Practice Problems (200 points total)

Pepe Olea and Nathaniel Mark

May 6, 2019

1 Very Short Answer Questions (15 Points - 5 each)

In these questions, please write a 1-2 sentence answer to the prompt or check the correct box.

Question 1.1 What is the difference between a parametric model and a non-parametric model? Be precise.

Question 1.2 Say that $p > \tilde{p}$ What is largest?

☐ $ln(SSR(p)/T)$

☐ $ln(SSR(\tilde{p})/T)$

☐ BIC(p)

☐ Not enough information to tell.

Question 1.3 Assume we are in a regression framework and I write "Define Q_{95} as the 95th quantile of the distribution of $\hat{\beta}_2$. Our parametric bootstrap estimate of Q_{95} converges to Q_{95} as $L \to \infty$ and $N \to \infty$."
If N is the sample size, what is L?

2 Short Answer Questions (62.5 Points - 12.5 each)

In these questions, please write a 3 - 6 sentence answer to the prompt or follow its directions.

Question 2.1 What is the primary purpose of the parametric bootstrap procedure? When would we use it?

Question 2.2 Take the "Dumb Information Criterion" (DIC):

$$DIC(p) = ln(\frac{SSR(p)}{T}) + ln(p)$$

prove that $\hat{p} = argmin_p DIC(p)$ is not consistent.

Hint: An estimator, $\hat{p}$, is inconsistent if $P(\hat{p} = p*)$ converges to 0 for any set of data generating process in our model. This is because the definition of convergence states that $\hat{p}$ is consistent if $P(\hat{p} = p*)$ converges to 1 for all true parameters (data generating processes) in the model.

Question 2.3 Take the model

$$X_i = \mu + \epsilon_i$$

$$\epsilon_i \sim N(0,1) i.i.d.$$

with parameter $\mu \in \mathbb{R}$ and loss function

$$L(a, \mu) = (a - \mu)^2$$

You observe a dataset $\{X_1, ..., X_n\}$ Answer the following questions:

a) Is the decision rule $d^1(x) = \frac{1}{N} \sum X_i - \frac{1}{2}$ an admissible decision rule? Why or why not?

b) Is the decision rule $d^2(x) = 12$ an admissible decision rule? Why or why not?

c) Is the decision rule $d^3(x) = Random(\{X_1, ..., X_n\})$ an admissible decision rule? Why or why not?

$Random(\{X_1, ..., X_n\})$ is defined as a function that takes a random value (with uniform probabilities) from the dataset and defines that value as the action.

Question 2.4 If the posterior distribution of μ given x in a bayesian framework is

$$N(\frac{\sigma^2}{\sigma^2 + n}m + \frac{n}{\sigma^2 + n}\bar{X}, \frac{\sigma^2}{\sigma^2 + n})$$

a) What do you think m is likely to be?

b) What is the Bayesian estimator of μ if the loss function is $L(a, \mu) = (a - \mu)^2$?

c) What is the Bayesian estimator of μ if the loss function is $L(a, \mu) = (a - \mu)^2 - 3\theta$?
where θ is a constant.

Question 2.5 Take the AR(p) model

$$X_t = \phi X_{t-1} + \epsilon_t$$

$$\epsilon_t \sim N(0, \sigma^2)$$

Define $Z_t = X_{t-1}\epsilon_t$.
Show that $E[Z_t Z_{t-1}] = 0$

Long Answer Questions

In these problems,

1. Fill in the blanks.
2. Justify your answers in a separate sheet (imagine you have blue books).

Please read each of the questions very carefully and provide clean concise answers.

NOTE: If for some reason filling in the blanks seems to confusing for you, just write your own answer in a separate piece of paper.

3 Maximum Likelihood Estimation of σ^2 (25 Points)

In class we derived the Likelihood Function of the Gaussian AR(1) model:

$$y_t = \phi y_{t-1} + \epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0, \sigma^2), \quad \text{i.i.d..} \tag{1}$$

Define the Y_1-conditional likelihood of $(Y_2, \ldots, Y_T)$ as:

$$f(Y_2, \ldots, Y_T | Y_1; \phi, \sigma^2) \tag{2}$$

Subquestion 1 (12.5 points) Write down the Y_1 - conditional Likelihood Function for the Gaussian AR(1) model.

