HIGH-DIMENSIONAL INFERENCE ROBUST TO THE LACK OF MODEL SPARSITY

Jelena Bradic (joint with a PhD student Yinchu Zhu)

www.jelenabradic.net
Assistant Professor
Department of Mathematics
University of California, San Diego
jbradic@ucsd.edu

Setup

Consider a **high dimensional linear regression setting,**

$$\mathsf{Y} = \mathsf{X}\boldsymbol{\gamma}^* + \mathsf{Z}\beta^* + \boldsymbol{\varepsilon}, \tag{1}$$

where $\mathsf{Z} \in \mathbb{R}^n$ and $\mathsf{X} \in \mathbb{R}^{n \times p}$ are the design matrices, $p \gg n$, $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ is the error term independent of the design with $\mathbb{E}(\boldsymbol{\varepsilon}) = 0$ and $\mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top) = \sigma_\varepsilon^2 \mathbb{I}_n$, and $\boldsymbol{\gamma}^*$ and β^* are unknown model parameters.

We focus on the problem of testing single entries of the model parameter, namely the following hypothesis:

$$H_0 : \beta^* = \beta_0, \qquad \text{versus} \qquad H_1 : \beta^* \neq \beta_0. \tag{2}$$

Sparsity assumption: $\|\boldsymbol{\gamma}^*\|_0 := s_\gamma \ll n$ and for inference procedures is such that $s_\gamma \log p / \sqrt{n} \to 0$ as $n \to \infty$.

The sparsity condition is also the backbone of various inference methods for the testing problem (2), including approaches based on consistent variable selection, such as Fan and Li (2001), Wasserman and Roeder (2009), Chatterjee et al. (2013) and recent advances, such as Buhlmann (2013), Zhang and Zhang (2014) (ZZ), Van de Geer et al. (2014) (VBRD), Javanmard and Montanari (2014a) (JM), Belloni et al. (2014) (BCH) and Ning and Liu (2014) (NL).

However, the sparsity condition might not be satisfied in practice. For contemporary large-scale datasets, the sparsity assumption could be quite hard to hold, see e.g., Pritchard (2001), Ward (2009), Jin and Ke (2014), and Hall et al. (2014). Indeed, Kraft and Hunter (2009), Donoho and Jin (2015), Janson et al. (2015) and Dicker (2016) provide evidence suggesting that such an ideal assumption might not be valid.

What happens if we apply sparsity-based methods when the underlying model parameter is not sparse? Can we obtain misleading and spurious results ?

Example 1

⋆ Assume: $\mathbf{X} = \mathbb{I}_p$, ε_i are i.i.d. with $\mathcal{N}(0,1)$ and such that for $a \in [-10, 10]$

$$\beta^* = 0 \qquad \text{and} \qquad \boldsymbol{\gamma}^* = ap^{-1/2}\mathbf{1}_p,$$

⋆ We consider the "de-biasing" approach as formulated in [**?**] Let $\boldsymbol{\pi}^* = (\beta^*, {\boldsymbol{\gamma}^*}^\top)^\top \in \mathbb{R}^{p+1}$ and $\mathbf{W} = (\mathbf{Z}, \mathbf{X}) \in \mathbb{R}^{n \times (p+1)}$. The debiased estimator is then defined $\widetilde{\boldsymbol{\pi}} = \widehat{\boldsymbol{\pi}} + \mathbb{I}_{p+1} W^\top (\mathbf{Y} - \mathbf{W}\widehat{\boldsymbol{\pi}})/n$

