_n \)-ball in \( F \) around \( f_0 \).

The finite covering entropy part of condition (4) can be derived from assumption (12). If we are willing to strengthen the assumption on the \( \beta \)-mixing type of dependence, we can also obtain stochastic equicontinuity on the bootstrapped moments, a result derived
in Radulović (1996). Similarly to before, define the real-valued bootstrap counterparts of the mappings on $\mathcal{F}$: $P_n^*(f) = \frac{1}{n} \sum_{i=1}^{n} f(X_i^*)$; and $v_n^*(f) = n^{1/2}(P_n^* - P_n)(f)$. Here, we shall use the MBB procedure to generate $X^*$, according to definition 2.

**Assumption 13.** Assume $\{X_i\}_{i=-\infty}^{\infty}$ is a strictly stationary sequence of random variables such that $k^{q(p-2)}(\log k)^{\frac{q(p-1)}{p-2}} \beta_k \to 0$ for the $p$ in assumption(12). Further assume that for some $q > p/(p-2)$ such that $k^{q} > k^{q} \beta_k = O(1)$. Also assume that in MBB procedure, the number of blocks $b(n) \to \infty$ and $b(n) = O(n^\rho)$ for some $0 < \rho < \frac{p-2}{2(p-1)}$.

The following proposition is due to Radulović (1996).

**Proposition 2.** Suppose assumptions 12 to 13 hold, and that the bootstrap sample is generated according to the MBB procedure. Then, for any positive sequence $\delta_n = o(1)$, we have both:

$$\sup_{f \in \mathcal{F}} |v_n(f) - v_n(f_0)| \overset{P}{\to} 0$$

and

$$\sup_{f \in \mathcal{F}} |v_n^*(f) - v_n^*(f_0)| \overset{P}{\to} 0.$$
Table 1: Mean, Standard Deviation and MSE for \((\hat{\alpha}_1, \hat{\alpha}_2, \hat{\beta})\)

<table>
<thead>
<tr>
<th>Sample Size</th>
<th>100</th>
<th>200</th>
<th>500</th>
<th>1000</th>
<th>2000</th>
<th>5000</th>
<th>10000</th>
</tr>
</thead>
<tbody>
<tr>
<td>mean</td>
<td>1.2877</td>
<td>1.0735</td>
<td>1.0289</td>
<td>1.0125</td>
<td>1.0067</td>
<td>1.0029</td>
<td>1.0016</td>
</tr>
<tr>
<td>(\alpha_1) std</td>
<td>2.7294</td>
<td>0.2254</td>
<td>0.1170</td>
<td>0.0765</td>
<td>0.0541</td>
<td>0.0341</td>
<td>0.0239</td>
</tr>
<tr>
<td>MSE</td>
<td>7.5322</td>
<td>0.0562</td>
<td>0.0145</td>
<td>0.0060</td>
<td>0.0030</td>
<td>0.0012</td>
<td>0.0006</td>
</tr>
<tr>
<td>(\alpha_2) mean</td>
<td>1.2895</td>
<td>1.0772</td>
<td>1.0311</td>
<td>1.0135</td>
<td>1.0076</td>
<td>1.0027</td>
<td>1.0015</td>
</tr>
<tr>
<td>std</td>
<td>2.7242</td>
<td>0.2277</td>
<td>0.1203</td>
<td>0.0783</td>
<td>0.0543</td>
<td>0.0339</td>
<td>0.0238</td>
</tr>
<tr>
<td>MSE</td>
<td>7.5051</td>
<td>0.0578</td>
<td>0.0154</td>
<td>0.0063</td>
<td>0.0030</td>
<td>0.0012</td>
<td>0.0006</td>
</tr>
<tr>
<td>(\beta) mean</td>
<td>2.9645</td>
<td>3.1052</td>
<td>3.0031</td>
<td>2.9791</td>
<td>2.9927</td>
<td>2.9951</td>
<td>2.9988</td>
</tr>
<tr>
<td>std</td>
<td>3.6053</td>
<td>2.6605</td>
<td>1.9326</td>
<td>1.2366</td>
<td>0.1115</td>
<td>0.0594</td>
<td>0.0398</td>
</tr>
<tr>
<td>MSE</td>
<td>12.9994</td>
<td>7.0891</td>
<td>3.7350</td>
<td>1.5297</td>
<td>0.0125</td>
<td>0.0036</td>
<td>0.0016</td>
</tr>
</tbody>
</table>

