MODERN STATISTICAL METHODS Part III
Example Sheet 1 (of 4) RDS/Michaelmas 2017

1. Consider minimising the following objective involving response $Y \in \mathbb{R}^n$ and design matrix $X \in \mathbb{R}^{n\times p}$ over $(\mu, \beta) \in \mathbb{R} \times \mathbb{R}^p$:

$$\|Y - \mu\mathbf{1} - X\beta\|_2^2 + J(\beta).$$

Here $J : \mathbb{R}^p \to \mathbb{R}$ is an arbitrary penalty function. Suppose $\bar{X}_k = 0$ for $k = 1, \ldots, p$. Assuming that a minimiser $(\hat{\mu}, \hat{\beta})$ exists, show that $\hat{\mu} = \bar{Y}$. Now take $J(\beta) = \lambda\|\beta\|_2^2$ so we have the ridge regression objective. Show that

$$\hat{\beta} = (X^TX + \lambda I)^{-1}X^TY.$$

From here onwards, whenever we refer to ridge regression, we will assume X has had its columns mean-centred.

2. Let $X \in \mathbb{R}^{n\times p}$ $(n > p)$ be a centred data matrix with (thin) SVD $X = UDV^T$. We saw in lectures that the first principal component was $D_{11}U_1 = XV_1$. V_1 is known as the first *loading vector*. 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).

3. Consider performing ridge regression when $Y = X\beta^0 + \varepsilon$, where $X \in \mathbb{R}^{n\times p}$ has full column rank, and $\mathrm{Var}(\varepsilon) = \sigma^2 I$. Let the SVD of X be UDV^T and write $U^TX\beta^0 = \gamma$. Show that

$$\frac{1}{n}\mathbb{E}\|X\beta^0 - X\hat{\beta}_\lambda^R\|_2^2 = \frac{1}{n}\sum_{j=1}^{p}\left(\frac{\lambda}{\lambda + D_{jj}^2}\right)^2\gamma_j^2 + \frac{\sigma^2}{n}\sum_{j=1}^{p}\frac{D_{jj}^4}{(\lambda + D_{jj}^2)^2}.$$

Now suppose the size of the signal is n, so $\|X\beta^0\|_2^2 = n$. For what γ is the mean squared prediction error above minimised? For what γ is it maximised?

4. Show that the ridge regression estimates can be obtained by ordinary least squares regression on an augmented data set with $\sqrt{\lambda}I$ added to the bottom of X (where I here is $p \times p$), and p zeroes added to the end of the response Y.

5. In the following, assume that forming AB where $A \in \mathbb{R}^{a\times b}$, $B \in \mathbb{R}^{b\times c}$ requires $O(abc)$ computational operations, and that if $M \in \mathbb{R}^{d\times d}$ is invertible, then forming M^{-1} requires $O(d^3)$ operations.

   (a) Suppose we wish to apply ridge regression to data $(Y, X) \in \mathbb{R}^n \times \mathbb{R}^{n\times p}$ with $n \gg p$. A complication is that the data is split into m separate datasets of size $n/m \in \mathbb{N}$,

$$Y = \begin{pmatrix} Y^{(1)} \\ \vdots \\ Y^{(m)} \end{pmatrix} \qquad X = \begin{pmatrix} X^{(1)} \\ \vdots \\ X^{(m)} \end{pmatrix},$$

with each dataset located on a different server. Moving large amounts of data between servers is expensive. Explain how one can produce ridge estimates $\hat{\beta}_\lambda$ by communicating only $O(p^2)$ numbers from each server to some central server. What is the total order of the computation time required at each server, and at the central server for your approach?

(b) Now suppose instead that $p \gg n$ and it is instead the variables that are split across m servers, so each server has only a subset of $p/m \in \mathbb{N}$ variables for each observation, and some central server stores Y. Explain how one can obtain the fitted values $X\hat{\beta}_\lambda$ communicating only $O(n^2)$ numbers from each server to the central server. What is the total order of the computation time required at each server, and at the central server for your approach?