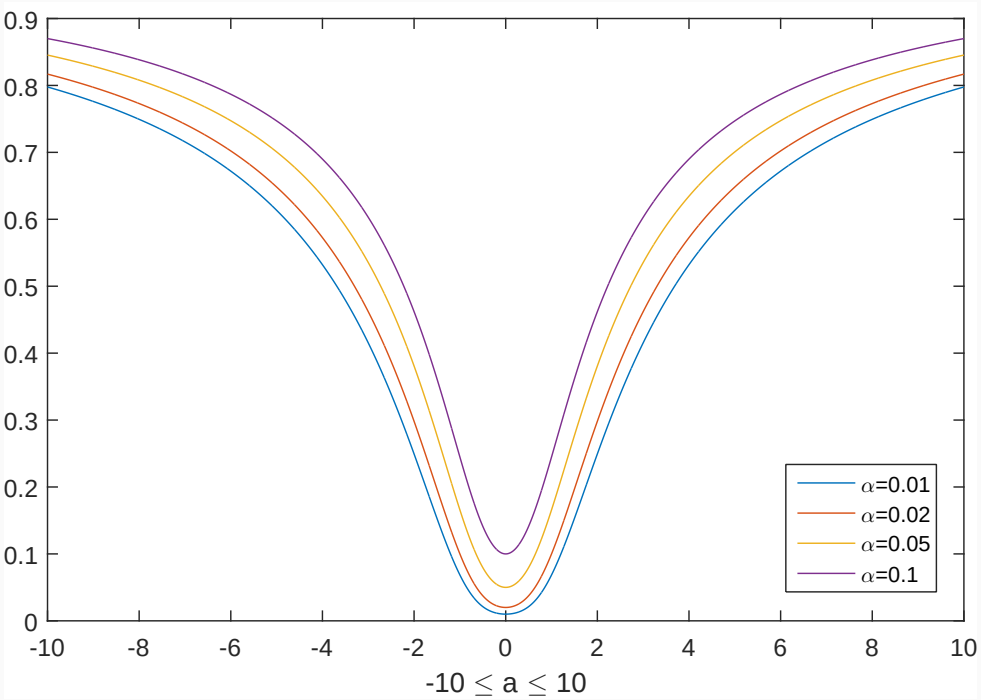
⋆ Wald test rejects the hypothesis whenever $|\widetilde{\boldsymbol{\pi}}_1| > \Phi^{-1}(1 - \alpha/2)/\sqrt{n}$.

Theorem

In the above setup, we have $\lim_{n\to\infty} P\left(|\widetilde{\boldsymbol{\pi}}_1| > \Phi^{-1}(1-\alpha/2)/\sqrt{n}\right) = F(\alpha, a)$, *where* $F(\alpha, a) = 2 - 2\Phi\left[\Phi^{-1}\left(1 - \frac{\alpha}{2}\right)/\sqrt{1+a^2}\right]$.

Rejection Probability

Figure: Plot of the asymptotic Type I error of Wald test



The horizontal axis denotes a and the vertical axis denotes $F(\alpha, a)$.

Goals of our work

- To develop sparsity-robust tests for the hypothesis (2)
  We say that a test is sparsity-robust if the Type I error is asymptotically bounded by the nominal level, regardless of whether or not $\boldsymbol{\gamma}^*$ is sparse.
- We propose a new test, *CorrT*, that has Type I error asymptotically equal to the nominal level without any assumption on the sparsity of $\boldsymbol{\gamma}^*$.
- Our methodology is based on the idea of exploiting the implication of the null hypothesis. Instead of directly estimating the parameter under testing, we test a moment condition that is equivalent to the null hypothesis.
- Moreover, whenever the sparsity condition holds, our method is shown to be optimal and matches existing sparsity-based methods in terms of Type II errors. Therefore, the proposed method, compared to sparsity-based tests, is more robust and does not lose efficiency.

Jelena Bradic (joint with a PhD student Yinchu Zhu)

www.jelenabradic.net Assistant Professor Department of Mathematics University of California San Diego jbradic@ucsd.edu

We observe that a new variable $\mathsf{V} := \mathsf{Y} - \mathsf{Z}\beta_0$ satisfies a linear model

$$\mathsf{V} = \mathsf{X}\boldsymbol{\gamma}^* + \mathsf{e}, \qquad \mathsf{e} = \mathsf{Z}(\beta^* - \beta_0) + \boldsymbol{\varepsilon}.$$

We formally introduce a model to account for the dependence between X and Z:

$$\mathsf{Z} = \mathsf{X}\boldsymbol{\theta}^* + \mathsf{u}, \qquad i = 1, \ldots, n. \tag{3}$$

where $\boldsymbol{\theta}^* \in \mathbb{R}^p$ is sparse and $\mathsf{u} \in \mathbb{R}^n$ is independent of X with mean zero and variance $\mathbb{E}(\mathsf{u}\mathsf{u}^\top) = \sigma_u^2 \mathbb{I}_n$.

