High-Dimensional Data Analysis and Statistical Inference **April 12, 2015**

Homework 1: High-dimensional regression

Student: Jun Yu *ID : 1401110054*

1. *Gaussian tail bound* Let $Z \sim N(0, \sigma^2)$. Show that

$$\sup_{t>0}(P(Z \geq t)e^{t^2/(2\sigma^2)}) = \frac{1}{2}.$$

Solution:

Note that

$$P(Z \geq t)e^{t^2/(2\sigma^2)} = \int_t^{+\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2-t^2}{2\sigma^2}} dx.$$

After derivation calculus of t to above equation, we know $P(Z \geq t)e^{t^2/(2\sigma^2)}$ is a decreasing function of t for $t \geq 0$. So the desired result follows from the fact that $P(Z \geq t)e^{t^2/(2\sigma^2)}$ achieved its upper bound when $t = 0$.

2. *Irrepresentable condition* Consider the covariance matrix $\Sigma = (\sigma_{ij})$ with an autoregressive Toeplitz structure: $\sigma_{ij} = \rho^{|i-j|}$ for all i and j with $0 < |\rho| < 1$. Show that the irrepresentable condition

$$\|\Sigma_{S^cS}(\Sigma_{SS})^{-1}sgn(\beta_S^*)\|_\infty \leq \alpha < 1$$

holds and identify the constant α.

Solution:

Without loss of generality, let us assume $x_j \sim_{i.i.d} N(0; \Sigma), j = 1, \cdots, n$ are i.i.d random variables. Then the power decay design implies an AR(1) model where

$$\begin{aligned}
x_{j1} &= \eta_{j1} \\
x_{j2} &= \rho x_{j1} + (1-\rho^2)^{1/2}\eta_{j2} \\
&\vdots \\
x_{jp} &= \rho x_{j(p-1)} + (1-\rho^2)^{1/2}\eta_{jp}
\end{aligned}$$

where η_{ij} are i.i.d N(0,1) random variables. Thus, the predictors follow a Markov Chain:

$$x_{j1} \rightarrow x_{j2} \rightarrow \cdots \rightarrow x_{jp}$$

Now let

$$I_1 = i : \beta_i \neq 0; I_2 = i : \beta_i = 0$$

for $\forall k \in I_2$, assume

$$k_l = \{i : i < k\} \bigcap I_1; k_h = \{i : i > k\} \bigcap I_1.$$

Then by the Markov property, we have

$$x_{jk} \perp x_{jg} | (x_{jk_l}, x_{jk_h})$$

for $j = 1, \cdots, n$ and $\forall g \in I_1 / \{k_l, k_h\}$. Therefore to check Strong Irrepresentable Condition for x_{jk} we only need to consider x_{jk_l} and x_{jk_h} since the rest of the entries are zero by the conditional independence. To further simplify, we assume $\rho \geq 0$. Now regressing x_{jk} on (x_{jk_l}, x_{jk_h}) we get

$$cov(\begin{pmatrix} x_{jk_l} \\ x_{jk_h} \end{pmatrix})^{-1} cov(x_{jk}, \begin{pmatrix} x_{jk_l} \\ x_{jk_h} \end{pmatrix}) = \begin{pmatrix} \frac{\rho^{k_l-k} - \rho^{k-k_l}}{\rho^{k_l-k_h} - \rho^{k_h-k_l}} \\ \frac{\rho^{k_h-k} - \rho^{k-k_h}}{\rho^{k_l-k_h} - \rho^{k_h-k_l}} \end{pmatrix}$$

Then sum of both entries follow

$$\frac{\rho^{k_l-k} - \rho^{k-k_l}}{\rho^{k_l-k_h} - \rho^{k_h-k_l}} + \frac{\rho^{k_h-k} - \rho^{k-k_h}}{\rho^{k_l-k_h} - \rho^{k_h-k_l}} = 1 - \frac{(1 - \rho^{k_l-k})(1 - \rho^{k-k_h})}{1 + \rho^{k_l-k_h}} \leq 1 - frac(1 - c_k)^2 2$$

where c_k is a constant which can be got from mean value inequality. Therefore Strong Irrepresentable Condition holds entry-wise, and $\alpha = 1 - \frac{(1-\max(c_k))^2}{2}$.
This is the Corollary 3 in Zhao and Yu(2006)[1] , more detail can be found in this paper.

