1. (a) Why are confidence intervals important? How can they be computed and under what conditions? Provide an example to illustrate their interpretation.

> **Solution:** Confidence intervals (CI) quantify the uncertainty of predictions, offering a range where the true value is expected to lie with a given probability.
>
> For example a forecast with 95% CI [42, 48] implies that the true value is expected within this range 95% of the time.
>
> When the residuals are normally distributed, the confidence interval can be computed with
>
> $$CI = \hat{x}_{t+h|t} \pm z_{1-\alpha/2}\hat{\sigma}_h$$
>
> where $z_{1-\alpha/2}$ is the critical value of the normal distribution and $\hat{\sigma}_h$ is an estimate of the standard deviation of the h-step forecast.
>
> When the residuals are uncorrelated and have constant variance, the confidence interval can be computed using bootstrapping.

(b) When is it more appropriate to use scaled performance metrics versus scale-dependent performance metrics? Explain the differences and provide examples to illustrate scenarios where each is most suitable.

> **Solution:** Scale-Dependent Metrics: MSE, RMSE. Used when comparing errors within the same dataset or when scale is meaningful. For example, forecasting energy consumption in kWh for a single region.
>
> Scaled Metrics: MAPE, SMAPE. Useful for comparing errors across datasets with different scales. For example, comparing sales forecasting errors for regions with different currencies.

(c) How does the approach to splitting time series data differ between classification tasks and forecasting tasks? Explain the reasoning behind these differences and provide an example to illustrate the appropriate splitting method for each task.

> **Solution:** Classification: Random splitting is acceptable as classification is performed on time series level. For example, splitting ECG recordings into random train-test sets for binary classification.
>
> Forecasting: Chronological order must be kept while ensuring future values are not used to predict past values. Splitting is performed using cutoff timesteps. For example, using the first 80% of time-ordered sales data to train a sales forecasting model that is evaluated on the remaining 20%.

(d) The ETS model for Holt-Winters method is a generalization of the ETS model for simple exponential smoothing. Should it then be always preferred? Discuss the trade-offs and the scenarios where each model is more appropriate.

Solution: Holt-Winters can captures trends and seasonality, while Simple Exponential Smoothing cannot.

SES is simpler with fewer parameters to estimate and may underfit, while Holt-Winters is more complex therefore more computationally intensive and risks overfitting.

Thus, selection should be based on the time series properties: use Holt-Winters for complex patterns (presence of trend and seasonality) and SES for simpler cases where the mean is constant.

2. Consider the stochastic process $X_t = -2t + W_t + \frac{1}{2}W_{t-1}$ where $W_t \sim \mathcal{N}(0, \sigma^2)$.

   (a) Compute the mean and auto-covariance and determine whether X_t is stationary.

Solution:

- **Mean**

$$\begin{aligned} E[X_t] &= E[-2t + W_t + \frac{1}{2}W_{t-1}] \\ &= E[-2t] + E[W_t] + \tfrac{1}{2}E[W_{t-1}] \\ &= -2t + 0 + 0 = -2t. \end{aligned}$$

  Since $E[X_t]$ depends on t, the mean is time-dependent.

- **Auto-covariance** Define $Y_t = W_t + \frac{1}{2}W_{t-1}$. Notice that the constant term $-2t$ does not affect covariances, so

$$\begin{aligned} cov(X_{t+h}, X_t) &= cov(-2(t+h) + Y_{t+h}, -2t + Y_t) = cov(Y_{t+h}, Y_t) \\ &= cov(W_{t+h} + \tfrac{1}{2}W_{t+h-1}, W_t + \tfrac{1}{2}W_{t-1}) \\ &= cov(W_{t+h}, W_t) + \tfrac{1}{2}cov(W_{t+h}, W_{t-1}) \\ &+ \tfrac{1}{2}cov(W_{t+h-1}, W_t) + \tfrac{1}{4}cov(W_{t+h-1}, W_{t-1}) \\ &= \begin{cases} \frac{5}{4}\sigma^2, & h = 0, \\ \frac{1}{2}\sigma^2, & h = \pm 1, \\ 0, & \text{otherwise.} \end{cases} \end{aligned}$$

- **Stationarity** A process is (weakly) stationary if its mean and auto-covariance are independent of time. Here, although $cov(X_{t+h}, X_t)$ depends only on h, the mean $E[X_t] = -2t$ is time-dependent. Therefore, X_t is *not* stationary.

