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

$$\frac{1}{2n}\|X(\beta^0 - \hat{\beta})\|_2^2 \leq \frac{1}{n}\varepsilon^T X(\hat{\beta} - \beta^0) + \lambda\|\beta^0\|_1 - \lambda\|\hat{\beta}\|_1.$$

Show that in fact we can improve this to

$$\frac{1}{n}\|X(\beta^0 - \hat{\beta})\|_2^2 \leq \frac{1}{n}\varepsilon^T X(\hat{\beta} - \beta^0) + \lambda\|\beta^0\|_1 - \lambda\|\hat{\beta}\|_1.$$

Solution: We start with the KKT conditions for the Lasso

$$\frac{1}{n}X^T(Y - X\hat{\beta}) = \lambda\hat{\nu},$$

where, writing $\hat{S} = \{k : \hat{\beta}_k \neq 0\}$, we have $\text{sgn}(\hat{\beta}_{\hat{S}} = \hat{\nu}_{\hat{S}}$, and also $\|\hat{\nu}\|_\infty \leq 1$. Now we multiply (both sides) by ${\beta^0}^T - \hat{\beta}^T$. Note that $\hat{\beta}^T\hat{\nu} = \hat{\beta}_{\hat{S}}^T\text{sgn}(\hat{\beta}_{\hat{S}}) = \|\hat{\beta}\|_1$. Furthermore, by Hölder's inequality, $|{\beta^0}^T\hat{\nu}| \leq \|\beta^0\|_1\|\hat{\nu}\|_\infty \leq \|\beta^0\|_1$. Substituting $Y = X\beta^0 + \varepsilon - \bar{\varepsilon}\mathbf{1}$ yields

$$\frac{1}{n}\|X(\beta^0 - \hat{\beta})\|_2^2 + \frac{1}{n}\varepsilon^T X(\beta^0 - \hat{\beta}) \leq \lambda\|\beta^0\|_1 - \lambda\|\hat{\beta}\|_1.$$

Rearranging gives the result.

An alternative solution is as follows. Let Q denote the Lasso objective as in lectures. We have $Q(\hat{\beta}) \leq Q((1-t)\beta^0 + t\hat{\beta})$ for all t. Thus

$$\frac{1}{2n}\|X\beta^0 - X\hat{\beta} + \varepsilon\|_2^2 + \lambda\|\hat{\beta}\|_1 \leq \frac{1}{2n}\|t(X\beta^0 - X\hat{\beta}) - \varepsilon\|_2^2 + t\lambda\|\hat{\beta}\|_1 + (1-t)\|\beta^0\|_1.$$

Dividing by $1 - t$ and rearranging we have

$$\frac{1+t}{2n}\|X\beta^0 - X\hat{\beta}\|_2^2 + \lambda\|\hat{\beta}\|_1 \leq \frac{1}{n}\varepsilon^T X(\hat{\beta} - \beta^0) + \lambda\|\beta^0\|_1$$

for all $t < 1$. Letting $t \uparrow 1$ then gives the result.

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

$$\frac{1}{n}\|X(\beta^0 - \hat{\beta})\|_2^2 \leq \frac{9A^2\log(p)}{4\phi^2}\frac{\sigma^2 s}{n}.$$

Solution: We follow the proof of Theorem 23, but starting with the improved "basic inequality" in the previous question. We arrive at

$$\frac{2}{n}\|X(\beta^0 - \hat{\beta})\|_2^2 + \lambda\|\hat{\beta}_N - \beta_N^0\|_1 \leq 3\lambda\|\hat{\beta}_S - \beta_S^0\|_1.$$

Using the compatibility condition, the RHS is at most

$$\|\hat{\beta}_S - \beta^0_S\|_1 \leq \frac{\sqrt{s}\|X(\beta^0 - \hat{\beta})\|_2/\sqrt{n}}{\phi}.$$

Substituting this into the previous inequality, we get

$$\frac{2}{n}\|X(\beta^0 - \hat{\beta})\|_2^2 + \lambda\|\hat{\beta}_N - \beta^0_N\|_1 \leq \frac{3\lambda\sqrt{s}\|X(\beta^0 - \hat{\beta})\|_2/\sqrt{n}}{\phi}$$

whence

$$\frac{1}{n}\|X(\beta^0 - \hat{\beta})\|_2^2 \leq \frac{9\lambda^2 s}{4\phi^2}$$

as required.

