cemmap

centre for microdata methods and practice

Bootstrap based asymptotic refinements for high-dimensional nonlinear models

Joel L. Horowitz
Ahnaf Rafi

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP06/23

Economic
and Social
Research Council

BOOTSTRAP BASED ASYMPTOTIC REFINEMENTS FOR HIGH-DIMENSIONAL NONLINEAR MODELS

Joel L. Horowitz
and
Ahnaf Rafi
Department of Economics, Northwestern University [*]

March 2023

Abstract

We consider penalized extremum estimation of a high-dimensional, possibly nonlinear model that is sparse in the sense that most of its parameters are zero but some are not. We use the SCAD penalty function, which provides model selection consistent and oracle efficient estimates under suitable conditions. However, asymptotic approximations based on the oracle model can be inaccurate with the sample sizes found in many applications. This paper gives conditions under which the bootstrap, based on estimates obtained through SCAD penalization with thresholding, provides asymptotic refinements of size $O\left(n^{-2}\right)$ for the error in the rejection (coverage) probability of a symmetric hypothesis test (confidence interval) and $O\left(n^{-1}\right)$ for the error in rejection (coverage) probability of a one-sided or equal tailed test (confidence interval). The results of Monte Carlo experiments show that the bootstrap can provide large reductions in errors in coverage probabilities. The bootstrap is consistent, though it does not necessarily provide asymptotic refinements, even if some parameters are close but not equal to zero. Random-coefficients logit and probit models and nonlinear moment models are examples of models to which the procedure applies.

Keywords: extremum estimation, nonlinear models, high-dimensional inference, bootstrap based confidence intervals, asymptotic refinements

[*] We thank Jian Huang, Soumendra Lahiri, and Whitney Newey for helpful comments and discussions.

1 Introduction

This paper is about using the bootstrap to obtain asymptotic refinements for inference about the sparse but possibly high-dimensional parameter θ_0 that is estimated by a thresholded version of the penalized extremum estimator

$$\widetilde{\theta}_n = \underset{\theta \in \Theta_n}{\operatorname{argmin}} \left[Q_n \left(\chi_n, \theta \right) + p_{\lambda_n}(\theta) \right]. \tag{1.1}$$

In this equation, Θ_n is a parameter set, χ_n is a random sample of size n from the distribution of a random vector, Q_n is a known function such as minus a log-likelihood function, p_{λ_n} is a penalty function, and is λ_n is a penalization parameter. In contrast to most of the large literature on high-dimensional estimation, we do not assume that θ is the vector of parameters of a linear or generalized linear model, the vector of coefficients of a linear combination of covariates (linear index), or the vector of coefficients of a linear approximation to a nonlinear function. Instead, Q_n (or $-Q_n$) is the objective function of a general extremum estimator, such as a maximum likelihood estimator; linear or nonlinear regression estimator; instrumental variables estimator of a linear or nonlinear model; or generalized method of moments (GMM) estimator. The random coefficients logit and probit models are examples of widely used models that are neither generalized linear models, linear index models, or easily approximated by a linear combination of functions of their covariates. A non-separable, nonlinear demand model with a possibly endogenous price variable is another example. Maximum likelihood estimation of random coefficients logit or probit models and GMM estimation of a demand model are among the estimators that are accommodated by the methods presented in this paper.

If θ_0 has a fixed dimension and is point identified, then conditions under which $\sqrt{n}(\widetilde{\theta}_n - \theta_0)$ is asymptotically normally distributed without penalization are well known. See, for example, Amemiya (1985), among many other references. However, the asymptotic normal approximation can be inaccurate with the sample sizes found in many applications. Under conditions that are satisfied in many applications, the bootstrap provides asymptotic refinements for confidence intervals and hypothesis tests. See, for example, Hall (1992) and Horowitz (2001), among other references. The resulting reductions in the differences between true and nominal coverage and rejection probabilities (errors in coverage and rejection probabilities or ECPs and ERPs) can be large. This paper gives conditions under which the same asymptotic refinements can be obtained in penalized estimation of nonlinear models. We assume that θ_0 is sparse, meaning that most of its components are zero but some are non-zero. We carry out inference about a non-zero component or linear combination of non-zero components. We give conditions under which bootstrap asymptotic refinements have the same order of magnitude that they would have if they were obtained from estimation of the oracle model (the model in which it is known a priori which components are non-zero and which are zero). For example, the error in the coverage (rejection) probability of a symmetrical confidence interval (hypothesis test) is $O\left(n^{-2}\right)$. We also give conditions under which the bootstrap is consistent, though it does not necessarily provide asymptotic refinements, even if some non-zero components of θ_0 are close but not equal to zero. Under these conditions, there is not a risk of inconsistency due to violation of the exact sparsity assumption.

Chatterjee and Lahiri (2010, 2011) give conditions under which the bootstrap based on LASSO (Tibshirani (1996)) estimators of high-dimensional linear models is consistent. Chatterjee and Lahiri (2013) give conditions under which the bootstrap provides asymptotic refinements for confidence intervals and hypothesis tests based on adaptive LASSO (ALASSO, Zou (2006)) estimators of high-dimensional linear models. Das et al. (2022) give conditions under which the bootstrap provides asymptotic refinements for symmetrical tests and confidence intervals in penalized high dimensional linear models with a variety of penalty functions. The asymptotic refinements provided in the present paper are of the same or higher order than those of Chatterjee and Lahiri (2013) and Das et al. (2022) but are for models that may be nonlinear.

We use the SCAD penalty function (Antoniadis and Fan (2001), Fan and Li (2001)), which avoids penalization bias of estimates of the non-zero components of θ_0. Fan and Li (2001) and Fan and Peng (2004) give conditions under which a penalized maximum likelihood estimator with the SCAD penalty function is oracle efficient, meaning that the centered and scaled estimates of non-zero components of θ_0 have the same asymptotic distribution that they would have if it were known a priori which components are zero. We consider the more general estimator (1.1).

To the best of our knowledge, this paper is the first to obtain its order of asymptotic refinements for high dimensional models that may be nonlinear. Not surprisingly, achieving these refinements requires assumptions that are stronger than those in much of the recent literature. We assume that the number of parameters is less than n, though it may be an increasing function of n, and that the number of non-zero components of θ_0 is fixed as n increases. These assumptions are motivated by applications in the social sciences, where a model may have many parameters, but the total number of parameters is less than the sample size and few have substantial effects on the dependent variable. In a random coefficients logit or probit model, for example, the parameters include the means and variances of the coefficients, but some coefficients may be non-stochastic, in which case the corresponding variances are zero. As in Chatterjee and Lahiri (2013) and Das et al. (2022), we require the non-zero parameters to be sufficiently far from zero, though the distance of these parameters from zero can decrease as the sample size increases. This ensures that “large” parameters can be distinguished from “small” ones with sufficiently high probability. It is not possible to obtain asymptotic refinements of the order obtained here without making such a distinction, though the bootstrap is consistent even if some nonzero parameters are “small.” The appendix presents precise statements of our assumptions and their justifications.

The results in this paper are most closely related to those of Chatterjee and Lahiri (2013) and Das et al. (2022), who show that suitable versions of the residual and permutation bootstraps provide asymptotic refinements for test statistics based on a large class of penalized estimators of the coefficients of a linear mean-regression model. See also Das et al. (2019). These papers obtain their results by carrying out higher-order expansions of the distributions of the relevant statistics. This paper uses a different approach. We give conditions under which a combination of penalized estimation with the SCAD penalty function and hard thresholding causes differences between the penalized, thresholded parameter estimate and the infeasible oracle estimate to converge to zero very rapidly. Consequently, the estimate obtained from penalization and thresholding can be treated as if it were the oracle estimate. It suffices to consider only the properties of the bootstrap applied to the oracle model, which are well known. It is not necessary to carry out higher-order expansions of the distribution of the penalized estimator.

The literature on high-dimensional estimation is very large. Here, we mention only few references that are most relevant to the present paper. Tibshirani (1996) introduces the LASSO. Knight and Fu (2000); Candes and Tao (2007); Huang et al. (2008); Zhang and Huang (2008); Belloni and Chernozhukov (2011); and Bühlmann and Van De Geer (2011) describe properties of the LASSO and related penalized estimators of linear mean- and quantile-regression models. Zou (2006) introduces the ALASSO and describes its properties. Fan and Li (2001) and Fan and Peng (2004) describe properties of SCAD-penalized least squares and maximum likelihood estimators. Belloni et al. (2014); van de Geer et al. (2014); Zhang and Zhang (2014); Chernozhukov et al. (2018); Lu et al. (2017); Wang et al. (2020); and Yu et al. (2020) describe methods for first order asymptotic inference in high-dimensional models. Bühlmann (2015) and Dezeure et al. (2015) provide reviews. Bach (2009); Chatterjee and Lahiri (2010, 2011, 2013); Javanmard and Montanari (2018); Minnier et al. (2011); Camponovo (2015, 2020); Dezeure et al. (2017); Zhang and Cheng (2017); Wang et al. (2018); Das et al. (2019); Liu et al. (2020); and Das et al. (2022) describe bootstrap methods for high-dimensional estimators. Chatterjee and Lahiri (2011) apply hard thresholding to the LASSO estimator of a linear model. They show that the residual bootstrap consistently estimates the distribution of the t-statistic this model but do not obtain asymptotic refinements or treat nonlinear models. Meinshausen and Yu (2009) and Bühlmann and Van De Geer (2011) also discuss thresholding the LASSO.

The remainder of this paper is organized as follows. Section 2 describes our method and its properties. This section also treats the case in which some parameters are close but not equal to zero. Section 3 presents the results of a Monte Carlo investigation of the numerical behavior of the method, and Section 4 presents an empirical example of the application of the method. Section 5 presents conclusions. Regularity conditions are in the appendix. The proofs of theorems and certain auxiliary results are in the online supplement.

2 The Method

Section 2.1 defines notation that is used in the remainder of this paper. Section 2.2 presents the bootstrap method and its properties. Section 2.3 treats the case in which some parameters are close but not equal to zero.

2.1 Notation

Let $\chi_n = \{X_i : i = 1, \ldots, n\}$ be an independent random sample of the random vector X. To accommodate the possibility that the dimension of the target parameter may increase as n increases, we use the notation θ_{0n} for the target parameter and Θ_n for the parameter set. Define $p_n = \dim(\theta_{0n})$. Let A_0 and $\overline{A}_0$, respectively denote the sets of indices of the non-zero and zero components of θ_{0n}. Let θ_{0n} have p_0 non-zero components. This number is fixed as n increases. We assume without further loss of generality that the first p_0 components of θ_{0n} are the non-zero ones. Thus, $A_0 = \{1, \ldots, p_0\}$, and $\overline{A}_0 = \{p_0 + 1, \ldots, p_n\}$. Let θ_{0nA_0} be the $p_0 \times 1$ vector of non-zero components of θ_{0n}. Then $\theta'_{0n} = \left(\theta'_{0nA_0}, 0'_{p_n - p_0}\right)$, where $0_{p_n-p_0}$ denotes a $(p_n - p_0) \times 1$ vector of zeros. Denote a generic element of Θ_n by $\theta'_n = \left(\theta'_{nA_0}, \theta'_{n\overline{A}_0}\right)$. Write $Q_n(\chi_n, \theta_n)$ as $Q_n\left(\chi_n, \theta_{nA_0}, \theta_{n\overline{A}_0}\right)$ when it is necessary to distinguish between components whose indices are in A_0 and components whose indices are in $\overline{A}_0$. Let θ_{nj} $(j = 1, \ldots, p_n)$ denote the j^{th} component of any vector $\theta_n \in \Theta_n$. The penalty parameter depends on n and therefore, is denoted by λ_n. The SCAD penalty function is

$$p_{\lambda_n}(\theta_n) = \lambda_n \sum_{j=1}^{p_n} \widetilde{p}_{\lambda_n}(|\theta_{nj}|),$$

where the function $\widetilde{p}_{\lambda_n}$ is defined by its derivative

$$\widetilde{p}'_{\lambda_n}(v) = I(v \leq \lambda_n) + \frac{\max\{a\lambda_n - v, 0\}}{(a-1)\lambda_n} I(v > \lambda_n); \quad a > 2, v > 0.$$

Let $\widetilde{\theta}_n$ denote the penalized extremum estimator defined in (1.1) with the SCAD penalty function. Define the thresholded estimator $\widehat{\theta}_n$ as the $p_n \times 1$ vector whose j^{th} component is

$$\widehat{\theta}_{nj} = \widetilde{\theta}_{nj} I\left(\widetilde{\theta}_{nj} \geq \tau_n\right) \tag{2.1}$$

where $\tau_n \ll \lambda_n$ is a thresholding parameter. Define the sets

$$\widehat{A}_0 = \left\{j = 1, \ldots, p_n : \left|\widehat{\theta}_{nj}\right| > 0\right\},$$

$$\overline{\widehat{A}}_0 = \left\{ j = 1, \ldots, p_n : \left| \widehat{\theta}_{nj} \right| = 0 \right\}.$$

Let $\widehat{\theta}_{n\widehat{A}_0}$ and $\widehat{\theta}_{n\overline{\widehat{A}}_0}$, respectively, denote the vectors of non-zero and zero components of $\widehat{\theta}_n$. Define

$$\Theta_n^O = \left\{ \theta_n \in \Theta_n : \theta_{n\overline{A}_0} = 0_{p_n - p_0} \right\},$$

and

$$\Theta_n^{PO} = \left\{ \theta_n \in \Theta_n : \theta_{n\overline{\widehat{A}}_0} = 0_{\left|\overline{\widehat{A}}_0\right|} \right\},$$

where $\left|\overline{\widehat{A}}_0\right|$ is the number of elements in $\overline{\widehat{A}}_0$ and $0_{\left|\overline{\widehat{A}}_0\right|}$ is a $\left|\overline{\widehat{A}}_0\right| \times 1$ vector of zeros. The infeasible oracle estimator of θ_{0n} is $\widehat{\theta}_n^O = \left(\widehat{\theta}_{nA_0}^O, 0_{p_n - p_0} \right)$, where

$$\widehat{\theta}_n^O = \underset{\theta_n \in \Theta_n^O}{\operatorname{argmin}}\, Q_n \left(\chi_n, \theta_n \right). \tag{2.2}$$

This is the estimator obtained by setting $\theta_{n\overline{A}_0} = 0_{p_n - p_0}$ and choosing θ_{nA_0} to minimize the unpenalized objective function Q_n. Define the pseudo-oracle estimator $\widehat{\theta}_n^{PO}$ by

$$\widehat{\theta}_n^{PO} = \underset{\theta_n \in \Theta_n^{PO}}{\operatorname{argmin}}\, Q_n \left(\chi_n, \theta_n \right). \tag{2.3}$$

This is the estimator obtained by setting $\theta_{n\overline{\widehat{A}}_0} = 0_{\left|\overline{\widehat{A}}_0\right|}$ and choosing $\theta_{n\widehat{A}_0}$ to minimize the unpenalized objective function.

