STATISTICAL LEARNING II Rajen D. Shah
Example Sheet 1 (of 2) 2020

1. Show that if $\lambda \geq \lambda_{\max} := \|X^T Y\|_\infty / n$, then $\hat{\beta}^{\mathrm{L}}_\lambda = 0$.

2. Show that when the columns of X are orthogonal (so necessarily $p \leq n$) and scaled to have ℓ_2-norm $\sqrt{n}$, the kth component of the Lasso estimator is given by
$$\hat{\beta}^L_{\lambda,k} = (|\hat{\beta}^{\mathrm{OLS}}_k| - \lambda)_+ \mathrm{sgn}(\hat{\beta}^{\mathrm{OLS}}_k)$$
where $(\cdot)_+ = \max(0, \cdot)$. What is the corresponding estimator if the ℓ_1 penalty $\|\beta\|_1$ in the Lasso objective is replaced by the ℓ_0 penalty $\|\beta\|_0 := |\{k : \beta_k \neq 0\}|$?

3. Let $Y = X\beta^0 + \varepsilon - \bar{\varepsilon}\mathbf{1}$ and let $S = \{k : \beta^0 \neq 0\}$, $N := \{1, \ldots, p\} \setminus S$. Without loss of generality assume $S = \{1, \ldots, |S|\}$. Assume that X_S has full column rank and let $\Omega = \{\|X^T\varepsilon\|_\infty / n \leq \lambda_0\}$. Show that, when $\lambda > \lambda_0$, if the following two conditions hold
$$\sup_{\tau : \|\tau\|_\infty \leq 1} \|X_N^T X_S (X_S^T X_S)^{-1} \tau\|_\infty < \frac{\lambda - \lambda_0}{\lambda + \lambda_0}$$
$$(\lambda + \lambda_0) \|\{(\tfrac{1}{n} X_S^T X_S)^{-1}\}_k\|_1 < |\beta^0_k| \qquad \text{for } k \in S,$$
then on Ω the (unique) Lasso solution satisfies $\mathrm{sgn}(\hat{\beta}^L_\lambda) = \mathrm{sgn}(\beta^0)$.

4. Find the KKT conditions for the group Lasso.

5. (a) Consider the Lasso and let $\hat{E}_\lambda = \{k : \frac{1}{n}|X_k^T(Y - X\hat{\beta}^L_\lambda)| = \lambda\}$ be the so-called equicorrelation set at λ. Suppose that $\mathrm{rank}(X_{\hat{E}_\lambda}) = |\hat{E}_\lambda|$ for all $\lambda > 0$. Argue that the Lasso solution is unique for all $\lambda > 0$.

   (b) Under the assumptions above, let $\hat{\beta}^L_{\lambda_1}$ and $\hat{\beta}^L_{\lambda_2}$ be two Lasso solutions at different values of the regularisation parameter. Suppose that $\mathrm{sgn}(\hat{\beta}^L_{\lambda_1}) = \mathrm{sgn}(\hat{\beta}^L_{\lambda_2})$. Show that then for all $t \in [0, 1]$,
$$t\hat{\beta}^L_{\lambda_1} + (1 - t)\hat{\beta}^L_{\lambda_2} = \hat{\beta}^L_{t\lambda_1 + (1-t)\lambda_2}.$$
   Hint: Check the KKT conditions.

   (c) Conclude that the solution path $\lambda \mapsto \hat{\beta}^L_\lambda$ is piecewise linear with a finite number of knots (points λ where the solution path is not linear at λ) and these occur when the sign of the Lasso solution changes.

6. When proving the theorems on the prediction error of the Lasso, we started with the so-called basic inequality that
$$\frac{1}{2n}\|X(\beta^0 - \hat{\beta})\|_2^2 \leq \frac{1}{n}\varepsilon^T X(\hat{\beta} - \beta^0) + \lambda\|\beta^0\|_1 - \lambda\|\hat{\beta}\|_1.$$
Show that in fact we can improve this to
$$\frac{1}{n}\|X(\beta^0 - \hat{\beta})\|_2^2 \leq \frac{1}{n}\varepsilon^T X(\hat{\beta} - \beta^0) + \lambda\|\beta^0\|_1 - \lambda\|\hat{\beta}\|_1.$$

7. Under the assumptions of Theorem 21 on the prediction and estimation properties of the Lasso under a compatibility condition, show that, with probability $1 - 2p^{-(A^2/8-1)}$, we have
$$\frac{1}{n}\|X(\beta^0 - \hat{\beta})\|_2^2 \leq \frac{9A^2 \log(p)}{4\phi^2} \frac{\sigma^2 s}{n}.$$