Table 2: Mean, Standard Deviation and MSE for \(\sqrt{N}[(\hat{\alpha}_1, \hat{\alpha}_2, \hat{\beta}) - (\alpha_1, \alpha_2, \beta)]\)

<table>
<thead>
<tr>
<th>Sample Size</th>
<th>100</th>
<th>200</th>
<th>500</th>
<th>1000</th>
<th>2000</th>
<th>5000</th>
<th>10000</th>
</tr>
</thead>
<tbody>
<tr>
<td>mean</td>
<td>2.8772</td>
<td>1.0390</td>
<td>0.6467</td>
<td>0.3966</td>
<td>0.2978</td>
<td>0.2022</td>
<td>0.1570</td>
</tr>
<tr>
<td>(\alpha_1) std</td>
<td>27.2936</td>
<td>3.1881</td>
<td>2.6172</td>
<td>2.4199</td>
<td>2.4216</td>
<td>2.4096</td>
<td>2.3861</td>
</tr>
<tr>
<td>MSE</td>
<td>753.2219</td>
<td>11.2436</td>
<td>7.2680</td>
<td>6.0131</td>
<td>5.9526</td>
<td>5.8472</td>
<td>5.7180</td>
</tr>
<tr>
<td>mean</td>
<td>2.8948</td>
<td>1.0918</td>
<td>0.6959</td>
<td>0.4265</td>
<td>0.3401</td>
<td>0.1903</td>
<td>0.1491</td>
</tr>
<tr>
<td>(\alpha_2) std</td>
<td>27.2421</td>
<td>3.2199</td>
<td>2.6892</td>
<td>2.4776</td>
<td>2.4286</td>
<td>2.3981</td>
<td>2.3837</td>
</tr>
<tr>
<td>MSE</td>
<td>750.5123</td>
<td>11.5595</td>
<td>7.7163</td>
<td>6.3202</td>
<td>6.0137</td>
<td>5.7870</td>
<td>5.7044</td>
</tr>
<tr>
<td>mean</td>
<td>-0.3552</td>
<td>1.4884</td>
<td>0.0697</td>
<td>-0.6618</td>
<td>-0.3272</td>
<td>-0.3487</td>
<td>-0.1172</td>
</tr>
<tr>
<td>(\beta) std</td>
<td>36.0529</td>
<td>37.6246</td>
<td>43.2146</td>
<td>43.1059</td>
<td>4.9870</td>
<td>4.2035</td>
<td>3.9832</td>
</tr>
<tr>
<td>MSE</td>
<td>1299.9351</td>
<td>1417.8261</td>
<td>1867.5048</td>
<td>1529.7064</td>
<td>24.9777</td>
<td>17.7907</td>
<td>15.8800</td>
</tr>
</tbody>
</table>
Table 3: Mean, Standard Deviation and MSE for fast-bootstrap estimates

