Homework Problems

Econ 722

Spring 2019

Instructions

This document contains problems to accompany the lecture material in Econ 722. The first nine problems are theoretical, while the last five are computational. As part of your grade for the course, you must submit solutions *eight* of these problems: five theoretical and three computational. Subject to this constraint, you may choose any problems that you prefer. Your full set of solutions must be type-written, and submitted in pdf form to canvas by 11:59pm on May 1st, 2019. You will be graded both on the completeness and correctness of your answers, and the clarity of your writing and exposition. Along with your pdf write-up, make sure to submit all the source code required to replicate your solutions to the computational exercises. This code *must be written in an open-source programming language*. R, Python, Julia, and octave are all permitted but Matlab is *not*. During the course of the semester, you will also present one of your solutions to the class during lecture. This presentation will be graded separately from the final write-up of homework problems that you submit at the end of the semester. See the syllabus for details.

Theoretical Problems

1. Let Θ be a discrete set and π be a prior distribution that gives strictly positive probability to each element of Θ. Show that if $\widehat{\theta}$ is a Bayes rule with respect to π, it is admissible.

2. Derive the KL divergence from a $N(\mu_0, \sigma_0^2)$ distribution to a $N(\mu_1, \sigma_1^2)$ distribution.

3. Suppose we observe a random sample $\{(\mathbf{x}_t, y_t)\}_{t=1}^T$ from some population and decide to predict y from $\mathbf{x}$ using the following linear model:

$$y_t = \mathbf{x}_t'\beta + \varepsilon_t$$

Let $\widehat{\beta}$ denote the ordinary least squares estimator of β based on $\{(\mathbf{x}_t, y_t)\}_{t=1}^T$. Now suppose that we observe a *second* random sample $\{(\widetilde{\mathbf{x}}_t, \widetilde{y}_t)\}_{t=1}^T$ from the sample population that is *independent* of the first. Show that

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}(y_t - \mathbf{x}_t'\widehat{\beta})^2\right] \leq \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}(\widetilde{y}_t - \widetilde{\mathbf{x}}_t'\widehat{\beta})^2\right]$$

In other words, show that the in-sample squared prediction error is an overly optimistic estimator of the out-of-sample squared prediction error.

4. In this question you'll derive a computational shortcut for leave-one-out cross-validation in the special case of least-squares estimation. (The same basic idea holds for any linear smoother.) Let $\widehat{\beta}$ be the full-sample least squares estimator, and $\widehat{\beta}_{(t)}$ be the estimator that leaves out observation t. Similarly, let $\widehat{y}_t = \mathbf{x}_t'\widehat{\beta}$ and $\widehat{y}_{(t)} = \mathbf{x}_t'\widehat{\beta}_{(t)}$.

   (a) Let X be a $T \times p$ design matrix with full column rank, and define

   $$A = X'X = \sum_{t=1}^{T}\mathbf{x}_t\mathbf{x}_t' = \mathbf{x}_t\mathbf{x}_t' + \sum_{k\neq t}\mathbf{x}_k\mathbf{x}_k' = A_{(t)} + \mathbf{x}_t\mathbf{x}_t'$$

   Show that

   $$A^{-1} = A_{(t)}^{-1} - \frac{A_t^{-1}\mathbf{x}_t\mathbf{x}_t'A_{(t)}^{-1}}{1 + \mathbf{x}_t'A_{(t)}^{-1}\mathbf{x}_t}$$

   where you may assume that $A_{(t)}$ is also of rank p.

   (b) Let $\{h_1, \ldots, h_T\} = diag\{\mathbf{I}_T - X(X'X)^{-1}X'\}$. Show that

   $$h_t = 1 - \mathbf{x}_t'A^{-1}\mathbf{x}_t = \frac{1}{1 + \mathbf{x}_t'A_{(t)}^{-1}\mathbf{x}_t}$$

   (c) Let $\mathbf{w} = \sum_{k\neq t}\mathbf{x}_k y_k$. Now, note that we can write $\widehat{\beta} = (A_{(t)} + \mathbf{x}_t\mathbf{x}_t')^{-1}(\mathbf{w} + \mathbf{x}_t y_t)$ and $\mathbf{x}_t'\widehat{\beta}_{(t)} = \mathbf{x}_t'A_{(t)}^{-1}\mathbf{w}$. Use these facts along with the results you proved in the preceding parts to show that $(y_t - \widehat{y}_{(t)}) = (y_t - \widehat{y}_t)/h_t$.

