الجمهورية الجزائرية الديمقراطية الشعبية
RÉPUBLIQUE ALGÉRIENNE DÉMOCRATIQUE ET POPULAIRE

وزارة التعليم العالي والبحث العلمي
MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE SCIENTIFIQUE

جامعة عمار ثليجي - الأغواط
UNIVERSITÉ AMAR TELIDJI - LAGHOUAT

كلية العلوم
FACULTÉ DES SCIENCES

قسم الرياضيات
DÉPARTEMENT DE MATHÉMATIQUES

Mémoire de Master

Domaine : Mathématiques et Informatique
Filière : Mathématiques
Option : Analyse fonctionnelle et applications

Présenté par: **NEBEG Noussiba**

Thème

Estimation of Vector Autoregressive Models

Devant le jury composé de:

Nawel ABDESSELAM	M.C.B	Université de Laghouat	Président
Mohamed BASSOUDI	M.C.A	Université de Laghouat	Encadreur
Ahcene BOUKEHILA	M.C.A	Université de Laghouat	Examinateur

Année Universitaire: 2024 / 2025

⁂ *Acknowledgments* ⁂

First and foremost, I praise and thank Allah (SWT), who granted me the strength, wisdom, and perseverance to complete this memory. His blessings have been my constant support throughout this journey.

I would like to express my deepest gratitude to my esteemed supervisor, **Dr. Mohammed BASSOUDI**. It has been both an honor and privilege to be among your students and to benefit from your profound knowledge and guidance. Your exceptional pedagogical approach and human qualities have been a true inspiration to me. Your kindness, patience, and constant encouragement have left an indelible mark on my academic development.

I am sincerely grateful to the members of the examination committee for their valuable time and consideration in evaluating this work. Their insightful feedback and constructive criticism have greatly contributed to improving this research.

My sincere thanks go to all faculty members of the Mathematics Department for their academic support and for fostering a stimulating intellectual environment. In particular, I would like to express special appreciation to : **Dr. Fares YAZID, Dr. Rahmoune Abdelaziz, Dr. Nawal ABDESSELAM, Dr. Djamel OUCHENANE, and Dr. Ameur YAGOUB, Dr. Ahcene BOUKEHILA** for their generous assistance, valuable advice, and moral support throughout my studies.

Finally, words cannot adequately express my gratitude to my beloved family - especially my parents and sisters - for their unconditional love, endless patience, and unwavering support during both challenging and rewarding moments of this academic endeavor. Their encouragement has been my greatest strength.

⁂ *Dedications* ⁂

My university journey has come to an end after a lifetime of hardship and fatigue. Here I am, concluding my graduation project with energy and enthusiasm. I dedicate this graduation to the light of my eyes, my generous parents, who devoted their entire lives to me. They planted letters in me, which produced numbers. They taught me that hope is a difficult equation, but that its solution lies in patience and faith.

To my grandparents for their unconditional love and support throughout my life. Their wise counsel and wisdom have been a source of inspiration. To my brother Ibrahim, the joy of my life.

To my sisters, who have always been able to motivate me, even in moments of doubt. Thank you so much for their encouragement and moral support.I will never forget my sister's children :Khadidja, Yagoub, and Abed.

And to my lifelong friend, Teggar Fatima Zohra, who walked with me the path of success.

And to my fiancé Mahdi, my partner in every step.

To all my family and friends.

NEBEG Noussiba

ملخص

الهدفُ من هذه الدراسة هو استخدام النماذج الذاتية للانحدار المتجه لتحليل وتوقّع العلاقات الديناميكية بين عدة متغيّرات زمنية، باستعمال تقنيات رياضية وإحصائية .

الكلمات المفتاحية : النماذج الذاتية الانحدار المتجهة ؛ تحليل السلاسل الزمنية ؛ تقدير النماذج ؛ نموذج .LASSO

RÉSUMÉ

Cette étude est vise à explorer l'utilisation des modèles autorégressifs vectoriels (VAR) pour analyser et prévoir les relations dynamiques entre plusieurs variables temporelles, en utilisant les techinques mathématique et statistiques.

Mots clés : modèles autorégressifs vectoriels ; analyse des séries temporelles ; estimation des modèles ; modèle LASSO

ABSTRACT

This study aims to explore the use of Vector Autoregressive (VAR) models to analyze and predict dynamic relationships between multiple time series variables, using mathematical and statistical techniques.

Key words : vector autoregressive models ; time series analysis ; model estimation ; LASSO model.

Table des matières

Notation and symbols

	X_t	Time series random variables at time t.
	ϵ_t	White noise error term at time t.
	ε_t	Vector of error terms.
	μ_t	Mean at time t.
	$\gamma(h)$	Autocovariance at lag h.
	$\rho(h)$	Autocorrelation at lag h.
	$\Psi(h)$	Partial autocorrelation at lag h.
	Σ_u	Covariance matrix of errors.
	ϕ_i	Coefficient matrix at lag i.
	β	Vector of regression coefficients.
	$\tilde{\beta}$	Estimated coefficient vector.
	$\tilde{\Sigma}_u$	Estimated error covariance matrix.
	$X'X$	Gram matrix.
	λ	Regularization parameter.
A	AIC	Akaike Information Criterion.
	$ARIMA$	Autoregressive Integrated Moving Average.
B	BIC	Bayesian Information Criterion.
D	DS	Deffernce-Stationary.
L	$LASSO$	Least Absolute Shrinkage and Selection Operator.
M	ML	Maximun Likelihood.
	MLE	Maximun Likelihood Estimation.
	MLR	Multiple Linear Regression.
O	OLS	Ordinary Least Squares.
R	RSS	Residual Sum of Squares.
	$RMSE$	Root Mean Squared Error.
T	TS	Trend-Stationary.
V	$VAR(p)$	Vector autoregressive model of order p.
	$VARMA(p,q)$	Vector ARMA model.
	$VMA(q)$	Vector moving average model of order q.

List of contributions

In the setting of this memory, we have realized the following scientific contributions :

1. A comprehensive comparative study of estimation techniques for Vector Autoregressive (VAR) models, including Ordinary Least Squares (OLS), Maximum Likelihood Estimation (MLE), and penalized methods such as LASSO and Ridge regression.
2. Derivation and analysis of the theoretical properties of OLS and MLE estimators in the context of VAR models, highlighting their conditions for consistency, efficiency, and asymptotic behavior
3. Implementation and evaluation of regularization techniques (LASSO and Ridge) in VAR estimation, with a focus on their ability to handle multicollinearity and model selection in high-dimensional settings.
4. A detailed comparison between LASSO and Ridge estimators based on bias-variance trade-offs, sparsity, and predictive performance, supported by theoretical arguments and numerical simulations.
5. Development of numerical examples and simulations illustrating the performance and limitations of each estimation method in finite samples.

General introduction

In a world characterized by increasing interdependence among economic variables, joint modeling of multivariate time series has become essential for understanding economic dynamics. Governments, financial institutions, and businesses rely on accurate forecasts and robust analytical tools to guide their decisions. In this context, Vector Autoregressive (VAR) models occupy a central position in econometrics and applied statistics, particularly for analyzing dynamic interactions between multiple economic variables such as Gross Domestic Product (GDP), inflation rate, exchange rate, and energy consumption.

Introduced by Christopher Sims in 1980, VAR models provide a flexible framework that allows each variable to be modeled as a linear function of lagged values of all variables in the system, including itself. This approach enables structural analysis without imposing strong theoretical restrictions, while facilitating economic interpretation through tools such as Impulse Response Functions (IRF) and Forecast Error Variance Decomposition. While traditional univariate models, such as ARIMA models, are limited to a single time series, economic and financial phenomena are inherently multivariate. They require tools capable of capturing cross-effects and feedback between variables. The VAR approach thus establishes itself as an essential method in the dynamic analysis of multivariate time series data.

However, despite their popularity, several challenges persist : selecting the optimal number of lags, stationarity of the series, efficient parameter estimation, and correct interpretation of results. The objective of this memory is to study and compare different VAR parameter estimation methods - particularly Ordinary Least Squares (OLS), Maximum Likelihood Estimation (MLE), as well as penalized approaches like LASSO and Ridge - in both applied and simulated frameworks.

This work aims to thoroughly explore VAR model estimation methods from both theoretical and practical perspectives. The specific objectives are to : present the theoretical foundations of VAR models and their validity conditions ; examine different parameter estimation methods, notably OLS, MLE, LASSO and Ridge ; apply these methods to simulated data in a controlled experimental setting ; compare estimator performance using statistical criteria such as RMSE ; interpret results through tools like Impulse Response Functions (IRF) and variance decomposition ; and provide recommendations for choosing estimation methods based on application context.

This memory consists of four chapters. In the first chapter, we develop the theoretical framework

for time series study, presenting formal definitions, properties of VAR models, as well as stationarity and identification conditions, stationarity, invertibility, autocorrelation and cross-correlation. We also present classical multivariate linear models, namely VMA(q), VAR(p), and VARMA(p,q) models. In the second chapter, we examine the theoretical framework for least squares estimation in VAR models, followed by Lasso and ridge regression. The third chapter describes the practical implementation of estimation methods on simulated data, with emphasis on R programming and results analysis. A simulation study is conducted to compare the performance of different estimation methods under various conditions. Finally, we present general conclusions and future perspectives.

Chapter 1

Introduction to Vector Autoregressive (VAR) Models

Contents

1.1 General Aspects of Time Series

The intrinsic nature of a time series is such that observed values are generally dependent, and the objective is to identify and model the structure of temporal dependence. In this dissertation, we primarily focus on linear time series $\{X_t, t \in \mathbb{Z}\}$, where the observation at the current time is assumed to be the result of time-invariant linear filtering of a stationary white noise. We also consider non-stationary series, which can be rendered stationary by differentiating the series a sufficient number of times.

1.1.1 Time Series

The theory of time series is a combination of two concepts : probabilistic and statistical. The probabilistic concept involves studying the characteristics of the random variables X_t. The statistical problem is to determine the characteristics of the distributions of the time series X_t, for observations $X_1, X_2, \ldots, X_n$ at times $t = 1, 2, \ldots, n$. The resulting statistical model serves, on the one hand, to understand the stochastic system and, on the other hand, to predict the future (i.e., $X_{n+1}, X_{n+2}, \ldots$).

