MODERN STATISTICAL METHODS Part III
Example Sheet 4 (of 4) RDS/Michaelmas 2018

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
   $$p + \max_{U: S+U \succ 0,\, \|U\|_\infty \le \lambda} \log\det(S+U) \le Q(\hat{\Omega}).$$
   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
   $$\hat{\Sigma} = \underset{W: W \succ 0,\, \|W - S\|_\infty \le \lambda}{\mathrm{argmin}} -\log\det(W). \tag{1}$$
   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):
   $$\begin{aligned} &\text{minimise} && -\alpha + \beta^T \hat{\Sigma}_{-j,-j}^{-1} \beta, \\ &\text{such that} && \|\beta - S_{-j,j}\|_\infty \le \lambda,\ |\alpha - S_{jj}| \le \lambda. \end{aligned}$$
   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 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.*