6. Prove Proposition 4 in our notes. *Hint: For part (ii) it may help to consider the eigendecompositions of positive semi-definite matrices $K^{(1)}$ and $K^{(2)}$ derived from kernels k_1 and k_2 in the form $K^{(1)} = PDP^T = \sum_{i=1}^n P_i P_i^T D_{ii}$ for example.*

7. Let $\mathcal{X} = \{x \in \mathbb{R}^d : \|x\|_2 < 1\}$. Show that $k(x, x') = (1 - x^T x')^{-\alpha}$ defined on $\mathcal{X} \times \mathcal{X}$, where $\alpha > 0$, is a kernel.

8. Suppose we have a matrix of predictors $X \in \mathbb{R}^{n \times p}$ where $p \gg n$. Explain how to obtain the fitted values of the following ridge regression using the kernel trick:

$$\text{Minimise over } \beta \in \mathbb{R}^p, \theta \in \mathbb{R}^{p(p-1)/2}, \gamma \in \mathbb{R}^p,$$
$$\sum_{i=1}^n \left(Y_i - \sum_{k=1}^p X_{ik}\beta_k - \sum_{k=1}^p \sum_{j=1}^{k-1} X_{ik}X_{ij}\theta_{jk} - \sum_{k=1}^p X_{ik}^2 \gamma_k \right)^2 + \lambda_1 \|\beta\|_2^2 + \lambda_2 \|\theta\|_2^2 + \lambda_3 \|\gamma\|_2^2.$$

Note we have indexed θ with two numbers for convenience.

9. Let $\hat{\alpha}$ be a minimiser of $\|Y - K\alpha\|_2^2 + \lambda \alpha^T K \alpha$ over α, with K being a kernel matrix as usual (i.e. symmetric positive semi-definite). Show that $K\hat{\alpha} = K(K + \lambda)^{-1} Y$.

10. Consider minimising

$$c(Y, X, f(x_1) + \mu, \ldots, f(x_n) + \mu) + J(\|f\|_{\mathcal{H}}^2)$$

over $f \in \mathcal{H}$ and $\mu \in \mathbb{R}$ where $\mathcal{H}$ is an RKHS. Here c is an arbitrary loss function and J is strictly increasing. Let k be the reproducing kernel of $\mathcal{H}$. Show that any minimiser $\hat{g}(\cdot) = \hat{f}(\cdot) + \hat{\mu}$ may be written as

$$\hat{g}(\cdot) = \hat{\mu} + \sum_{i=1}^n \hat{\alpha}_i k(\cdot, x_i)$$

where $\hat{\alpha}_i \in \mathbb{R}$ for $i = 1, \ldots, n$.

11. 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 $K = \Phi\Phi^T$.

12. This question proves a result needed for Theorem 7 in our notes. Let $\mathcal{H}$ be a RKHS of functions on $\mathcal{X}$ with reproducing kernel k and suppose $f^0 \in \mathcal{H}$. Let $x_1, \ldots, x_n \in \mathcal{X}$ and let K be the kernel matrix $K_{ij} = k(x_i, x_j)$. Show that

$$\Big(f^0(x_1), \ldots, f^0(x_n) \Big)^T = K\alpha,$$

for some $\alpha \in \mathbb{R}^n$ and moreover that $\|f^0\|_{\mathcal{H}}^2 \geq \alpha^T K \alpha$.

13. Show from first principles that the Sobolev kernel is indeed a (positive definite) kernel.

MODERN STATISTICAL METHODS Part III
Example Sheet 2 (of 4) RDS/Michaelmas 2017

1. Let $x, x' \in \mathbb{R}^p$ and let $\psi \in \{-1, 1\}^p$ be a random vector with independent components taking the values $-1, 1$ each with probability $1/2$. Show that $\mathbb{E}(\psi^T x \psi^T x') = x^T x'$. Construct a random feature map $\hat{\phi} : \mathbb{R}^p \to \mathbb{R}$ such that $\mathbb{E}\{\hat{\phi}(x)\hat{\phi}(x')\} = (x^T x')^2$.

