

Modern Statistical Methods — Example Sheet 3

Lucas Riedstra

November 22, 2020

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.

Question 1. *When proving the theorems on the prediction error of the Lasso, we started with the so-called basic inequality that*

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

Show that in fact we can improve this to

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

Proof. By the KKT conditions, we there exists $\hat{\nu} \in \mathbb{R}^p$ with $\|\hat{\nu}\|_\infty \leq 1$ and $\langle \hat{\nu}, \hat{\beta} \rangle = \|\hat{\beta}\|_1$ such that

$$\begin{aligned} \frac{1}{n} X^\top (Y - X\hat{\beta}) &= \lambda \hat{\nu} \\ \frac{1}{n} X^\top (X(\beta^0 - \hat{\beta}) + \varepsilon) &= \lambda \hat{\nu} \\ \frac{1}{n} X^\top X(\beta^0 - \hat{\beta}) &= -\frac{1}{n} X^\top \varepsilon + \lambda \hat{\nu} \\ \frac{1}{n} (\beta^0 - \hat{\beta})^\top X^\top X(\beta^0 - \hat{\beta}) &= -\frac{1}{n} (\beta^0 - \hat{\beta})^\top X^\top \varepsilon + \lambda (\beta^0 - \hat{\beta})^\top \hat{\nu} \\ \frac{1}{n} \|X(\beta^0 - \hat{\beta})\|_2^2 &= \frac{1}{n} \varepsilon^\top X(\hat{\beta} - \beta^0) + \lambda \langle \hat{\nu}, \beta^0 \rangle - \lambda \|\hat{\beta}\|_1, \end{aligned}$$

and now plugging in $\langle \beta^0, \hat{\nu} \rangle \leq \|\beta^0\|_1 \|\hat{\nu}\|_\infty \leq \|\beta^0\|_1$ yields the result. \square

Question 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*

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

Proof. We have

$$\begin{aligned}
\frac{1}{\lambda n} \|X(\beta^0 - \hat{\beta})\|_2^2 &\leq \frac{1}{2\lambda n} \left\langle \frac{2X^\top \varepsilon}{n}, (\hat{\beta} - \beta^0) \right\rangle + \|\beta^0\|_1 - \|\hat{\beta}\|_1 \\
&\leq \frac{1}{2} \|\hat{\beta} - \beta^0\|_1 + \|\beta^0\|_1 - \|\hat{\beta}\|_1 \\
&= \frac{1}{2} (\|\hat{\beta}_S - \beta_S^0\|_1 + \|\hat{\beta}_N\|_1) + \|\beta_S^0\|_1 - \|\hat{\beta}_S\|_1 - \|\hat{\beta}_N\|_1 \\
&= \frac{1}{2} \|\hat{\beta}_S - \beta_S^0\|_1 + \|\beta_S^0\|_1 - \|\hat{\beta}_S\|_1 - \frac{1}{2} \|\hat{\beta}_N\|_1 \\
&\stackrel{\star}{\leq} \frac{3}{2} \|\hat{\beta}_S - \beta_S^0\|_1 - \frac{1}{2} \|\hat{\beta}_N\|_1 \\
&\leq \frac{3}{2} \|\hat{\beta}_S - \beta_S^0\|_1,
\end{aligned}$$

and note that from \star we conclude $\|\hat{\beta}_N\|_1 \leq 3\|\hat{\beta}_S - \beta_S^0\|_1$, and therefore $\hat{\beta} - \beta^0$ can be plugged into the compatibility constant. We find

$$\frac{1}{n} \|X(\beta^0 - \hat{\beta})\|_2^2 \leq \frac{3\lambda\sqrt{s}}{2\varphi\sqrt{n}} \|X(\beta^0 - \hat{\beta})\|_2 \implies \frac{1}{\sqrt{n}} \|X(\beta^0 - \hat{\beta})\|_2 \leq \frac{3\lambda\sqrt{s}}{2\varphi},$$