We notice that

$$\mathbb{E}\left[(\mathsf{V} - \mathsf{X}\boldsymbol{\gamma}^*)^\top (\mathsf{Z} - \mathsf{X}\boldsymbol{\theta}^*)\right] / n = \sigma_u^2(\beta^* - \beta_0).$$

Hence, solving the inference problem (2) is equivalent to testing

$$H_0 : \mathbb{E}\left[(\mathsf{V} - \mathsf{X}\boldsymbol{\gamma}^*)^\top (\mathsf{Z} - \mathsf{X}\boldsymbol{\theta}^*)\right] = 0, \quad \text{vs} \quad H_1 : \mathbb{E}\left[(\mathsf{V} - \mathsf{X}\boldsymbol{\gamma}^*)^\top (\mathsf{Z} - \mathsf{X}\boldsymbol{\theta}^*)\right] \neq 0. \tag{4}$$

⋆ We define the following estimator

$$\widehat{\gamma} = \widetilde{\gamma}(\widehat{\sigma}_\gamma)\mathbf{1}\{\mathcal{S}_\gamma \neq \emptyset\} + \widetilde{\gamma}(2n^{-1}\|V\|_2)\mathbf{1}\{\mathcal{S}_\gamma = \emptyset\} \tag{5}$$

where $\widehat{\sigma}_\gamma = \arg\max\{\sigma : \sigma \in \mathcal{S}_\gamma\}$ and the set $\mathcal{S}_\gamma$ is defined as

$$\mathcal{S}_\gamma = \left\{\sigma \geq \sqrt{\rho_n}\|V\|_2/\sqrt{n} : \quad 1.5\sigma \geq n^{-1/2}\|V - X\widetilde{\gamma}(\sigma)\|_2 \geq 0.5\sigma\right\}. \tag{6}$$

and

$$\begin{aligned} \widetilde{\gamma}(\sigma) \quad &:= \quad \arg\min_{\gamma\in\mathbb{R}^p}\|\gamma\|_1 \\ &s.t. \quad \|n^{-1}X^\top(V - X\gamma)\|_\infty \leq \eta_0\sigma \\ &\qquad \|V - X\gamma\|_\infty \leq \|V\|_2/\log^2 n \\ &\qquad n^{-1}V^\top(V - X\gamma) \geq \rho_n n^{-1}\|V\|_2^2. \end{aligned} \tag{7}$$

for $\eta_0 = \frac{1.1\Phi^{-1}(1-p^{-1}n^{-1})\sqrt{\max_{1\leq j\leq p} n^{-1}\sum_{i=1}^n x_{i,j}^2}}{n^{1/2}}$, $\rho_n = 0.01/\sqrt{\log n}$.

→ When the estimation target fails to be sparse, the estimator is stable;
→ when the estimation target is sparse, the estimator automatically achieves consistency
→ does not require knowledge of the noise level.

CorrT test statistic

We propose to consider the following correlation test (CorrT) statistic

$$T_n(\beta_0) = \frac{n^{-1/2}(\mathsf{V} - \mathsf{X}\widehat{\boldsymbol{\gamma}})^\top(\mathsf{Z} - \mathsf{X}\widehat{\boldsymbol{\theta}})}{\widehat{\sigma}_\varepsilon \widehat{\sigma}_u}, \tag{8}$$

where $\widehat{\sigma}_\varepsilon = \|\mathsf{V} - \mathsf{X}\widehat{\boldsymbol{\gamma}}\|_2/\sqrt{n}$ and $\widehat{\sigma}_u = \|\mathsf{Z} - \mathsf{X}\widehat{\boldsymbol{\theta}}\|_2/\sqrt{n}$.

