Assessing Uncertainty in High-dimensional Regression Models
A summary of van de Geer et al. (2013)

Chen Shizhe

Department of Biostatistics
University of Washington

May 30, 2013

Outline

- Motivation: marginal/conditional associations.
- Testing procedure: a bias-corrected estimator.
- Properties: Theoretical and numerical.
- Discussion.

Associations: Marginal v.s. Conditional

$$\text{Genes: } \{A^*\}\ \{B_1^*, B_2, B_3\}\ \{C_1, C_2, C_3\}$$

(An example from Grazier G'Sell et al. 2013.)

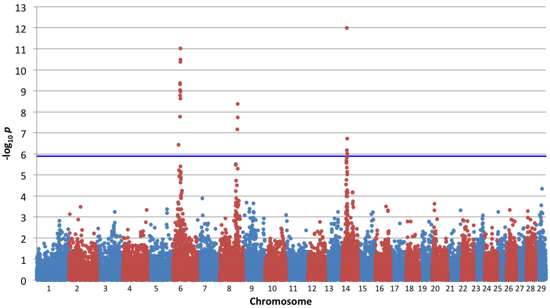


Figure: A Manhattan plot from Nishimura et al. (2012).
$-\log(0.05) \approx 1.3$

Our goal

Consider a linear model:

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

We want to find:

- A p-value for $H_0 : \beta_1^* = 0$ v.s. $H_a : \beta_1^* \neq 0$.
- A p-value for $H_0 : \beta_j^* = 0 \ \forall j \in G$ v.s. $H_a : \exists j \in G, \beta_j^* \neq 0$.
- A $(1 - \alpha)$ confidence interval for β_1^*.

Related work

- Variable selections: Correct recover of support requires the "beta-min" assumption (Meinshausen and Bühlmann 2006, Wainwright 2009, Negahban et al. 2010).
- A significance test for the lasso (Lockhart, Taylor, Tibshirani, and Tibshirani, 2013).
- Bootstrap and subsampling method. In low-dimensional setting, bootstraps for lasso are proposed by Chatterjee and Lahiri (2011), Sartori (2011). For Ridge regression, see Crivelli et al. (1995) and Cule et al. (2011).

Another line of research...

- Zhang and Zhang (2011), and Bühlmann (2012) provided hypothesis testing procedures on high-dimensional linear models.
- Javanmard and Montanari (2013) provided a minimax test for linear models.
- van de Geer et al. (2013) claimed to reach the semiparametric efficiency bound, and the method be generalized to ℓ_1-penalized GLMs and ridge regressions.

Review of linear models

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

When $p < n$, the MLE is

$$\hat{\underset{\sim}{\beta}} = \underset{\underset{\sim}{\beta} \in \mathbb{R}^p}{\operatorname{argmin}} (\| \underset{\sim}{Y} - \mathbf{X}\underset{\sim}{\beta} \|_2^2 / 2n). \tag{3}$$

The Karush-Kuhn-Tucker condition is then

$$-\mathbf{X}^T(\underset{\sim}{Y} - \mathbf{X}\hat{\underset{\sim}{\beta}}) = \underset{\sim}{0}. \tag{4}$$

Let $\hat{\mathbf{\Sigma}} \triangleq n^{-1}\mathbf{X}^T\mathbf{X}$. Further assume that $\hat{\mathbf{\Sigma}}$ is invertible, with $\hat{\mathbf{\Theta}} = \hat{\mathbf{\Sigma}}^{-1}$. Hence,

$$\hat{\underset{\sim}{\beta}} = \frac{1}{n}\hat{\mathbf{\Theta}}\mathbf{X}^T\underset{\sim}{Y} = \underset{\sim}{\beta}^* + \frac{1}{n}\hat{\mathbf{\Theta}}\mathbf{X}^T\underset{\sim}{\epsilon}. \tag{5}$$