and squaring both sides gives

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

□

Question 3. Let $Y = X\beta^0 + \varepsilon - \varepsilon\mathbf{1}$ and let $S = \{k \mid \beta^0 \neq 0\}$, $N := \{1, \dots, p\} \setminus S$. Without loss of generality assume $S = \{1, \dots, |S|\}$. Assume that X has full column rank and let $\Omega = \{\|X^\top \varepsilon\|_\infty / n \leq \lambda_0\}$. Show that, when $\lambda > \lambda_0$, if the following two conditions hold

$$\begin{aligned}
\sup_{\|\tau\|_\infty \leq 1} \|X_N^\top X_S (X_S^\top X_S)^{-1} \tau\|_\infty &< \frac{\lambda - \lambda_0}{\lambda + \lambda_0} \\
(\lambda + \lambda_0) \left\| \left\{ \left(\frac{1}{n} X_S^\top X_S \right)^{-1} \right\}_k \right\|_1 &< |\beta_k^0| \quad \text{for } k \in S,
\end{aligned}$$

then on Ω the (unique) Lasso solution satisfies $\text{sgn}(\hat{\beta}_\lambda^L) = \text{sgn}(\beta^0)$.

Proof. Following the proof of theorem 22 in the lecture notes, the KKT conditions become

$$\frac{1}{n} \begin{pmatrix} X_S^\top X_S & X_S^\top X_N \\ X_N^\top X_S & X_N^\top X_N \end{pmatrix} \begin{pmatrix} \beta_S^0 - \hat{\beta}_S \\ -\hat{\beta}_N \end{pmatrix} + \frac{1}{n} \begin{pmatrix} X_S^\top \\ X_N^\top \end{pmatrix} \varepsilon = \lambda \begin{pmatrix} \hat{\nu}_S \\ \hat{\nu}_N \end{pmatrix}.$$

We first prove there exists a Lasso solution $(\hat{\beta}_S, 0)$ which satisfies the KKT conditions. Solving the top equation block for $\hat{\beta}_S$ gives

$$\hat{\beta}_S = \beta_S^0 - \left(\frac{1}{n} X_S^\top X_S \right)^{-1} \left(\lambda \text{sgn}(\beta_S^0) - \frac{1}{n} X_S^\top \varepsilon \right). \quad (1)$$

To prove that $\text{sgn}(\beta_S^0) = \text{sgn}(\hat{\beta}_S)$ holds, note that for $k \in S$ we have

$$\begin{aligned} \left| \left\{ \left(\frac{1}{n} X_S^\top X_S \right)^{-1} \left(\lambda \text{sgn}(\beta_S^0) - \frac{1}{n} X_S^\top \varepsilon \right) \right\}_k \right| &= \left| \left(\frac{1}{n} X_S^\top X_S \right)_k^{-1} \left(\lambda \text{sgn}(\beta_S^0) - \frac{1}{n} X_S^\top \varepsilon \right) \right| \\ &\leq \left\| \left(\frac{1}{n} X_S^\top X_S \right)_k^{-1} \right\|_1 \left\| \lambda \text{sgn}(\beta_S^0) - \frac{1}{n} X_S^\top \varepsilon \right\|_\infty \\ &< \frac{|\beta_k^0|}{\lambda + \lambda_0} \cdot (\lambda + \lambda_0) = |\beta_k^0|, \end{aligned}$$

since $\left\| \frac{1}{n} X_S^\top \varepsilon \right\|_\infty \leq \left\| \frac{1}{n} X^\top \varepsilon \right\|_\infty \leq \lambda_0$. Plugging this result into eq. (1) shows that $\text{sgn}(\hat{\beta}_S) = \text{sgn}(\beta_S^0)$.