Finally, let $T\left(\widehat{\theta}_{n\widehat{A}_0}^{PO} \right)$ be a statistic based on $\widehat{\theta}_{n\widehat{A}_0}^{PO}$ for testing a hypothesis about a smooth scalar function of θ_{0nA_0}. For example, T might be a symmetrical t-statistic for testing a hypothesis about the j^{th} component of θ_{0nA_0}. Denote the hypothesized value of this component by $\theta_{0nA_0,j}$. Then

$$T\left(\widehat{\theta}_{n\widehat{A}_0}^{PO} \right) = \frac{\left| \widehat{\theta}_{n\widehat{A}_0,j}^{PO} - \theta_{0nA_0,j} \right|}{s^{PO}}, \tag{2.4}$$

where s^{PO} is a standard error, and $\theta_{0nA_0,j}$ is the hypothesized value. Let $T\left(\widehat{\theta}_{nA_0}^O \right)$ be the same statistic based on the oracle estimate $\widehat{\theta}_{nA_0}^O$. Then, the foregoing statistic is

$$T\left(\widehat{\theta}_{nA_0}^O \right) = \frac{\left| \widehat{\theta}_{nA_0,j}^O - \theta_{0nA_0,j} \right|}{s^O},$$

where s^O is a standard error. Let $\widehat{c}_\alpha^{PO}$ be the α-level critical value of $T\left(\widehat{\theta}^{PO}_{n\widehat{A}_0}\right)$ that is obtained by the bootstrap procedure described in Section 2.2. Let $\widehat{c}_\alpha^{O}$ be the α-level critical value of $T\left(\widehat{\theta}^{O}_{nA_0}\right)$ that would be obtained through the conventional bootstrap if A_0 were known.

2.2 Description and Properties of the Method

The method proposed in this paper consists of the following steps.

1. Obtain the penalized estimator $\widetilde{\theta}_n$ from (1.1) with the SCAD penalty function.

2. Obtain the thresholded estimator $\widehat{\theta}_n$ from (2.1).

3. Obtain the pseudo-oracle estimator $\widehat{\theta}^{PO}_{n\widehat{A}_0}$ from (2.3).

4. Obtain bootstrap samples by sampling the data randomly with replacement (not the residual bootstrap). Obtain the critical value $\widehat{c}_\alpha^{PO}$ by using the conventional bootstrap methods with $\widehat{\theta}^{PO}_{n\widehat{A}_0}$ treated as if it were true oracle estimator $\widehat{\theta}^{O}_{nA_0}$.

If A_0 were known and the values of the parameters were fixed, an α-level critical value $\widehat{c}_\alpha^{O}$ could be obtained by applying conventional bootstrap methods such as those described by Hall (1992) and Horowitz (2001) to the oracle estimator. The null hypothesis being tested would be rejected at the nominal level α if $T\left(\widehat{\theta}^{O}_{nA_0}\right) > \widehat{c}_\alpha^{O}$. The difference between the nominal and true probabilities of rejecting a correct null hypothesis (the ERP) would be

$$\mathrm{ERP}^O = \left| P\left[T\left(\widehat{\theta}^{O}_{nA_0}\right) > \widehat{c}_\alpha^{O} \right] - \alpha \right|.$$

ERP^O is $O\left(n^{-2}\right)$ for a the Studentized symmetrical statistic $T\left(\widehat{\theta}^{O}_{nA_0}\right)$. It is typically $O\left(n^{-c}\right)$, where $c = 1/2, 1, 3/2, 2$, depending on the hypothesis and the test statistic, for non-Studentized statistics or statistics for one-sided or equal-tailed hypothesis tests. We assume that the conventional bootstrap provides the same order of refinements if some non-zero parameters approach zero at the rate specified in Assumption 2 (ii) of the appendix. This paper concentrates on the Studentized, symmetrical statistic in (2.4), but the results presented here apply after obvious modifications to the other statistics.

The method of this paper replaces the unknown A_0 with $\widehat{A}_0$ and applies conventional bootstrap methods to $\widehat{\theta}^{PO}_{n\widehat{A}_0,j}$ as if $\widehat{A}_0$ were non-stochastic. This works because $P\left(\widehat{A}_0 \neq A_0\right)$ approaches zero very rapidly. The precise result is given by the following theorem.

Theorem 2.1. *Let Assumptions 1-5 in the appendix hold. Then*

$$P\left(\widehat{A}_0 \neq A_0\right) = o\left(n^{-2}\right) \tag{2.5}$$

as $n \to \infty$. In addition

$$\left|P\left[T\left(\widehat{\theta}^{PO}_{n\widehat{A}_0}\right) > \widehat{c}^{PO}_{\alpha}\right] - P\left[T\left(\widehat{\theta}^{O}_{nA_0}\right) > \widehat{c}^{O}_{\alpha}\right]\right| = o\left(n^{-2}\right). \tag{2.6}$$

It follows from Theorem 2.1 that

$$\text{ERP}^{PO} = \left|P\left[T\left(\widehat{\theta}^{PO}_{n\widehat{A}_0}\right) > \widehat{c}^{PO}_{\alpha}\right] - \alpha\right| = O\left(n^{-2}\right). \tag{2.7}$$

Therefore, a test based on the feasible statistic $T\left(\widehat{\theta}^{PO}_{n\widehat{A}_0}\right)$ and a feasible bootstrap critical value $\widehat{c}^{PO}_{\alpha}$ is equivalent up to $O\left(n^{-2}\right)$ to a test based on the infeasible statistic $T\left(\widehat{\theta}^{O}_{nA_0}\right)$ and infeasible bootstrap critical value $\widehat{c}^{O}_{\alpha}$. Moreover, $\widehat{\theta}^{PO}_{n\widehat{A}_0}$ is an oracle efficient estimator of θ_{0n}. A confidence interval for a smooth scalar function of θ_{0nA_0} is the set of values of the function that are not rejected by the hypothesis test. Therefore, (2.7) also applies to the error in the coverage probability (ECP) of a confidence interval.

2.3 Small Parameters

In this section, we assume that some or all components of $\theta_{0n\overline{A}_0}$ are non-zero but small in the sense that $\left\|\theta_{0n\overline{A}_0}\right\|_1 = o\left(\tau_n\right)$. We use the following additional notation. Let θ_{0nA_0}

denote the $p_0 \times 1$ vector of components of that are "large" in the sense that $|\theta_{0nA_0,j}| \gg \lambda_n$ for all $j = 1, \ldots, p_0$. Define $\widehat{\theta}^{PT}_{nA_0}$ to be the parameter estimate obtained from the unpenalized pseudo-true model

$$\widehat{\theta}^{PT}_{nA_0} = \underset{\theta_{nA_0}}{\operatorname{argmin}}\, Q_n\left(\chi_n, \theta_{nA_0}, 0_{\overline{A}_0}\right). \tag{2.8}$$

This is the pseudo-true estimate of the large parameters that would be obtained if A_0 were known and all the components of the parameters with indices in $\overline{A}_0$ were set equal to zero.

The following theorem gives conditions under which the bootstrap consistently estimates the asymptotic distribution of

$$T\left(\widehat{\theta}^{PT}_{nA_0}\right) = \frac{\widehat{\theta}^{PT}_{nA_0,j} - \theta_{0nA_0,j}}{s^{PT}},$$

where s^{PT} is a standard error. $T\left(\widehat{\theta}^{PT}_{nA_0}\right)$ is a t-statistic for testing a hypothesis about $\theta_{0nA_0,j}$ (the true parameter value, not a pseudo-true value) when A_0 is known. Under the assumptions of the theorem, the bootstrap estimates the distribution of $T\left(\widehat{\theta}^{PT}_{nA_0}\right)$ consistently when A_0 is unknown. Equivalently, the bootstrap provides a confidence interval for $\theta_{0nA_0,j}$ with asymptotically correct coverage probability.

Theorem 2.2. *Let Assumptions 2(ii), 3(i) and 1S-3S of the appendix hold. Let $\widehat{\theta}^{PT}_{n\widehat{A}_0}$ be the estimate of θ_{0nA_0} obtained from the unpenalized pseudo-true model with $\widehat{A}_0$ in place of A_0:*

$$\widehat{\theta}^{PT}_{n\widehat{A}_0} = \underset{\theta_{n\widehat{A}_0}}{\operatorname{argmin}}\, Q_n\left(\chi_n, \theta_{n\widehat{A}_0}, 0_{\overline{\widehat{A}}_0}\right),$$

where $\widehat{A}_0$ is defined following (2.1). Denote the bootstrap sample by χ^*_n and the bootstrap estimate based on the unpenalized pseudo-true model by

$$\theta^{*PT}_{n\widehat{A}_0} = \underset{\theta_{n\widehat{A}_0}}{\operatorname{argmin}}\, Q_n\left(\chi^*_n, \theta_{n\widehat{A}_0}, 0_{\overline{\widehat{A}}_0}\right).$$

Define the statistic T based on $\theta^{*PT}_{n\widehat{A}_0}$ by

$$T\left(\theta^{*PT}_{n\widehat{A}_0}\right) = \frac{\theta^{*PT}_{n\widehat{A}_0,j} - \widehat{\theta}^{PT}_{n\widehat{A}_0,j}}{s^{*PT}}$$

where s^{*PT} is the bootstrap standard error obtained when $\widehat{A}_0$ used in place of A_0. Then

$$\sup_{z\in\mathbb{R}}\left|P^*\left[T\left(\theta^{*PT}_{n\widehat{A}_0}\right)\le z\right]-P\left[T\left(\widehat{\theta}^{PT}_{nA_0}\right)\le z\right]\right|\xrightarrow{\text{p}}0.$$

3 Monte Carlo Experiments

This section reports the results of a Monte Carlo investigation of the finite-sample performance of the penalization and thresholding method. Section 3.1 describes the computational algorithm. Section 3.2 describes the investigation.

3.1 Computational Algorithm

The algorithm estimates θ_{0n} iteratively. Let $\widetilde{\theta}^0_n$ denote the starting value; $t=1,2,\ldots$ index iterations; and $\widetilde{\theta}^t_n$ denote the estimate of θ_{0n} at iteration t. We use the local linear approximation of the SCAD penalty function of Zou and Li (2008). We obtain $\widetilde{\theta}^{t+1}_n$ from $\widetilde{\theta}^t_n$ by using the coordinate descent method of Friedman et al. (2007) and Friedman et al. (2010) to solve

$$\widetilde{\theta}^{t+1}_n=\underset{\theta_n\in\Theta_n}{\operatorname{argmin}}\, Q_n\left(\theta_n\right)+\sum_{j=1}^{p_n}\widetilde{p}'_{\lambda_n}\left(\left|\widetilde{\theta}^t_{nj}\right|\right)\left|\theta_{nj}\right|.$$

To speed the computation, we iterate only over non-zero components at a given t and implement updates to the set of non-zero components as in Zhao et al. (2018).

3.2 Monte Carlo Results

We estimate the parameter θ_{0n} of the binary logit model

$$P(Y=1|X)=\frac{\exp\left(\theta'_{0n}X\right)}{1+\exp\left(\theta'_{0n}X\right)}\tag{3.1}$$

where $Y \in \{0,1\}$; $X \sim \mathcal{N}(0_{p_n}, \Sigma_n)$; $\Sigma_{n,j\ell} = 0.3^{|j-\ell|}$ for $j, \ell = 1, \ldots, p_n$; $p_n = n/10$, $p_n = n/2$ or $p_n = 3n/4$; $p_0 = 15$; $\dim(\theta_{0n}) = p_n \times 1$; and

$$\theta_{0n} = (4, -1.5, -3, 1.9, 2.6, 4, -1.5, -3, 1.9, 2.6, 4, -1.5, -3, 1.9, 2.6, 0, \ldots, 0).$$

$-Q_n(\chi_n, \theta_n)$ is the log-likelihood function for estimating θ_n. The penalty parameter was selected to minimize the BIC criterion

$$\min_{\lambda} Q_n\left(\chi_n, \widetilde{\theta}_n(\lambda)\right) + C_n \left|S_n(\lambda)\right| \log n, \tag{3.2}$$

where $S_n(\lambda)$ is the set of indices of non-zero components of the penalized MLE $\widetilde{\theta}_n(\lambda)$ with penalty parameter λ but without thresholding, $|S_n(\lambda)|$ is the cardinality of $S_n(\lambda)$, and $C_n = 1$ or $\log\log p_n$. The concavity parameter in the SCAD penalty function was $a = 3.7$ (Fan and Li 2001), and the thresholding parameter was $\tau_n = n^{-1/8} a \lambda_n$. There were 500 Monte Carlo replications and 2000 bootstrap replications per experiment.

Tables 8-13 show empirical coverage probabilities of nominal one-sided and symmetrical 0.90 intervals for $\theta_{0n,1}$ and $\theta_{0n,2}$ obtained the following six ways. Section B.4 of the online supplement presents additional results.

1. Unpenalized MLE of the full model with first-order asymptotic critical values.

2. Unpenalized MLE of the full model with bootstrap critical values.

3. Unpenalized MLE of the infeasible oracle model with first-order asymptotic critical values.

4. Unpenalized MLE of the infeasible oracle model with bootstrap critical values.

5. Penalized and thresholded MLE with first-order asymptotic critical values.

6. Penalized and thresholded MLE with bootstrap critical values.

The tables show that the empirical coverage probabilities obtained with the full model are far from the nominal coverage probabilities with either first-order asymptotic or bootstrap-based critical values. The empirical coverage probabilities are especially far from the nominal probabilities when $p_n = n/2$ and $p_n = 3n/4$. This is because the estimated Hessian and outer product matrices with these values of p_n are nearly singular. Consequently, random sampling errors in the inverse of the estimated Hessian (or outer product), which is used for Studentization, are very large. The empirical coverage probabilities obtained with the oracle model and bootstrap-based critical values are close to the nominal probabilities, but the oracle model is unknown and infeasible in applications. The empirical coverage probabilities obtained with the penalized, thresholded estimator and bootstrap-based critical values are close to the nominal probabilities and are not sensitive to the choice of C_n when $n \geq 1000$.

4 Empirical Example

Gentzkow et al. (2019) investigated the relation between party affiliation and two-word phrases (bigrams) spoken by members of Congress. We use a subset of their data to estimate a binary logit model of the probability of a member's party affiliation (Democrat or Republican) conditional on phrases that the member has used. Our data consist of observations on 4319 members of Congress from 2001-2016. The covariates are the number of times a member used each of 2441 phrases as well as 60 variables describing characteristics of the member (e.g. state represented, gender). Thus, the logit model has 2501 covariates in total. Most of the phrases are used roughly equally often by Democrats and Republicans. These phrases are unlikely to be useful for predicting party affiliation, thereby justifying an assumption of sparsity or approximate sparsity.

The model is as in (3.1), where $Y = 1$ if a member is a Republican, $Y = 0$ if the member is a Democrat, and X is the 2501×1 vector of covariates. The SCAD penalty and thresholding parameters were selected as described in Section 3.1. Penalized estimation with $C_n = 1$ in (3.2) resulted in selection of 106 phrases out of the initial 2441. Penalized estimation with $C_n = \log \log p_n$ resulted in selection of 59 phrases, 56 of which are among the 106 selected with $C_n = 1$.

Table 7 shows point estimates of the coefficients of several example phrases and nominal 90% first-order asymptotic and bootstrap-based confidence intervals for the coefficients. The point estimates and confidence intervals are not highly sensitive to the choice of C_n.

