AEA Papers and Proceedings 2021, 111: 616–620
https://doi.org/10.1257/pandp.20211036

The Proximal Bootstrap for Finite-Dimensional Regularized Estimators†

By Jessie Li*

We propose a computationally efficient bootstrap procedure to conduct pointwise asymptotically valid inference for a large class of $\sqrt{n}$-consistent estimators with nonstandard asymptotic distributions for which standard bootstrap procedures are known to be inconsistent. The application we consider in this paper is finite-dimensional regularized estimators, such as the lasso (Tibshirani 1996), ℓ_1-norm regularized quantile regression (Belloni and Chernozhukov 2011), ℓ_1-norm support vector regression (Zhu et al. 2004, Bai et al. 2019), and trace regression via nuclear norm regularization (Koltchinskii, Lounici, and Tsybakov 2011, Moon and Weidner 2018). Another application that will be explored in a subsequent paper is constrained optimization problems with a possibly nonsmooth and nonconvex objective function and a finite number of either estimated or fixed inequality and/or equality constraints, and where the true parameter can lie on the boundary of the constraint set (Andrews 1999, 2000, 2002a).

Motivated by the optimization literature and recent contributions in computationally efficient bootstrap procedures (e.g., Forneron and Ng 2020), our proximal bootstrap estimator can be expressed as the solution to a convex optimization problem and efficiently computed starting from an initial $\sqrt{n}$-consistent estimator using built-in and freely available software. Additionally, when the sample Hessian is proportional to the identity matrix, the proximal bootstrap has a closed-form solution. In the case of a smooth sample objective function and no regularization, the proximal bootstrap is very similar to the k-step bootstrap (for $k = 1$) proposed by Davidson and MacKinnon (1999) and investigated further by Andrews (2002b).

The consistency of the proximal bootstrap relies on a scaling sequence (labeled α_n in this paper) that converges to zero at a slower-than-$\sqrt{n}$ rate. The purpose of the slower-than-$\sqrt{n}$ rate is to offset the estimation error from the initial $\sqrt{n}$-consistent estimator. The purpose of α_n is similar to that of ϵ_n in the numerical bootstrap in Hong and Li (2020). However, we want to emphasize that the proximal bootstrap is a different procedure than the numerical bootstrap because it solves a different optimization problem. The proximal bootstrap works only for $\sqrt{n}$-consistent estimators but is typically more computationally efficient than the numerical bootstrap.

Section I reviews the concept of proximal mappings from the optimization literature. Section II contains all of the theoretical results demonstrating consistency of the proximal bootstrap for finite-dimensional regularized estimators. Section III concludes. The online Appendix contains the proof of consistency; provides the specific form of the proximal bootstrap estimator for the lasso, ℓ_1-norm support vector regression (of which ℓ_1-norm regularized quantile regression is a special case), and trace regression via nuclear norm regularization; and also contains a Monte Carlo simulation for the lasso.

I. Proximal Mappings

Given an Euclidean space $\mathcal{D}$ and a function $r : \mathcal{D} \mapsto \mathbb{R}$, the proximal mapping of r is the operator given by

$$prox_r(z) = \underset{\beta \in \mathcal{D}}{\arg\min} \left\{ r(\beta) + \frac{1}{2} \| \beta - z \|_2^2 \right\}$$

for any $z \in \mathcal{D}$.

*Department of Economics, University of California, Santa Cruz (email: jeqli@ucsc.edu). I would like to thank Alex Torgovitsky and the participants of the UChicago Econometrics workshop and the 2021 ASSA session Optimization-Conscious Econometrics for helpful comments and suggestions.

†Go to https://doi.org/10.1257/pandp.20211036 to visit the article page for additional materials and author disclosure statement(s).

Given a function $r:\mathcal{D}\mapsto\mathbb{R}$ and a symmetric positive definite matrix H, the scaled proximal mapping of r is the operator given by, for $\|\beta-z\|_H^2 = (\beta-z)'H(\beta-z)$,

$$prox_{H,r}(z) = \operatorname*{argmin}_{\beta\in\mathcal{D}}\left\{ r(\beta)+\frac{1}{2}\|\beta-z\|_H^2\right\} \quad \text{for any } z\in\mathcal{D}.$$

When r is a proper closed and convex function, then $prox_r(z)$ is a singleton for any $z\in\mathcal{D}$ (Beck 2017, theorem 6.3). The same can be said for $prox_{H,r}(z)$ (Lee, Sun, and Saunders 2014).

The proximal map often has a closed-form solution. For instance, the proximal mapping of the ℓ_1-norm is given by

$$\begin{aligned} prox_{\lambda\|\cdot\|_1}(z) &= \arg\min_{\beta}\left\{\lambda\|\beta\|_1+\frac{1}{2}\|\beta-z\|_2^2\right\} \\ &= \operatorname{sign}(z)\max\{|z|-\lambda,0\} \\ &= (z-\lambda)^+-(z+\lambda)^-, \end{aligned}$$

where $x^+\equiv\max(x,0)$ and $x^-\equiv-\min(x,0)$.

Although it is rarely the case that the scaled proximal map has a closed-form solution, it can still be efficiently computed as the solution to a convex optimization problem if r is convex. Additionally, Friedlander and Goh (2017) show that for certain r that have a "quadratic support" representation (which is satisfied for many functions such as the ℓ_1 norm, the ℓ_2 norm, and indicators on polyhedral cones), the scaled proximal map can be written as a quadratic optimization problem over conic constraints.

II. Proximal Bootstrap

A. *Notation*

Consider a random sample $X_1,X_2,\ldots,X_n$ of independent draws from a probability measure P on a sample space $\mathcal{X}$. Define the empirical measure $P_n\equiv(1/n)\sum_{i=1}^n\delta_{X_i}$, where δ_x is the measure that assigns mass 1 at x and 0 everywhere else. Denote the bootstrap empirical measure by P_n^*, which can refer to the multinomial, wild, or other exchangeable bootstraps. Weak convergence is defined in the sense of Kosorok (2007): $Z_n\rightsquigarrow Z$ in the metric space $(\mathcal{D},d)$ if and only if $\sup_{f\in BL_1}|E^*f(Z_n)-Ef(Z)|\to 0$, where BL_1 is the space of functions $f:\mathcal{D}\mapsto\mathbb{R}$ with Lipschitz norm bounded by 1. Conditional weak convergence is also defined in the sense of Kosorok (2007): $Z_n\overset{\mathcal{P}}{\underset{\mathcal{W}}{\rightsquigarrow}}Z$ in the metric space $(\mathcal{D},d)$ if and only if $\sup_{f\in BL_1}|E_{\mathcal{W}}f(Z_n)-Ef(Z)|\xrightarrow{p}0$ and $E_{\mathcal{W}}f(Z_n)^*-E_{\mathcal{W}}f(Z_n)_*\xrightarrow{p}0$ for all $f\in BL_1$, where BL_1 is the space of functions $f:\mathcal{D}\mapsto\mathbb{R}$ with Lipschitz norm bounded by 1, $E_{\mathcal{W}}$ denotes expectation with respect to the bootstrap weights $\mathcal{W}$ conditional on the data, and $f(Z_n)^*$ and $f(Z_n)_*$ denote measurable majorants and minorants with respect to the joint data (including the weights $\mathcal{W}$). Let $X_n^*=o_P^*(1)$ if the law of X_n^* is governed by P_n and if $P_n(|X_n^*|>\epsilon)=o_P(1)$ for all $\epsilon>0$. Also define $M_n^*=O_p^*(1)$ (hence also $O_p(1)$) if $\lim_{m\to\infty}\limsup_{n\to\infty}P(P_n(M_n^*>m)>\epsilon)\to 0,\forall\epsilon>0$.

