DS-GA 1018: Homework 2

Due Monday October 15$^{\text{th}}$ at 5:00 pm

Problem 1 (20 points): Consider a causal AR(2) process of the form:

$$X_t = \phi_2 X_{t-2} + W_t, \tag{1}$$

with $0 < \phi_2 < 1$ and $W_t \sim \mathcal{N}(0, \sigma_w^2)$.

i. (6 points): Assume that we have observations $\{x_1, x_2\}$. Derive the mean and variance of a future observation x_t with $t > 2$. (*Hint: you'll need your solutions from Problem 4 of Homework 1.*)

The **mean** can be derived using the recurrence relation w.r.t odd terms and even terms and we could get:

$$\mathbb{E}[X_t] = \begin{cases} \phi_2^{\frac{t-1}{2}} x_1 & \text{t is odd} \\ \phi_2^{\frac{t}{2}-1} x_2 & \text{t is even} \end{cases}$$

For the **variance**, we can solve for the recurrence relation w.r.t odd terms and event terms and we could get:

$$\mathbb{V}[X_t] = \begin{cases} \sigma_W^2 \sum_{m=0}^{\frac{t-3}{2}} \phi_2^{2m} & \text{t is odd} \\ \sigma_W^2 \sum_{m=0}^{\frac{t}{2}-2} \phi_2^{2m} & \text{t is even} \end{cases}$$

ii. (2 points): What is the mean for $t = 3$ and $t = 4$. Explain the intuition behind this result.

We fit into the formula given in problem 1.1 and get $\mathbb{E}[X_3] = \phi_2 x_1$ and $\mathbb{E}[X_4] = \phi_2 x_2$. The intuition would be **The mean of future time steps in an AR(2) process is directly influenced by past values(2-step earlier in particular), with earlier observations playing a key role in determining future outcomes.** But as time grows, the influence from the past observations decays.

iii. (2 points): What is the covariance for $t = 3$ and $t = 4$. Explain the intuition behind this result.

We fit into the formula given in problem 1.1 and get $\mathbb{V}[X_3] = \sigma_W^2$ and $\mathbb{V}[X_4] = \sigma_W^2$ and the covariance between X_3 and X_4 is $Cov[X_3, X_4|x_1, x_2] = Cov[W_3, W_4] = 0$. The intuition would be **The variance of future time steps will be increasingly influenced past values. At first, the variance are just some white noise, but as time grows, variance will become bigger until convergence(if $0 < \phi_2 < 1$).** Also since X_3 and X_4 only depend on constants X_1 and X_2 and the randomness only comes from noisy terms W_3 and W_4, which produces zero covariance.

iv. (2 points): What is the mean and variance of x_t as $t \to \infty$? Explain the intuition behind this result.

From problem 1.1, we know that the **mean** will decay to **0**.
For the **variance**, we have $\forall h > 0$:

$$\mathbb{E}\left[X_t x_{t-h}\right] = \phi_2 \mathbb{E}\left[X_{t-2} X_{t-h}\right] + \mathbb{E}\left[W_t X_{t-h}\right]$$

Since we have proved in homework 1 that this process is causal, so by independency we have

$$\gamma(h) = \phi_2 \gamma(h-2) \quad \forall h > 0$$

Moreover, when $h = 0$, we will have

$$\mathbb{E}\left[X_t^2\right] = \phi_2 \mathbb{E}\left[X_{t-2} X_t\right] + \mathbb{E}\left[W_t X_t\right] \tag{2}$$

$$\gamma(0) = \phi_2 \gamma(-2) + \sigma_W^2 \tag{3}$$

Since we know $\gamma(h) = \gamma(-h)$, we have the following equations:

$$\begin{cases} \gamma(2) = \phi_2 \gamma(0) \\ \gamma(0) = \phi_2 \gamma(-2) + \sigma_W^2 = \phi_2 \gamma(2) + \sigma_W^2 \end{cases}$$

Solving for this we will get $\gamma(0) = \frac{\sigma_W^2}{1-\phi_2^2} = \sigma_W^2(1 + \phi_2^2 + \phi_2^4 + \cdots)$ as $t \to \infty$.
Alternatively, using the formula from problem 1.1 we know that as $t \to \infty$, the variance is just $\mathbb{V}[X_t] = \sigma_W^2(1 + \phi_2^2 + \phi_2^4 + \cdots) = \frac{\sigma_W^2}{1-\phi_2^2}$.

v. (6 points): Assume that we have observations x_1. Derive the mean vector and covariance matrix of a future set of observations $\{x_t, x_{t+1}\}$ with $t > 1$.