Answer: Using the definition of conditional density we have for any $t > 2$:

$$f(Y_2, \ldots, Y_t | Y_1; \phi, \sigma^2) = \boxed{3.1)\ f\Big(\quad \Big| \qquad\qquad ; \phi, \sigma^2 \Big)} f\Big(Y_2, \ldots, Y_{t-1} \Big| Y_1; \phi, \sigma^2\Big).$$

This implies that (after substituting recursively) we can write (2) as:

$$\Pi_{t=2}^{T} \boxed{3.2)\ f\Big(\quad \Big| \qquad\qquad ; \phi, \sigma^2 \Big)}$$

We also know that in the AR(1) model: $Y_t | Y_1, Y_2, \ldots, Y_{t-1} \sim \boxed{3.3)\ \mathcal{N}\Big(\qquad , \qquad \Big)}$.

Therefore, using the definition of the normal p.d.f.:

$$f(Y_2, \ldots, Y_T | Y_1; \phi, \sigma^2) = \frac{1}{\boxed{\Big(3.4) \qquad \Big)}^{T-1}} \exp \boxed{\Big(3.5) \qquad\qquad \Big)}$$

Subquestion 2 (12.5 points) Define the Conditional Maximum Likelihood (CML) estimator of (ϕ, σ^2) as the value that maximizes the likelihood associated to (2). We have already shown that:

$$\widehat{\phi}_{\text{CML}} = \widehat{\phi}_{\text{OLS}} = \sum_{t=2}^{T} Y_t Y_{t-1} \Big/ \sum_{t=2}^{T} Y_{t-1}^2. \tag{3}$$

Derive the conditional maximum likelihood estimator of σ^2 for the AR(1) model (HINT: Take the derivative of the conditional log-likelihood function with respect to σ^2 and set it equal to zero). How does this estimator compare to the OLS estimator for the variance?

Answer: The conditional log-likelihood function $\ln f(Y_2, \ldots, Y_T | Y_1; \phi, \sigma^2)$ equals:

$$-\frac{T-1}{2}\ln(2\pi) - \boxed{\left(\text{3.6)} \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \right)}$$

The first-order conditions with respect to σ^2 give:

$$\boxed{\left(\text{3.7)} \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \right)} = 0$$

We have already shown that $\widehat{\phi}_{\text{CML}}$ is given by (3). Therefore, $\widehat{\sigma}^2_{\text{CML}}$ should satisfy:

$$-\frac{T-1}{2}\frac{1}{\widehat{\sigma}^2_{\text{CML}}} + \boxed{\left(\text{3.8)} \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \right)} = 0$$

This implies that:

$$\widehat{\sigma}^2_{\text{CML}} = \frac{1}{T-1}\boxed{\left(\text{3.9)} \qquad\qquad\qquad\qquad\qquad \right)} = \boxed{\left(\qquad\qquad \right)}\widehat{\sigma}^2_{\text{OLS}}.$$

4 Parametric Bootstrap (32.5 points)

Let $\widehat{\phi}_{\text{CML}}$ and $\widehat{\sigma}^2_{\text{CML}}$ denote the values of ϕ and σ^2 that maximize the conditional maximum likelihood in (2). In this question I want you to use the parametric bootstrap to approximate the distribution of these estimators. We will do this in two steps.

Subquestion 1 (12.5 points): Write a python function called `ar1draws` that takes as inputs the values of $(\phi, \sigma, Y_1, T, I)$ with output equal to I draws of the time series data:

$$(Y_2, Y_3, \ldots, Y_T)$$

distributed according to $f(Y_2, \ldots, Y_T | Y_1; \phi, \sigma^2)$. The output should be a matrix of dimension $I \times T$ with each row containing a draw from $(Y_1, Y_2, \ldots, Y_T)$.

Answer: One possible solution is as follows:

```
import numpy as np

def ar1draws(phi, sigmasq, Yinit = 0, T, I):

    e = [ (4.1)            ]   % I times T draws from N(0, sigmasq)

    Y = np.zeros([I, T]);       %Initialize the values of Y

    Y[:, 0] = [ (4.2)            ];      % Initial Condition

for t in range(1, T):

          [4.3)      ] = phi * [4.4)      ] + e[:,t]
return Y;
```

Subquestion 2 (12.5 points): Suppose that you have a function that computes $\widehat{\phi}_{\text{CML}}$ and $\widehat{\sigma}^2_{\text{CML}}$. Let this function be called `CML` and suppose that it takes as input a given data set:

$$y = [Y_1, \ldots, Y_T]$$

To be more specific, assume that that the syntax of the function is of the form:

```
phiCMLd, sigma2CML = CML(y)
```

Write a python function `ParamBootsAR1` that takes as inputs $(\widehat{\phi}_{\text{CML}}, \widehat{\sigma}_{\text{ML}}, Y_1, T)$ and generates B bootstrap draws of $\widehat{\phi}_{\text{CML}}$ and $\widehat{\sigma}_{\text{ML}}$.

