Assessing Uncertainty in High-dimensional Regression Models
Part II

Chen Shizhe

Department of Biostatistics
University of Washington

May 7, 2013

Review

- Marginal associations v.s. conditional associations.
- Reasons for using penalized regressions on high-dimensional data.
- Current attempts to make statistical inference on high-dimensional regressions.

Our goal

$$\underset{\sim}{Y} = \mathbf{X}\underset{\sim}{\beta}^* + \underset{\sim}{\epsilon} = \beta_1^* \underset{\sim}{X}_{(1)} + \mathbf{X}_{(-1)}\underset{\sim}{\beta}^*_{-1} + \underset{\sim}{\epsilon}, \quad \underset{\sim}{\epsilon} \sim N_n(\underset{\sim}{0}, \sigma_\epsilon^2 \mathbf{I}_n). \quad (1)$$

We want to find:

- The p-value for $H_0 : \beta_1^* = 0$ v.s. $H_a : \beta_1^* \neq 0$.
- A $(1 - \alpha)$ confidence interval for β_1^*.

The method in van de Geer et al. (2013)

$$\hat{\underset{\sim}{\beta}} = \underset{\underset{\sim}{\beta} \in \mathbb{R}^p}{\operatorname{argmin}} (\| \underset{\sim}{Y} - \mathbf{X} \underset{\sim}{\beta} \|_2^2 / (2n) + \lambda \| \underset{\sim}{\beta} \|_1). \tag{2}$$

The Karush-Kuhn-Tucker conditions are

$$-\mathbf{X}^T (\underset{\sim}{Y} - \mathbf{X} \hat{\underset{\sim}{\beta}}) + \lambda \hat{\underset{\sim}{\tau}} = \underset{\sim}{0}, \tag{3}$$

$$\| \hat{\underset{\sim}{\tau}} \|_\infty \leq 1, \text{ and } \hat{\tau}_j = \operatorname{sgn} (\hat{\beta}_j) \text{ if } \hat{\beta}_j \neq 0. \tag{4}$$

Note: The sub-gradient for $f(x) = |x|$ is

$$\frac{\partial f}{\partial x} = \begin{cases} 1 & x > 0 \\ \tau, \ \ \tau \in [0, 1] & x = 0 \\ -1 & x < 0. \end{cases}$$

Using the KKT condition, we have

$$n^{-1}\mathbf{X}^T\mathbf{X}(\hat{\underset{\sim}{\beta}} - \underset{\sim}{\beta}^*) + \lambda\hat{\underset{\sim}{\tau}} = \mathbf{X}^T\underset{\sim}{\epsilon}/n. \quad (5)$$

Now assume we have a $\hat{\mathbf{\Theta}}$ that is a "relaxed form" of an inverse of $\hat{\mathbf{\Sigma}} \triangleq n^{-1}\mathbf{X}^T\mathbf{X}$. Multiplying $\hat{\mathbf{\Theta}}$ on both sides of (5) gives:

$$\hat{\underset{\sim}{\beta}} - \underset{\sim}{\beta}^* + \hat{\mathbf{\Theta}}\lambda\hat{\underset{\sim}{\tau}} = \hat{\mathbf{\Theta}}\mathbf{X}^T\underset{\sim}{\epsilon}/n - \underset{\sim}{\Delta}, \quad (6)$$

where $\underset{\sim}{\Delta} = (\hat{\mathbf{\Theta}}\hat{\mathbf{\Sigma}} - \mathbf{I}_p)(\hat{\underset{\sim}{\beta}} - \underset{\sim}{\beta}^*)$.
Recall that:

$$\lambda\hat{\underset{\sim}{\tau}} = \mathbf{X}^T(\underset{\sim}{Y} - \mathbf{X}\hat{\underset{\sim}{\beta}}), \quad (7)$$

then let

$$\hat{\underset{\sim}{b}} = \hat{\underset{\sim}{\beta}} + \hat{\mathbf{\Theta}}\mathbf{X}^T(\underset{\sim}{Y} - \mathbf{X}\hat{\underset{\sim}{\beta}})/n. \quad (8)$$