<table>
<thead>
<tr>
<th>Sample Size</th>
<th>100</th>
<th>200</th>
<th>500</th>
<th>1000</th>
<th>2000</th>
<th>5000</th>
<th>10000</th>
</tr>
</thead>
<tbody>
<tr>
<td>mean</td>
<td>14.5373</td>
<td>1.5004</td>
<td>1.0546</td>
<td>1.0247</td>
<td>1.0126</td>
<td>1.0052</td>
<td>1.0027</td>
</tr>
<tr>
<td>$\alpha_1$ std</td>
<td>101.0498</td>
<td>4.7923</td>
<td>0.1230</td>
<td>0.0781</td>
<td>0.0547</td>
<td>0.0343</td>
<td>0.0239</td>
</tr>
<tr>
<td>MSE</td>
<td>10394.3270</td>
<td>23.2167</td>
<td>0.0181</td>
<td>0.0067</td>
<td>0.0032</td>
<td>0.0012</td>
<td>0.0006</td>
</tr>
<tr>
<td>mean</td>
<td>14.9813</td>
<td>1.5008</td>
<td>1.0569</td>
<td>1.0256</td>
<td>1.0136</td>
<td>1.0050</td>
<td>1.0027</td>
</tr>
<tr>
<td>$\alpha_2$ std</td>
<td>104.2516</td>
<td>4.8161</td>
<td>0.1263</td>
<td>0.0802</td>
<td>0.0549</td>
<td>0.0341</td>
<td>0.0239</td>
</tr>
<tr>
<td>MSE</td>
<td>11063.8677</td>
<td>23.4453</td>
<td>0.0192</td>
<td>0.0071</td>
<td>0.0032</td>
<td>0.0012</td>
<td>0.0006</td>
</tr>
<tr>
<td>mean</td>
<td>3.3752</td>
<td>2.7205</td>
<td>3.1420</td>
<td>2.9289</td>
<td>2.9914</td>
<td>2.9950</td>
<td>2.9988</td>
</tr>
<tr>
<td>$\beta$ std</td>
<td>7.8970</td>
<td>13.4796</td>
<td>5.4209</td>
<td>1.8550</td>
<td>0.1326</td>
<td>0.0628</td>
<td>0.0414</td>
</tr>
<tr>
<td>MSE</td>
<td>62.5034</td>
<td>181.7785</td>
<td>29.4063</td>
<td>3.4459</td>
<td>0.0177</td>
<td>0.0040</td>
<td>0.0017</td>
</tr>
</tbody>
</table>

Results are reported for $E_{\text{bootstrap}}(\alpha_1^*, \alpha_2^*, \beta^*)$ - Fast Bootstrap [$\alpha^*$ was obtained by bootstrapping the first stage; $\beta^*$ was obtained by inverting the fast bootstrap statistic].

Table 4: Mean, Standard Deviation and MSE for conventional bootstrap estimates

<table>
<thead>
<tr>
<th>Sample Size</th>
<th>100</th>
<th>200</th>
<th>500</th>
<th>1000</th>
<th>2000</th>
<th>5000</th>
<th>10000</th>
</tr>
</thead>
<tbody>
<tr>
<td>mean</td>
<td>14.5373</td>
<td>1.5004</td>
<td>1.0546</td>
<td>1.0247</td>
<td>1.0126</td>
<td>1.0052</td>
<td>1.0027</td>
</tr>
<tr>
<td>$\alpha_1$ std</td>
<td>101.0498</td>
<td>4.7923</td>
<td>0.1230</td>
<td>0.0781</td>
<td>0.0547</td>
<td>0.0343</td>
<td>0.0239</td>
</tr>
<tr>
<td>MSE</td>
<td>10394.3270</td>
<td>23.2167</td>
<td>0.0181</td>
<td>0.0067</td>
<td>0.0032</td>
<td>0.0012</td>
<td>0.0006</td>
</tr>
<tr>
<td>mean</td>
<td>14.9813</td>
<td>1.5008</td>
<td>1.0569</td>
<td>1.0256</td>
<td>1.0136</td>
<td>1.0050</td>
<td>1.0027</td>
</tr>
<tr>
<td>$\alpha_2$ std</td>
<td>104.2516</td>
<td>4.8161</td>
<td>0.1263</td>
<td>0.0802</td>
<td>0.0549</td>
<td>0.0341</td>
<td>0.0239</td>
</tr>
<tr>
<td>MSE</td>
<td>11063.8677</td>
<td>23.4453</td>
<td>0.0192</td>
<td>0.0071</td>
<td>0.0032</td>
<td>0.0012</td>
<td>0.0006</td>
</tr>
<tr>
<td>mean</td>
<td>2.9978</td>
<td>2.9626</td>
<td>2.9711</td>
<td>2.9752</td>
<td>2.9743</td>
<td>2.9902</td>
<td>2.9971</td>
</tr>
<tr>
<td>$\beta$ std</td>
<td>0.3769</td>
<td>0.3386</td>
<td>0.2989</td>
<td>0.2148</td>
<td>0.1334</td>
<td>0.0591</td>
<td>0.0389</td>
</tr>
<tr>
<td>MSE</td>
<td>0.1420</td>
<td>0.1161</td>
<td>0.0902</td>
<td>0.0468</td>
<td>0.0185</td>
<td>0.0036</td>
<td>0.0015</td>
</tr>
</tbody>
</table>