Answer: One possible solution is as follows:

```
def ParamBootsAR(phihat, sigmasqhat, Y1 = 0, To, B):

    Y = ar1draws( 4.5) , 4.6) , Yinit = Y1, T
= To, 2.7) )

    phis = np.zeros([B, 1])

    sigma2s = np.zeros([B, 1])

    for b in range(B):

        phi, sigma = CML(Y[b,:])

        phis[b] = phi

        sigma2s[b] = sigma

    return phis, sigma2s;
```

Subquestion 3 (7.5 points):

Explain how you would calculate a 95% confidence interval for $\widehat{\phi}_{\text{CML}}$ using the output of the ParamBootsAR function in 3 - 6 sentences.

5 Bayesian Model Selection (15 points)

In this question, consider the case where we have assumed a statistical model

$$X_i = \epsilon_i$$

$$\epsilon_i \sim N(\mu, 1)$$

There are two possible model types:

$$M_1 : \mu = 0$$

$$M_2 : \mu \in [0, 2]$$

and prior over parameters and model types:

$$\pi(\mu, M) = \pi(\mu|M)\pi(M)$$

$$\pi(M) = .5 \text{ for all M}$$

$$\pi(\mu|M_2) = .5 \text{ for all } \mu \text{ between 0 and 2}$$

That is, the prior is uniformly distributed from 0 to 2.

The *posterior odds* of the two models M_1 and M_2 is defined as

$$\text{Posterior Odds} = \frac{\pi(M_1|x)}{\pi(M_2|x)}$$

In the next three subquestions, we derive the posterior odds.

Subquestion 1 (Fill-in the blank)

We know that the posterior probability of model M is given by:

$$\pi(M|x) = \frac{f(x|M)\boxed{5.1)\quad\quad}}{f^*(x)}$$

where $f(x|M)$ can be derived from integrating a function of our statistical model and priors. That is,

$$f(x|M) = \int \boxed{5.2)\quad\quad\quad\quad\quad\quad\quad\quad} d\mu$$

Subquestion 2 (Proof)
Prove that

$$\text{Posterior Odds} = \frac{f(x|M_1)}{f(x|M_2)}$$

Subquestion 2 (Fill-in the blank)
In this section, we derive $f(x|M_1)$ and $f(x|M_2)$
For model one:

$$f(x|M_1) = \prod_i^n \phi(X_i)$$

where $\phi(x)$ is the pdf of the standard normal distribution. For model two:

$$f(x|M_2) = \int \boxed{5.3)\quad\quad\quad\quad} d\mu$$

where $\phi(x)$ is the pdf of the standard normal distribution.

6 Model Selection for AR(p) models (25 points)

Consider the function

$$AIC(p) \equiv \ln\left(\frac{1}{T}\sum_{t=p+1}^{T}(y_t - \widehat{\mu} - \sum_{j=1}^{p}\widehat{\phi}_j y_{t-j})^2\right) + (p+1)\frac{2}{T}$$

The number of lags selected by the *Akaike Information Criterion*, $\widehat{p}$, minimize AIC(p) for values $p \in \{0, 1, 2 \ldots \bar{p}\}$. Take $0 < p^* < \bar{p}$. We can show (using the same arguments we used in lecture) that:

$$\lim_{T\to\infty} \mathbb{P}_{p^*}(\widehat{p} < p^*) = 0. \tag{4}$$

Taking equation (4) as given, is it true that:

$$\lim_{T\to\infty} \mathbb{P}_{p^*}(\widehat{p} > p^*) > 0?$$

(HINT: Use the fact that for any $c > 0$, $0 < \lim_{T\to\infty} \mathbb{P}_{p^*}(\text{Wald}(p^*) < c) < 1$.)
Answer: Suppose not. Then $\lim_{T\to\infty} \mathbb{P}_{p^*}(\widehat{p} > p^*) = 0$. Since

$$\mathbb{P}_{p^*}(\widehat{p} < p^*) + \boxed{\left(\text{4.1)} \qquad\qquad \right)} + \mathbb{P}_{p^*}(\widehat{p} > p^*) = 1,$$

then we should have:

$$\lim_{T\to\infty} \mathbb{P}_{p^*}(\widehat{p} = p^*) = \boxed{\text{4.2)} \qquad\qquad}.$$