We must show that with this choice of $\hat{\beta}$, the bottom block is satisfied, i.e.,

$$\lambda \geq \left\| \frac{1}{n} X_N^\top X_S (\beta_S^0 - \hat{\beta}_S) + \frac{1}{n} X_N^\top \varepsilon \right\|_\infty = \left\| X_N^\top X_S (X_S^\top X_S)^{-1} \left(\lambda \text{sgn}(\beta_S^0) - \frac{1}{n} X_S^\top \varepsilon \right) + \frac{1}{n} X_N^\top \varepsilon \right\|_\infty = (*).$$

Note that

$$\left\| \lambda \text{sgn}(\beta_S^0) - \frac{1}{n} X_S^\top \varepsilon \right\|_\infty \leq \lambda + \lambda_0,$$

and plugging that into the previous equation yields

$$(*) \leq (\lambda + \lambda_0) \cdot \sup_{\|\tau\|_\infty \leq 1} \left\| X_N^\top X_S (X_S^\top X_S)^{-1} \tau \right\|_\infty + \lambda_0 \leq \lambda - \lambda_0 + \lambda_0 = \lambda,$$

which concludes the proof.

TODO: Uniqueness????

□

Question 4. Find the KKT conditions for the group Lasso.

Proof. Recall the objective function in group Lasso is the convex function

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

Note that the subdifferential of the 2-norm is given by

$$\partial \|x\|_2 = \begin{cases} \{x/\|x\|_2\}, & x \neq 0, \\ \{v : \|v\|_2 \leq 1\}, & x = 0. \end{cases}$$

Furthermore, note that the subdifferential of $\frac{1}{2n} \|Y - X\beta\|_2^2$ is $\{-\frac{1}{n} X^\top (Y - X\beta)\}$. We therefore find that 0 lies in the subdifferential if and only if, for $j = 1, \dots, q$, we have

$$\begin{cases} (\frac{1}{n} X^\top (Y - X\beta))_{G_j} = \lambda m_j \beta_{G_j} / \|\beta_{G_j}\|_2, & \text{if } \beta_{G_j} \neq 0, \\ \left\| (\frac{1}{n} X^\top (Y - X\beta))_{G_j} \right\|_2 \leq \lambda m_j, & \text{if } \beta_{G_j} = 0. \end{cases}$$

□

Question 5. (a) Show that

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

where

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

Show that the unique ϑ maximising G is $\vartheta^* = (Y - X\hat{\beta}_\lambda^L)/n$.

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 $\tilde{\vartheta}$ be such that $\|X^\top \tilde{\vartheta}\|_\infty \leq \lambda$. Explain why if

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

then we know that $\hat{\beta}_{\lambda,k}^L = 0$. By considering $\tilde{\vartheta} = Y\lambda/(n\lambda_{\max})$, show that $\hat{\beta}_{\lambda,k}^L = 0$ if

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

Proof. (a) As in the hint, we write the Lasso objective problem as

$$\min_{\substack{(\beta, z) \in \mathbb{R}^p \times \mathbb{R}^n \\ z - X\beta = 0}} \frac{1}{2n} \|Y - z\|_2^2 + \lambda\|\beta\|_1.$$

The Lagrangian is now

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

and the dual function is given by

$$\tilde{f}(\vartheta) = \inf_{(\beta, z) \in \mathbb{R}^p \times \mathbb{R}^n} L(z, \beta, \vartheta) = \inf_{\beta \in \mathbb{R}^p} (\lambda\|\beta\|_1 - \vartheta^\top X\beta) + \inf_{z \in \mathbb{R}^n} \left(\frac{1}{2n} \|Y - z\|_2^2 + \vartheta^\top z \right).$$

For the first term: note that $\vartheta^\top X\beta \leq \|X^\top \vartheta\|_\infty \|\beta\|_1$, with equality for β suitably chosen. Therefore, the first term has infimum $-\infty$ if $\|X^\top \vartheta\|_\infty > \lambda$, and otherwise has infimum 0 when setting $\beta = 0$.

