ARIMA modeling of a time series

Linear Time Series

Enzo MORAN
Gabin SIMONNET

Contents

Introduction

The objective of this project is to model a macroeconomic variable as a time series and to forecast its short-term evolution using the software `R` (the full code is included in the file `gsimonnet_emoran.zip`).

1 The dataset

1.1 Description and modifications (Q1)

The Industrial Production Index (IPI) is a statistical tool that tracks the monthly evolution of industrial activity in France. It covers the secondary sector, including factories, and is calculated by Insee using data from sectoral surveys conducted by Insee, SDES, SSP, and professional organizations.

We choose the IPI corresponding to the industrial production index for agricultural and forestry machinery (available here: INSEE website). We use the monthly index, ranging from January 1990 to February 2025.

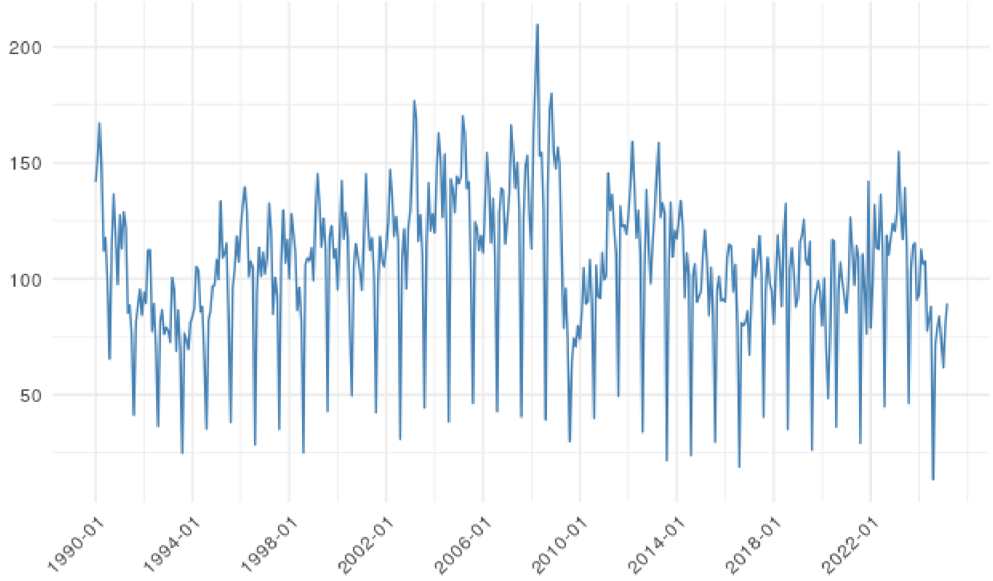

Figure 1: Industrial production index for agricultural and forestry machinery from January 1990 to February 2025

A first overview highlights strong signs of non-stationarity of the series. We suspect a seasonality trend with a periodicity of 12, which is confirmed by the representation of the complete autocorrelation functions of the series (cf Figure 2 below).

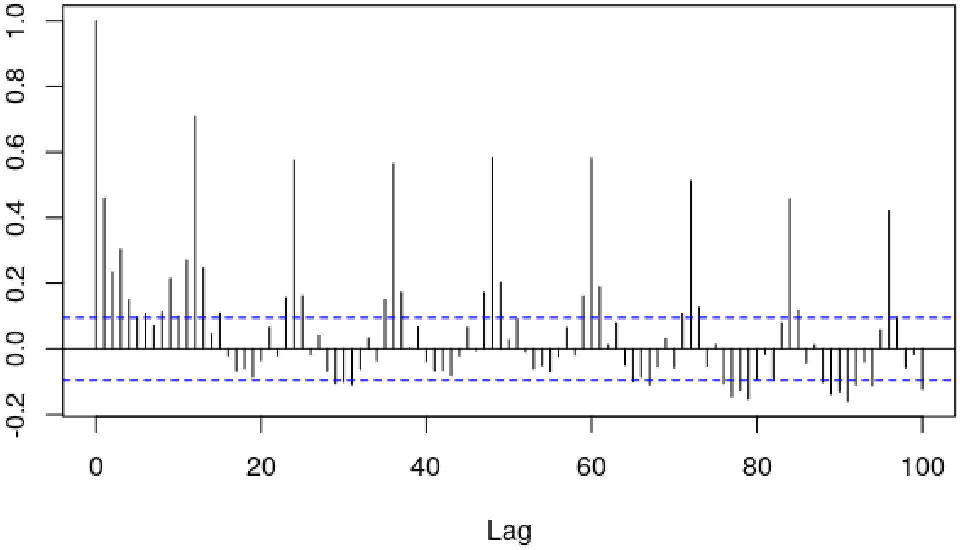

Figure 2: Autocorrelation Function (ACF) of the Industrial Production Index

Although we could use a SARIMA model, we choose to work with the seasonally and calendar-adjusted version of the series (CVS-CJO) also provided by Insee, and retain this version for the remainder of the analysis.