1.1.2 Analysis of Time Series

The term *time series* refers both to actual chronological series and to a theoretical sequence of random variables indexed by time ($t \in T$), which serves to model the former.

Definition 1.1.1. A time series is a sequence of repeated observations corresponding to different dates, or a set of values representing the evolution of a phenomenon over time. Generally, the observations of a phenomenon are equidistant from one another (discrete time, $t \in \mathbb{N}, \mathbb{Z}, \ldots$) ; time can correspond to a day, a month, a year, etc.

For example, in fields such as finance, one may cite, among others.

- The daily value of the Dollar in Euro at the opening of the stock market.
- Monthly unemployment data.
- Stock prices, etc,

However, in other fields (such as physics), observations are recorded continuously, with the index t taking values in an interval of $\mathbb{R}$.

Definition 1.1.2. [12]

A time (or chronological) series is a sequence of observations $x_1, x_2, \ldots, x_n$ indexed by time. It is assumed to be a realization of a process X, that is, a sequence $\{X_i\}$ of random variables.

Definition 1.1.3. The study of time series in statistics corresponds to regularly spaced observations over time.

1.1.3 Modeling a Time Series

Time series modeling is based on the classical decomposition, known as "Persons' decomposition", which relies on the following four components :

1. **Trend (T_t)** : Long-term movement (long period).
2. **Seasonality (S_t)** : Periodic function of time (short period).
3. **Cycle (C_t)** : Business cycle, periodic fluctuation (medium term).
4. **Residual (ϵ_t)** : Irregular component, corresponding to the concept of deviation from the model.

In general, a model representing the studied time series can be proposed by combining the four elements mentioned above. For this purpose, there are three types of models : the first is the adjustment model of additive or multiplicative form as follows :

- **Additive model** :
$$X_t = T_t + S_t + C_t + \epsilon_t$$
- **Multiplicative model** :
$$X_t = T_t \cdot S_t \cdot C_t + \epsilon_t$$

The second type is the model in which we assume that X_t is a function of its past values and a random disturbance.

$$X_t = f(X_{t-1}, X_{t-2}, \dots \epsilon_t)$$

In this class, we can cite the models AR, MA, ARMA, ARIMA, SARIMA, ...

In this category of models, the random variable X_t is expressed as a function of another variable Y_t and a random disturbance ϵ_t.

$$X_t = f(Y_t, \epsilon_t)$$

Either Y_t is deterministic or random ; in the latter case, the processes $(Y_t)_t$ and $(\epsilon_t)_t$ have certain properties of independence or lack of correlation. These models are basic models that we essentially consider to link them. We thus have two particular cases of explanatory models :

- **Static explanatory model :** where the variables Y_t do not contain past values of X_t and the ϵ_t are independent of each other.
- **Dynamic explanatory model :** where the ϵ_t are autocorrelated and the Y_t contain past values of X_t.

1.1.4 Linear Series

Definition 1.1.4. A series X_t is said to be linear if it can be written as :

$$X_t = \mu + \sum_{i=-\infty}^{+\infty} \psi_i \epsilon_{t-i}$$

where $\epsilon_t \sim BB(0, \delta_\epsilon^2)$, $\psi_0 = 1$, and the sequence (ψ_i) is absolutely summable, meaning :

$$\sum_{i=-\infty}^{+\infty} |\psi_i| < \infty$$

A series (X_t) is said to be linear and causal if it is linear with $\psi_i = 0$ for $i < 0$:

$$X_t = \mu + \sum_{i=-\infty}^{+\infty} \psi_i \epsilon_{t-i} \tag{1.1}$$

It will be assumed that a linear series is stationary. The study of non-causal series leads to non-intuitive results that are difficult to use.

1.1.5 Stationary Model

Representation of time-dependent random phenomena. Let (Ω, F, P) be a probability space, and let T be a non-empty index set (for example : $\mathbb{N}, \mathbb{Z}, \mathbb{R}_+, \ldots$, etc.). Let X_t be a function from $T \times \Omega$, associated with each pair (t, ω), the process $X_t(\omega)$, where ϵ denotes the state space of the process. Hence :

i) For $t \in T$, $X_t(\omega)$ is a random variable.

ii) For $\omega \in \Omega$, $X_t(\omega)$ is a trajectory.

Definition 1.1.5. [4] Stochastic process

A stochastic process defined on T, denoted $(X_t(\omega))_{t \in T}$ or simply $(X_t)_t$, is a collection of random variable X_t of with values in $\mathbb{R}$, in such a way that at each element $t \in T$ is associated with a random variable X_t. We thus have two cases :

i) A discrete-time process. If T is discrete ($T \subseteq \mathbb{Z}$),

ii) A continuous-time process if T is continuous ($T \subseteq \mathbb{R}$),

As a consequence, we are interested in stochastic models, whose elements X_t of the time series $(X_t)_t$ are considered as random variables. Subsequently, we designated by a model, the stochastic process that models the time series.

Generally, the variables of a series $(X_t)_t$ are neither independent nor identically distributed. The means, variances and covariances of these variables depend on their positions in the series. In particular, if we assume that $(X_t)_t$ is square integrable

$$(i.e. E[X_t^2] < \infty, \forall t \in T)$$

Then ;

$$E(X_t) = \mu_t, var(X_t) = \delta_t^2$$

$$Cov(X_t, X_{t-h}) = E[(X_t X_{t-h}) - E(X_t)E(X_{t-h})], h \in \mathbb{Z}$$

1.2 Autocovariance and Autocorrelation

Definition 1.2.1. *Autocovariance function*

The autocovariance function of a time series $(X_t)_t$ is a sequence $(\gamma_x(h))_{h\in\mathbb{Z}}$, where it is an even function, positive semi-definite, i.e.,

$$\gamma_x(h) = cov(X_t, X_{t-h})$$

$$= \sum_{i=1}^{N}\sum_{j=1}^{N} a_i a_j \gamma(t_i - t_j) \geqslant 0$$

$$|\gamma_x(h)| \leqslant \gamma_x(0) = var(X_t), \quad h \in \mathbb{Z}$$

Remark 1.2.1. The function $\gamma_x(h)$ is even, i.e. :

$$\gamma_x(h) = \gamma_x(-h)$$

Definition 1.2.2. *Autocorrelation Function*

Similarly, we define a sequence $(\rho_x(h))_{h\in\mathbb{Z}}$ called the autocorrelation function of the series $(X_t)_t$:

$$\rho_x(h) = corr(X_t, X_{t-h})$$

$$= \frac{cov(X_t, X_{t-h})}{\sqrt{var(X_t)} \cdot \sqrt{var(X_{t-h})}} = \frac{\rho_x(h)}{\rho_x(0)}$$

It is an even function, positive semi-definite, i.e.,

$$\sum_{i=1}^{N}\sum_{j=1}^{N} a_i a_j \rho(t_i - t_j) \geq 0$$

$$|\rho_x(h)| \leq \rho_x(0) = 1, h \in \mathbb{Z}$$

Remark 1.2.2. This function $\rho_x(.)$ takes values in $[-1, 1]$ and $\rho_x(0) = 1$.

Definition 1.2.3. The autocorrelation matrix of the vector $(X_t, X_{t-1}, \ldots, X_{t-h+1})$ is :

$$\mathbf{R(h)} = \begin{pmatrix} 1 & \rho(1) & \rho(2) & \cdots & \rho(h-1) \\ \rho(1) & 1 & \rho(1) & \cdots & \rho(h-2) \\ \rho(2) & \rho(1) & 1 & \cdots & \rho(h-3) \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ \rho(h-1) & \rho(h-2) & \rho(h-3) & \cdots & 1 \end{pmatrix}.$$

1.2.1 The Partial Autocorrelation Function

The partial autocorrelation measures the correlation between (X_t) and X_{t-h} while excluding the influence of variables prior to X_{t-h}. Thus, it can be shown that the partial autocorrelation function of a process$(X_t)_{t\in\mathbb{Z}}$ is given by [11] :

$$\Psi_X(h) = \frac{|R^*(h)|}{|R(h)|}, \quad \text{for all } h. \tag{1.2}$$

or

$$\mathbf{R}^*(\mathbf{h}) = \begin{pmatrix} 1 & \rho(1) & \rho(2) & \cdots & \rho(h-2) & \rho(1) \\ \rho(1) & 1 & \rho(1) & \cdots & \rho(h-3) & \rho(2) \\ \rho(2) & \rho(1) & 1 & \cdots & \rho(h-4) & \rho(3) \\ \cdot & \cdot & \cdot & \cdots & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdots & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdots & \cdot & \cdot \\ \rho(h-3) & \rho(h-4) & \rho(h-5) & \cdots & \rho(1) & \rho(h-2) \\ \rho(h-2) & \rho(h-3) & \rho(h-4) & \cdots & 1 & \rho(h-1) \\ \rho(h-1) & \rho(h-2) & \rho(h-3) & \cdots & \rho(1) & \rho(h) \end{pmatrix}.$$

and

$$\mathbf{R}(\mathbf{h}) = \begin{pmatrix} 1 & \rho(1) & \rho(2) & \cdots & \rho(h-2) & \rho(h-1) \\ \rho(1) & 1 & \rho(1) & \cdots & \rho(h-3) & \rho(h-2) \\ \cdot & \cdot & \cdot & \cdots & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdots & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdots & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdots & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdots & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdots & \cdot & \cdot \\ \rho(h-1) & \rho(h-2) & \rho(h-3) & \cdot & \rho(1) & 1 \end{pmatrix}.$$

Here, $|R(h)|$ is the determinant of a square matrix $R(h)$. Thus, the first three autocorpartial relations are determined by relations

$$\Psi(1) = \rho(1)$$

$$\Psi(2) = \frac{\rho(2) - \rho(1)^2}{1 - \rho(1)^2}$$

$$\Psi(3) = \frac{\rho(1)^3 - \rho(1)\rho(2)(2 - \rho(2)) + \rho(3)(1 - \rho(1)^2)}{1 - \rho(2)^2 - 2\rho(1)^2(1 - \rho(2))}$$

1.2.2 Graphs for time series

Chronogram : The study of a time series begins with the examination of its chronoprogram. He gives it an overall life, shows certain aspects. Such as possible breaks, a change in the dynamics of the series.