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 $\mathrm{sgn}(\hat{\beta}^{\mathrm{L}}_\lambda) = \mathrm{sgn}(\beta^0)$.
Solution: Suppressing the dependence of $\hat{\beta}^{\mathrm{L}}_\lambda$ on λ and dropping the superscript L for ease of notation, we can write the KKT conditions as

$$\frac{1}{n}\begin{pmatrix} X_S^T X_S & X_S^T X_N \\ X_N^T X_S & X_N^T X_N \end{pmatrix}\begin{pmatrix} \beta_S - \hat{\beta}_S \\ -\hat{\beta}_N \end{pmatrix} + \frac{1}{n}\begin{pmatrix} X_S^T\varepsilon \\ X_N^T\varepsilon \end{pmatrix} = \lambda\begin{pmatrix} \hat{\nu}_S \\ \hat{\nu}_N \end{pmatrix}, \tag{1}$$

where $\|\hat{\nu}\|_\infty \leq 1$ and writing $\hat{S} = \{k : \hat{\beta}_k \neq 0\}$, we have $\mathrm{sgn}(\hat{\beta}_{\hat{S}}) = \hat{\nu}_{\hat{S}}$. Now if we do have $\mathrm{sgn}(\hat{\beta}) = \mathrm{sgn}(\beta^0)$, it must be the case that (considering the first block of (1)),

$$\frac{1}{n}X_S^T X_S(\beta_S - \hat{\beta}_S) + \frac{1}{n}X_S^T\varepsilon = \lambda\mathrm{sgn}(\beta^0_S),$$

and, substituting this into the second block of (1),

$$X_N^T X_S (X_S^T X_S)^{-1}\{\lambda\mathrm{sgn}(\beta^0_S) - \tfrac{1}{n}X_S^T\varepsilon\} + \tfrac{1}{n}X_N^T\varepsilon = \lambda\hat{\nu}_N.$$

Now we work on Ω and claim that

$$(\hat{\beta}_S, \hat{\nu}_S) = (\beta^0_S - (\tfrac{1}{n}X_S^T X_S)^{-1}\{\lambda\mathrm{sgn}(\beta^0_S) - \tfrac{1}{n}X_S^T\varepsilon\},\ \mathrm{sgn}(\beta^0_S)),$$
$$(\hat{\beta}_N, \hat{\nu}_N) = (0,\ [X_N^T X_S (X_S^T X_S)^{-1}\{\lambda\mathrm{sgn}(\beta^0_S) - \tfrac{1}{n}X_S^T\varepsilon\} + \tfrac{1}{n}X_N^T\varepsilon]/\lambda),$$

satisfy (1). We first check that $\mathrm{sgn}(\hat{\beta}_S) = \mathrm{sgn}(\beta^0_S)$. This holds because

$$\begin{aligned} |[(\tfrac{1}{n}X_S^T X_S)^{-1}\{\lambda\mathrm{sgn}(\beta^0_S) - \tfrac{1}{n}X_S^T\varepsilon\}]_k| &\leq \|\{(\tfrac{1}{n}X_S^T X_S)^{-1}\}_k\|_1\{\lambda\|\mathrm{sgn}(\beta^0_S)\|_\infty + \|\tfrac{1}{n}X_S^T\varepsilon\|_\infty\} \\ &\leq \|\{(\tfrac{1}{n}X_S^T X_S)^{-1}\}_k\|_1(\lambda + \lambda_0). \end{aligned}$$

