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
   $$Q(\Omega) = -\log\det(\Omega) + \mathrm{tr}(S\Omega) + \lambda\|\Omega\|_1$$
   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* $\mathrm{tr}(U(\Omega - \Theta))$.
   Solution: We know that for all symmetric $U \in \mathbb{R}^{p \times p}$,
   $$\min_{\Omega, \Theta \succ 0} \{-\log\det(\Omega) + \mathrm{tr}(S\Omega) + \lambda\|\Theta\|_1 + \mathrm{tr}(U(\Omega - \Theta))\} \leq Q(\hat{\Omega}^L).$$
   Subdifferentiating the LHS, we see that for a minimiser of the LHS, Ω^*, Θ^* we have $S + U = \Omega^{*,-1}$ provided $S + U \succ 0$, and $U = \lambda\Gamma$ where $\|\Gamma\|_\infty \leq 1$ and $\Gamma_{jk} = \mathrm{sgn}(\Theta^*_{jk})$ when $\Theta^*_{jk} \neq 0$. The latter implies that $\mathrm{tr}(\Theta^* U) = \lambda\|\Theta^*\|_1$. Thus we get that the LHS is
   $$\log\det(S+U) + \mathrm{tr}((S+U)(S+U)^{-1}) = \log\det(S+U) + p,$$
   and as the inequality is true for all U, we may take the maximum over U.

   (b) Suppose that U^* is the unique maximiser of the LHS. Show that $\hat{\Omega} = (S + U^*)^{-1}$.
   Solution: The KKT conditions for the original objective Q tell us that
   $$\hat{\Omega}^{-1} - S = \lambda\hat{\Gamma}$$
   where $\|\hat{\Gamma}\|_\infty \leq 1$ and $\Gamma_{jk} = \mathrm{sgn}(\Omega_{jk})$. Note that $\|\hat{\Omega}^{-1} - S\|_\infty \leq \lambda$, so this is a feasible value of U. We will show that it is the optimal U. We see that
   $$\mathrm{tr}(\hat{\Omega}(\hat{\Omega}^{-1} - S)) = \lambda\|\Omega\|_1$$
   so
   $$\begin{aligned} Q(\hat{\Omega}) &= \log\det(\hat{\Omega}^{-1}) + \mathrm{tr}(S\hat{\Omega}) + \mathrm{tr}(\hat{\Omega}(\hat{\Omega}^{-1} - S)) \\ &= \log\det(\hat{\Omega}^{-1}) + p. \end{aligned}$$
   This show that taking $U = \hat{\Omega}^{-1} - S$ gives equality. So by uniqueness we must have $U^* = \hat{\Omega}^{-1} - S$.

(c) Now consider

$$\hat{\Sigma} = \underset{W: W \succ 0,\, \|W - S\|_\infty \leq \lambda}{\operatorname{argmin}} - \log\det(W). \tag{1}$$

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.)
Solution: Follows from noting that

$$\log\det(W) = \log(W_{k,k,} - W_{k,-k} W^{-1}_{-k,-k} W_{-k,k}) + \log\det(W_{-k,-k}).$$

2. Explain why if P is faithful to a DAG $\mathcal{G}$ then it also satisfies causal minimality w.r.t. $\mathcal{G}$.
Solution: Removing an edge from $\mathcal{G}$ will introduce an extra conditional independency in P, but this contradicts the fact that $\mathcal{G}$ contains all conditional independencies in P. Therefore P cannot be Markov w.r.t. a proper subgraph of $\mathcal{G}$.

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.
Solution: Let P be faithful to $\mathcal{G}_1$. Note that P is Markov w.r.t. $\mathcal{G}_2$. We have

$$\begin{aligned} j,\, k \text{ not adjacent in } \mathcal{G}_2 &\Rightarrow \exists \text{ set } S \text{ such that } j \text{ and } k \text{ are } d\text{-separated by } S \text{ in } \mathcal{G}_2 \quad \text{(Prop. 34)} \\ &\Rightarrow Z_j \perp\!\!\!\perp Z_k | Z_S \quad \text{(Markov property)} \\ &\Rightarrow j \text{ and } k \text{ are } d\text{-separated by } S \text{ in } \mathcal{G}_1 \quad \text{(faithfulness)} \\ &\Rightarrow j,\, k \text{ not adjacent in } \mathcal{G}_1 \quad \text{(Prop. 34)}. \end{aligned}$$

Now repeat the argument with $\mathcal{G}_1$ and $\mathcal{G}_2$ swapping places to show that $\mathcal{G}_1$ and $\mathcal{G}_2$ have the same skeletons. The argument for the v-structures is similar but uses Prop. 35 instead of Prop. 34.