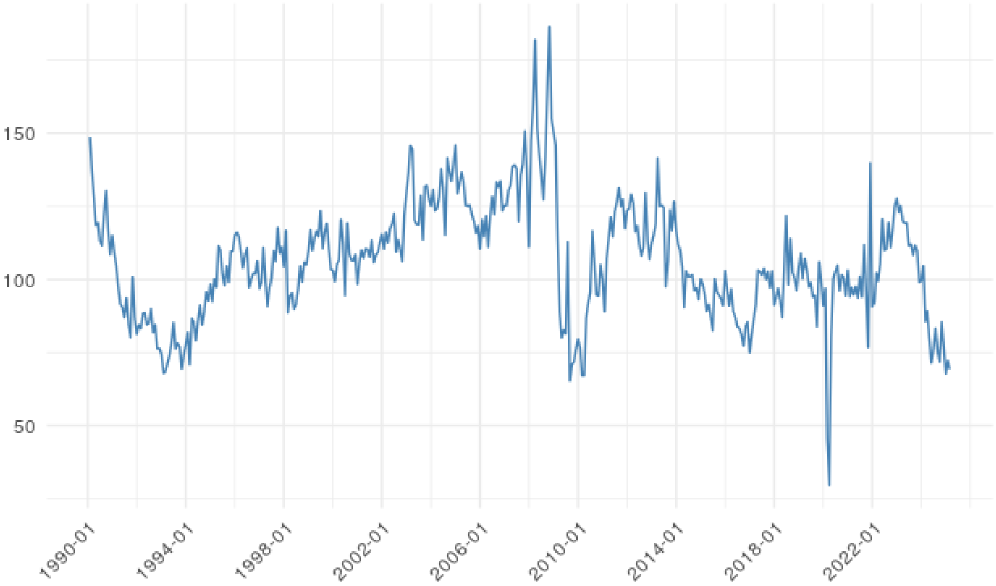

Figure 3: IPI for agricultural and forestry machinery from January 1990 to February 2025, CVS-CJO

Still, the representation given in Figure 3 suggests that the CVS-CJO series may exhibit non-stationary behavior. This observation is supported by the autocorrelation function (ACF), presented in Appendix A, which displays a very slow decay and strong persistence, both typical characteristics of non-stationary processes. However, these visual indicators alone are insufficient to conclude that the series is integrated. Therefore, we apply formal statistical tests to confirm this hypothesis. If non-stationarity is confirmed, we will transform the series, typically through differencing,in order to satisfy the stationarity condition required for ARMA modeling.

1.2 Stationarity (Q2)

In the following, we denote the series by X_t for notational convenience.

Stationarity is a cornerstone of ARMA models. Therefore, before fitting any such models, it is essential to verify whether the series satisfies this property.

To select the most appropriate specification for unit root testing, we first need to determine if the series exhibits a deterministic trend. Although a visual inspection of the series (see Figure 3) does not reveal any obvious trend, we formally assess this by regressing the series on time. As shown in Table 1, the coefficient on time is not statistically significant, confirming the absence of a trend, whereas the constant term is clearly significant. Consequently, we perform unit root tests with a specification that includes an intercept but excludes a trend.

Table 1: Linear regression on t and a constant

Coefficients	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	108.101587	1.950168	55.432	<2e-16
time	-0.008922	0.007971	-1.119	0.264

To proceed, we apply the Augmented Dickey-Fuller (ADF) test to assess the stationarity of the series. However, the optimal number of lags to include in the ADF regression is not known beforehand. Choosing an inappropriate lag length may result in autocorrelated residuals, which would invalidate the test. Therefore, we perform the ADF test for various lag values k and verify for each specification the absence of autocorrelation in the residuals using the Ljung-Box test.