3. *Model size of the Lasso* Assume what we have done in class. Prove inequality (7.9) in Bickel, Ritov and Tsybakov (2009): with probability tending to 1, the Lasso estimator $\hat{\beta}_L$ satisfies

$$\mathcal{M}(\hat{\beta}_L) \leq \frac{64\phi_{max}}{\kappa^2(s,3)} s$$

where $\mathcal{M}(\beta) = \sum_{j=1}^{p} I(\beta_j \neq 0)$ and ϕ_{max} is the maximal eigenvalue of $\mathbf{X^T X}/n$

Solution:
In class we have proof the following theorem:

Theorem (Bickel,Ritov and Tsybakov (2009) Theorem 7.2)**.** *Let W_i be independent $N(0, \sigma^2)$ random variables with $\sigma^2 > 0$. Let all the diagonal elements of the matrix $\mathbf{X^T X}/n$ be equal to 1, and let $\mathcal{M}(\beta^*) \leq s$, where $1 \leq s \leq M, n \geq 1, M \geq 2$. Let Assumption RE(s, 3) be satisfied. Consider the Lasso estimator $\hat{\beta}_L$ defined by*

$$\hat{\beta}_L = \arg\min_{\beta \in \mathbb{R}^M} \{\frac{1}{n} \|\boldsymbol{y} - X\beta\|_2^2 + 2r\|\beta\|_1\}$$

with $r = A\sigma\sqrt{\frac{\log M}{n}}$ and $A > 2\sqrt{2}$.Then, with probability at least $1 - M^{1-A^2/8}$, we have

$$\|X(\hat{\beta}_L - \beta^*)\|_2^2 \leq \frac{16A^2}{\kappa^2(s,3)} \sigma^2 s \log M$$

And we use the lemma B.1 in Bickel,Ritov and Tsybakov (2009) for the linear model case, we know:

Theorem (Bickel,Ritov and Tsybakov (2009) Lemma B.1)**.** *Let W_i be independent $N(0,\sigma^2)$ random variables with $\sigma^2 > 0$. Let the Lasso estimator $\hat{\beta}_L$ defined by*

$$\hat{\beta}_L = \arg\min_{\beta\in\mathbb{R}^M}\{\frac{1}{n}\|\boldsymbol{y} - X\beta\|_2^2 + 2r\|\beta\|_1\}$$

with $r = A\sigma\sqrt{\frac{\log M}{n}}$ and $A > 2\sqrt{2}$.Then, with probability at least $1 - M^{1-A^2/8}$, we have

$$\mathcal{M}(\hat{\beta}_L) \leq 4\phi_{\max}(\|X(\hat{\beta}_L - \beta^*)\|_n^2/r^2),$$

where $\|X(\hat{\beta}_L - \beta^*)\|_n = \sqrt{\frac{1}{n}\sum_{i=1}^n(\hat{\beta}_L - \beta^*)^2}$.

The result is obvious.

The more details and proof of the above theorems can be found in Bickel,Ritov and Tsybakov (2009)[2].

4. *Residual of the Dantzig selector* For the linear regression model, show that, with probability tending to 1, the residual $\delta = \hat{\beta}_D - \beta^*$ of the Dantzig selector $\hat{\beta}_D$ satisfies $\|\delta_{S^c}\|_1 \leq \|\delta_S\|_1$.

   Solution:

   Note the definition of the Dantzig selector is

$$\min_{\beta\in\mathbb{R}^p}\|\beta\|_1; \qquad \text{subject to } \|\boldsymbol{X}^T(y - X\boldsymbol{\beta})\|_\infty \leq r$$

   here r is a constant.