The second term has z -gradient $-\frac{1}{n}(Y - z) + \vartheta$, and equating that to 0 gives $z = Y - n\vartheta$. Plugging this into the second term gives

$$\begin{aligned} \frac{1}{2n} \|n\vartheta\|_2^2 + \vartheta^\top Y - n\vartheta^\top \vartheta &= \vartheta^\top Y - \frac{n}{2} \vartheta^\top \vartheta \\ 0 &= \vartheta^\top Y - \frac{n}{2} \vartheta^\top \vartheta + \frac{1}{2n} Y^\top Y - \frac{1}{2n} Y^\top Y \\ &= \frac{1}{2n} Y^\top Y - \frac{1}{2n} (Y - n\vartheta)^\top (Y - n\vartheta) \\ &= \frac{1}{2n} \|Y\|_2^2 - \frac{1}{2n} \|Y - n\vartheta\|_2^2 = G(\vartheta). \end{aligned}$$

Therefore, the dual function is given by

$$\tilde{f}(\vartheta) = \begin{cases} -\infty & \text{if } \|X^\top \vartheta\|_\infty > \lambda, \\ G(\vartheta) & \text{if } \|X^\top \vartheta\|_\infty \leq \lambda. \end{cases}$$

The optimal value of the dual problem is therefore

$$d^* = \max_{\vartheta \in \mathbb{R}^n} \tilde{f}(\vartheta) = \max_{\vartheta: \|X^\top \vartheta\|_\infty \leq \lambda} G(\vartheta),$$

which equals the optimal value of the primal problem. This proves the first claim.

By the KKT conditions we have $\|X^\top \vartheta\|_\infty = \frac{1}{n} \|X^\top (Y - X\hat{\beta})\|_\infty \leq \lambda$, and plugging in ϑ^* in G gives

$$\begin{aligned} G(\vartheta^*) &= \frac{1}{2n} \left(\|Y\|_2^2 - \|X\hat{\beta}\|_2^2 \right) = \frac{1}{2n} \left(\|Y - X\hat{\beta}\|_2^2 + 2Y^\top X\hat{\beta} - 2\|X\hat{\beta}\|_2^2 \right) \\ &= \frac{1}{2n} \|Y - X\hat{\beta}\|_2^2 + \hat{\beta}^\top \left\{ \frac{1}{n} X^\top (Y - X\hat{\beta}) \right\} \\ &\stackrel{*}{=} \frac{1}{2n} \|Y - X\hat{\beta}\|_2^2 + \lambda \|\beta\|_1, \end{aligned}$$

where \star follows from the KKT conditions. This shows that ϑ^* maximises G over the objective set. **TODO:** Uniqueness of ϑ^* ?

(b) Clearly

$$\max_{\vartheta: G(\vartheta) \geq G(\tilde{\vartheta})} |X_k^\top \vartheta| < \lambda \implies |X_k^\top \vartheta^*| < \lambda.$$

We compute

$$X_k^\top \vartheta^* = (X^\top \vartheta^*)_k = \left(\frac{1}{n} X^\top (Y - X\hat{\beta}) \right)_k = \lambda \hat{\nu}_k,$$

where $\hat{\nu}$ is from the KKT conditions, and we know that $\hat{\beta}_{\lambda,k}^L \neq 0 \implies |\hat{\nu}_k| = 1$. However, we have

$$\lambda > |X_k^\top \vartheta^*| = \lambda |\hat{\nu}_k| \implies |\hat{\nu}_k| < 1 \implies \hat{\beta}_{\lambda,k}^L = 0.$$

Now let $\tilde{\vartheta} = Y\lambda/(n\lambda_{\max})$. It is easily checked that $\|X^\top \tilde{\vartheta}\|_\infty = \lambda$. Now, clearly we have

$$G(\vartheta) \geq G(\tilde{\vartheta}) \implies \|Y - n\vartheta\|_2^2 \leq \|Y - n\tilde{\vartheta}\|_2^2 = \left(1 - \frac{\lambda}{\lambda_{\max}}\right)^2 \|Y\|_2^2 = \left(\frac{\lambda_{\max} - \lambda}{\lambda_{\max}}\right)^2 \|Y\|_2^2.$$

We find that for $G(\vartheta) \geq G(\tilde{\vartheta})$ we have

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

and from the previous observation we conclude $\hat{\beta}_{\lambda,k}^L = 0$ (note that we used $\|X_k\|_2 = \sqrt{n}$). \square

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

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

Conclude that the solution path $\lambda \mapsto \hat{\beta}_\lambda^L$ 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.