This iterative procedure is implemented via the `dfTestvalid()` function provided in our code (see `gsimonnet_emoran.zip`). We select the smallest number of lags for which the residuals of the ADF regression pass the Ljung-Box test at the conventional significance level ($\alpha = 5\%$). Using this approach, we find that 11 lags are required in the ADF test to ensure the absence of residual autocorrelation.

Table 2: ADF test on X_t with 11 lag

Lag Order	Dickey-Fuller Statistic	P-value
11	-3.4024	0.0121

Since the p-value is below the 5% significance level, the test rejects the null hypothesis of a unit root, suggesting that the series is likely stationary. However, this conclusion is subject to some doubts. The visual inspection of the series does not clearly support stationarity. Moreover, the ADF test required including up to 11 lags to address autocorrelation in the residuals (which is relatively high and calls the robustness of the result into question). To gain further insight, we also conduct Phillips-Perron and KPSS tests.

Table 3: Tests on X_t

Test	H_0	Statistic	P-value
ADF (lag order: 11)	unit root	-3.4024	0.0121
Phillips-Perron (lag: 5)	unit root	-48.338	0.0100
KPSS (lag: 5)	stationarity	0.67423	0.01589

The KPSS test rejects the null hypothesis of stationarity at the 5% significance level, while both the ADF and Phillips-Perron tests reject the null hypothesis of a unit root. This conflicting evidence suggests ambiguity regarding the true nature of the series. As a result, we consider working with the first-differenced series: $Y_t := \Delta X_t = X_t - X_{t-1}$.

We then apply the same procedure to the first-differenced series. The linear regression of the differenced series indicates that the coefficients associated with both the time trend and the constant are not statistically significant (see Table 4). Therefore, we proceed with an ADF test without constant or trend. The optimal number of lags selected is 4, which is sufficient to eliminate residual autocorrelation. All stationarity tests applied to the differenced series (including ADF, Phillips-Perron, and KPSS) consistently indicate that the series is stationary (see Table 5).

Table 4: Linear regression of Y_t on t and a constant

Coefficients	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	- 0.0432746	1.0718485	-0.040	0.968
time	-0.0004688	0.0043915	-0.107	0.915

Table 5: Tests on Y_t

Test	H_0	Statistic	P-value
ADF (lag order: 4)	unit root	-11,1537	0.01
Phillips-Perron (lag: 5)	unit root	-444,73	0.01
KPSS (lag: 5)	stationarity	0.035426	0.1

1.3 Graphical representation (Q3)

We introduce the series and its differentiated version.