B. *Finite-Dimensional Regularized Estimators*

We first consider $\sqrt{n}$-consistent estimators $\hat{\beta}_n$ that minimize an objective function that can be written as the sum of two functions: the random, possibly nonconvex, nonsmooth loss function $\hat{Q}_n(\beta)$ and the penalty function $(\lambda_n/\sqrt{n})\,r(\beta)$, where $r:\mathbb{R}^d\mapsto\mathbb{R}$ is a typically convex but nonsmooth deterministic function, and $(\lambda_n/\sqrt{n})=o(1)$. We assume d is fixed. Formally,

$$\hat{\beta}_n = \operatorname*{argmin}_{\beta\in\mathbb{R}^d}\left\{\hat{Q}_n(\beta)+\frac{\lambda_n}{\sqrt{n}}r(\beta)\right\}.$$

We propose a proximal bootstrap estimator $\hat{\beta}_n^*$ that can be efficiently computed using standard, built-in optimization routines starting from an initial $\sqrt{n}$-consistent estimator $\sqrt{n}(\bar{\beta}_n-\beta_0)=O_p(1)$, where $\beta_0=\operatorname{argmin}_{\beta\in\mathbb{R}^d}Q(\beta)$. One possible $\bar{\beta}_n$ is $\hat{\beta}_n$, but sometimes there are more computationally efficient estimators. For some $\alpha_n\to 0$ and $\alpha_n\sqrt{n}\to\infty$,

$$\hat{\beta}_n^*=prox_{\bar{H}_n,\alpha_n\lambda_n r(\cdot)}\Big(\bar{\beta}_n-\alpha_n\sqrt{n} \times\bar{H}_n^{-1}\big(\hat{l}_n^*(\bar{\beta}_n)-\hat{l}_n(\bar{\beta}_n)\big)\Big).$$

Here, $\hat{l}_n(\bar{\beta}_n)$ is a consistent estimate of $l(\beta_0)=\partial Q(\beta_0)/\partial\beta$, where $Q(\beta)$ is a lower semicontinuous function that is twice differentiable at β_0, and $\sup_{\beta\in K}|\hat{Q}_n(\beta)-Q(\beta)|=o_p(1)$ for every compact subset K of $\mathbb{R}^d$. In the case

where $\hat{Q}_n(\beta)$ is differentiable, $\hat{l}_n(\beta)$ can simply be the Jacobian of $\hat{Q}_n(\beta)$. More generally, to handle nondifferentiable $\hat{Q}_n(\beta)$, $\hat{l}_n(\beta)$ is a subgradient of $\hat{Q}_n(\beta)$. Note, $\hat{l}_n^*(\bar{\beta}_n)$ is a bootstrap analog of $\hat{l}_n(\bar{\beta}_n)$ using the multinomial, wild, or other exchangeable bootstraps; $\bar{H}_n$ is a consistent, symmetric, positive definite estimate of the population Hessian $H_0 = \partial^2 Q(\beta_0)/\partial\beta\partial\beta'$.

If $\bar{H}_n = (1/c)I_d$ for some constant c, then $\hat{\beta}_n^*$ reduces down to an unscaled proximal map, which often has a closed-form solution:

$$\begin{aligned}
&\hat{\beta}_n^* \\
&= prox_{\frac{1}{c}I_d,\alpha_n\lambda_n r(\cdot)}\Big(\bar{\beta}_n - \alpha_n\sqrt{n}\,cI_d\big(\hat{l}_n^*(\bar{\beta}_n) - \hat{l}_n(\bar{\beta}_n)\big)\Big) \\
&= \operatorname*{argmin}_{\beta}\Big\{\alpha_n\lambda_n r(\beta) \\
&\qquad + \frac{1}{2c}\|\beta - \bar{\beta}_n \\
&\qquad\quad + c\alpha_n\sqrt{n}\big(\hat{l}_n^*(\bar{\beta}_n) - \hat{l}_n(\bar{\beta}_n)\big)\|^2\Big\} \\
&= \operatorname*{argmin}_{\beta}\Big\{c\alpha_n\lambda_n r(\beta) \\
&\qquad + \frac{1}{2}\|\beta - \bar{\beta}_n \\
&\qquad\quad + c\alpha_n\sqrt{n}\big(\hat{l}_n^*(\bar{\beta}_n) - \hat{l}_n(\bar{\beta}_n)\big)\|^2\Big\} \\
&= prox_{c\alpha_n\lambda_n r(\cdot)}\Big(\bar{\beta}_n - c\alpha_n\sqrt{n}\big(\hat{l}_n^*(\bar{\beta}_n) - \hat{l}_n(\bar{\beta}_n)\big)\Big).
\end{aligned}$$

Even if there is no closed form for $\hat{\beta}_n^*$, it is still the solution to a convex optimization problem assuming $r(\beta)$ is convex:

$$\begin{aligned}
\hat{\beta}_n^* &= \operatorname*{argmin}_{\beta}\Big\{\alpha_n\lambda_n r(\beta) \\
&\qquad + \frac{1}{2}\|\beta - \bar{\beta}_n \\
&\qquad\quad + \alpha_n\sqrt{n}\,\bar{H}_n^{-1}\big(\hat{l}_n^*(\bar{\beta}_n) \\
&\qquad\qquad - \hat{l}_n(\bar{\beta}_n)\big)\|_{\bar{H}_n}^2\Big\} \\
&= \operatorname*{argmin}_{\beta}\Big\{\alpha_n\lambda_n r(\beta) \\
&\qquad + \alpha_n\sqrt{n}\big(\hat{l}_n^*(\bar{\beta}_n) - \hat{l}_n(\bar{\beta}_n)\big)' \\
&\qquad\quad \times (\beta - \bar{\beta}_n) + \frac{1}{2}\|\beta - \bar{\beta}_n\|_{\bar{H}_n}^2\Big\}.
\end{aligned}$$

Furthermore, for certain types of $r(\beta)$, we can use proposition 4.1 in Friedlander and Goh (2017) to efficiently compute the proximal bootstrap by solving a quadratic optimization problem over conic constraints. For example, if $r(\beta) = \|\beta\|_1$,

$$\begin{aligned}
\hat{\beta}_n^* &= \bar{H}_n^{-1}\Big(\bar{H}_n\Big(\bar{\beta}_n - \alpha_n\sqrt{n}\,\bar{H}_n^{-1} \\
&\qquad \times \big(\hat{l}_n^*(\bar{\beta}_n) - \hat{l}_n(\bar{\beta}_n)\big)\Big) \\
&\qquad - \alpha_n\lambda_n\gamma^*\Big),
\end{aligned}$$

$$\begin{aligned}
\gamma^* &= \operatorname*{argmin}_{\gamma\in\{\gamma:\|\gamma\|_\infty\le 1\}} \frac{\alpha_n\lambda_n}{2}\gamma'\bar{H}_n^{-1}\gamma \\
&\qquad - \Big(\bar{\beta}_n - \alpha_n\sqrt{n}\,\bar{H}_n^{-1} \\
&\qquad\quad \times \big(\hat{l}_n^*(\bar{\beta}_n) - \hat{l}_n(\bar{\beta}_n)\big)\Big)'\gamma.
\end{aligned}$$

Remark 1: In the case of $r(\beta) = 0$, smooth $\hat{Q}_n(\beta)$, and $\bar{\beta}_n$ that satisfies $\hat{l}_n(\bar{\beta}_n) = 0$, the proximal bootstrap is similar to the k-step bootstrap (for $k = 1$) proposed by Davidson and MacKinnon (1999) and investigated further by Andrews (2002b), except with an additional scaling factor of $\alpha_n\sqrt{n}$:

$$\begin{aligned}
\hat{\beta}_n^* &= \bar{\beta}_n - \alpha_n\sqrt{n}\,\bar{H}_n^{-1}\big(\hat{l}_n^*(\bar{\beta}_n) - \hat{l}_n(\bar{\beta}_n)\big) \\
&= \bar{\beta}_n - \alpha_n\sqrt{n}\,\bar{H}_n^{-1}\hat{l}_n^*(\bar{\beta}_n).
\end{aligned}$$

If $\alpha_n = 1/\sqrt{n}$ in this case, then the proximal bootstrap coincides with the one-step bootstrap.