   (d) Suppose that we wanted to carry out leave-one-out cross-validation under squared error loss:

   $$CV_1 = \frac{1}{T}\sum_{t=1}^{T}(y_t - \widehat{y}_{(t)})^2$$

   In light of the preceding parts, explain how we could carry out this calculation *without* explicitly calculating $\widehat{\beta}_{(t)}$ for each observation t.

5. This question asks you to derive some simple results for concerning influence functions.

   (a) A functional that takes the form $\mathbf{T}(G) = \int_{-\infty}^{\infty} u(z)\, \mathrm{d}G(z)$ for some function u is called a *linear functional*. Derive the influence function of a linear functional.

   (b) The mean μ of a distribution G can be expressed as a linear functional. Using part (a), show that the influence function of the mean equals $y - \mu$.

   (c) Let $\mathbf{T}$ be a $\mathbb{R}$-valued functional that depends on two *other* $\mathbb{R}$-valued functionals $\mathbf{T}_1$ and $\mathbf{T}_2$ according to $\mathbf{T}(G) = h\left(\mathbf{T}_1(G), \mathbf{T}_2(G)\right)$ where h is a continuously differentiable function from $\mathbb{R}^2$ to $\mathbb{R}$. Derive an expression for the influence function $\psi(G, y)$ of $\mathbf{T}$ in terms of h and the influence functions $\psi_1(G, y), \psi_2(G, y)$ of $\mathbf{T}_1, \mathbf{T}_2$. Hint: the influence function is defined as a limit but is equivalent to a partial derivative.

   (d) Use parts (a)–(c) to show that the influence function of the *variance* σ^2 of a distribution equals $(y - \mu)^2 - \sigma^2$.

6. This question asks you to fill in some of the missing details from the example comparing AIC and BIC in Lecture #4. Suppose that $Y_1, \ldots, Y_T \sim$ iid $N(\mu, 1)$. Let $\ell_T(\mu)$ denote the sample log-likelihood function evaluated at μ, where $\mathrm{Var}(Y_i) = 1$ is assumed known.

   (a) Show that $\sum_{t=1}^{T}(Y_t - \mu)^2 = T(\bar{Y} - \mu)^2 + T\widehat{\sigma}^2$ and use this result to establish that $\ell_T(\mu) = \text{Constant} - \dfrac{T}{2}(\bar{Y} - \mu)^2$.

   (b) Suppose that g is a $N(\mu, 1)$ density while h is a $N(0, 1)$ density. Show that $KL(g; h) = \mu^2/2$. (If you like, you can simply apply the formula from Problem Set #1, although a direct argument is very simple.)

   (c) Let $Z \sim N(0, 1)$, and $X = \mathbf{1}\{A\}$ where $A = \left\{|\sqrt{T}\mu + Z| \geq \sqrt{d_T}\right\}$. Show that

$$\mathbb{E}\left\{\left[\left(\sqrt{T}\mu + Z\right)X - \sqrt{T}\mu\right]^2\right\} = \mathbb{P}(A)\,\mathbb{E}\left[Z^2 | X = 1\right] + \left[1 - \mathbb{P}(A)\right]T\mu^2$$

   (d) Continuing from the preceding part, argue that the conditional density of Z given $X = 1$ is $\mathbf{1}(A)\varphi(z)/\mathbb{P}(A)$. Using this, along with $\mathbb{E}[Z^2] = 1$, show that

$$\mathbb{P}(A)\mathbb{E}[Z^2 | X = 1] = 1 - \int_a^b z^2 \varphi(z)\, dz$$

   (e) Continuing from the preceding part, show that

$$\int_a^b z^2 \varphi(z)\, dz = a\varphi(a) - b\varphi(b) + \Phi(b) - \Phi(a)$$

(f) Combine the three preceding parts and calculate $\mathbb{P}(A)$ to show that

$$R(\mu, \widehat{\mu}) = 1 + [b\phi(b) - a\phi(a)] + (T\mu^2 - 1)[\Phi(b) - \Phi(a)]$$

where $a = -\sqrt{d_T} - \sqrt{T}\mu$ and $b = \sqrt{d_T} - \sqrt{T}\mu$.