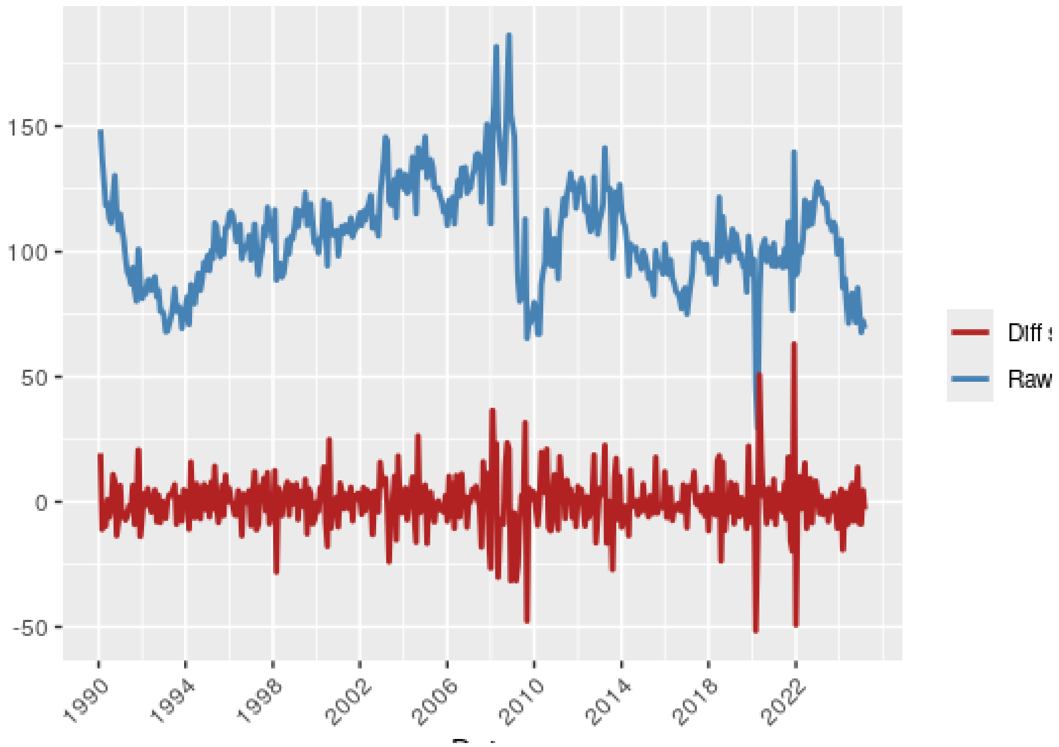

Figure 4: IPI for agricultural and forestry machinery from January 1990 to February 2025, before and after transformation

2 ARMA models

2.1 Identification, estimation and diagnostic checking (Q4)

We aim to model the series using an ARIMA$(p, 1, q)$ model, drawing inspiration from the Box-Jenkins methodology. This approach follows three main stages : identification, estimation, and diagnostic.

In the identification phase, we begin by analyzing the autocorrelation function (ACF) and the partial autocorrelation function (PACF). These diagnostic tools help us identify suitable maximum lag orders for the autoregressive ($p_{\max}$) and moving average ($q_{\max}$) components.

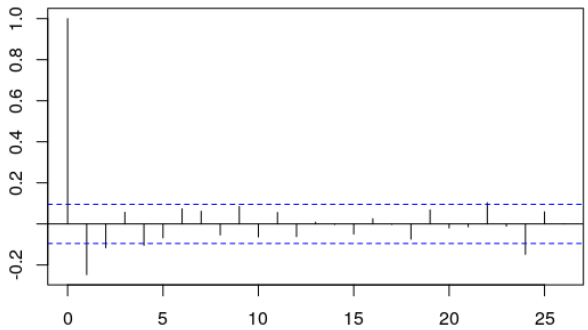

Figure 5: ACF of Y_t

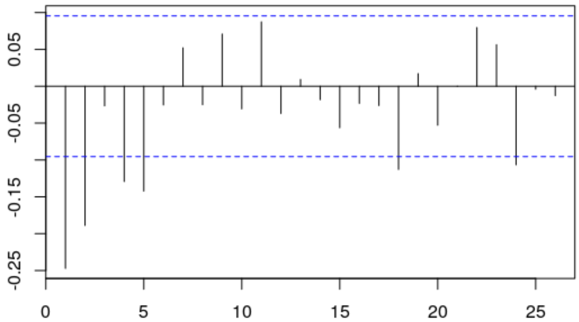

Figure 6: PACF of Y_t

A quick analysis of Figures 5 and 6 shows that the autocorrelation and the partial autocorrelation exceed the $\pm 1.96/\sqrt{n}$ confidence bounds at lag 2 and at lag 5. This indicates that the autocorrelation is significant up to lag 2, and the partial autocorrelation up to lag 5. Based on these observations, we choose $p^* = 5$ and $q^* = 2$. Higher lags are unlikely to significantly improve the model so they are not included. Consequently, the series can be modeled using an ARIMA$(p, 1, q)$ with $p \leq p^*$ and $q \leq q^*$.

To select the most appropriate ARIMA$(p, 1, q)$ model within the specified bounds, we implement a systematic procedure reflected in our code (see `gsimonnet_emoran.zip`). For each candidate model such that $p \leq p^*$ and $q \leq q^*$, we estimate the parameters using the `arima()` function. We then assess the statistical significance of the autoregressive and moving average coefficients using a custom `signif()` function, which computes the corresponding t-statistics and p-values.

Simultaneously, the residuals are evaluated using the `Qtests()` function, which performs the Ljung-Box test over multiple lags (up to 24) to detect any remaining autocorrelation. The `modelchoice()` function consolidates these checks, flagging models as valid when the highest-order coefficients are statistically significant and the residuals show no signs of autocorrelation. This combined estimation and diagnostic process is applied systematically to all candidate models through the `armamodelchoice()` routine.

As a result, three models are identified as both well-adjusted and valid: ARIMA$(5, 1, 0)$, ARIMA$(4, 1, 2)$, and ARIMA$(4, 1, 1)$.

