41500: High-Dimensional Statistics **Winter 2020**

Debiased Lasso

Lecturer: Chao Gao *Scribe: Xinze Li*

In previous lectures, we obtain estimators of precision matrix. Now it's time for us to do inference, which is a harder problem.

1 Debiased Lasso

First suppose that

$$y = X\beta + z, \quad z \sim \mathcal{N}(0, I_n) \tag{1}$$

Also suppose that $X = (X_1, \cdots, X_n)^\top$, $X_i \in \mathbb{R}^p$ and that X_i are i.i.d. drawn from multivariate normal distribution $\mathcal{N}(0, \Omega^{-1})$, where $\Omega = \Sigma$ is the precision matrix. And so we could write the ordinary least square estimator if $n > p$

$$\widehat{\beta}_{OLS} = \left(X^\top X\right) X^\top y \sim \mathcal{N}\left(\beta, \left(X^\top X\right)^{-1}\right) \tag{2}$$

Also suppose that $\widetilde{\beta}$ is the Lasso estimator

$$\widetilde{\beta} = \arg\max_{\beta} \left(\|y - X\beta\|^2 - \lambda \|\beta\|_1\right) \tag{3}$$

We know that $\mathbf{E}y = X\beta$ and thus $\mathbf{E}X^\top y = X^\top X\beta$. And so

$$\widetilde{\beta} - \widehat{\beta}_{OLS} = \left(X^\top X\right)^{-1} \left(X^\top X\widetilde{\beta} - X^\top y\right) \tag{4}$$

is approximately the bias. A tautology from this equation is

$$\begin{aligned} \widehat{\beta} &= \widetilde{\beta} + \left(X^\top X\right)^{-1} \left(X^\top y - X^\top X\widetilde{\beta}\right) \\ &= \widetilde{\beta} + \left(\frac{1}{n}X^\top X\right)^{-1} \cdot \frac{1}{n}\left(X^\top y - X^\top X\widetilde{\beta}\right) \end{aligned} \tag{5}$$

So now our idea is to replace $\left(\frac{1}{n}X^\top X\right)^{-1}$ by a matrix close to the precision matrix. Let's call this approximation matrix M, and now the in general we could write

$$\widehat{\beta} = \widetilde{\beta} + M \cdot \frac{1}{n}\left(X^\top y - X^\top X\widetilde{\beta}\right) \tag{6}$$

This idea first appeard in the article [Zhang and Zhang, 2011]. It is also rediscovered in article [Javanmard and Montanari, 2015] and [van de Geer et al., 2014]. The results in the latter two papers are a subset of [Zhang and Zhang, 2011], which is hard to read.

1.1 Known Covariance

If Σ the covariance matrix is known, then a natural attempt is to plug in $M = \Omega = \Sigma^{-1}$. Suppose $\widehat{\Sigma} = \frac{1}{n} X^\top X$ is the sample covariance matrix, we have

$$\begin{aligned} \widehat{\beta} &= \widetilde{\beta} + \frac{1}{n}\Omega X^\top \left(y - X\widetilde{\beta}\right) \\ &= \widetilde{\beta} + \frac{1}{n}\Omega X^\top \left(X\beta + z - X\widetilde{\beta}\right) \\ &= \widetilde{\beta} + \frac{1}{n}\Omega X^\top X\beta + \frac{1}{n}\Omega X^\top z - \frac{1}{n}\Omega X^\top X\widetilde{\beta} \\ &= \beta + \frac{1}{n}\Omega X^\top z + \left(I_p - \Omega\widehat{\Sigma}\right)\left(\widetilde{\beta} - \beta\right) \end{aligned} \tag{7}$$

Rearranging the order gives

$$\sqrt{n}\left(\widehat{\beta} - \beta\right) = \frac{1}{\sqrt{n}}\Omega X^\top z + \sqrt{n}\left(I_p - \Omega\widehat{\Sigma}\right)\left(\widetilde{\beta} - \beta\right) \tag{8}$$

Since inference is applied on each component, let's analyze each component of this vector.

$$\sqrt{n}\left(\widehat{\beta}_j - \beta_j\right) = \frac{1}{\sqrt{n}}\Omega_j^\top X^\top z + \sqrt{n}\left(e_j - \widehat{\Sigma}\Omega_j\right)^\top \cdot \left(\widetilde{\beta} - \beta\right) \tag{9}$$

Clearly the first term of the RHS is Gaussian given X.

$$\frac{1}{\sqrt{n}}\Omega_j X^\top z | X \sim \mathcal{N}\left(0, \Omega_j^\top \widehat{\Sigma}\Omega_j\right) \tag{10}$$