The **mean vector** can be derived using the recurrence relation w.r.t odd terms and even terms and referring to question 1.1 could get:

$$\mathbb{E}\begin{bmatrix} X_t \\ X_{t+1} \end{bmatrix} = \begin{cases} \begin{bmatrix} \phi_2^{\frac{t-1}{2}} x_1 \\ 0 \end{bmatrix} & \text{t is odd} \\ \begin{bmatrix} 0 \\ \phi_2^{\frac{t}{2}} x_1 \end{bmatrix} & \text{t is even} \end{cases}$$

where $t > 1$. The reason why $\mathbb{E}[X_t|x_1] = 0$ is that we only have one observation so that $\mathbb{E}[X_2|x_1] = \phi_2 \mathbb{E}[X_0] + 0 = 0$. As $t \to \infty$ we have $\mathbb{E}\begin{bmatrix} X_t \\ X_{t+1} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$.
For the **covariance matrix**, we can solve for the recurrence relation w.r.t odd terms and event terms and we could get the following:
When t is even,

$$\begin{aligned} \operatorname{Var}[X_2 \mid x_1] &= \phi_2^2 \operatorname{Var}[X_0] + \sigma_w^2 && (4) \\ &= \phi_2^2 \gamma(0) + \sigma_w^2 && (5) \\ &= \phi_2^2 \left(\frac{\sigma_w^2}{1-\phi_2^2} \right) + \sigma_w^2 && (6) \\ &= \frac{\sigma_w^2}{1-\phi_2^2} && (7) \end{aligned}$$

and $\mathbb{V}[X_t|x_1] = \frac{\sigma_w^2}{1-\phi_2^2}$ when $t = 2k (k \geq 1)$, thus:

$$\mathbf{Cov}\begin{bmatrix} X_t \\ X_{t+1} \end{bmatrix} = \begin{cases} \begin{bmatrix} \sigma_W^2 \sum_{m=0}^{\max\{\frac{t-3}{2},0\}} \phi_2^{2m} & 0 \\ 0 & \frac{\sigma_W^2}{1-\phi_2^2} \end{bmatrix} & \text{t is odd} \\ \begin{bmatrix} \frac{\sigma_W^2}{1-\phi_2^2} & 0 \\ 0 & \sigma_W^2 \sum_{m=0}^{\max\{\frac{t}{2}-2,0\}} \phi_2^{2m} \end{bmatrix} & \text{t is even} \end{cases}$$

where $t > 1$. The reason why $Cov(X_t, X_{t+1}) = 0$ is that we can derive the expression for $\gamma(h)$, which is

$$r(h) = c_1 \cdot 0.4^h + c_2 \cdot (-0.4)^h$$

and that from problem 1.4 we know the initial condition:

$$\begin{cases} \gamma(1) = 0 \\ \gamma(0) = \frac{\sigma_W^2}{1-\phi_2^2} \end{cases}$$

where it gives $c_1 = c_2 = \frac{\sigma_W^2}{2-2\phi_2^2}$, thus when $h = 1$, $\gamma(h) = 0$, indicating that the $\mathbf{Cov}(X_t, X_{t+1}) = 0, \forall t > 0$.

vi. (2 points): What is the mean vector and covariance matrix of $\{x_t, x_{t+1}\}$ for $t = 2$? Explain the intuition behind this result.

Fitting in the value we get:

1. **Mean Vector**: $\mathbb{E}\begin{bmatrix} X_2 \\ X_3 \end{bmatrix} = \begin{bmatrix} 0 \\ \phi_2 x_1 \end{bmatrix}$

2. **Covariance Matrix**: $\begin{bmatrix} \frac{\sigma_w^2}{1-\phi_2^2} & 0 \\ 0 & \sigma_W^2 \end{bmatrix}$

The intuition would be that:

1. For X_2, since making its prediction relies on x_0 which is unobserved, so the best estimate for mean is just the mean of X_0 and the mean of noisy term, which is also zero. The variance of X_2 includes the variance of X_0(which is huge since it is unobserved) and that of the noisy term.

2. For X_3, since it relies on x_1 and we have observed it. so the variance of its prediction is low compared with X_2. The mean of X_3 directly comes from $\phi_2 x_1$(since the mean of noise term is zero).

3. For the covariance between X_2 and X_3. Since the prediction of X_2 and X_3 follows totally different path, for X_2 it is $X_0 \to X_2$, for X_3 it is $X_1 \to X_3$, so they are independent, which confirms that the covariance is 0 given x_1.