5 Conclusions

Empirical research in economics, among other fields, often involves estimation of the parameters of a nonlinear model in which the number of parameters may be a large fraction of the sample size but most parameters are zero or close to zero. In such settings, the accuracy of inference about large parameters can be improved greatly through the use of penalized estimation methods that reduce the number of parameters that must be estimated. However, inference is usually based on asymptotic approximations that can be highly inaccurate in finite samples. Under suitable conditions, the bootstrap provides asymptotic refinements that increase the accuracy of inference, but the usual conditions, such as those of Hall (1992), are not satisfied in penalized estimation. This paper has described a method for obtaining bootstrap asymptotic refinements in penalized estimation of nonlinear models. The refinements are of the same order as those that would be achieved with the oracle model if it were known. The results of Monte Carlo experiments show that with samples of the sizes encountered in much applied research, the method can achieve large reductions in the errors of the coverage probabilities of confidence intervals. The bootstrap is consistent even if the sparsity assumption needed to obtain the asymptotic refinements reported here is not satisfied. An empirical example has illustrated the method's practical usefulness.

Table 1: Logit Coverage Probabilities for $\theta_{0n,1}$, Nominal level 0.90, $p_n = n/10$

		Full Model		Oracle Model		P-Oracle Model $C_n = 1$		P-Oracle Model $C_n = \log \log p_n$	
n	Interval	Asymp.	Boot.	Asymp.	Boot.	Asymp.	Boot.	Asymp.	Boot.
500	Lower 1-sided	0.982	0.850	0.688	0.892	0.678	0.884	0.708	0.894
1000		0.990	0.880	0.782	0.938	0.764	0.930	0.782	0.938
2000		0.976	0.842	0.820	0.906	0.812	0.906	0.820	0.906
4000		0.984	0.852	0.842	0.898	0.838	0.894	0.840	0.898
500	Upper 1-sided	1.000	0.830	0.988	0.902	0.964	0.882	0.912	0.818
1000		1.000	0.886	0.970	0.902	0.972	0.902	0.970	0.902
2000		1.000	0.824	0.956	0.886	0.956	0.888	0.956	0.886
4000		1.000	0.834	0.912	0.872	0.912	0.872	0.912	0.872
500	Symmetrical	0.982	0.780	0.796	0.890	0.764	0.866	0.734	0.828
1000		0.990	0.852	0.882	0.938	0.870	0.930	0.882	0.938
2000		0.976	0.820	0.878	0.912	0.876	0.910	0.878	0.912
4000		0.984	0.824	0.882	0.898	0.878	0.896	0.882	0.898

Table 2: Logit Coverage Probabilities for $\theta_{0n,1}$, Nominal level 0.90, $p_n = n/2$

		Full Model		Oracle Model		P-Oracle Model		P-Oracle Model	
						$C_n = 1$		$C_n = \log\log p_n$	
n	Interval	Asymp.	Boot.	Asymp.	Boot.	Asymp.	Boot.	Asymp.	Boot.
500	Lower 1-sided	1.000	0.000	0.664	0.906	0.572	0.798	0.730	0.918
1000		1.000	0.000	0.778	0.906	0.732	0.876	0.778	0.912
2000		1.000	0.000	0.812	0.916	0.786	0.902	0.812	0.916
4000		1.000	0.000	0.858	0.908	0.844	0.896	0.858	0.910
500	Upper 1-sided	1.000	1.000	0.968	0.884	0.954	0.878	0.780	0.692
1000		1.000	1.000	0.950	0.884	0.954	0.896	0.948	0.888
2000		1.000	1.000	0.972	0.914	0.974	0.918	0.972	0.916
4000		1.000	1.000	0.938	0.880	0.940	0.882	0.938	0.882
500	Symmetrical	1.000	0.000	0.806	0.892	0.688	0.772	0.668	0.738
1000		1.000	0.000	0.836	0.902	0.812	0.868	0.834	0.898
2000		1.000	0.000	0.896	0.918	0.878	0.912	0.896	0.916
4000		1.000	0.000	0.898	0.912	0.890	0.902	0.898	0.920

Table 3: Logit Coverage Probabilities for $\theta_{0n,1}$, Nominal level 0.90, $p_n = 3n/4$

		Full Model		Oracle Model		P-Oracle Model $C_n = 1$		P-Oracle Model $C_n = \log\log p_n$	
n	Interval	Asymp.	Boot.	Asymp.	Boot.	Asymp.	Boot.	Asymp.	Boot.
500	Lower 1-sided	1.000	0.000	0.660	0.886	0.558	0.760	0.740	0.894
1000		1.000	0.000	0.702	0.904	0.656	0.850	0.702	0.904
2000		1.000	0.000	0.818	0.894	0.786	0.874	0.816	0.898
4000		1.000	0.000	0.832	0.904	0.822	0.892	0.874	0.910
500	Upper 1-sided	1.000	1.000	0.970	0.894	0.958	0.902	0.726	0.618
1000		1.000	1.000	0.898	0.898	0.976	0.904	0.970	0.904
2000		1.000	1.000	0.900	0.900	0.974	0.916	0.962	0.902
4000		1.000	1.000	0.906	0.906	0.966	0.916	0.930	0.905
500	Symmetrical	1.000	0.000	0.798	0.880	0.660	0.748	0.584	0.654
1000		1.000	0.000	0.830	0.894	0.768	0.846	0.828	0.896
2000		1.000	0.000	0.868	0.904	0.856	0.884	0.868	0.906
4000		1.000	0.000	0.874	0.902	0.862	0.892	0.874	0.896

Table 4: Logit Coverage Probabilities for $\theta_{0n,2}$, Nominal level 0.90, $p_n = n/10$

		Full Model		Oracle Model		P-Oracle Model $C_n = 1$		P-Oracle Model $C_n = \log\log p_n$	
n	Interval	Asymp.	Boot.	Asymp.	Boot.	Asymp.	Boot.	Asymp.	Boot.
500	Lower 1-sided	0.996	0.810	0.958	0.910	0.948	0.908	0.886	0.864
1000		0.994	0.824	0.972	0.916	0.974	0.916	0.972	0.916
2000		0.996	0.818	0.938	0.902	0.940	0.902	0.938	0.902
4000		0.994	0.778	0.916	0.894	0.918	0.894	0.916	0.894
500	Upper 1-sided	0.994	0.842	0.760	0.924	0.726	0.890	0.672	0.822
1000		0.994	0.806	0.816	0.938	0.810	0.932	0.816	0.938
2000		0.986	0.814	0.836	0.908	0.836	0.908	0.836	0.908
4000		0.988	0.832	0.870	0.914	0.866	0.912	0.868	0.912
500	Symmetrical	0.996	0.786	0.868	0.912	0.846	0.890	0.788	0.828
1000		0.998	0.750	0.904	0.944	0.896	0.938	0.904	0.944
2000		0.998	0.752	0.904	0.912	0.904	0.912	0.904	0.912
4000		0.994	0.762	0.902	0.912	0.900	0.912	0.900	0.912

Table 5: Logit Coverage Probabilities for $\theta_{0n,2}$, Nominal level 0.90, $p_n = n/2$

		Full Model		Oracle Model		P-Oracle Model $C_n = 1$		P-Oracle Model $C_n = \log\log p_n$	
n	Interval	Asymp.	Boot.	Asymp.	Boot.	Asymp.	Boot.	Asymp.	Boot.
500	Lower 1-sided	1.000	0.948	0.708	0.902	0.602	0.824	0.638	0.806
1000		1.000	0.992	0.784	0.904	0.748	0.888	0.782	0.900
2000		1.000	1.000	0.830	0.910	0.814	0.898	0.830	0.910
4000		1.000	1.000	0.814	0.874	0.810	0.864	0.814	0.874
500	Upper 1-sided	1.000	0.052	0.972	0.908	0.984	0.918	0.912	0.858
1000		1.000	0.008	0.956	0.902	0.956	0.910	0.956	0.902
2000		1.000	0.000	0.954	0.902	0.956	0.910	0.954	0.902
4000		1.000	0.000	0.942	0.904	0.948	0.906	0.942	0.904
500	Symmetrical	1.000	0.000	0.808	0.902	0.712	0.826	0.662	0.750
1000		1.000	0.000	0.800	0.908	0.834	0.888	0.846	0.904
2000		1.000	0.000	0.892	0.916	0.882	0.902	0.892	0.916
4000		1.000	0.000	0.870	0.880	0.866	0.872	0.870	0.880

Table 6: Logit Coverage Probabilities for $\theta_{0n,2}$, Nominal level 0.90, $p_n = 3n/4$

		Full Model		Oracle Model		P-Oracle Model $C_n = 1$		P-Oracle Model $C_n = \log\log p_n$	
n	Interval	Asymp.	Boot.	Asymp.	Boot.	Asymp.	Boot.	Asymp.	Boot.
500	Lower 1-sided	1.000	0.948	0.692	0.922	0.051	0.788	0.672	0.844
1000		1.000	0.992	0.768	0.924	0.696	0.882	0.768	0.924
2000		1.000	1.000	0.796	0.892	0.772	0.896	0.796	0.892
4000		1.000	1.000	0.862	0.914	0.846	0.900	0.862	0.914
500	Upper 1-sided	1.000	0.052	0.974	0.898	0.982	0.922	0.916	0.824
1000		1.000	0.008	0.948	0.890	0.958	0.904	0.946	0.890
2000		1.000	0.000	0.950	0.904	0.954	0.920	0.950	0.904
4000		1.000	0.000	0.936	0.898	0.940	0.904	0.936	0.898
500	Symmetrical	1.000	0.000	0.822	0.912	0.638	0.786	0.704	0.790
1000		1.000	0.000	0.852	0.912	0.798	0.878	0.850	0.910
2000		1.000	0.000	0.876	0.904	0.848	0.880	0.876	0.904
4000		1.000	0.000	0.892	0.902	0.890	0.900	0.892	0.902

Table 7: Coefficients and Confidence Intervals in the Empirical Example

		P-Oracle Model			P-Oracle Model		
		$C_n = 1$			$C_n = \log \log p_n$		
Phrases	Interval	Coef.	Asymp.	Boot.	Coef.	Asymp.	Boot.
Federal regulation	Lower 1-sided	0.53	$(-\infty, 0.66)$	$(-\infty, 0.69)$	0.57	$(-\infty, 0.70)$	$(-\infty, 0.70)$
Medical liability		0.84	$(-\infty, 1.10)$	$(-\infty, 1.10)$	0.81	$(-\infty, 1.03)$	$(-\infty, 1.11)$
Gun violence		-0.64	$(-\infty, -0.50)$	$(-\infty, -0.42)$	-0.58	$(-\infty, -0.45)$	$(-\infty, -0.40)$
Death tax		0.61	$(-\infty, 0.77)$	$(-\infty, 0.91)$	0.71	$(-\infty, 0.86)$	$(-\infty, 0.99)$
Federal regulation	Upper 1-sided	0.53	$(0.41, \infty)$	$(0.39, \infty)$	0.57	$(0.47, \infty)$	$(0.45, \infty)$
Medical liability		0.84	$(0.57, \infty)$	$(0.46, \infty)$	0.81	$(0.59, \infty)$	$(0.42, \infty)$
Gun violence		-0.64	$(-0.79, \infty)$	$(-0.72, \infty)$	-0.58	$(-0.72, \infty)$	$(-0.67, \infty)$
Death tax		0.61	$(0.45, \infty)$	$(0.29, \infty)$	0.71	$(0.56, \infty)$	$(0.41, \infty)$
Federal regulation	Symmetrical	0.53	$(0.37, 0.69)$	$(0.34, 0.73)$	0.57	$(0.44, 0.71)$	$(0.41, 0.73)$
Medical liability		0.84	$(0.50, 1.18)$	$(0.41, 1.27)$	0.81	$(0.53, 1.09)$	$(0.43, 1.20)$
Gun violence		-0.64	$(-0.84, -0.45)$	$(-0.88, -0.41)$	-0.58	$(-0.76, -0.41)$	$(-0.78, -0.39)$
Death tax		0.61	$(0.40, 0.82)$	$(0.21, 1.00)$	0.71	$(0.52, 0.91)$	$(0.34, 1.09)$

References

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

Antoniadis, A. and Fan, J. (2001), 'Regularization of wavelet approximations', *Journal of the American Statistical Association* **96**(455), 939–967.

Bach, F. (2009), 'Model-Consistent Sparse Estimation through the Bootstrap', *arXiv preprint* **arXiv:0901.3202**.

Belloni, A. and Chernozhukov, V. (2011), 'ℓ_1-penalized quantile regression in high-dimensional sparse models', *The Annals of Statistics* **39**(1), 82–130.

Belloni, A., Chernozhukov, V. and Hansen, C. (2014), 'Inference on treatment effects after selection among high-dimensional controls', *The Review of Economic Studies* **81**(2), 608–650.

Bühlmann, P. (2015), Confidence intervals and tests for high-dimensional models: A compact review, *in* 'Modeling and Stochastic Learning for Forecasting in High Dimensions', Springer, pp. 21–34.

Bühlmann, P. and Van De Geer, S. (2011), *Statistics for High-Dimensional Data: Methods, Theory and Applications*, Springer Science & Business Media.

Camponovo, L. (2015), 'On the validity of the pairs bootstrap for lasso estimators', *Biometrika* **102**(4), 981–987.

Camponovo, L. (2020), 'Bootstrap inference for penalized GMM estimators with oracle properties', *Econometric Reviews* **39**(4), 362–372.

Candes, E. and Tao, T. (2007), ‘The Dantzig Selector: Statistical Estimation When p Is Much Larger than n’, *The Annals of Statistics* **35**(6), 2313–2351.

Chatterjee, A. and Lahiri, S. N. (2010), ‘Asymptotic properties of the residual bootstrap for lasso estimators’, *Proceedings of the American Mathematical Society* **138**(12), 4497–4509.

Chatterjee, A. and Lahiri, S. N. (2011), ‘Bootstrapping lasso estimators’, *Journal of the American Statistical Association* **106**(494), 608–625.

Chatterjee, A. and Lahiri, S. N. (2013), ‘Rates of convergence of the adaptive lasso estimators to the oracle distribution and higher order refinements by the bootstrap’, *The Annals of Statistics* **41**(3), 1232–1259.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W. and Robins, J. (2018), ‘Double/debiased machine learning for treatment and structural parameters’, *The Econometrics Journal* **21**(1), C1–C68.

Das, D., Chatterjee, A. and Lahiri, S. N. (2022), ‘Higher Order Accurate Symmetric Bootstrap Confidence Intervals in High Dimensional Penalized Regression’, *Department of Mathematics and Statistics, Washington University in St. Louis* **Working paper**.

Das, D., Gregory, K. and Lahiri, S. N. (2019), ‘Perturbation bootstrap in adaptive Lasso’, *The Annals of Statistics* **47**(4), 2080 – 2116.

Dezeure, R., Bühlmann, P., Meier, L. and Meinshausen, N. (2015), ‘High-Dimensional Inference: Confidence Intervals, p-Values and R-Software hdi’, *Statistical Science* **30**(4), 533 – 558.

Dezeure, R., Bühlmann, P. and Zhang, C.-H. (2017), ‘High-dimensional simultaneous inference with the bootstrap’, *TEST* **26**(4), 685–719.

Fan, J. and Li, R. (2001), 'Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties', *Journal of the American Statistical Association* **96**(456), 1348–1360.

Fan, J. and Lv, J. (2011), 'Nonconcave Penalized Likelihood With NP-Dimensionality', *IEEE Transactions on Information Theory* **57**(8), 5467–5484.