   (b) Transform the process to make it stationary without using decomposition methods, and compute the mean and auto-covariance of the resulting process.

Solution: This time series has a linear trend. We apply first-order differencing.

$$\begin{aligned}\nabla X_t = X_t - X_{t-1} &= -2t + W_t + \frac{1}{2}W_{t-1} - (-2(t-1) + W_{t-1} + \frac{1}{2}W_{t-2}) \\ &= -2 + W_t - \frac{1}{2}W_{t-1} - \frac{1}{2}W_{t-2}\end{aligned}$$

- **Mean**

$$\begin{aligned}E[\nabla X_t] &= E[-2 + W_t - \frac{1}{2}W_{t-1} - \frac{1}{2}W_{t-2}] \\ &= E[-2] + E[W_t] - \tfrac{1}{2}E[W_{t-1}] - \tfrac{1}{2}E[W_{t-2}] \\ &= -2 + 0 - 0 - 0 = -2.\end{aligned}$$

  $E[X_t]$ is constant and therefore, independent of time.

- **Auto-covariance** Define $Y_t = W_t - \frac{1}{2}W_{t-1} - \frac{1}{2}W_{t-2}$. Notice that the constant term -2 does not affect covariances, so

$$\begin{aligned}cov(\nabla X_{t+h}, \nabla X_t) &= cov(-2 + Y_{t+h}, -2 + Y_t) = cov(Y_{t+h}, Y_t) \\ &= cov(W_{t+h} - \tfrac{1}{2}W_{t+h-1} - \tfrac{1}{2}W_{t+h-2}, W_t - \tfrac{1}{2}W_{t-1} - \tfrac{1}{2}W_{t-2}) \\ &= cov(W_{t+h}, W_t) - \tfrac{1}{2}cov(W_{t+h}, W_{t-1}) - \tfrac{1}{2}cov(W_{t+h}, W_{t-2}) \\ &- \tfrac{1}{2}cov(W_{t+h-1}, W_t) + \tfrac{1}{4}cov(W_{t+h-1}, W_{t-1}) + \tfrac{1}{4}cov(W_{t+h-1}, W_{t-2}) \\ &- \tfrac{1}{2}cov(W_{t+h-2}, W_t) + \tfrac{1}{4}cov(W_{t+h-2}, W_{t-1}) + \tfrac{1}{4}cov(W_{t+h-2}, W_{t-2}) \\ &= \begin{cases} \frac{3}{2}\sigma^2, & h = 0, \\ -\frac{1}{4}\sigma^2, & h = \pm 1, \\ -\frac{1}{2}\sigma^2, & h = \pm 2, \\ 0, & \text{otherwise.} \end{cases}\end{aligned}$$

- **Stationarity** The differenced process is (weakly) stationary since its mean and auto-covariance are independent of time.

(c) Plot the theoretical auto-correlation function of both process and discuss any differences.

