STA 542 – Homework 2

Due: Wednesday, September 16, at the start of class

Credit. This homework is graded for effort rather than correctness. A substantive attempt addresses every assigned part, shows the reasoning behind each answer, and interprets any computational output. Full effort credit does not require a correct solution: show what you tried, including unsuccessful approaches, and identify precisely where you remain stuck.

Submission. Submit one ordered hard copy, secured with a staple or paper clip, with your name clearly written on the first page. Mathematical and statistical reasoning must be written legibly by hand. R code, numerical output, tables, and plots may be printed and appended, or sketched when exact output is unnecessary. Label every computational exhibit, refer to it in the handwritten response, and explain what it shows.

Problems

Problem 1. Learning the covariances needed for a drift interval. Let $(D_t)_{t\in\mathbb{Z}}$ be strictly stationary increments with finite second moments, mean μ, and ACVF $\gamma(h) := \mathrm{Cov}(D_t, D_{t+h})$. The mean μ is the fixed, unknown drift. Assume the covariance cutoff is known:

$$\gamma(h) = 0 \quad \text{for } |h| > 1.$$

We observe $D_1, \dots, D_T$, where $T \geq 2$, and estimate μ by $\bar{D}_T := T^{-1}\sum_{t=1}^{T} D_t$. All probabilities below are under this one fixed process law. Work by hand; a calculator may be used for the normal probability in part (c).

(a) First suppose $\gamma(0)$ and $\gamma(1)$ are known. Use Proposition 5.1 to find $v_T := \mathrm{Var}(\bar{D}_T)$. For $0 < \alpha < 1$, construct a Chebyshev confidence interval for μ with coverage at least $1-\alpha$. Write its coverage statement and explain what varies across repetitions.

(b) The covariances are now unknown. Identify the three fixed-window functions g whose averages estimate the mean, the second moment, and the lag-one raw product. For the remainder, assume (D_t) is g-ergodic for each of these functions, as in Definition 5.7. State the three probability limits and use them to prove

$$\widehat{\gamma}_T(h) := \frac{1}{T}\sum_{t=1}^{T-h}(D_t - \bar{D}_T)(D_{t+h} - \bar{D}_T) \xrightarrow{P} \gamma(h), \qquad h = 0, 1.$$

Show why estimating the mean in this expression has the required limit. *Hint: expand the centered products. For lag one, the two partial means omit D_T/T and D_1/T; bound their expected absolute values using $\mathbb{E}[|D_t|] \leq \sqrt{\mu^2 + \gamma(0)}$, then apply Markov's inequality.* Explain briefly, using Example 5.6, why mean consistency alone would not justify this argument.

(c) Assume additionally that $c := \gamma(0) + 2\gamma(1) > 0$. Define

$$\widehat{v}_T := \frac{1}{T}\max\left\{0, \widehat{\gamma}_T(0) + 2\left(1 - \frac{1}{T}\right)\widehat{\gamma}_T(1)\right\}.$$

Use part (b) to show $T\widehat{v}_T \xrightarrow{P} c$ and $\widehat{v}_T/v_T \xrightarrow{P} 1$. Explain the role of $c > 0$. *Hint: first show $Tv_T \to c$; this deterministic denominator is eventually bounded away from zero.*

To see why $\widehat{v}_T - v_T \xrightarrow{P} 0$ alone is insufficient, consider i.i.d. $N(\mu, 1)$ increments and the deliberately poor estimate $\widetilde{v}_T := 1/(20T)$. Show that its absolute error tends to zero, but replacing v_T by $\widetilde{v}_T$ in the 95% interval from part (a) gives coverage below 95%. Express that coverage using the standard normal CDF Φ.

(d) This part extends Workbook 5's fixed-lag result to an estimated interval. Define

$$\widehat{C}_T := \left[\bar{D}_T - \sqrt{\widehat{v}_T/\alpha},\ \bar{D}_T + \sqrt{\widehat{v}_T/\alpha}\right].$$

For fixed $0 < \varepsilon < 1$, split noncoverage according to whether $\widehat{v}_T < (1-\varepsilon)v_T$. Use Chebyshev's inequality to prove

$$\mathbb{P}\{\mu \notin \widehat{C}_T\} \le \mathbb{P}\left\{\frac{\widehat{v}_T}{v_T} < 1 - \varepsilon\right\} + \frac{\alpha}{1-\varepsilon}.$$

Let $T \to \infty$ and then $\varepsilon \downarrow 0$ to conclude $\liminf_{T\to\infty} \mathbb{P}\{\mu \in \widehat{C}_T\} \ge 1 - \alpha$. No independence between $\bar{D}_T$ and $\widehat{v}_T$ is assumed. Finish by explaining why consistency at each fixed lag would not, without further assumptions, justify replacing every covariance through lag $T-1$ in Proposition 5.1 by its sample estimate.

Problem 2. Does the previous wait help predict the next Old Faithful eruption?

