Exercises

1. (4 points) Consider the following *Moving Average of order 1* (MA(1)) process:

$$Y_k = \epsilon_k + \theta\epsilon_{k-1}, \quad \text{with } \epsilon_k \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma_\epsilon^2),$$

where θ is the moving average coefficient.

Answer the following questions:

(a) Derive the two-steps-ahead forecast density, $f_{Y_{k+2}|\mathcal{F}_k}$, where $\mathcal{F}_k = \{Y_k = y_k, Y_{k-1} = y_{k-1}, \ldots, Y_1 = y_1\}$.
Answer: For the two-steps-ahead forecast:

$$Y_{k+2} = \epsilon_{k+2} + \theta\epsilon_{k+1}.$$

Since ϵ_{k+1} and ϵ_{k+2} are independent of $\mathcal{F}_k$, the conditional mean and variance are:

$$\begin{aligned}\mathbb{E}[Y_{k+2} \mid \mathcal{F}_k] &= \mathbb{E}[\epsilon_{k+2}] + \theta\mathbb{E}[\epsilon_{k+1}] = \mu + \theta\mu = \mu(1+\theta),\\ \operatorname{Var}(Y_{k+2} \mid \mathcal{F}_k) &= \operatorname{Var}(\epsilon_{k+2}) + \theta^2 \operatorname{Var}(\epsilon_{k+1}) = \sigma_\epsilon^2 + \theta^2\sigma_\epsilon^2 = (1+\theta^2)\sigma_\epsilon^2.\end{aligned}$$

Therefore, the two-steps-ahead forecast density is:

$$Y_{k+2} \mid \mathcal{F}_k \sim \mathcal{N}\left(\mu(1+\theta),\ (1+\theta^2)\sigma_\epsilon^2\right).$$

(b) Derive the value of the two-steps-ahead optimal forecast $\widehat{y}_{k+2}^\star$ when the loss function is the squared error.
Answer: Under squared error loss, the optimal forecast is the conditional mean:

$$\widehat{y}_{k+2}^\star = \mathbb{E}[Y_{k+2} \mid \mathcal{F}_k] = \mu(1+\theta).$$

(c) Derive the value of the one-step-ahead optimal forecast $\widehat{y}_{k+1}^\star$ when the loss function is the squared error. *Hint*: $\mathcal{F}_0 = \emptyset$.
Answer: We want to compute:

$$\widehat{y}_{k+1}^\star = \mathbb{E}\left[Y_{k+1} \mid \mathcal{F}_k\right] = \mathbb{E}\left[\varepsilon_{k+1} + \theta\varepsilon_k \mid \mathcal{F}_k\right] = \mathbb{E}\left[\varepsilon_{k+1} \mid \mathcal{F}_k\right] + \theta\mathbb{E}\left[\varepsilon_k \mid \mathcal{F}_k\right].$$

On the one hand, we have that

$$\mathbb{E}\left[\varepsilon_{k+1} \mid \mathcal{F}_k\right] = \mathbb{E}\left[\varepsilon_{k+1}\right] = \mu.$$

On the other hand, since $Y_k = \varepsilon_k + \theta\varepsilon_{k-1}$, we have that $\varepsilon_k = Y_k - \theta\varepsilon_{k-1}$, so

$$\mathbb{E}\left[\varepsilon_k \mid \mathcal{F}_k\right] = \mathbb{E}\left[Y_k \mid \mathcal{F}_k\right] - \theta\mathbb{E}\left[\varepsilon_{k-1} \mid \mathcal{F}_k\right] = y_k - \theta\mathbb{E}\left[\varepsilon_{k-1} \mid \mathcal{F}_{k-1}\right].$$

Denoting $E_k = \mathbb{E}[\varepsilon_k \mid \mathcal{F}_k]$, from the chain of equalities above, we have the recursion

$$E_k = y_k - \theta E_{k-1},$$

with the initial condition:

$$E_0 = \mathbb{E}[\varepsilon_0 \mid \mathcal{F}_0] = \mathbb{E}[\varepsilon_0 \mid \emptyset] = \mathbb{E}[\varepsilon_0] = \mu.$$