Proof. Write $\hat{\beta}_i = \hat{\beta}_{\lambda_i}^L$, $\gamma = t\hat{\beta}_1 + (1-t)\hat{\beta}_2$, then we have

$$\begin{aligned}\frac{1}{n}X^\top(Y - X\gamma) &= t\frac{1}{n}X^\top(Y - X\hat{\beta}_1) + (1-t)\frac{1}{n}X^\top(Y - X\hat{\beta}_2) \\ &= t\lambda_1\hat{\nu}_1 + (1-t)\lambda_2\hat{\nu}_2.\end{aligned}$$

Note that $\hat{\nu}_1$ and $\hat{\nu}_2$ agree on \hat{E}_λ , and on the other points we clearly have $\|\hat{\nu}\|_\infty \leq 1$, so we conclude

$$\frac{1}{n}X^\top(Y - X\gamma) = (t\lambda_1 + (1-t)\lambda_2)\hat{\nu},$$

where $(\hat{\nu})_S = \text{sgn}(\gamma)_S$ and $\|\hat{\nu}\|_\infty \leq 1$. This shows that γ is a Lasso solution.

It follows that the solution path is piecewise linear with knots whenever the sign of the Lasso solution changes. Since the Lasso solution is 0 for λ large enough, there can only be finitely many sign changes, which concludes the proof. (TODO: elaborate on last step) \square

Question 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.

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

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

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

Proof. 1. The minimiser β^* is unique because the objective function is strictly convex: since $\alpha < 1$, the term $(1-\alpha)\|\beta\|_2^2/2$ is strictly convex, and therefore the sum with another convex function is also convex.

If $\beta_k^* \neq \beta_j^*$, then it is easily seen that replacing both β_k^* and β_j^* by $\frac{1}{2}(\beta_k^* + \beta_j^*)$ gives a smaller objective value (the term $\|Y - X\beta\|_2^2$ stays the same and the other two terms will be less by convexity) which contradicts the fact that β^* is a minimiser.

2. For simplicity, let $\beta_A := \hat{\beta}_A^{(m)}$ and $\beta_B := \hat{\beta}_B^{(m-1)}$, so our goal is to find

$$\arg \min_{\beta_k \in \mathbb{R}} g(\beta_k) \quad \text{where } g(\beta_k) = f(\beta_A, \beta_k, \beta_B).$$

Note that for some $h(\beta_A, \beta_B)$ we can write

$$g(\beta_k) = \frac{1}{2n}\|Y - (X_A\beta_A + X_B\beta_B + X_k\beta_k)\|_2^2 + \lambda\alpha|\beta_k| + \frac{\lambda(1-\alpha)}{2}\beta_k^2 + h(\beta_A, \beta_B),$$

and therefore the subdifferential of g in β_k becomes, using $X_k^\top X_k = \|X_k\|_2^2 = n$,

$$\begin{aligned}\partial g(\beta_k) &= -\frac{1}{n}X_k^\top(Y - X_A\beta_A - X_B\beta_B - X_k\beta_k) + \lambda\alpha\hat{\nu} + \lambda(1-\alpha)\beta_k \\ &= -\frac{1}{n}X_k^\top(Y - X_A\beta_A - X_B\beta_B) + \lambda\alpha\hat{\nu} + (1 + \lambda(1-\alpha))\beta_k,\end{aligned}$$

where $\hat{\nu} = \text{sgn}(\beta_k)$ if $\beta_k \neq 0$ and else $\hat{\nu} \in [-1, 1]$. Rewriting gives

$$0 \in \partial g(\beta_k) \iff \beta_k = \frac{\frac{1}{n} X^\top (Y - X_A \beta_A - X_B \beta_B) - \lambda \alpha \hat{\nu}}{1 + \lambda(1 - \alpha)}. \quad (2)$$