And for the second term, we use Holder's inequality

$$\left|\sqrt{n}\left(e_j - \widehat{\Sigma}\Omega_j\right)^\top \left(\widetilde{\beta} - \beta\right)\right| \le \sqrt{n}\left\|e_j - \widehat{\Sigma}\Omega_j\right\|_\infty \cdot \left\|\widetilde{\beta} - \beta\right\|_1 \tag{11}$$

For the latter part of the RHS $\left\|\widetilde{\beta} - \beta\right\|_1$, we know from properties of lasso estimator that w.h.p.

$$\left\|\widetilde{\beta} - \beta\right\|_1 \lesssim s\sqrt{\frac{\log p}{n}} \tag{12}$$

And for the former part of RHS, we have

$$\begin{aligned} &\left\|e_j - \widehat{\Sigma}\Omega_j\right\|_\infty \\ &= \max_{1 \le l \le p} \frac{1}{n}\left|\sum_{i=1}^{n}\left[\left(\Omega_j^\top X_i\right)\left(e_l^\top X_i\right) - \mathbf{E}\left(\Omega_j^\top X_i\right)\left(e_l^\top X_i\right)\right]\right| \\ &\le \sqrt{\frac{\log p}{n}} \end{aligned} \tag{13}$$

Details see HW5 Problem2 solution. So combining equation 13 and 12 and plugging the bound into 11 gives the following upper bound

$$\left|\sqrt{n}\left(e_j - \Omega_j^\top \hat{\Sigma}\right)\left(\widetilde{\beta} - \beta\right)\right| \lesssim s\frac{\log p}{\sqrt{n}} \tag{14}$$

Thus, if $\frac{s\log p}{\sqrt{n}} \to 0$ and $\Omega_j^\top \widehat{\Sigma}\Omega_j \gtrsim 1$ not degenerate w.h.p. (to be proved), then

$$\frac{\sqrt{n}\left(\widehat{\beta}_j - \beta_j\right)}{\sqrt{\Omega_j^\top \widehat{\Sigma}\Omega_j}} \rightsquigarrow \mathcal{N}(0,1) \tag{15}$$

1.2 Unknown Covariance

Following the process as in eqs. (7) to (11), suppose $M = (m_1, m_2, \cdots, m_p)^\top$ and that M depends only on X, we deduce that

$$\begin{aligned}
\widehat{\beta} &= \beta + \frac{1}{n}MX^\top z + \left(I_p - M\widehat{\Sigma}\right)\left(\widetilde{\beta} - \beta\right) \\
\sqrt{n}\left(\widehat{\beta} - \beta\right) &= \frac{1}{\sqrt{n}}MX^\top z + \sqrt{n}\left(I_p - M\widehat{\Sigma}\right)\left(\widetilde{\beta} - \beta\right) \\
\sqrt{n}\left(\widehat{\beta}_j - \beta_j\right) &= \frac{1}{\sqrt{n}}m_j^\top X^\top z + \sqrt{n}\left(e_j - \widehat{\Sigma}m_j\right)^\top\left(\widetilde{\beta} - \beta\right) \\
&\frac{1}{\sqrt{n}}m_j^\top X^\top z | X \sim \mathcal{N}\left(0, m_j^\top \widehat{\Sigma}m_j\right) \\
\left|\sqrt{n}\left(e_j - \widehat{\Sigma}m_j\right)^\top\left(\widetilde{\beta} - \beta\right)\right| &\leq \sqrt{n}\left\|e_j - \widehat{\Sigma}m_j\right\|_\infty \cdot \left\|\widetilde{\beta} - \beta\right\|_1
\end{aligned} \tag{16}$$

In order to obtain the normality, note that we should restrict $\left\|e_j - \widehat{\Sigma}m_j\right\|_\infty$ to a scale of $\sqrt{\frac{\log}{n}}$ such that $\left|\sqrt{n}\left(e_j - \widehat{\Sigma}m_j\right)\left(\widetilde{\beta} - \beta\right)\right| \lesssim s\frac{\log p}{\sqrt{n}}$ as before. In the same time, we would like the confidence interval be as short as possible. In order to do this, we introduce the following convex program (j is fixed)

$$\begin{aligned}
\min_{m \in \mathbb{R}^p} \quad & m^\top \widehat{\Sigma}m \\
s.t. \quad & \left\|e_j - \widehat{\Sigma}m_j\right\|_\infty \leq C\sqrt{\frac{\log p}{n}}
\end{aligned} \tag{17}$$