Under certain conditions, $\sqrt{n}\underset{\sim}{\Delta}$ is asymptotically negligible, then:

$$\sqrt{n}(\hat{\underset{\sim}{b}} - \underset{\sim}{\beta}^*) = \hat{\mathbf{\Theta}}\mathbf{X}^T\underset{\sim}{\epsilon} + o_P(1), \quad \hat{\mathbf{\Theta}}\mathbf{X}^T\underset{\sim}{\epsilon}|\mathbf{X} \sim N_p(\underset{\sim}{0}, \sigma_\epsilon^2\hat{\mathbf{\Theta}}\hat{\mathbf{\Sigma}}\hat{\mathbf{\Theta}}^T). \quad (9)$$

Finding $\hat{\Theta}$

Let $\hat{\underset{\sim}{\gamma}}_j = \arg\min(\|\underset{\sim}{X}_j - \mathbf{X}_{-j}\underset{\sim}{\gamma}\|_2^2/(2n) + \lambda_j\|\underset{\sim}{\gamma}\|_1)$.
Then define

$$\hat{\mathbf{C}} = \begin{pmatrix} 1 & -\hat{\gamma}_{1,2} & \cdots & -\hat{\gamma}_{1,p} \\ -\hat{\gamma}_{2,1} & 1 & \cdots & -\hat{\gamma}_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ -\hat{\gamma}_{p,1} & -\hat{\gamma}_{p,2} & \cdots & 1 \end{pmatrix}, \tag{10}$$

and also

$$\hat{\mathbf{T}}^2 = \mathrm{diag}(\hat{\tau}_1^2, \cdots, \hat{\tau}_p^2), \quad \hat{\tau}_j^2 = (\underset{\sim}{X}_j - \mathbf{X}_{-j}\hat{\underset{\sim}{\gamma}}_j)^T \underset{\sim}{X}_j/n \tag{11}$$

Finally,

$$\hat{\mathbf{\Theta}} = \hat{\mathbf{\Theta}}_{\text{Lasso}} = \hat{\mathbf{T}}^{-2}\hat{\mathbf{C}}. \tag{12}$$

A short summary

- We defined a new estimator for $\underset{\sim}{\beta}^*$:

$$\hat{\underset{\sim}{b}} = \hat{\underset{\sim}{\beta}} + \hat{\Theta}\mathbf{X}^T(\underset{\sim}{Y} - \mathbf{X}\hat{\underset{\sim}{\beta}})/n.$$

- And we claimed that the asymptotic distribution of $\hat{\underset{\sim}{b}}$ is

$$\sqrt{n}(\hat{\underset{\sim}{b}} - \underset{\sim}{\beta}^*) = \hat{\mathbf{\Theta}}\mathbf{X}^T\underset{\sim}{\epsilon} + o_P(1), \quad \hat{\mathbf{\Theta}}\mathbf{X}^T\underset{\sim}{\epsilon}|\mathbf{X} \sim N_p(\underset{\sim}{0}, \sigma_\epsilon^2\hat{\mathbf{\Theta}}\hat{\mathbf{\Sigma}}\hat{\mathbf{\Theta}}^T).$$

One theoretical justification

$$\underset{\sim}{Y} = \beta_1^* \underset{\sim}{X}_{(1)} + \mathbf{X}_{(-1)} \underset{\sim}{\beta}^*_{-1} + \underset{\sim}{\epsilon}, \quad \underset{\sim}{\epsilon} \sim N_n(\underset{\sim}{0}, \mathbf{I}_n). \tag{13}$$

It can be seen as a special case of

$$Y = \beta_1^* X_1 + K(Z) + \epsilon, \ \epsilon \sim N(0, \sigma_\epsilon^2). \tag{14}$$

Theorem (Theorem 2.3 in van de Geer et al. (2013))

Under certain conditions, the limiting variance of $\sqrt{n}(\hat{b}_1 - \beta_1^*)$ reaches the information bound. Furthermore, $\hat{b}_1$ is regular at the one-dimensional parametric sub-model with component β_1 and hence, $\hat{b}_1$ is asymptotically efficient for estimating β_1^0.