C. *Assumptions*

The first assumption is needed to show consistency of $\hat{\beta}_n$ for β_0.

ASSUMPTION 1: *(i)* $\hat{\beta}_n = \operatorname{argmin}_{\beta\in\mathbb{R}^d}\big\{\hat{Q}_n(\beta) + (\lambda_n/\sqrt{n})\,r(\beta)\big\}$ *is uniformly tight.* *(ii)* $\beta_0 = \operatorname{argmin}_{\beta\in\mathbb{R}^d} Q(\beta)$ *is unique, where* $Q(\beta)$ *is a lower semicontinuous function that is twice differentiable at* β_0 *and* $\sup_{\beta\in K}\big|\hat{Q}_n(\beta) - Q(\beta)\big| = o_p(1)$ *for every compact subset K of* $\mathbb{R}^d$.

The next assumption states that the objective function admits a uniform local quadratic approximation around $\sqrt{n}$ neighborhoods of β_0. It is needed to derive the asymptotic distribution of $\sqrt{n}\big(\hat{\beta}_n - \beta_0\big)$.

ASSUMPTION 2: *There exists a symmetric, positive definite* H_0 *and* $\sqrt{n}\big(\hat{l}_n(\beta_0) - l(\beta_0)\big) = O_p(1)$ *such that for any* $\delta_n \to 0$,

$$\sup_{\|h\| \le \sqrt{n}\delta_n} \Big| \Big(n\hat{Q}_n\Big(\beta_0 + \tfrac{h}{\sqrt{n}}\Big) - n\hat{Q}_n(\beta_0) - h'\sqrt{n}\big(\hat{l}_n(\beta_0) - l(\beta_0)\big) - \tfrac{1}{2}h'H_0h\Big)/(1+\|h\|^2)\Big| = o_p(1).$$

The next assumption is needed to show that $\sqrt{n}\big(\hat{l}_n(\beta_0) - l(\beta_0)\big)$ and $\sqrt{n}\big(\hat{l}^*_n(\beta_0) - \hat{l}_n(\beta_0)\big)$ have the same asymptotic distribution.

ASSUMPTION 3: *There exists a function* $g : \mathcal{X} \mapsto \mathbb{R}$ *indexed by a parameter* $\beta \in \mathbb{R}^d$ *such that for any* $\beta \in \mathbb{R}^d$, $\sqrt{n} \times \big(\hat{l}_n(\beta) - l(\beta)\big) = \sqrt{n}(P_n - P)g(\cdot,\beta) + o_p(1)$ *and* $\sqrt{n}\big(\hat{l}^*_n(\beta) - \hat{l}_n(\beta)\big) = \sqrt{n}(P^*_n - P_n)g(\cdot,\beta) + o^*_p(1)$, *where* $\lim_{n\to\infty} P\|g(\cdot,\beta_0)\|^2 \times \mathbf{1}\big(\|g(\cdot,\beta_0)\| > \epsilon\sqrt{n}\big) = 0$ *for each* $\epsilon > 0$.

The next assumption is needed to show stochastic equicontinuity of $\sqrt{n}\big(\hat{l}_n(\beta) - l(\beta)\big)$ and bootstrap equicontinuity results, which will be used to show $\sqrt{n}\big(\hat{l}^*_n(\bar{\beta}_n) - \hat{l}_n(\bar{\beta}_n)\big)$ and $\sqrt{n}\big(\hat{l}^*_n(\beta_0) - \hat{l}_n(\beta_0)\big)$ have the same asymptotic distribution.

ASSUMPTION 4: *(i)* $\mathcal{G}_R \equiv \big\{g(\cdot,\beta) - g(\cdot,\beta_0) : \|\beta - \beta_0\| \le R\big\}$ *is a Donsker class for some* $R > 0$, *and* $P\big(g(\cdot,\beta) - g(\cdot,\beta_0)\big)^2 \to 0$ *for* $\beta \to \beta_0$. *(ii)* $\lim_{\lambda\to\infty} \limsup_{n\to\infty} \sup_{t\ge\lambda} t^2 \times P\Big\{ \sup_{g(\cdot,\beta)\in\mathcal{G}_{\delta_n}} \big\| \frac{g(\cdot,\beta) - g(\cdot,\beta_0)}{1 + \sqrt{n}\,\|\beta - \beta_0\|} \big\| > t \Big\} = 0$ *for any* $\delta_n \to 0$.

Note, (i) will imply stochastic equicontinuity, which in combination with the envelope function integrability condition in (ii) will imply bootstrap equicontinuity. A sufficient condition for (ii) is $\sup_{g(\cdot,\beta)\in\mathcal{G}_{\delta_n}} \big\| \frac{g(\cdot,\beta) - g(\cdot,\beta_0)}{1 + \sqrt{n}\,\|\beta - \beta_0\|} \big\| \le C$ for some constant C.

The next assumption states that $r(\beta)$ is closed, convex, and Hadamard directionally differentiable at β_0. It is needed to derive the asymptotic distribution of $\sqrt{n}\big(\hat{\beta}_n - \beta_0\big)$.

ASSUMPTION 5: $r : \mathbb{R}^d \to \mathbb{R}$ *is a proper closed, convex function, and there is a continuous map* $r'_{\beta_0} : \mathbb{R}^d \to \mathbb{R}$ *such that for all* $h_n \to h \in \mathbb{R}^d$, $\lim_{\alpha_n \downarrow 0} \Big| \frac{r(\beta_0 + \alpha_n h_n) - r(\beta_0)}{\alpha_n} - r'_{\beta_0}(h) \Big| = 0$.

The next theorem demonstrates consistency of the proximal bootstrap by showing that the limiting distribution of $(\hat{\beta}^*_n - \hat{\beta}_n)/\alpha_n$ coincides with the limiting distribution of $\sqrt{n}\big(\hat{\beta}_n - \beta_0\big)$.

THEOREM 1: *Suppose Assumptions 1–5 are satisfied and* $\lambda_n \to \lambda_0 \in [0,\infty)$. *Then for any* $\bar{\beta}_n$ *such that* $\sqrt{n}\big(\bar{\beta}_n - \beta_0\big) = O_p(1)$, *for any* $\bar{H}_n$ *that is a consistent, symmetric, positive definite estimate of* H_0, *and for any sequence* α_n *such that* $\alpha_n \to 0$ *and* $\sqrt{n}\alpha_n \to \infty$, $\sqrt{n}\big(\hat{\beta}_n - \beta_0\big) \rightsquigarrow \mathcal{J}$ *and* $\big(\hat{\beta}^*_n - \hat{\beta}_n\big)/\alpha_n \overset{\mathcal{P}}{\underset{\mathcal{W}}{\rightsquigarrow}} \mathcal{J}$, *where* $\mathcal{J} = \operatorname{argmin}_{h\in\mathbb{R}^d} \big\{ \lambda_0\, r'_{\beta_0}(h) + h'W_0 + (1/2)h'H_0h \big\}$ *and* $W_0 \sim N\big(0, P\big(g(\cdot,\beta_0) - Pg(\cdot,\beta_0)\big)\big(g(\cdot,\beta_0) - Pg(\cdot,\beta_0)\big)'\big)$.