Note that this problem is certainly feasible since we have just proved that Ω_j is a feasible point in 13. Let m_j be the solution to 17, then as we previously stated, if $\frac{s\log p}{\sqrt{n}} \to 0$ and $m_j^\top \widehat{\Sigma}m_j \gtrsim 1$ not degenerate w.h.p. (to be proved), then

$$\frac{\sqrt{n}\left(\widehat{\beta}_j - \beta_j\right)}{\sqrt{m_j^\top \widehat{\Sigma}m_j}} \rightsquigarrow \mathcal{N}(0,1) \tag{18}$$

Now let's prove that $m_j^\top \widehat{\Sigma} m_j$ is indeed bounded away from zero. Suppose $\mu = C\sqrt{\frac{\log p}{n}}$. Since $\left\| e_j - \widehat{\Sigma} m_j \right\|_\infty \le \mu$, we have

$$\left| e_j^\top \widehat{\Sigma} m_j - 1 \right| \le \mu \tag{19}$$

And so

$$\begin{aligned} m^\top \Sigma m &\ge m^\top \widehat{\Sigma} m + c e_j^\top \widehat{\Sigma} m - c e_j^\top \widehat{\Sigma} m \\ &\ge m^\top \widehat{\Sigma} m + c(1-\mu) - c e_j^\top \widehat{\Sigma} m \end{aligned} \tag{20}$$

Taking minimum on the both sides gives

$$\begin{aligned} \min_{\text{m feasible}} m^\top \widehat{\Sigma} m &\ge c(1-\mu) + \min_{m \in \mathbb{R}^p} \left(m^\top \widehat{\Sigma} m - c e_j^\top \widehat{\Sigma} m \right) \\ &\ge c(1-\mu) - \frac{c^2}{4} e_j^\top \widehat{\Sigma} e_j \\ &\ge \frac{(1-\mu)^2}{\widehat{\Sigma}_{jj}} \end{aligned} \tag{21}$$

Now note that $\mu = C\sqrt{\frac{\log p}{n}}$ is small. $\widehat{\Sigma}_{jj} = \Sigma_{jj} - \left(\widehat{\Sigma}_{jj} - \Sigma_{jj} \right)$, Σ_{jj} is bounded away from zero, and $\widehat{\Sigma}_{jj} - \Sigma_{jj}$ is of order $O\left(\frac{1}{\sqrt{n}}\right)$ goes to zero. We conclude with the following theorem

Theorem 1.1. *If* $\frac{s \log p}{\sqrt{n}} \rightsquigarrow 0$*, then*

$$\frac{\sqrt{n}\left(\widehat{\beta}_j - \beta_j \right)}{\sqrt{m_j^\top \widehat{\Sigma} m_j}} \rightsquigarrow \mathcal{N}(0,1) \tag{22}$$

1.3 Estimate Noise Level

If $y = X\beta + \sigma z$, then we could use lasso residual (not debiased lasso) to estimate the noise level σ

$$\begin{aligned} \widehat{\sigma}^2 &= \frac{1}{n} \left\| y - X\widetilde{\beta} \right\|^2 \\ &= \frac{1}{n} \left\| X\beta - X\widetilde{\beta} + \sigma z \right\|^2 \\ &= \frac{\sigma^2}{n} \|z\|^2 + \frac{1}{n} \left\| X\left(\beta - \widetilde{\beta} \right) \right\|^2 + \frac{2\sigma}{n} z^\top X \left(\beta - \widetilde{\beta} \right) \end{aligned} \tag{23}$$

And from bound deduced in last lecture, we deduce that

$$\sqrt{n}\left(\widehat{\sigma}^2 - \sigma^2 \right) = \sigma^2 \frac{1}{\sqrt{n}} \sum_{i=1}^n \left(z_i^2 - 1 \right) + O_p\left(\frac{s \log p}{\sqrt{n}} \right) \tag{24}$$

Note that $\sigma^2 \frac{1}{\sqrt{n}} \sum_{i=1}^n \left(z_i^2 - 1 \right) \rightsquigarrow \mathcal{N}\left(0, 2\sigma^4\right)$. Thus we have

$$\frac{\sqrt{n}\left(\widehat{\sigma}^2 - \sigma^2 \right)}{\sqrt{2}\sigma^2} \rightsquigarrow \mathcal{N}(0,1) \tag{25}$$

References

[Javanmard and Montanari, 2015] Javanmard, A. and Montanari, A. (2015). De-biasing the lasso: Optimal sample size for gaussian designs.

[van de Geer et al., 2014] van de Geer, S., Bühlmann, P., Ritov, Y., and Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):11661202.

[Zhang and Zhang, 2011] Zhang, C.-H. and Zhang, S. S. (2011). Confidence intervals for low-dimensional parameters in high-dimensional linear models.