The first few values of the recursion are:

$$\begin{aligned}
E_1 &= y_1 - \theta\mu, \\
E_2 &= y_2 - \theta E_1 = y_2 - \theta(y_1 - \theta\mu) = y_2 - \theta y_1 + \theta^2\mu, \\
E_3 &= y_3 - \theta E_2 = y_3 - \theta(y_2 - \theta y_1 + \theta^2\mu) = y_3 - \theta y_2 + \theta^2 y_1 - \theta^3\mu, \\
&\vdots
\end{aligned}$$

Hence,

$$E_k = \sum_{i=0}^{k-1}(-\theta)^i y_{k-i} + (-\theta)^k\mu.$$

Therefore, we have that

$$\widehat{y}^{\star}_{k+1} = \mathbb{E}[Y_{k+1} \mid \mathcal{F}_k] = 0 + \theta E_k = \theta\sum_{i=0}^{k-1}(-\theta)^i y_{k-i} + \theta(-\theta)^k\mu.$$

Therefore, in the MA(1) model, the optimal one-step-ahead predictor at time k depends on the whole previous history $y_{1:k}$.

2. (6 points) Consider the following one-step-ahead forecast density:

$$f_{Y_{k+1}|\mathcal{F}_k}(y \mid \mathcal{F}_k) = \begin{cases} \frac{1}{y_k} & y \in [0, y_k] \\ 0 & \text{otherwise} \end{cases}$$

Answer the following questions:

(a) Write down an expression for the conditional expected loss for the squared error (SE) loss function. Your answer should be an expression not involving integrals.

Answer: To compute the conditional risk, we need to solve the following

integral:

$$\begin{aligned}\mathcal{R}\left(\widehat{y}_{k+1};\mathcal{F}_k\right) &= \int_{-\infty}^{\infty}\left(\widehat{y}_{k+1}-y\right)^2 f_{Y_{k+1}|\mathcal{F}_k}(y\mid\mathcal{F}_k)\,dy\\ &= \int_0^{y_k}\left(\widehat{y}_{k+1}-y\right)^2\frac{1}{y_k}\,dy\\ &= \frac{1}{y_k}\int_0^{y_k}\left(\widehat{y}_{k+1}^2+y^2-2\widehat{y}_{k+1}y\right)\,dy\\ &= \frac{1}{y_k}\left(\widehat{y}_{k+1}^2\int_0^{y_k}dy+\int_0^{y_k}y^2\,dy-2\widehat{y}_{k+1}\int_0^{y_k}y\,dy\right)\\ &= \widehat{y}_{k+1}^2+\frac{y_k^2}{3}-\widehat{y}_{k+1}y_k\end{aligned}$$

(b) Find the derivative of the conditional risk with respect $\widehat{y}_{k+1}$, force the derivative to be zero, and find the value $\widehat{y}_{k+1}^{\star}$ that makes the derivative equal to zero.

Answer: The derivative is:

$$\frac{\partial\mathcal{R}\left(\widehat{y}_{k+1};\mathcal{F}_k\right)}{\partial\widehat{y}_{k+1}} = 2\widehat{y}_{k+1}-y_k = 0.$$

Hence, the optimal predictor is:

$$\widehat{y}_{k+1}^{\star} = \frac{y_k}{2}.$$

(c) Derive an expression for this conditional expectation $\mathbb{E}[Y_{k+1}\mid\mathcal{F}_k]$. Write the integral equation of the conditional expectation and solve it explicitly.

Answer: The conditional expectation is:

$$\mathbb{E}[Y_{k+1}\mid\mathcal{F}_k] = \int_{-\infty}^{\infty} y f_{Y_{k+1}|\mathcal{F}_k}(y\mid\mathcal{F}_k)\,dy = \int_0^{y_k} y\frac{1}{y_k}\,dy = \frac{y_k}{2}.$$

(d) What is the relationship between your two previous answers? Justify this relationship.

Answer: They are the same because the conditional expectation is equal to the optimal predictor when the loss function of the squared error.