7. This question concerns the so-called "GMM-AIC" moment selection criterion from Andrews (1999) which takes the form

$$\text{GMM-AIC}(c) = J_T(c) - 2(|c| - p)$$

where $J_T(c)$ denotes the J-test statistic for the estimator based on the collection of moment restrictions indexed by c, $|c|$ denotes the number of moment conditions in specification c and p denotes the number of parameters. Characterize the asymptotic behavior of this criterion under the assumptions of the consistency theorem we proved in class. Hint: there are two cases, which parallel our proof from class.

8. The FMSC of DiTraglia (2016) is a moment selection criterion constructed by deriving the asymptotic MSE of an estimator of some "target parameter" μ, under local misspecification. A very similar idea can also be used for *model selection* in maximum likelihood models: this is the so-called "Focused Information Criterion" of Claeskens & Hjort (2003). In this question you will derive the simplest possible example of the FIC. This will require you to "get your hands dirty" with local asymptotics, so you may want to read the beginning of Chapter 4 from the lecture notes before attempting this problem. Consider a linear regression model with two regressors x and z

$$y_t = \theta x_t + \gamma z_t + \epsilon_t$$

where $\{(x_t, z_t, \epsilon_t)\}_{t=1}^T \sim$ iid with means $(0, 0, 0)$ and variances $(\sigma_x^2, \sigma_z^2, \sigma_\epsilon^2)$. For simplicity, assume the errors are homoskedastic. Our goal is to estimate θ with minimum MSE, and the model selection decision is whether or not to include z in the regression. Consider two estimators of θ: the "long" regression estimator $\widehat{\theta}$ calculated from $(\widehat{\theta}, \widehat{\gamma})' = \left\{[\mathbf{x}, \mathbf{z}]'[\mathbf{x}, \mathbf{z}]\right\}^{-1}[\mathbf{x}, \mathbf{z}]'\mathbf{y}$ and the "short" regression estimator $\widetilde{\theta} = (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{y}$. Since all random variables are mean zero, you do not have to include a constant.

(a) Suppose that γ is local to zero, in other words $\gamma = \delta/\sqrt{T}$. Under this assumption, derive the asymptotic distributions of $\sqrt{T}(\widetilde{\theta} - \theta)$ and

$$\sqrt{T}\begin{bmatrix} \widehat{\theta} - \theta \\ \widehat{\gamma} - 0 \end{bmatrix}.$$

Note that the limit distribution of $\widehat{\gamma}$ is centered around zero since $\delta/\sqrt{T} \to 0$ as $T \to \infty$. You should find that $\widetilde{\theta}$ has an asymptotic bias that depends on δ.

(b) Under what conditions does $\widetilde{\theta}$ have a lower AMSE than $\widehat{\theta}$? Note that your answer should depend on δ. Explain the intuition for your result.

(c) Propose an asymptotically unbiased estimator of δ constructed from $\sqrt{T}\widehat{\gamma}$.

(d) Combine steps (a) and (c) to propose asymptotically unbiased estimators of the AMSE of $\widehat{\theta}$ and $\widetilde{\theta}$.

(e) The FIC chooses the estimator with the lower estimated AMSE from step (d). How does this rule compare to AIC, BIC, Mallow's C_p, and a t-test of the null hypothesis $H: \gamma = 0$ at the $\alpha \times 100\%$ level? Comment briefly on any relationships you uncover.