Table 6: Table of BIC and AIC for different models

Model	BIC	AIC
ARIMA(5,1,0)	3203.077	3174.762
ARIMA(4,1,2)	3209.158	3176.798
ARIMA(4,1,1)	3206.148	3177.833

To choose between these three models, we compute both the Bayesian Information Criterion (BIC) and the Akaike Information Criterion (AIC) for each combination. The ARIMA$(5,1,0)$ model achieves the lowest AIC and BIC, and is therefore selected as the most appropriate model for the data.

2.2 Model (Q5)

As established earlier, Y_t follows an ARMA$(5,0)$ model which is equivalent to an AR(5) process and can be written as:

$$Y_t = \sum_{i=1}^{5} \phi_i Y_{t-i} + \varepsilon_t$$

The AR coefficients ϕ_i are estimated using Maximum Likelihood Estimation (MLE):

Table 7: Estimation of the AR(5) coefficients

Coefficient	Estimate	Std. Error
AR(1)	-0.3197	0.0483
AR(2)	-0.2297	0.0501
AR(3)	-0.9500	0.0511
AR(4)	-0.1715	0.0502
AR(5)	-0.1419	0.0484

This yields the final expression of the model:

$$Y_t = -0.3197 \cdot Y_{t-1} - 0.2297 \cdot Y_{t-2} - 0.9500 \cdot Y_{t-3} - 0.1715 \cdot Y_{t-4} - 0.1419 \cdot Y_{t-5} + \varepsilon_t$$

In the following, we return to the original series X_t to facilitate the next computations. Recall that X_t is the integrated version of Y_t, and we now express it as:

$$X_t = 0.6803 \cdot X_{t-1} + 0.0900 \cdot X_{t-2} - 0.7203 \cdot X_{t-3} + 0.7785 \cdot X_{t-4} + 0.0296 \cdot X_{t-5} + 0.1419 \cdot X_{t-6} + \varepsilon_t$$

3 Prediction

3.1 Confidence Regions for Future Values (Q6)

With the explicit form of our time series established, we can now proceed to forecast the next values. Specifically, we predict the next two observations, X_{T+1} and X_{T+2}, and provide the corresponding confidence intervals.

$$\begin{aligned}\hat{X}_{T+1|T} = {} & 0.6803 \cdot X_T + 0.0900 \cdot X_{T-1} - 0.7203 \cdot X_{T-2} \\ & + 0.7785 \cdot X_{T-3} + 0.0296 \cdot X_{T-4} + 0.1419 \cdot X_{T-5} = 69.563\end{aligned}$$

$$\begin{aligned}\hat{X}_{T+2|T} = {} & 0.6803 \cdot \hat{X}_{T+1|T} + 0.0900 \cdot X_T - 0.7203 \cdot X_{T-1} \\ & + 0.7785 \cdot X_{T-2} + 0.0296 \cdot X_{T-3} + 0.1419 \cdot X_{T-4} = 72.506\end{aligned}$$

We can now express the confidence region based on the forecast error variance, which we estimate using the empirical variance of the residuals σ(assumed to be constant).

Let us begin with X_{T+1}. Since $X_{T+1} = \hat{X}_{T+1|T} + \varepsilon_{T+1}$ and the prediction error comes only from the innovation ε_{T+1}, the forecast error variance is given by:

$$\mathrm{Var}(X_{T+1} - \hat{X}_{T+1|T}) = \mathrm{Var}(\varepsilon_{T+1}) = \sigma^2 \approx 10.23^2 = 104.66$$

We can then deduce the confidence region for X_{T+1} at the confidence level $1 - \alpha = 0.95$, which corresponds to:

$$\frac{(X_{T+1} - \hat{X}_{T+1|T})^2}{\sigma^2} \leq \chi^2_{0.95}(1)$$

where $\chi^2_{0.95}(1) = 3.841$ is the 95th percentile of the χ^2 distribution with one degree of freedom. Rewriting the inequality gives:

$$X_{T+1} \in \left[\hat{X}_{T+1|T} \pm \sqrt{\sigma^2 \cdot \chi^2_{0.95}(1)}\right]$$

Using $\hat{X}_{T+1|T} = 69.563$ and $\sigma = 10.23$, the confidence interval becomes:

$$X_{T+1} \in \left[69.563 \pm \sqrt{10.23^2 \cdot 3.841}\right] \quad \text{i.e.} \quad \boxed{X_{T+1} \in [49.51,\ 89.61]}$$