8. (a) Show that

$$\max_{\theta : \|X^T\theta\|_\infty \le \lambda} G(\theta) = \frac{1}{2n}\|Y - X\hat{\beta}^L_\lambda\|_2^2 + \lambda\|\hat{\beta}^L_\lambda\|_1,$$

where

$$G(\theta) = \frac{1}{2n}\|Y\|_2^2 - \frac{1}{2n}\|Y - n\theta\|_2^2.$$

Show that the unique θ maximising G is $\theta^* = (Y - X\hat{\beta}^L_\lambda)/n$. *Hint: Treat the Lasso optimisation problem as minimising* $\|Y - z\|_2^2/(2n) + \lambda\|\beta\|_1$ *subject to* $z - X\beta = 0$ *over* $(\beta, z) \in \mathbb{R}^p \times \mathbb{R}^n$ *and consider the Lagrangian.*

(b) Let $\tilde{\theta}$ be such that $\|X^T\tilde{\theta}\|_\infty \le \lambda$. Explain why if

$$\max_{\theta : G(\theta) \ge G(\tilde{\theta})} |X_k^T\theta| < \lambda,$$

then we know that $\hat{\beta}^L_{\lambda,k} = 0$. By considering $\tilde{\theta} = Y\lambda/(n\lambda_{\max})$ with $\lambda_{\max} = \|X^TY\|_\infty/n$, show that $\hat{\beta}^L_{\lambda,k} = 0$ if

$$\frac{1}{n}|X_k^TY| < \lambda - \frac{\|Y\|_2}{\sqrt{n}}\frac{\lambda_{\max} - \lambda}{\lambda_{\max}}.$$

9. The elastic net estimator in the linear model minimises

$$\frac{1}{2n}\|Y - X\beta\|_2^2 + \lambda(\alpha\|\beta\|_1 + (1-\alpha)\|\beta\|_2^2/2)$$

over $\beta \in \mathbb{R}^p$, where $\alpha \in [0, 1]$ is fixed.

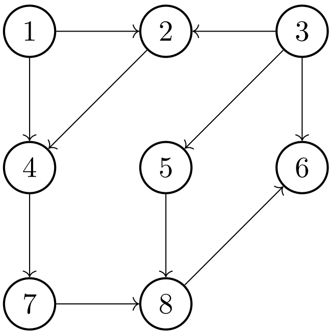
(a) Suppose X has two columns X_j and X_k that are identical and $\alpha < 1$. Explain why the minimising β^* above is unique and has $\beta^*_k = \beta^*_j$.

(b) Let $\hat{\beta}^{(0)}, \hat{\beta}^{(1)}, \ldots$ be the solutions from iterations of a coordinate descent procedure to minimise the elastic net objective. For a fixed variable index k, let $A = \{1, \ldots, k-1\}$ and $B = \{k+1, \ldots, p\}$. Show that for $m \ge 1$,

$$\hat{\beta}^{(m)}_k = \frac{S_{\lambda\alpha}\Big(n^{-1}X_k^T(Y - X_A\hat{\beta}^{(m)}_A - X_B\hat{\beta}^{(m-1)}_B)\Big)}{1 + \lambda(1-\alpha)},$$

where $S_t(u) = \mathrm{sgn}(u)(|u| - t)_+$ is the soft-thresholding operator.

10. For the following DAG $\mathcal{G}$



write down

(a) the descendants of 3;

(b) all sets of variables that d-separate 1 and 3;

(c) all sets of variables that d-separate $\{1, 4\}$ and 6;

(d) all the v-structures.

11. Let $Z = (Z_1, \ldots, Z_p)^T \in \{0, 1\}^p$ be a binary random vector with probability mass function given by

$$\mathbb{P}(Z_1 = z_1, \ldots, Z_p = z_p) = \exp\left(\Theta_{00} + \sum_{k=1}^{p} \Theta_{0k} z_k + \sum_{k=1}^{p} \sum_{j=1}^{k-1} \Theta_{jk} z_j z_k - \Phi(\Theta)\right)$$

where $\exp(-\Phi(\Theta))$ is a normalising constant. Show that

$$\text{logit}(\mathbb{P}(Z_k = 1 | Z_{-k} = z_{-k})) = \Theta_{0k} + \sum_{j:j<k} \Theta_{jk} z_j + \sum_{j:j>k} \Theta_{kj} z_j,$$

where $\text{logit}(q) = \log\{q/(1-q)\}$ for $q \in (0, 1)$. Conclude that, for $j < k$,

$$Z_j \perp\!\!\!\perp Z_k | Z_{-jk} \Longleftrightarrow \Theta_{jk} = 0.$$

Note that for discrete random variables we can replace the densities in our definition of conditional independence with probability mass functions (which are in any case densities with respect to counting measure). How might we go about estimating the Θ_{jk}?