ESE-4380 Midterm. NAME: Tony Wang (Jie) ; tonyw3 ; 83749364

1. (15 points) Short questions: **Provide justification** for all answers to receive full credit.

   (a) Under what conditions does weak-sense stationarity imply strong-sense stationarity?

   When the timeseries is a linear combination of gaussian distribution

   (b) Is the auto-covariance at lag zero, $C_{\mathcal{Y}}(k, 0)$, necessarily equal to the variance of the process $\mathcal{Y}$ if the process is not weak-sense stationary (WSS)?

   No, if the process is not WSS, then $C_y(k,0)$ may also depend on k rather than solely h $\Rightarrow C_y(k,0) \neq Var(\mathcal{Y})$

   (c) Describe the autocorrelation function (ACF) of a white noise process.

   A white noise process is WSS, its ACF is independent with time k

   (d) Consider the following process:

   $$Y_k = Y_{k-1} + \epsilon_k, \quad \text{where } \epsilon_k \sim \mathcal{N}(0, \sigma_\epsilon^2) \text{ and } Y_0 = 0.$$

   is not a WSS for all k

   Is this process weak-sense stationary (WSS) for all k? Justify your answer.

   ① Mean: $E[Y_k] = E[Y_{k-1}] + E[\varepsilon_k]$ ② Var: $Var[Y_k] = Var[Y_{k-1}] + 6_\varepsilon^2$

   $= E[Y_{k-1}] + 0$ so $Var[Y_k] = k 6_\varepsilon^2$ depends on k,

   So $E[Y_k] = 0$ for $k \in \mathbb{Z}$ $\boxed{\checkmark}$

   (e) True or False: If a stochastic process $\mathcal{X}$ Granger-causes $\mathcal{Y}$, then there is a true causation from $\mathcal{X}$ to $\mathcal{Y}$.

   True

   (f) What issue arises when calculating the empirical autocorrelation function (ACF) of an AR(1) process with a linear trend added? The empirical ACF uses a single realization of the path integral to estimate the theoretical autocorrelation.

   The empirical ACF may be drifted to infinity caused by integration on the extra linear trend

   (g) Consider a one-step-ahead forecast density that is symmetric around its mean μ, i.e., $f_{Y_{k+1}|\mathcal{F}_k}(\mu + y \mid \mathcal{F}_k) = f_{Y_{k+1}|\mathcal{F}_k}(\mu - y \mid \mathcal{F}_k)$ for all $y \in \mathbb{R}$. What is the optimal one-step-ahead predictor $\widehat{y}^*_{k+1}$ under the absolute error (AE) loss function? Justify your answer.

   $y^*_{k+1} \equiv f_{Y_{k+1}|F_k}(\mu | F_k)$, because ~~after~~ symmetry + AE means the peak is at $y = \mu$

   (h) What is the Gibbs sampler used for in Bayesian estimation?

   It's used to sample random process in gibbs method

(i) What is the most appropriate loss function when the one-step-ahead forecast distribution is uniform between -1 and 1?

Squared Error Loss

(j) What is one of the key advantages of the deep learning approach compared to traditional machine learning methods?

① More powerful, it can learn representation in data automatically thus having better accuracy in inference.

(k) What is one significant disadvantage of the traditional machine learning approach compared to the model-based approach?

It may overfit on the training set data, and ML ~~causes~~ cost more resource / data

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

Bayesian model-based one involves random variable to handle estimation

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

① Self-attention mechanism could obtain global info of data
② Multihead makes GPU-based multithread processing easily

(n) In the context of Maximum Likelihood Estimation (MLE), why do we maximize the likelihood (i.e., the probability density of the observed data given the parameters) rather than maximizing the probability of observing the data directly?