Fan, J. and Peng, H. (2004), 'Nonconcave Penalized Likelihood with a Diverging Number of Parameters', *The Annals of Statistics* **32**(3), 928–961.

Friedman, J. H., Hastie, T. and Tibshirani, R. (2010), 'Regularization Paths for Generalized Linear Models via Coordinate Descent', *Journal of Statistical Software; Vol 1, Issue 1 (2010)* .

Friedman, J., Hastie, T., Höfling, H. and Tibshirani, R. (2007), 'Pathwise Coordinate Optimization', *The Annals of Applied Statistics* **1**(2), 302–332.

Gentzkow, M., Shapiro, J. M. and Taddy, M. (2019), 'Measuring group differences in high-dimensional choices: Method and application to congressional speech', *Econometrica* **87**(4), 1307–1340.

Hall, P. (1992), *The Bootstrap and Edgeworth Expansion*, Springer Series in Statistics, Springer New York.

Horowitz, J. L. (2001), Chapter 52 The Bootstrap, Vol. 5 of *Handbook of Econometrics*, Elsevier, pp. 3159–3228.

Huang, J., Horowitz, J. L. and Ma, S. (2008), 'Asymptotic properties of bridge estimators in sparse high-dimensional regression models', *The Annals of Statistics* **36**(2), 587 – 613.

Javanmard, A. and Montanari, A. (2018), 'Debiasing the lasso: Optimal sample size for Gaussian designs', *The Annals of Statistics* **46**(6A), 2593 – 2622.

Knight, K. and Fu, W. (2000), 'Asymptotics for Lasso-Type Estimators', *The Annals of Statistics* **28**(5), 1356–1378.

Liu, H., Xu, X. and Li, J. J. (2020), 'A Bootstrap LASSO + Partial Ridge Method to Construct Confidence Intervals for Parameters in High-dimensional Sparse Linear Models', *Statistica Sinica* **30**(3), 1333–1355.

Lu, S., Liu, Y., Yin, L. and Zhang, K. (2017), 'Confidence intervals and regions for the lasso by using stochastic variational inequality techniques in optimization', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **79**(2), 589–611.

Meinshausen, N. and Yu, B. (2009), 'Lasso-type recovery of sparse representations for high-dimensional data', *The Annals of Statistics* **37**(1), 246 – 270.

Minnier, J., Tian, L. and Cai, T. (2011), 'A Perturbation Method for Inference on Regularized Regression Estimates', *Journal of the American Statistical Association* **106**(496), 1371–1382.

Pötscher, B. M. and Leeb, H. (2009), 'On the distribution of penalized maximum likelihood estimators: The LASSO, SCAD, and thresholding', *Journal of Multivariate Analysis* **100**(9), 2065–2082.

Tibshirani, R. (1996), 'Regression Shrinkage and Selection via the Lasso', *Journal of the Royal Statistical Society. Series B (Methodological)* **58**(1), 267–288.

van de Geer, S., Bühlmann, P., Ritov, Y. and Dezeure, R. (2014), 'On asymptotically optimal confidence regions and tests for high-dimensional models', *The Annals of Statistics* **42**(3), 1166–1202.

Wang, J., He, X. and Xu, G. (2020), 'Debiased inference on treatment effect in a high-dimensional model', *Journal of the American Statistical Association* **115**(529), 442–454.

Wang, L., Van Keilegom, I. and Maidman, A. (2018), 'Wild residual bootstrap inference for penalized quantile regression with heteroscedastic errors', *Biometrika* **105**(4), 859–872.

Yu, G., Yin, L., Lu, S. and Liu, Y. (2020), 'Confidence intervals for sparse penalized regression with random designs', *Journal of the American Statistical Association* **115**(530), 794–809.

Zhang, C.-H. and Huang, J. (2008), 'The Sparsity and Bias of the Lasso Selection in High-Dimensional Linear Regression', *The Annals of Statistics* **36**(4), 1567–1594.

Zhang, C.-H. and Zhang, S. S. (2014), 'Confidence intervals for low dimensional parameters in high dimensional linear models', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **76**(1), 217–242.

Zhang, X. and Cheng, G. (2017), 'Simultaneous Inference for High-Dimensional Linear Models', *Journal of the American Statistical Association* **112**(518), 757–768.

Zhao, T., Liu, H. and Zhang, T. (2018), 'Pathwise coordinate optimization for sparse learning: Algorithm and theory', *The Annals of Statistics* **46**(1), 180–218.

Zou, H. (2006), 'The adaptive lasso and its oracle properties', *Journal of the American Statistical Association* **101**(476), 1418–1429.

Zou, H. and Li, R. (2008), 'One-Step Sparse Estimates in Nonconcave Penalized Likelihood Models', *The Annals of Statistics* **36**(4), 1509–1533.

A Regularity conditions

A.1 Assumptions for Theorem 2.1

Assumption 1.

(i) $\chi_n = \{X_i : i = 1, \ldots, n\}$ is an independent random sample from the distribution of the random vector X.

(ii) For each n, Θ_n is a compact subset of $\mathbb{R}^{p_n}$.

(iii) There exists $C < \infty$ such that $\|\theta_n\|_1 \leq C$ for all $\theta_n \in \Theta_n$ and all n.

Assumption 2.

(i) $p_n = O\left(n^b\right)$ for some $0 \leq b < 1$.

(ii) $|\theta_{0nA_0,j}| \gg \lambda_n$ for each $j = 1, \ldots, p_0$.

(iii) $\lambda_n = \lambda_0 n^{-1/4+2\zeta}$, where $0 < \zeta < 1/8$ and $\lambda_0 > 0$ is a constant.

For the next set of assumptions, define the following quantities.

a. $\delta_n = n^{-d}$ and $m_n = \exp\left(n\delta_n\right)$, where $1 - 4\zeta < d < 1 - b$.

b. $\tau_n = \tau_0 n^{-1/4+\zeta}$, where τ_0 is a constant and $0 < \tau_0 < a\lambda_0$.

c. The set $V_{nv} = \{\theta_n : \|\theta_n - \theta_{0n}\|_1 \leq v\}$, where $v > 0$ is a constant.

d. $S_{n\infty}\left(\theta_n\right) = Q_{n\infty}\left(\theta_n\right) + p_{\lambda_n}\left(\theta_n\right)$, where $Q_{n\infty}$ is the non-stochastic function defined in Assumption 3 (i) below and p_{λ_n} is the SCAD penalty function.

Assumption 3.

(i) There are non-stochastic functions $Q_{n\infty}(\theta_n)$ and positive, finite constants A, ε_0, c, ℓ and n_0 such that

$$P\left[\sup_{\theta_n\in\Theta_n}\left|Q_n(\chi_n,\theta_n)-Q_{n\infty}(\theta_n)\right|>\varepsilon\right]\leq Am_n\exp\left(-cn\varepsilon^2\right)$$

for any $\ell\tau_n^2\leq\varepsilon\leq\varepsilon_0$ and $n\geq n_0$.

(ii) For each n such that $p_n>p_0$, $Q_{n\infty}$ has a (not necessarily unique) global minimum in Θ_n at a point $\left(\theta'_{0nA_0},0'_{p_n-p_0}\right)'\in\operatorname{int}(\Theta_n)$.

(iii) $Q_{n\infty}$ is weakly convex and twice continuously differentiable in a neighborhood of $\left(\theta'_{0nA_0},0'_{p_n-p_0}\right)'$.

(iv) There is a constant $\rho>0$ such that for any $v>0$ and any n,

$$\inf_{\left\|\theta_{nA_0}-\theta_{0nA_0}\right\|_2\geq v}\left[Q_{n\infty}(\theta_n)-Q_{n\infty}(\theta_{0n})\right]\geq\rho v^2.$$

For the next assumption, define

$$\begin{aligned}
H_{n,11}(\theta_n)&=\frac{\partial^2}{\partial\theta_{nA_0}\partial\theta'_{nA_0}}Q_{n\infty}(\theta_n),\\
H_{n,12}(\theta_n)&=\frac{\partial^2}{\partial\theta_{nA_0}\partial\theta'_{n\bar{A}_0}}Q_{n\infty}(\theta_n),\\
H_{n,21}(\theta_n)&=H_{n,12}(\theta_n)',\\
H_{n,22}(\theta_n)&=\frac{\partial^2}{\partial\theta_{n\bar{A}_0}\partial\theta'_{n\bar{A}_0}}Q_{n\infty}(\theta_n).
\end{aligned}$$

Note that $H_{n,12}$ is a $p_0\times(p_n-p_0)$ matrix. Let $\mu_n(\theta_n)$ denote the smallest eigenvalue of $H_{n,11}(\theta_n)$.

Assumption 4. There is a $v>0$ such that for all $\theta_n\in V_{nv}$,

(i) $\mu_n\geq\mu_0$ for all n and some $\mu_0>0$.

(ii) The components of $H_{n,21}(\theta_n) H_{n,11}^{-1}(\theta_n)$ are bounded for all n.

Assumption 5. The bootstrap provides asymptotic refinements through $O(n^{-2})$ for the quantity $P\left[T\left(\widehat{\theta}_{nA_0}^{O}\right) > \widehat{c}_{\alpha}^{O}\right]$. That is, $\left|P\left[T\left(\widehat{\theta}_{nA_0}^{O}\right) > \widehat{c}_{\alpha}^{O}\right] - \alpha\right| = O(n^{-2})$.

Assumption 1 specifies the sampling process and parameter set. Assumption 2(i) restricts the rate at which p_n can grow as n increases and rules out $p_n > n$. Chatterjee and Lahiri (2013) obtain bootstrap asymptotic refinements through $O_p\left(n^{-1/2}\right)$ with $p_n > n$ for a linear model. Fan and Lv (2011) give conditions under which $P\left(\widehat{\theta}_{n\bar{A}_0} = 0\right) = O(n^{-1})$ for a generalized linear model with $p_n > n$. This paper gives conditions under which $P\left(\widehat{A}_0 = A_0\right) = o(n^{-2})$ and the bootstrap achieves refinements through $O(n^{-2})$ for a large class of nonlinear models that contains but is not restricted to linear and generalized linear models. Assumption 2(ii) allows the components of θ_{0nA_0} to be small, but they must be larger than random sampling error. Keeping non-zero coefficients sufficiently far from zero is necessary to obtain model selection consistency and asymptotic refinements. See, for example, Pötscher and Leeb (2009); Bühlmann and Van De Geer (2011); Chatterjee and Lahiri (2011, 2013); Fan and Lv (2011); Das et al. (2019); and Das et al. (2022). Assumption 2(iii) specifies the rate of convergence to zero of the penalization parameter. Assumption 3(i) is a high-level restriction on the objective function Q_n. In typical applications, $Q_{n\infty} = E[Q_n(\chi_n, \theta_n)]$. Proposition 1 in Section B.3 of the online supplement gives conditions under which Assumption 3(i) is satisfied. Section B.3 also presents examples of models and objective functions that satisfy these conditions, including log-likelihood functions and objective functions of GMM estimation. Assumption 3(ii)-(iv) and 4(i)-(ii) place restrictions on the shapes of the functions and ensure that the non-zero parameter vector is identified. Assumption 5 applies to the oracle model. Hall (1992) gives conditions under which Assumption 5 holds when the non-zero parameters have fixed values. We assume that Assumption 5 holds when the non-zero parameters satisfy Assumption 2(ii).

A.2 Assumptions for Theorem 2.2 but not Theorem 2.1

Assumption 1S. $\left\|\theta_{0n\overline{A}_0}\right\|_1 = o\left(\tau_n\right)$ as $n \to \infty$.

Assumption 2S.

1. $\mu_n\left(\theta_n\right)$ is bounded away from 0 for all sufficiently large n uniformly over θ_n in a neighborhood of θ_{0n}.
2. The components of $H_{n,12}\left(\theta_n\right)$ are bounded for all n uniformly over θ_n in a neighborhood of θ_{0n}.

Assumption 3S.

(i) The bootstrap consistently estimates the asymptotic distribution of $T\left(\widehat{\theta}^{PT}_{nA_0}\right) = \left(\widehat{\theta}^{PT}_{nA_0,j} - \theta_{0nA_0,j}\right)/s^{PT}$. That is,

$$\sup_{z\in\mathbb{R}}\left|P^*\left[T\left(\theta^{*PT}_{nA_0}\right)\le z\right] - P\left[T\left(\widehat{\theta}^{PT}_{nA_0}\right)\le z\right]\right| \xrightarrow{p} 0.$$

(ii) $P\left(\sqrt{n}\left(\widehat{\theta}^{PT}_{nA_0} - \theta_{0nA_0}\right)\le z\right)$ is a continuous function of z.

(iii) $\sqrt{n}\cdot s^{PT}$ converges in probability to a non-stochastic, finite and positive constant.

Section B presents the proof of Theorem 2.1. Section C presents the proof of Theorem 2.2. Section D presents auxiliary results. Section E presents additional Monte Carlo results.

B Proof Of Theorem 2.1

Assumptions 1-5 from the paper hold throughout this section. Define

$$S_n\left(\chi_n,\theta_n\right) = Q_n\left(\chi_n,\theta_n\right) + p_{\lambda_n}\left(\theta_n\right)$$

and

$$S_{n\infty}\left(\chi_n,\theta_n\right)=Q_{n\infty}\left(\chi_n,\theta_n\right)+p_{\lambda_n}\left(\theta_n\right),$$

where p_{λ_n} is the SCAD penalty function.