The asymptotic distribution

Theorem (Theorem 2.2 in van de Geer et al. (2013))

For the linear model in (1) with Gaussian error $\underset{\sim}{\epsilon} \sim N_n(\underset{\sim}{0}, \sigma_\epsilon^2 \mathbf{I}_n)$. Assume (A2) and the sparsity assumption hold, when using the Lasso for nodewise regression in (8) with $\lambda_j = \lambda_{\max} \asymp \sqrt{\log(p)/n}$, $\forall j$ and the Lasso in (2) with $\lambda \asymp \sqrt{\log(p)/n}$. Then:

$$\begin{aligned} &\sqrt{n}(\underset{\sim}{\hat{b}}_{Lasso} - \underset{\sim}{\beta}^0) = \underset{\sim}{W}_n + \underset{\sim}{\Delta}_n, \\ &\underset{\sim}{W}_n | \mathbf{X} \sim N_p(\underset{\sim}{0}, \sigma_\epsilon^2 \boldsymbol{\Omega}), \ \boldsymbol{\Omega}_n = \hat{\boldsymbol{\Theta}} \hat{\boldsymbol{\Sigma}} \hat{\boldsymbol{\Theta}}^T, \\ &\|\underset{\sim}{\Delta}_n\|_\infty = o_P(1). \end{aligned} \tag{15}$$

Furthermore, $\|\boldsymbol{\Omega}_n - \boldsymbol{\Sigma}^{-1}\|_\infty = o_P(1)$ as $n \to \infty$.

Assumptions

Assumption (Sparsity)

$s_0 = o(n^{1/2}/\log(p))$ *and* $s_j \leq s_{\max} = o(n/\log(p))$.

Assumption (A2)

The rows of X *are i.i.d. realization from a Gaussian distribution* P_X *whose p-dimensional covariance matrix* Σ *has smallest eigenvalue* $\Lambda^2_{\min} \geq L > 0$, *and* $\|\boldsymbol{\Sigma}\|_\infty \triangleq \max_{j,k} |\boldsymbol{\Sigma}_{jk}| = O(1)$.

Simulation study (Bühlmann, 2012)

We let the first s_0 elements of $\underset{\sim}{\beta}^*$ to be b_0, and draw each column of $\mathbf{X}$ from $N_n(\underset{\sim}{0}, \mathbf{I}_n)$. Each model were replicated 500 times. For each replicate, we draw a vector $\underset{\sim}{Y}$ from $N_n(\mathbf{X}\underset{\sim}{\beta}^*, \mathbf{I}_n)$. The parameters in this study are:

- $p = 500$.
- $n \in \{100, 499\}$.
- $s_0 \in \{3, 15\}$.
- $b_0 \in \{0.25, 0.5, 1\}$.
- $\lambda \in \{0.1, 0.5, 1, 2, 4\}$.

The considered type I error is $(p - s_0)^{-1} \sum_{\{j:\beta_j^*=0\}} \mathbb{1}_{[p_j \leq 0.05]}$, and the power $s_0^{-1} \sum_{\{j:\beta_j^* \neq 0\}} \mathbb{1}_{[p_j \leq 0.05]}$.

Simulation results

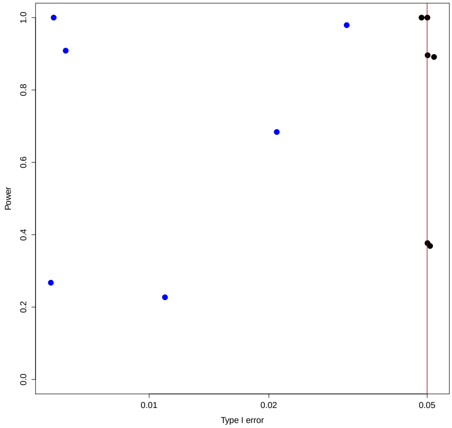

Figure: Power v.s. Type I error, $\lambda = 1$. Colours: $n = 100$, $n = 499$.

