HIGH-DIMENSIONAL STATISTICS APTS
Assessment material RDS/July 2017

You may wish to pick out 3 or 4 questions from those given below. In all of the following, assume that the design matrix $\mathbf{X} \in \mathbb{R}^{n\times p}$ has had its columns mean-centred and then scaled to have ℓ_2-norm $\sqrt{n}$.

Ridge regression

1. Consider performing ridge regression when $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}^0 + \boldsymbol{\varepsilon} - \mathbf{1}\bar{\varepsilon}$ (so the response has been centred), where $\mathbf{X} \in \mathbb{R}^{n\times p}$ has full column rank, and $\mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$. Let the (thin) SVD of $\mathbf{X}$ be $\mathbf{UDV}^T$ and write $\mathbf{U}^T\mathbf{X}\boldsymbol{\beta}^0 = \boldsymbol{\gamma}$. Show that

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

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

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

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

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

   with each dataset located on a different server. Moving large amounts of data between servers is expensive. Explain how one can produce ridge estimates $\hat{\boldsymbol{\beta}}^{\mathrm{R}}_\lambda$ by communicating only $p(p+1)$ 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 $\mathbf{Y}$. Explain how one can obtain the fitted values $\mathbf{X}\hat{\boldsymbol{\beta}}^{\mathrm{R}}_\lambda$ communicating only 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?

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

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

   Note we have indexed $\boldsymbol{\theta}$ with two numbers for convenience.

The Lasso

4. (a) When proving results on the Lasso, we started with the so-called basic inequality that
$$\frac{1}{2n}\|\mathbf{X}(\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}})\|_2^2 \leq \frac{1}{n}\boldsymbol{\varepsilon}^T\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0) + \lambda\|\boldsymbol{\beta}^0\|_1 - \lambda\|\hat{\boldsymbol{\beta}}\|_1.$$
Show that in fact we can improve this to
$$\frac{1}{n}\|\mathbf{X}(\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}})\|_2^2 \leq \frac{1}{n}\boldsymbol{\varepsilon}^T\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0) + \lambda\|\boldsymbol{\beta}^0\|_1 - \lambda\|\hat{\boldsymbol{\beta}}\|_1. \quad (1)$$
Hint: Start from the KKT conditions for the Lasso.

(b) Under the assumptions of Theorem 7 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
$$\frac{1}{n}\|\mathbf{X}(\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}})\|_2^2 \leq \frac{9A^2\log(p)}{4\phi^2}\frac{\sigma^2 s}{n}.$$

5. Consider once more the setup of Theorem 7 and its proof. Let $\hat{S} = \{j : \hat{\beta}_j \neq 0\}$ and set $\hat{s} = |\hat{S}|$. Show that on the event $\Omega = \{2\|\mathbf{X}^T\boldsymbol{\varepsilon}\|_\infty/n \leq \lambda\}$, for any non-empty subset B of $\hat{S}$, we have
$$\frac{1}{n}\mathrm{sgn}(\hat{\boldsymbol{\beta}}_B)^T\mathbf{X}_B^T\mathbf{X}(\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}}) \geq \frac{\lambda|B|}{2}. \quad (2)$$
Let κ_m^2 be the maximum eigenvalue of $\mathbf{X}_M^T\mathbf{X}_M/n$ over all $M \subset \{1,\ldots,p\}$ with $|M| = m$. Let
$$m^* = \min\{m \geq 1 : m > 9\kappa_m^2 s/\phi^2\},$$
with $m^* = \infty$ if there does not exist any m satisfying the condition defining the set above. Prove that on Ω, we have $\hat{s} < m^*$. *Hint: First try to obtain an upper bound on the LHS of* (2) *involving* $\kappa_{|B|}$ *by making use of the* (1) *and the Cauchy–Schwarz inequality.*

By considering the minimality of m^*, show furthermore that on Ω, we have $\hat{s} \leq 9\kappa_{m^*}^2 s/\phi^2$. [In words, the number of non-zero coefficients of the Lasso is of the same order as the number of true non-zeroes.]

Graphical modelling

7. Let $\mathbf{Z} = (Z_1,\ldots,Z_p)^T \in \{0,1\}^p$ be a binary random vector with probability mass function given by
$$\mathbb{P}(Z_1 = z_1,\ldots,Z_p = z_p) = \exp\left(\Theta_{00} + \sum_{k=1}^{p}\Theta_{0k}z_k + \sum_{k=1}^{p}\sum_{j=1}^{k-1}\Theta_{jk}z_jz_k - \Phi(\boldsymbol{\Theta})\right)$$
where $\exp(-\Phi(\boldsymbol{\Theta}))$ is a normalising constant. Show that
$$\mathrm{logit}(\mathbb{P}(Z_k = 1|\mathbf{Z}_{-k} = \mathbf{z}_{-k})) = \Theta_{0k} + \sum_{j:j<k}\Theta_{jk}z_j + \sum_{j:j>k}\Theta_{kj}z_j,$$
where $\mathrm{logit}(q) = \log\{q/(1-q)\}$ for $q \in (0,1)$. Conclude that, for $j < k$,
$$Z_j \perp\!\!\!\perp Z_k|\mathbf{Z}_{-jk} \iff \Theta_{jk} = 0.$$
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}?

High-dimensional inference

In the following questions, suppose there are m null hypotheses being tested, $H_1, \ldots, H_m$. 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.

8. Consider the closed testing procedure applied to m hypotheses and 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* $\{\phi_{I_0} = 0\}$. This question is based on the *cherry picking* procedure of Goeman and Solari [2011].

9. Suppose we have a family of intersection hypotheses $H_I : I \in \mathcal{I}$ that is hierarchical in the sense that for any $I,\ J \in \mathcal{I}$, we either have $I \cap J = \emptyset$ or $I \subseteq J$ or $J \subseteq I$. Suppose that for each H_I, $I \in \mathcal{I}$ we have a p-value p_I. Define the adjusted p-value of H_I to be

$$p_I^{\text{adj}} = \max_{J : J \in \mathcal{I},\, J \supseteq I} \frac{m}{|J|} p_J.$$

   Consider the procedure [Meinshausen, 2008] that rejects all hypotheses H_I for which $p_I^{\text{adj}} \leq \alpha$. Let $\mathcal{I}_0$ be the subset of $\mathcal{I}$ consisting of true intersection hypotheses, so

$$\mathcal{I}_0 = \{I \in \mathcal{I} : I \subseteq I_0\}.$$

   Show that with probability at least $1 - \alpha$, this procedure makes no false rejections, so no intersection hypothesis indexed by $\mathcal{I}_0$ is rejected. *Hint: Consider the set* $\mathcal{T}_0 \subseteq \mathcal{I}$ *of* $I \in \mathcal{I}_0$ *that are maximal, that is*

$$\mathcal{T}_0 = \{I \in \mathcal{I}_0 :\ \textit{if } J \in \mathcal{I}_0 \textit{ then } J \subseteq I \textit{ or } J \cap I = \emptyset\}.$$

10. 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 [Benjamini and Yekutieli, 2001] 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 $1/R$.

References

Y. Benjamini and D. Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, 29:1165–1188, 2001.

J. J. Goeman and A. Solari. Multiple testing for exploratory research. *Statistical Science*, pages 584–597, 2011.

N. Meinshausen. Hierarchical testing of variable importance. *Biometrika*, 95:265–278, 2008.