3. (4 points) Consider the following AR(1) model:

$$Y_k = \theta Y_{k-1}+\epsilon_k, \quad \text{with } \epsilon_k \overset{\text{iid}}{\sim} \mathcal{N}(0,1) \text{ and } Y_0 = y_0.$$

where θ is an unknown parameter we want to estimate. You are provided with L observations of the time series $y_{1:L} = (y_1, y_2, \ldots, y_L)$.

Answer the following questions:

(a) Write down the likelihood function $\mathscr{L}(\theta; y_{1:L})$ for the AR(1) model using the causal factorization in (3).

Answer: The likelihood function is:

$$\mathscr{L}(\theta; y_{1:L}) = f_{Y_{1:L}|\mathcal{F}_0}(y_{1:L} \mid \mathcal{F}_0) = \prod_{k=0}^{L-1} f_{Y_{k+1}|\mathcal{F}_k}(y_{k+1} \mid \mathcal{F}_k),$$

where

$$f_{Y_{k+1}|\mathcal{F}_k}(y_{k+1} \mid \mathcal{F}_k) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(y_{k+1} - \theta y_k)^2}{2}\right).$$

Therefore, the likelihood function can be expressed as:

$$\mathscr{L}(\theta; y_{1:L}) = (2\pi)^{-L/2} \exp\left(-\sum_{k=0}^{L-1} \frac{(y_{k+1} - \theta y_k)^2}{2}\right).$$

(b) Derive the log-likelihood function, defined as $\log \mathscr{L}(\theta; y_1, y_2, \ldots, y_L)$, where $\log(\cdot)$ is the natural logarithm.

Answer: The log-likelihood function is:

$$\log \mathscr{L}(\theta; y_{1:L}) = -\frac{L}{2} \log(2\pi) - \frac{1}{2} \sum_{k=0}^{L-1} (y_{k+1} - \theta y_k)^2.$$

(c) Compute the derivative of the log-likelihood function with respect to θ.

Answer: To compute the derivative of the log-likelihood with respect to θ, we differentiate each term:

$$\frac{\partial}{\partial \theta} \log \mathscr{L}(\theta; y_{1:L}) = -\frac{1}{2} \sum_{k=0}^{L-1} \frac{\partial}{\partial \theta} (y_{k+1} - \theta y_k)^2.$$

Using the chain rule, we compute the derivative of the squared term:

$$\frac{\partial}{\partial \theta} (y_{k+1} - \theta y_k)^2 = 2 (y_{k+1} - \theta y_k) (-y_k),$$

Substituting this into the expression for the derivative of the log-likelihood:

$$\frac{\partial}{\partial \theta} \log \mathscr{L}(\theta; y_{1:L}) = \sum_{k=0}^{L-1} (y_{k+1} - \theta y_k) y_k.$$

(d) Set the derivative equal to zero and solve for θ to obtain the Maximum Likelihood Estimate (MLE) $\theta^\star$ as a function of the observations.

Answer: To find the Maximum Likelihood Estimate $\theta^\star$, we set the derivative of the log-likelihood function equal to zero:

$$\frac{\partial}{\partial\theta}\log\mathscr{L}(\theta; y_{1:L}) = \sum_{k=0}^{L-1}(y_{k+1} - \theta y_k)\, y_k = 0.$$

Expand the summation:

$$\sum_{k=0}^{L-1} y_{k+1}y_k - \theta\sum_{k=0}^{L-1} y_k^2 = 0.$$

Rearrange this equation to solve for $\theta^\star$:

$$\theta^\star\sum_{k=0}^{L-1} y_k^2 = \sum_{k=0}^{L-1} y_{k+1}y_k.$$

Thus, the MLE for θ is:

$$\theta^\star = \frac{\sum_{k=0}^{L-1} y_{k+1}y_k}{\sum_{k=0}^{L-1} y_k^2}.$$

(e) Interpret the MLE expression in terms of the autocorrelation of the observed data.

Answer: The MLE $\theta^\star$ can be interpreted as an estimate of the lag-1 autocorrelation of the time series. In the context of the AR(1) model, this means that $\theta^\star$ represents the strength of the linear relationship between consecutive observations.