Solution: We compute the autocorrelation function directly, using

$$\rho(h) = \frac{\gamma(h)}{\gamma_(0)}$$

Hence,

$$\rho_X(h) = \begin{cases} 1, & h = 0, \\ \frac{2}{5}, & h = \pm 1, \\ 0, & |h| \geq 2. \end{cases}$$

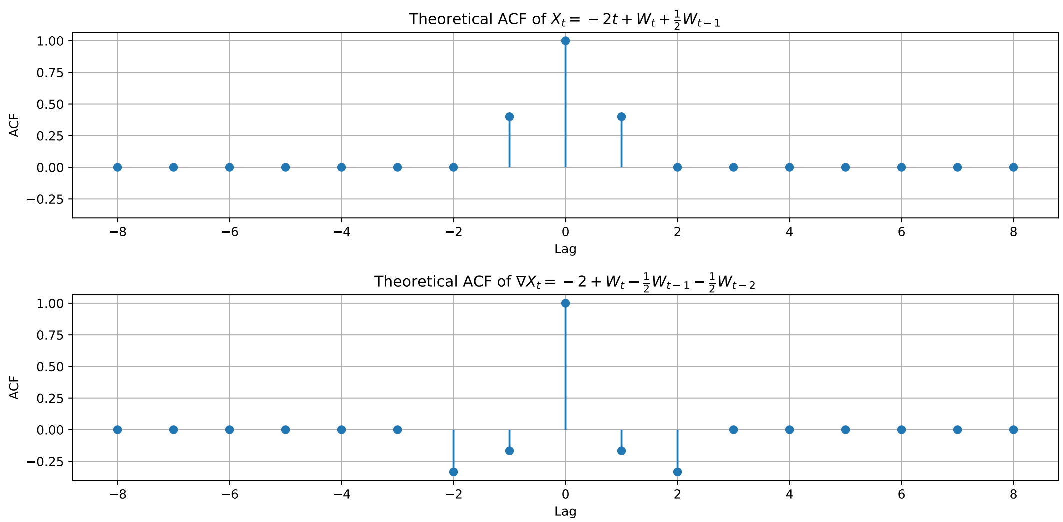
Only overlapping white-noise indices contribute to the covariance; normalization by $\gamma(0)$ gives a symmetric ACF with positive short-lag correlation.

For the differenced process we get

$$\rho_{\nabla X}(h) = \begin{cases} 1, & h = 0, \\ -\frac{1}{6}, & h = \pm 1, \\ -\frac{1}{3}, & h = \pm 2, \\ 0, & |h| \geq 3. \end{cases}$$

Differencing removes deterministic trends but can introduce artificial autocorrelation even when the original noise sequence is uncorrelated. This happens because each differenced value shares common noise terms with its neighbors (e.g., W_t appears in both ∇X_t and ∇X_{t+1}), creating mechanical dependencies not present in the underlying process. These correlations are purely a by-product of the differencing operation, not genuine temporal dependence in the data-generating process. Hence, while differencing ensures stationarity, it can distort the short-term correlation structure and should be applied with care.

This is visible in the following figure:



3. Consider the ARMA model $X_t = X_{t-1} - \frac{1}{4}X_{t-2} + W_t - \frac{1}{4}W_{t-1}$ where $W_t \sim \mathcal{N}(0, \sigma^2)$.

   (a) Determine the order (p,q) of the model and derive the characteristic polynomials for the autoregressive (AR) and moving average (MA) components.

Solution: The model consists of both autoregressive (AR) and moving average (MA) components:

- The AR terms involve X_{t-1} and X_{t-2}, indicating an autoregressive order of $p = 2$.
- The MA terms involve W_t and W_{t-1}, indicating a moving average order of $q = 1$.

Hence, the model is an ARMA$(2, 1)$ process.

- **AR characteristic polynomial:**

$$X_t - X_{t-1} + \frac{1}{4}X_{t-2} = W_t - \frac{1}{4}W_{t-1}.$$

This corresponds to the characteristic equation:

$$\phi(B) = 1 - B + \frac{1}{4}B^2.$$

- **MA characteristic polynomial:**

$$W_t - \frac{1}{4}W_{t-1}.$$

Thus, the MA characteristic polynomial is:

$$\theta(B) = 1 - \frac{1}{4}B.$$

(b) Assess whether the model is stationary.

Solution: To assess stationarity, we examine the autoregressive (AR) characteristic polynomial. A necessary and sufficient condition for stationarity in an ARMA model is that the roots of the characteristic equation must lie outside the unit circle in the complex plane.