Lemma 1. *Let $\{\theta_n\}$ be any sequence with $\theta_n\in\Theta_n$ for each n. For sufficiently large n, $S_{n\infty}\left(\theta_n\right)\geq S_{n\infty}\left(\theta_{0n}\right)$ with equality holding only if $\theta_n=\theta_{0n}$.*

Proof. Define $\delta_{nA_0,j}=\theta_{nA_0,j}-\theta_{0nA_0,j}$. Then,

$$\begin{aligned}S_{n\infty}\left(\theta_n\right)-S_{n\infty}\left(\theta_{0n}\right)&=Q_{n\infty}\left(\theta_n\right)-Q_{n\infty}\left(\theta_{0n}\right)\\&\quad+\lambda_n\sum_{j=1}^{p_0}\left[\widetilde{p}_{\lambda_n}\left(\left|\theta_{0nA_0,j}+\delta_{nA_0,j}\right|\right)-\widetilde{p}_{\lambda}\left(\left|\theta_{0nA_0,j}\right|\right)\right]\\&\quad+\lambda_n\sum_{j=p_0+1}^{p_n}\left[\widetilde{p}_{\lambda_n}\left(\left|\theta_{n\overline{A}_0,j}\right|\right)\right]\\&\geq Q_{n\infty}\left(\theta_n\right)-Q_{n\infty}\left(\theta_{0n}\right)\\&\quad-\frac{1}{2}p_0(1+a)\lambda_n^2\\&\quad+\lambda_n\sum_{j=p_0+1}^{p_n}\left[\widetilde{p}_{\lambda_n}\left(\left|\theta_{n\overline{A}_0,j}\right|\right)\right].\end{aligned}$$

If $\left|\theta_{nA_0,j}-\theta_{0nA_0,j}\right|\gg\lambda_n$ for some $j=1,\ldots,p_0$, then $\left[Q_{n\infty}\left(\theta_n\right)-Q_{n\infty}\left(\theta_{0n}\right)\right]\gg\rho\lambda_n^2$ for all sufficiently large n by Assumption 3(iv). Therefore,

$$S_{n\infty}\left(\theta_n\right)-S_{n\infty}\left(\theta_{0n}\right)\gg\lambda_n^2+\lambda_n\sum_{j=p_0+1}^{p_n}\widetilde{p}_{\lambda_n}\left(\left|\theta_{n\overline{A}_0,j}\right|\right),$$

and $S_{n\infty}\left(\theta_n\right)-S_{n\infty}\left(\theta_{0n}\right)>0$ for all sufficiently large n. If $\left|\theta_{nA_0,j}-\theta_{0nA_0,j}\right|\leq c\lambda_n$ for some $c>0$ then by Assumption 2(ii)

$$\widetilde{p}_{\lambda_n}\left(\left|\theta_{nA_0,j}\right|\right)-\widetilde{p}_{\lambda_n}\left(\left|\theta_{0nA_0,j}\right|\right)=0$$

for all $j=1,\ldots,p_0$ and sufficiently large n. Therefore,

$$S_{n\infty}\left(\theta_n\right)-S_{n\infty}\left(\theta_{0n}\right)=Q_{n\infty}\left(\theta_n\right)-Q_{n\infty}\left(\theta_{0n}\right)+\lambda_n\sum_{j=p_0+1}^{p_n}\left[\widetilde{p}_{\lambda_n}\left(\left|\theta_{n\overline{A}_0,j}\right|\right)\right]$$

$$\geq \lambda_n \sum_{j=p_0+1}^{p_n} \left[\widetilde{p}_{\lambda_n} \left(\left| \theta_{n\overline{A}_0,j} \right| \right) \right].$$

The lemma follows from the observation that $\widetilde{p}_{\lambda_n} \left(\left| \theta_{n\overline{A}_0,j} \right| \right) > 0$ if $\theta_{n\overline{A}_0,j} \neq \theta_{0n\overline{A}_0,j}$. □

Define

$$\mathcal{N}_n = \left\{ \theta \in \Theta_n : \left\| \theta - \theta_{0n} \right\|_1 < \tau_n \right\}$$

and

$$\varepsilon_n = \inf_{\theta \in \Theta_n \setminus \mathcal{N}_n} S_{n\infty}(\theta) - S_{n\infty}(\theta_{0n}).$$

It follows from Lemma 3 below that $\varepsilon_n > 0$ for all sufficiently large n. Let B_n be the event

$$\sup_{\theta \in \Theta_n} \left| S_n(\chi_n, \theta) - S_{n\infty}(\theta) \right| < \varepsilon_n / 2.$$

Lemma 2. *$\widehat{A}_0 = A_0$ for all sufficiently large n if B_n occurs.*

Proof. It follows from the definition of B_n that

$$B_n \implies S_{n\infty}\left(\widetilde{\theta}_n\right) - \frac{\varepsilon_n}{2} < S_n\left(\chi_n, \widetilde{\theta}_n\right) \tag{B.1}$$

and

$$B_n \implies S_n(\chi_n, \theta_{0n}) - \frac{\varepsilon_n}{2} < S_{n\infty}(\theta_{0n}). \tag{B.2}$$

By the definition of $\widetilde{\theta}_n$, $S_n\left(\chi_n, \widetilde{\theta}_n\right) \leq S_n(\chi_n, \theta_{0n})$. Therefore,

$$B_n \implies S_{n\infty}\left(\widetilde{\theta}_n\right) - \frac{\varepsilon_n}{2} < S_n(\chi_n, \theta_{0n}). \tag{B.3}$$

Combining (B.2) and (B.3) yields

$$B_n \implies S_{n\infty}\left(\widetilde{\theta}_n\right) - \varepsilon_n < S_{n\infty}(\theta_{0n})$$

and

$$B_n \implies S_{n\infty}\left(\widetilde{\theta}_n\right) - S_{n\infty}(\theta_{0n}) < \varepsilon_n.$$

Therefore,

$$
\begin{aligned}
B_n &\implies \left\|\widetilde{\theta}_n - \theta_{0n}\right\|_1 < \tau_n, \\
B_n &\implies \left\|\widetilde{\theta}_{n\overline{A}_0}\right\|_1 < \tau_n,
\end{aligned}
$$

and

$$B_n \implies \left|\widehat{\theta}_{nj}\right| > 2\tau_n$$

for each $j = 1, \dots, p_0$ and sufficiently large n. Moreover,

$$B_n \implies \left|\widetilde{\theta}_{nj}\right| \leq \tau_n \text{ and } \left|\widehat{\theta}_{nj}\right| = 0 \text{ for all } j = p_0 + 1, \dots, p_n.$$

It follows from the definition of τ_n that $B_n \implies \widehat{A}_0 = A_0$. □

Lemma 3. *There is a finite constant C_ε such that for all sufficiently large n,*

$$\varepsilon_n \geq C_\varepsilon \tau_n^2.$$

Proof. The proof follows from showing that $\|\theta_{A_0} - \theta_{0nA_0}\|_1 = \tau_n$ and $\theta_{\overline{A}_0} = 0$ satisfy the Karush-Kuhn-Tucker (KKT) conditions for

$$\varepsilon_n = \inf_{\theta \in \Theta_n \backslash \mathcal{N}_n} S_{n\infty}(\theta) - S_{n\infty}(\theta_{0n})$$

if n is sufficiently large.

We have

$$
\begin{aligned}
S_{n\infty}(\theta) - S_{n\infty}(\theta_{0n}) = Q_{n\infty}(\theta) - Q_{n\infty}(\theta_{0n}) + \lambda_n \sum_{j=1}^{p_0} \left[\widetilde{p}_{\lambda_n}(|\theta_j|) - \widetilde{p}_{\lambda_n}(|\theta_{0n,j}|)\right] \\
+ \lambda_n \sum_{j=p_0+1}^{p_n} \widetilde{p}_{\lambda_n}(|\theta_j|).
\end{aligned}
\tag{B.4}
$$

Define

$$h(\theta_n) = \|\theta_{nA_0} - \theta_{0nA_0}\|_1$$

$$= \sum_{j=1}^{p_0} \left[(\theta_{n,j} - \theta_{0n,j}) I (\theta_{n,j} - \theta_{0n,j} \geq 0) - (\theta_{n,j} - \theta_{0n,j}) I (\theta_{n,j} - \theta_{0n,j} < 0) \right].$$

Then

$$\frac{\partial h}{\partial \theta_{nj}} (\theta_{nj}) = s_j,$$

where

$$s_j = \begin{cases} 1 & \theta_{n,j} - \theta_{0n,j} \geq 0 \text{ and } j \in A_0 \\ -1 & \theta_{n,j} - \theta_{0n,j} < 0 \text{ and } j \in A_0 \\ 0 & j \in \overline{A}_0 \end{cases}$$

Set $s = (s_1, \ldots, s_{p_n})$. If $\|\theta_{nA_0} - \theta_{0nA_0}\| = \tau_n$ and $j \in A_0$, then $\widetilde{p}_{\lambda_n} (\theta_{nj}) - \widetilde{p}_{\lambda_n} (\theta_{0nj})$. If, in addition, $\theta_{n\overline{A}_0} = 0$, the KKT conditions for (B.4) are

$$\frac{\partial Q_{n\infty}}{\partial \theta_{nj}} (\theta_n) + v s_j = 0 \tag{B.5}$$

if $j = 1, \ldots, p_0$, where v is a Lagrangian multiplier, and

$$\left| \frac{\partial Q_{n\infty}}{\partial \theta_{nj}} (\theta_n) \right| \leq \lambda_n \tag{B.6}$$

if $j = p_0 + 1, \ldots, p_n$. By a Taylor series expansion,

$$\frac{\partial Q_{n\infty}}{\partial \theta_n} (\theta_n) = \left[\frac{\partial^2 Q_{n\infty}}{\partial \theta_n \partial \theta_n'} \left(\check{\theta}_n \right) \right] (\theta_n - \theta_{0n})$$

where $\check{\theta}_n$ is the Taylor series intermediate point and may be different in different occurrences. Therefore, (B.5) and (B.6) can be written as

$$H_{n11} \left(\check{\theta}_n \right) (\theta_{nA_0} - \theta_{0nA_0}) + v s_{A_0} = 0, \tag{B.7}$$

$$\left| \left[H_{n21} \left(\check{\theta}_n \right) (\theta_{nA_0} - \theta_{0nA_0}) \right]_j \right| \leq \lambda_n \text{ for every } j = p_0 + 1, \ldots, p_n, \tag{B.8}$$

where $s_{A_0}' = (s_1, \ldots, s_{p_0})$. By (B.7) and Assumption 4(i),

$$(\theta_{nA_0} - \theta_{0nA_0}) = -v H_{n11}^{-1} \left(\check{\theta}_n \right) s_{A_0}$$

so $\|\theta_{nA_0} - \theta_{0nA_0}\|_1 = \tau_n$ implies that

$$|v| \sum_{j=1}^{p_0} \left| \left(H_{n11}^{-1}\left(\check{\theta}_n\right) s_{A_0} \right)_j \right| = \tau_n.$$

Define

$$C_{nA_0} = \sum_{j=1}^{p_0} \left| \left(H_{n11}^{-1}\left(\check{\theta}_n\right) s_{A_0} \right)_j \right|.$$

Then $|v| = \tau_n / C_{nA_0}$ and

$$\left| \left(\theta_{nA_0} - \theta_{0nA_0} \right)_j \right| = \frac{\tau_n}{C_{nA_0}} \left| \left(H_{n11}^{-1}\left(\check{\theta}_n\right) s_{A_0} \right)_j \right|$$

for each $j = 1, \ldots, p_0$. Inequality (B.8) is

$$C_{nA_0}^{-1} \left| H_{n21}\left(\check{\theta}_n\right) H_{n11}^{-1}\left(\check{\theta}_n\right) s_{A_0} \right| \leq \frac{\lambda_n}{\tau_n}.$$

By Assumption 4(ii), this holds for all sufficiently large n because $\tau_n \ll \lambda_n$ and C_{nA_0} is bounded away from 0 for all n. It follows that the KKT conditions are satisfied.

Now

$$\begin{aligned} Q_{n\infty}\left(\theta_{nA_0}, 0_{p_n - p_0}\right) - Q_{n\infty}\left(\theta_{0n}\right) &= \frac{1}{2}\left(\theta_{nA_0} - \theta_{0nA_0}\right)' H_{n11}\left(\check{\theta}_n\right)\left(\theta_{nA_0} - A_{0nA_0}\right) \\ &= \frac{1}{2}\left(\frac{\tau_n}{C_{nA_0}}\right)^2 s_{A_0}' H_{n11}^{-1}\left(\check{\theta}_n\right) s_{A_0} \geq \frac{1}{2\mu_0}\left(\frac{\tau_n}{C_{nA_0}}\right)^2. \end{aligned}$$

Therefore,

$$\varepsilon_n \geq \frac{1}{2\mu_0}\left(\frac{\tau_n}{C_{nA_0}}\right)^2.$$

Set $C_\varepsilon = \min_n \left[1/\left(2\mu_0 C_{nA_0}^2\right) \right]$. □

Proof of Theorem 2.1. The proof consists of proving equations (2.5) and (B).

Equation (2.5): Note that

$$S_n\left(\chi_n, \widehat{\theta}_n\right) - S_{n\infty}\left(\chi_n, \widehat{\theta}_n\right) = Q_n\left(\chi_n, \widehat{\theta}_n\right) - Q_{n\infty}\left(\chi_n, \widehat{\theta}_n\right).$$

Equation (2.5) now follows from Assumption 3(i), Lemma 2 and Lemma 3.

Equation (B): By the definition of $\widehat{c}_\alpha^{PO}$,

$$P\left[T\left(\widehat{\theta}^{PO}_{n\widehat{A}_0}\right) > \widehat{c}_\alpha^{PO}\middle|\widehat{A}_0 = A_0\right] = P\left[T\left(\widehat{\theta}^{O}_{nA_0}\right) > \widehat{c}_\alpha^{O}\right].$$

By (2.5),

$$\begin{aligned}
P\left[T\left(\widehat{\theta}^{PO}_{n\widehat{A}_0}\right) > \widehat{c}_\alpha^{PO}\right] &= P\left[T\left(\widehat{\theta}^{PO}_{n\widehat{A}_0}\right) > \widehat{c}_\alpha^{PO}\middle|\widehat{A}_0 = A_0\right] P\left(\widehat{A}_0 = A_0\right) \\
&\quad + P\left[T\left(\widehat{\theta}^{PO}_{n\widehat{A}_0}\right) > \widehat{c}_\alpha^{PO}\middle|\widehat{A}_0 \neq A_0\right] P\left(\widehat{A}_0 \neq A_0\right) \\
&= P\left[T\left(\widehat{\theta}^{PO}_{n\widehat{A}_0}\right) > \widehat{c}_\alpha^{PO}\middle|\widehat{A}_0 = A_0\right] \\
&\quad + \left\{\begin{array}{c} P\left[T\left(\widehat{\theta}^{PO}_{n\widehat{A}_0}\right) > \widehat{c}_\alpha^{PO}\middle|\widehat{A}_0 \neq A_0\right] \\ -P\left[T\left(\widehat{\theta}^{PO}_{n\widehat{A}_0}\right) > \widehat{c}_\alpha^{PO}\middle|\widehat{A}_0 = A_0\right] \end{array}\right\} \times P\left(\widehat{A}_0 \neq A_0\right) \\
&= \ P\left[T\left(\widehat{\theta}^{PO}_{n\widehat{A}_0}\right) > \widehat{c}_\alpha^{PO}\middle|\widehat{A}_0 = A_0\right] + o\left(n^{-2}\right) \\
&= P\left[T\left(\widehat{\theta}^{O}_{nA_0}\right) > \widehat{c}_\alpha^{O}\right] + o\left(n^{-2}\right).
\end{aligned}$$

Therefore,

$$\left|P\left[T\left(\widehat{\theta}^{PO}_{n\widehat{A}_0}\right) > \widehat{c}_\alpha^{PO}\right] - P\left[T\left(\widehat{\theta}^{O}_{nA_0}\right) > \widehat{c}_\alpha^{O}\right]\right| = o\left(n^{-2}\right).$$

□

C Proof of Theorem 2.2

Lemma 4. *Let Assumptions 2(ii), 3(i) and 1S-3S hold. Then*

(i) $S_{n\infty}$ *is minimized at* $\left(\check{\theta}_{0nA_0}, 0_{\overline{A}_0}\right)$ *for some* $\check{\theta}_{0nA_0}$.

(ii) $\left\|\check{\theta}_{0nA_0} - \theta_{0nA_0}\right\|_1 = o\left(\lambda_n\right)$.

(iii) $\widehat{\theta}_n - \check{\theta}_{0n} \xrightarrow{\text{p}} 0$ *as* $n \to \infty$ *and* $\lim_{n\to\infty} P\left(\widehat{A}_0 = A_0\right) = 1$.