We perform the exact same computation to determine the confidence region for X_{T+2}. We now have:

$$X_{T+2} = \hat{X}_{T+2|T} + (\varepsilon_{T+2} + \beta_1 \varepsilon_{T+1})$$

Thus, the prediction error arises both from the new innovation ε_{T+2} and from the propagation of the previous error $\beta_1 \varepsilon_{T+1}$. Assuming that the ε_t are uncorrelated, the forecast error variance is given by:

$$\mathrm{Var}(X_{T+2} - \hat{X}_{T+2|T}) = \mathrm{Var}(\varepsilon_{T+2} + \beta_1 \varepsilon_{T+1}) = \sigma^2 + \beta_1^2 \sigma^2 \approx 10.23^2 + 0.6803^2 \cdot 10.23^2 \approx 153.0871$$

We can then write the confidence region for X_{T+2} at the 95% confidence level as:

$$\frac{(X_{T+2} - \hat{X}_{T+2|T})^2}{\mathrm{Var}(X_{T+2} - \hat{X}_{T+2|T})} \leq \chi^2_{0.95}(1)$$

which implies:

$$X_{T+2} \in \left[\hat{X}_{T+2|T} \pm \sqrt{\mathrm{Var}(X_{T+2} - \hat{X}_{T+2|T}) \cdot \chi^2_{0.95}(1)} \right]$$

Substituting the values $\hat{X}_{T+2|T} = 72.506$ and $\mathrm{Var}(X_{T+2} - \hat{X}_{T+2|T}) \approx 153.0871$, we obtain:

$$X_{T+2} \in \left[72.506 \pm \sqrt{153.0871 \cdot 3.841} \right] \quad \text{i.e.,} \quad \boxed{X_{T+2} \in [48.26,\ 96.76]}$$

3.2 Hypothesis (Q7)

The main assumptions underlying our confidence region computations are the normality and the absence of autocorrelation in the residuals. In this section, we assess the validity of these assumptions to confirm the reliability of our previous results.

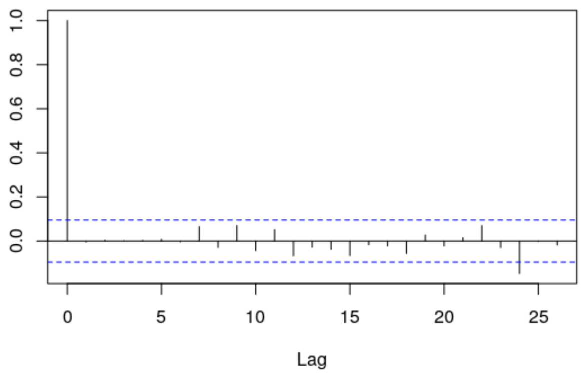

Figure 7: ACF of the residuals

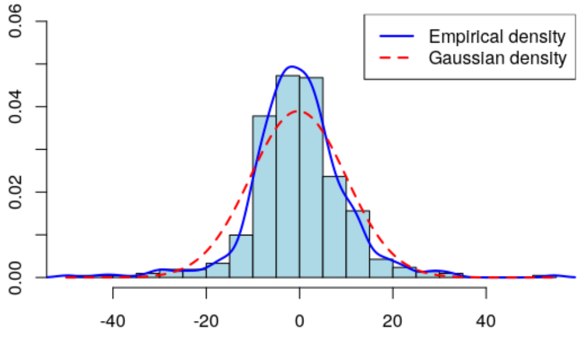

Figure 8: Distribution of the residuals compared to a Gaussian density

The ACF plot of the residuals (Figure 7) shows no significant autocorrelation beyond lag 0, indicating that the residuals can be considered uncorrelated. See Appendix B for additional residual diagnostics.

Moreover, the histogram of the residuals (Figure 8) closely aligns with a Gaussian distribution, suggesting approximate normality.

Taken together, these observations support the validity of the assumptions made for constructing the confidence intervals. Hence, the results obtained for the forecast regions can be considered reliable.