Rearrange the equation:

$$\frac{1}{4}B^2 - B + 1 = 0.$$

Applying the quadratic formula $B = \frac{-b\pm\sqrt{b^2-4ac}}{2a}$, where $a = \frac{1}{4}$, $b = -1$, and $c = 1$,

$$B = \frac{1 \pm \sqrt{(-1)^2 - 4 \cdot \frac{1}{4} \cdot 1}}{2 \cdot \frac{1}{4}} = \frac{1 \pm \sqrt{1-1}}{\frac{1}{2}} = \frac{1 \pm 0}{\frac{1}{2}} = 2.$$

The only root is $B = 2$, which is outside the unit circle ($|B| > 1$).

Since all roots of the characteristic equation lie outside the unit circle, the ARMA(2,1) model is **stationary**.

(c) Evaluate whether the model is invertible and provide an explanation of the concept.

Solution: Invertibility refers to the ability to express a moving average (MA) process uniquely and stably as an infinite autoregressive (AR) process based on past observations. This property is important because it ensures that the current value of the process can be represented in terms of past shocks (innovations) without ambiguity or instability.

In an ARMA model, invertibility applies to the MA part. A necessary and sufficient condition for invertibility is that all roots of the MA characteristic equation lie outside the unit circle in the complex plane.

For the given model, rearrange the MA equation:

$$1 - \tfrac{1}{4}B = 0$$

which gives the root $B = 4$.

Since $|B| > 1$, the root lies outside the unit circle. Therefore, the MA part satisfies the invertibility condition, and the ARMA(2,1) model is **invertible**.

(d) Considering the realization $\{2, \frac{3}{2}, \frac{9}{8}\}$, compute the one, two, three step-ahead forecasts assuming the process starts at $t = 1$ with initial conditions $x_0 = w_0 = 0$.

Solution: The given ARMA(2,1) model is:

$$X_t = X_{t-1} - \frac{1}{4}X_{t-2} + W_t - \frac{1}{4}W_{t-1}.$$

Compute the past noise terms

For $t = 1$:

$$x_1 = x_0 - \frac{1}{4}x_{-1} + w_1 - \frac{1}{4}w_0 \Rightarrow w_1 = 2.$$

For $t = 2$:

$$x_2 = x_1 - \frac{1}{4}x_0 + w_2 - \frac{1}{4}w_1.$$

$$\frac{3}{2} = 2 - \frac{1}{4}(0) + w_2 - \frac{1}{4}(2) \Rightarrow w_2 = 0.$$

For $t = 3$:

$$x_3 = x_2 - \frac{1}{4}x_1 + w_3 - \frac{1}{4}w_2.$$

$$\frac{9}{8} = \frac{3}{2} - \frac{1}{4}(2) + w_3 - \frac{1}{4}(0) \Rightarrow w_3 = \frac{1}{8}.$$

Forecasts

The general forecast equation is

$$\hat{X}_{t+h|t} = X_{t+h-1|t} - \frac{1}{4}X_{t+h-2|t} + \hat{W}_{t+h|t} - \frac{1}{4}\hat{W}_{t+h-1|t},$$

where $\hat{W}_{t+h|t} = 0$ for $h \geq 1$.

(a) One-step-ahead forecast

$$\hat{x}_{4|3} = x_3 - \frac{1}{4}x_2 + 0 - \frac{1}{4}w_3.$$

Substitute values:

$$\hat{x}_{4|3} = \frac{9}{8} - \frac{1}{4}\left(\frac{3}{2}\right) - \frac{1}{4}\left(\frac{1}{8}\right) = \frac{23}{32}.$$