III. Conclusion

We have proposed a computationally efficient proximal bootstrap estimator that consistently estimates the limiting distribution of $\sqrt{n}$-consistent estimators for which the standard bootstrap is known to be inconsistent. This paper has considered the application to finite-dimensional regularized estimators; an application to constrained estimators will be explored in a subsequent paper.

REFERENCES

Andrews, Donald W.K. 1999. "Estimation When a Parameter Is on a Boundary." *Econometrica* 67 (6): 1341–83.

Andrews, Donald W.K. 2000. "Inconsistency of the Bootstrap When a Parameter Is on the Boundary of the Parameter Space." *Econometrica* 68 (2): 399–405.

Andrews, Donald W.K. 2002a. "Generalized Method of Moments Estimation When a Parameter Is on a Boundary." *Journal of Business & Economic Statistics* 20 (4): 530–44.

Andrews, Donald W.K. 2002b. "Higher-Order Improvements of a Computationally Attractive

k-Step Bootstrap for Extremum Estimators." *Econometrica* 70 (1): 119–62.

Bai, Yuehao, Hung Ho, Guillaume A. Pouliot, and Joshua K.C. Shea. 2019. "Inference for Support Vector Regression under l1 Regularization." Unpublished.

Beck, Amir. 2017. *First-Order Methods in Optimization.* Philadelphia: SIAM.

Belloni, Alexandre, and Victor Chernozhukov. 2011. "ℓ1-penalized Quantile Regression in High-Dimensional Sparse Models." *Annals of Statistics* 39 (1): 82–130.

Davidson, Russell, and James G. MacKinnon. 1999. "Bootstrap Testing in Nonlinear Models." *International Economic Review* 40 (2): 487–508.

Forneron, Jean-Jacques, and Serena Ng. 2020. "Inference by Stochastic Optimization: A Free-Lunch Bootstrap." https://arxiv.org/abs/2004.09627.

Friedlander, Michael P., and Gabriel Goh. 2017. "Efficient Evaluation of Scaled Proximal Operators." *Electronic Transactions on Numerical Analysis* 46: 1–22.

Hong, Han, and Jessie Li. 2020. "The Numerical Bootstrap." *Annals of Statistics* 48 (1): 397–412.

Koltchinskii, Vladimir, Karim Lounici, and Alexandre B. Tsybakov. 2011. "Nuclear-Norm Penalization and Optimal Rates for Noisy Low-Rank Matrix Completion." *Annals of Statistics* 39 (5): 2302–29.

Kosorok, Michael R. 2007. *Introduction to Empirical Processes and Semiparametric Inference.* Berlin: Springer.

Lee, Jason D., Yuekai Sun, and Michael A. Saunders. 2014. "Proximal Newton-Type Methods for Minimizing Composite Functions." *SIAM Journal on Optimization* 24 (3): 1420–43.

Moon, Hyungsik Roger, and Martin Weidner. 2018. "Nuclear Norm Regularized Estimation of Panel Regression Models." Available on arXiv at 1810.10987v1.

Tibshirani, Robert. 1996. "Regression Shrinkage and Selection via the Lasso." *Journal of the Royal Statistical Society: Series B (Methodological)* 58 (1): 267–88.

Zhu, Ji, Saharon Rosset, Trevor Hastie, and Rob Tibshirani. 2004. "1-norm Support Vector Machines." In *NIPS'03: Proceedings of the 16th International Conference on Nueral Information Processing Systems*, edited by Sebastian B. Thrun and Lawrence K. Saul, 49–56. Cambridge, MA: MIT Press.

Online Appendix to The Proximal Bootstrap for Finite-Dimensional Regularized Estimators

Jessie Li

January 20, 2021

1 Examples of Regularized Estimators

1.1 LASSO

$$\hat{\beta}_n = \arg\min_{\beta} \left\{ \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - x_i'\beta\right)^2 + \frac{\lambda_n}{\sqrt{n}} \left\|\beta\right\|_1 \right\}$$

$$r'_{\beta_0}(h) = \sum_{j=1}^{p} \left(h_j sign\left(\beta_{0j}\right) 1\left(\beta_{0j} \neq 0\right) + \left|h_j\right| 1\left(\beta_{0j} = 0\right)\right)$$

$$\hat{\beta}_n^* = \arg\min_{\beta} \alpha_n \lambda_n \left\|\beta\right\|_1 + \alpha_n \sqrt{n} \left(\hat{l}_n^*\left(\bar{\beta}_n\right) - \hat{l}_n\left(\bar{\beta}_n\right)\right)' \left(\beta - \bar{\beta}_n\right) + \frac{1}{2} \left\|\beta - \bar{\beta}_n\right\|_{\bar{H}_n}^2$$

$$l\left(\beta_0\right) = -E\left[x_i\left(y_i - x_i'\beta\right)\right], \quad \hat{l}_n\left(\bar{\beta}_n\right) = -\frac{1}{n}\sum_{i=1}^{n} x_i\left(y_i - x_i'\bar{\beta}_n\right)$$

$$H_0 = E\left[x_i x_i'\right], \quad \bar{H}_n = \frac{1}{n}\sum_{i=1}^{n} x_i x_i'$$

Examples of $\hat{l}_n^*\left(\bar{\beta}_n\right)$ include the multinomial and wild bootstrap analogs of $\hat{l}_n\left(\bar{\beta}_n\right)$:

$$\hat{l}_n^*\left(\bar{\beta}_n\right) = -\frac{1}{n}\sum_{i=1}^{n} x_i^*\left(y_i^* - x_i^{*'}\bar{\beta}_n\right), \quad \hat{l}_n^*\left(\bar{\beta}_n\right) = -\frac{1}{n}\sum_{i=1}^{n}\left(\xi_i - \bar{\xi}\right) x_i\left(y_i - x_i'\bar{\beta}_n\right)$$

where ξ_i are i.i.d. variables with variance 1 and finite 3rd moment and $\bar{\xi} = \frac{1}{n}\sum_{i=1}^{n}\xi_i$.

If $\bar{H}_n = \frac{1}{c} I_d$, $\hat{\beta}_n^*$ has a closed form solution:

$$\begin{aligned}\hat{\beta}_n^* &= prox_{c\alpha_n\lambda_n\|\cdot\|_1}\left(\bar{\beta}_n - c\alpha_n\sqrt{n}\left(\hat{l}_n^*\left(\bar{\beta}_n\right) - \hat{l}_n\left(\bar{\beta}_n\right)\right)\right) \\ &= \left(\bar{\beta}_n - c\alpha_n\sqrt{n}\left(\hat{l}_n^*\left(\bar{\beta}_n\right) - \hat{l}_n\left(\bar{\beta}_n\right)\right) - c\alpha_n\lambda_n\right)^+ - \left(\bar{\beta}_n - c\alpha_n\sqrt{n}\left(\hat{l}_n^*\left(\bar{\beta}_n\right) - \hat{l}_n\left(\bar{\beta}_n\right)\right) + c\alpha_n\lambda_n\right)^-\end{aligned}$$

where $x^+ \equiv \max(x, 0)$ and $x^- \equiv -\min(x, 0)$.