3.3 Forecast of X_{T+1} and X_{T+2} (Q8)

We have validated the assumptions underlying our confidence intervals and we can now fully rely on the forecast expressions derived earlier. Below, we display the predictions for X_{T+1} and X_{T+2} along with their corresponding confidence intervals:

$$\hat{X}_{T+1|T} = 69.563 \quad \text{and} \quad \hat{X}_{T+2|T} = 72.506$$

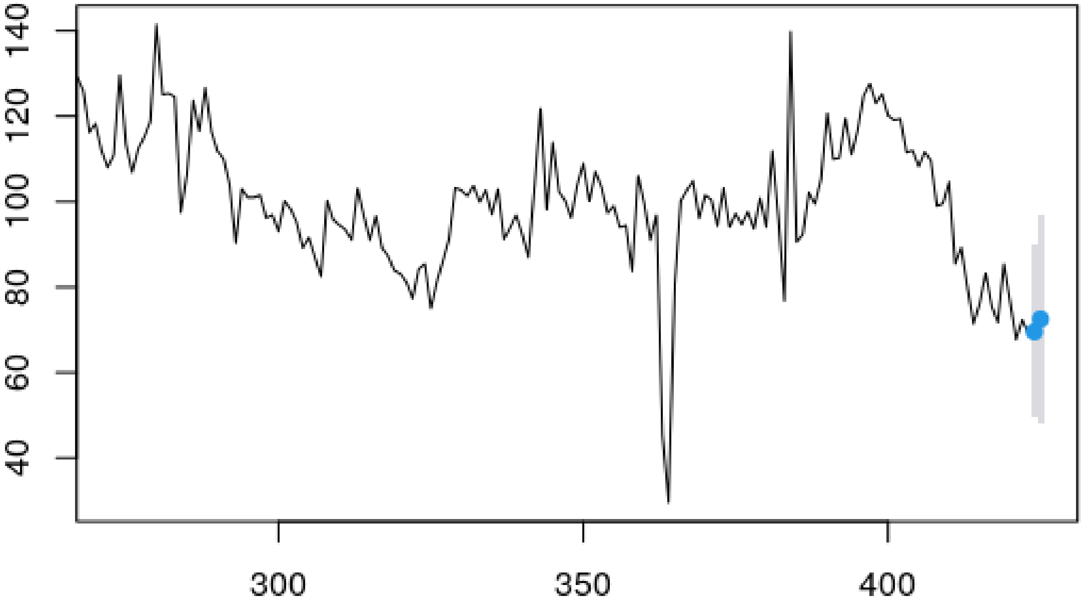

Figure 9: Forecast and confidence intervals for X_{T+1} and X_{T+2}

3.4 Completion of X_T using another stationary series (Q9)

We now consider a secondary stationary series Y_t, observed from $t = 1$ to T, with the additional assumption that Y_{T+1} becomes available before X_{T+1}.

We investigate whether this early observation of Y_{T+1} can enhance the forecast of X_{T+1}. This situation corresponds to *instantaneous Granger causality*, which is defined as:

$$Y_{T+1} \text{ instantaneously causes } X_{T+1} \iff \mathbb{E}[X_{T+1} \mid \mathcal{F}_T, Y_{T+1}] \neq \mathbb{E}[X_{T+1} \mid \mathcal{F}_T],$$

where $\mathcal{F}_T = \sigma(\{X_t, Y_t\}, t \leq T)$ denotes the information set generated by all past observations. To ensure this notion is meaningful, we assume that the joint process (X_t, Y_t) is stationary.

To illustrate this mechanism, we consider a bivariate stationary VAR(1) model:

$$\begin{pmatrix} X_t \\ Y_t \end{pmatrix} = A \begin{pmatrix} X_{t-1} \\ Y_{t-1} \end{pmatrix} + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(0, \Sigma),$$

where

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix}.$$

Without access to Y_{T+1}, the best linear forecast of X_{T+1} given past information is:

$$\widehat{X}_{T+1|T} = a_{11}X_T + a_{12}Y_T.$$

However, once Y_{T+1} is observed, we have:

$$X_{T+1} = a_{11}X_T + a_{12}Y_T + \varepsilon_{X,T+1}, \quad Y_{T+1} = a_{21}X_T + a_{22}Y_T + \varepsilon_{Y,T+1}.$$

If the innovations $\varepsilon_{X,T+1}$ and $\varepsilon_{Y,T+1}$ are correlated (i.e., $\rho \neq 0$), then Y_{T+1} contains additional information about $\varepsilon_{X,T+1}$, and hence about X_{T+1} beyond what is captured by $\mathcal{F}_T$.