The R dataset `MASS::geyser` contains 299 consecutive records from Old Faithful in Yellowstone National Park, collected August 1–15, 1985. Use its `waiting` column: each entry is the interval, in minutes, from the start of the preceding eruption to the start of the recorded eruption.[1] Let X_t denote this interval before eruption t. At the start of eruption t, X_t is observed and the next interval X_{t+1} is to be predicted.

Model (X_t) as a weakly stationary univariate series with unknown mean μ, ACVF γ, and $\gamma(0) > 0$. Proposition 3.8 of Workbook 3 gives the best affine predictor of X_{t+1} from X_t alone:

$$m(x) := \mu + a(x - \mu), \qquad a := \frac{\gamma(1)}{\gamma(0)}.$$

Use this formula without rederiving it. The task is to estimate the rule from one path using Workbook 5, then assess its forecasts on later waits.

Run the preparation code below and use only `train` until part (c). Keep the observations in their recorded order. The first 200 waits form the training record; the remaining 99 will be used for evaluation.

```
data(geyser, package = "MASS")
x <- geyser$waiting
train <- x[1:200]
```

(a) Using only `train`, compute the sample mean and the sample ACVF at lags zero and one, using the divisor-T definition in Workbook 5, Definition 5.5. Replace the population quantities in $m(x)$ by these estimates to define $\widehat{a}_T$ and $\widehat{m}_T(x)$. Report a table with the training mean, both sample covariances, and the fitted slope. Give the predicted next interval after a 55-minute wait and after an 85-minute wait.

Plot the 199 training pairs (x_t, x_{t+1}), $t = 1, \ldots, 199$, with the preceding wait on the horizontal axis. Label both axes in minutes and add the fitted line $\widehat{m}_{200}(x)$ and the horizontal line at the training mean. Explain the fitted slope's sign in one sentence.

(b) For the limiting argument, let T grow and assume that (X_t) is ACVF-learnable through lag one, as in Workbook 5, Definition 5.8. Use Proposition 5.9 to show

$$\widehat{a}_T \xrightarrow{P} a, \qquad \widehat{m}_T(x) \xrightarrow{P} m(x) \quad \text{for each fixed } x \in \mathbb{R}.$$

Define $\widehat{a}_T = 0$ when $\widehat{\gamma}_T(0) = 0$. Explain how $\gamma(0) > 0$ is used and why this convention has no asymptotic effect. *Hint: first work on the event $\widehat{\gamma}_T(0) \geq \gamma(0)/2$, whose probability tends to one.*

[1] Azzalini and Bowman (1990), *A Look at Some Data on the Old Faithful Geyser.* See the MASS dataset documentation.

(c) Now use the 99 later waits. At each step, the actual preceding wait becomes available before the next forecast is made. Keep the training mean and fitted slope fixed throughout this evaluation.

```
test_index <- 201:length(x)
previous <- x[test_index - 1]
observed <- x[test_index]
```

For each $t = 201, \ldots, 299$, compare the constant forecast $\bar{x}_{200}$ with the fitted forecast $\widehat{m}_{200}(x_{t-1})$. Compute each rule's mean squared error,

$$\frac{1}{99} \sum_{t=201}^{299} (x_t - \widehat{x}_{t|t-1})^2.$$

Report both values and the percentage reduction in mean squared error from using the fitted rule instead of the constant rule. These are successive one-step forecasts: use the observed preceding wait each time, rather than feeding earlier predictions back into the rule.

(d) In three or four sentences, explain what the later observations say about using the previous wait to predict the next start in this historical record. Does part (b) imply that the next waiting time can eventually be predicted with arbitrarily small mean squared error? Use the population forecast error from Workbook 3, Proposition 3.8,

$$\mathbb{E}\big[(X_{t+1} - m(X_t))^2\big] = \gamma(0) - \frac{\gamma(1)^2}{\gamma(0)},$$

in your explanation, and identify when this error is positive.

Problem 3. What would help you understand this homework better?

Look back over your work on Problems 1 and 2. Identify three points that you would like to understand better, using the categories below. Each response should refer to a specific part of this homework.

(a) **A mathematical step.** Choose a calculation or proof step that you found difficult or could not fully justify. This might concern expanding the sample covariance, bounding an endpoint term, or justifying a convergence statement.

(b) **A statistical idea.** Choose a statistical idea used in your solution that remains unclear. For example, what does the confidence interval's probability statement mean, or why might a consistent estimate still be insufficient for constructing an interval? The idea should arise in this homework, even if it also applies to independent data.

(c) **A time-series idea.** Choose something about dependence or learning from one observed path that you would like to understand better. This could arise when estimating covariances in Problem 1 or learning a forecast rule from waiting times in Problem 2.

For each response, write a short paragraph addressing the following:

- Identify the subpart and formulate your own question. Explain what you understood or could do, and where your reasoning became uncertain.
- Suggest how you could explore the question using a specific result, proof, example, or exercise from Workbooks 1–5. Describe a possible activity, such as working through a simpler case or discussing your attempted reasoning with a TA.
- Describe a short calculation, explanation, or example you could attempt without notes to check your understanding.

These are possible approaches to further study; you are not being asked to complete additional work for this assignment.