Likelihood is the summary of all probabilities given initial condition, it could better ~~imp~~ indicate the global optimal, which may be local

(o) Provide the definition of the conditional risk, $\mathcal{R}(\widehat{y}_{k+1}; \mathcal{F}_k) = \mathbb{E}[\ell(Y_{k+1}, \widehat{y}_{k+1}) \mid \mathcal{F}_k]$ in the form of an integral. in prob...

$$R(\hat{y}_{k+1}; F_k) = E[\ell(Y_{k+1}, \hat{Y}_{k+1}) \mid F_k]$$

$$= E\left[\prod_{k=1}^{L} P(y_k \mid F_k)\right]$$

$$= \int_{F_k}^{F_{k+1}} P(y_k ; F_k)$$

2. (15 points) Programming questions:

   (a) What type of stochastic process does the following Python function generate? What do the parameters param1 and param2 represent in the process?

```
def generate_process(param1, param2, n):
    process = np.zeros(n)
    meas = np.random.normal(0, param2, n)
    for t in range(1, n):
        process[t] = param1 * process[t-1] + meas[t]
    return process
```

$Y_k = \phi Y_{k-1} + \varepsilon_n$

Answer: AR(1) ; param1 is ϕ for the AR(1)

Param2 is σ^2 of white noise

variance

   (b) Describe the type of stochastic process generated by the following Python function, and explain the role of the parameters in the process.

```
def generate_process(L=1000, C=2, A=None, AC=None):
    if A is None:
        A = np.array([[0.3, 0.5], [0.2, 0.1]])
    if AC is None:
        AC = np.eye(C)
    meas = np.random.multivariate_normal(np.zeros(C), AC, L)
    data = np.zeros((L, C))
    for t in range(1, L):
        data[t] = A @ data[t-1] + meas[t]
    df = pd.DataFrame(data, columns=[f'Y{i+1}' for i in range(C)])
    return df
```

L: length of ~~param~~ data

C: matrix dim

A: ~~State transit~~

State transition matrix

AC: (c×c) gaussian matrix

Mutivariate Random Process

Answer: Vector Autoregression process with lag (1)

(2×2 matrix)

   (c) The following Python code generates a time series based on a certain stochastic process. Identify the type of process generated by this function and explain the meaning of each parameter in the context of this process.

```
def generate_time_series(param1, param2, param3, n):
    process = np.zeros(n)
    meas_values = np.random.normal(0, param3, n)
    for t in range(1, n):
        process[t] = param1 * process[t-1] + param2 * meas[t-1] + meas[t]
    return process
```

ε_k is normal distribute

$\varepsilon_k \overset{iid}{\sim} N(0, \text{Param}^3)$

$Y_k = \phi Y_{k-1} + \theta \varepsilon_{k-1} + \varepsilon_k$

Answer:

MAAR(1)

Param1 is ϕ for random process

Param2 is θ for ε_{k-1} term

Param3 is σ^2 variance of white noise

3. (10 points) Below is a matrix representing the autocovariance matrix of a stationary multivariate random process with two components: $\mathcal{Z}^1$ and $\mathcal{Z}^2$. Information about the autocorrelation functions and cross-correlation functions between the processes are shown for lag values less than 3 in the following table:

Lag	Autocorrelation: $R_{\mathcal{Z}^1}(h)$ Value	Cross-correlation: $R_{\mathcal{Z}^1\mathcal{Z}^2}(h)$ Value	Cross-correlation: $R_{\mathcal{Z}^2\mathcal{Z}^1}(h)$ Value	Autocorrelation: $R_{\mathcal{Z}^2}(h)$ Value
0	1.0	0.7	0.7	1.0
1	0.6	0.35	0.4	0.5
2	0.3	0.15	0.05	0.25
3	0.15	0.05	0.01	0.1

Assuming that the variance of both $\mathcal{Z}^1$ and $\mathcal{Z}^2$ is 2, answer the following questions:

(a) Compute $\mathrm{Cov}(\mathcal{Z}^2_k, \mathcal{Z}^2_{k+3})$.

$$\mathrm{Cov}(Z_k^2 Z_{k+3}^2) = \sigma_{z^2}^2 R_{z^2}(h), \quad h=3$$

$$= 2^2 \times 0.1 = 0.4$$

(b) Compute $\mathrm{Cov}(\mathcal{Z}^1_k, \mathcal{Z}^2_{k+3})$.