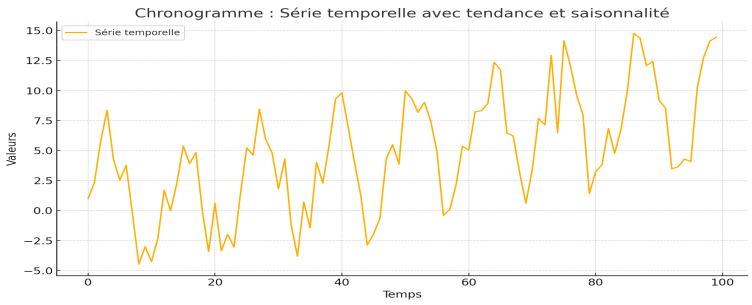

FIGURE 1.1 – Chronogram : Time series with trend and seasonality

Correlogram : A correlogram is the graphical representation of the function autocorrelation, which is a concept related to correlation, it is not a calculation between two different chronicles but between the series and herself at different offsets in the time allowing to detect internal connections within the series.

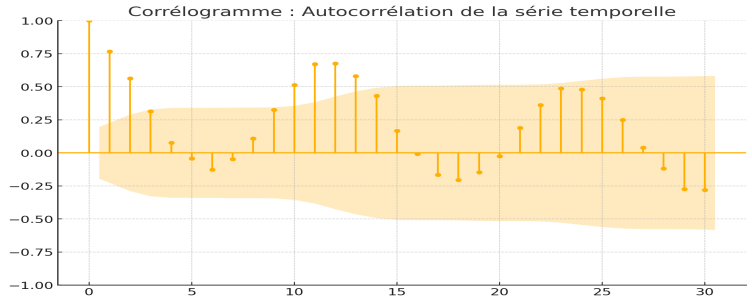

FIGURE 1.2 – Correlogram : Analysis of the autocorrelation of the time series

1.2.3 White Noise Processes

Definition 1.2.4. [1]

A stochastic process is a collection of random variables $\{X_t, t \in I\}$, all defined on a probability space (Ω, F, P). Where I can be $\mathbb{N}$ or $\mathbb{Z}$.

Definition 1.2.5. White noise belongs to the class of stationary processes. Specifically, $(\epsilon_t)_{t\in\mathbb{Z}}$ is white noise if it satisfies the following conditions :

$$\begin{cases} E(\epsilon_t) = 0, & \forall t \\ E(\epsilon_t^2) = \sigma_\epsilon^2, & \forall t \\ \text{cov}(\epsilon_t, \epsilon_{t+h}) = 0, & \forall t, \forall h \neq 0 \end{cases}$$

Remark 1.2.3. Consequently, the behavior of white noise at time t has no influence on its behavior at time $t+h$. We speak of Gaussian white noise when $\epsilon_1, \epsilon_2, \cdots$ are i.i.d. (independent and identically distributed).

Definition 1.2.6. [4] [Spectral density]

The spectral density, denoted as f, of a time series $(X_t)_t$is a function defined on $\mathbb{R}$ by :

$$f(\lambda) = \frac{1}{2\pi} \sum_{h\in\mathbb{Z}} \gamma(h) \exp(ih\lambda), \quad \forall\lambda \in \mathbb{R}$$

Property of White Noise

1. The spectral density of white noise is constant in λ .
2. Any stationary process with a constant spectral density is white noise .

Proposition 1.2.1. *Let $(\epsilon_t)_t$ be a white noise process that is independently and identically distributed (i.i.d.), centered, and with variance $\sigma^2 < \infty$.*

The autocovariance function of white noise,$\gamma_\epsilon(h)$, and its spectral density $f_\varepsilon(\lambda)$ are given by :

1. *For this white noise process, we have :*

$$\gamma_\varepsilon(h) = Cov(\varepsilon_t, \varepsilon_{t+h}) = E[\varepsilon_t\varepsilon_{t+h}] = \begin{cases} \sigma^2, & \text{if } h = 0 \\ 0, & \text{otherwise.} \end{cases}$$

Thus,

$$f_\varepsilon(\lambda) = \frac{1}{2\pi} \sum_{h\in\mathbb{Z}} \gamma_\varepsilon(h) \exp(ih\lambda) = \frac{\sigma^2}{2\pi}$$

(which is constant in λ).

2. *Conversely, if $f_\varepsilon \equiv C^{te}$*

$$\gamma_\varepsilon(h) = \int_{-\pi}^{\pi} \exp(-ih\lambda) f_\varepsilon(h) d\lambda$$

$$= \int_{-\pi}^{\pi} Cte \exp(-ih\lambda) d\lambda = \begin{cases} \gamma_\varepsilon(h), & \text{if } h = 0 \\ 0, & \text{otherwise} \end{cases}$$

Therefore, for the process to be white noise, it suffices that it is stationary, which is assumed by hypothesis.

1.2.4 Stationary Processes

Stationarity is a characteristic of a time series that implies the behavior of the series does not depend on time. Specifically, a time series $(X_t)_{t\in\mathbb{Z}}$ is said to be stable if it does not exhibit seasonal trends, upward trends, or downward trends. More formally, we distinguish between two types of stationarity : **Strongly stationary** and **Weakly stationary** [8] .

Definition 1.2.7. [6] [Strongly stationary series]

A time series$(X_t)_t$ is strongly (or strictly) stationary if and only if :

$$(X_{t_1}, X_{t_2}, ..., X_{t_n}) \overset{D}{=} (X_{t_1+k}, X_{t_2+k}, ..., X_{t_n+k})$$

This means that the distribution of the vector $(X_{t_1}, X_{t_2}, ..., X_{t_n})$ is identical to that of $(X_{t_1+k}, X_{t_2+k}, ..., X_{t_n+k})$ for any subset $\{t_1, t_2, ..., t_n\} \subseteq T, \forall k \in \mathbb{N}$.

Definition 1.2.8. [6] [Weakly stationary series]

A time series $(X_t)_t$ is weakly stationary (or second-order stationary) if and only if :

1. $\mathbb{E}[X_t^2] < \infty$, $\forall t \in T$;
2. $\mathbb{E}[X_t] = \mathbb{E}[X_s] = \mu$, $\forall s, t \in T$;
3. $\mathrm{Cov}(X_s, X_t) = \mathrm{Cov}(X_{s+k}, X_{t+k})$, $\forall s, t \in T$, $\forall k \in \mathbb{N}$.

The existence of the autocovariance function for a stationary time series (often called weakly stationary) is guaranteed by the following proposition.

Proposition 1.2.2. *If the time series $(X_t)_t$ is stationary, then there exists a function $\gamma_X : \mathbb{Z} \to \mathbb{R}$, such that the autocovariances depend only on the difference between observations :*

$$Cov(X_s, X_t) = \gamma_X(|t - s|), \quad \forall s, t \in \mathbb{Z}.$$

Proposition 1.2.3. *Let $r \in \mathbb{Z}$. Since $(X_t)_t$ is stationary, then for all $s, t \in \mathbb{Z}$:*

First case : *If $s \leq t$,*

$$Cov(X_s, X_t) = Cov(X_{s+r-s}, X_{t+r-s}) = Cov(X_r, X_{r+t-s}), \quad \text{if } s < r.$$

Second case : *If $s > t$, then :*

$$Cov(X_s, X_t) = Cov(X_t, X_s) = Cov(X_r, X_{r+s-t}).$$

Thus, for all $s, t \in \mathbb{Z}$:

$$Cov(X_s, X_t) = Cov(X_r, X_{r+|t-s|}) = \gamma_X(|t - s|).$$

Remark 1.2.4.

1. If $(X_t)_t$ is square-integrable for all $t \in T$, then strong stationarity implies weak stationarity.
2. If the series $(X_t)_t$ is Gaussian, then weak stationarity implies strong stationarity.
3. A white noise process is defined as the sequence $(\epsilon_t)_t$ of i.i.d. random variables, centered, and with variance σ^2, i.e.,

$$\mathrm{Var}(\epsilon_t) = E[\epsilon_t^2] = \sigma^2.$$

In this case, it is clear that a white noise process is a second-order stationary process, with :

$$\gamma_\epsilon(0) = \sigma^2 \quad \text{and} \quad \gamma_\epsilon(h) = 0, \quad \forall h \neq 0.$$

Trend-Stationary Process (TS)

This process exhibits non-stationarity of a deterministic nature and is used to remove the trend. It is expressed as follows :

$$Y_t = \alpha + \beta t + \epsilon_t \tag{1.3}$$

where ϵ_t represents the model error. It is clear that Y_t is not stationary, since its first-order moment :

$$E(Y_t) = \alpha + \beta t \tag{1.4}$$

depends on time t.

Difference-Stationary Process (DS)

The Difference-Stationary (DS) process exhibits non-stationarity of a stochastic nature and is used to remove seasonality. It is also known as a **Random Walk**. This process is written as follows :

$$Y_t = Y_{t-1} + \beta + \epsilon_t \tag{1.5}$$

where $\beta \in \mathbb{R}$ and ϵ_t represents the model error. A series $\{Y_t, \forall t \in \mathbb{Z}\}$ is said to be of order d (order of integration) if the filtered process $(1-L)^d$ is stationary. Stationarity is used as a tool for time series analysis. Raw data from any given series are often transformed to achieve stationarity. For example, financial series are often seasonal and depend on a non-stationary price level. In the following sections, we will focus on stationary processes.