4. Suppose P is faithful to a DAG $\mathcal{G}$. Show that the moral graph of $\mathcal{G}$ is the CIG.
Solution: Consider two non-adjacent nodes j, k in $\mathcal{G}$ which do not have a common child. We claim that j and k are d-separated by $\{j,k\}^c$. Indeed any path with 3 or more edges will be blocked by $\{j,k\}^c$, and the only path of two edges that cannot be blocked is a v-structure. Conversely, two non-adjacent nodes j, k with a common child cannot be d-separated by $\{j,k\}^c$, so by faithfulness, we know $Z_k \not\perp\!\!\!\perp Z_j | Z_{-jk}$. Clearly two adjacent nodes j, k can also not be d-separated (by any set at all) and so again $Z_k \not\perp\!\!\!\perp Z_j | Z_{-jk}$. We therefore see that we don't have an edge between j and k in the moral graph iff. $Z_j \perp\!\!\!\perp Z_k | Z_{-jk}$.

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. $\mathcal{G}$ and $Z \sim P$ then for any node k

$$Z_k \perp\!\!\!\perp Z_{\mathrm{nd}(k)\setminus \mathrm{pa}(k)} | Z_{\mathrm{pa}(k)}.$$

Solution: We need to show k and $\mathrm{nd}(k) \setminus \mathrm{pa}(k)$ are d-separated by $\mathrm{pa}(k)$. Every path from k to $\mathrm{nd}(k) \setminus \mathrm{pa}(k)$ starting $k \leftarrow \cdots$ will be blocked. All other paths must contain a collider, since we cannot have a directed path from k leading to $\mathrm{nd}(k)$ by definition of $\mathrm{nd}(k)$. But the first collider along the path cannot have a descendant in $\mathrm{pa}(k)$ since this would imply a cycle (c.f. Prop. 34).

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 $Z_{-k}|Z_k$.
Solution: By the Markov factorisation property, the LHS is

$$\prod_{j \neq k} f(z_j | z_{\mathrm{pa}(j)})$$

which is the same as the RHS.

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

7. Show that

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

Solution: It is easy to see that the LHS equals $\|X\hat{\theta}_j\|_2^2/n$, and the result follows from the equation in the middle of page 53 in the notes.

8. Show that

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

Solution: Let us rewrite $X_j = Y$, $X_{-j} = X$, $\hat{\gamma}^{(j)} = \hat{\beta}$, $\lambda_j = \lambda$ for notational simplicity. Also let $\hat{S} = \{k : \hat{\beta}_k \neq 0\}$. We have

$$\begin{aligned}\frac{1}{n}Y^T(Y - X\hat{\beta}) &= \frac{1}{n}(Y - X\hat{\beta} + X\hat{\beta})^T(Y - X\hat{\beta}) \\ &= \frac{1}{n}\|Y - X\hat{\beta}\|_2^2 + \frac{1}{n}\hat{\beta}^T X^T(Y - X\hat{\beta}).\end{aligned}$$

Also $X_{\hat{S}}^T(Y - X\hat{\beta})/n = \lambda \mathrm{sgn}(\hat{\beta}_{\hat{S}})$, so we see the final term above is $\lambda\|\hat{\beta}\|_1$.