(b) Two-step-ahead forecast

$$\hat{x}_{5|3} = \hat{x}_{4|3} - \frac{1}{4}x_3 + 0 - \frac{1}{4}(0).$$

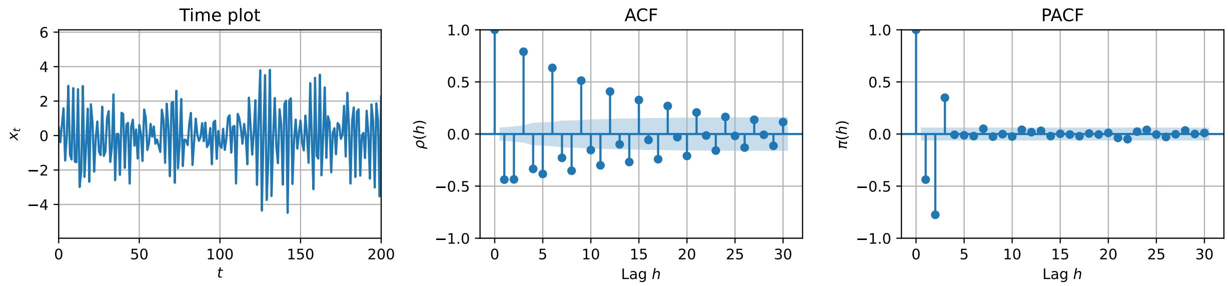
$$\hat{x}_{5|3} = \frac{23}{32} - \frac{1}{4}\left(\frac{9}{8}\right) = \frac{7}{16}.$$

(c) Three-step-ahead forecast

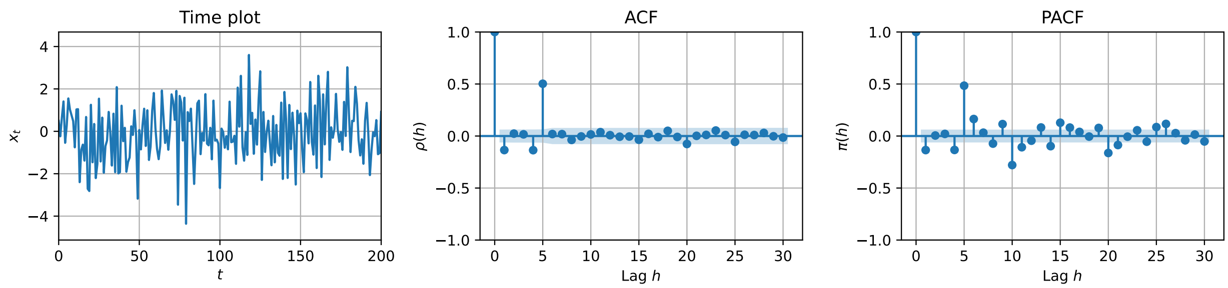
$$\hat{x}_{6|3} = \hat{x}_{5|3} - \frac{1}{4}\hat{x}_{4|3} + 0 - \frac{1}{4}(0).$$

$$\hat{x}_{6|3} = \frac{7}{16} - \frac{1}{4}\left(\frac{23}{32}\right) = \frac{33}{128}.$$

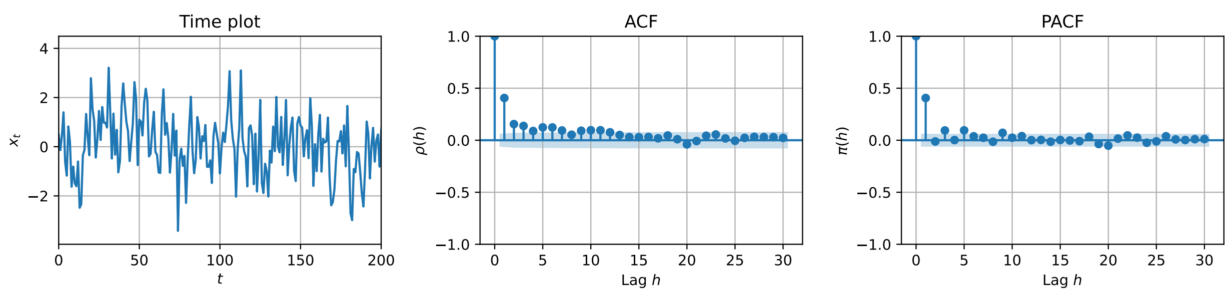
(e) Determine the appropriate order (p,q) for an ARMA model to fit each time series shown in fig. 1, using the provided sample autocorrelation function (ACF) and partial autocorrelation function (PACF). Justify your choices.



