MODERN STATISTICAL METHODS Part III
Example Sheet 3 (of 4) Michaelmas 2019

In all of the below, assume that any design matrices X are $n \times p$ and have their columns centred and then scaled to have ℓ_2-norm $\sqrt{n}$, and that any responses $Y \in \mathbb{R}^n$ are centred.

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

2. Under the assumptions of Theorem 23 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
$$\frac{1}{n}\|X(\beta^0-\hat\beta)\|_2^2 \le \frac{9A^2\log(p)}{4\phi^2}\frac{\sigma^2 s}{n}.$$

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 $\Omega=\{\|X^T\varepsilon\|_\infty/n \le \lambda_0\}$. Show that, when $\lambda>\lambda_0$, if the following two conditions hold
$$\sup_{\tau:\|\tau\|_\infty\le 1}\|X_N^T X_S(X_S^T X_S)^{-1}\tau\|_\infty < \frac{\lambda-\lambda_0}{\lambda+\lambda_0}$$
$$(\lambda+\lambda_0)\|\{(\tfrac{1}{n}X_S^T X_S)^{-1}\}_k\|_1 < |\beta_k^0| \qquad \text{for } k \in S,$$
then on Ω the (unique) Lasso solution satisfies $\mathrm{sgn}(\hat\beta_\lambda^{\mathrm{L}})=\mathrm{sgn}(\beta^0)$.

4. Find the KKT conditions for the group Lasso.

5. (a) Show that
$$\max_{\theta:\|X^T\theta\|_\infty\le\lambda} G(\theta)=\frac{1}{2n}\|Y-X\hat\beta_\lambda^{\mathrm{L}}\|_2^2+\lambda\|\hat\beta_\lambda^{\mathrm{L}}\|_1,$$
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 $\theta^*=(Y-X\hat\beta_\lambda^{\mathrm{L}})/n$. *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 $\hat\beta_{\lambda,k}^{\mathrm{L}}=0$. By considering $\tilde\theta = Y\lambda/(n\lambda_{\max})$ with $\lambda_{\max}=\|X^TY\|_\infty/n$, show that $\hat\beta_{\lambda,k}^{\mathrm{L}}=0$ if
$$\frac{1}{n}|X_k^TY|<\lambda-\frac{\|Y\|_2}{\sqrt{n}}\frac{\lambda_{\max}-\lambda}{\lambda_{\max}}.$$

6. Consider the Lasso and let $\hat{E}_\lambda = \{k : \frac{1}{n}|X_k^T(Y - X\hat{\beta}^{\mathrm{L}}_\lambda)| = \lambda\}$ be the equicorrelation set at λ. Suppose that $\mathrm{rank}(X_{\hat{E}_\lambda}) = |\hat{E}_\lambda|$ for all $\lambda > 0$, so the Lasso solution is unique for all $\lambda > 0$. Let $\hat{\beta}^{\mathrm{L}}_{\lambda_1}$ and $\hat{\beta}^{\mathrm{L}}_{\lambda_2}$ be two Lasso solutions at different values of the regularisation parameter. Suppose that $\mathrm{sgn}(\hat{\beta}^{\mathrm{L}}_{\lambda_1}) = \mathrm{sgn}(\hat{\beta}^{\mathrm{L}}_{\lambda_2})$. Show that then for all $t \in [0,1]$,

$$t\hat{\beta}^{\mathrm{L}}_{\lambda_1} + (1-t)\hat{\beta}^{\mathrm{L}}_{\lambda_2} = \hat{\beta}^{\mathrm{L}}_{t\lambda_1 + (1-t)\lambda_2}.$$

Hint: Check the KKT conditions. Conclude that the solution path $\lambda \mapsto \hat{\beta}^{\mathrm{L}}_\lambda$ 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.

7. 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 $m \geq 1$,

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

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

8. Theorem 28 in the notes assumes that X is an $n \times d$ matrix with i.i.d. rows $x_1, \ldots, x_n$ of mean zero and covariance Σ, which satisfy the sub-Gaussian condition

$$\mathbb{E}(e^{\lambda\langle x_1, v\rangle}) \leq e^{\frac{\lambda^2\sigma^2}{2}} \qquad \text{for all } \lambda > 0, v \in S^{d-1}.$$

Then, the empirical covariance matrix $\hat{\Sigma} = n^{-1}X^TX$, satisfies, for some constant C,

$$\mathbb{P}\left(\frac{\|\hat{\Sigma} - \Sigma\|_{op}}{\sigma^2} \geq C\left(\frac{d+\delta}{n} \vee \sqrt{\frac{d+\delta}{n}}\right)\right) \leq e^{-\delta} \qquad \text{for all } \delta > 0.$$

Now suppose that the rows of X have distribution $N(\mu, \Sigma)$ with non-zero mean and $\|\Sigma\|_{op} = \sigma^2$. Prove a similar deviation bound for the maximum likelihood estimator $\hat{\Sigma} = n^{-1}\sum_{i=1}^n (x_i - \overline{X})(x_i - \overline{X})^T$, where $\overline{X} = n^{-1}\sum_{i=1}^n x_i$.

9. Let $X \in \mathbb{R}^{n \times p}$ $(n > p)$ be a centred data matrix with (thin) SVD $X = UDV^T$. Let the first *principal component* be $u^{(1)} = D_{11}U_1$, and the first *loading vector* be $v^{(1)} = V_1$. We may define the kth principal component $u^{(k)}$ and loading vector $v^{(k)}$ for $k > 1$ inductively as follows.

$$v^{(k)} \text{ maximises } \|Xv\|_2 \text{ over } v \in \mathbb{R}^p \text{ with constraints}$$
$$\|v\|_2 = 1 \text{ and } {u^{(j)}}^T Xv = 0 \text{ for all } j < k;$$
$$u^{(k)} = Xv^{(k)}.$$

Suppose that $D_{11}, \ldots, D_{pp}$ are all distinct. Show that $v^{(k)} = V_k$ and $u^{(k)} = D_{kk}U_k$ (up to an arbitrary sign).

10. Suppose we wish to obtain the principal components of the (not necessarily centred) matrix $\Phi \in \mathbb{R}^{n \times d}$. Explain how we can recover the principal components given only $K = \Phi\Phi^T$.