2. Let $\mathcal{X}$ be the set of all subsets of $\{1, \ldots, p\}$ and let $z, z' \in \mathcal{X}$. Let k be the Jaccard similarity kernel. Let π be a random permutation of $\{1, \ldots, p\}$. Let $M = \min\{\pi(k) : k \in z\}$, $M' = \min\{\pi(k) : k \in z'\}$. Show that

$$\mathbb{P}(M = M') = k(z, z')$$

when $z, z' \neq \emptyset$. Now let $\psi \in \{-1, 1\}^p$ be a random vector with i.i.d. components taking the values -1 or 1, each with probability $1/2$. By considering $\mathbb{E}(\psi_M \psi_{M'})$ show that the Jaccard similarity kernel is indeed a kernel. Explain how we can use the ideas above to approximate kernel ridge regression with Jaccard similarity, when n is very large (you may assume that none of the data points are the empty set).

3. Consider the logistic regression model where we assume $Y_1, \ldots, Y_n \in \{-1, 1\}$ are independent and

$$\log\left(\frac{\mathbb{P}(Y_i = 1)}{\mathbb{P}(Y_i = -1)}\right) = x_i^T \beta^0.$$

Show that the maximum likelihood estimate $\hat{\beta}$ minimises

$$\sum_{i=1}^{n} \log\{1 + \exp(-Y_i x_i^T \beta)\}$$

over $\beta \in \mathbb{R}^p$.

4*. Consider the following algorithm for model selection when we have a response $Y \in \mathbb{R}^n$ and matrix of predictors $X \in \mathbb{R}^{n \times p}$.

   (a) First centre Y and all the columns of X. Initialise the current model $M \subseteq \{1, \ldots, p\}$ to be $\emptyset$ and set the current residual R to be Y.

   (b) Find the variable k^* in M^c having the highest correlation in absolute value with the current residual R. Set M to be $M \cup \{k^*\}$. Replace R with the residual from regressing R on X_{k^*}. Further replace each variable in M^c with the residual from regressing itself on X_{k^*}.

   (c) Continue the previous step until $R = 0$.

   Show that this algorithm is equivalent to forward selection. *Hint: Use induction on the iteration m of the algorithm. Consider strengthening the natural inductive hypothesis that the model at iteration m is the same as that selected after m steps of forward selection.*

5. Show that if W is mean-zero and sub-Gaussian with parameter σ, then $\mathrm{Var}(W) \leq \sigma^2$.

6. Verify Hoeffding's lemma for the special case where W is a Rademacher random variable, so W takes the values $-1, 1$ each with probability $1/2$.

7. (a) Let $W \sim \chi^2_d$. Show that

$$\mathbb{P}(|W/d - 1| \geq t) \leq 2e^{-dt^2/8}$$

for $t \in (0, 1)$. You may use the facts that the mgf of a χ^2_1 random variable is $1/\sqrt{1-2\alpha}$ for $\alpha < 1/2$, and $e^{-\alpha}/\sqrt{1-2\alpha} \leq e^{2\alpha^2}$ when $|\alpha| < 1/4$.

(b) Let $A \in \mathbb{R}^{d\times p}$ have i.i.d. standard normal entries. Fix $u \in \mathbb{R}^p$. Use the result above to conclude that

$$\mathbb{P}\left(\left|\frac{\|Au\|_2^2}{d\|u\|_2^2} - 1\right| \geq t\right) \leq 2e^{-dt^2/8}.$$

(c) Suppose we have (data) $u_1, \ldots, u_n \in \mathbb{R}^p$ (note each u_i is a vector), with p large and $n \geq 2$. Show that for a given $\epsilon \in (0, 1)$ and $d > 16\log(n/\sqrt{\epsilon})/t^2$, each data point may be compressed down to $u_i \mapsto Au_i/\sqrt{d} = w_i$ whilst approximately preserving the distances between the points:

$$\mathbb{P}\left(1 - t \leq \frac{\|w_i - w_j\|_2^2}{\|u_i - u_j\|_2^2} \leq 1 + t \text{ for all } i, j \in \{1, \ldots, n\},\ i \neq j\right) \geq 1 - \epsilon.$$

This is the famous Johnson–Lindenstrauss Lemma.