We can distinguish three cases:

- (a) If $\frac{1}{n} X^\top (Y - X_A \beta_A - X_B \beta_B) > \lambda \alpha$, we can set $\hat{\nu} = 1$ in eq. (2) (and indeed β_k will be strictly positive).
- (b) If $\frac{1}{n} X^\top (Y - X_A \beta_A - X_B \beta_B) < -\lambda \alpha$, we can set $\hat{\nu} = -1$ in eq. (2) (and indeed β_k will be strictly negative).
- (c) If $\left| \frac{1}{n} X^\top (Y - X_A \beta_A - X_B \beta_B) \right| \leq \lambda \alpha$, we can choose $\hat{\nu} \in [-1, 1]$ such that the expression in eq. (2) becomes 0, and therefore we can choose $\beta_k = 0$.

It is easily checked that these three cases can be combined into the formula

$$\beta_k = \frac{S_{\lambda \alpha}(\frac{1}{n} X^\top (Y - X_A \beta_A - X_B \beta_B))}{1 + \lambda(1 - \alpha)}.$$

□

Question 8. Assume X is an $n \times d$ matrix with i.i.d. rows with $N(\mu, \Sigma)$ distribution and $\|\Sigma\|_{\text{op}} = \sigma^2$. Prove a deviation bound similar to theorem 28 for the maximum likelihood estimator $\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{X})(x_i - \bar{X})^\top$. The bound should hold for $\delta > d$.

Proof. ???

□

Question 9. Let $X \in \mathbb{R}^{n \times p}$ ($n > p$) be a centred data matrix with (thin) SVD $X = UDV^\top$. 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 k th principal component $u^{(k)}$ and loading vector $v^{(k)}$ for $k > 1$ inductively as follows:

$$v^{(k)} := \arg \max_{\substack{\|v\|_2=1, \\ Xv \perp \{u^{(1)}, \dots, u^{(k-1)}\}}} \|Xv\|_2, \quad u^{(k)} := Xv^{(k)}.$$

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

Proof. We can write $X = UDV^\top = \sum_{i=1}^p D_{ii}U_iV_i^\top$. Now, expand any $v \in S^{p-1}$ in the basis defined by the columns of V , so $v = \sum_{i=1}^p \alpha_i V_i$ (we have $\sum_i \alpha_i^2 = 1$). Then we have

$$Xv = \sum_{i=1}^p \sum_{j=1}^p D_{ii} \alpha_j U_i V_i^\top V_j = \sum_{i=1}^p \alpha_i D_{ii} U_i, \quad \|Xv\|_2^2 = \sum_{i=1}^p \alpha_i^2 D_{ii}^2.$$

Now, it is clear that

$$Xv \perp \{u^{(1)}, \dots, u^{(k-1)}\} \xLeftrightarrow{\text{IH}} Xv \perp \{U_1, \dots, U_{k-1}\} \iff \alpha_1, \dots, \alpha_{k-1} = 0.$$

Subject to the constraint $\alpha_1, \dots, \alpha_{k-1} = 0$ and $\sum_i \alpha_i^2 = 1$, it is clear that we can maximise $\|Xv\|_2^2$ by choosing $\alpha_k = \pm 1$ and $\alpha_{k+1}, \dots, \alpha_p = 0$, and then we obtain $v^{(k)} = \pm V_k$ and $u^{(k)} = Xv^{(k)} = \pm D_{kk}U_k$. □

Question 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 $K = \Phi\Phi^\top$.

Proof. If Φ were centred, we could simply compute the eigenvectors U_1, \dots, U_p of K with nonzero eigenvalues $\lambda_1, \dots, \lambda_p$, and the principal components would be $\sqrt{\lambda_1}U_1, \dots, \sqrt{\lambda_p}U_p$. □