Proof. Part (i): The proof consists of showing that for all sufficiently large n, the KKT conditions for minimizing $S_{n\infty}$ are satisfied by $\theta_n = \left(\check{\theta}_{0nA_0}, 0_{\overline{A}_0}\right)$. By a Taylor series expansion,

$$Q_{n\infty}\left(\theta_n\right) = Q_{n\infty}\left(\theta_{0n}\right) + \frac{1}{2}\left(\theta_n - \theta_{0n}\right)' H_n\left(\overline{\theta}_n\right)\left(\theta_n - \theta_{0n}\right),$$

where $H_n := \frac{\partial^2 Q_n}{\partial\theta_n \partial\theta_n'}$ and $\overline{\theta}_n$ is the Taylor series intermediate point. Define $\overline{H}_n = H_n\left(\overline{\theta}_n\right)$. Then

$$\begin{aligned}&\left(\theta_n - \theta_{0n}\right)'\overline{H}_n\left(\theta_n - \theta_{0n}\right) \\ &\quad = \left[\left(\theta_{nA_0} - \theta_{0nA_0}\right)', \left(\theta_{n\overline{A}_0} - \theta_{0n\overline{A}_0}\right)'\right]\begin{bmatrix}\overline{H}_{n11} & \overline{H}_{n12} \\ \overline{H}_{n21} & \overline{H}_{n22}\end{bmatrix}\begin{bmatrix}\theta_{nA_0} - \theta_{0nA_0} \\ \theta_{n\overline{A}_0} - \theta_{0n\overline{A}_0}\end{bmatrix}\end{aligned}$$

and

$$\begin{aligned}\frac{\partial}{\partial\theta_{nA_0}}Q_{n\infty}\left(\theta_n\right) &= \overline{H}_{n11}\left(\theta_{nA_0} - \theta_{0nA_0}\right) + \overline{H}_{n12}\left(\theta_{n\overline{A}_0} - \theta_{0n\overline{A}_0}\right), \\ \frac{\partial}{\partial\theta_{n\overline{A}_0}}Q_{n\infty}\left(\theta_n\right) &= \overline{H}_{n21}\left(\theta_{nA_0} - \theta_{0nA_0}\right) + \overline{H}_{n22}\left(\theta_{n\overline{A}_0} - \theta_{0n\overline{A}_0}\right).\end{aligned}$$

The KKT conditions with $|\theta_{nA_0,j}| \gg \lambda_n$ for all $j = 1, \ldots, p_0$ and $\theta_{n\overline{A}_0} = 0$ are

$$\frac{\partial}{\partial\theta_{nA_0}}Q_{n\infty}\left(\theta_n\right) = \overline{H}_{n11}\left(\theta_{nA_0} - \theta_{0nA_0}\right) - \overline{H}_{n12}\theta_{0n\overline{A}_0} = 0 \tag{C.1}$$

$$-\lambda_n \leq \frac{\partial}{\partial\theta_{n\overline{A}_0}}Q_{n\infty}\left(\theta_n\right) = \overline{H}_{n21}\left(\theta_{nA_0} - \theta_{0nA_0}\right) - \overline{H}_{n22}\theta_{0n\overline{A}_0} \leq \lambda_n \tag{C.2}$$

component-wise. If $\theta_{n\overline{A}_0} = 0$, (C.1) gives

$$\theta_{nA_0} - \theta_{0nA_0} = \overline{H}_{n11}^{-1}\overline{H}_{n12}\theta_{0n\overline{A}_0}.$$

Therefore, by Assumption 2S,

$$\left|\theta_{nA_0,j} - \theta_{0nA_0,j}\right| \leq Mo\left(\tau_n\right) = o\left(\lambda_n\right) \tag{C.3}$$

for some $M < \infty$ and all $j \in A_0$. Condition (C.2) is

$$\left|\overline{H}_{n21}\left(\theta_{nA_0} - \theta_{0nA_0}\right) - \overline{H}_{n22}\theta_{0n\overline{A}_0}\right| \leq \lambda_n, \tag{C.4}$$

component-wise. This inequality is satisfied for all sufficiently large n because both terms on its left hand side are $o\left(\tau_n\right)$. The result follows from (C.3) and (C.4).

Part (ii): This follows from part (i) and (C.3).

Part (iii): The conclusion and proof of Lemma 2 remain unchanged after replacing θ_{0n} with $\check{\theta}_{0n} = \left(\check{\theta}_{0nA_0}, 0_{p_n - p_0}\right)$. Therefore, it follows from Assumption 3(i) that $\widehat{\theta}_n - \check{\theta}_{0n} \xrightarrow{\text{p}} 0$ and $P\left(\widehat{A}_0 = A_0\right) \to 1$ as $n \to \infty$. □

Part (i) of Lemma 4 shows that the penalization and thresholding procedure drives the small non-zero parameters to zero and replaces the true values of the large parameters with the pseudo-true values $\check{\theta}_{0nA_0}$. Part (ii) shows that the true and pseudo-true parameter values differ by $o\left(\lambda_n\right)$. Part (iii) of Lemma 4 shows that the penalization and thresholding procedure estimates the parameters of the pseudo-true model consistently and discriminates correctly between large and small parameters as $n \to \infty$.

Proof of Theorem 2.2. By Assumption 3S and part (iii) of Lemma 4, it suffices to prove that

$$\sup_z \left| P^* \left(\frac{\theta^{*PT}_{nA_0,j} - \widehat{\theta}^{PT}_{nA_0,j}}{s^{*PT}_{A_0}} \le z \right) - P \left(\frac{\widehat{\theta}^{PT}_{nA_0,j} - \theta_{0nA_0,j}}{s^{PT}} \le z \right) \right| \xrightarrow{p} 0$$

as $n \to \infty$, where $s^{*PT}_{A_0}$ is the bootstrap standard error obtained when $\widehat{A}_0$ is replaced with A_0. Now

$$\begin{aligned} P \left(\frac{\widehat{\theta}^{PT}_{nA_0,j} - \theta_{0nA_0,j}}{s^{PT}} \le z \right) &= P \left(\frac{\widehat{\theta}^{PT}_{nA_0,j} - \check{\theta}_{0nA_0,j}}{s^{PT}} + \frac{\check{\theta}_{0nA_0,j} - \theta_{0nA_0,j}}{s^{PT}} \le z \right) \\ &= P \left(\frac{\widehat{\theta}^{PT}_{nA_0,j} - \check{\theta}_{0nA_0,j}}{s^{PT}} \le z \right) + o(1) \end{aligned}$$

uniformly over $-\infty < z < \infty$. Therefore,

$$\begin{aligned} &\sup_z \left| P^* \left(\frac{\theta^{*PT}_{nA_0,j} - \widehat{\theta}^{PT}_{nA_0,j}}{s^{*PT}_{A_0}} \le z \right) - P \left(\frac{\widehat{\theta}^{PT}_{nA_0,j} - \theta_{0nA_0,j}}{s^{PT}} \le z \right) \right| \\ &= \sup_z \left| P^* \left(\frac{\theta^{*PT}_{nA_0,j} - \widehat{\theta}^{PT}_{nA_0,j}}{s^{*PT}_{A_0}} \le z \right) - P \left(\frac{\widehat{\theta}^{PT}_{nA_0,j} - \check{\theta}_{0nA_0,j}}{s^{PT}} \le z \right) \right| + o(1) \end{aligned}$$

and the result follows from Assumption 3S. □

D Sufficient Conditions for Assumption 3(i)

Proposition 1. *For each $\theta \in [0,1]^{p_n}$ let $g(X,\theta)$ be a measurable (real-valued) function of the possibly possibly vector-valued random element X. Assume:*

(i) $\{X_i : i = 1, \ldots, n\}$ *is an independent random sample from the distribution of* X.

(ii) $\theta \in \Theta_n = [0,1]^{p_n}$, $\|\theta\|_1 \leq C$ *for some* $C < \infty$, *every* n *and* $p_n = n^b$ *for some* $b < 1$.

(iii) $E[g(X,\theta)] = 0$ *and* $E\left[g(X,\theta)^2\right] \leq \sigma_g^2$ *for some constant* $\sigma_g^2 < \infty$, *all* $\theta \in \Theta_n$ *and all* n.

(iv) There is a constant $K_g < \infty$ *not depending on* n *such that for each* $\ell = 3, 4, \ldots$, $E\left[|g(X,\theta)|^\ell\right] \leq \ell! \sigma_g^2 K_g^{\ell-2}$ *for all* $\theta \in \Theta_n$ *and all* n.

(v) For each n, *there is a function* $M_n(X)$ *such that*

$$|g(X,\theta_1) - g(X,\theta_2)| \leq M_n(X) \|\theta_1 - \theta_2\|_1$$

for all $\theta_1, \theta_2 \in \Theta_n$. *Moreover, there are finite constants* M^* *and* K_M *not depending on* n *such that* $|E[M_n(X)]| \leq M^*$ *and* $E\left[|M_n(X) - E[M_n(X)]|^\ell\right] \leq \ell! \sigma_{M_n}^2 K_M^{\ell-2}$ *for every* $\ell = 3, 4, \ldots$ *and each* n, *where* $\sigma_{M_n}^2 = \mathrm{Var}[M_n(X)]$ *and for some finite and positive* m_0, M_0, *and* $m_0 \leq \sigma_{M_n}^2 \leq M_0$.

Then, there are finite constants $c > 0$, $\varepsilon_0 > 0$, $n_0 > 0$ *and* $\ell > 0$ *such that*

$$P\left[\sup_{\theta \in \Theta_n} \left| \frac{1}{n} \sum_{i=1}^{n} g(X_i,\theta) \right| > \varepsilon \right] \leq 3 m_n \exp\left(-cn\varepsilon^2\right)$$

for all $n > n_0$ *if* $\ell \tau_n^2 \leq \varepsilon < \varepsilon_0$.

Proof. Define m_n as in Section A.1 of the appendix. Divide Θ_n into m_n hypercubic cells whose edges have lengths $d_n = m_n^{-\frac{1}{p_n}}$. Denote the cells by Θ_{nj} for $j \in \{1, \ldots, m_n\}$. Let $\theta_n^{(j)}$ be a point in the interior of Θ_{nj}. Let $\varepsilon > 0$ be given. Define

$$P_n = P\left[\sup_{\theta \in \Theta_n} \left| \frac{1}{n} \sum_{i=1}^{n} g\left(X_i, \theta\right) \right| > \varepsilon \right].$$

Then

$$\begin{aligned}
P_n &= P\left[\max_{1 \le j \le m_n} \left\{ \sup_{\theta \in \Theta_{nj}} \left| \frac{1}{n} \sum_{i=1}^{n} g\left(X_i, \theta\right) \right| \right\} > \varepsilon \right] \\
&= P\left[\max_{1 \le j \le m_n} \left\{ \sup_{\theta \in \Theta_{nj}} \left| \frac{1}{n} \sum_{i=1}^{n} g\left(X_i, \theta_n^{(j)}\right) + \frac{1}{n} \sum_{i=1}^{n} \left[g\left(X_i, \theta\right) - g\left(X_i, \theta_n^{(j)}\right) \right] \right| \right\} > \varepsilon \right] \\
&\le P\left[\max_{1 \le j \le m_n} \left\{ \left| \frac{1}{n} \sum_{i=1}^{n} g\left(X_i, \theta_n^{(j)}\right) \right| + \sup_{\theta \in \Theta_{nj}} \left| \frac{1}{n} \sum_{i=1}^{n} \left[g\left(X_i, \theta\right) - g\left(X_i, \theta_n^{(j)}\right) \right] \right| \right\} > \varepsilon \right] \\
&\le P\left[\max_{1 \le j \le m_n} \left| \frac{1}{n} \sum_{i=1}^{n} g\left(X_i, \theta_n^{(j)}\right) \right| > \frac{\varepsilon}{2} \right] \\
&\quad + P\left[\max_{1 \le j \le m_n} \sup_{\theta \in \Theta_{nj}} \left| \frac{1}{n} \sum_{i=1}^{n} \left[g\left(X_i, \theta\right) - g\left(X_i, \theta_n^{(j)}\right) \right] \right| > \frac{\varepsilon}{2} \right] \\
&\equiv P_{n1} + P_{n2}.
\end{aligned}$$

Consider P_{n1}. By Bernstein's inequality

$$P_{n1} = P\left[\max_{1 \le j \le m_n} \left| \frac{1}{n} \sum_{i=1}^{n} g\left(X_i, \theta_n^{(j)}\right) \right| > \frac{\varepsilon}{2} \right] \le 2 m_n \exp\left(-\frac{n \varepsilon^2}{32 \sigma_g^2} \right)$$

if $\varepsilon < \frac{4\sigma_g^2}{K_g}$.

Now consider P_{n2}. By assumption (v)

$$\begin{aligned}
P_{n2} &= P\left[\max_{1 \le j \le m_n} \sup_{\theta \in \Theta_{nj}} \left| \frac{1}{n} \sum_{i=1}^{n} \left[g\left(X_i, \theta\right) - g\left(X_i, \theta_n^{(j)}\right) \right] \right| > \frac{\varepsilon}{2} \right] \\
&\le P\left[\max_{1 \le j \le m_n} \sup_{\theta \in \Theta_{nj}} \frac{1}{n} \sum_{i=1}^{n} \left| g\left(X_i, \theta\right) - g\left(X_i, \theta_n^{(j)}\right) \right| > \frac{\varepsilon}{2} \right] \\
&\le P\left[\max_{1 \le j \le m_n} \sup_{\theta \in \Theta_{nj}} \frac{1}{n} \sum_{i=1}^{n} M_n\left(X_i\right) \left\| \theta - \theta_n^{(j)} \right\|_1 > \frac{\varepsilon}{2} \right]
\end{aligned}$$

$$
\begin{aligned}
&\leq m_n P\left[d_n p_n \frac{1}{n}\sum_{i=1}^{n} M_n\left(X_i\right) > \frac{\varepsilon}{2}\right] \\
&= m_n P\left[\frac{1}{n}\sum_{i=1}^{n} M_n\left(X_i\right) > \frac{\varepsilon}{2 d_n p_n}\right].
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
P_{n2} &\leq m_n P\left[\frac{1}{n}\sum_{i=1}^{n}\left[M_n\left(X_i\right) - E\left[M_n(X)\right]\right] > \frac{\varepsilon}{2 d_n p_n} - E\left[M_n(X)\right]\right] \\
&\leq m_n P\left[\frac{1}{n}\sum_{i=1}^{n}\left[M_n\left(X_i\right) - E\left[M_n(X)\right]\right] > \frac{\varepsilon}{2 d_n p_n} - M^*\right].
\end{aligned}
$$

By Bernstein's inequality,

$$
\begin{aligned}
P_{n2} &\leq m_n \exp\left\{-\frac{n\left[\left(\varepsilon/\left(2 d_n p_n\right)\right) - M^*\right]^2}{4 M_0 + 2 K_M\left[\left(\varepsilon/\left(2 d_n p_n\right)\right) - M^*\right]}\right\} \\
&\leq m_n \exp\left\{-\frac{n\left[\left(\varepsilon/\left(2 d_n p_n\right)\right)\right]^2}{4 M_0 + 2 K_M\left[\left(\varepsilon/\left(2 d_n p_n\right)\right)\right]}\right\}.
\end{aligned}
$$

if $\varepsilon > 4 d_n p_n M^*$. Now let $\varepsilon > 4 d_n p_n \max\left\{M^*, M_0/K_M\right\}$. Then

$$
P_{n2} \leq 2 m_n \exp\left[-\frac{n\varepsilon^2}{32 K_M d_n p_n}\right].
$$

Let n_0 be the smallest value of n such that

$$
\left[4 d_n p_n \max\left\{M^*, M_0/K_M\right\}\right]^{\frac{1}{2}} < \ell\tau_n^2 < \sigma_g^2/\left(K_M d_n p_n\right)
$$

for some $\ell > 0$. Then if $n > n_0$ and $\ell\tau_n^2 \leq \varepsilon < \sigma_g^2/\left(K_M d_n p_n\right)$,

$$
P_{n2} \leq 2 m_n \left(-\frac{n\varepsilon^2}{32\sigma_g^2}\right).
$$

Set

$$
\varepsilon_0 = \min\left\{\frac{4\sigma_g^2}{K_g}, \frac{\sigma_g^2}{K_M d_{n_0} p_{n_0}}\right\}
$$

and $c = \frac{1}{32\sigma_g^2}$. □

D.1 Examples of Models that Satisfy Assumption 3(i)

