STATISTICAL LEARNING II Rajen D. Shah
Example Sheet 2 (of 2) 2020

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

   (a) Let
   $$Q(\Omega) = -\log\det(\Omega) + \operatorname{tr}(S\Omega) + \lambda\|\Omega\|_1$$
   be the graphical Lasso objective with $\hat{\Omega} = \operatorname{argmin}_{\Omega \succ 0} Q(\Omega)$ assumed unique. Consider the following version of the graphical Lasso objective:
   $$\min_{\Omega,\Theta \succ 0}\{-\log\det(\Omega) + \operatorname{tr}(S\Omega) + \lambda\|\Theta\|_1\}$$
   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* $\operatorname{tr}(U(\Omega - \Theta))$.

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

   (c) Now consider
   $$\hat{\Sigma} = \operatorname*{argmin}_{W: W \succ 0,\, \|W - S\|_\infty \leq \lambda} -\log\det(W). \tag{1}$$
   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):
   $$\begin{aligned} \text{minimise} \qquad & -\alpha + \beta^T \hat{\Sigma}^{-1}_{-j,-j}\beta, \\ \text{such that} \qquad & \|\beta - S_{-j,j}\|_\infty \leq \lambda,\ |\alpha - S_{jj}| \leq \lambda. \end{aligned}$$
   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 $\operatorname{nd}(k)$, by
   $$\operatorname{nd}(k) = V \setminus (\operatorname{de}(k) \cup \{k\})$$
   Show that if P is global Markov w.r.t. P and $Z \sim P$ then for any node k
   $$Z_k \perp\!\!\!\perp Z_{\operatorname{nd}(k)} | Z_{\operatorname{pa}(k)}.$$

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, let all quantities be as defined in Section 5 of the lecture notes concerning the debiased Lasso.

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

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

9. Prove that $\mathbb{P}(\Lambda_n) \to 1$, where the sequence of events Λ_n is defined in the proof of Theorem 36. *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 treat $X_j^T\varepsilon$ as a sum of i.i.d. products of (sub)-Gaussian random variables.*