Stat243: Problem Set 7, Due Monday November 25

November 8, 2019

This covers Units 9 and 10.

- It's due at 2 pm on Monday November 25, **both submitted as a PDF to bCourses as well as committed to your Github repository**.
- Please note my comments in the syllabus about when to ask for help and about working together. In particular, **please give the names of any other students that you worked with on the problem set and indicate in comments any ideas or code you borrowed from another student.**
- The formatting requirements are the same as previous problem sets. Note that for any mathematical derivations, you are welcome to write these out by hand (neatly!), but if you do so, you must scan/take a picture of your derivation and place the scan/picture into a PDF in correct numerical order by problem number with the solutions for the other problems.

Problems

1. The goal of this problem is to think carefully about the design and interpretation of simulation studies, which we'll talk about in Unit 10. In particular, we'll work with Cao et al. (2015), an article in the Journal of the Royal Statistical Society, Series B, which is a leading statistics journal. The article is available as *cao_etal_2015.pdf* under the *ps* directory on Github. Read Section 1, Section 2.1, and Section 4 of the article.

   You don't need to understand their method for fitting the regression [i.e., you can treat it as some black box algorithm] or the theoretical development. In particular, you don't need to know what an estimating equation is - you can think of it as an alternative to maximum likelihood or to least squares for estimating the parameters of the statistical model. Equation 3 on page 759 is analogous to taking the sum of squares for a regression model and differentiating with respect to β. To find $\hat{\beta}$ one sets the equation equal to zero and solves for β. As far as the kernel, its role is to weight each pair of observation and covariate value. This downweights pairs where the covariate is measured at a very different time than the observation.

   Briefly (a few sentences for each of the three questions below) answer the following questions.

   (a) What are the goals of their simulation study and what are the metrics that they consider in assessing their method?

   (b) What choices did the authors have to make in designing their simulation study? What are the key aspects of the data generating mechanism that might affect their assessment of their method?

(c) Consider their tables reporting the simulation results (Tables 1-3). For a method to be a good method, what would one want to see numerically in these columns?

In Section on November 26, we'll talk in more detail about this simulation study.

2. Suppose I have a statistical method that estimates a regression coefficient and its standard error. As in the Cao et al. paper, I develop a simulation study and have m=1000 simulated datasets that each give me an estimate of the coefficent and its standard error. How would I determine if the statistical method properly characterizes the uncertainty of the estimated regression coefficient? Note your answer could be as simple as a sentence or two describing what quantities to consider. You can also consider Tables 1-3 in the Cao et al. (2015) paper and answer this question specifically in that setting.

3. The following calculation arises in solving a least squares regression problem where the coefficients are subject to an equality constraint, in particular, we want to minimize $(Y - X\beta)^\top(Y - X\beta)$ with respect to β subject to the m constraints $A\beta = b$ for an m by p matrix A. (Each row of A represents a constraint that that linear combination of β equals the corresponding element of b.)
Solving this problem is a form of optimization called quadratic programming. Some derivation using the Lagrange multiplier approach gives the following solution:

$$\hat{\beta} = C^{-1}d + C^{-1}A^\top(AC^{-1}A^\top)^{-1}(-AC^{-1}d + b)$$

where $C = X^\top X$ and $d = X^\top Y$. X is n by p.

(a) Describe how you would implement this in pseudo-code, taking account of the principles discussed in class in terms of matrix inverses and factorizations

(b) Write an R function to efficiently compute $\hat{\beta}$, taking account of the principles discussed in class in terms of matrix inverses and factorizations. Note: you can use any of R's matrix manipulation functions that you want - I'm not expecting you to code up any algorithms from scratch.
Note: in reality a very efficient solution is only important when the number of regression coefficients, p, is large.

4. Two-stage least squares (2SLS) is a way of implementing a causal inference method called instrumental variables that is commonly used in economics. Consider the following set of regression equations:

$$\begin{aligned}\hat{X} &= Z(Z^\top Z)^{-1}Z^\top X \\ \hat{\beta} &= (\hat{X}^\top\hat{X})^{-1}\hat{X}^\top y\end{aligned}$$

which can be interpreted as regressing y on X after filtering such that we only retain variation in X that is correlated with the instrumental variable Z. An economics graduate student asked how he could compute $\hat{\beta}$ if Z is 60 million by 630, X is 60 million by 600, and y is 60 million by 1, but both Z and X are sparse matrices.

(a) Describe briefly why I can't do this calculation in two steps as given in the equations, even if I use the techniques for OLS discussed in class for each stage.