Theorem 1.2.1 (Wold's)**.** *[2, 5],*

If $(X_t)_{t\in\mathbb{Z}}$ is a centered and second-order stationary process, then it can be decomposed as follows :

$$X_t = \sum_{j=0}^{+\infty} \psi_j \epsilon_{t-j} + V_t, \quad t \in \mathbb{Z} \tag{1.6}$$

where :

1. *$\psi_0 = 1$ and $\sum_{j=0}^{+\infty} \psi_j^2 < \infty$, with $\psi_j \in \mathbb{R}$.*
2. *$(\epsilon_t)_{t\in\mathbb{Z}}$ is the white noise associated with $(X_t)_{t\in\mathbb{Z}}$.*
3. *$(V_t)_{t\in\mathbb{Z}}$ is a deterministic process.*
4. *$Cov(\epsilon_t, V_s) = 0$, for all $s, t \in \mathbb{Z}$.*

This result is mainly theoretical, as the coefficients ψ_j are usually unknown in practice. Some models involve an infinite number of terms.

Lag Operator

It is often necessary to consider a variable as a function of its past values. Therefore, it is convenient to define an operator that transforms a variable X_t into its past value. This is the **lag operator**, denoted by the letter L, and defined as :

$$LX_t = X_{t-1}, \quad \text{and} \quad L^k X_t = X_{t-k}$$

This operator is used in polynomials, for example :

$$B(L) = \beta_0 + \beta_1 L + \beta_2 L^2 + \cdots + \beta_q L^q$$

Thus :

$$B(L)X_t = \beta_0 X_t + \beta_1 X_{t-1} + \beta_2 X_{t-2} + \cdots + \beta_q X_{t-q}$$

Common operations such as **addition, multiplication, division, and inversion** can be applied to the set of lag polynomials with the same properties as entire series.

Two particular values of lag polynomials are worth noting : $B(0)$ gives the value of the first coefficient of the polynomial, i.e., its constant term. $B(1)$ provides the sum of the coefficients of the same polynomial.

Finally, the operator $1 - L$ plays a special role as it allows computing the first difference of a time series :

$$(1 - L) = X_t - X_{t-1}$$

Properties of the Lag Operator

1. $L^j X_{t-j} = X_{t-j}$
2. $L^0 X_t = X_t$
3. If $X_t = c \in \mathbb{R}$ for all $t \in \mathbb{Z}$, then $L^j X_t = L^j c = c$ for all $j \in \mathbb{Z}$
4. $L^j(L^k X_t) = L^{j+k} X_t = X_{t-(j+k)}$
5. $L^{-j} X_t = X_{t+j}$
6. $(L^j + L^k)X_t = L^j X_t + L^k X_t = X_{t-j} + X_{t-k}$
7. If $|a| < 1$, then :

$$(1 - aL)^{-1} X_t = \sum_{j=0}^{\infty} a^j X_{t-j}$$

Definition 1.2.9. Moving Average

A moving average, centered with length $P\,(P < T)$ of the series $\{x_t, t = 1, \ldots, T\}$, consists of successive averages calculated based on the parity of P according to the following formulas :

- First case, P is odd, $P = 2m + 1$:

$$M_P(t) = \frac{1}{P} \sum_{k=-m}^{+m} x_{t+k}$$

There are $(T - P + 1)$ centered moving averages of odd length P.

- Second case, P is even, $P = 2m$:

$$M_P(t) = \frac{1}{P} \left(\frac{x_{t-m}}{2} + \sum_{k=-m+1}^{m-1} x_{t+k} + \frac{x_{t+m}}{2} \right)$$

The centered moving average $M_{2m}(t)$ appears as a weighted average of the values in the series surrounding the date t, with weighting coefficients equal to $\frac{1}{2P}$ for the two extreme values x_{t-m} and x_{t+m}, and equal to $\frac{1}{P}$ for the $(P - 2)$ intermediate values from x_{t-m+1} to x_{t+m-1}. Thus, it consists of $(P + 1)$ terms :

Values	x_{t-m}	x_{t-m+1}	$\dots$	x_t	$\dots$	x_{t+m-1}	x_{t+m}
Weights	$\frac{1}{2P}$	$\frac{1}{P}$	$\dots$	$\frac{1}{P}$	$\dots$	$\frac{1}{P}$	$\frac{1}{2P}$

TABLE 1.1 – Table of values and weights for the centered moving average

There are $(T - P)$ centered moving averages of even length P. For simplicity, given that the length P of the moving average is fixed, we will now denote Y_t as the centered moving average of length P at date t.

1.3 The VAR(p) Model

A VAR model with k variables and p lags, denoted as VAR(p), can be expressed in matrix form as follows :

$$X_t = a + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \phi_3 X_{t-3} + \cdots + \phi_p X_{t-p} + \varepsilon_t \tag{1.7}$$

Where :

$$X_t = \begin{bmatrix} x_t^1 \\ x_t^2 \\ \vdots \\ x_t^k \end{bmatrix}; \quad \phi_i = \begin{bmatrix} \phi_{1i}^1 & \phi_{1i}^2 & \phi_{1i}^3 & \phi_{1i}^k \\ \phi_{2i}^1 & \phi_{2i}^2 & \phi_{2i}^3 & \phi_{2i}^k \\ \vdots & \vdots & \vdots & \vdots \\ \phi_{ki}^1 & \phi_{ki}^2 & \phi_{ki}^3 & \phi_{ki}^k \end{bmatrix}, \quad i = 1, \dots, p; a = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_k \end{bmatrix} \quad \text{and } \varepsilon_t = \begin{bmatrix} \varepsilon_t^1 \\ \varepsilon_t^2 \\ \vdots \\ \varepsilon_t^k \end{bmatrix} \tag{1.8}$$

X_t : Vector of K non-stationary endogenous variables. a : Matrix of coefficients to be estimated. ε_t : Vector of error terms.

Note that the components of ε_t are uncorrelated with each other, and $\Sigma_\varepsilon = E(\xi_t \xi_t')$ denotes the $(k \times k)$ variance-covariance matrix of the errors, which are white noise with constant variance and zero mean.

1.3.1 Stationarity Condition

A VAR(p) model is stationary if it satisfies the following conditions :

- $E(X_t) = \mu \ \forall t$,
- $Var(X_t) < \infty$,
- $\Gamma(t+h) = cov(X_t, X_{t+h}) = E[(X_t - \mu)(X_{t-h} - \mu)']$, $\forall t$.

The determinant : $\det(1 - \phi_1 Z - \phi_2 Z^2 - \phi_3 Z^3 - \cdots - \phi_p Z^p)$ has its roots outside the unit circle, where Z is the lag operator.

1.3.2 Determining the Order of a VAR Model

The order of a VAR model is determined using two information criteria : the Akaike Information Criterion (AIC) and the Schwartz Information Criterion (SIC).

AIC $(p) = \ln(\det|\Sigma_\varepsilon|) + \frac{2k^2 p}{n}$

SIC $(p) = \ln(\det|\Sigma_\varepsilon|) + \frac{k^2 \ln(n)}{n}$

Where :

- k : Number of variables in the system ;
- n : Number of observations ;
- p : Number of lags ;
- Σ_ε : Maximum likelihood estimator of the variance-covariance matrix of the residuals of the VAR(p) model.

These criteria can be used to select the optimal number of lags in the VAR model. The smaller the information criterion value, the better the model.

1.3.3 Vector Error Correction Model (VECM)

The VECM is a model that captures the adjustments leading to a long-term equilibrium. It integrates both short-term and long-term dynamics.

Definition 1.3.1. Consider a vector of variables (X_t) integrated of order 1. The idea behind vector error correction models is to consider relationships of the form :

$$\Delta X_t = \lambda z_{t-1} + \beta_1 \Delta X_{t-1} + \beta_2 \Delta X_{t-2} + \cdots + \beta_p \Delta X_{t-p} + \varepsilon_t \tag{1.9}$$

Where :

λ : Speed of adjustment matrix toward the long-term target (Reduction in the growth rate at time t. If this coefficient is not significant and not negative, there is no return-to-equilibrium phenomenon).

z_{t-1} : Measures the disequilibrium between the cointegrated variables, representing the error correction term, which accounts for long-term equilibrium.

Remark 1.3.1. To use a VECM representation, the variables must be integrated of order 1.

Chapter 2

Estimation Methods for VAR Models

Contents

Estimating Vector Autoregressive (VAR) models involves selecting appropriate methods to determine the parameters of the model. VAR models are widely used in econometrics and time series analysis to capture the linear interdependencies among multiple time series. Below are the key estimation methods for VAR models :

2.1 Ordinary Least Squares (OLS)

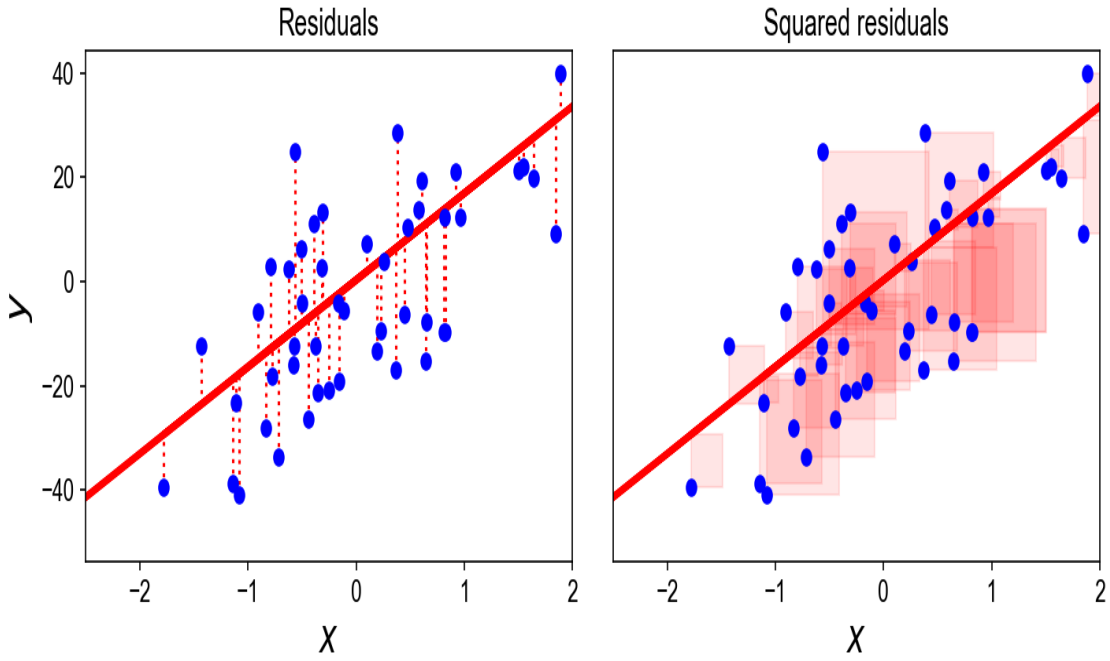

FIGURE 2.1 – OLS estimation schematic

2.1.1 Multiple Linear Regression Model and OLS

2.1.1.1 Multiple Linear Regression Model (MLR) ; General Form

Definition 2.1.1. The goal is to predict and/or explain the values of a quantitative variable Y based on the values of p variables $X_1, \ldots, X_p$. In this context, we aim to "explain Y using $X_1, \ldots, X_p$". Here, Y is referred to as the "dependent variable" (or "response variable"), and $X_1, \ldots, X_p$ are called the "independent variables" (or "explanatory variables"). To achieve this, we have data consisting of n observations of $(Y, X_1, \ldots, X_p)$, denoted as :