ℓ_1-penalized regressions

$$\hat{\underset{\sim}{\beta}} = \underset{\underset{\sim}{\beta} \in \mathbb{R}^p}{\operatorname{argmin}} (\| \underset{\sim}{Y} - \mathbf{X} \underset{\sim}{\beta} \|_2^2 / (2n) + \lambda \| \underset{\sim}{\beta} \|_1). \tag{6}$$

The KKT condition is

$$-\mathbf{X}^T (\underset{\sim}{Y} - \mathbf{X} \hat{\underset{\sim}{\beta}}) + \lambda \hat{\underset{\sim}{\tau}} = \underset{\sim}{0}, \; \hat{\tau}_j = \text{ SGN } (\hat{\beta}_j) \tag{7}$$

Now assume we have a $\hat{\mathbf{\Theta}}$ that is a "relaxed form" of an inverse of $\hat{\mathbf{\Sigma}} \triangleq n^{-1} \mathbf{X}^T \mathbf{X}$.

$$\hat{\underset{\sim}{\beta}} - \underset{\sim}{\beta}^* + \hat{\mathbf{\Theta}} \lambda \hat{\underset{\sim}{\tau}} = \hat{\mathbf{\Theta}} \mathbf{X}^T \underset{\sim}{\epsilon} / n - \underset{\sim}{\Delta}, \tag{8}$$

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

$$\hat{\underset{\sim}{b}} = \hat{\underset{\sim}{\beta}} + \hat{\mathbf{\Theta}} \mathbf{X}^T (\underset{\sim}{Y} - \mathbf{X} \hat{\underset{\sim}{\beta}}) / n. \tag{9}$$

Finding $\hat{\boldsymbol{\Theta}}$: sparse inverse covariance estimation

Let $\hat{\underset{\sim}{\gamma}}_j = \underset{\underset{\sim}{\beta} \in \mathbb{R}^p}{\operatorname{argmin}} (\| \underset{\sim}{X}_j - \mathbf{X}_{-j} \underset{\sim}{\gamma} \|_2^2 / (2n) + \lambda_j \| \underset{\sim}{\gamma} \|_1)$.

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

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

Finally,

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

A short summary

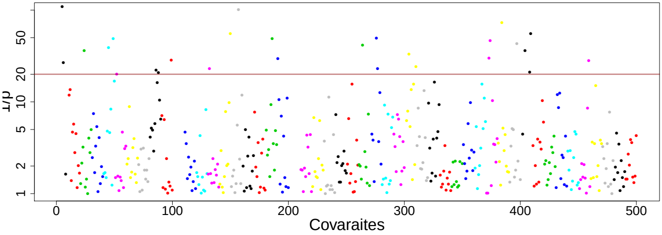
- We defined a new estimator for $\underset{\sim}{\beta}^*$:

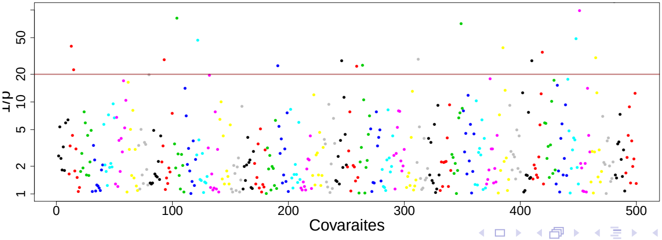
$$\hat{\underset{\sim}{b}} = \hat{\underset{\sim}{\beta}} + \hat{\Theta}\mathbf{X}^T(\underset{\sim}{Y} - \mathbf{X}\hat{\underset{\sim}{\beta}})/n.$$

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

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

One preliminary result:





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 Restricted Eigenvalues and the sparsity assumption hold. When using* $\lambda_j = \lambda_{\max} \asymp \sqrt{\log(p)/n}$, $\forall j$, *and* $\lambda \asymp \sqrt{\log(p)/n}$, *we have:*

$$\begin{aligned} \sqrt{n}(\underset{\sim}{\hat{b}}_{Lasso} - \underset{\sim}{\beta}^0) &= \underset{\sim}{W}_n + \underset{\sim}{\Delta}_n, \\ \underset{\sim}{W}_n | \mathbf{X} \sim N_p(\underset{\sim}{0}, \sigma_\epsilon^2 \mathbf{\Omega}),\ \mathbf{\Omega}_n &= \hat{\mathbf{\Theta}} \hat{\mathbf{\Sigma}} \hat{\mathbf{\Theta}}^T, \\ \|\underset{\sim}{\Delta}_n\|_\infty &= o_P(1). \end{aligned} \tag{13}$$

Furthermore, $\|\mathbf{\Omega}_n - \mathbf{\Sigma}^{-1}\|_\infty = o_P(1)$ *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 (Restricted eigenvalue)

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

... but why?