Problem 2 (10 points): Consider the generalization of ARCH(1) model given by:

$$R_t = \delta + Y_t \tag{8}$$
$$Y_t = \sigma_t W_t, \quad W_t \sim N(0,1) \tag{9}$$
$$\sigma_t^2 = \alpha_0 + \alpha_1 Y_{t-1}^2, \tag{10}$$

where $\alpha_0 > 0$, $1 > \alpha_1 > 0$, and δ is a constant value.

i. (2 points): Derive the mean μ_{R_t}.

$\mathbb{E}[Y_t] = \mathbb{E}[\sqrt{\alpha_0 + \alpha_1 \cdot Y_{t-1}^2} W_t] = \mathbb{E}[W_t]\mathbb{E}[\sqrt{\alpha_0 + \alpha_1 \cdot Y_{t-1}^2}] = 0$ since Y_{t-1} is only related to the W_m where $m \leq t-1$ and that $W_x \perp W_y, \forall x \neq y$.
So $\mathbb{E}[R_t] = \delta + \mathbb{E}[Y_t] = \delta$.

ii. (6 points): Derive the covariance $\gamma_{R_t,R_{t+h}}$.

When $h > 0$ we have:

$$\begin{aligned}
\gamma_{R_t,R_{t+h}} &= \mathbb{E}[(R_t - \delta)(R_{t+h} - \delta)) \\
&= \mathbb{E}[Y_t Y_{t+h}] \\
&= \mathbb{E}[\sigma_t W_t \sigma_{t+h} W_{t+h}] \\
&= \mathbb{E}[W_t]\mathbb{E}[\sigma_t \sigma_{t+h} W_{t+h}] \\
&= 0
\end{aligned}$$

where we rely on the fact that $W_x \perp W_y, \forall x \neq y$.
When $h = 0$ we have:

$$\begin{aligned}
\gamma_{R_t,R_t} &= \mathbb{E}[Y_t^2] - \mathbb{E}[Y_t]^2 \\
&= \mathbb{E}[(\delta + Y_t)(\delta + Y_t)] - 0 \\
&= \mathbb{E}[Y_t^2] = \mathbb{E}[\alpha_0 + \alpha_1 Y_{t-1}^2] \\
&= \alpha_0 + \alpha_1 \mathbb{E}[Y_{t-1}^2]
\end{aligned}$$

Solving the recurrence we have

$$\mathbb{V}[R_t] = \mathbb{E}[Y_t]^2 = \frac{\alpha_0}{1 - \alpha_1}$$

given that $0 < \alpha_1 < 1$. Thus we have $\gamma_{R_t,R_{t+h}} = \begin{cases} \frac{\alpha_0}{1-\alpha_1} & h = 0 \\ 0 & h \neq 0 \end{cases}$

iii. (2 points): Is R_t a (weak) stationary process? Justify your answer quantitatively.

Yes, R_t **is a weak stationary process.** From the previous questions, we know that the mean is constant ,the aurocorrelation/covariance function only depends on h and the variance is finite($\frac{\alpha_0}{1-\alpha_1}$), so the process is weakly stationary.

Problem 3 (10 points): Consider the latent space model we presented in class defined by:

$$\mathbf{z}_t = \mathbf{A}\mathbf{z}_{t-1} + \mathbf{w}_t \tag{11}$$
$$\mathbf{x}_t = \mathbf{C}\mathbf{z}_t + \mathbf{v}_t \tag{12}$$

where the latent space $\mathbf{z}$ is has dimension d and the data $\mathbf{x}$ has dimension n. Our noise is being drawn from $\mathbf{w}_t \sim \mathcal{N}(0, \mathbf{Q})$ and $\mathbf{v}_t \sim \mathcal{N}(0, \mathbf{R})$.

i. (4 points): Assume that, $n = d$ and that we have $C = \alpha\mathbb{I}$ and $R = \sigma_v^2\mathbb{I}$. Write the mean $\mu_{t|t} = \mu_{\mathbf{z}_t|\mathbf{z}_{t-1},\mathbf{x}_t}$ and covariance $\Sigma_{t|t}$ in terms of the mean $\mu_{t|t-1} = \mu_{\mathbf{z}_t|\mathbf{z}_{t-1}}$ and covariance $\Sigma_{t|t-1}$. Simplify as much as possible.