$$(y_1, x_{1,1}, \ldots, x_{p,1}), (y_2, x_{1,2}, \ldots, x_{p,2}), \ldots, (y_n, x_{1,n}, \ldots, x_{p,n}). \tag{2.1}$$

These observations are typically organized in a table :

Y	X_1	$\dots$	X_p
y_1	$x_{1,1}$	$\dots$	$x_{p,1}$
y_2	$x_{1,2}$	$\dots$	$x_{p,2}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
y_n	$x_{1,n}$	$\dots$	$x_{p,n}$

If a linear relationship between Y and $X_1, \dots, X_p$ is plausible, we can consider the multiple linear regression model (MLR). Its general form is :

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon,$$

where :

- $\beta_0, \dots, \beta_p$ are unknown real coefficients,
- ε is a quantitative variable with a mean of zero, independent of $X_1, \dots, X_p$, representing a sum of random and multifactorial errors (e.g., measurement errors, unpredictable effects, omitted variables, etc.).

The β coefficients of the true, but unknown model are estimated by the OLS regression model, yielding $\tilde{\beta}$ coefficients, by minimising the residual sum of squares (RSS) :

$$\mathrm{RSS} = \sum_{i=1}^{n} (y_i - \tilde{y_i})^2 = \sum_{i=1}^{n} (y_i - X_i \tilde{\beta})^2 \tag{2.2}$$

2.1.1.2 Modeling the Variables

Definition 2.1.2. The variables are modeled as real random variables (defined on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$), keeping the same notations by convention. Based on these, the MLR model is characterized by : for all $i \in \{1, \dots, n\}$,

- $(x_{1,i}, \dots, x_{p,i})$ is a realization of the real random vector $(X_1, \dots, X_p)$,
- given that $(X_1, \dots, X_p) = (x_{1,i}, \dots, x_{p,i})$, y_i is a realization of

  $$Y_i = \beta_0 + \beta_1 x_{1,i} + \cdots + \beta_p x_{p,i} + \varepsilon_i,$$

  where ε_i is a random variable independent of $X_1, \dots, X_p$ with $\mathbb{E}(\varepsilon_i) = 0$.

2.1.1.3 Matrix Form of the MLR Model

Definition 2.1.3. The MLR model can then be written in matrix form as :

$$Y = X\beta + \varepsilon,$$

where

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \quad X = \begin{pmatrix} 1 & x_{1,1} & \cdots & x_{p,1} \\ 1 & x_{1,2} & \cdots & x_{p,2} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1,n} & \cdots & x_{p,n} \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

2.1.1.4 Ordinary Least Squares Estimator ; A Central Result

Theorem 2.1.1. *Let $\|\cdot\|$ denote the Euclidean norm : for any column vector x, $\|x\|^2 = x^t x =$ sum of the squares of the components of x. Starting from the MLR model written in matrix form :*

$$Y = X\beta + \varepsilon,$$

an ordinary least squares (OLS) estimator $\hat{\beta}$ of β satisfies :

$$\tilde{\beta} \in \arg\min_{\beta\in\mathbb{R}^{p+1}} \|Y - X\beta\|^2.$$

We assume that X is of full column rank, there exists no non-zero column vector x with $p+1$ components such that $Xx = \mathbf{0}$ (this ensures the existence of $(X^tX)^{-1}$). Then, $\tilde{\beta}$ is unique and is given by the formula :

$$\tilde{\beta} = (X^tX)^{-1}X^tY. \tag{2.3}$$

Proposition 2.1.1. *Let*

$$f(\beta) = \|Y - X\beta\|^2, \quad \beta \in \mathbb{R}^{p+1}.$$

Since $\tilde{\beta} \in \arg\min_{\beta\in\mathbb{R}^{p+1}} f(\beta)$, $\tilde{\beta}$ is an extremum of $f(\beta)$, and

$$\tilde{\beta} \text{ extremum of } f(\beta) \implies \frac{\partial}{\partial\beta_j} f(\tilde{\beta}) = 0, \quad j \in \{0,\dots,p\}.$$

Let us simplify the expression for $f(\beta)$. Using the formulas : $(A+B)^t = A^t + B^t$ and $(AB)^t = B^tA^t$, we have

$$f(\beta) = \|Y - X\beta\|^2 = (Y - X\beta)^t(Y - X\beta) = (Y^t - (X\beta)^t)(Y - X\beta)$$

$$= (Y^t - \beta^tX^t)(Y - X\beta) = Y^tY - Y^tX\beta - \beta^tX^tY + \beta^tX^tX\beta.$$

Since $Y^tX\beta$ is the multiplication of a row vector Y^t by a column vector $X\beta$, it is a real number. Consequently, it is equal to its transpose ; we have

$$Y^tX\beta = (Y^tX\beta)^t = (X\beta)^t(Y^t)^t = \beta^tX^tY.$$

Thus,

$$f(\beta) = Y^tY - 2\beta^tX^tY + \beta^tX^tX\beta.$$

For all $j \in \{0,\dots,p\}$, let us determine the partial derivative $\frac{\partial}{\partial\beta_j} f(\beta)$. Let e_j be the column vector with $p+1$ components, where all components are zero except the $(j+1)$-th component, which is 1. Using the formula :

$$(u(x)v(x))' = u'(x)v(x) + u(x)v'(x),$$

we have

$$\frac{\partial}{\partial\beta_j}f(\beta) = \frac{\partial}{\partial\beta_j}\left(Y^tY - 2\beta^tX^tY + \beta^tX^tX\beta\right)$$
$$= \frac{\partial}{\partial\beta_j}(Y^tY) - 2\frac{\partial}{\partial\beta_j}(\beta^tX^tY) + \frac{\partial}{\partial\beta_j}(\beta^tX^tX\beta)$$
$$= 0 - 2e_j^tX^tY + e_j^tX^tX\beta + \beta^tX^tXe_j.$$

Since $e_j^tX^tX\beta$ is the multiplication of a row vector $e_j^tX^t$ by a column vector $X\beta$, it is a real number. Consequently, it is equal to its transpose ; we have

$$e_j^tX^tX\beta = (e_j^tX^tX\beta)^t = (X\beta)^t(e_j^tX^t)^t = \beta^tX^tXe_j.$$

Therefore,

$$\frac{\partial}{\partial\beta_j}f(\beta) = -2e_j^tX^tY + 2e_j^tX^tX\beta.$$

It follows that

$$\frac{\partial}{\partial\beta_j}f(\hat{\beta}) = 0 \implies -2e_j^tX^tY + 2e_j^tX^tX\hat{\beta} = 0 \implies e_j^tX^tX\hat{\beta} = e_j^tX^tY.$$

Since this holds for all $j \in \{0,\dots,p\}$ and $e_j^tX^tX\hat{\beta}$ computes the j-th row of the matrix $X^tX\hat{\beta}$, it follows that

$$\frac{\partial}{\partial\beta_j}f(\hat{\beta}) = 0, \quad j \in \{0,\dots,p\} \implies X^tX\hat{\beta} = X^tY.$$

Since $(X^tX)^{-1}$ exists, the equality $(X^tX)^{-1}X^tX = I_{p+1}$ implies

$$X^tX\hat{\beta} = X^tY \implies (X^tX)^{-1}X^tX\hat{\beta} = (X^tX)^{-1}X^tY \implies \hat{\beta} = (X^tX)^{-1}X^tY.$$

In conclusion, we have

$$\hat{\beta} \text{ extremum of } f(\beta) \implies \hat{\beta} = (X^tX)^{-1}X^tY.$$

It remains to show that $\hat{\beta}$ is indeed a minimum for $f(\beta)$. To do this, we compute the Hessian matrix

$$H(f) = \left(\frac{\partial^2}{\partial\beta_j\partial\beta_k}f(\beta)\right)_{(j,k)\in\{0,\dots,p\}^2}$$

and show that it is positive definite : for any non-zero column vector x with $p+1$ components, we have $x^tH(f)x > 0$. For all $(j,k) \in \{0,\dots,p\}^2$, we have

$$\frac{\partial^2}{\partial\beta_j\partial\beta_k}f(\beta) = \frac{\partial}{\partial\beta_k}\left(\frac{\partial}{\partial\beta_j}f(\beta)\right) = \frac{\partial}{\partial\beta_k}\left(-2e_j^tX^tY + 2e_j^tX^tX\beta\right)$$

$$= -2\frac{\partial}{\partial\beta_k}(e_j^tX^tY) + 2\frac{\partial}{\partial\beta_k}(e_j^tX^tX\beta) = 0 + 2e_j^tX^tXe_k = 2e_j^tX^tXe_k.$$

Therefore,

$$H(f) = \left(2e_j^t X^t X e_k\right)_{(j,k)\in\{0,\dots,p\}^2} = 2X^tX.$$

For any non-zero vector $x = \begin{pmatrix} x_0 \\ \vdots \\ x_p \end{pmatrix}$*, since* X *is of full column rank, we have*

$$x^tH(f)x = x^t(2X^tX)x = 2x^tX^tXx = 2(Xx)^t(Xx) = 2\|Xx\|^2 > 0.$$

Thus, $H(f)$ *is positive definite, and* $\hat{\beta}$ *is indeed a minimum for* $f(\beta)$*. We conclude that*

$$\tilde{\beta} \in \arg\min_{\beta\in\mathbb{R}^{p+1}} \|Y - X\beta\|^2 \iff \hat{\beta} = (X^tX)^{-1}X^tY.$$

2.1.1.5 OLS Estimator of β_j

Theorem 2.1.2. *The OLS estimator* $\hat{\beta}$ *of* $\beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix}$ *is written as* $\hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \vdots \\ \hat{\beta}_p \end{pmatrix}$.