Example 1. Penalized least squares estimation of a linear model. The model is

$$Y_i = \sum_{j=1}^{p_n} \theta_{0n,j} X_{ij} + U_i = X_i' \theta_{0n} + U_i; \quad i = 1, \ldots, n \tag{D.1}$$

where the X_{ij}'s are random variables, $\|X_i\|_1 \leq M$ for some $M < \infty$, all $i = 1, \ldots, n$ and all $j = 1, \ldots, p_n$; the U_i's are independent and identically distributed sub-Gaussian random variables; $E\left[U_i | X_i\right] = 0$ and $E\left[U_i^2\right] = \sigma^2$ for all $i = 1, \ldots, n$. Assume that $\|\theta\|_1 \leq M$ for all $\theta \in \Theta_n$ and all n. Also assume that $p_n = O\left(n^b\right)$ for some $0 \leq b < 1$. Let

$$Q_n\left(\chi_n, \theta\right) = \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - X_i' \theta\right)^2 = \frac{1}{n} \sum_{i=1}^{n} \left[U_i - X_i'\left(\theta - \theta_{0n}\right)\right]^2,$$

and

$$Q_{n\infty}(\theta) = E\left[Q_n\left(\chi_n, \theta\right)\right] = \sigma^2 + \left(\theta - \theta_{0n}\right)' \Sigma_{XX} \left(\theta - \theta_{0n}\right),$$

where $\Sigma_{XX} = E\left[XX'\right]$. In the notation of Proposition 1, the function g here is

$$g(X, U, \theta) = \left(U^2 - \sigma^2\right) - 2UX'\left(\theta - \theta_{0n}\right) + \left(\theta - \theta_{0n}\right)'\left(XX' - \Sigma_{XX}\right)\left(\theta - \theta_{0n}\right).$$

We show that model (D.1) satisfies the conditions of Proposition 1. Conditions (i), (ii), and (iii) are satisfied by the definition of the model and by the arguments below for condition (iv) with $\ell = 2$. Condition (iv): Let $X(j)$ denote the j^{th} component of X. Then,

$$\begin{aligned} g(X, U, \theta) \leq & \left|U^2 - \sigma^2\right| + 2|U| \left|X'\left(\theta - \theta_{0n}\right)\right| \\ & + \left|\left(\theta - \theta_{0n}\right)'\left(XX' - \Sigma_{XX}\right)\left(\theta - \theta_{0n}\right)\right|, \end{aligned}$$

$$\begin{aligned} \left|\left(\theta - \theta_{0n}\right)' X\right| \leq & \left|\sum_{j=1}^{p_n} \left(\theta_{n,j} - \theta_{0n,j}\right) X(j)\right| \\ \leq & \sum_{j=1}^{p_n} \left|\theta_{n,j} - \theta_{0n,j}\right| |X(j)| \\ \leq & M \left\|\theta - \theta_{0n}\right\|_1 \end{aligned}$$

$$
\begin{aligned}
&\leq 2M^2,\\
(\theta-\theta_{0n})' XX' (\theta-\theta_{0n}) &\leq 4M^4,\\
(\theta-\theta_{0n})' \Sigma_{XX} (\theta-\theta_{0n}) &\leq 4M^4,
\end{aligned}
$$

so that

$$|g(X,U,\theta)| \leq \left|U^2-\sigma^2\right| + 4|U|M^2 + 8M^4.$$

Therefore,

$$
\begin{aligned}
E\left[|g(X,U,\theta)|^\ell\right] \leq& E\left[\left|\left|U^2-\sigma^2\right| + 4|U|M^2 + 8M^4\right|^\ell\right]\\
&\leq 2^\ell E\left[\left|U^2-\sigma^2\right|^\ell\right] + 2^\ell E\left[\left(4M^2|U| + 8M^4\right)^\ell\right]\\
&\leq 2^\ell E\left[\left|U^2-\sigma^2\right|^\ell\right] + 2^{4\ell}M^{2\ell}E\left[|U|^\ell\right] + 2^{2\ell}\left(8M^4\right)^\ell.
\end{aligned}
$$

The first and second terms on the right-hand side of the inequality satisfy condition (iv) because U is sub-Gaussian and $U^2-\sigma^2$ is sub-exponential. The third term satisfies condition (iv) because it is a constant.

Condition (v): $g(X,U,\theta)$ is continuously differentiable with respect to θ. Therefore,

$$|g(X,U,\theta_2) - g(X,U,\theta_1)| \leq \left|\frac{\partial}{\partial\theta} g\left(X,U,\widetilde{\theta}\right)(\theta_2-\theta_1)\right|.$$

where $\widetilde{\theta}$ is the Taylor series intermediate point.

$$
\begin{aligned}
\frac{\partial}{\partial\theta} g\left(X,U,\widetilde{\theta}\right) &= -2\left[X'U - \left(\widetilde{\theta}-\theta_{0n}\right)'(XX'-\Sigma_{XX})\right]\\
\left|\frac{\partial}{\partial\theta} g\left(X,U,\widetilde{\theta}\right)(\theta_2-\theta_1)\right| &\leq 2M^2|U| + 8M^4.
\end{aligned}
$$

Condition (v) is satisfied because U is sub-Gaussian.

Example 2. A binary logit model with normally distributed random coefficients. Let $\{Y_i, X_i : i = 1, \ldots, n\}$ be independently and identically distributed realizations of the binary random variable Y supported in $\{0,1\}$ and the $d_X \times 1$ random vector X. Let X_{ij} denote the j^{th} component of X_i for each (i,j), and assume that $\|X\|_1 \leq M$ for some $M < \infty$. Define $\theta = \text{vec}(\beta, C)$ where C is a $d_X \times d_X$ Cholesky factorization matrix. The model is

$$P\left(Y_i = 1 | X_i, \theta\right) = \int \left\{ \frac{\exp\left[(\beta + C\varepsilon)' X_i\right]}{1 + \exp\left[(\beta + C\varepsilon)' X_i\right]} \right\} f(\varepsilon) \mathrm{d}\varepsilon, \tag{D.2}$$

where f is the $\mathcal{N}\left(0, I_{d_X}\right)$ probability density function. In the notation of Proposition 1, $g(x,\theta) = P\left(Y = 1 | X = x, \theta\right)$. Assume that $g(x,\theta)$ is bounded away from 0 and 1 for all $x \in \text{support}(X)$ and all $\theta \in \Theta_n$. The log-likelihood function for estimating θ is

$$Q_n\left(\chi_n, \theta\right) = \frac{1}{n} \sum_{i=1}^{n} \left\{ Y_i \log g\left(X_i, \theta\right) + \left(1 - Y_i\right) \log\left[1 - g\left(X_i, \theta\right)\right] \right\}.$$

Let

$$\begin{aligned} Q_{n\infty}(\theta) &= E\left[P(Y=1|X) \cdot \log g(X,\theta) + P(Y=0|X) \log\left[1 - g(X,\theta)\right]\right] \\ &= E\left[g\left(X, \theta_{0n}\right) \cdot \log g(X,\theta) + \left[1 - g\left(X, \theta_{0n}\right)\right] \log\left[1 - g(X,\theta)\right]\right] \end{aligned}$$

where $\theta_{0n} \in \Theta_n$. We show that model (D.2) satisfies the conditions of Proposition 1.

Conditions (i) and (iii) are satisfied by the definition of the model. Conditions (iii) and (iv) are satisfied because $\left|Y_i \log g\left(X_i, \theta\right) + \left(1 - Y_i\right) \log\left[1 - g\left(X_i, \theta\right)\right]\right|$ is bounded for all $i = 1, \ldots, n$.

Condition (v): This condition is satisfied if $\left|g\left(X, \theta_2\right) - g\left(X, \theta_1\right)\right| \leq M_g \left\|\theta_2 - \theta_1\right\|_1$, where $M_g < \infty$ is a constant. To establish this inequality, define

$$\pi(x, \varepsilon, \theta) = \frac{\exp\left[(\beta + C\varepsilon)' x\right]}{1 + \exp\left[(\beta + C\varepsilon)' x\right]}.$$

Then

$$\frac{\partial}{\partial \beta_j} \pi = X_j \pi (1 - \pi)$$

and

$$\frac{\partial}{\partial C_{jk}} \pi = \varepsilon_j X_k \pi (1 - \pi).$$

A Taylor series expansion gives

$$\begin{aligned}
|g(X,\theta_2) - g(X,\theta_1)| &\leq \int |\pi(X,\varepsilon,\theta_2) - \pi(X,\varepsilon,\theta_1)| \, f(\varepsilon)\, \mathrm{d}\varepsilon \\
&\leq \int \sum_{j=1}^{\dim(\theta)} \left| \frac{\partial}{\partial \theta_j} \pi\left(X,\varepsilon,\widetilde{\theta}\right) \right| |\theta_{1j} - \theta_{2j}| \, f(\varepsilon)\, \mathrm{d}\varepsilon \\
&= \sum_{j=1}^{\dim(\theta)} |\theta_{1j} - \theta_{2j}| \int \left| \frac{\partial}{\partial \theta_j} \pi\left(X,\varepsilon,\widetilde{\theta}\right) \right| f(\varepsilon)\, \mathrm{d}\varepsilon,
\end{aligned}$$

where $\widetilde{\theta}$ is the Taylor series intermediate point. Let $\mu = (2/\pi)^{\frac{1}{2}}$ denote the first absolute moment of the $\mathcal{N}(0,1)$ distribution. Because $\pi(1-\pi) \leq 0.25$ and $|X_{ij}| \leq M$,

$$\begin{aligned}
|g(X,\theta_2) - g(X,\theta_1)| &\leq 0.25M \int \left(\sum_{j=1}^{d_X} |\beta_{2j} - \beta_{1j}| + \sum_{j,k=1}^{d_X} |\varepsilon_j| \, |C_{2jk} - C_{1jk}| \right) f(\varepsilon) \mathrm{d}\varepsilon \\
&\leq 0.25M \left\| \beta_2 - \beta_1 \right\|_1 + 0.25M\mu \sum_{j,k=1}^{d_X} |C_{2jk} - C_{1jk}| \\
&\leq 0.25M \left\| \beta_2 - \beta_1 \right\|_1 + 0.25M \sum_{j,k=1}^{d_X} |C_{2jk} - C_{1jk}| .
\end{aligned}$$

Therefore,

$$|g(X,\theta_2) - g(X,\theta_1)| \leq 0.25M \left\| \theta_2 - \theta_1 \right\|_1 .$$

Set $M_g = 0.25M$.

Example 3. A Generalized Method of Moments Estimator with a Fixed Weight Matrix. Let

$$Q_n(\chi_n, \theta) = \left[\frac{1}{n} \sum_{i=1}^{n} g(X_i, \theta) \right]' \Omega \left[\frac{1}{n} \sum_{i=1}^{n} g(X_i, \theta) \right]$$

and

$$Q_{n\infty}(\theta) = E[g(X,\theta)]' \Omega E[g(X,\theta)],$$

where g is a $q \times 1$ vector-valued function and Ω is a $q \times q$ positive definite symmetrical matrix of finite constants. Assume each component of g satisfies the conditions of Proposition 1. Then Assumption 3(i) is satisfied.

Example 4. A Generalized Method of Moments Estimator with the Continuous Updating Estimate of the Asymptotically Optimal Weight Matrix.

Use the notation of Example 3 and assume that $\Omega_0(\theta) = \mathrm{Var}[g(X,\theta)]$ is positive definite with bounded eigenvalues for all $\theta \in \Theta_n$ and n. Also assume that the components of $\Omega_0(\theta)$ satisfy $|\Omega_{0,ij}(\theta)| \leq M$ for some $M < \infty$ and all $\theta \in \Theta_n$, i, j and n. Then

$$Q_n\left(\chi_n, \theta\right) = \left[\frac{1}{n}\sum_{i=1}^{n} g\left(X_i, \theta\right)\right]' \Omega_n^{-1}(\theta) \left[\frac{1}{n}\sum_{i=1}^{n} g\left(X_i, \theta\right)\right]$$

where

$$\begin{aligned} \Omega_n(\theta) &= \frac{1}{n}\sum_{i=1}^{n} g\left(X_i, \theta\right) g\left(X_i, \theta\right)' - \overline{g}_n(\theta)\overline{g}_n(\theta) \\ \overline{g}_n(\theta) &= \frac{1}{n}\sum_{i=1}^{n} g\left(X_i, \theta\right) \end{aligned}$$

and

$$Q_{n\infty}(\theta) = E[g(X,\theta)]'\Omega_0^{-1}(\theta)E[g(X,\theta)].$$

Let each component of g satisfy the conditions of Proposition 1. Then Assumption 3(i) holds.

E Logit Confidence Intervals

This section presents confidence intervals for the logit model of Section 3 averaged over Monte Carlo replications. Confidence intervals for the full model are not shown because, as Tables 1-6 show, the differences between their true and nominal coverage probabilities are very large in most cases.