Next

$$\begin{aligned}\|X_N^T X_S(X_S^T X_S)^{-1}\{\lambda \mathrm{sgn}(\beta_S^0) - \tfrac{1}{n}X_S^T\varepsilon\} + \tfrac{1}{n}X_N^T\varepsilon\|_\infty &\leq \|X_N^T X_S(X_S^T X_S)^{-1}\{\lambda \mathrm{sgn}(\beta_S^0) - \tfrac{1}{n}X_S^T\varepsilon\}\|_\infty + \lambda_0 \\ &\leq (\lambda + \lambda_0) \sup_{\tau:\|\tau\|_\infty \leq 1} \|X_N^T X_S(X_S^T X_S)^{-1}\tau\|_\infty + \lambda_0 \\ &< \lambda,\end{aligned}$$

by the first assumption given in the question. Thus the KKT conditions are satisfied. Because we have the strict inequality $\|\hat{\nu}_N\|_\infty < 1$, S is the equicorrelation set. Since X_S has full column rank, we know the Lasso solution is unique.

4. Find the KKT conditions for the group Lasso.
Solution: For $G \subset \{1, \ldots, p\}$, consider the function $\beta \mapsto \|\beta_G\|_2$. The subdifferential of this function at a β with $\beta_G \neq 0$ is singleton a vector v with $v_{G^c} = 0$ and

$$v_G = \frac{\beta_G}{\|\beta_G\|_2}.$$

We claim that the subdifferential when $\beta_G = 0$ is $\{v : v_{G^c} = 0 \text{ and } \|v_G\|_2 \leq 1\}$. Indeed, if $v_{G^c} \neq 0$ then taking y with $y_G = 0$, $y_{G^c} - \beta_{G^c} = v_{G^c}$, we have

$$0 = \|y_G\|_2 < v^T(y - \beta) = \|v_{G^c}\|_2^2.$$

Now if $v_{G^c} = 0$ and $\|v_G\|_2 \leq 1$, then

$$\|y_G\|_2 \geq \|v_G\|_2\|y_G\|_2 \geq v^T y.$$

Conversely, if $\|v_G\|_2 > 1$, then taking $y_G = v_G$ (and $y_{G^c} = 0$), we have

$$\|y_G\|_2 < \|v_G\|_2\|y_G\|_2 = v^T y.$$

Since the subdifferential of a sum of convex functions is the set sum of the subdifferentials of the individual functions, we see that the KKT conditions for the group Lasso objective

$$\frac{1}{2n}\|Y - X\beta\|_2^2 + \lambda \sum_{j=1}^q m_j \|\beta_{G_j}\|_2,$$

are that $\hat{\beta}$ is a minimiser if and only if, for $j = 1, \ldots, q$,

$$\frac{1}{n}X_{G_j}^T(Y - X\hat{\beta}) = \lambda m_j \hat{\nu}_{G_j},$$

where $\hat{\nu} \in \mathbb{R}^p$ is such that $\|\hat{\nu}_{G_j}\|_2 \leq 1$, and if $\hat{\beta}_{G_j} \neq 0$ then

$$\hat{\nu}_{G_j} = \frac{\hat{\beta}_{G_j}}{\|\hat{\beta}_{G_j}\|_2}.$$

5. (a) Show that

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

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 $\theta^* = (Y - X\hat{\beta}^{\mathrm{L}}_\lambda)/n$. *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.*

Solution: Taking the hint, we write the Lagrangian for the Lasso problem

$$L(\beta, z, \theta) = \frac{1}{2n}\|Y - z\|_2^2 + \lambda\|\beta\|_1 + \theta^T(z - X\beta).$$

The minimising β and z, β^* and z^* satisfy

$$\begin{aligned}\frac{1}{n}(Y - z^*) &= \theta,\\ \lambda\nu^* &= X^T\theta,\end{aligned}$$

provided θ is such that $\|\nu^*\|_\infty \leq 1$ so $\|X^T\theta\|_\infty \leq \lambda$. Substituting into the Lagrangian and using the fact that ${\beta^*}^T X^T\theta = \lambda {\beta^*}^T\nu^* = \lambda\|\beta\|_1$ we get

$$L(\beta^*, z^*, \theta) = \frac{n}{2}\|\theta\|_2^2 + \theta^T(Y - n\theta) = G(\theta),$$

provided $\|X^T\theta\|_\infty \leq \lambda$. Thus we have that

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

To get equality we take $\theta = \theta^* = (Y - X\hat{\beta}_\lambda)/n$ (dropping the superscript L for the Lasso solution for clarity) for then

$$G(\theta^*) = \frac{1}{2n}\|Y - X\hat{\beta}_\lambda\|_2^2 + \frac{1}{n}\hat{\beta}_\lambda^T X^T(Y - X\hat{\beta}^{\mathrm{L}}_\lambda),$$

the final term equalling $\lambda\|\hat{\beta}_\lambda\|_1$ by the KKT conditions. Uniqueness of the maximiser follows from the facts that $-G$ is strictly convex and $\{\theta : \|X^T\theta\|_\infty \leq \lambda\}$ is a convex set.