In the following questions assume that $X \in \mathbb{R}^{n\times p}$ has had its columns centred and scaled to have ℓ_2-norm $\sqrt{n}$, and that $Y \in \mathbb{R}^n$ is also centred.

8. Show that any two Lasso solutions when $\lambda > 0$ must have the same ℓ_1-norm.

9. A *convex combination* of a set of points $S = \{v_1, \ldots, v_m\} \subseteq \mathbb{R}^{d'}$ is any point of the form

$$\alpha_1 v_1 + \cdots + \alpha_m v_m,$$

where $\alpha_j \in \mathbb{R}$ and $\alpha_j \geq 0$ for $j = 1, \ldots, m$, and $\sum_{j=1}^m \alpha_j = 1$. Carathéodory's Lemma states that if S is in a subspace of dimension d, any v that is a convex combination of points in S can be expressed as a convex combination of $d + 1$ points from S i.e. there exist $j_1, \ldots, j_{d+1} \in \{1, \ldots, m\}$ and non-negative reals $\alpha_1, \ldots, \alpha_{d+1}$ summing to 1 with

$$v = \alpha_1 v_{j_1} + \cdots + \alpha_{d+1} v_{j_{d+1}}.$$

With this knowledge, show that for any value of λ, there is always a Lasso solution with no more than n non-zero coefficients.

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

11. 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 $\|\beta\|_0 := |\{k : \beta_k \neq 0\}|$?

MODERN STATISTICAL METHODS Part III
Example Sheet 3 (of 4) RDS/Michaelmas 2017

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

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

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

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

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

$$\sup_{\tau:\|\tau\|_\infty \leq 1} \|X_N^T X_S (X_S^T X_S)^{-1}\tau\|_\infty < \frac{\lambda - \lambda_0}{\lambda + \lambda_0}$$
$$(\lambda + \lambda_0)\|\{(\tfrac{1}{n}X_S^T X_S)^{-1}\}_k\|_1 < |\beta_k^0| \qquad \text{for } k \in S,$$

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

4. Find the KKT conditions for the group Lasso.

5. (a) Show that

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

   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 $\theta^* = (Y - X\hat{\beta}_\lambda^L)/n$. *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 $\|X^T\tilde{\theta}\|_\infty \leq \lambda$. Explain why if

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

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

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

6. Consider the Lasso and let $\hat{E}_\lambda = \{k : \frac{1}{n}|X_k^T(Y - X\hat{\beta}_\lambda^L)| = \lambda\}$ 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 $t \in [0, 1]$,

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

Hint: Check the KKT conditions. 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.

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.

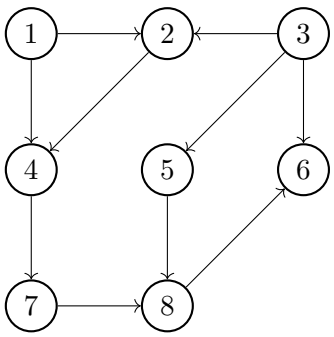
   (a) Suppose X has two columns X_j and X_k that are identical and $\alpha < 1$. Explain why the minimising β^* above is unique and has $\beta_k^* = \beta_j^*$.

   (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 $m \geq 1$,

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

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

8. 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.

9. 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 $j < k$,

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

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}?