1.2 ℓ_1-norm support vector regression

The ℓ_1-norm support vector regression (SVR) estimator of Zhu et al. (2004) is similar to the ℓ_1 penalized quantile regression estimator of Belloni and Chernozhukov (2011):

$$\hat{\beta}_n = \arg\min \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(\rho_\tau \left(y_i - x_i' \beta \right) - \kappa \right)^+ + \frac{\lambda_n}{\sqrt{n}} \left\| \beta \right\|_1 \right\}$$

The objective uses a relaxed version of the check function:

$$\begin{aligned}
&\left(\rho_\tau \left(y_i - x_i' \beta \right) - \kappa \right)^+ \\
&= \left(\left\{ (1-\tau) \, 1 \left(y_i - x_i' \beta \leqslant 0 \right) + \tau 1 \left(y_i - x_i' \beta > 0 \right) \right\} \left| y_i - x_i' \beta \right| - \kappa \right)^+ \\
&= \begin{cases} \left(- (1-\tau) \left(y_i - x_i' \beta \right) - \kappa \right) 1 \left(- (1-\tau) \left(y_i - x_i' \beta \right) - \kappa > 0 \right) & , y_i - x_i' \beta \leqslant 0 \\ \left(\tau \left(y_i - x_i' \beta \right) - \kappa \right) 1 \left(\tau \left(y_i - x_i' \beta \right) - \kappa > 0 \right) & , y_i - x_i' \beta > 0 \end{cases} \\
&= \begin{cases} \left(- (1-\tau) \left(y_i - x_i' \beta \right) - \kappa \right) 1 \left(y_i < x_i' \beta - \frac{\kappa}{1-\tau} \right) & , y_i - x_i' \beta \leqslant 0 \\ \left(\tau \left(y_i - x_i' \beta \right) - \kappa \right) 1 \left(y_i > x_i' \beta + \frac{\kappa}{\tau} \right) & , y_i - x_i' \beta > 0 \end{cases} \\
&= \left(\tau \left(y_i - x_i' \beta \right) - \kappa \right) 1 \left(y_i > x_i' \beta + \frac{\kappa}{\tau} \right) - \left((1-\tau) \left(y_i - x_i' \beta \right) + \kappa \right) 1 \left(y_i < x_i' \beta - \frac{\kappa}{1-\tau} \right)
\end{aligned}$$

The proximal bootstrap estimator is

$$\hat{\beta}_n^* = \arg\min_{\beta} \alpha_n \lambda_n \left\| \beta \right\|_1 + \alpha_n \sqrt{n} \left(\hat{l}_n^* \left(\bar{\beta}_n \right) - \hat{l}_n \left(\bar{\beta}_n \right) \right)' \left(\beta - \bar{\beta}_n \right) + \frac{1}{2} \left\| \beta - \bar{\beta}_n \right\|_{\bar{H}_n}^2$$

$\hat{l}_n \left(\bar{\beta}_n \right)$ is a consistent estimate of $l \left(\beta_0 \right)$ using $\bar{\beta}_n$:

$$\begin{aligned}
l \left(\beta_0 \right) &= -E \left[x_i \left(\tau 1 \left(y_i > x_i' \beta_0 + \frac{\kappa}{\tau} \right) - (1-\tau) \, 1 \left(y_i < x_i' \beta_0 - \frac{\kappa}{1-\tau} \right) \right) \right] \\
\hat{l}_n \left(\bar{\beta}_n \right) &= -\frac{1}{n} \sum_{i=1}^{n} x_i \left(\tau 1 \left(y_i > x_i' \bar{\beta}_n + \frac{\kappa}{\tau} \right) - (1-\tau) \, 1 \left(y_i < x_i' \bar{\beta}_n - \frac{\kappa}{1-\tau} \right) \right)
\end{aligned}$$

The population Hessian and its consistent estimate using $\bar{\beta}_n$ are given by

$$\begin{aligned}
H_0 &= E \left[x_i x_i' \left(\tau f_{y|x} \left(x_i' \beta_0 + \frac{\kappa}{\tau} \right) + (1-\tau) \, f_{y|x} \left(x_i' \beta_0 - \frac{\kappa}{1-\tau} \right) \right) \right] \\
\bar{H}_n &= \frac{1}{n} \sum_{i=1}^{n} x_i x_i' \left(\tau \hat{f}_{y|x} \left(x_i' \bar{\beta}_n + \frac{\kappa}{\tau} \right) + (1-\tau) \, \hat{f}_{y|x} \left(x_i' \bar{\beta}_n - \frac{\kappa}{1-\tau} \right) \right)
\end{aligned}$$

An example of $\hat{f}_{y|x} \left(y \right)$ is $\frac{1}{n} \sum_{j=1}^{n} K_h \left(y \right)$, where $K_h \left(y \right) = \frac{1}{h} K \left(y/h \right)$ and $K \left(u \right)$ is a kernel function that is symmetric around 0 and integrates to 1.

1.3 Trace Regression via Nuclear Norm Regularization

$$\hat{\Theta}_n = \arg\min_{\Theta \in \mathbb{R}^{d_1 \times d_2}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(y_i - tr \left(\Theta' X_i \right) \right) + \lambda_n \left\| \Theta \right\|_* \right\}$$

where $\|\Theta\|_* = \sum_{j=1}^{d_1 \wedge d_2} \sigma_j(\Theta)$ is the nuclear norm of Θ, and $\sigma_j(\Theta)$ is the jth largest singular value of Θ.

$$\hat{\Theta}_n^* = \arg\min_{\Theta} \alpha_n \lambda_n \|\Theta\|_* + \alpha_n \sqrt{n} \left(\hat{l}_n^* \left(\bar{\Theta}_n \right) - \hat{l}_n \left(\bar{\Theta}_n \right) \right)' \left(\Theta - \bar{\Theta}_n \right) + \frac{1}{2} \left\| \Theta - \bar{\Theta}_n \right\|_{\bar{H}_n}^2$$

$$\hat{l}_n \left(\bar{\Theta}_n \right) = -\frac{1}{n} \sum_{i=1}^{n} X_i \left(y_i - tr \left(\bar{\Theta}_n' X_i \right) \right), \quad \bar{H}_n = \frac{1}{n} \sum_{i=1}^{n} X_i X_i'$$

$$r'_{\Theta_0}(h) = \sum_{j=1}^{d_1 \wedge d_2} \left(h_j 1 \left(\sigma_j \left(\Theta_0 \right) \neq 0 \right) + |h_j| 1 \left(\sigma_j \left(\Theta_0 \right) = 0 \right) \right)$$

In the case of $\bar{H}_n = \frac{1}{c} I_{d_1}$, the proximal bootstrap has a closed form:

$$\hat{\Theta}_n^* = prox_{c\alpha_n \lambda_n \|\cdot\|_*} \left(\bar{\Theta}_n - c\alpha_n \sqrt{n} \left(\hat{l}_n^* \left(\bar{\Theta}_n \right) - \hat{l}_n \left(\bar{\Theta}_n \right) \right) \right) = U \Sigma_{c\alpha_n \lambda_n} V^T$$

where $\Sigma_{c\alpha_n \lambda_n} = \text{diag} \left\{ \max \left(\Sigma_1 - c\alpha_n \lambda_n, 0 \right), \max \left(\Sigma_2 - c\alpha_n \lambda_n, 0 \right), ..., \max \left(\Sigma_{d_1 \wedge d_2} - c\alpha_n \lambda_n, 0 \right) \right\}$, and for $j = 1...d_1 \wedge d_2$, Σ_j are the singular values of $\bar{\Theta}_n - c\alpha_n \sqrt{n} \left(\hat{l}_n^* \left(\bar{\Theta}_n \right) - \hat{l}_n \left(\bar{\Theta}_n \right) \right)$.