Take any $\tilde{p} > p^*$. Note that:

$$\begin{aligned}
\mathbb{P}_{p^*}(\widehat{p} = p^*) &= \mathbb{P}_{p^*}(\boxed{\left(\text{4.3)} \qquad\qquad \right)} \quad \text{for all } p \in \{0, 1, \ldots, \bar{p}\}, p \neq p^*) \\
&\leq \mathbb{P}_{p^*}(\text{AIC}(p^*) \boxed{\text{4.4)} \qquad\qquad} \text{AIC}(\tilde{p})) \\
&\quad \text{(since } P(A \cap B) \leq P(A)) \\
&= \mathbb{P}_{p^*}(\ln(1 + (\text{SSR}(p^*) - \text{SSR}(\tilde{p}))/\text{SSR}(\tilde{p})) < 2(\tilde{p} - p^*)/T) \\
&= \mathbb{P}_{p^*}(\ln(1 + \text{Wald}(p^*)/T) < 2(\tilde{p} - p^*)/T) \\
&\approx \mathbb{P}_{p^*}(\text{Wald}(p^*) < 2(\tilde{p} - p^*), \quad \text{(using a Taylor expansion))}
\end{aligned}$$

Which implies that

$$\boxed{\text{4.5)} \qquad\qquad} = \lim_{T\to\infty} \mathbb{P}_{p^*}(\widehat{p} = p^*) \leq \lim_{T\to\infty} \boxed{\left(\text{4.6)} \qquad\qquad \right)})$$

Since $2(\tilde{p} - p^*) > 0$, then:

$$1 \leq \lim_{T\to\infty} \boxed{\left(4.7) \qquad\qquad \right)} < \boxed{\left(4.8) \qquad\qquad \right)}.$$

A contradiction.

7 Bayesian Decision Making (25 points)

How would you set the problem of forecasting y_{t+k} as a Bayesian Decision Problem un quadratic loss? What is the the forecast that a Bayesian would use?

Answer: To set-up the problem we need to think about what is the *data*, what is the *statistical model* that we are working with, what is the *parameter space*, what is the *action space*, and finally what is the *loss function*.

DATA: The data in this problem is a univariate time series of size T:

$$(Y_1, Y_2, \ldots, Y_T).$$

STATISTICAL MODEL AND PARAMETER SPACE: The statistical model is a Gaussian AR(1):

$$Y_t = \phi Y_{t-1} + \epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0, \sigma^2), \quad \epsilon_t \text{ i.i.d.}$$

We will assume that σ^2 is known, so that the only parameter of the model is $\boxed{\text{5.1)} \qquad\qquad}$. This means that the parameter space is—in principle— $\boxed{\text{5.2)} \qquad\qquad}$. We will introduce some notation. Let f_ϕ denote:

$$f(Y_1, \ldots, Y_T | \phi)$$

and let f_ϕ^k denote

$$f(Y_1, \ldots, Y_T, Y_{T+k} | \phi).$$

ACTIONS: We are interested in generating a forecast for Y_{T+k}. A forecast is a function:

$$a : (Y_1, \ldots, Y_T) \to \mathbb{R}$$

This is slightly different to what we have done in class as actions are not scalars: they are functions.

Loss Function: A loss function is a function of $\boxed{\text{5.3)} \quad\quad\quad}$ and $\boxed{\text{5.3)} \quad\quad\quad}$. In this example, one possible loss to evaluate forecasts of Y in k periods in the future is the following:

$$\mathcal{L}(\boxed{\text{5.4)} \quad\quad\quad}, \boxed{\text{5.4)} \quad\quad\quad}) = \mathbb{E}_{f_\phi^k}[(Y_{T+k} - a(Y_1, \ldots, Y_T))^2]. \quad (5)$$

We can simplify the expression for the loss function. We have shown that:

$$Y_{T+k} = \phi^k Y_T + \eta_{T+k},$$

where η_{T+k} is some error term that is mean zero (even conditional on the information available at time T) and whose distribution depends on ϕ. Note that we can write the loss as:

$$\begin{aligned} \mathcal{L}(\boxed{\text{5.4)} \quad\quad}, \boxed{\text{5.4)} \quad\quad}) &= \mathbb{E}_{f_\phi^k}[(Y_{T+k} - \phi^k Y_T + \phi^k Y_T - a(Y_1, \ldots, Y_T))^2] \\ &= \mathbb{E}_{f_\phi^k}[(\eta_{T+k} + \phi^k Y_T - a(Y_1, \ldots, Y_T))^2] \\ &= \mathrm{Var}_\phi(\eta_{T+k}) + \boxed{\left(\text{5.5)} \mathbb{E}_{f_\phi} \quad\quad\quad\quad\quad\quad \right)}. \end{aligned}$$

This means that if ϕ were known, we would like to forecast Y_{T+k} using simply $\phi^k Y_T$.