(a)



(b)



(c)

Figure 1: Time series realizations.

Solution:

(a) The ACF displays an exponential decay, while the PACF has a cut-off at $h = 3$. This behavior is typical of an AR(3) model. An appropriate model to fit the time series would be an ARMA(3,0).

(b) The ACF has a cut-off at $h = 5$, while the PACF displays an exponential decay. This behavior is typical of an MA(5) model. An appropriate model to fit the time series would be an ARMA(0,5).

(c) The ACF shows a dominant spike at $h = 1$, while the PACF shows a drop-off at $h = 1$ followed by a strong decay pattern. This pattern suggests that both autoregressive and moving average components are present. This is characteristic of an ARMA(1,1) process.

4. Consider the ETS(A,N,N) model defined as

$$x_t = \ell_{t-1} + \epsilon_t,$$

$$\ell_t = \ell_{t-1} + \alpha\epsilon_t,$$

where $\epsilon_t \sim \mathcal{N}(0, \sigma^2)$. The observed values are

$$x_1 = 10, \quad x_2 = 12, \quad x_3 = 13.$$

Assume the initial level is $\ell_0 = 9$.

(a) Compute the level updates ℓ_1, ℓ_2, and ℓ_3 using the smoothing parameter $\alpha = 0.2$.

Solution:

1. **Compute ℓ_1:**
$$\epsilon_1 = x_1 - \ell_0 = 10 - 9 = 1.$$
$$\ell_1 = \ell_0 + \alpha\epsilon_1 = 9 + 0.2 \cdot 1 = 9.2.$$

2. **Compute ℓ_2:**
$$\epsilon_2 = x_2 - \ell_1 = 12 - 9.2 = 2.8.$$
$$\ell_2 = \ell_1 + \alpha\epsilon_2 = 9.2 + 0.2 \cdot 2.8 = 9.2 + 0.56 = 9.76.$$

3. **Compute ℓ_3:**
$$\epsilon_3 = x_3 - \ell_2 = 13 - 9.76 = 3.24.$$
$$\ell_3 = \ell_2 + \alpha\epsilon_3 = 9.76 + 0.2 \cdot 3.24 = 9.76 + 0.648 = 10.408.$$

(b) Repeat the computations for ℓ_1, ℓ_2, and ℓ_3 using $\alpha = 0.8$.

Solution:

1. **Compute ℓ_1:**
$$\epsilon_1 = x_1 - \ell_0 = 10 - 9 = 1.$$
$$\ell_1 = \ell_0 + \alpha\epsilon_1 = 9 + 0.8 \cdot 1 = 9 + 0.8 = 9.8.$$

2. **Compute ℓ_2:**
$$\epsilon_2 = x_2 - \ell_1 = 12 - 9.8 = 2.2.$$
$$\ell_2 = \ell_1 + \alpha\epsilon_2 = 9.8 + 0.8 \cdot 2.2 = 9.8 + 1.76 = 11.56.$$

3. **Compute ℓ_3:**
$$\epsilon_3 = x_3 - \ell_2 = 13 - 11.56 = 1.44.$$
$$\ell_3 = \ell_2 + \alpha\epsilon_3 = 11.56 + 0.8 \cdot 1.44 = 11.56 + 1.152 = 12.712.$$

(c) Compare the previous results and discuss how the choice of the smoothing parameter affects the model behavior.