10. Let $Z \sim N_p(\mu, \Sigma)$ with Σ positive definite. Prove that if the distribution of Z is pairwise Markov with respect to an undirected graph $\mathcal{G}$, then it is also global Markov with respect to $\mathcal{G}$. *Hint: First consider $A \cup B \cup S = \{1, \ldots, p\}$; then argue that the general case follows.*

MODERN STATISTICAL METHODS Part III
Example Sheet 4 (of 4) RDS/Michaelmas 2016

1. In this question we will outline an algorithm to compute the graphical Lasso.

   (a) Let
   $$Q(\Omega) = -\log\det(\Omega) + \mathrm{tr}(S\Omega) + \lambda\|\Omega\|_1$$
   be the graphical Lasso objective with $\hat{\Omega} = \underset{\Omega \succ 0}{\mathrm{argmin}}\, Q(\Omega)$ assumed unique. Consider the following version of the graphical Lasso objective:
   $$\min_{\Omega, \Theta \succ 0} \{-\log\det(\Omega) + \mathrm{tr}(S\Omega) + \lambda\|\Theta\|_1\}$$
   subject to $\Omega = \Theta$. By introducing the Lagrangian for this objective, show that
   $$p + \max_{U : S+U \succ 0,\, \|U\|_\infty \leq \lambda} \log\det(S+U) \leq Q(\hat{\Omega}).$$
   Here $\|U\|_\infty = \max_{j,k} |U_{jk}|$ and p is the number of columns in the underlying data matrix X. *Hint: Write the additional term in the Lagrangian as* $\mathrm{tr}(U(\Omega - \Theta))$.

   (b) Suppose that U^* is the unique maximiser of the LHS. Show that $\hat{\Omega} = (S + U^*)^{-1}$.

   (c) Now consider
   $$\hat{\Sigma} = \underset{W : W \succ 0,\, \|W - S\|_\infty \leq \lambda}{\mathrm{argmin}} \; -\log\det(W). \tag{1}$$
   By using the formula for the determinant in terms of Schur complements, show that $(\hat{\Sigma}_{jj}, \hat{\Sigma}_{-j,j}) = (\alpha^*, \beta^*)$, where (α^*, β^*) solve the following optimisation problem over (α, β):
   $$\begin{aligned} &\text{minimise} && -\alpha + \beta^T \hat{\Sigma}^{-1}_{-j,-j} \beta, \\ &\text{such that} && \|\beta - S_{-j,j}\|_\infty \leq \lambda,\ |\alpha - S_{jj}| \leq \lambda. \end{aligned}$$
   Conclude that $\alpha^* = S_{jj} + \lambda$. ($\beta^*$ can be found by standard quadratic programming techniques, or by converting the optimisation to a standard Lasso optimisation problem; thus we can perform block coordinate descent on the optimisation problem in (1), updating a row and corresponding column of W at each iteration.)

2. Explain why if P is faithful to a DAG $\mathcal{G}$ then it also satisfies causal minimality w.r.t. $\mathcal{G}$.

3. Show that two DAGs $\mathcal{G}_1$ and $\mathcal{G}_2$ are Markov equivalent only if they have the same skeleton and v-structures. You may assume that for every DAG $\mathcal{G}$ there is a distribution P which is faithful to it.

4. Suppose P is faithful to a DAG $\mathcal{G}$. Show that the moral graph of $\mathcal{G}$ is the CIG.

5. In a DAG $\mathcal{G} = (V, E)$, define the set of *non-descendants* of a node k, written $\mathrm{nd}(k)$, by
   $$\mathrm{nd}(k) = V \setminus (\mathrm{de}(k) \cup \{k\})$$
   Show that if P is global Markov w.r.t. P and $Z \sim P$ then for any node k
   $$Z_k \perp\!\!\!\perp Z_{\mathrm{nd}(k)} | Z_{\mathrm{pa}(k)}.$$

6. Consider an SEM for $Z \in \mathbb{R}^p$ where Z has a joint density f (w.r.t. a product measure). Suppose that Z_k has no parents. Show that

$$f(z|do(Z_k = z_k)) = f(z_{-k}|z_k).$$

Here the LHS is the joint density of Z in the new SEM where we have replaced the structural equation involving Z_k with $Z_k = z_k$, and the RHS is the conditional density of $Z_{-k}|Z_k$.

