SJSU SAN JOSÉ STATE UNIVERSITY

Time Series Data

Part 2

1

SJSU SAN JOSÉ STATE UNIVERSITY

Agenda

- Importance of Stationarity
- Autocorrelation and Partial Autocorrelation
  - Basics on autocorrelation & partial autocorrelation functions (ACF & PACF)
  - Differences between ACF & PACF
- Time Series Models
  - AR & MA processes
  - ARIMA
  - SARIMA
- Model Evaluation and Forecasting
  - Model evaluation (AIC, BIC etc)
  - Forecasting using ARIMA models

2

SJSU SAN JOSÉ STATE UNIVERSITY

Importance of Stationarity

Predictive Modeling Assumptions

- Many statistical models (e.g., ARIMA, SARIMA) assume stationarity for optimal performance.
- Non-stationary data can cause trend dominance, leading to overfitting and inaccurate forecasts.

Simplified Analysis

- Makes it easier to identify and interpret relationships within the series.
- Consistent statistical inferences over time.

Enhanced Forecast Accuracy

- Stabilizes patterns to achieve better fitting models.
- Reliable parameter estimations and prediction intervals.

Time-Invariant Relationships

- Ensures consistent variable relationships over time, essential for multivariate models like VAR (Vector Autoregression)

3

SJSU SAN JOSÉ STATE UNIVERSITY

Autocorrelation and Partial Autocorrelation

4

SJSU SAN JOSÉ STATE UNIVERSITY

Autocorrelation

- Autocorrelation measures the correlation of a time series with its own past values.

$$r_k = \frac{\sum_{t=k+1}^{n}(X_t - \bar{X})(X_{t-k} - \bar{X})}{\sum_{t=1}^{n}(X_t - \bar{X})^2}$$

Where r_k is the autocorrelation at lag k, X_t is the time series, and $\bar{X}$ is the mean of the series.

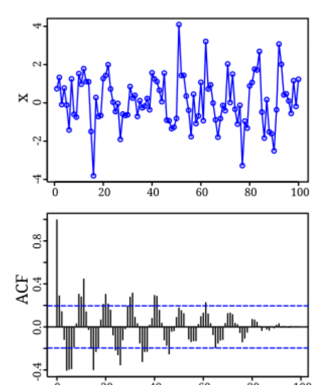
- Quantifies how the current value of the series is related to its lagged (previous) values.
- Used for
  - Identifying repeating patterns, such as seasonality or periodic behavior
  - Diagnosing time-series models like AR, MA, or ARIMA

5

SJSU SAN JOSÉ STATE UNIVERSITY

Autocorrelation Function (ACF)

- Autocorrelation Function (ACF) shows the correlation between a time series and its lagged versions for multiple lags.
- The ACF plot displays the autocorrelation values at different lags.
- Plot helps determine the significant lags that contribute to the structure of the series.
  - A slow decay in ACF suggests possible non-stationarity
  - Significant spikes at specific lags in the ACF can indicate seasonality or lag-based dependencies



6

Partial Autocorrelation Function (PACF)

- Partial Autocorrelation Function (PACF) measures the correlation between a time series and its lagged values, removing the influence of intermediate lags.
- It is a key tool for identifying the order of autoregressive (AR) terms in a model.

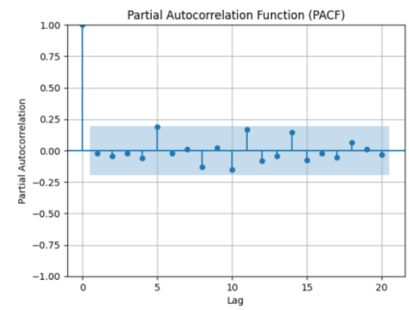
$$X_t = \phi_{k1}X_{t-1} + \phi_{k2}X_{t-2} + \cdots + \phi_{kk}X_{t-k} + \epsilon_t$$

Where:

- X_t: Value of the series at time t.
- ϕ_{kj}: Partial autocorrelation coefficient at lag k.
- ϵ_t: Residual (error term) at time t.

In this equation:

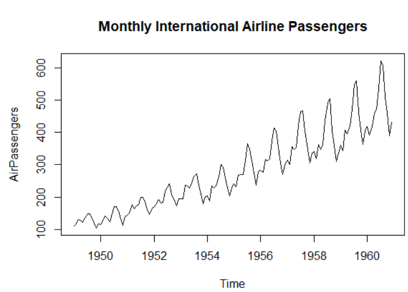
- The ϕ_{kk} term represents the partial autocorrelation at lag k, obtained after removing the influence of the intermediate lags $1, 2, \ldots, k-1$.