- Why sparsity?
- Why restrict the minimum eigenvalue?
- Why Gaussian design?

Simulation design

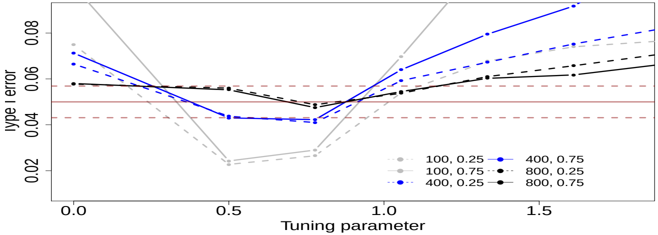
Some setups:

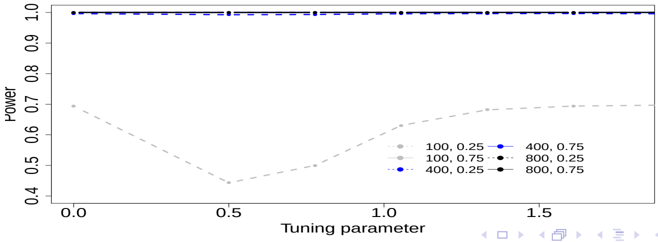
- $\underset{\sim}{\beta}^* = (b_0, b_0, b_0, b_0, b_0, 0, ...0)$.
- $X_i \sim_{\text{iid}} N_n(\underset{\sim}{0}, \mathbf{\Sigma})$, where $\mathbf{\Sigma}$ is a block-diagonal matrix. $(m_1 = 40)$
- $\underset{\sim}{Y} \sim N_n(\mathbf{X}\underset{\sim}{\beta}^*, \mathbf{I}_n)$. $(m_2 = 100)$.

Parameters in this study are:

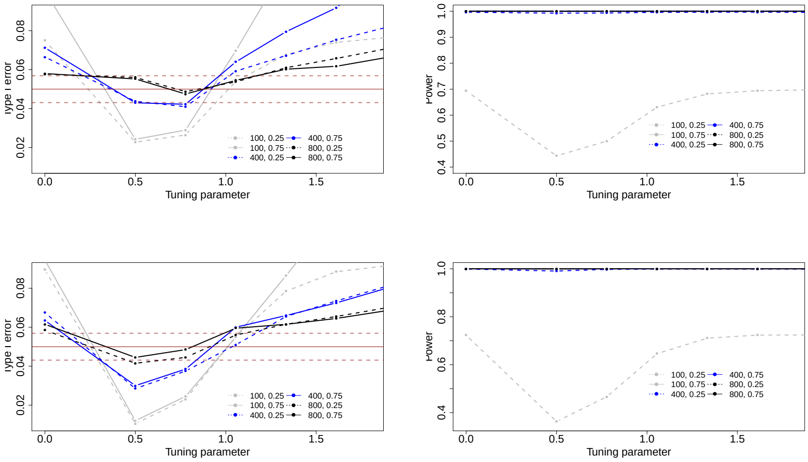
- $p \in \{1000, 2000\}$.
- $n \in \{100, 400, 800\}$.
- $\rho \in \{0, 0.4\}$
- $b_0 \in \{0.25, 0.75\}$.
- λ ranges from 0.5 to 3, and one chosen by BIC.