9. Consider a regression model of the form $y_t = \mathbf{x}_t'\beta + \epsilon_t$ where $\mathbf{x}_t$ is $(p \times 1)$ and satisfies $T^{-1}\sum_{t=1}^{T} \mathbf{x}_t\mathbf{x}_t' = \mathbf{I}_p$ and $\epsilon_t \overset{iid}{\sim} N(0, \sigma^2)$ where σ^2 is finite. You may treat the regressors as *fixed* rather than random in your derivations.

(a) What is the MLE for β in this setting? Derive its finite-sample distribution. Is the MLE consistent for β?

(b) Derive a closed-form expression for the Ridge Regression estimator of β in this setting, expressed in terms of the MLE and the shrinkage parameter λ. Use this result to write out the finite-sample distribution of the Ridge Estimator.

Important Note: It will be helpful to factor a T from λ in your derivation. When you do this, you can still call the re-scaled shrinkage parameter λ rather than λ/T. Remember: we are free to choose λ so its precise scaling is irrelevant. Writing things this way will make your expression for the Ridge estimator match the (slightly different) example from the lecture notes and should help to avoid confusion in the parts that follow below.

(c) First, suppose that we choose a fixed positive value for λ. Explain why the corresponding Ridge estimator will *not* be consistent for β as $T \to \infty$ in this case. Next suppose that, instead of fixing λ we decide to allow it to change with sample size. State sufficient conditions on the sequence λ_T to ensure that the Ridge estimator is consistent.

(d) Derive a closed-form expression for the LASSO estimator in this example, expressed in terms of the MLE and the shrinkage parameter λ.

Important Note: As in the Ridge example above, and for the same reason, if you encounter the term $T\lambda$ you can forget about the T and just call this λ.

(e) For sufficiently large λ, LASSO shrinks some coefficients all the way to zero and hence can be used to carry out variable selection. Derive an exact finite sample expression for the probability that LASSO decides to "exclude" regressor j, that is $P(\widehat{\beta}_j^{Lasso} = 0)$. Explain the intuition with the help of one or more plots.

(f) Now suppose we allow the LASSO shrinkage parameter to depend on sample size. Prove that $\lim_{T\to\infty} \lambda_T = 0$ implies that the probability of LASSO excluding a relevant regressor, i.e. one with a non-zero coefficient, converges to zero.

(g) Now consider the case of an *irrelevant regressor*, i.e. $\beta_j = 0$. What is the probability that LASSO excludes such a regressor? If we allow λ_T to depend on sample size, what condition on the *rate* at which $\lambda_T \to 0$ is required to ensure that LASSO excludes irrelevant regressors with probability approaching one in the limit? What happens if $\lambda_T \to 0$ at a slower rate?

(h) Combining the two preceding parts gives a *consistency* result for LASSO: provided that λ_T converges to zero at a sufficiently fast rate, all relevent regressors are selected and all irrelevant regressors excluded with probability approaching one in the limit. Crucially, however, this result depended on β_j being *fixed*. Suppose instead that we consider a sequence of local parameter values $\beta_{j,T} = \delta/\sqrt{T}$ where δ is a constant. When $\delta \neq 0$, this captures in asymptotic form the idea of "small but nonzero" coefficients. How do the results of the preceding parts change under these asymptotics? Discuss your findings.

Computational Problems

10. Let $X \sim N(\theta, 1)$ where $\theta \in \Theta = [-m, m]$ $m > 0$. Further define:

$$\pi(\theta) = \begin{cases} 1/2, & \theta = -m \\ 0, & -m < \theta < m \\ 1/2, & \theta = m \end{cases}$$

(a) Show that $\widehat{\theta}(X) = m\tanh(mX)$ is the Bayes rule with respect to π under squared error loss. Recall that $\tanh(z) = (e^z - e^{-z})/(e^z + e^{-z})$.

(b) Write code to calculate the risk function $R(\theta, \widehat{\theta})$ and Bayes risk $r(\pi, \widehat{\theta})$ numerically. I suggest numerical integration rather than a simulation-based approximation.

(c) For each value of $m \in \{0.5, 0.75, 1, 1.25\}$ plot the following:

(i) X vs. $\widehat{\theta}(X)$ for $X \in [-3, 3]$, with the 45-degree line indicated in red.