Hence, a test with nominal size $\alpha \in (0, 1)$ rejects the hypothesis (2) if and only if $|T_n(\beta_0)| > \Phi^{-1}(1 - \alpha/2)$.

Why does this work ?

We can show, without assuming sparsity of $\boldsymbol{\gamma}^*$, that

$$n^{-1/2}(\mathsf{V} - \mathsf{X}\widehat{\boldsymbol{\gamma}})^\top(\mathsf{Z} - \mathsf{X}\widehat{\boldsymbol{\theta}}) = n^{-1/2}\,(\mathsf{V} - \mathsf{X}\widehat{\boldsymbol{\gamma}})^\top\,\mathsf{u} + O_P(\sqrt{\log p}\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_1),$$

where under the null hypothesis, the first term on the right hand side has zero expectation and the second term vanishes fast enough.

Let $\sigma > 0$, and denote with $\widetilde{\gamma}(\sigma) = \widehat{b}^{+}(\sigma) - \widehat{b}^{-}(\sigma)$, where $\widehat{b}(\sigma) = (\widehat{b}^{+}(\sigma)^{\top}, \widehat{b}^{-}(\sigma)^{\top})^{\top}$ is defined as the solution to the following parametric right hand side linear program

$$\widehat{b}(\sigma) = \underset{b \in \mathbb{R}^{2p}}{\arg\max}\, M_1^{\top} b \qquad \text{subject to} \qquad M_2 b \leq M_3 + M_4 \sigma, \tag{9}$$

where the matrices $M_1 - M_4$ are taken to be

$$M_1 = -\mathbf{1}_{2p \times 1}\,, \; M_2 = \begin{pmatrix} n^{-1}X^{\top}X & -n^{-1}X^{\top}X \\ -n^{-1}X^{\top}X & n^{-1}X^{\top}X \\ X & -X \\ -X & X \\ n^{-1}V^{\top}X & -n^{-1}V^{\top}X \end{pmatrix} \in \mathbb{R}^{(2p+2n+1)\times 2p},$$

$$M_3 = \begin{pmatrix} n^{-1}X^{\top}V \\ -n^{-1}X^{\top}V \\ V + \mathbf{1}_{n\times 1}\|V\|_2/\log^2 n \\ -V + \mathbf{1}_{n\times 1}\|V\|_2/\log^2 n \\ (1-\rho_n)n^{-1}V^{\top}V \end{pmatrix} \in \mathbb{R}^{2p+2n+1}, M_4 = \begin{pmatrix} \eta_0 \mathbf{1}_{p\times 1} \\ \eta_0 \mathbf{1}_{p\times 1} \\ \mathbf{0}_{n\times 1} \\ \mathbf{0}_{n\times 1} \\ 0 \end{pmatrix} \in \mathbb{R}^{2p+2n+1}.$$

Jelena Bradic (joint with a PhD student Yinchu Zhu)

www.jelenabradic.net Assistant Professor Department of Mathematics University of California San Diego jbradic@ucsd.edu

Condition

Let $\mathsf{W} = (\mathsf{Z}, \mathsf{X})$ *and* $\mathbf{w}_i = (z_i, \mathbf{x}_i^\top)^\top$. *The matrix* $\mathbf{\Sigma}_W = \mathbb{E}[\mathsf{W}^\top \mathsf{W}]/n \in \mathbb{R}^{p \times p}$ *satisfies that* $\kappa_1 \leq \sigma_{\min}(\mathbf{\Sigma}_W) \leq \sigma_{\max}(\mathbf{\Sigma}_W) \leq \kappa_2$. *The vectors* $\mathbf{\Sigma}_W^{-1/2}\mathbf{w}_1$ *are centered with sub-Gaussian norms upper bounded by* κ_3 *and* $\mathbb{E}|\varepsilon_1|^{2+\delta} \leq \kappa_4$. *Moreover,* $\log p = o\left(n^{\delta/(2+\delta)} \wedge n\right)$.

→ For the designs, it is standard to impose well-behaved covariance matrices and sub-Gaussian properties.