2 Proof of Theorem 1

Assumption 1 implies that the conditions of part 2 of Corollary 3.2.3 of van der Vaart and Wellner (1996) are satisfied, and therefore $\hat{\beta}_n \xrightarrow{p} \beta_0 = \arg\min_{\beta \in \mathbb{R}^d} Q(\beta)$. To derive its asymptotic distribution, use the centered and scaled parameter $h = \sqrt{n} (\beta - \beta_0)$:

$$\begin{aligned}
\sqrt{n} \left(\hat{\beta}_n - \beta_0 \right) &= \arg\min_h \left\{ n\hat{Q}_n \left(\beta_0 + \frac{h}{\sqrt{n}} \right) - n\hat{Q}_n \left(\beta_0 \right) + \lambda_n \sqrt{n} r \left(\beta_0 + \frac{h}{\sqrt{n}} \right) \right\} \\
&= \arg\min_h \left\{ h' \sqrt{n} \left(\hat{l}_n \left(\beta_0 \right) - l \left(\beta_0 \right) \right) + \frac{1}{2} h' H_0 h + \lambda_n \left(\frac{r \left(\beta_0 + \frac{h}{\sqrt{n}} \right) - r \left(\beta_0 \right)}{1/\sqrt{n}} \right) + o_p(1) \right\} \\
&\rightsquigarrow \arg\min_h \left\{ \lambda_0 r'_{\beta_0}(h) + h' W_0 + \frac{1}{2} h' H_0 h \right\}
\end{aligned}$$

The second line is due to the uniform in h local quadratic expansion of $n\hat{Q}_n \left(\beta_0 + \frac{h}{\sqrt{n}} \right) - n\hat{Q}_n \left(\beta_0 \right)$, which follows from assumption 2. The last line follows from the following arguments. Assumption 3 implies the Lindeberg Condition is satisfied and $\sqrt{n} (P_n - P) g(\cdot, \beta_0) \rightsquigarrow W_0$. Assumption 5 implies $\frac{r \left(\beta_0 + \frac{h}{\sqrt{n}} \right) - r(\beta_0)}{1/\sqrt{n}} \to r'_{\beta_0}(h)$ for each $h \in \mathbb{R}^d$ and that $r'_{\beta_0}(h)$ is a convex function of h. Since $h' \sqrt{n} \left(\hat{l}_n \left(\beta_0 \right) - l \left(\beta_0 \right) \right) + \frac{1}{2} h' H_0 h + \lambda_n \left(\frac{r \left(\beta_0 + \frac{h}{\sqrt{n}} \right) - r(\beta_0)}{1/\sqrt{n}} \right)$ is a convex function of h , pointwise convergence implies uniform convergence over compact sets $K \subset \mathbb{R}^d$ (Pollard (1991)). Therefore,

$$n\hat{Q}_n \left(\beta_0 + \frac{h}{\sqrt{n}} \right) - n\hat{Q}_n \left(\beta_0 \right) + \lambda_n \sqrt{n} r \left(\beta_0 + \frac{h}{\sqrt{n}} \right) - \lambda_n \sqrt{n} r \left(\beta_0 \right) \rightsquigarrow h' W_0 + \frac{1}{2} h' H_0 h + \lambda_0 r'_{\beta_0}(h)$$

as a process indexed by h in the space of bounded functions $\ell^\infty(K)$ for any compact $K \subset \mathbb{R}^d$. Convexity implies $\lambda_0 r'_{\beta_0}(h) + h'W_0 + \frac{1}{2}h'H_0h$ has a unique minimum, so by the argmin continuous mapping theorem (Theorem 3.2.2 in van der Vaart and Wellner (1996)), $\sqrt{n}\left(\hat{\beta}_n - \beta_0\right) \rightsquigarrow \mathcal{J}$.

Now we show $\hat{\beta}_n^* \xrightarrow{p} \beta_0$. Since $\alpha_n \to 0$ and $\alpha_n\lambda_n \to 0$ imply $\alpha_n\lambda_n r(h+\beta_0) = o(1)$ and $\alpha_n\sqrt{n}\bar{H}_n\left(\hat{l}_n^*(\bar{\beta}_n) - \hat{l}_n(\bar{\beta}_n)\right) = o_p^*(1)$,

$$\begin{aligned}
\hat{\beta}_n^* - \beta_0 &= \arg\min_h \left\{ \alpha_n\lambda_n r(h+\beta_0) + \frac{1}{2}\left\| h + \beta_0 - \bar{\beta}_n + \alpha_n\sqrt{n}\bar{H}_n^{-1}\left(\hat{l}_n^*(\bar{\beta}_n) - \hat{l}_n(\bar{\beta}_n)\right)\right\|_{\bar{H}_n}^2 \right\} \\
&= \arg\min_h \left\{ \frac{1}{2}h'H_0h + h'H_0\left(\beta_0 - \bar{\beta}_n\right) + \frac{1}{2}\left\|\beta_0 - \bar{\beta}_n\right\|_{H_0}^2 \right\} + o_p(1) \\
&= \bar{\beta}_n - \beta_0 + o_p(1) = o_p(1)
\end{aligned}$$

The second line follows from convexity of the proximal bootstrap objective function, which implies the difference between $\alpha_n\lambda_n r(h+\beta_0) + \frac{1}{2}\left\| h + \beta_0 - \bar{\beta}_n + \alpha_n\sqrt{n}\bar{H}_n^{-1}\left(\hat{l}_n^*(\bar{\beta}_n) - \hat{l}_n(\bar{\beta}_n)\right)\right\|_{\bar{H}_n}^2$ and $\frac{1}{2}\left\|h + \beta_0 - \bar{\beta}_n\right\|_{H_0}^2 = \frac{1}{2}h'H_0h + h'H_0\left(\beta_0 - \bar{\beta}_n\right) + \frac{1}{2}\left\|\beta_0 - \bar{\beta}_n\right\|_{H_0}^2$ converges uniformly in probability to zero over any compact subset of $\mathbb{R}^d$.

To derive $\hat{\beta}_n^*$'s asymptotic distribution, first note that because $\sqrt{n}\left(\hat{\beta}_n - \beta_0\right) = O_p(1)$ and $\sqrt{n}\alpha_n \to \infty$,

$$\frac{\hat{\beta}_n^* - \hat{\beta}_n}{\alpha_n} = \frac{\hat{\beta}_n^* - \beta_0}{\alpha_n} - \frac{\sqrt{n}\left(\hat{\beta}_n - \beta_0\right)}{\sqrt{n}\alpha_n} = \frac{\hat{\beta}_n^* - \beta_0}{\alpha_n} + o_p(1)$$

It therefore suffices to show that $\frac{\hat{\beta}_n^* - \beta_0}{\alpha_n} \underset{\mathbb{W}}{\overset{\mathbb{P}}{\rightsquigarrow}} \mathcal{J}$. To do this, use the centered and scaled parameter $h = (\beta - \beta_0)/\alpha_n$:

$$\begin{aligned}
\frac{\hat{\beta}_n^* - \beta_0}{\alpha_n} &= \arg\min_h \left\{ \alpha_n\lambda_n r(\beta_0 + \alpha_n h) + \alpha_n\sqrt{n}\left(\hat{l}_n^*(\bar{\beta}_n) - \hat{l}_n(\bar{\beta}_n)\right)'\left(\beta_0 - \bar{\beta}_n + \alpha_n h\right) + \frac{1}{2}\left\|\beta_0 - \bar{\beta}_n + \alpha_n h\right\|_{\bar{H}_n}^2 \right\} \\
&= \arg\min_h \left\{ \lambda_n\left(\frac{r(\beta_0 + \alpha_n h) - r(\beta_0)}{\alpha_n}\right) + \sqrt{n}\left(\hat{l}_n^*(\bar{\beta}_n) - \hat{l}_n(\bar{\beta}_n)\right)'\left(\frac{\beta_0 - \bar{\beta}_n}{\alpha_n} + h\right) + \frac{1}{2}\left\|\frac{\beta_0 - \bar{\beta}_n}{\alpha_n} + h\right\|_{\bar{H}_n}^2 \right\} \\
&= \arg\min_h \left\{ \lambda_n\left(\frac{r(\beta_0 + \alpha_n h) - r(\beta_0)}{\alpha_n}\right) + h'\sqrt{n}\left(\hat{l}_n^*(\bar{\beta}_n) - \hat{l}_n(\bar{\beta}_n)\right) + \frac{1}{2}h'\bar{H}_n h + o_p^*(1) \right\} \\
&\underset{\mathbb{W}}{\overset{\mathbb{P}}{\rightsquigarrow}} \arg\min_h \left\{ \lambda_0 r'_{\beta_0}(h) + h'W_0 + \frac{1}{2}h'H_0h \right\}
\end{aligned}$$