4. (5 points) Consider the AR(1) model with uniform noise:

$$Y_k = \theta Y_{k-1} + \epsilon_k, \quad \text{with } \epsilon_k \overset{\text{iid}}{\sim} \texttt{Unif}(-1,1) \text{ and } Y_0 = 1,$$

where θ is the unknown parameter we want to estimate and $\texttt{Unif}(a,b)$ is the uniform distribution for $a < b$, i.e.,

$$X \sim \texttt{Unif}(a,b) \Rightarrow f_X(x) = \begin{cases} \dfrac{1}{b-a}, & x \in [a,b], \\ 0, & \text{otherwise.} \end{cases}$$

In the following questions, you are required to perform two steps of Bayesian parameter estimation assuming the following prior uniform distribution for the unknown parameter: $\theta \sim \texttt{Unif}(-1,1)$, i.e., we are 100% certain that the parameter θ is between -1 and 1.

(a) Compute the likelihood function $\mathscr{L}(\theta; y_0 = y_1 = 1)$.
Answer: After observing $Y_0 = 1$, we have that:

$$(Y_1 \mid \theta, Y_0 = 1) \sim \text{Unif}(-1 + \theta, 1 + \theta),$$

i.e.,

$$f_{Y_1|\theta,\mathcal{F}_0}(y_1 \mid \theta, Y_0 = 1) = \begin{cases} 1/2, & y_1 \in [-1 + \theta, 1 + \theta], \\ 0, & \text{otherwise.} \end{cases}$$

Notice that $y_1 = 1$ and $y_1 \in [-1 + \theta, 1 + \theta]$ implies that $\theta \in [0, 2]$; hence, we have the following likelihood function:

$$\mathscr{L}(\theta; y_0 = 1, y_1 = 1) = \begin{cases} 1/2, & \theta \in [0, 2], \\ 0, & \text{otherwise.} \end{cases}$$

(b) Compute the posterior distribution $f_{\theta|\mathcal{F}_1}$ given $\mathcal{F}_1 = \{Y_0 = Y_1 = 1\}$.
Answer: The posterior distribution is proportional to the product of the likelihood and the prior distribution, i.e.,

$$f_{\theta|\mathcal{F}_1} \propto \mathscr{L}(\theta; y_0 = 1, y_1 = 1) \cdot f_\theta(\theta).$$

Since the likelihood is a uniform distribution supported on $[0, 2]$ and the prior is $\theta \sim \texttt{Unif}(-1, 1)$, we have that the posterior is uniformly distributed on $[0, 1]$, i.e.,

$$f_{\theta|\mathcal{F}_1} = \begin{cases} 1, & \theta \in [0, 1], \\ 0, & \text{otherwise,} \end{cases}$$

where the value of $f_{\theta|\mathcal{F}_1}$ for $\theta \in [0, 1]$ is chosen such that the resulting function is a density function that integrates to 1. This result implies that after we observe $Y_1 = 1$, we are 100% certain that $\theta \in [0, 1]$.

(c) Compute the likelihood function $\mathscr{L}(\theta; y_0 = y_1 = 1, y_2 = -1/2)$.
Answer: After observing $Y_2 = -1/2$, we have that:

$$(Y_2 \mid \theta, Y_0 = Y_1 = 1) \sim \text{Unif}(-1 + \theta, 1 + \theta),$$

i.e.,

$$f_{(Y_2|\theta,\mathcal{F}_1)}(y_2 \mid \theta, Y_0 = Y_1 = 1) = \begin{cases} 1/2, & y_2 \in [-1 + \theta, 1 + \theta], \\ 0, & \text{otherwise.} \end{cases}$$

Since $y_2 = -1/2$, the new likelihood function is:

$$\mathscr{L}(\theta; y_0 = 1, y_1 = 1, y_2 = -1/2) = \begin{cases} 1/2, & \theta \in [-3/2, 1/2], \\ 0, & \text{otherwise.} \end{cases}$$

(d) Compute the new posterior distribution $f_{\theta|\mathcal{F}_2}$.