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

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

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

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

Solution: θ^* must be in the set $\{\theta : G(\theta) \geq G(\tilde{\theta})\}$ so if the inequality in the question is true, then we know $|X_k^T(Y - X\hat{\beta}_\lambda)|/n < \lambda$ whence by the KKT conditions for the Lasso, $\hat{\beta}_{\lambda,k}$ must be zero. With the given choice of $\tilde{\theta}$, we know $\|X^T\tilde{\theta}\|_\infty \leq \lambda$.
We now need to show that when

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

and $G(\theta) \geq G(Y\lambda/(n\lambda_{\max}))$, we have $|X_k^T\theta| < \lambda$. Note the condition $G(\theta) \geq G(Y\lambda/(n\lambda_{\max}))$ is equivalent to

$$\|Y - n\theta\|_2 \leq (1 - \lambda/\lambda_{\max})\|Y\|_2.$$

Now, under the conditions above,

$$\begin{aligned}|X_k^T\theta| &= |X_k^T(\theta - Y/n + Y/n)| \\ &\leq |X_k^T(\theta - Y/n)| + |X_k^T Y|/n \\ &< \|X_k\|_2(1-\lambda/\lambda_{\max})\|Y\|_2/n + \lambda - \frac{\|Y\|_2}{\sqrt{n}}\frac{\lambda_{\max}-\lambda}{\lambda_{\max}} \\ &= \lambda,\end{aligned}$$

using the fact that $\|X_k\|_2 = \sqrt{n}$ to get the final equality.

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.
Solution: We need only check that $t\hat{\beta}^{\mathrm{L}}_{\lambda_1} + (1-t)\hat{\beta}^{\mathrm{L}}_{\lambda_2}$ satisfies the KKT conditions for the Lasso at $t\lambda_1 + (1-t)\lambda_2$. To ease notation, let us write $\hat{\beta}^{(j)} = \hat{\beta}^{\mathrm{L}}_{\lambda_j}$, $j = 1, 2$. Now we know that for $j = 1, 2$,

$$\frac{1}{n}X^T(Y - X\hat{\beta}^{(j)}) = \lambda_j\hat{\nu}^{(j)}$$

where, writing $S = \{k : \hat{\beta}^{(1)} \neq 0\}$, $\hat{\nu}^{(1)}_S = \hat{\nu}^{(2)}_S = \mathrm{sgn}(\hat{\beta}^{(1)}_S)$, and $\|\hat{\nu}^{(j)}\|_\infty \leq 1$. Thus

$$\begin{aligned}\frac{1}{n}X^T[Y - X\{t\hat{\beta}^{(1)} + (1-t)\hat{\beta}^{(2)}\}] &= t\frac{1}{n}X^T(Y - X\hat{\beta}^{(1)}) + (1-t)\frac{1}{n}X^T(Y - X\hat{\beta}^{(2)}) \\ &= t\lambda_1\hat{\nu}^{(1)} + (1-t)\lambda_2\hat{\nu}^{(2)}.\end{aligned}$$

Now the indices of the nonzero components of $t\hat{\beta}^{(1)} + (1-t)\hat{\beta}^{(2)}$ are S and

$$\mathrm{sgn}(t\hat{\beta}^{(1)}_S + (1-t)\hat{\beta}^{(2)}_S) = \hat{\nu}^{(1)}_S = \hat{\nu}^{(2)}_S = \frac{t\lambda_1\hat{\nu}^{(1)}_S + (1-t)\lambda_2\hat{\nu}^{(2)}_S}{t\lambda_1 + (1-t)\lambda_2}.$$