We have used $\frac{\beta_0 - \bar{\beta}_n}{\alpha_n} = \frac{\sqrt{n}\left(\beta_0 - \bar{\beta}_n\right)}{\sqrt{n}\alpha_n} = o_p(1)$, $\bar{H}_n \xrightarrow{p} H_0$, the assumption of directional differentiability of $r(\beta)$ at β_0, and the following arguments. Assumption 4(i) says $\mathcal{G}_R \equiv \{g(\cdot,\beta) - g(\cdot,\beta_0) : \|\beta - \beta_0\| \leqslant R\}$ is a Donsker class for some $R > 0$, and $P(g(\cdot,\beta) - g(\cdot,\beta_0))^2 \to 0$ for $\beta \to \beta_0$. By Lemma 3.3.5 of van der Vaart and Wellner (1996), $\sqrt{n}(P_n - P)g(\cdot,\beta)$ is stochastically equicontinuous, which implies

$$\left\|\sqrt{n}(P_n - P)\left(g\left(\cdot,\bar{\beta}_n\right) - g(\cdot,\beta_0)\right)\right\| = o_p\left(1 + \sqrt{n}\left\|\bar{\beta}_n - \beta_0\right\|\right) = o_p(1)$$

Stochastic equicontinuity and the envelope integrability condition in assumption 4(ii) imply that the assumptions of Lemma 4.2 in Wellner and Zhan (1996) are satisfied. Therefore, $\sqrt{n}\left(P_n^* - P_n\right) g\left(\cdot, \beta\right)$ is bootstrap equicontinuous, which implies

$$\left\|\sqrt{n}\left(P_n^* - P_n\right)\left(g\left(\cdot, \bar{\beta}_n\right) - g\left(\cdot, \beta_0\right)\right)\right\| = o_p^*\left(1 + \sqrt{n}\left\|\bar{\beta}_n - \beta_0\right\|\right) = o_p^*(1)$$

Therefore, $h'\sqrt{n}\left(\hat{l}_n^*\left(\bar{\beta}_n\right) - \hat{l}_n\left(\bar{\beta}_n\right)\right) = h'\sqrt{n}\left(P_n^* - P_n\right) g\left(\cdot, \beta_0\right) + h'\sqrt{n}\left(P_n^* - P_n\right)\left(g\left(\cdot, \bar{\beta}_n\right) - g\left(\cdot, \beta_0\right)\right) + o_p^*(1) \underset{\mathbb{W}}{\overset{\mathbb{P}}{\rightsquigarrow}} h'W_0$. By convexity, pointwise convergence implies uniform convergence over compact sets $K \subset \mathbb{R}^d$, so

$$\lambda_n\left(\frac{r\left(\beta_0 + \alpha_n h\right) - r\left(\beta_0\right)}{\alpha_n}\right) + h'\sqrt{n}\left(\hat{l}_n^*\left(\bar{\beta}_n\right) - \hat{l}_n\left(\bar{\beta}_n\right)\right) + \frac{1}{2}h'\bar{H}_n h \underset{\mathbb{W}}{\overset{\mathbb{P}}{\rightsquigarrow}} \lambda_0 r'_{\beta_0}(h) + h'W_0 + \frac{1}{2}h'H_0 h$$

as a process indexed by h in the space of bounded functions $\ell^\infty(K)$ for any compact $K \subset \mathbb{R}^d$. $\frac{\hat{\beta}_n^* - \beta_0}{\alpha_n} \underset{\mathbb{W}}{\overset{\mathbb{P}}{\rightsquigarrow}} \mathcal{J}$ follows from the bootstrap version of the argmin continuous mapping theorem (see Lemma 14.2 in Hong and Li (2020)). ■

Monte Carlo Simulation for Finite-dimensional Lasso

We consider the following data generating process:

$$y_i = x_i'\beta_0 + \epsilon_i, \quad \beta_0 = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \end{pmatrix}', x_i \sim N\left(0, I_5 + 0.5\left(\iota\iota' - I_5\right)\right), \epsilon_i \sim N(0,1)$$

We compute the Lasso estimator $\hat{\beta}_n = \arg\min_\beta \left\{\frac{1}{2n}\sum_{i=1}^n \left(y_i - x_i'\beta\right)^2 + \frac{\lambda_n}{\sqrt{n}}\|\beta\|_1\right\}$ using the CVX modeling software in Matlab developed by Grant and Boyd (2009). The proximal bootstrap estimator $\hat{\beta}_n^* = \arg\min_\beta \alpha_n\lambda_n\|\beta\|_1 + \alpha_n\sqrt{n}\left(\hat{l}_n^*\left(\bar{\beta}_n\right) - \hat{l}_n\left(\bar{\beta}_n\right)\right)'\left(\beta - \bar{\beta}_n\right) + \frac{1}{2}\left\|\beta - \bar{\beta}_n\right\|_{\bar{H}_n}^2$, for $\bar{\beta}_n = \hat{\beta}_n$, $\bar{H}_n = \frac{1}{n}\sum_{i=1}^n x_i x_i'$, $\hat{l}_n\left(\bar{\beta}_n\right) = -\frac{1}{n}\sum_{i=1}^n x_i\left(y_i - x_i'\bar{\beta}_n\right)$, and $\hat{l}_n^*\left(\bar{\beta}_n\right) = -\frac{1}{n}\sum_{i=1}^n x_i^*\left(y_i^* - x_i^{*\prime}\bar{\beta}_n\right)$, is computed using the fminunc Matlab function so that we can run the code in parallel (the current version of CVX does not support parallel for loops). We also tried using the fmincon Matlab function, and the results were the same.

We consider five different sample sizes $n \in \{100, 500, 1000, 5000, 10000\}$, three different α_n's for each n: $\alpha_n \in \left\{n^{-1/3}, n^{-1/4}, n^{-1/6}\right\}$, and two choices of $\lambda_n \in \{0.1, 0.5\}$. We use 5000 bootstrap iterations and 2000 Monte Carlo simulations. Empirical coverage frequencies for equal-tailed nominal 95% confidence intervals $\left[\hat{\beta}_n - \frac{c_{97.5}}{\sqrt{n}}, \hat{\beta}_n - \frac{c_{2.5}}{\sqrt{n}}\right]$, where c_τ is the τ-th percentile of $\frac{\hat{\beta}_n^* - \hat{\beta}_n}{\alpha_n}$, and average interval lengths are reported in tables 1-3. Although the proximal bootstrap undercovers for smaller sample sizes, it achieves coverage very close to 95% for sufficiently large n.