(b) Figure out how to rewrite the equations such that you can actually calculate $\hat{\beta}$ on a computer without a huge amount of memory. You can assume that any matrix multiplications involving sparse matrices can be done on the computer (e.g., using the spam package in R). Describe the specific steps of how you would do this and/or write out in pseudo-code.

Notes: (1) The product of two sparse matrices is not (in general) sparse and would not be sparse in this case. (2) As discussed in Section 6.2 of Unit 9, there are R packages (and software packages more generally) for efficiently storing (to save memory) and efficiently doing matrix manipulations (to save computation time) with sparse matrices.

5. Details of the Cholesky decomposition presented in Unit 9. Work out the operation count (total number of multiplications plus divisions) for the Cholesky decomposition, including the constant c, not just the order, for terms involving n^3 or n^2 (e.g., $5n^3/2 + 8n^2$, not $O(n^3)$). You can ignore the square root and any additions/subtractions. You can ignore pivoting for the purpose of this problem. Remember not to count any steps that involve multiplying by 0 or 1. Compare your result to that given in the notes.

6. **(Extra credit)** In class we saw that the condition number when solving a system of equations, $Ax = b$, is the ratio of the largest and smallest magnitude eigenvalues of A. Show that $\|A\|_2$ (i.e., the matrix norm induced by the usual L2 vector norm; see Section 1.6 of Unit 9) is the largest of the absolute values of the eigenvalues of A for symmetric A. To do so, find the following quantity,

$$\|A\|_2 = \sup_{z:\|z\|_2=1} \sqrt{(Az)^\top Az}.$$

   If you're not familiar with the notion of the supremum (the *sup* here), just think of it as the maximum. It accounts for situations such as trying to find the maximum of the numbers in the open interval (0,1). The max is undefined in this case since there is always a number closer to 1 than any number you choose, but the *sup* in this case is 1.

   Hints: when you get to having the quantity $\Gamma^\top z$ for orthogonal Γ, set $y = \Gamma^\top z$ and show that if $\|z\|_2 = 1$ then $\|y\|_2 = 1$. Finally, if you have the quantity $y^\top Dy$, think about how this can be rewritten given the form of D and think intuitively about how to maximize it if $\|y\|_2 = 1$.

7. **(Extra credit)** In Unit 9 we discussed that having a 0 as an eigenvalue of a covariance matrix amounts to having a constraint (one of the eigenvectors has zero weight). In Section 2.4, I say a bit about having a zero eigenvalue of a precision matrix, where a precision matrix is the inverse of the covariance matrix.

   (a) What is the relationship between the eigenvalues of a covariance matrix, Σ, and the eigenvalues of the corresponding precision matrix, Σ^{-1}?

   (b) Consider the following autoregressive style model:

$$y_i \sim N(y_{i-1}, \sigma^2)$$

   The likelihood (joint distribution for $(y_1, \ldots, y_n)$) is

$$\prod_{i=2}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(y_i - y_{i-1})^2\right)$$

   which gives us the sum of squares:

$$\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - y_{i-1})^2$$

   We can equivalently represent the model as

$$(y_1, \ldots, y_n) = Y \sim N(0, \sigma^2\Sigma)$$

where $Q = \Sigma^{-1}$ is the precision matrix and looks like this (for the specific case of $n = 5$):

$$Q = \begin{pmatrix} 1 & -1 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 \\ 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & -1 & 1 \end{pmatrix}$$

Show that with the joint distribution based on the multivariate representation, $Y \sim N(0, \sigma^2 Q^{-1})$, you get the same sum of squares as above.

(c) Show that if you add a constant value to every y_i, the sum of squares is unchanged. This makes sense because we can see the sum of squares as involving contrasts (differences) of adjacent data values. The interpretation is that this model says nothing about the overall level of the y_i values, only about their contrasts.

(d) For the $n = 100$ case, create Q and find the eigendecomposition in R and show that one of the eigenvalues is 0 and that the corresponding eigenvector is a vector with each value equal to $1/\sqrt{n}$.

(e) To generate a random vector $Y = \Gamma\Lambda^{1/2}z$, where Γ and Λ have the eigenvectors and eigenvalues of Σ, we would need to invert Q. But we can't do that because an eigenvalue is 0. Instead, we would use the pseudo-inverse described at the start of Section 2.4. For the case of $n = 100$ carry out this algorithm and generate and plot a small number of random vectors Y from this autoregressive model. You should see that the mean of each vector Y is 0, which is consistent with having imposed a constraint by using the pseudo-inverse.