Using the properties of the multivariate normal distribution, the conditional expectation becomes:

$$\mathbb{E}[X_{T+1} \mid \mathcal{F}_T, Y_{T+1}] = a_{11}X_T + a_{12}Y_T + \rho\frac{\sigma_X}{\sigma_Y} \cdot \varepsilon_{Y,T+1},$$

which shows explicitly how the forecast is improved when $\rho \neq 0$.

To assess this effect empirically, one can estimate the VAR(1) model, compute the residuals $\widehat{\varepsilon}_{X,t}$ and $\widehat{\varepsilon}_{Y,t}$, and test the null hypothesis $H_0 : \rho = 0$ using a Student or Wald test.

Conclusion: Observing Y_{T+1} before X_{T+1} improves the forecast of X_{T+1} if and only if the innovation terms $\varepsilon_{X,T+1}$ and $\varepsilon_{Y,T+1}$ are correlated.

A Appendix A: Autocorrelation Function (ACF) of the CVS-CJO series

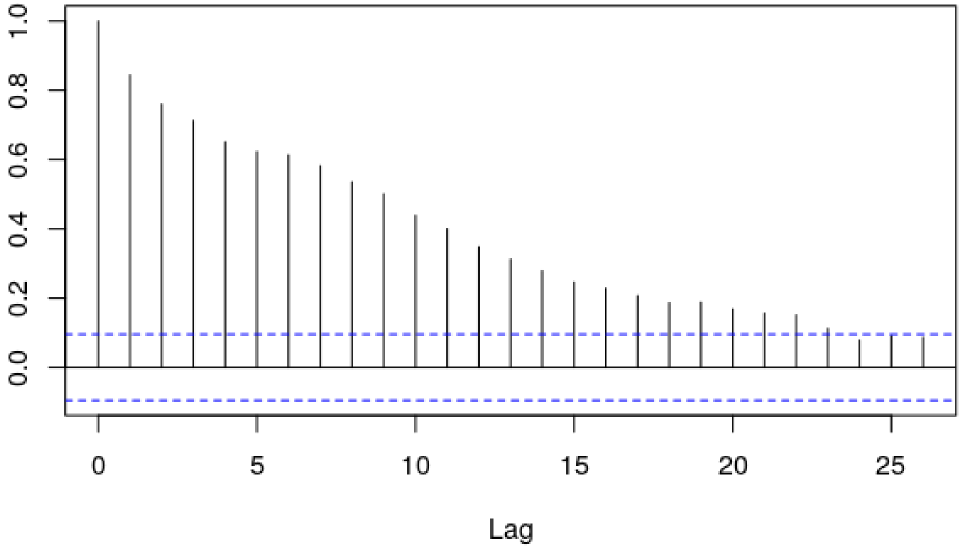

Figure 10: Autocorrelation Function (ACF) of the CVS-CJO Industrial Production Index

B Appendix B: Residuals over time

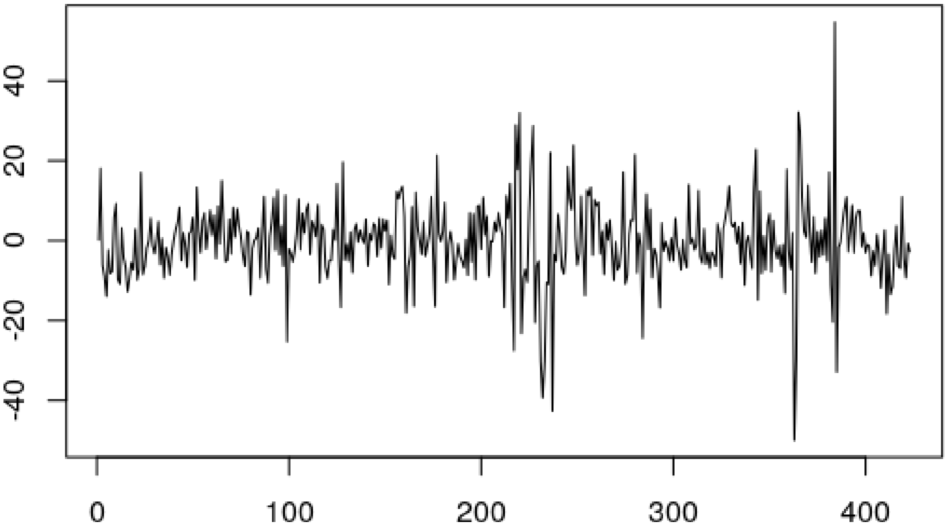

Figure 11: Residuals over time