Condition

$\|\boldsymbol{\gamma}^*\|_2 \leq \kappa_5$ *and* $s_\theta = o\left(\sqrt{n/\log n}/\log p\right)$, *where* $s_\theta = \|\boldsymbol{\theta}^*\|_0$.

→ The assumption on s_θ imposes sparsity in the first row of the precision matrix $\mathbf{\Sigma}_W$ and the rate for s_θ is stronger than the conditions in BCH and NL imposing $o(\sqrt{n}/\log p)$ and in VBRD imposing $o(n/\log p)$.

Theorem

Let Conditions 1 and 2 hold. If $\log p = o\left(n^{\delta/(2+\delta)} \wedge n\right)$, *then under* H_0

$$\forall \alpha \in (0,1), \lim_{n\to\infty} \mathbb{P}\left(|T_n(\beta_0)| > \Phi^{-1}(1-\alpha/2)\right) = \alpha.$$

→ Theorem 2 formally establishes that the new CorrT test is asymptotically exact in testing $\boldsymbol{\beta}^* = \boldsymbol{\beta}_0$. In particular, CorrT is robust to dense γ^* in the sense that even under dense γ^*, our procedure does not generate false positive results.

Remark

Our result is theoretically intriguing as it overcomes limitations of the "inference based on estimation" principle. This principle relies on accurate estimation, which is challenging, if not impossible, for non-sparse models. To see the difficulty, consider the full model parameter $\boldsymbol{\pi}^* := (\beta^*, \boldsymbol{\gamma}^{*\top})^\top \in \mathbb{R}^{p+1}$. *The minimax rate of estimation in terms of* ℓ_2*-loss for parameters in a* $(p+1)$*-dimensional* ℓ_q*-ball with* $q \in [0,1]$ *and of radius* r_n *is* $r_n(n^{-1}\log p)^{1-q/2}$.

Theorem 2 says that CorrT is valid even when $\boldsymbol{\pi}^*$ *cannot be consistently estimated.*

For example, suppose that $\log p \asymp n^c$ *for a constant* $c > 0$ *and* $\boldsymbol{\pi}^*_j = 1/\sqrt{p}$ *for* $j = 1, \ldots, p+1$. *Notice that, as* $n \to \infty$, $\|\boldsymbol{\pi}^*\|_q(n^{-1}\log p)^{1-q/2} \to \infty$ *for any* $q \in [0,1]$, *suggesting potential failure in estimation. Due to this difficulty, it appears quite unrealistic to expect valid inference in individual components of* $\boldsymbol{\pi}^*$ *based on accurate estimates for* $\boldsymbol{\pi}^*$. *In fact, the debiasing technique does not perform well in this case as discussed in Example 1.*

Sparsity-adaptive property

We say that a procedure for testing the hypothesis (2) is sparsity-adaptive if

(i) this procedure does not require knowledge of s_γ,

(ii) provides valid inference under any s_γ and

(iii) achieves efficiency with sparse $\boldsymbol{\gamma}^*$.

We now show the third property, efficiency under sparse $\boldsymbol{\gamma}^*$. To formally discuss our results, we consider testing $H_0 : \beta^* = \beta_0$ versus

$$H_{1,h} : \ \beta^* = \beta_0 + h/\sqrt{n}. \tag{10}$$

where $h \in \mathbb{R}$ is a fixed constant.

Theorem

Let Conditions 1 and 2 hold. Suppose that $s_\gamma = o\left(n/\log(p \vee n)\right)$ *and* $\sigma_u/\sigma_\varepsilon \to \kappa_0$ *for some constant* $\kappa_0 > 0$*. Then, under* $H_{1,h}$ *in (10),*

$$P\left(|T_n(\beta_0)| > \Phi^{-1}(1-\alpha/2)\right) \to \Psi(\alpha, h),$$

where $\Psi(h, \kappa_0, \alpha) = 2 - \Phi\left(\Phi^{-1}(1-\alpha/2) + h\kappa_0\right) - \Phi\left(\Phi^{-1}(1-\alpha/2) - h\kappa_0\right)$.