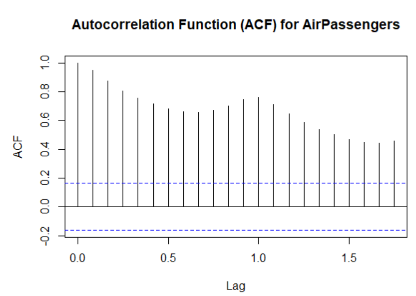


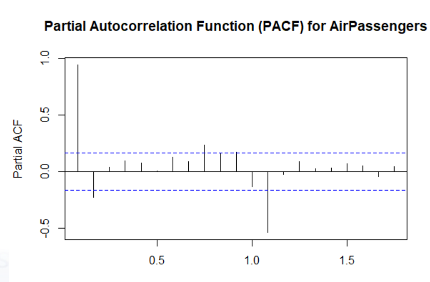
7

ACF vs PACF

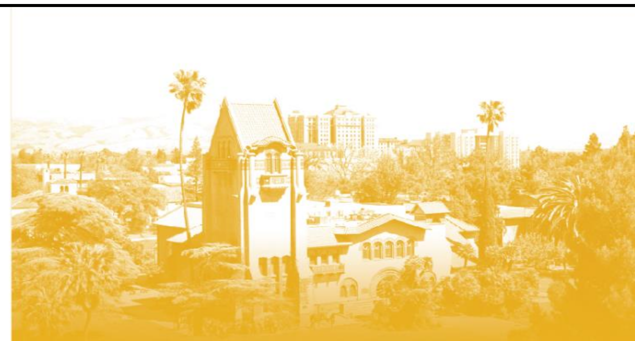
- Monthly International Airline Passengers:







8



Time Series Models

9

Autoregressive Process (AR)

- Model based on the relationship between current value and lagged values:

$$X_t = c + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + \epsilon_t$$

Where:

- X_t: Current value.
- ϕ_p: AR coefficients.
- p: Number of lags.

- This is a stationary process if abs(ϕ) < 1.
- Examples:
  - AR(1) $X_t = c + \phi_1 X_{t-1}$
  - AR(2) $X_t = c + \phi_1 X_{t-1} + \phi_2 X_{t-2}$

10

SJSU SAN JOSÉ STATE UNIVERSITY

Moving Average Process (MA)

- Model based on past error terms (residuals) to refine predictions:

$$X_t = c + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \cdots + \theta_q \epsilon_{t-q}$$

- This is a stationary process regardless of values of θ.
- Examples:
  - MA(1) $X_t = c + \epsilon_t + \theta_1 \epsilon_{t-1}$
  - MA(2) $X_t = c + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2}$

Where:

- ϵ_t: Error term.
- θ_q: MA coefficients.
- q: Number of lags.

11

SJSU SAN JOSÉ STATE UNIVERSITY

Autoregressive Moving Average (ARMA)

Autoregressive (AR)

Model based on the relationship between current value and lagged values

$$X_t = c + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + \epsilon_t$$

Moving Average (MA)

Uses past error terms (residuals) to refine predictions

$$X_t = c + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \cdots + \theta_q \epsilon_{t-q}$$

ARMA Model Equation:

$$X_t = c + \sum_{i=1}^{p} \phi_i X_{t-i} + \epsilon_t + \sum_{j=1}^{q} \theta_j \epsilon_{t-j}$$

Where:

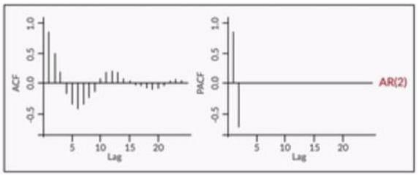
- X_t: Current value of the series.
- ϕ_i: Coefficients of the AR terms.
- θ_j: Coefficients of the MA terms.
- ϵ_t: Error (random noise).
- p: Number of AR terms.
- q: Number of MA terms.