Results are reported for $E_{\text{bootstrap}}(\alpha_1^*, \alpha_2^*, \beta^*)$ for conventional bootstrap.
Table 5: Mean, Standard Deviation and MSE for fast-bootstrap bias corrected estimates

<table>
<thead>
<tr>
<th>Sample Size</th>
<th>100</th>
<th>200</th>
<th>500</th>
<th>1000</th>
<th>2000</th>
<th>5000</th>
<th>10000</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha_1$</td>
<td>mean</td>
<td>-11.9619</td>
<td>0.6466</td>
<td>1.0032</td>
<td>1.0004</td>
<td>1.0008</td>
<td>1.0006</td>
</tr>
<tr>
<td></td>
<td>std</td>
<td>100.9696</td>
<td>4.5941</td>
<td>0.1116</td>
<td>0.0750</td>
<td>0.0536</td>
<td>0.0339</td>
</tr>
<tr>
<td></td>
<td>MSE</td>
<td>10362.8785</td>
<td>21.2309</td>
<td>0.0125</td>
<td>0.0056</td>
<td>0.0029</td>
<td>0.0012</td>
</tr>
<tr>
<td>$\alpha_2$</td>
<td>mean</td>
<td>-12.4023</td>
<td>0.6537</td>
<td>1.0054</td>
<td>1.0014</td>
<td>1.0017</td>
<td>1.0004</td>
</tr>
<tr>
<td></td>
<td>std</td>
<td>104.1809</td>
<td>4.6333</td>
<td>0.1146</td>
<td>0.0767</td>
<td>0.0538</td>
<td>0.0337</td>
</tr>
<tr>
<td></td>
<td>MSE</td>
<td>11033.2819</td>
<td>21.5872</td>
<td>0.0132</td>
<td>0.0059</td>
<td>0.0029</td>
<td>0.0011</td>
</tr>
<tr>
<td>$\beta$</td>
<td>mean</td>
<td>2.7742</td>
<td>3.4900</td>
<td>2.8642</td>
<td>3.0292</td>
<td>2.9940</td>
<td>2.9951</td>
</tr>
<tr>
<td></td>
<td>std</td>
<td>7.1746</td>
<td>13.5140</td>
<td>4.2968</td>
<td>2.6226</td>
<td>0.1220</td>
<td>0.0577</td>
</tr>
<tr>
<td></td>
<td>MSE</td>
<td>51.5254</td>
<td>182.8681</td>
<td>18.4813</td>
<td>6.8789</td>
<td>0.0149</td>
<td>0.0034</td>
</tr>
</tbody>
</table>

Table 6: Mean, Standard Deviation and MSE for conventional bootstrap bias corrected estimates

<table>
<thead>
<tr>
<th>Sample Size</th>
<th>100</th>
<th>200</th>
<th>500</th>
<th>1000</th>
<th>2000</th>
<th>5000</th>
<th>10000</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha_1$</td>
<td>mean</td>
<td>-11.9619</td>
<td>0.6466</td>
<td>1.0032</td>
<td>1.0004</td>
<td>1.0008</td>
<td>1.0006</td>
</tr>
<tr>
<td></td>
<td>std</td>
<td>100.9696</td>
<td>4.5941</td>
<td>0.1116</td>
<td>0.0750</td>
<td>0.0536</td>
<td>0.0339</td>
</tr>
<tr>
<td></td>
<td>MSE</td>
<td>10362.8785</td>
<td>21.2309</td>
<td>0.0125</td>
<td>0.0056</td>
<td>0.0029</td>
<td>0.0012</td>
</tr>
<tr>
<td>$\alpha_2$</td>
<td>mean</td>
<td>-12.4023</td>
<td>0.6537</td>
<td>1.0054</td>
<td>1.0014</td>
<td>1.0017</td>
<td>1.0004</td>
</tr>
<tr>
<td></td>
<td>std</td>
<td>104.1809</td>
<td>4.6333</td>
<td>0.1146</td>
<td>0.0767</td>
<td>0.0538</td>
<td>0.0337</td>
</tr>
<tr>
<td></td>
<td>MSE</td>
<td>11033.2819</td>
<td>21.5872</td>
<td>0.0132</td>
<td>0.0059</td>
<td>0.0029</td>
<td>0.0011</td>
</tr>
<tr>
<td>$\beta$</td>
<td>mean</td>
<td>2.9311</td>
<td>3.2479</td>
<td>3.0352</td>
<td>2.9830</td>
<td>3.0111</td>
<td>3.0000</td>
</tr>
<tr>
<td></td>
<td>std</td>
<td>7.1345</td>
<td>5.3097</td>
<td>3.8377</td>
<td>2.4843</td>
<td>0.1483</td>
<td>0.0651</td>
</tr>
<tr>
<td></td>
<td>MSE</td>
<td>50.9063</td>
<td>28.2541</td>
<td>14.7288</td>
<td>6.1718</td>
<td>0.0221</td>
<td>0.0042</td>
</tr>
</tbody>
</table>
Table 7: Coverage Properties of Confidence Intervals