Table 1: Proximal Bootstrap Coverage Frequencies and Interval Lengths for $\alpha_n = n^{-1/3}$

	$\lambda_n = 0.1$					$\lambda_n = 0.5$				
n	100	500	1000	5000	10000	100	500	1000	5000	10000
	0.940	0.940	0.945	0.957	0.951	0.933	0.933	0.938	0.958	0.950
	(0.489)	(0.222)	(0.157)	(0.070)	(0.050)	(0.450)	(0.204)	(0.145)	(0.065)	(0.046)
	0.922	0.944	0.946	0.946	0.947	0.919	0.940	0.942	0.950	0.949
	(0.458)	(0.209)	(0.147)	(0.066)	(0.047)	(0.308)	(0.143)	(0.101)	(0.045)	(0.032)
	0.935	0.945	0.942	0.953	0.954	0.934	0.944	0.939	0.953	0.945
	(0.459)	(0.208)	(0.147)	(0.066)	(0.047)	(0.308)	(0.143)	(0.101)	(0.046)	(0.032)
	0.933	0.935	0.948	0.953	0.949	0.936	0.938	0.940	0.945	0.945
	(0.456)	(0.208)	(0.147)	(0.066)	(0.047)	(0.306)	(0.142)	(0.101)	(0.045)	(0.032)
	0.929	0.947	0.953	0.939	0.950	0.936	0.949	0.951	0.938	0.945
	(0.457)	(0.208)	(0.148)	(0.066)	(0.047)	(0.306)	(0.143)	(0.102)	(0.045)	(0.032)

Table 2: Proximal Bootstrap Coverage Frequencies and Interval Lengths for $\alpha_n = n^{-1/4}$

	$\lambda_n = 0.1$					$\lambda_n = 0.5$				
n	100	500	1000	5000	10000	100	500	1000	5000	10000
	0.930	0.940	0.945	0.957	0.952	0.888	0.935	0.940	0.958	0.952
	(0.485)	(0.222)	(0.157)	(0.070)	(0.050)	(0.425)	(0.204)	(0.145)	(0.065)	(0.046)
	0.921	0.944	0.946	0.946	0.948	0.921	0.942	0.943	0.950	0.950
	(0.458)	(0.209)	(0.147)	(0.066)	(0.047)	(0.306)	(0.143)	(0.101)	(0.045)	(0.032)
	0.936	0.945	0.943	0.953	0.954	0.934	0.945	0.939	0.953	0.946
	(0.458)	(0.208)	(0.147)	(0.066)	(0.047)	(0.307)	(0.143)	(0.101)	(0.045)	(0.032)
	0.933	0.935	0.948	0.954	0.950	0.933	0.939	0.940	0.947	0.944
	(0.455)	(0.208)	(0.147)	(0.066)	(0.047)	(0.304)	(0.142)	(0.101)	(0.045)	(0.032)
	0.929	0.947	0.953	0.939	0.950	0.938	0.950	0.952	0.938	0.945
	(0.456)	(0.208)	(0.147)	(0.066)	(0.047)	(0.304)	(0.143)	(0.101)	(0.045)	(0.032)

Table 3: Proximal Bootstrap Coverage Frequencies and Interval Lengths for $\alpha_n = n^{-1/6}$

	$\lambda_n = 0.1$					$\lambda_n = 0.5$				
n	100	500	1000	5000	10000	100	500	1000	5000	10000
	0.913	0.934	0.946	0.958	0.952	0.784	0.902	0.929	0.958	0.953
	(0.462)	(0.220)	(0.157)	(0.070)	(0.050)	(0.349)	(0.190)	(0.143)	(0.065)	(0.046)
	0.921	0.944	0.946	0.946	0.948	0.919	0.941	0.943	0.951	0.950
	(0.457)	(0.208)	(0.147)	(0.066)	(0.047)	(0.302)	(0.142)	(0.101)	(0.045)	(0.032)
	0.934	0.946	0.943	0.953	0.954	0.930	0.944	0.939	0.953	0.946
	(0.458)	(0.208)	(0.147)	(0.066)	(0.047)	(0.303)	(0.142)	(0.101)	(0.045)	(0.032)
	0.933	0.936	0.949	0.953	0.950	0.933	0.937	0.941	0.948	0.945
	(0.455)	(0.207)	(0.147)	(0.066)	(0.047)	(0.300)	(0.142)	(0.101)	(0.045)	(0.032)
	0.928	0.948	0.954	0.940	0.950	0.939	0.951	0.952	0.938	0.947
	(0.456)	(0.208)	(0.147)	(0.066)	(0.047)	(0.301)	(0.142)	(0.101)	(0.045)	(0.032)

We also compare the proximal bootstrap to the standard multinomial bootstrap estimator $\hat{\beta}_n^{**} = \arg\min_{\beta} \left\{ \frac{1}{2n} \sum_{i=1}^{n} \left(y_i^* - x_i^{*\prime} \beta \right)^2 + \frac{\lambda_n}{\sqrt{n}} \|\beta\|_1 \right\}$. Empirical coverage frequencies for equal-tailed nominal 95% confidence intervals $\left[\hat{\beta}_n - \frac{d_{97.5}}{\sqrt{n}}, \hat{\beta}_n - \frac{d_{2.5}}{\sqrt{n}} \right]$, where d_τ is the τ-th percentile of $\sqrt{n} \left(\hat{\beta}_n^{**} - \hat{\beta}_n \right)$, and average interval lengths are reported in table 4. We use 5000 bootstrap iterations and 2000 Monte Carlo simulations. Interestingly, for the case of $\lambda_n = 0.1$, the standard bootstrap coverage frequencies are close to the nominal level. The surprisingly good coverage of the standard bootstrap under certain DGPs is also documented in section 6.2 of Chatterjee and Lahiri (2011). However, when we use $\lambda_n = 0.5$, the standard bootstrap undercovers for the nonzero parameter and overcovers for the zero parameters. Additionally, the standard bootstrap confidence intervals are on average wider than the proximal bootstrap confidence intervals.

Table 4: Standard Bootstrap Coverage Frequencies and Interval Lengths

	$\lambda_n = 0.1$					$\lambda_n = 0.5$				
n	100	500	1000	5000	10000	100	500	1000	5000	10000
	0.947	0.944	0.945	0.961	0.953	0.915	0.917	0.914	0.926	0.920
	(0.509)	(0.224)	(0.158)	(0.071)	(0.050)	(0.474)	(0.211)	(0.149)	(0.067)	(0.047)
	0.950	0.960	0.959	0.966	0.965	0.982	0.986	0.983	0.987	0.991
	(0.477)	(0.211)	(0.149)	(0.067)	(0.047)	(0.329)	(0.150)	(0.107)	(0.048)	(0.034)
	0.963	0.961	0.961	0.969	0.966	0.987	0.991	0.985	0.989	0.990
	(0.478)	(0.211)	(0.149)	(0.067)	(0.047)	(0.328)	(0.151)	(0.107)	(0.048)	(0.034)
	0.959	0.955	0.964	0.967	0.964	0.989	0.982	0.986	0.990	0.991
	(0.474)	(0.211)	(0.149)	(0.067)	(0.047)	(0.327)	(0.151)	(0.107)	(0.048)	(0.034)
	0.958	0.967	0.966	0.954	0.965	0.987	0.993	0.989	0.989	0.993
	(0.476)	(0.211)	(0.149)	(0.067)	(0.047)	(0.327)	(0.150)	(0.107)	(0.048)	(0.034)

References

Belloni, Alexandre and Victor Chernozhukov, "ℓ_1-penalized quantile regression in high-dimensional sparse models," *The Annals of Statistics*, 2011, *39* (1), 82–130. 2

Chatterjee, Arindam and Soumendra Nath Lahiri, "Bootstrapping lasso estimators," *Journal of the American Statistical Association*, 2011, *106* (494), 608–625. 7

Grant, Michael and Stephen Boyd, "CVX: Matlab software for disciplined convex programming," 2009. 5

Hong, Han and Jessie Li, "The numerical bootstrap," *The Annals of Statistics*, 2020, *48* (1), 397–412. 5

Pollard, David, "Asymptotics for least absolute deviation regression estimators," *Econometric Theory*, 1991, *7* (2), 186–199. 3

van der Vaart, AW and Jon A Wellner, *Weak Convergence and Empirical Processes*, Springer, 1996. 3, 4

Wellner, Jon A and Yihui Zhan, "Bootstrapping Z-estimators," *University of Washington Department of Statistics Technical Report*, 1996, *308*. 5

Zhu, Ji, Saharon Rosset, Robert Tibshirani, and Trevor J Hastie, "1-norm support vector machines," in "Advances in neural information processing systems" 2004, pp. 49–56. 2