*** ARMA can be used ONLY when the series is stationary**

12

SJSU SAN JOSÉ STATE UNIVERSITY

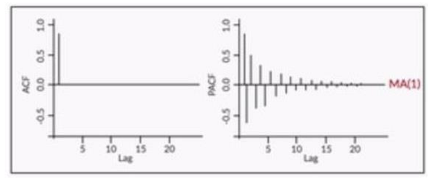
ACF vs PACF Revisited



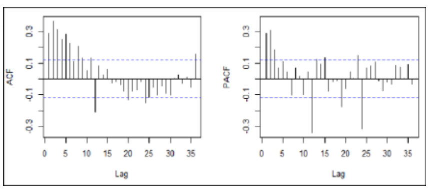
ACF and PACF plots for AR model

Case 1 - Autoregressive (AR) Process
ACF: Decays exponentially
PACF: Sharp cutoff after lag *p* (order of AR process).

Case 2 - Moving Average (MA) Process
ACF: Sharp cutoff after lag *q* (order of MA process).
PACF: Decays exponentially.



ACF and PACF plots for MA model



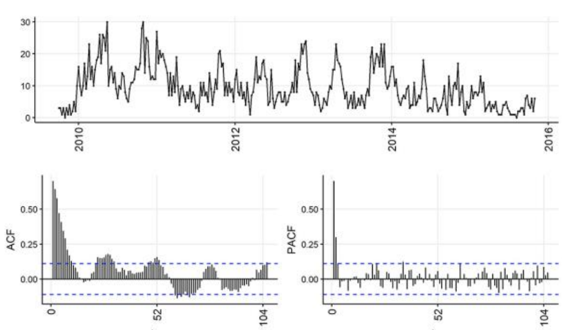
ACF and PACF correlogram for ARMA model

Case 3 - Mixed ARMA Process
ACF and PACF: Both decay gradually without a sharp cutoff

13

SJSU SAN JOSÉ STATE UNIVERSITY

ACF vs PACF Revisited



ACF	PACF
Measures correlation between series and all past lags.	Measures correlation between series and a specific lag, removing intermediate lags' effects.
Identifies patterns, seasonality, and moving average (MA) terms.	Determines the number of autoregressive (AR) terms.
Gradual decay for AR processes.	Sharp cutoff for AR processes.
Diagnosing MA components in ARIMA.	Diagnosing AR components in ARIMA.

14

SJSU SAN JOSÉ STATE UNIVERSITY

Autoregressive Integrated Moving Average (ARMA)

Autoregressive (AR)

Model based on the relationship between current value and lagged values

$$X_t = c + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + \epsilon_t$$

Integrated (I)

Represents the differencing step to make the series stationary

$$X'_t = X_t - X_{t-1}$$

Moving Average (MA)

Uses past error terms (residuals) to refine predictions

$$X_t = c + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \cdots + \theta_q \epsilon_{t-q}$$

ARIMA Model Equation:

$$Y_t = X_t - X_{t-1} \quad (\text{for } d = 1)$$

$$Y_t = c + \sum_{i=1}^{p} \phi_i Y_{t-i} + \epsilon_t + \sum_{j=1}^{q} \theta_j \epsilon_{t-j}$$

When to use ARIMA

- When the series is **non-stationary** but lacks clear seasonality
- Differencing helps remove trends or cyclic behavior.

15

SJSU SAN JOSÉ STATE UNIVERSITY

Seasonal Autoregressive Integrated Moving Average (SARIMA)

Non-Seasonal Components

- $\phi_p(B)$: Non-seasonal autoregressive (AR) operator of order p.
- $\theta_q(B)$: Non-seasonal moving average (MA) operator of order q.
- $(1-B)^d$: Non-seasonal differencing of order d.

Seasonal Components

- $\Phi_P(B^m)$: Seasonal autoregressive (SAR) operator of order P.
- $\Theta_Q(B^m)$: Seasonal moving average (SMA) operator of order Q.
- $(1-B^m)^D$: Seasonal differencing of order D.

Other Terms

- B: Backshift operator (e.g., $BX_t = X_{t-1}$).
- m: Seasonal period (e.g., $m = 12$ for annual seasonality in monthly data).
- ϵ_t: White noise (error term).