→ Theorem 3 establishes the local power of CorrT. It turns out that this local power matches that of existing sparsity-based methods, such as VBRD, NL and BCH, that are shown to be efficient.

Jelena Bradic (joint with a PhD student Yinchu Zhu)

www.jelenabradic.net Assistant Professor Department of Mathematics University of California San Diego jbradic@ucsd.edu

Setting

LTD Light-tailed design: $N(0, \Sigma_{(\rho)})$ with the (i, j) entry of $\Sigma_{(\rho)}$ being $\rho^{|i-j|}$.

HTD Heavy-tailed design: each row of W is generated as $\Sigma_{(\rho)}^{1/2} U$, where $U \in \mathbb{R}^n$ contains i.i.d random variables of Student's t-distribution with 3 degrees of freedom normalized to have variance one. (the third moment does not exist.)

The error term $\varepsilon \in \mathbb{R}^n$ contains i.i.d random variables from either $N(0, 1)$ (light-tailed error, or LTE) or Student's t-distribution with 6 degrees of freedom normalized to have variance one (heavy-tailed error, or HTE).

We set

$$\pi_j^* = \begin{cases} 2/\sqrt{n} & 2 \leq j \leq 4 \\ 0 & j > \max\{s, 4\} \\ U(0, 4)/\sqrt{n} & \text{otherwise.} \end{cases}$$

We test the hypothesis

$$H_0 : \pi_3^* = 2/\sqrt{n} + h.$$

Table: Size properties ($h = 0$)

	LTD + LTE, $\rho = 0$			LTD + LTE, $\rho = -\frac{1}{2}$			HTD + HTE, $\rho = 0$		
	CorrT	Debias	Score	CorrT	Debias	Score	CorrT	Debias	Score
$s = 1$	0.03	0.05	0.04	0.05	0.04	0.05	0.06	0.04	0.02
$s = 3$	0.06	0.05	0.05	0.06	0.06	0.05	0.05	0.11	0.03
$s = 5$	0.09	0.09	0.09	0.07	0.11	0.10	0.07	0.04	0.04
$s = 10$	0.01	0.03	0.03	0.03	0.05	0.03	0.06	0.05	0.03
$s = 20$	0.08	0.12	0.11	0.03	0.06	0.06	0.03	0.12	0.04
$s = 50$	0.07	0.16	0.17	0.04	0.10	0.12	0.02	0.09	0.09
$s = 100$	0.05	0.29	0.28	0.01	0.15	0.14	0.05	0.20	0.21
$s = n$	0.04	0.35	0.33	0.04	0.27	0.27	0.04	0.38	0.38
$s = p$	0.07	0.54	0.52	0.04	0.39	0.40	0.05	0.57	0.53
	LTD + HTE, $\rho = 0$			LTD + HTE, $\rho = -\frac{1}{2}$			HTD + LTE, $\rho = 0$		
	CorrT	Debias	Score	CorrT	Debias	Score	CorrT	Debias	Score
$s = 1$	0.03	0.05	0.04	0.04	0.04	0.02	0.06	0.05	0.05
$s = 3$	0.06	0.05	0.05	0.11	0.06	0.06	0.03	0.07	0.04
$s = 5$	0.09	0.09	0.09	0.05	0.06	0.05	0.06	0.11	0.07
$s = 10$	0.01	0.03	0.03	0.03	0.04	0.03	0.09	0.11	0.10
$s = 20$	0.08	0.12	0.11	0.06	0.11	0.10	0.05	0.13	0.06
$s = 50$	0.07	0.16	0.17	0.07	0.16	0.15	0.06	0.19	0.14
$s = 100$	0.05	0.29	0.28	0.05	0.33	0.26	0.05	0.24	0.22
$s = n$	0.04	0.35	0.33	0.05	0.43	0.41	0.05	0.40	0.31
$s = p$	0.07	0.54	0.52	0.06	0.51	0.50	0.06	0.53	0.51

Power curves

Figure: Light-tailed errors