$$\mathrm{Cov}(Z_k^1 Z_{k+3}^2) = \sigma_{z^1}\sigma_{z^2} R_{z_1 z^2}(h=3)$$

$$= 2 \times 2 \times 0.05$$

$$= 0.2$$

(c) Compute $\mathrm{Cov}(\mathcal{Z}^2_k, \mathcal{Z}^1_{k+3})$.

$$\mathrm{Cov}(Z_k^2, Z_{k+3}^1) = \sigma_{z^2}\sigma_{z^1} R_{z_2 z_1}(h=3)$$

$$= 2 \times 2 \times 0.01$$

$$= 0.04$$

4. (15 points) Consider a fair coin. At each time step k, you flip the coin. If it lands heads, you assign the value $N_k = 1$; if it lands tails, you assign the value $N_k = -1$. Based on this random process $\mathcal{N}$, define the following stochastic process:

$$X_k = \alpha X_{k-1} + N_k, \quad \text{with } X_0 = 1 \text{ and } |\alpha| < 1,$$

where $\alpha \in \mathbb{R}$ is a constant.

$1 > \alpha > 0$

$P_{N_k} = \begin{cases} 0.5, & N_k = 1 \\ 0.5, & -1 \end{cases}$

$E[N_k] = 0.5 \times 1 + 0.5 \times -1 = 0$

(a) Compute the theoretical mean $E[X_k]$ of the process for all $k \in \mathbb{N}$.

$E[X_k] = E[\alpha X_{k-1} + N_k]$
$= E[\alpha X_{k-1}] + E[N_k]$
$= \alpha E[X_{k-1}] + 0$

So $E[X_k] = 0$ for $k \in N$

(b) Compute the theoretical variance $\text{Var}(X_k)$ of the process for all $k \in \mathbb{N}$. *Hint*: Feel free to leave your answer as a summation. *Hint*: The variance of the coin-flip process is 1.

$Var(X_k) = Var(\alpha X_{k-1} + N_k)$, $Var(N_k) = 1$ as hint. and proved in HW2
$= Var(\alpha X_{k-1}) + Var(N_k)$
$= \alpha^2 Var(X_{k-1}) + 1$

let $V_k = Var(X_k)$,
then $V_k = \alpha^2 V_{k-1} + 1$
$V_{k-1} = \alpha^2 V_{k-2} + 1$
$\vdots$
$V_2 = \alpha^2 V_1 + 1$
$V_1 = \alpha^2 V_0 + 1 = 1$

So $V_k = \sum_{i=1}^{k-1} \alpha^{2i} + 1$

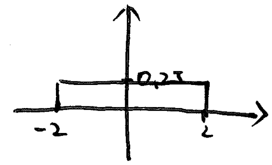
$\Rightarrow V_k = \alpha^2 \left(\frac{1-\alpha^{k-1}}{1-\alpha}\right) + 1$

(c) For any time step k, construct a 100% confidence interval for X_k that is as narrow as possible. For example, the interval $(-\infty, \infty)$ is not a valid answer. *Hint*: Feel free to define the interval using summations.

Within k step, we know
X_k can reach: highest as $1+k$, with prob $= 0.5^k$
lowest as $1-k$ with prob $= 0.5^k$

$\Rightarrow$ ~~$X_k \in$~~ 100% conf interval is $[1-k, 1+k]$

5. (15 points) Consider the following model with uniform noise:



$$Y_k = \beta Y_{k-1} + \epsilon_k, \quad \text{with } \epsilon_k \overset{\text{iid}}{\sim} \texttt{Unif}(-2, 2) \text{ and } Y_0 = 2,$$

where β is the unknown parameter we want to estimate, and $\texttt{Unif}(a, b)$ is the uniform distribution for $a < b$, i.e., $X \sim \texttt{Unif}(a, b) \Rightarrow f_X(x) = \frac{1}{b-a}$ for $x \in [a, b]$; and 0 otherwise. In the following questions, you are required to perform one step of Bayesian parameter estimation assuming the following prior uniform distribution for β, i.e., $\beta \sim \texttt{Unif}(0, 2)$.