(ii) θ vs. $R(\theta, \widehat{\theta})$ for $\theta \in [-m, m]$ with $r(\pi, \widehat{\theta})$ indicated as a red horizontal line.

(d) Explain your findings from part (c) above. How does $\widehat{\theta}$ compare to the MLE? For which, if any, of the values of $m \in \{0.5, 0.75, 1, 1.25\}$ is $\widehat{\theta}$ minimax?

11. Let $X \sim N(\theta, I)$ where θ is a p-vector for $p \geq 3$ and I is the $(p \times p)$ identity matrix. For this problem, we showed in class that the maximum likelihood estimator for θ, $\widehat{\theta} = X$ is inadmissible, as it is dominated by the James-Stein estimator:

$$\widehat{\theta}^{JS} = \widehat{\theta}\left(1 - \frac{p-2}{\widehat{\theta}'\widehat{\theta}}\right)$$

I also argued, without proof, that the James-Stein estimator is itself inadmissible, as it is dominated by the so-called "positive-part" James-Stein estimator, namely

$$\widetilde{\theta}^{JS} = \widehat{\theta}\left[\max\left\{1 - \frac{p-2}{\widehat{\theta}'\widehat{\theta}}, 0\right\}\right]$$

This estimator takes its name from the fact that, unlike the plain-vanilla James-Stein estimator, it can never shrink "past" zero and hence cannot have a different sign than the MLE. Design and carry out a simulation experiment comparing the risk of $\widehat{\theta}$, $\widehat{\theta}^{JS}$, and $\widetilde{\theta}^{JS}$ under squared error loss. Your results should be based on 10,000 simulation replications over a range of values for $p \geq 3$ and different configurations of the true mean vector θ. Write a brief summary of your results, accompanied by tables and or figures, as needed.

12. Consider a collection of AR(p) models for $p = 1, 2, \ldots, 6$. In this question you will choose the lag order p using AIC, BIC, and cross-validation under two different true data generating processes:

$$\begin{aligned} \text{DGP1:} \quad & y_t = 0.7y_{t-1} + \varepsilon_t \\ \text{DGP2:} \quad & z_t = \varepsilon_t + 0.6\varepsilon_{t-1} \end{aligned}$$

where $\varepsilon_t \sim$ iid N$(0, 1)$ for $t = 1, \ldots, T$ and $T = 100$. Note that DGP1 is among the candidate AR(p) specifications under consideration while DGP2 is not. To answer this

question, you will need to consult some papers from the shared Dropbox folder for the course: Burman, Chow & Nolan (1994); Racine (2000), Ng & Perron (2005); and Bergmeir, Hyndman & Koo (2015). In all of calculations below, carry out estimation via least-squares (the conditional maximum likelihood estimator). Note that when you estimate an AR model in this fashion, you will need to drop the first p observations in your sample, meaning that the different AR models will be use different sample sizes.

(a) Carry out a simulation study to calculate the one-step-ahead predictive MSE of each of the six AR specifications under both data generating processes. Briefly discuss your findings. In particular, you will need to carry out the following steps:

(i) Generate $\varepsilon_0, \varepsilon_1, \ldots, \varepsilon_{100}, \varepsilon_{101} \sim$ iid $N(0,1)$.

(ii) Set $y_0 = 0$ and $y_t = 0.7y_{t-1} + \varepsilon_t$ for $t = 1, 2, \ldots, 100, 101$.

(iii) Set $z_t = \varepsilon_t + 0.6\varepsilon_{t-1}$ for $t = 1, 2, \ldots, 100, 101$. (For z_1 you will need to use ε_0.)

(iv) Fit AR(p) models for $p = 1, 2, \ldots, 6$ via conditional maximum likelihood to $\{y_1, \ldots, y_{100}\}$. For each lag-length p construct an out-of-sample forecast $\widehat{y}_{101}(p)$ of y_{101} based on your fitted model.