9. Prove that $\mathbb{P}(\Lambda_n) \to 1$, where the sequence of events Λ_n is defined in the proof of Theorem 36.
Solution: We need to show that

(a) $\mathbb{P}(\{\phi^2_{\hat{\Sigma},s_{\max}} \geq c_{\min}/2\} \cup_j \{\phi^2_{\hat{\Sigma}_{-j,-j},s_j} \geq c_{\min}/2\}) \to 1$,

(b) $\mathbb{P}(\cup_j\{2\|X^T_{-j}\varepsilon^{(j)}\|_\infty/n > \lambda\} \cup \{2\|X^T\varepsilon\|_\infty/n > \lambda\}) \to 0$,

(c) $\mathbb{P}(\cup_j\{\Omega_{jj}\|\varepsilon^{(j)}\|_2^2/n < 1 - 4\sqrt{\log(p)/n}\}) \to 0$,

for A sufficiently large, where $\lambda = A\sqrt{\log(p)/n}$. Consider (a) first. Firstly, we know that $\min_m \phi^2_{\Sigma,m} \geq c_{\min}$ (see the discussion preceding Theorem 21). Clearly also $\min_m \phi^2_{\Sigma_{-j,-j},m} \geq c_{\min}$.

From Lemma 22, we know that on the event

$$\Xi_n = \{\max_{jk} |\hat{\Sigma}_{jk} - \Sigma_{jk}| \leq c_{\min}/(32 s_{\max})\}$$

we have in particular that $\phi^2_{\hat{\Sigma},s} \geq \phi^2_{\Sigma,s}/2 \geq c_{\min}/2$, and also $\phi^2_{\hat{\Sigma}_{-j,-j},s_j} \geq \phi^2_{\Sigma_{-j,-j},s_j}/2 \geq c_{\min}/2$ for all j. Theorem 25 then shows that $\mathbb{P}(\Xi_n) \to 1$ provided $s_{\max}\sqrt{\log(p)/n} \to 0$ (which is true by assumption).

For (c), note that $\|\varepsilon^{(j)}\|_2^2\Omega_{jj} \sim \chi^2_n$. From question 8 (a) on example sheet 2, we know that if $W \sim \chi^2_d$, then $\mathbb{P}(W/d \leq 1 - t) \leq e^{-dt^2/8}$. Thus by a union bound, we have

$$\mathbb{P}(\cup_j\{\Omega_{jj}\|\varepsilon^{(j)}\|_2^2/n < 1 - 4\sqrt{\log(p)/n}\}) \leq p\exp(-2\log(p)) = 1/p \to 0.$$

Finally, we turn to (b). We know that $X_{ij}\varepsilon_i$ satisfies Bernstein's condition with parameter $(8\Sigma_{jj}\sigma, 4\Sigma_{jj}\sigma)$. Thus noting that $0 < c_{\min} < \Sigma_{jj} \leq 1$,

$$p\mathbb{P}\left(\left|\sum_{i=1}^n X_{ij}\varepsilon_i\right| \geq \lambda\right) \leq 2\exp\left(\frac{-n\lambda^2}{2(64\Sigma^2_{jj}\sigma^2 + 4\Sigma_{jj}\sigma\lambda)} + \log p\right) \leq 2p^{-A^2/c_1+1}$$

for some constant $c_1 > 0$ and λ sufficiently small (so n is sufficiently large). Similarly noting that $1 \leq \Omega_{jj} \leq c^{-1}_{\min}$,

$$p^2\mathbb{P}\left(\left|\sum_{i=1}^n X_{ik}\varepsilon_i^{(j)}\right| \geq \lambda\right) \leq 2\exp\left(\frac{-n\lambda^2}{2(64\Omega^{-2}_{jj}\sigma^2 + 4\Omega^{-1}_{jj}\sigma\lambda)} + 2\log p\right) \leq 2p^{-A^2/c_2+1}$$

for $k \neq j$, some constant $c_2 > 0$ and λ sufficiently small. Thus by taking A sufficiently large, we have

$$\begin{aligned}&\mathbb{P}(\cup_j\{2\|X^T_{-j}\varepsilon^{(j)}\|_\infty/n > \lambda\} \cup \{2\|X^T\varepsilon\|_\infty/n > \lambda\})\\ \leq &\sum_{j=1}^p \mathbb{P}(2|X_j^T\varepsilon| > \lambda) + \sum_j\sum_{j\neq k}\mathbb{P}(2|X_k^T\varepsilon^{(j)}| > \lambda) \to 0.\end{aligned}$$