Thus, for all $j \in \{0,\dots,p\}$*, the* $(j+1)$*-th component of* $\hat{\beta}$*, denoted* $\hat{\beta}_j$*, is the OLS estimator of* β_j*.*

2.1.1.6 Estimator of the Mean Value

Theorem 2.1.3. *The mean value of* Y *when* $(X_1,\dots,X_p) = (x_1,\dots,x_p) = x$ *is the unknown real number :*

$$y_x = \mathbb{E}(Y \mid \{X_1,\dots,X_p\} = x) = \beta_0 + \beta_1x_1 + \cdots + \beta_px_p.$$

An estimator of y_x *is :*

$$\tilde{Y}_x = \tilde{\beta}_0 + \tilde{\beta}_1x_1 + \cdots + \tilde{\beta}_px_p.$$

By defining $x_\bullet = (1, x_1,\dots,x_p)$*, we have* $y_x = x_\bullet\beta$ *and* $\tilde{Y}_x = x_\bullet\tilde{\beta}$*.*

2.1.1.7 Point Estimates

Theorem 2.1.4. *A point estimate of* β *is the realization* b *of* $\hat{\beta}$ *corresponding to the data :*

$$b = (X^tX)^{-1}X^ty,$$

where $y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$*. We can write* b *as* $b = \begin{pmatrix} b_0 \\ \vdots \\ b_p \end{pmatrix}$*. Thus, for all* $j \in \{0,\dots,p\}$*, the* $(j+1)$*-th component of* b*, denoted* b_j*, is a point estimate of* β_j*.*

Let $x_\bullet = (1, x_1, \ldots, x_p)$. A point estimate of $y_x = x_\bullet\beta$ is the realization d_x of $\hat{Y}_x = x_\bullet\hat{\beta}$ corresponding to the data :

$$d_x = x_\bullet b = b_0 + b_1x_1 + \cdots + b_px_p.$$

We say that d_x is the predicted value of Y when $(X_1, \ldots, X_p) = x$.

2.2 Maximum Likelihood Estimation

2.2.1 The Likelihood Function

Assuming that the distribution of the process is known, maximum likelihood **ML** estimation is an alternative to **LS** estimation. We will consider **ML** estimation under the assumption that the VAR(p) processy_t is Gaussian.More precisely, Estimation of Vector Autoregressive Processes

$$\mathbf{u} = \mathrm{vec}(U) = \begin{bmatrix} \mathbf{u}_1 \\ \vdots \\ \mathbf{u}_T \end{bmatrix} \sim \mathcal{N}(\mathbf{0}, I_T \otimes \Sigma_u). \tag{2.4}$$

In other words,the probability density of $\mathbf{u}$ is :

$$f_{\mathbf{u}}(\mathbf{u}) = \frac{1}{(2\pi)^{KT/2}} |I_T \otimes \Sigma_u|^{-1/2} \exp\left[-\tfrac{1}{2}\mathbf{u}'(I_T \otimes \Sigma_u^{-1})\mathbf{u}\right]. \tag{2.5}$$

Moreover,

$$\mathbf{u} = \begin{bmatrix} I_K & 0 & \cdots & 0 & \cdots & \cdots & 0 \\ -A_1 & I_K & 0 & \cdots & \cdots & 0 & \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \\ -A_p & -A_{p-1} & \cdots & I_K & 0 & & \\ 0 & -A_p & \ddots & \vdots & \vdots & & \\ \vdots & \vdots & \ddots & \vdots & \vdots & & \\ 0 & 0 & \cdots & -A_p & \cdots & \cdots & I_K \end{bmatrix} (\mathbf{y} - \boldsymbol{\mu}^*) + \begin{bmatrix} -A_1 & -A_2 & \cdots & -A_p \\ -A_2 & -A_3 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ -A_p & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix} (Y_0 - \boldsymbol{\mu}), \tag{2.6}$$

$$
\begin{aligned}
\ln l(\boldsymbol{\mu}, \boldsymbol{\alpha}, \Sigma_u) &= -\frac{KT}{2}\ln(2\pi) - \frac{T}{2}\ln|\Sigma_u| \\
&\quad - \frac{1}{2}[\mathbf{y} - \boldsymbol{\mu}^* - (X' \otimes I_K)\boldsymbol{\alpha}]'(I_T \otimes \Sigma_u^{-1})[\mathbf{y} - \boldsymbol{\mu}^* - (X' \otimes I_K)\boldsymbol{\alpha}] \\
&= -\frac{KT}{2}\ln 2\pi - \frac{T}{2}\ln|\Sigma_u| \\
&\quad - \frac{1}{2}\sum_{t=1}^{T}\left[(y_t - \mu) - \sum_{i=1}^{p} A_i(y_{t-i} - \mu)\right]' \times \Sigma_u^{-1}\left[(y_t - \mu) - \sum_{i=1}^{p} A_i(y_{t-i} - \mu)\right] \\
&= -\frac{KT}{2}\ln 2\pi - \frac{T}{2}\ln|\Sigma_u| \\
&\quad - \frac{1}{2}\sum_{t}\left(y_t - \sum_i A_i y_{t-i}\right)' \Sigma_u^{-1}\left(y_t - \sum_i A_i y_{t-i}\right) \\
&\quad + \mu'\left(I_K - \sum_i A_i\right)' \Sigma_u^{-1}\sum_t\left(y_t - \sum_i A_i y_{t-i}\right) \\
&\quad - \frac{T}{2}\mu'\left(I_K - \sum_i A_i\right)' \Sigma_u^{-1}\left(I_K - \sum_i A_i\right)\mu \\
&= -\frac{KT}{2}\ln 2\pi - \frac{T}{2}\ln|\Sigma_u| - \frac{1}{2}\mathrm{tr}[(Y^0 - AX)'\Sigma_u^{-1}(Y^0 - AX)] \qquad (2.7)
\end{aligned}
$$

where :

- $Y^0 := (y_1 - \mu, \ldots, y_T - \mu)$
- $A := (A_1, \ldots, A_p)$ These different expressions of the log-likelihood function will be useful in the following.

2.2.2 The ML Estimators

In order to determine the **ML** estimators of μ, α, and Σ_u, the system of first order partial derivatives is needed :

$$
\begin{aligned}
\frac{\partial \ln l}{\partial \mu} &= \left(I_K - \sum_i A_i\right)' \Sigma_u^{-1}\sum_t\left(y_t - \sum_i A_i y_{t-i}\right) \\
&\quad - T\left(I_K - \sum_i A_i\right)' \Sigma_u^{-1}\left(I_K - \sum_i A_i\right)\mu \\
&= [I_K - A(j \otimes I_K)]'\Sigma_u^{-1}\left[\sum_t (y_t - \mu - AY_{t-1}^0)\right], \qquad (2.8)
\end{aligned}
$$

where :

- Y_t^0
- $j := (1, \ldots, 1)'$ is a $(p \times 1)$ vector of ones

$$\frac{\partial \ln l}{\partial \boldsymbol{\alpha}} = (X \otimes I_K)(I_T \otimes \Sigma_u^{-1})[\mathbf{y} - \boldsymbol{\mu}^* - (X' \otimes I_K)\boldsymbol{\alpha}]$$
$$= (X \otimes \Sigma_u^{-1})(\mathbf{y} - \boldsymbol{\mu}^*) - (XX' \otimes \Sigma_u^{-1})\boldsymbol{\alpha}, \tag{2.9}$$
$$\frac{\partial \ln l}{\partial \Sigma_u} = -\frac{T}{2}\Sigma_u^{-1} + \frac{1}{2}\Sigma_u^{-1}(Y^0 - AX)(Y^0 - AX)'\Sigma_u^{-1}. \tag{2.10}$$

Equating to zero gives the system of normal equations which can be solved for the estimators :

$$\tilde{\boldsymbol{\mu}} = \frac{1}{T}\left(I_K - \sum_i \tilde{A}_i\right)^{-1} \sum_t \left(y_t - \sum_i \tilde{A}_i y_{t-i}\right), \tag{2.11}$$
$$\tilde{\boldsymbol{\alpha}} = \left((\tilde{X}\hat{X}')^{-1}\tilde{X} \otimes I_K\right)(\mathbf{y} - \tilde{\boldsymbol{\mu}}^*), \tag{2.12}$$
$$\tilde{\Sigma}_u = \frac{1}{T}(\tilde{Y}^0 - \tilde{A}\hat{X})(\tilde{Y}^0 - \tilde{A}\hat{X})', \tag{2.13}$$

where $\tilde{X}$ and $\tilde{Y}^0$ are obtained from X and Y^0 by replacing μ with $\tilde{\mu}$.