<table>
<thead>
<tr>
<th>Sample Size</th>
<th>Confidence Level</th>
<th>100</th>
<th>200</th>
<th>500</th>
<th>1000</th>
<th>2000</th>
<th>5000</th>
<th>10000</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>a. Fast Bootstrap</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>99%</td>
<td>0.9840</td>
<td>0.9910</td>
<td>0.9890</td>
<td>0.9880</td>
<td>0.9930</td>
<td>0.9950</td>
<td>0.9910</td>
</tr>
<tr>
<td></td>
<td>95%</td>
<td>0.9500</td>
<td>0.9600</td>
<td>0.9530</td>
<td>0.9620</td>
<td>0.9590</td>
<td>0.9610</td>
<td>0.9530</td>
</tr>
<tr>
<td></td>
<td>90%</td>
<td>0.9120</td>
<td>0.9250</td>
<td>0.9200</td>
<td>0.9220</td>
<td>0.9240</td>
<td>0.9130</td>
<td>0.9180</td>
</tr>
<tr>
<td></td>
<td>b. Conventional Bootstrap</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>99%</td>
<td>0.9660</td>
<td>0.9910</td>
<td>0.9960</td>
<td>0.9970</td>
<td>0.9940</td>
<td>0.9720</td>
<td>0.9780</td>
</tr>
<tr>
<td></td>
<td>95%</td>
<td>0.9210</td>
<td>0.9590</td>
<td>0.9800</td>
<td>0.9840</td>
<td>0.9730</td>
<td>0.9320</td>
<td>0.9240</td>
</tr>
<tr>
<td></td>
<td>90%</td>
<td>0.8470</td>
<td>0.9100</td>
<td>0.9470</td>
<td>0.9380</td>
<td>0.9290</td>
<td>0.8740</td>
<td>0.8660</td>
</tr>
<tr>
<td></td>
<td>b. Standard CLT Approach</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>99%</td>
<td>0.9880</td>
<td>0.9910</td>
<td>0.9870</td>
<td>0.9870</td>
<td>0.9880</td>
<td>0.9890</td>
<td>0.9890</td>
</tr>
<tr>
<td></td>
<td>95%</td>
<td>0.9580</td>
<td>0.9640</td>
<td>0.9520</td>
<td>0.9480</td>
<td>0.9580</td>
<td>0.9460</td>
<td>0.9500</td>
</tr>
<tr>
<td></td>
<td>90%</td>
<td>0.9180</td>
<td>0.9350</td>
<td>0.9150</td>
<td>0.9140</td>
<td>0.9120</td>
<td>0.9020</td>
<td>0.9020</td>
</tr>
</tbody>
</table>

Table 8: Average Number of Minutes Needed for each Simulation of Monte-Carlo Experiment

<table>
<thead>
<tr>
<th>Sample Size</th>
<th>Bootstrap Method</th>
<th>100</th>
<th>200</th>
<th>500</th>
<th>1000</th>
<th>2000</th>
<th>5000</th>
<th>10000</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Fast</td>
<td>0.26</td>
<td>0.33</td>
<td>0.68</td>
<td>1.25</td>
<td>2.39</td>
<td>6.68</td>
<td>13.72</td>
</tr>
<tr>
<td></td>
<td>Conventional</td>
<td>6.49</td>
<td>7.54</td>
<td>10.74</td>
<td>28.57</td>
<td>27.83</td>
<td>73.63</td>
<td>168.91</td>
</tr>
</tbody>
</table>

The simulations were performed in MATLAB(R) using a Unix Server running 8 parallel processes with 8-core CPUs.