   From the definition we know $\|\hat{\beta}_D\| \leq \|\beta^*\|$, hence

$$\sum_{j\in S}|\beta_j^* - \hat{\beta}_{Dj}| \geq \sum_{j\in S}|\beta_j^*| - \sum_{j\in S}|\hat{\beta}_{Dj}| \geq \sum_{j\in S^C}|\hat{\beta}_{Dj}|$$

   where the first inequality follows by triangle inequality and second inequality follows by the fact $\|\hat{\beta}_D\| \leq \|\beta^*\|$.

5. *Dirichlet - multinomial regression* Derive the log-likelihood function of the Dirichlet - multinomial regression model (eq. (7) in Chen and Li (2013)).

   Solution:

   Note the joint Dirichlet-multinomial (DM) distribution has the density:

$$f_{DM}(y_1,\cdots,y_q;\gamma) = \binom{\gamma_+}{y}\frac{\Gamma(y_+ + 1)\Gamma(\gamma_+)}{\Gamma(y_+ + \gamma_+)}\prod_{j=1}^{q}\frac{\Gamma(y_j + \gamma_j)}{\Gamma(\gamma_j)\Gamma(y_j + 1)}$$

   where $y_+ = \sum_{j=1}^q y_j$ and $\gamma_+ = \sum_{j=1}^q \gamma_j$. We also note the link function is

$$\gamma_j(\mathbf{x}^i, \boldsymbol{\beta}^j) = \exp(\alpha_j + \sum_{k=1}^{p}\beta_{jk}x_{ik}).$$

Substituting the link function into DM probability function and ignoring the part that does not involve the parameters, we get the desired result.

6. *Coordinate descent with nonconvex penalties* In the proof of Theorem 4 in Mazumder, Friedman and Hastie (2011), the authors arrived at inequality (A.13):

$$Q(\beta^m) - Q(\beta^{m+1}) \geq \theta \|\beta^{m+1} - \beta^m\|_2^2,$$

from which they concluded that the sequence $\{\beta^k\}$ converges. Prove or disprove this claim.

Solution:

Suppose $\beta^{m+1} - \beta^m = \frac{1}{m}$, than the sequence $\{\beta^k\}$ satisfied the condition in Mazumder, Friedman and Hastie (2011), but it's not converges, due to the divergence of series $\{\frac{1}{n}\}$.

More condition should be add to claim the convergence of coordinate descent,which can be found in Lin and Lv(2013)[3].

7. *ADMM for group Lasso* Derive the ADMM algorithm in scaled form for the group Lasso problem

$$\min_{\beta}\{\frac{1}{2}\|\mathbf{y} - \mathbf{x}\boldsymbol{\beta}\|_2^2 + \lambda \sum_{g=1}^{G} \|\boldsymbol{\beta}_g\|_2\}$$

where $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \cdots, \boldsymbol{\beta}_G^T)^T$.

Solution:

In ADMM form, the lasso problem can be written as

$$\begin{aligned} &\text{minimise } f(\beta) + g(\gamma) \\ &\text{subject to } \beta_g - \tilde{\gamma}_g = 0, \ g = 1, \cdots, G \end{aligned}$$

with local variables β_g and global variable γ. Here, $\tilde{\gamma}$ is the global variable γ's idea of what the local variable β_g should be, and is given by a linear function of γ.

Hence, the ADMM algorithm is

$$\begin{aligned} \beta^{k+1} &:= (\mathbf{x^T x} + \rho \mathbf{I})^{\mathbf{-1}}(\mathbf{x^T y} + \rho(\gamma^{\mathbf{k}} - \mathbf{u^k})) \\ \tilde{\gamma}^{k+1} &:= S_{\lambda/\rho}(\beta^{k+1} + u^k) \\ u^{k+1} &:= u^k + \beta^{k+1} - \gamma^{k+1} \end{aligned}$$

where γ-update is block soft thresholding

$$\tilde{\gamma}_i^{k+1} = S_{\lambda/\rho}(\beta_g^{k+1} + u^k), g = 1, \cdots, G,$$

which defined as $S_\kappa : \boldsymbol{R}^m \to \boldsymbol{R}^m$

$$S_\kappa(a) = (1 - \kappa/\|a\|_2)_+ a,$$

with $S_\kappa(0) = 0$.