Solution: Effect of Smoothing Parameter α:

- $\alpha = 0.2$ (small smoothing parameter):
  - The updates to the level ℓ_t are more gradual.
  - The model places greater weight on historical levels and less on the most recent observation.
  - This leads to a smoother, less responsive model.
- $\alpha = 0.8$ (large smoothing parameter):
  - The updates to the level ℓ_t are larger and more responsive to recent observations.
  - The model places more weight on the most recent observation, making it more sensitive to changes in the data.
  - This results in a model that quickly adapts to new information but may also be more volatile.

The choice of α determines the trade-off between responsiveness and stability in the model. A smaller α smooths the level updates, making the model less sensitive to noise or short-term fluctuations. In contrast, a larger α prioritizes recent observations, allowing the model to adapt quickly to changes in the time series, at the expense of stability.

(d) Taking $\alpha = 0.2$, compute the forecasts $\hat{x}_{4|3}$ and $\hat{x}_{10|3}$.

Solution: In the ETS(A,N,N) model, the forecast for any future time $t+h$ is given by:

$$\hat{x}_{t+h|t} = \ell_t.$$

- $\hat{x}_{4|3}$:
$$\hat{x}_{4|3} = \ell_3 = 10.408.$$
- $\hat{x}_{10|3}$:
$$\hat{x}_{10|3} = \ell_3 = 10.408.$$

For the ETS(A,N,N) model, forecasts are constant for all future time steps and equal to the most recent level:

$$\hat{x}_{4|3} = 10.408, \quad \hat{x}_{10|3} = 10.408.$$

(e) The forecast variance of the model is given by:

$$\hat{\sigma}_h^2 = \hat{\sigma}_e^2 \left(1 + \alpha^2(h-1)\right),$$

with $\hat{\sigma}_e^2$ the residuals variance. Calculate the 95% confidence intervals for both forecasts assuming normally distributed residuals. Explain the difference.

Solution: For normal residuals, the 95% confidence intervals is computed as:

$$\mathrm{CI}(\hat{x}_{t+h|t}) = \hat{x}_{t+h|t} \pm 1.96 \cdot \hat{\sigma}_{h|t}.$$

(a) $\hat{x}_{4|3}$:
$$\hat{\sigma}_4^2 = \hat{\sigma}_e^2 \left(1 + \alpha^2(4-1)\right) = \hat{\sigma}_e^2 \left(1 + \alpha^2 \cdot 3\right).$$
$$\mathrm{CI}_{4|3} = 10.408 \pm 1.96 \cdot \sqrt{\hat{\sigma}_e^2 \left(1 + \alpha^2 \cdot 3\right)}.$$

(b) $\hat{x}_{10|3}$:
$$\hat{\sigma}_{10}^2 = \hat{\sigma}_e^2 \left(1 + \alpha^2(10-1)\right) = \hat{\sigma}_e^2 \left(1 + \alpha^2 \cdot 9\right).$$
$$\mathrm{CI}_{10|3} = 10.408 \pm 1.96 \cdot \sqrt{\hat{\sigma}_e^2 \left(1 + \alpha^2 \cdot 9\right)}.$$

Difference:

- The forecast variance grows over time because the observed value is influenced by the accumulation of all intermediate error terms.
- The forecast variance depends on the smoothing parameter α and increases with the forecast horizon h, as the additional term $\alpha^2(h-1)$ grows linearly.

> - For shorter horizons ($h = 4$), the confidence interval is narrower since the impact of the additional term is small.
> - For longer horizons ($h = 10$), the confidence interval is much wider, reflecting the increased uncertainty over time.

(f) An ETS(A,N,N) model was fitted to monthly data with a fixed smoothing parameter α. Interpret each of the residual analysis plots shown in fig. 2. Suggest specific ways to improve performance based on your observations.