Furthermore, by the triangle inequality,

$$\|t\lambda_1\hat{\nu}^{(1)} + (1-t)\lambda_2\hat{\nu}^{(2)}\|_\infty \leq t\lambda_1\|\hat{\nu}^{(1)}\|_\infty + (1-t)\lambda_2\|\hat{\nu}^{(2)}\|_\infty \leq t\lambda_1 + (1-t)\lambda_2.$$

Thus the pair

$$\left(t\hat{\beta}^{(1)} + (1-t)\hat{\beta}^{(2)},\ \frac{t\lambda_1\hat{\nu}^{(1)} + (1-t)\lambda_2\hat{\nu}^{(2)}}{t\lambda_1 + (1-t)\lambda_2}\right)$$

satisfies the KKT conditions at $t\lambda_1 + (1-t)\lambda_2$. Since there are 3^p sign patterns a Lasso solution can take (each component can be either positive, negative or equal to 0), there are a finite number of knots.

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 $\beta_k^* = \beta_j^*$.
Solution: The minimum is unique as the objective above is strictly convex, and existence can be shown via the same argument used to show the existence of Lasso solutions. Suppose then that β^* is the unique minimiser. Let $\beta' = \beta^*$ in all components except $\beta_j' = \beta_k^*$ and $\beta_k' = \beta_j^*$. The objective is strictly convex in β so $\beta'/2 + \beta^*/2$ has an objective value at least as large small as that of β^*, so $\beta' = \beta^*$ by uniqueness.

(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 $m \geq 1$,

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

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

$$\hat{\beta}_k^{(m)} = \underset{\beta\in\mathbb{R}}{\mathrm{argmin}}\ \{\|Y - X_A\hat{\beta}_A^{(m)} - X_B\hat{\beta}_B^{(m-1)} - \beta X_k\|_2^2/(2n) + \lambda(\alpha|\beta| + (1-\alpha)\beta^2/2)\}$$

The minimiser $\hat{\beta}_k^{(m)}$ must satisfy the subgradient optimality condition:

$$-\frac{1}{n}X_k^T(Y - X_A\hat{\beta}_A^{(m)} - X_B\hat{\beta}_B^{(m-1)}) + \hat{\beta}_k^{(m)} + \lambda(1-\alpha)\hat{\beta}_k^{(m)} + \lambda\alpha\hat{\nu} = 0,$$

where $\hat{\nu} \in [-1,1]$ and if $\hat{\beta}_k^{(m)} \neq 0$, $\hat{\nu} = \mathrm{sgn}(\hat{\beta}_k^{(m)})$. Rearranging, we have

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

and we may check that the given expression for $\hat{\beta}_k^{(m)}$ satisfies this. Note that $\hat{\beta}_k^{(m)}$ is the unique minimiser as the objective is strictly convex.

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 $\overline{X} = n^{-1}\sum_{i=1}^n x_i$.

Solution: In the more general case, letting J be the orthogonal projection matrix with all entries equal to $1/n$,

$$\begin{aligned}
\|\hat{\Sigma} - \Sigma\|_{op} &= \|n^{-1}(X - JX)^T(X - JX) - \Sigma\|_{op} \\
&= \|n^{-1}X^T(I - J)X - \Sigma\|_{op} \\
&= \|n^{-1}(X - 1\mu^T)^T(I - J)(X - 1\mu^T) - \Sigma\|_{op} \\
&= \|n^{-1}(X - 1\mu^T)^T(X - 1\mu^T) - \Sigma - n^{-1}(X - 1\mu^T)^T J(X - 1\mu^T)\|_{op} \\
&\leq \|n^{-1}(X - 1\mu^T)^T(X - 1\mu^T) - \Sigma\|_{op} + \|(\mu - \overline{X})(\mu - \overline{X})^T\|_{op} \\
&= \|n^{-1}(X - 1\mu^T)^T(X - 1\mu^T) - \Sigma\|_{op} + \|\mu - \overline{X}\|_2^2. \qquad (2)
\end{aligned}$$