8. *Bayesian elastic net* Recall that the elastic net penalty is $\lambda_1\|\beta\|_1 + \lambda_2\|\beta\|_2^2$. Derive a hierarchical representation of the Bayesian elastic net similar to eq. (5) in Park and Casella (2008).

   Solution:

   Under Park and Casella's assumptions, solving the en problem is equivalent to finding the marginal posterior mode of $\beta|y$ when the prior distribution of β is given by

$$\pi(\beta) \propto \exp\{-\lambda_1\|\beta\|_1 - \lambda_2\|\beta\|_2^2\}.$$

   We also use the improper prior density

$$\pi(\sigma^2) \propto \frac{1}{\sigma^2}.$$

   Based on the discussion above, we have the following hierarchial model.

$$\begin{aligned} \boldsymbol{y}|\beta,\sigma^2 &\sim N(\boldsymbol{X}\beta,\sigma^2 I_n), \\ \beta|\sigma^2 &\sim \exp\{-\frac{1}{2\sigma^2}(\lambda_1\|\beta\|_1 + \lambda_2\|\beta\|_2^2)\}, \\ \sigma^2 &\sim \frac{1}{\sigma^2}. \end{aligned}$$

   Firstly, note that

$$\exp\{-\frac{1}{2\sigma^2}(\lambda_1\|\beta\|_1 + \lambda_2\|\beta\|_2^2)\} = \prod_{j=1}^{p} \exp\{-\frac{1}{2\sigma^2}(\lambda_1|\beta_j| + \lambda_2\beta_j^2)\}.$$

   Secondly, using (4) in Park and Casella (2008), we have

$$\exp(-\frac{\lambda_1}{2\sigma^2}|\beta_j|) \propto \int_0^\infty \frac{1}{\sqrt{s}}\exp(-\frac{\beta_j^2}{2s})\exp(-\frac{\lambda_1^2}{8\sigma^4}s)ds.$$

   Then, we have

$$\exp\{-\frac{1}{2\sigma^2}(\lambda_1\|\beta\|_1 + \lambda_2\|\beta\|_2^2)\} \propto \int_1^\infty \sqrt{\frac{t}{t-1}}\exp(-\frac{\beta_j^2\lambda_2 t}{2\sigma^2(t-1)})t^{-1/2}\exp(-\frac{\lambda_1^2 t}{8\lambda_2\sigma^2})dt.$$

   This implies that we can treat$\beta_j|\sigma^2$ as a mixture of normal distributions, $N(0,\sigma^2(t-1)/(\lambda_2 t))$, where the mixing distribution is over the variance $\sigma^2(t-1)/(\lambda_2 t)$ and is given by a truncated gamma distribution with shape parameter $1/2$, scale parameter $8\lambda_2\sigma^2/\lambda_1^2$ and support $(1,\infty)$,which we denote as trG(1/2,$8\lambda_2\sigma^2/\lambda_1^2$).

   Hence,we have the following hierarchical model.

$$\begin{aligned} \boldsymbol{y}|\beta,\sigma^2 &\sim N(\boldsymbol{X}\beta,\sigma^2 I_n), \\ \beta|\sigma^2,\boldsymbol{\tau} &\sim \prod N(0,\frac{\sigma^2(\tau_j-1)}{\lambda_2\tau_j}), \\ \boldsymbol{\tau} &\sim \prod_{j=1}^{p} trG(1/2, 8\lambda_2\sigma^2/\lambda_1^2), \\ \sigma^2 &\sim \frac{1}{\sigma^2}. \end{aligned}$$

References

[1] P. Zhao and B. Yu, "On model selection consistency of lasso," *The Journal of Machine Learning Research*, vol. 7, pp. 2541–2563, 2006.

[2] P. J. Bickel, Y. Ritov, and A. B. Tsybakov, "Simultaneous analysis of lasso and dantzig selector," *The Annals of Statistics*, pp. 1705–1732, 2009.

[3] W. Lin and J. Lv, "High-dimensional sparse additive hazards regression," *Journal of the American Statistical Association*, vol. 108, no. 501, pp. 247–264, 2013.