2.2.3 Properties of the ML Estimators

The ML estimators μ and α are identical to the LS estimators. Thus, $\tilde{\boldsymbol{\mu}}$ and $\tilde{\boldsymbol{\alpha}}$ are consistent estimators if $\mathbf{y}_t$ is a stationary, stable Gaussian VAR(p) process, and $\sqrt{T}(\tilde{\mu} - \mu)$ and $\sqrt{T}(\tilde{\alpha} - \alpha)$ are asymptotically normally. The information matrix is :

$$I(\boldsymbol{\delta}) = -E\left[\frac{\partial^2 \ln l}{\partial \boldsymbol{\delta} \partial \boldsymbol{\delta}'}\right], \tag{2.14}$$

where $\boldsymbol{\delta}' := (\boldsymbol{\mu}', \boldsymbol{\alpha}', \boldsymbol{\sigma}')$with $\boldsymbol{\sigma} := \text{vech}(\Sigma_u)$ Note that vech is a column stacking operator that stacks only the elements on and below the main diagonal of Σ_u. It is related to the vec operator by the $(\frac{1}{2}K(K+1) \times K^2)$ elimination matrix $\boldsymbol{L}_K$, that is, $\text{vech}(\Sigma_u) = \boldsymbol{L}_K \text{vec}(\Sigma_u)$ or, defining $\boldsymbol{\omega} := \text{vec}(\Sigma_u)$, $\boldsymbol{\sigma} = \boldsymbol{L}_K\omega$ For instance, for $K = 3$,

$$\omega = \text{vec}(\Sigma_u) = \text{vec}\begin{bmatrix}\sigma_{11} & \sigma_{12} & \sigma_{13}\\ \sigma_{12} & \sigma_{22} & \sigma_{23}\\ \sigma_{13} & \sigma_{23} & \sigma_{33}\end{bmatrix} = (\sigma_{11}, \sigma_{12}, \sigma_{13}, \sigma_{12}, \sigma_{22}, \sigma_{23}, \sigma_{13}, \sigma_{23}, \sigma_{33})'$$

and,

$$\boldsymbol{\sigma} = \text{vech}(\Sigma_u) = \mathbf{L}_3\boldsymbol{\omega} = \begin{bmatrix}\sigma_{11}\\ \sigma_{12}\\ \sigma_{13}\\ \sigma_{22}\\ \sigma_{23}\\ \sigma_{33}\end{bmatrix}.$$

where $\boldsymbol{\delta}$ contains only the unique elements of Σ_u.

The asymptotic covariance matrix of the**ML** estimator $\tilde{\boldsymbol{\delta}}$ is :

$$\lim_{T\to\infty}\left[\frac{\mathcal{I}(\boldsymbol{\delta})}{T}\right]^{-1}.$$

In order to determine this matrix, we need the second order partial derivatives of the log-likelihood. From (2.8) to (2.10) we get

$$\frac{\partial^2 \ln l}{\partial \boldsymbol{\mu} \partial \boldsymbol{\mu}'} = -T \left(I_K - \sum_i A_i \right)' \Sigma_u^{-1} \left(I_K - \sum_i A_i \right), \tag{2.15}$$

$$\frac{\partial^2 \ln l}{\partial \boldsymbol{\alpha} \partial \boldsymbol{\alpha}'} = - \left(XX' \otimes \Sigma_u^{-1} \right), \tag{2.16}$$

$$\begin{aligned} \frac{\partial^2 \ln l}{\partial \boldsymbol{\omega} \partial \boldsymbol{\omega}'} = & \frac{T}{2} \left(\Sigma_u^{-1} \otimes \Sigma_u^{-1} \right) - \frac{1}{2} \left(\Sigma_u^{-1} \otimes \Sigma_u^{-1} UU' \Sigma_u^{-1} \right) \\ & - \frac{1}{2} \left(\Sigma_u^{-1} UU^\top \Sigma_u^{-1} \otimes \Sigma_u^{-1} \right), \end{aligned} \tag{2.17}$$

where $\boldsymbol{\omega} = \mathrm{vec}(\Sigma_u)$.

$$\begin{aligned} \frac{\partial^2 \ln l}{\partial \boldsymbol{\mu} \partial \boldsymbol{\alpha}'} = & - \left[I_K - (j' \otimes I_K) A' \right] \Sigma_u^{-1} \sum_t Y_{t-1}^0 \otimes I_K \\ & - \left(\sum_t \mathbf{u}_t' \Sigma_u^{-1} \otimes I_K \right) (I_K \otimes j' \otimes I_K) \frac{\partial \mathrm{vec}(A')}{\partial \boldsymbol{\alpha}'}, \end{aligned} \tag{2.18}$$

$$\frac{\partial^2 \ln l}{\partial \boldsymbol{\omega} \partial \boldsymbol{\mu}'} = -\frac{1}{2} \left(\Sigma_u^{-1} \otimes \Sigma_u^{-1} \right) \left[(I_K \otimes U) \frac{\partial \mathrm{vec}(U')}{\partial \boldsymbol{\mu}'} + (U \otimes I_K) \frac{\partial \mathrm{vec}(U)}{\partial \boldsymbol{\mu}'} \right]. \tag{2.19}$$

and

$$\frac{\partial^2 \ln l}{\partial \boldsymbol{\omega} \partial \boldsymbol{\alpha}'} = -\frac{1}{2} \left(\Sigma_u^{-1} \otimes \Sigma_u^{-1} \right) \left[(I_K \otimes UX') \frac{\partial \mathrm{vec}(A')}{\partial \boldsymbol{\alpha}'} + (UX' \otimes I_K) \right]. \tag{2.20}$$

It is obvious from (2.18) that

$$\lim_{T \to \infty} T^{-1} E \left(\frac{\partial^2 \ln l}{\partial \boldsymbol{\mu} \partial \boldsymbol{\alpha}'} \right) = 0,$$

since $E \left(\frac{1}{T} \sum_t Y_{t-1}^0 \right) \to 0$. Furthermore, From (2.19), we have :

$$E \left(\frac{\partial^2 \ln l}{\partial \boldsymbol{\omega} \partial \boldsymbol{\mu}'} \right) = 0,$$

because $E(U) = 0$ and $\partial \mathrm{vec}(U')/\partial \boldsymbol{\mu}'$ is constant.

From (2.20) :

$$\lim_{T \to \infty} T^{-1} E \left(\frac{\partial^2 \ln l}{\partial \boldsymbol{\omega} \partial \boldsymbol{\alpha}'} \right) = 0,$$

because $E(UX'/T) \to 0$. Thus, $\lim_{T \to \infty} \mathcal{I}(\boldsymbol{\delta})/T$ is block diagonal and we get the asymptotic distributions of $\boldsymbol{\mu}, \boldsymbol{\alpha}$ and $\boldsymbol{\sigma}$ as follows. Multiplying minus the inverse of(2.15) by T gives the asymptotic covariance matrix of the **ML** estimator for the mean vectorμ, that is,

$$\sqrt{T}(\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \left(I_K - \sum_{i=1}^{p} A_i\right)^{-1} \Sigma_u \left(I_K - \sum_{i=1}^{p} A_i'\right)^{-1}\right).$$

Hence, $\tilde{\mu}$ has the same asymptotic distribution as$\bar{y}$. In other words, the two estimators forμ are asymptotically equivalent and, under the present conditions, this fact implies that $\bar{y}$ is asymptotically efficient because the **ML** estimator is asymptotically efficient. The asymptotic equivalence of $\tilde{\mu}$ and $\tilde{y}$ can also be seen from (2.11). Taking the limit of T^{-1} times the expectation of minus (2.16) gives$\Gamma_Y(0) \otimes \Sigma_u^{-1}$. Note that $E(XX'/T)$ is not strictly equal to $\Gamma_Y(0)$ because we have assumed fixed initial values $y_{-p+1}, \ldots, y_0$. However, asymptotically, as T goes to infinity, the impact of the initial values vanishes. Thus, we get

$$\sqrt{T}(\tilde{\boldsymbol{\alpha}} - \boldsymbol{\alpha}) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \Gamma_Y(0)^{-1} \otimes \Sigma_u\right).$$

Of course, this result also follows from the equivalence of the **ML** and **LS** estimators. From (2.17) and $E(UU') = T\Sigma_u$:

$$E\left(\frac{\partial^2 \ln l}{\partial \boldsymbol{\omega} \partial \boldsymbol{\omega}'}\right) = -\frac{T}{2}(\Sigma_u^{-1} \otimes \Sigma_u^{-1}).$$

Let $\boldsymbol{D}_K$ be the $(K^2 \times \frac{1}{2}K(K+1))$ duplication matrix so that : $\boldsymbol{\omega} = \mathbf{D}_K \boldsymbol{\sigma}$, we get

$$\frac{\partial^2 \ln l}{\partial \boldsymbol{\sigma} \partial \boldsymbol{\sigma}'} = \frac{\partial \boldsymbol{\omega}'}{\partial \boldsymbol{\sigma}} \frac{\partial^2 \ln l}{\partial \boldsymbol{\omega} \partial \boldsymbol{\omega}'} \frac{\partial \boldsymbol{\omega}}{\partial \boldsymbol{\sigma}'} = \mathbf{D}_K' \frac{\partial^2 \ln l}{\partial \boldsymbol{\omega} \partial \boldsymbol{\omega}'} \mathbf{D}_K,$$

leading to :

$$\sqrt{T}(\tilde{\boldsymbol{\sigma}} - \boldsymbol{\sigma}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \Sigma_{\tilde{\sigma}}),$$

where :

$$\begin{aligned} \Sigma_{\tilde{\sigma}} = -TE\left(\frac{\partial^2 \ln l}{\partial \boldsymbol{\sigma} \partial \boldsymbol{\sigma}'}\right)^{-1} &= 2\left[\mathbf{D}_K'(\Sigma_u^{-1} \otimes \Sigma_u^{-1})\mathbf{D}_K\right]^{-1} \\ &= 2\mathbf{D}_K^{+}(\Sigma_u \otimes \Sigma_u)\mathbf{D}_K^{+'}. \end{aligned} \tag{2.21}$$

Here $\boldsymbol{D}_K^{+} = (\mathbf{D}_K'\boldsymbol{D}_K)^{-1}\mathbf{D}_K'$ is the Moore-Penrose inverse of the duplication matrix $\boldsymbol{D_K}$