Now, the first term on the right hand side satisfies the deviation bound of Theorem 28 as $X - 1\mu^T$ is a matrix with i.i.d. rows of zero mean and covariance Σ, with the sub-Gaussian property. The second term is the squared error of the mean estimator $\overline{X}$. The vector $\mu - \overline{X}$ has distribution $N(0, n^{-1}\Sigma)$. Thus letting $Z = \sqrt{n}\Sigma^{-1/2}(\mu - \overline{X})$, with distribution $N(0, I)$, we have

$$\|\mu - \overline{X}\|_2^2 \stackrel{d}{=} (n^{-1/2}\Sigma^{1/2}Z)^T(n^{-1/2}\Sigma^{1/2}Z)^T \stackrel{d}{=} n^{-1}Z^T\Sigma Z \stackrel{d}{=} n^{-1}\sum_{i=1}^{d} Z_i^2\sigma_i^2,$$

where σ_i^2 is the ith eigenvalue of Σ and the final equality follows from the orthogonal invariance of $N(0, I)$. As σ^2 is the largest eigenvalue of Σ, $n^{-1}\sum_{i=1}^{d} Z_i^2\sigma_i^2 \leq n^{-1}\sigma^2\sum_{i=1}^{d} Z_i^2$. So the mean squared error has a χ_d^2 distribution scaled by σ^2/n. If $\eta \sim \chi_d^2$, for any $\lambda < 1/2$

$$\mathbb{E}[e^{\lambda(\eta - d)}] = e^{-\lambda d}(1 - 2\lambda)^{-d/2}.$$

A Chernoff inequality leads to the tail bound $\mathbb{P}(\eta \geq (\sqrt{d + \delta} \vee d + \delta)) \leq e^{-\delta}$, which implies

$$\mathbb{P}\left(\frac{\|\mu - \overline{X}\|_2^2}{\sigma^2} \geq n^{-1}(\sqrt{d + \delta} \vee d + \delta)\right) \leq e^{-\delta}.$$

Finally, a union bound for each term on the right of (2) leads to a deviation bound as the one in Theorem 28.

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.

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

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).

Solution: Let $k \geq 2$. Given that $v^{(j)} = V_j$ and $u^{(j)} = D_{jj}U_j$ for $j = 1, \ldots, k-1$, we shall show that $v^{(k)} = V_k$ and $u^{(k)} = D_{kk}U_k$. Let $v \in \mathbb{R}^p$ have $\|v\|_2 = 1$ and ${u^{(j)}}^T Xv = 0$ for all $j < k$ and let $w = V^T v$. Then $U_j^T UDV^T v = (Dw)_j = 0$ for all $j < k$ whence $w_j = 0$ for all $j < k$. Now

$$\|Xv\|_2^2 = w^T D^2 w = \sum_{j \geq k} w_j^2 D_{jj}^2 \leq D_{kk}^2 \|w\|_2^2 = D_{kk}^2,$$

with equality if and only if $w_k = \pm 1$ with all other entries equal to zero. Thus $v^{(k)} = \pm V_k$ and $u^{(k)} = \pm D_{kk} U_k$.

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 $K = \Phi\Phi^T$.
Solution: Let $\tilde{\Phi} = \Phi - J\Phi$ where J has all entries equal to $1/n$ be a centred version of Φ. The principal components will be the columns of $\tilde{U}\tilde{D}$ where these matrices come from the SVD of $\tilde{\Phi} = \tilde{U}\tilde{D}\tilde{V}^T$. Now

$$\begin{aligned}\tilde{K} &:= \tilde{\Phi}\tilde{\Phi}^T \\ &= (\Phi - J\Phi)(\Phi - J\Phi)^T \\ &= K - JK - KJ + JKJ.\end{aligned}$$

Thus we can recover the principal components by computing the eigendecomposition of $\tilde{K}$ which can be constructed from K as above. [We have skated over the fact that eigendecompositions and hence principal components are not unique (e.g. we can always take the negative). We can have more serious non-uniqueness issues when $\tilde{K}$ is the identity matrix, for example (unlikely of course). In general though, principal components obtained through this approach using $\tilde{K}$ will still be valid principal components, but they may not necessarily agree with principal components from an SVD of Φ.]