The baseline graph

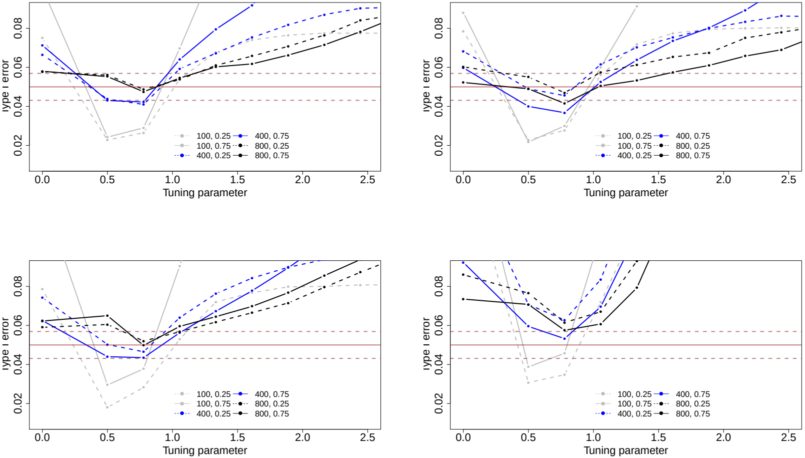




$p = 2000$


Type I error
Power
Tuning parameter
100, 0.25
100, 0.75
400, 0.25
400, 0.75
800, 0.25
800, 0.75

$\rho = 0.4$



Type I error
Tuning parameter
100, 0.25
100, 0.75
400, 0.25
400, 0.75
800, 0.25
800, 0.75

Summary

- The estimation/testing procedure.
- Theoretical justifications.
- Some simulation results.

Discussion

For general linear models, we need to consider

- $\hat{\beta}$ (Van de Geer, 2008),
- $\hat{\Sigma} \triangleq -\ddot{\ell}$: new assumptions.

$\hat{\mathbf{\Theta}}$: "the relaxed inverse",

- Sparse matrix: estimate Θ using other methods, e.g. Glasso (Friedman et al., 2008).
- Matrix of different types.

In application,

- Sample size.
- Tuning parameter.

Reference

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

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

Ana Crivelli, Luis Firinguetti, Rosa Montaño, and Margarita Muñóz. Confidence intervals in ridge regression by bootstrapping the dependent variable: a simulation study. *Communications in Statistics-Simulation and Computation*, 24(3):631–652, 1995.

Erika Cule, Paolo Vineis, and Maria De Iorio. Significance testing in ridge regression for genetic data. *BMC bioinformatics*, 12(1):372, 2011.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

Max Grazier G'Sell, Trevor Hastie, and Robert Tibshirani. False variable selection rates in regression. 2013.

Adel Javanmard and Andrea Montanari. Hypothesis testing in high-dimensional regression under the gaussian random design model: Asymptotic theory. *arXiv preprint arXiv:1301.4240*, 2013.

Richard Lockhart, Jonathan Taylor, Ryan Tibshirani, and Robert Tibshirani. A significance test for the lasso. *arXiv preprint arXiv:1301.7161*, 2013.

N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34(3):1436–1462, 2006.

Sahand Negahban, Pradeep Ravikumar, Martin J Wainwright, and Bin Yu. A unified framework for high-dimensional analysis of *m*-estimators with decomposable regularizers. *arXiv preprint arXiv:1010.2731*, 2010.

Shota Nishimura, Toshio Watanabe, Kazunori Mizoshita, Ken Tatsuda, Tatsuo Fujita, Naoto Watanabe, Yoshikazu Sugimoto, and Akiko Takasuga. Genome-wide association study identified three major qtl for carcass weight including the plag1-chchd7 qtn for stature in japanese black cattle. *BMC genetics*, 13(1):40, 2012.

Samantha Sartori. *Penalized Regressions: Bootstrap confidence intervals and variable selection for high-dimensional data sets*. PhD thesis, Università degli Studi di Milano, 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.

Sara A Van de Geer. High-dimensional generalized linear models and the lasso. *The Annals of Statistics*, 36(2): 614–645, 2008.

M.J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using¡ formula formulatype=. *Information Theory, IEEE Transactions on*, 55(5):2183–2202, 2009.