(v) Do the same as in the preceding step for z: fit AR(p) models for $p = 1, \ldots, 6$ using $\{z_1, \ldots, z_{100}\}$ and construct an out-of-sample forecast $\widehat{z}_{101}(p)$ of z_{101}.

(vi) For each AR(p) model and each DGP, calculate the squared forecast error: $[y_{101} - \widehat{y}_{101}(p)]^2$ and $[z_{101} - \widehat{z}_{101}(p)]^2$.

(vii) Repeat the above steps 10,000 times, storing the squared forecast errors for each DGP and AR lag length in each replication. Use the sample mean of the squared forecast errors across replications to approximate one-step-ahead sample predictive MSE.

(b) Based on the discussion in Ng and Perron (2005), what are the complications in defining AIC and BIC for AR(p) models? On the basis of their simulation results, what formulas do you suggest using for AIC and BIC in this setting?

(c) Based on Burman, Chow & Nolan (1994); Racine (2000); and Bergmeir, Hyndman & Koo (2015) what are the complications in applying cross-validation to AR(p) models? How do you suggest using cross-validation to select the AR lag order?

(d) Given your choices in parts (b) and (c), carry out a simulation study with 10,000 replications comparing AIC, BIC and cross-validation under each of the two DGPs.

For each DGP, calculate the fraction of replications in which a particular criterion (AIC, BIC, or Cross-Validation) selects each lag order. Briefly discuss your findings.

13. This question and the one that follows it are based on Stock and Watson (JBES, 2012) "Generalized Shrinkage Methods for Forecasting Using Many Predictors." You can download the paper, a supplemental appendix, data and replication files from Mark Watson's webpage at `https://www.princeton.edu/~mwatson/publi.html`

    (a) Read the Stock and Watson (2012) paper. Provide a brief (one or two paragraph) summary of the main findings in the paper.

    (b) Familiarize yourself with the data set. The transformed series which Stock and Watson use in their estimation are posted at `http://ditraglia.com/econ722/SW2012data.csv`. Take a look at Section B of the Supplemental Appendix to understand how the raw data (you can find them in the replication zip file on Mark's webpage) are transformed into the data that I have posted. Compare the GDP growth series in the Stock and Watson data set to a GDP growth series that you construct from FRED data.

    (c) Compute the first 20 principal components from 108 lower-level disaggregates.[1] How much variation in each of the 108 series is explained by the first or by the first two principal components? Think carefully about how to present this information in an efficient manner.

14. In this question you will construct diffusion index forecasts of real GDP (GDP 251), consumption (GDP252), and real government consumption expenditures (GDP265) following Stock & Watson (2012). The data you will need for this exercise are posted at `http://ditraglia.com/econ722/SW2012data.csv`. For each series and each part of this question you will construct one-step-ahead, pseudo-out-of-sample forecasts based on a rolling window of the most recent 100 observations and use RMSE to compare the different forecasting methods. (The procedure is detailed in section 3.1 of the paper.) Note that you will *not* use cross-validation in this question.

    (a) First try to replicate Stock and Watson's (2012) one-step-ahead RMSE results for the AR(4) model and the OLS model, both relative to the DFM5. These appear in

[1] Although the paper gives the total number of disaggregates as 109, there in fact only 108 in the replication dataset. This agrees with the count based on table B.1 of the Supplementary Appendix.

Table S-8 in the online supplemental appendix and are described in section 4.1 of the paper.

(b) An AR(4) model may be somewhat too complicated to serve as a reasonable benchmark for the DFM5. Augment your results from the preceding part by adding an AR(1) model as well as two model-selection based AR forecasting procedures: one using AIC and another using BIC. How do the results compare? Do they differ across the three series?

(c) Although Stock & Watson (2012) explore a number of shrinkage estimators in their paper, Ridge and Lasso are not among them. Try using Ridge and Lasso rather than OLS for the forecasting regression based on the first 50 PCs. Remember: since the design is orthogonal, there is a simple closed form for both Lasso and Ridge. Experiment with a variety of values for the shrinkage parameters. Try to find values that give similar performance to the DFM5. Is there a level of shrinkage that beats the best AR benchmark forecast? How do your results compare across the series?