Solution: We know from the lecture that

$$\begin{aligned}
\boldsymbol{\mu}_{t|t} &= \boldsymbol{\mu}_{t|t-1} + \boldsymbol{K}_t \left(\boldsymbol{x}_t - \boldsymbol{C}\boldsymbol{\mu}_{t|t-1}\right) \\
\boldsymbol{\Sigma}_{t|t} &= \boldsymbol{\Sigma}_{t|t-1} - \boldsymbol{K}_t\boldsymbol{C}\boldsymbol{\Sigma}_{t|t-1} \\
\boldsymbol{K}_t &= \boldsymbol{\Sigma}_{t|t-1}\boldsymbol{C}^T \left(\boldsymbol{C}\boldsymbol{\Sigma}_{t|t-1}\boldsymbol{C}^T + \boldsymbol{R}\right)^{-1}
\end{aligned}$$

so we can substitute $C = \alpha\mathbb{I}$ and $R = \sigma_v^2\mathbb{I}$ and get:

1. $$K_t = \alpha\Sigma_{t|t-1}\left(\alpha^2\Sigma_{t|t-1} + \sigma_v^2 I\right)^{-1}$$

2. $$\mu_{t|t} = \mu_{t|t-1} + \alpha\Sigma_{t|t-1}\left(\alpha^2\Sigma_{t|t-1} + \sigma_v^2 I\right)^{-1}\left(x_t - \alpha\mu_{t|t-1}\right)$$

3. $$\Sigma_{t|t} = \Sigma_{t|t-1} - \alpha\Sigma_{t|t-1}\left(\alpha^2\Sigma_{t|t-1} + \sigma_v^2 I\right)^{-1}\alpha\Sigma_{t|t-1}$$

ii. (2 points): What happens in the limit $\sigma_v \to 0$? Demonstrate the answer quantitatively and explain the intuition behind this limit.

Solution: Since $\Sigma_{t|t-1}$ is symmetric, so by spectral theorem we can always diagonalize it to be $\Sigma_{t|t-1} = U\Lambda U^\top$ and in this case we have:

1.
$$\begin{aligned}\mu_{t|t} &= \mu_{t|t-1} + \alpha U \Lambda \operatorname{diag}\left(\frac{1}{\alpha^2\lambda_i + \sigma_v^2}\right) U^T \left(x_t - \alpha\mu_{t|t-1}\right) \\ &= \mu_{t|t-1} + \alpha * \left(\sum_{i=1}^{d} \frac{\lambda_i}{\alpha^2\lambda_i + \sigma_v^2} \vec{u}_i \vec{u}_i^\top\right) \left(x_t - \alpha\mu_{t|t-1}\right)\end{aligned}$$

2.
$$\begin{aligned}\Sigma_{t|t} &= \Sigma_{t|t-1} - \alpha^2 U\Lambda \left(\alpha^2\Lambda + \sigma_v^2 I\right)^{-1} U^T \Sigma_{t|t-1} \\ &= U(\Lambda - \Lambda(\Lambda + \sigma_v^2 \cdot I)^{-1}\Lambda)U^\top \\ &= \sum_{i=1}^{d} \frac{\lambda_i \sigma_v^2}{\lambda_i + \sigma_v^2} \vec{u}_i \vec{u}_i^\top \\ &= \sum_{i=1}^{d} \frac{\lambda_i}{\frac{\lambda_i}{\sigma_v^2} + 1} \vec{u}_i \vec{u}_i^\top\end{aligned}$$

Thus when $\sigma_v \to 0$ we will have:

1. $u_{t|t} \to u_{t|t-1} + \alpha\Sigma_{t|t-1}(\alpha^2\Sigma_{t|t-1})^{-1}(x_t - \alpha\mu_{t|t-1}) = \frac{1}{\alpha}x_t$,
2. $\Sigma_{t|t} \to 0$

. The intuition would be:

1. Perfect Observation: When $\sigma_v^2 \to 0$, the observation noise goes to zero, meaning the observation x_t is considered perfect and noiseless. The Kalman filter will fully trust the observation and set the updated state equal to the observation.
2. No Prior Information: Since the observation is perfect, the filter ignores the prior (the previous state estimate $\mu_{t|t-1}$) and directly uses the observation to update the state.
3. Uncertainty Reduction: The uncertainty (covariance) is drastically reduced, reflecting the high confidence the filter has in the noiseless observation.

iii. (2 points): What happens in the limit $\sigma_v \to \infty$? Demonstrate the answer quantitatively and explain the intuition behind this limit.