Table 8: Logit Confidence Intervals for $\theta_{0n,1}$, Nominal level 0.90, $p_n = n/10$

		Oracle		P-Oracle		P-Oracle	
				$C_n = 1$		$C_n = \log\log p_n$	
n	Interval	Asymp.	Boot.	Asymp.	Boot.	Asymp.	Boot.
500	Lower 1-sided	$(-\infty, 5.74)$	$(-\infty, 5.02)$	$(-\infty, 5.72)$	$(-\infty, 5.01)$	$(-\infty, 5.54)$	$(-\infty, 4.87)$
1000		$(-\infty, 4.88)$	$(-\infty, 4.61)$	$(-\infty, 4.91)$	$(-\infty, 4.63)$	$(-\infty, 4.88)$	$(-\infty, 4.61)$
2000		$(-\infty, 4.52)$	$(-\infty, 4.40)$	$(-\infty, 4.53)$	$(-\infty, 4.40)$	$(-\infty, 4.52)$	$(-\infty, 4.40)$
4000		$(-\infty, 4.34)$	$(-\infty, 4.28)$	$(-\infty, 4.34)$	$(-\infty, 4.28)$	$(-\infty, 4.34)$	$(-\infty, 4.29)$
500	Upper 1-sided	$(3.79, \infty)$	$(3.29, \infty)$	$(3.78, \infty)$	$(3.28, \infty)$	$(3.67, \infty)$	$(3.20, \infty)$
1000		$(3.73, \infty)$	$(3.48, \infty)$	$(3.75, \infty)$	$(3.50, \infty)$	$(3.73, \infty)$	$(3.48, \infty)$
2000		$(3.75, \infty)$	$(3.64, \infty)$	$(3.76, \infty)$	$(3.64, \infty)$	$(3.75, \infty)$	$(3.64, \infty)$
4000		$(3.81, \infty)$	$(3.75, \infty)$	$(3.81, \infty)$	$(3.75, \infty)$	$(3.81, \infty)$	$(3.75, \infty)$
500	Symmetrical	$(3.51, 6.01)$	$(3.27, 6.25)$	$(3.50, 6.00)$	$(3.27, 6.23)$	$(3.40, 5.80)$	$(3.18, 6.02)$
1000		$(3.56, 5.05)$	$(3.46, 5.15)$	$(3.58, 5.08)$	$(3.48, 5.18)$	$(3.56, 5.05)$	$(3.46, 5.15)$
2000		$(3.64, 4.63)$	$(3.61, 4.66)$	$(3.65, 4.64)$	$(3.61, 4.67)$	$(3.64, 4.63)$	$(3.61, 4.66)$
4000		$(3.73, 4.41)$	$(3.72, 4.42)$	$(3.73, 4.41)$	$(3.72, 4.43)$	$(3.73, 4.41)$	$(3.72, 4.42)$

Table 9: Logit Confidence Intervals for $\theta_{0n,1}$, Nominal level 0.90, $p_n = n/2$

		Oracle		P-Oracle $C_n = 1$		P-Oracle $C_n = \log\log p_n$	
n	Interval	Asymp.	Boot.	Asymp.	Boot.	Asymp.	Boot.
500	Lower 1-sided	$(-\infty, 5.75)$	$(-\infty, 5.04)$	$(-\infty, 6.06)$	$(-\infty, 5.25)$	$(-\infty, 5.15)$	$(-\infty, 4.57)$
1000		$(-\infty, 4.89)$	$(-\infty, 4.61)$	$(-\infty, 4.95)$	$(-\infty, 4.66)$	$(-\infty, 4.88)$	$(-\infty, 4.60)$
2000		$(-\infty, 4.52)$	$(-\infty, 4.40)$	$(-\infty, 4.55)$	$(-\infty, 4.42)$	$(-\infty, 4.52)$	$(-\infty, 4.40)$
4000		$(-\infty, 4.35)$	$(-\infty, 4.29)$	$(-\infty, 4.36)$	$(-\infty, 4.39)$	$(-\infty, 4.35)$	$(-\infty, 4.29)$
500	Upper 1-sided	$(3.80, \infty)$	$(3.29, \infty)$	$(3.95, \infty)$	$(3.41, \infty)$	$(3.44, \infty)$	$(3.02, \infty)$
1000		$(3.73, \infty)$	$(3.48, \infty)$	$(3.77, \infty)$	$(3.51, \infty)$	$(3.72, \infty)$	$(3.48, \infty)$
2000		$(3.75, \infty)$	$(3.64, \infty)$	$(3.77, \infty)$	$(3.65, \infty)$	$(3.75, \infty)$	$(3.64, \infty)$
4000		$(3.82, \infty)$	$(3.76, \infty)$	$(3.83, \infty)$	$(3.77, \infty)$	$(3.82, \infty)$	$(3.76, \infty)$
500	Symmetrical	$(3.52, 6.03)$	$(3.28, 6.28)$	$(3.66, 6.35)$	$(3.40, 6.61)$	$(3.20, 5.40)$	$(3.00, 5.59)$
1000		$(3.56, 5.05)$	$(3.46, 5.15)$	$(3.60, 5.11)$	$(3.49, 5.22)$	$(3.56, 5.05)$	$(3.46, 5.15)$
2000		$(3.64, 4.63)$	$(3.61, 4.17)$	$(3.66, 4.65)$	$(3.63, 4.69)$	$(3.64, 4.63)$	$(3.61, 4.67)$
4000		$(3.74, 4.43)$	$(3.73, 4.44)$	$(3.75, 4.44)$	$(3.72, 4.43)$	$(3.74, 4.45)$	$(3.73, 4.44)$

Table 10: Logit Confidence Intervals for $\theta_{0n,1}$, Nominal level 0.90, $p_n = 3n/4$

		Oracle		P-Oracle $C_n = 1$		P-Oracle $C_n = \log\log p_n$	
n	Interval	Asymp.	Boot.	Asymp.	Boot.	Asymp.	Boot.
500	Lower 1-sided	$(-\infty, 5.81)$	$(-\infty, 5.06)$	$(-\infty, 6.39)$	$(-\infty, 5.48)$	$(-\infty, 5.17)$	$(-\infty, 4.57)$
1000		$(-\infty, 4.89)$	$(-\infty, 4.61)$	$(-\infty, 5.03)$	$(-\infty, 4.73)$	$(-\infty, 4.89)$	$(-\infty, 4.61)$
2000		$(-\infty, 4.56)$	$(-\infty, 4.44)$	$(-\infty, 4.59)$	$(-\infty, 4.47)$	$(-\infty, 4.56)$	$(-\infty, 4.44)$
4000		$(-\infty, 4.33)$	$(-\infty, 4.27)$	$(-\infty, 4.44)$	$(-\infty, 4.28)$	$(-\infty, 4.33)$	$(-\infty, 4.27)$
500	Upper 1-sided	$(3.83, \infty)$	$(3.32, \infty)$	$(4.12, \infty)$	$(3.55, \infty)$	$(3.45, \infty)$	$(3.03, \infty)$
1000		$(3.73, \infty)$	$(3.48, \infty)$	$(3.83, \infty)$	$(3.56, \infty)$	$(3.73, \infty)$	$(3.48, \infty)$
2000		$(3.78, \infty)$	$(3.67, \infty)$	$(3.81, \infty)$	$(3.69, \infty)$	$(3.78, \infty)$	$(3.67, \infty)$
4000		$(3.80, \infty)$	$(3.74, \infty)$	$(3.81, \infty)$	$(3.75, \infty)$	$(3.80, \infty)$	$(3.74, \infty)$
500	Symmetrical	$(3.55, 6.09)$	$(3.31.6.33)$	$(3.79, 6.71)$	$(3.54, 6.96)$	$(3.20, 5.41)$	$(3.02, 6.00)$
1000		$(3.56, 5.05)$	$(3.46, 5.15)$	$(3.65, 5.20)$	$(3.54, 5.32)$	$(3.56, 5.05)$	$(3.46, 5.15)$
2000		$(3.67, 4.67)$	$(3.63, 4.71)$	$(3.70, 4.70)$	$(3.66, 4.69)$	$(3.67, 4.67)$	$(3.63, 4.71)$
4000		$(3.72, 4.40)$	$(3.71, 4.41)$	$(3.73, 4.41)$	$(3.72, 4.43)$	$(3.72, 4.40)$	$(3.72, 4.41)$

Table 11: Logit Confidence Intervals for $\theta_{0n,2}$, Nominal level 0.90, $p_n = n/10$

		Oracle		P-Oracle $C_n = 1$		P-Oracle $C_n = \log\log p_n$	
n	Interval	Asymp.	Boot.	Asymp.	Boot.	Asymp.	Boot.
500	Lower 1-sided	$(-\infty, -1.27)$	$(-\infty, -1.04)$	$(-\infty, -1.33)$	$(-\infty, -1.10)$	$(-\infty, -1.43)$	$(-\infty, -1.20)$
1000		$(-\infty, -1.30)$	$(-\infty, -1.20)$	$(-\infty, -1.31)$	$(-\infty, -1.20)$	$(-\infty, -1.30)$	$(-\infty, -1.20)$
2000		$(-\infty, -1.35)$	$(-\infty, -1.30)$	$(-\infty, -1.35)$	$(-\infty, -1.30)$	$(-\infty, -1.35)$	$(-\infty, -1.30)$
4000		$(-\infty, -1.38)$	$(-\infty, -1.36)$	$(-\infty, -1.38)$	$(-\infty, -1.36)$	$(-\infty, -1.38)$	$(-\infty, -1.36)$
500	Upper 1-sided	$(-2.29, \infty)$	$(-2.03, \infty)$	$(-2.37, \infty)$	$(-2.10, \infty)$	$(-2.48, \infty)$	$(-2.21, \infty)$
1000		$(-1.93, \infty)$	$(-1.83, \infty)$	$(-1.94, \infty)$	$(-1.84, \infty)$	$(-1.93, \infty)$	$(-1.83, \infty)$
2000		$(-1.77, \infty)$	$(-1.73, \infty)$	$(-1.77, \infty)$	$(-1.73, \infty)$	$(-1.77, \infty)$	$(-1.73, \infty)$
4000		$(-1.67, \infty)$	$(-1.65, \infty)$	$(-1.67, \infty)$	$(-1.65, \infty)$	$(-1.67, \infty)$	$(-1.65, \infty)$
500	Symmetrical	$(-2.44, -1.12)$	$(-2.53, -1.03)$	$(-2.51, -1.18)$	$(-2.61, -1.08)$	$(-2.62, -1.28)$	$(-2.72, -1.19)$
1000		$(-2.02, -1.21)$	$(-2.06, -1.18)$	$(-2.03, -1.22)$	$(-2.07, -1.18)$	$(-2.02, -1.21)$	$(-2.05, -1.76)$
2000		$(-1.83, -1.29)$	$(-1.84, -1.27)$	$(-1.83, -1.29)$	$(-1.84, -1.28)$	$(-1.83, -1.29)$	$(-1.84, -1.27)$
4000		$(-1.72, -1.34)$	$(-1.72, -1.33)$	$(-1.72, -1.34)$	$(-1.72, -1.33)$	$(-1.72, -1.34)$	$(-1.72, -1.33)$

Table 12: Logit Confidence Intervals for $\theta_{0n,2}$, Nominal level 0.90, $p_n = n/2$

		Oracle		P-Oracle $C_n = 1$		P-Oracle $C_n = \log\log p_n$	
n	Interval	Asymp.	Boot.	Asymp.	Boot.	Asymp.	Boot.
500	Lower 1-sided	$(-\infty, -1.26)$	$(-\infty, -1.03)$	$(-\infty, -1.44)$	$(-\infty, -1.18)$	$(-\infty, -1.62)$	$(-\infty, -1.39)$
1000		$(-\infty, -1.29)$	$(-\infty, -1.19)$	$(-\infty, -1.31)$	$(-\infty, -1.20)$	$(-\infty, -1.30)$	$(-\infty, -1.20)$
2000		$(-\infty, -1.33)$	$(-\infty, -1.29)$	$(-\infty, -1.34)$	$(-\infty, -1.29)$	$(-\infty, -1.33)$	$(-\infty, -1.29)$
4000		$(-\infty, -1.38)$	$(-\infty, -1.36)$	$(-\infty, -1.38)$	$(-\infty, -1.36)$	$(-\infty, -1.38)$	$(-\infty, -1.36)$
500	Upper 1-sided	$(-2.28, \infty)$	$(-2.02, \infty)$	$(-2.55, \infty)$	$(-2.24, \infty)$	$(-2.67, \infty)$	$(-2.40, \infty)$
1000		$(-1.92, \infty)$	$(-1.82, \infty)$	$(-1.95, \infty)$	$(-1.84, \infty)$	$(-1.93, \infty)$	$(-1.83, \infty)$
2000		$(-1.76, \infty)$	$(-1.71, \infty)$	$(-1.77, \infty)$	$(-1.72, \infty)$	$(-1.76, \infty)$	$(-1.71, \infty)$
4000		$(-1.67, \infty)$	$(-1.65, \infty)$	$(-1.68, \infty)$	$(-1.66, \infty)$	$(-1.67, \infty)$	$(-1.65, \infty)$
500	Symmetrical	$(-2.43, -1.11)$	$(-2.53, -1.01)$	$(-2.72, -1.28)$	$(-2.83, -1.17)$	$(-2.82, -1.47)$	$(-2.92, -1.37)$
1000		$(-2.01, -1.20)$	$(-2.05, -1.17)$	$(-2.04, -1.22)$	$(-2.07, -1.18)$	$(-2.02, -1.21)$	$(-2.06, -1.18)$
2000		$(-1.82, -1.27)$	$(-1.83, -1.26)$	$(-1.83, -1.28)$	$(-1.83, -1.27)$	$(-1.82, -1.27)$	$(-1.83, -1.26)$
4000		$(-1.72, -1.34)$	$(-1.72, -1.33)$	$(-1.72, -1.34)$	$(-1.72, -1.34)$	$(-1.72, -1.33)$	$(-1.72, -1.33)$

Table 13: Logit Confidence Intervals for $\theta_{0n,2}$, Nominal level 0.90, $p_n = 3n/4$

		Oracle		P-Oracle		P-Oracle	
				$C_n = 1$		$C_n = \log\log p_n$	
n	Interval	Asymp.	Boot.	Asymp.	Boot.	Asymp.	Boot.
500	Lower 1-sided	$(-\infty, -1.30)$	$(-\infty, -1.07)$	$(-\infty, -1.52)$	$(-\infty, -1.25)$	$(-\infty, -1.58)$	$(-\infty, -1.36)$
1000		$(-\infty, -1.29)$	$(-\infty, -1.19)$	$(-\infty, -1.33)$	$(-\infty, -1.22)$	$(-\infty, -1.29)$	$(-\infty, -1.19)$
2000		$(-\infty, -1.35)$	$(-\infty, -1.31)$	$(-\infty, -1.36)$	$(-\infty, -1.32)$	$(-\infty, -1.35)$	$(-\infty, -1.31)$
4000		$(-\infty, -1.37)$	$(-\infty, -1.35)$	$(-\infty, -1.38)$	$(-\infty, -1.36)$	$(-\infty, -1.37)$	$(-\infty, -1.35)$
500	Upper 1-sided	$(-2.35, \infty)$	$(-2.07, \infty)$	$(-2.71, \infty)$	$(-2.34, \infty)$	$(-2.62, \infty)$	$(-2.35, \infty)$
1000		$(-1.92, \infty)$	$(-1.82, \infty)$	$(-1.98, \infty)$	$(-1.87, \infty)$	$(-1.92, \infty)$	$(-1.82, \infty)$
2000		$(-1.78, \infty)$	$(-1.74, \infty)$	$(-1.80, \infty)$	$(-1.75, \infty)$	$(-1.78, \infty)$	$(-1.74, \infty)$
4000		$(-1.67, \infty)$	$(-1.65, \infty)$	$(-1.67, \infty)$	$(-1.65, \infty)$	$(-1.67, \infty)$	$(-1.65, \infty)$
500	Symmetrical	$(-2.50, -1.16)$	$(-2.60, -1.06)$	$(-2.88, -1.35)$	$(-2.99, -1.24)$	$(-2.77, -1.43)$	$(-2.86, -1.34)$
1000		$(-2.01, -1.20)$	$(-2.05, -1.17)$	$(-2.07, -1.23)$	$(-2.11, -1.20)$	$(-2.01, -1.20)$	$(-2.05, -1.17)$
2000		$(-1.84, -1.29)$	$(-1.86, -1.28)$	$(-1.86, -1.30)$	$(-1.87, -1.29)$	$(-1.84, -1.29)$	$(-1.86, -1.28)$
4000		$(-1.71, -1.33)$	$(-1.71, -1.33)$	$(-1.71, -1.34)$	$(-1.72, -1.33)$	$(-1.71, -1.33)$	$(-1.71, -1.33)$