(a) Compute the likelihood function $\mathscr{L}(\beta; y_0 = 2, y_1 = 0)$.

$\mathcal{L}(\beta; y_0 = 2, y_1 = 0)$ is where we only observe $Y_0 = 2$

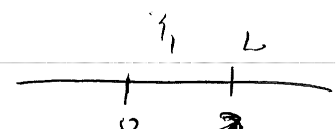
$= \cancel{E[Y_k \mid Y_0 = 2]} =$

$\beta \sim \text{Unif}(0, 2)$

$\Rightarrow Y_k \sim \text{Unif}(-2+\beta, 2+\beta)$

$\mathcal{L}(\beta; y_0 = 2, y_1 = 0) = \text{Unif}(-2, 4)$

(b) Compute the posterior distribution $f_{\beta|\mathcal{F}_1}$ given $\mathcal{F}_1 = \{Y_0 = 2, Y_1 = 0\}$.



Based on (a), we now observe $Y_1 = 0$

$$f_{\beta|F_1} = \begin{cases} \ldots, & \ldots < \beta \sim \text{Unif}(0.5, 1.5) \\ 0, & \text{otherwise} \end{cases}$$

(c) Interpret the implication of $f_{\beta|\mathcal{F}_1}$ on the possible values of β.

$f_{\beta|F_1}$ means β is uniformly sample on $[0, 2]$, with Y_0 & Y_1 fixed the ~~observati~~ Performing

$y_k^1 = \frac{1}{2} y_{k-1}^1 + \frac{1}{3} y_{k-2}^2 + n_k^1$

$y_k^2 = \frac{1}{2} y_{k-1}^2 + \frac{1}{3} y_{k-1}^1 + n_k^2$

6. (15 points) Consider the following VAR(2) process:

$$\begin{bmatrix} y_k^1 \\ y_k^2 \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{bmatrix} \begin{bmatrix} y_{k-1}^1 \\ y_{k-1}^2 \end{bmatrix} + \begin{bmatrix} 0 & \frac{1}{3} \\ \frac{1}{3} & 0 \end{bmatrix} \begin{bmatrix} y_{k-2}^1 \\ y_{k-2}^2 \end{bmatrix} + \begin{bmatrix} n_k^1 \\ n_k^2 \end{bmatrix},$$

where $n_k^1 \sim \mathcal{N}(0,1)$, $n_k^2 \sim \mathcal{N}(1,2)$, and initial conditions are given by $y_0^1 = 1$, $y_0^2 = 2$, $y_{-1}^1 = 2$, $y_{-1}^2 = 1$.

We would like to build an equivalent VAR(1) process of the form:

$$\mathbf{X}_k = A\mathbf{X}_{k-1} + \varepsilon_k.$$

(a) What should be the dimension of the vector $\mathbf{X}_k$ in the equivalent VAR(1) process?

~~2x1~~ 4x1

(b) Explain how you would assign values to the entries of the vector $\mathbf{X}_k$.

$$X_k = \begin{bmatrix} y_k^1 \\ y_{k-1}^1 \\ y_k^2 \\ y_{k-1}^2 \end{bmatrix} \quad \text{as} \quad X_{k-1} = \begin{bmatrix} y_{k-1}^1 \\ y_{k-2}^1 \\ y_{k-1}^2 \\ y_{k-2}^2 \end{bmatrix} \qquad \mathcal{E}_k = \begin{bmatrix} n_k^1 \\ 0 \\ n_k^2 \\ 0 \end{bmatrix}$$

(c) Provide an expression for the autoregressive matrix A in the VAR(1) process.