Solution: Since $\Sigma_{t|t-1}$ is symmetric, so by spectral theorem we can always diagonalize it to be $\Sigma_{t|t-1} = U\Lambda U^\top$ and in this case we have:

1.
$$\begin{aligned}\mu_{t|t} &= \mu_{t|t-1} + \alpha U \Lambda \operatorname{diag}\left(\frac{1}{\alpha^2\lambda_i + \sigma_v^2}\right) U^T \left(x_t - \alpha\mu_{t|t-1}\right) \\ &= \mu_{t|t-1} + \alpha * \left(\sum_{i=1}^{d} \frac{\lambda_i}{\alpha^2\lambda_i + \sigma_v^2}\vec{u}_i\vec{u}_i^\top\right)\left(x_t - \alpha\mu_{t|t-1}\right)\end{aligned}$$

2.
$$\begin{aligned}\Sigma_{t|t} &= \Sigma_{t|t-1} - \alpha^2 U\Lambda\left(\alpha^2\Lambda + \sigma_v^2 I\right)^{-1} U^T \Sigma_{t|t-1} \\ &= U(\Lambda - \Lambda(\Lambda + \sigma_v^2 \cdot I)^{-1}\Lambda)U^\top \\ &= \sum_{i=1}^{d} \frac{\lambda_i \sigma_v^2}{\lambda_i + \sigma_v^2}\vec{u}_i\vec{u}_i^\top \\ &= \sum_{i=1}^{d} \frac{\lambda_i}{\frac{\lambda_i}{\sigma_v^2} + 1}\vec{u}_i\vec{u}_i^\top\end{aligned}$$

Thus when $\sigma_v \to \infty$ we will have:

1. $K_t \to 0$
2. $u_{t|t} \to u_{t|t-1} + \alpha * \vec{0} = u_{t|t-1}$
3. $\Sigma_{t|t} \to \sum_{i=1}^{d} \frac{\lambda_i}{1}\vec{u}_i\vec{u}_i^\top = \Sigma_{t|t-1}$

. The intuition would be:

1. Observation Noise is Extremely High: When $\sigma_v^2 \to \infty$, the observation noise dominates, meaning that the observations x_t become highly unreliable. The filter essentially ignores the observation because it assumes the observation is too noisy to provide any meaningful information.

2. No Update from Observation: The Kalman gain K_t goes to zero, so the new observation does not influence the updated mean or covariance. The system behaves as if there were no new observation, and the state estimate is entirely based on the prior distribution.

iv. (2 points): What happens in the limit $\alpha \to 0$? Demonstrate the answer quantitatively and explain the intuition behind this limit.

Solution: Based on the analysis before, when $\alpha \to 0$, we will have:

1. $K_t \to 0$
2. $u_{t|t} \to u_{t|t-1} + \alpha * \vec{0} = u_{t|t-1}$
3. $\Sigma_{t|t} \to \Sigma_{t|t-1}$

The intuition would be:

1. As $\alpha \to 0$, the observation model $C = \alpha I$ approaches zero, meaning the observations x_t provide no useful information for updating the estimate of the state.
2. Since the Kalman gain K_t goes to zero, the covariance update no longer incorporates any information from the observation, leaving the updated covariance $\Sigma_{t|t}$ unchanged from the prior covariance $\Sigma_{t|t-1}$.
3. Intuition: The filter becomes purely predictive, relying entirely on the prior distribution (i.e., the previous state estimate) and ignoring any new information from the observations.

Problem 4 (5 points): Consider a modified version of our latent space model that depends on a set of **observed** values $\mathbf{y}_t$ as follows:

$$\mathbf{z}_t = \mathbf{A}\mathbf{z}_{t-1} + \mathbf{B}\mathbf{y}_t + \mathbf{w}_t, \tag{13}$$

otherwise all the other components are identical to those described in Problem 3.

i. (5 points): Derive how this new process changes the filtering step of our Kalman filtering.

The whole process of deriving $\mu_{t|t}$ and $\Sigma_{t|t}$ is the same as the lecture notes where:

$$\begin{aligned}
\boldsymbol{\mu}_{t|t} &= \boldsymbol{\mu}_{t|t-1} + \boldsymbol{K}_t \left(\boldsymbol{x}_t - \boldsymbol{C}\boldsymbol{\mu}_{t|t-1} \right) \\
\boldsymbol{\Sigma}_{t|t} &= \boldsymbol{\Sigma}_{t|t-1} - \boldsymbol{K}_t \boldsymbol{C} \boldsymbol{\Sigma}_{t|t-1} \\
\boldsymbol{K}_t &= \boldsymbol{\Sigma}_{t|t-1} \boldsymbol{C}^T \left(\boldsymbol{C} \boldsymbol{\Sigma}_{t|t-1} \boldsymbol{C}^T + \boldsymbol{R} \right)^{-1}
\end{aligned}$$

The only thing that's changing is the process of predicting $\mu_{t|t-1} = \mu_{z_t|x_{1:t-1}}$ where $\mu_{t|t-1} = A\mu_{t-1|t-1} + By_t$ instead of $\mu_{t|t-1} = A\mu_{t-1|t-1}$ given that y_t is observed value.