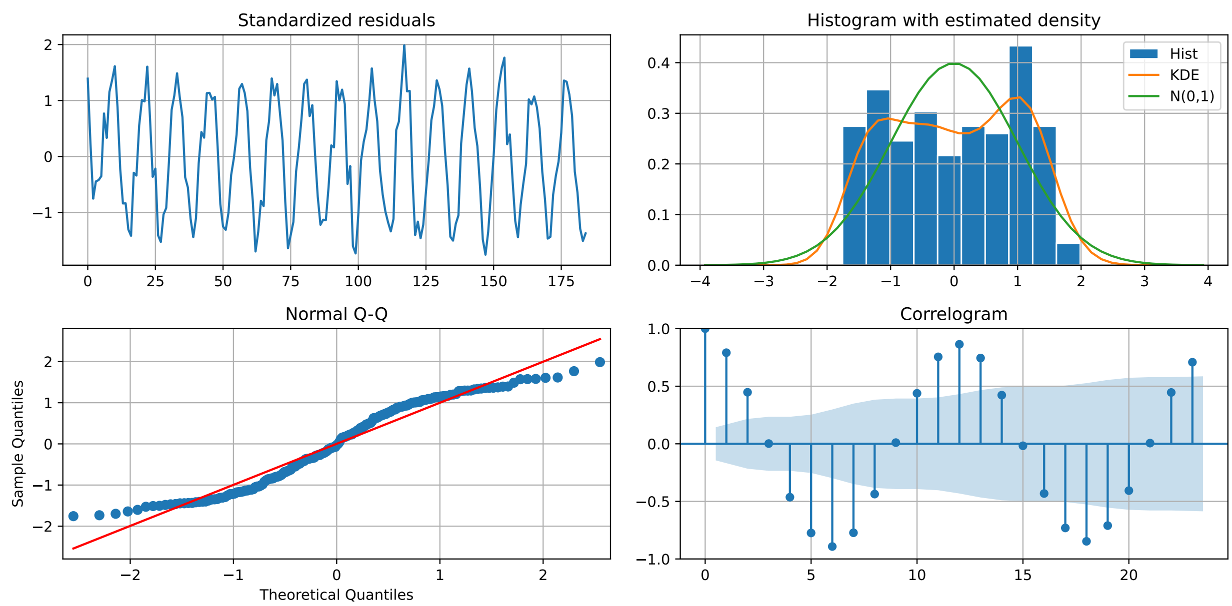


Figure 2: Residual analysis plots.

Solution: Standardized Residuals Plot:

The standardized residuals should exhibit a random scatter around zero, without obvious patterns. If there is a visible trend or seasonality, it suggests that the model is not fully capturing the structure of the data.

Observations: While the residuals fluctuate symmetrically around zero they display regular seasonal pattern indicating that an ETS(A,N,A) would be more suitable.

2. Histogram with Estimated Density:

The histogram should resemble a normal distribution centered at zero. Any skewness or excess kurtosis (fat tails) indicates non-normality.

Observations: The histogram appears bimodal, indicating that the residuals do not follow a normal distribution. This suggests that the model may be misspecified or that the data contain structural breaks or unmodeled regimes. As a result, statistical inference (e.g., confidence intervals and hypothesis tests based on normality) may be unreliable, and alternative models should be considered.

3. Normal Q-Q Plot:

The residuals should lie close to the diagonal reference line if they follow a normal distribution. Deviations in the tails indicate non-normality, which could lead to unreliable prediction intervals.

Observations: The tails clearly diverge from the diagonals. Extreme values occur less frequently than expected under a normal model. As a result, the model's inference and forecast uncertainty estimates may be distorted, and transformations or alternative error distributions may be more appropriate.

4. Correlogram:

Residuals should not exhibit significant autocorrelation. Significant spikes in the autocorrelation function (ACF) suggest that there is structure left in the data, meaning the model does not fully capture the dependence in the series.

Observations: There are significant non-zero autocorrelations. They follows a seasonal pattern suggesting that a simple ETS(A,N,N) model may not be sufficient.

Conclusion:

The residual analysis suggests that the ETS(A,N,N) model does not fully capture all patterns in the data. Further refinement, such as adding trend, seasonality, or using a hybrid approach with ARIMA, may improve the model's performance.