Non-Seasonal MA (q):

$$\theta_q(B) = 1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q$$

Non-Seasonal AR (p):

$$\phi_p(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p$$

Seasonal MA (Q):

$$\Theta_Q(B^m) = 1 + \Theta_1 B^m + \Theta_2 B^{2m} + \cdots + \Theta_Q B^{Qm}$$

Seasonal AR (P):

$$\Phi_P(B^m) = 1 - \Phi_1 B^m - \Phi_2 B^{2m} - \cdots - \Phi_P B^{Pm}$$

SARIMA Model Equation

$$\Phi_P(B^m)\phi_p(B)(1-B)^d(1-B^m)^D X_t = \Theta_Q(B^m)\theta_q(B)\epsilon_t$$

where

- (p, d, q): Non-seasonal ARIMA terms.
- (P, D, Q): Seasonal ARIMA terms.
- m: Seasonal period (e.g., $m = 12$ for monthly seasonality).

When to use SARIMA

- When the time series exhibits clear **seasonal patterns** (e.g., monthly sales data)

16

SJSU SAN JOSÉ STATE UNIVERSITY

Time Series Models Summary

	ARMA	ARIMA	SARIMA
Model Notation	*ARMA*(p,q)	*ARIMA*(p,d,q)	*SARIMA*(p,d,q)×(P,D,Q,s)
Description	Combines autoregressive (AR) and moving average (MA) components for stationary series	Extends ARMA by adding differencing to handle non-stationary series	Extends ARIMA by including seasonal components for handling seasonal data
Components	- AR (p): Lagged values of the series. - MA (q): Lagged errors.	- AR (p): Lagged values of the series. - MA (q): Lagged errors. - Differencing (d): To achieve stationarity.	- Non-seasonal: p,d,q. - Seasonal: P,D,Q,s (e.g., quarterly data: s=4).
Stationarity	Series must be stationary.	Handles non-stationary series through differencing (d).	Handles both non-stationary and seasonal series through seasonal differencing (D)
Seasonality	Does not model seasonality	Cannot explicitly model seasonality	Explicitly models seasonality with seasonal terms (P,D,Q,s)
Complexity	Simpler	Moderate complexity	Higher complexity due to seasonal parameters
Data Examples	- Stock prices (short term) - Temperature anomalies	- GDP or inflation trends. - Long-term forecasting in financial data.	- Retail sales (e.g., seasonal spikes during holidays). - Energy consumption forecasting.

17

SJSU SAN JOSÉ STATE UNIVERSITY

Model Evaluation and Forecasting with ARIMA Models

18

SJSU SAN JOSÉ STATE UNIVERSITY

Model Evaluation

Ensures that the selected model balances **fit** and **complexity**

Helps **avoid overfitting or underfitting** in time series analysis

Key Metrics for Model Evaluation

- Akaike Information Criterion (AIC)
- Bayesian Information Criterion (BIC)
- Root Mean Squared Error (RMSE)
- Mean Absolute Error (MAE)
- Ljung-Box Test (Residual Diagnostics)

19

SJSU SAN JOSÉ STATE UNIVERSITY

Model Evaluation Metrics

	What does it measure?	Formula	Result Interpretation
Akaike Information Criterion (AIC)	Relative quality of a statistical model for a given dataset	AIC=2k−2ln(L) Where k: Number of model parameters. L: Maximum likelihood of the model.	Lower AIC is better
Bayesian Information Criterion (BIC)	Similar to AIC but adds a stronger penalty for the number of parameters	BIC=kln(n)−2ln(L) Where k: Number of model parameters. L: Maximum likelihood of the model. n: Number of observations.	Lower BIC is better
Root Mean Squared Error (RMSE)	Standard deviation of prediction errors	$\mathrm{RMSE} = \sqrt{\frac{\sum_{t=1}^{n}(y_t - \hat{y}_t)^2}{n}}$	A smaller RMSE indicates better model performance RMSE is more sensitive to outliers because errors are squared
Mean Absolute Error (MAE)	Averages the absolute errors in predictions	$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\lvert y_i - \hat{y}_i\rvert$	A smaller MAE indicates better model performance It is less sensitive to outliers compared to RMSE
Ljung-Box Test (Residual Diagnostics)	Typically applied to residuals from a fitted time series model (e.g., ARIMA or SARIMA) to check whether the model adequately removes autocorrelation	$Q = n(n+2)\sum_{k=1}^{m}\frac{\hat{r}_k^2}{n-k}$ Where: • Q: Test statistic following a chi-square distribution. • n: Total number of observations. • m: Number of lags being tested. • $\hat{r}_k$: Sample autocorrelation at lag k.	Null Hypothesis (H_0): The residuals are uncorrelated (no significant autocorrelation). Residuals behave like white noise. Alternative Hypothesis (H_a): The residuals are correlated (presence of significant autocorrelation).

20