In the following questions, suppose there are m null hypotheses being tested, $H_1, \ldots, H_m$, and let $p_1, \ldots, p_m$ be the associated p-values, and let $p_{(1)} \leq \cdots \leq p_{(m)}$ be the ordered p values (so (i) is the index of the ith smallest p-value). Further let I_0 be the set of true null hypotheses.

7. Show that if all null hypotheses are true, then the FDR is equivalent to the FWER.

8. Show that the definition of Holm's procedure as the closed testing procedure with the local tests as the Bonferroni test is equivalent to the the step-down procedure definition.

9. The Benjamini–Hochberg procedure allows us to control the FDR when the p-values of true null hypotheses are independent of each other, and independent of the false null hypotheses. The following variant of the method, known as the Benjamini–Yekutieli procedure allows us to control the FDR under arbitrary dependence of the p-values, and works as follows. Define

$$\gamma_m = 1 + \frac{1}{2} + \cdots + \frac{1}{m}.$$

Let $\hat{k} = \max\{i : p_{(i)} \leq \alpha i/(m\gamma_m)\}$ and reject $H_{(1)}, \ldots, H_{(\hat{k})}$. First show that the FDR of this procedure satisfies

$$\text{FDR} = \sum_{i \in I_0} \mathbb{E}\left(\frac{1}{R} \mathbb{1}_{\{p_i \leq \alpha R/(m\gamma_m)\}} \mathbb{1}_{\{R>0\}}\right).$$

Now go on to prove that $\text{FDR} \leq \alpha m_0/m \leq \alpha$. *Hint: Verify that that for any $r \in \mathbb{N}$ we have*

$$\frac{1}{r} = \sum_{j=1}^{\infty} \frac{\mathbb{1}_{\{j \geq r\}}}{j(j+1)},$$

and use this to replace $1/R$.

10. Consider the closed testing procedure applied to m hypotheses $H_1, \ldots, H_m$. Let $\mathcal{R}$ be the collection of all $I \subseteq \{1, \ldots, m\}$ for which for all $J \supseteq I$, the local test $\phi_J = 1$. Now suppose that (perhaps after having looked at the results of the ϕ_I), we decide we want to reject a set of hypotheses indexed by $B \subseteq \{1, \ldots, m\}$. Let

$$t_\alpha(B) = \max\{|I| : I \subseteq B,\ I \notin \mathcal{R}\}.$$

Show that $\{0, 1, \ldots, t_\alpha(B)\}$ gives a $1 - \alpha$ confidence set for the number of false rejections in B. That is, show that

$$\mathbb{P}(|B \cap I_0| > t_\alpha(B)) \leq \alpha,$$

and that this is true no matter how B is chosen. *Hint: Argue by working on the event $\{\phi_{I_0} = 0\}$.*

In the following questions, let all quantities be as defined in Section 4.3 of the lecture notes concerning the debiased Lasso.

11. Show that
$$(\hat{\Theta}\hat{\Sigma}\hat{\Theta}^T)_{jj} = \frac{1}{n}\|X_j - X_{-j}\hat{\gamma}^{(j)}\|_2^2/\hat{\tau}j^4.$$

12. Show that
$$\frac{1}{n}X_j^T(X_j - X_{-j}\hat{\gamma}^{(j)}) = \frac{1}{n}\|X_j - X_{-j}\hat{\gamma}^{(j)}\|_2^2 + \lambda_j\|\hat{\gamma}^{(j)}\|_1.$$

13. Prove that $\mathbb{P}(\Lambda_n) \to 1$, where the sequence of events Λ_n is defined in the proof of Theorem 40. *Hint: Note that here the design matrix X has not been centred and scaled. Therefore to control the probability of $\mathbb{P}(2\|X^T\varepsilon\|_\infty/n \leq \lambda)$ it may help to use the arguments of Theorem 25 i.e. treating $X_j^T\varepsilon$ as a sum of i.i.d. products of (sub)-Gaussian random variables.*