Answer: The posterior distribution is computed using the previous posterior as the new prior, i.e.,

$$f_{\theta|\mathcal{F}_2} \propto \mathscr{L}(\theta; y_0 = y_1 = 1, y_2 = -1/2) \cdot f_{\theta|\mathcal{F}_1}(\theta \mid Y_0 = 1, Y_1 = 1).$$

The previous posterior is uniformly distributed in $[0, 1]$ and the new likelihood is uniformly distributed in $[-3/2, 1/2]$, hence, the new posterior distribution is uniformly distributed in the intersection of both intervals, i.e.,

$$f_{\theta|\mathcal{F}_2} = \begin{cases} 2, & \theta \in [0, 1/2], \\ 0, & \text{otherwise.} \end{cases}$$

This implies that, after the second observation $Y_2 = -1/2$, we are 100% certain that $\theta \in [0, 1/2]$.

5. Answer the following short questions (justify your answers):

(a) What is the most appropriate loss function when the one-step-ahead forecast distribution is heavy-tailed?

Answer: The most appropriate loss function in this case would be the absolute error. Heavy-tailed distributions are more sensitive to outliers, and using absolute error is more robust to outliers compared to the squared error loss, which gives larger penalties to extreme deviations.

(b) What is the most appropriate loss function when the one-step-ahead forecast distribution is Gaussian?

Answer: The most appropriate loss function is the squared error loss. This is because the maximum likelihood estimate for the mean of a Gaussian distribution minimizes the squared error loss.

(c) What are the key differences between the model-based approach and the Bayesian model-based approach?

Answer: In the model-based approach, parameters are treated as fixed but unknown quantities and are estimated using techniques such as maximum likelihood or least squares. In the Bayesian model-based approach, parameters are treated as random variables with prior distributions, and inference is performed using the posterior distribution of these parameters given the data, typically using techniques such as Markov Chain Monte Carlo (MCMC).

(d) What are the advantages and disadvantages of the traditional machine learning approach compared to the model-based approach?

Answer: Advantages: Traditional machine learning methods, such as random forests or boosting, often require fewer assumptions about the

underlying data-generating process and can perform well on a variety of tasks without explicitly specifying a model. They are flexible and generally perform well with high-dimensional data. **Disadvantages:** They often lack interpretability, and their predictions can be hard to explain. These models can also require large datasets to perform well and may overfit if not properly regularized.

(e) What are the advantages and disadvantages of the deep learning approach compared to the traditional machine learning approach?

Advantages: Deep learning approaches, such as neural networks, are powerful for capturing complex patterns in large datasets. They can automatically learn features from data, reducing the need for manual feature engineering.

Disadvantages: Deep learning models typically require large amounts of labeled data and high computational resources for training. They also suffer from interpretability issues and are often seen as *black-box* models.

(f) How can long-term dependencies be incorporated into Convolutional Neural Networks (CNNs)?

Answer: Long-term dependencies can be incorporated into CNNs by increasing the receptive field or subsampling the time series.

(g) What are the advantages of Transformers compared to vanilla RNNs? Vanilla RNNs do not include LSTMs or GRUs.

Answer: Transformers use a self-attention mechanism that allows them to capture relationships between distant time steps more effectively.

6. (From 1A—3 points) Consider a Markov chain $\mathcal{X} = \{X_1, X_2, \ldots\}$ with state space $\mathcal{S} = \{1, 2, 3\}$ and transition matrix:

$$P = \begin{bmatrix} 0 & 0.8 & * \\ 0.1 & * & 0.3 \\ 0.2 & 0.8 & * \end{bmatrix}.$$

Answer the questions below and feel free to use numerical software if you need to compute matrix multiplications:

(a) Fill in the unknown entries, denoted by $*$, in the transition matrix.

Answer: Since the rows of a Markov transition matrix must sum to 1, we can fill in the unknown entries. Thus, the completed transition matrix is:

$$P = \begin{bmatrix} 0 & 0.8 & 0.2 \\ 0.1 & 0.6 & 0.3 \\ 0.2 & 0.8 & 0 \end{bmatrix}.$$

(b) Find the probability that $X_2 = s_3$ given that $X_0 = s_1$.

Answer: The probability that $X_2 = s_3$ given $X_0 = s_1$ is computed by finding the probability of transitioning from state s_1 to s_3 in two steps. This can be written as:

$$P(X_2 = s_3 \mid X_0 = s_1) = [P^2]_{1,3} = 0.24,$$

where P^2 is the square of the transition matrix P.