$$\cancel{A = \begin{bmatrix} \frac{1}{2} & 0 & 0 & \frac{1}{3} \\ 0 & 0 & 0 & 0 \\ 0 & \frac{1}{3} & 0 & \frac{1}{2} \\ 0 & 0 & 0 & 0 \end{bmatrix}} \qquad A = \begin{bmatrix} \frac{1}{2} & 0 & 0 & \frac{1}{3} \\ 0 & 0 & 0 & 0 \\ 0 & \frac{1}{3} & 0 & \frac{1}{2} \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

(d) Provide the mean μ and covariance matrix Σ of the noise vector ε_k in the VAR(1) process.

$$\vec{\mu} = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} \qquad \Sigma = \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}$$

(e) Provide the initial condition $\mathbf{X}_0$ for the VAR(1) process.

$$X_0 = \begin{bmatrix} 1 \\ 2 \\ 2 \\ 1 \end{bmatrix}$$

7. (15 points) Consider a second-order Markov chain modeling the daily weather with two possible states, $\mathcal{S} = \{s_1 = \texttt{Hot}, s_2 = \texttt{Cold}\}$. To transform this process into a first-order Markov chain, we introduce a set of *compound states*, as defined as follows:

$$\mathcal{S}^2 = \{(\texttt{Hot},\texttt{Hot}), \quad (\texttt{Hot},\texttt{Cold}), \quad (\texttt{Cold},\texttt{Hot}), \quad (\texttt{Cold},\texttt{Cold})\},$$

$(H, H) \to (*, H)$

where the pairs of states are temporally ordered, i.e., the first entry in the compound state represents the weather the day before the second entry. In what follows, we analyze the resulting first-order Markov chain with these four compound states.

(a) The transition matrix of the lifted first-order Markov chain is given by:

$$P = \begin{bmatrix} * & * & 0.5 & * \\ 0.5 & * & * & * \\ * & 0.5 & * & * \\ * & * & * & 0.5 \end{bmatrix}.$$

$$\begin{bmatrix} 0.5 & 0 & 0.5 & 0 \\ 0.5 & 0 & 0 & 0.5 \\ 0 & 0.5 & 0.5 & 0 \\ 0 & 0.5 & 0 & 0.5 \end{bmatrix}$$

Complete the entries marked with $*$, ensuring the matrix remains stochastic. *Hints*: Recall that a stochastic matrix must have non-negative entries and each row must sum to 1. Additionally, certain entries are constrained to be zero due to the structure of the augmented Markov chain.

$$P = \begin{bmatrix} 0 & 0.5 & 0.5 & 0 \\ 0.5 & 0 & 0 & 0.5 \\ 0 & 0.5 & 0 & 0.5 \\ 0.5 & 0 & 0 & 0.5 \end{bmatrix}$$

in order to ensure

① no overlap (same) row

② each row sum to 1

(b) Given that the augmented Markov chain starts in state $X_0 = (\texttt{Hot},\texttt{Hot})$, compute the probability that $X_2 = (\texttt{Cold},\texttt{Hot})$ after two time steps. *Hint*: $\pi_k = (P^\top)^k \pi_0$.

$X_0 (H, H) \to X_1 (H, H) \to X_2 (C, H)$

only 1 way

$P = 0.5 \times 0.5$

$= 0.25$

(c) Starting from $X_0 = (\texttt{Hot}, \texttt{Hot})$, compute the distribution vector π_2, which represents the state probabilities at time step 2. *Hint*: $\pi_k = (P^\top)^k \pi_0$.

$\pi_2 =$

$X_0 = (H, H)$

$H_1 = (H,H)$ or $H_1 (C,H)$

$H_2 (H,H)$ $H (C,H)$ $H (H,C)$ fail

(d) In the original second-order Markov chain, what is the probability that a day is Hot when the previous two days are Hot?

$(H,H)(H,H) \to (H,H)$

$P_{\text{obs}}(X_2 = H \mid X_1 = H, X_0 = H)$

$= P(X_2 = H) = 0.5$

because markov chains depends on and only on current state