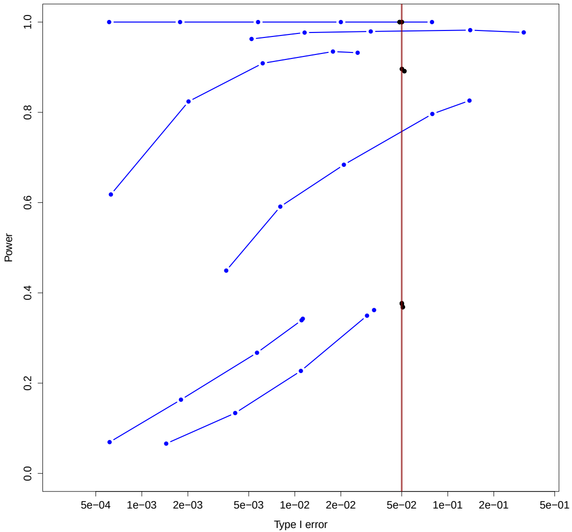

Figure: PvT: $\lambda = 0.1, 0.5, 1, 2, 4$. Colours: $n = 100$, $n = 499$.

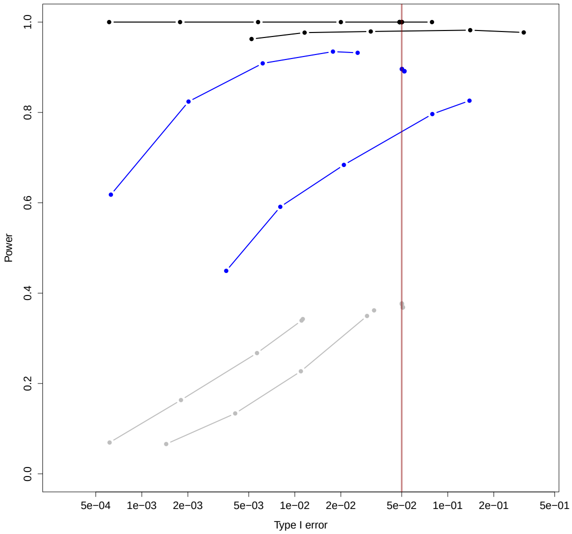

Figure: PvT: $\lambda = 0.1, 0.5, 1, 2, 4$. Colours: $b_0 = 0.25$, $b_0 = 0.5$, $b_0 = 1$.

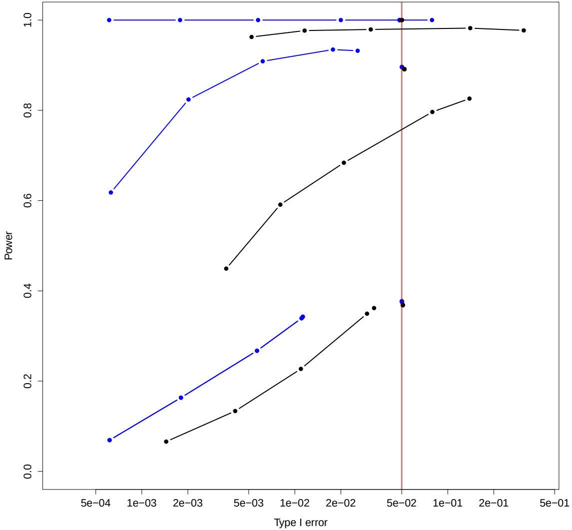

Figure: PvT: $\lambda = 0.1, 0.5, 1, 2, 4$. Colours: $s_0 = 3$, $s_0 = 15$.

Summary II

- The estimation procedure
- A theoretical justification and the asymptotic distribution.
- Some simulation results.

What's left?

- Using scaled lasso to estimate $\hat{\sigma}_\epsilon$ (Sun and Zhang, 2011).
- Regression models with non-Gaussian design, and generalized linear models.
- More simulations.
- ...and all those proofs.

Reference

Peter Bühlmann. Statistical significance in high-dimensional linear models. *arXiv preprint arXiv:1202.1377*, 2012.

Tingni Sun and Cun-Hui Zhang. Scaled sparse linear regression. *arXiv preprint arXiv:1104.4595*, 2011.

Sara van de Geer, Peter Bühlmann, and Ya'acov Ritov. On asymptotically optimal confidence regions and tests for high-dimensional models. *arXiv preprint arXiv:1303.0518*, 2013.