(c) Given that $X_0 = s_1$, compute the distribution vector $\boldsymbol{\pi}_2$, representing the state probabilities at time step 2.

Answer: The distribution vector $\boldsymbol{\pi}_2$ gives the probabilities of being in each state at time step 2, starting from $X_0 = s_1$. The state probabilities at time step 1 are given by:

$$\boldsymbol{\pi}_2 = (P^\intercal)^2 \boldsymbol{\pi}_0,$$

where $\boldsymbol{\pi}_0 = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix}^\intercal$ since we start in state 1. Therefore, the distribution vector at time step 2 is:

$$\boldsymbol{\pi}_2 = \begin{bmatrix} 0.12 \\ 0.64 \\ 0.24 \end{bmatrix}.$$

7. (From 1A—3 points) What is the maximum number of non-zero entries in the transition matrix P of a higher-order Markov chain with states $\mathcal{S} = \{s_1, s_2, \ldots, s_H\}$ and memory m (i.e., order $m-1$? Your answer should be a function of H and m.

Answer: The state transition matrix of the higher-order Markov chain is a square matrix of dimensions $H^{m+1} \times H^{m+1}$, corresponding to all possible compound states $(X_k, X_{k-1}, \ldots, X_{k-m})$. However, not all entries in this matrix can be non-zero. A compound state $(X_k, X_{k-1}, \ldots, X_{k-m})$ can only transition to a state of the form $(X_{k+1}, X_k, X_{k-1}, \ldots, X_{k-m+1})$, where X_{k+1} can take any of the H possible values.

Thus, each row of the transition matrix has, at most, H non-zero entries. Given that there are H^{m+1} rows in total, the maximum number of non-zero entries in the transition matrix is $H \times H^{m+1} = H^{m+2}$.

8. (From 1A—4 points) Below is a matrix representing the autocovariance matrix of a random process with two components: $\mathcal{Y}^1$ and $\mathcal{Y}^2$. Information about the autocorrelation functions and cross-correlation functions between the processes are shown for lag values less than 3 in the following table:

Autocorrelation: $R_{\mathcal{Y}^1}(h)$	
Lag	Value
0	1.0
1	0.5
2	0.2
3	0.1

Cross-correlation: $R_{\mathcal{Y}^1\mathcal{Y}^2}(h)$	
Lag	Value
0	0.8
1	0.4
2	0.2
3	0.1

Cross-correlation: $R_{\mathcal{Y}^2\mathcal{Y}^1}(h)$	
Lag	Value
0	0.8
1	0.3
2	0.2
3	0.05

Autocorrelation: $R_{\mathcal{Y}^2}(h)$	
Lag	Value
0	1.0
1	0.4
2	0.1
3	0.05

Assuming that the variance of $\mathcal{Y}^1$ and $\mathcal{Y}^2$ are 1 and 2, respectively, answer the following questions:

(a) Calculate the covariance between $\mathcal{Y}^1$ and $\mathcal{Y}^2$ at lag 0.

Answer: The covariance is:

$$R_{\mathcal{Y}^1\mathcal{Y}^2}(0) \cdot \sigma_{\mathcal{Y}^1} \cdot \sigma_{\mathcal{Y}^2} = 0.8\sqrt{2}$$

(b) Compute $R_{\mathcal{Y}^1}(-2)$.

Answer: For stationary processes, the autocorrelation function is symmetric, meaning:

$$R_{\mathcal{Y}^1}(h) = R_{\mathcal{Y}^1}(-h).$$

Thus,

$$R_{\mathcal{Y}^1}(-2) = R_{\mathcal{Y}^1}(2) = 0.2.$$

(c) Calculate $R_{\mathcal{Y}^1\mathcal{Y}^2}(-1)$?

Answer: For stationary processes, the cross-correlation function has the property:

$$R_{\mathcal{Y}^1\mathcal{Y}^2}(-h) = R_{\mathcal{Y}^2\mathcal{Y}^1}(h).$$

Thus,

$$R_{\mathcal{Y}^1\mathcal{Y}^2}(-1) = R_{\mathcal{Y}^2\mathcal{Y}^1}(1) = 0.3.$$