Bayesian Forecasting: To solve the forecasting problem, a Bayesian decision maker postulates a *prior* on ϕ. Let $\pi(\phi)$ denote the prior and $\pi(\phi|Y_1, \ldots, Y_T)$ denote the posterior. In class we have shown that the Bayesian chooses the action a that minimizes posterior loss:

$$\mathbb{E}_{\pi(\phi|Y_1,\ldots,Y_T)}[\mathcal{L}(a, \phi)] = \int_\phi \mathcal{L}(a, \phi)\pi(\phi|Y_1, \ldots, Y_T)d\phi$$

Since the choice of action does not affect $\mathrm{Var}_\phi(\eta_{T+k})$, minimizing the posterior loss is equivalent to:

$$\min_{a} \int_{\phi} \mathbb{E}_{f_{\phi}} \boxed{\left[\text{5.6)} \qquad\qquad\qquad\qquad \right]} \pi(\phi|Y_1, \ldots, Y_T) d\phi \tag{6}$$

This is still a complicated problem, since a is a function. We can simplify this problem by defining $\mathbf{Y} = (Y_1, \ldots, Y_T)$ and noting that:

$$\mathbb{E}_{f_{\phi}}[(\phi^k Y_T - a(\mathbf{Y}))^2] = \int_{Y} (\phi^k Y_T - a(\mathbf{Y}))^2 f(\mathbf{Y}|\phi) d\mathbf{Y}$$

which implies that posterior loss can be written as:

$$\int_{\phi} \Big(\int_{\mathbf{Y}} (\phi^k Y_T - a(\mathbf{Y}))^2 f(\mathbf{Y}|\phi) d\mathbf{Y} \Big) \pi(\phi|Y_1, \ldots, Y_T) d\phi. \tag{7}$$

Changing the order of integration we get:

$$\int_{\mathbf{Y}} \Big(\int_{\phi} (\phi^k Y_T - a(\mathbf{Y}))^2 \pi(\phi|Y_1, \ldots, Y_T) d\phi \Big) f(\mathbf{Y}|\phi) d\mathbf{Y}.$$

This means that the Bayesian forecast can be defined as the function $a(\mathbf{Y})$ such that for each realization $\mathbf{Y}$ it minimizes:

$$\int_{\phi} \boxed{\left(\text{5.7)} \qquad\qquad\qquad\qquad \right)} d\phi = \mathbb{E}_{\pi}[(\phi^k Y_T - a(\mathbf{Y}))^2 | Y_1, \ldots, Y_T]$$

But then, we can proceed as in class and do:

$$
\begin{aligned}
\mathbb{E}_\pi[(\phi^k Y_T - a(\mathbf{Y}))^2 | Y_1, \ldots, Y_T] &= \mathbb{E}_\pi[(\phi^k Y_T - \mathbf{E}[\phi^k Y_T|\mathbf{Y}] + \mathbf{E}[\phi^k Y_T|\mathbf{Y}] - a(\mathbf{Y}))^2|\mathbf{Y}] \\
&= \mathbb{E}_\pi[(\phi^k Y_T - \mathbf{E}[\phi^k Y_T|\mathbf{Y}])^2|\,\mathbf{Y}] \\
&+ \mathbb{E}_\pi[(\mathbf{E}[\phi^k Y_T|\mathbf{Y}] - a(\mathbf{Y}))^2|\,\mathbf{Y}] \\
&+ \boxed{5.8)\qquad\qquad\qquad\qquad\qquad\qquad} \\
&= \mathbb{E}_\pi[(\phi^k Y_T - \mathbf{E}[\phi^k Y_T|\mathbf{Y}])^2|\,\mathbf{Y}] \\
&+ \mathbb{E}_\pi[(\mathbf{E}[\phi^k Y_T|\mathbf{Y}] - a(\mathbf{Y}))^2|\,\mathbf{Y}]
\end{aligned}
$$

And therefore, the Bayesian forecast for a prior π is:

$$
a^*(Y_1, \ldots, Y_T) = \boxed{5.9)\qquad\qquad\qquad\qquad\qquad\qquad} \tag{8}
$$