Proposition 2.2.1 (Asymptotic Properties of ML Estimators)**.** *Let y_t be a stationary, stable Gaussian VAR(p) process. Then the ML estimators $\tilde{\boldsymbol{\mu}}$, $\tilde{\boldsymbol{\alpha}}$, and $\tilde{\boldsymbol{\sigma}} = vech(\tilde{\Sigma}_u)$ given in* (2.11)-(2.13)*are consistent and*

$$\sqrt{T}\begin{bmatrix} \tilde{\boldsymbol{\mu}} - \boldsymbol{\mu} \\ \tilde{\boldsymbol{\alpha}} - \boldsymbol{\alpha} \\ \tilde{\boldsymbol{\sigma}} - \boldsymbol{\sigma} \end{bmatrix} \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \Sigma_{\tilde{\mu}} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \Sigma_{\tilde{\alpha}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \Sigma_{\tilde{\sigma}} \end{bmatrix}\right).$$

So that $\tilde{\mu}$ is asymptotically independent of $\tilde{\alpha}$ and $\tilde{\Sigma}_u$ and $\tilde{\alpha}$ is asymptotically independent of $\tilde{\mu}$ and $\tilde{\Sigma}_u$. The covariance matrices are

$$\Sigma_{\tilde{\mu}} = \left(I_K - \sum_i A_i\right)^{-1} \Sigma_u \left(I_K - \sum_i A_i'\right)^{-1}, \tag{2.22}$$

$$\Sigma_{\tilde{\alpha}} = \Gamma_Y(0)^{-1} \otimes \Sigma_u, \tag{2.23}$$

$$\Sigma_{\tilde{\sigma}} = 2\mathbf{D}_K^+(\Sigma_u \otimes \Sigma_u)\mathbf{D}_K^{+\prime}. \tag{2.24}$$

They may be estimated consistently by replacing the unknown quantities by their **ML** estimators and estimating $\Gamma_Y(0)$ by $\tilde{X}\tilde{X}'/T$.

2.3 Penalized Methods (LASSO and Ridge) Regression

A penalised regression method is essentially a method of shrinking down a subsection of the β coefficients of the OLS regression model, in order to reduce the impact of features that are not as relevant to the model. Penalised regression methods are therefore sometimes known as 'shrinkage' methods, which force the regression model to shrink its coefficients towards 0 due the 'penalty' term imposed on its coefficients.

Recall that ordinary least squares (OLS) Regression selects predicted values β in order to minimize the residual sum of squares (RSS) :

$$\text{RSS} = \sum_{i=1}^{n}(y_i - X_i\tilde{\beta})^2 \tag{2.25}$$

Non-OLS regression selects coefficients in order to minimise a similar objective function. Specifically, penalised regression adds a penalty term (also known as a regularisation term or shrinkage term),

$$\lambda\|\beta\|_p$$

- $\|\beta\|_p$ is the p-norm of the coefficients :$\sum_i(|\beta_i|^p)^{\frac{1}{p}}$
- $\lambda > 0$ is a hyper-parameter, in this case known as the tuning parameter, defining how harshly the coefficients are penalised.

The aim is to now fit a penalised regression model to minimise the regularisation cost function :

$$\sum_{i=1}^{n}(y_i - X_i\tilde{\beta})^2 + \lambda\|\beta\|_p \tag{2.26}$$

Yielding penalised regression coefficients,

$$\tilde{\beta}_p = \arg\min_{\beta}\left(\|y - X\beta\|_2^2 + \lambda\|\beta\|_p^p\right) \tag{2.27}$$

According to [9] and [3] as cited in [10], penalized regression introduces a constraint in the equation model for having too many variables in the model. A coefficient that is close to zero or equal to zero will be assigned to the less contribute variables in the model. Thus, it allows the development of a linear regression model that is penalized.

Shrinkage means that the coefficients are reduced towards zero compared to the OLS parameter estimates. This is called regularization. Since the lowest possible estimate for a coefficient is zero, some – but not all - of the regularization models may be used for parameter selection (more about this later.)

Penalised Regression Methods

This report will focus on 2 commonly used penalised regression methods :

- **LASSO regression :** $L_1 = \sum_i |\beta_i|$ penalty term.
- **Ridge regression :** $L_2(\sum_i \|\beta_i\|_2^2)^{\frac{1}{2}}$ penalty term.

2.3.1 LASSO Regression

The LASSO (least absolute selection and shrinkage operator), first proposed by Robert Tibshirani [14] is the other main regularisation method. Where ridge doesn't set any coefficients exactly to 0, the L_1-penalty imposed by the lasso means that it can, in fact, perform **variable selection** in the linear model. This feature selection property is a key feature in correcting multicollinearity [13].

Definition 2.3.1. The lasso estimates are defined as :

$$\tilde{\beta}_{LASSO} = \arg\min \|y - X\beta\|_2^2 + \lambda\|\beta\|_1 \tag{2.28}$$

which minimise the quantity

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda \sum |\beta_j| \tag{2.29}$$

Though there is no closed-form expression available to calculate LASSO coefficients, the LASSO shares similar advantages with ridge regression :

- Can deal with $n < p$ problems (where the number of observations n is less than the number of predictors p).
- Shrinks coefficients to reduces the impact of predictor variables that are not relevant to the response.
- Decreases variance by introducing bias, leading to better generalization performance (on unseen data).

The plot below shows lasso regression coefficients against the shrinkage penalty. Again, each curve represents one of the 29 variables. As a result of the alternate shrinkage penalty, the plot shows a different picture of how parameter estimates become zero as we increase λ.

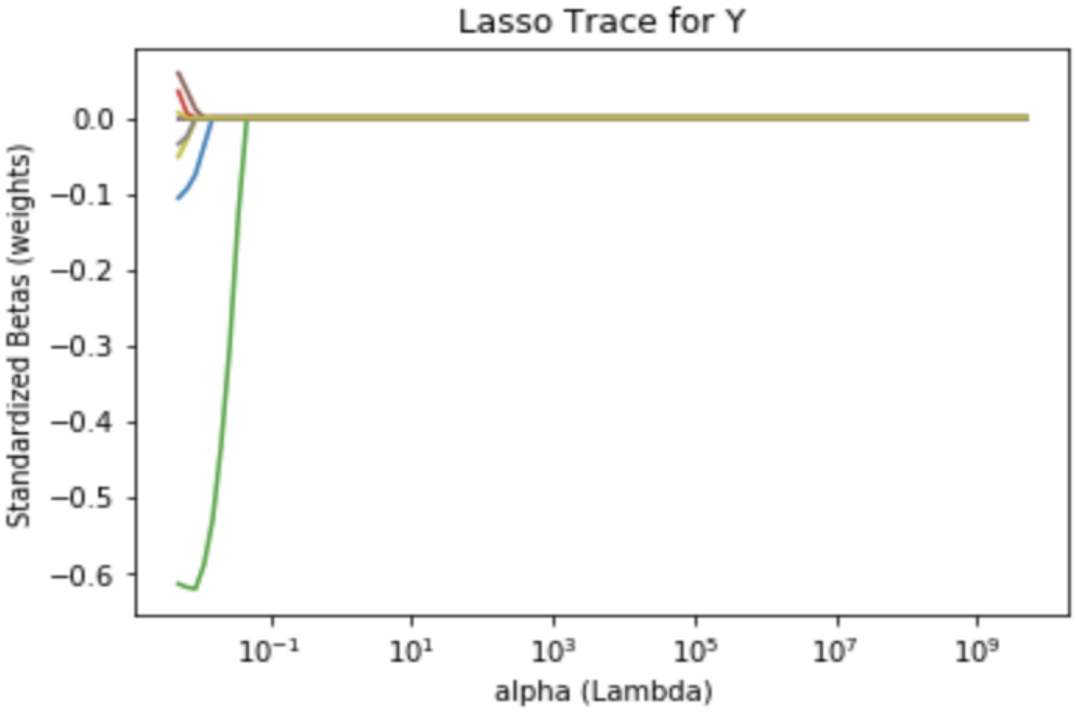

FIGURE 2.2 – Lasso Coefficient Trace Plot for Y vs. Lambda (a)

2.3.2 Ridge Regression

Ridge regression solves some of the shortcomings of linear regression. Ridge regression is an extension of the OLS method with an additional constraint. The OLS estimates are unconstrained, and might exhibit a large magnitude, and therefore large variance. In ridge regression, the coefficients are applied a penalty, so that they are shrunk towards zero, this also having the effect of reducing the variance and hence, the prediction error. Similar to the OLS approach, we choose the ridge coefficients to minimize a penalized residual sum of squares (RSS). As opposed to OLS, ridge regression provides biased estimators which have a low variance [7].

Definition 2.3.2. Ridge regression, first introduced by Hoerl and Kennard, 1970 [7] employs the L2 regularisation term, in order to penalise the squares of the regression coefficients. Firstly, the setup of the environment within which ridge regression can be performed.

- Start with fixed independent covariates (predictor variables) $\mathbf{x}_i \in \mathbb{R}^p, i = 1, \ldots, n$.
- Observe $y_i = f(\mathbf{x}_i) + \epsilon_i$, $i = 1, \ldots, n$.
- $f : \mathbb{R}^p \to \mathbb{R}$ unknown.
- $\mathrm{Var}[\epsilon_i] = \sigma^2$.

Ridge regression is similar to least squares but shrinks estimated coefficients toward zero for predictors deemed less relevant to the response. Given a response vector $y \in \mathbb{R}^n$ and predictor matrix $X^{n\times p}$, the ridge coefficients are defined as :

$$\tilde{\beta}_{ridge} = \arg\min_{\beta}(\|y - X\beta\|_2^2) + \lambda\|\beta\|_2^2 \tag{2.30}$$

Hence, we can interpret the ridge regression as yielding ridge coefficients that minimise the corresponding cost function :

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda\sum_{j=1}^{p}\beta_j^2 \tag{2.31}$$

The plot below shows ridge regression coefficients against the shrinkage penalty. Each curve represents one of the 29 variables. The left part of the plot shows OLS estimates, and lambda starts shrinking the parameter estimates at a differential rate as one moves towards the right. At the right-hand side of the plot all estimates become zero and there are no parameters in the model. In fact, coefficient estimates become zero before λ reaches the value 10.

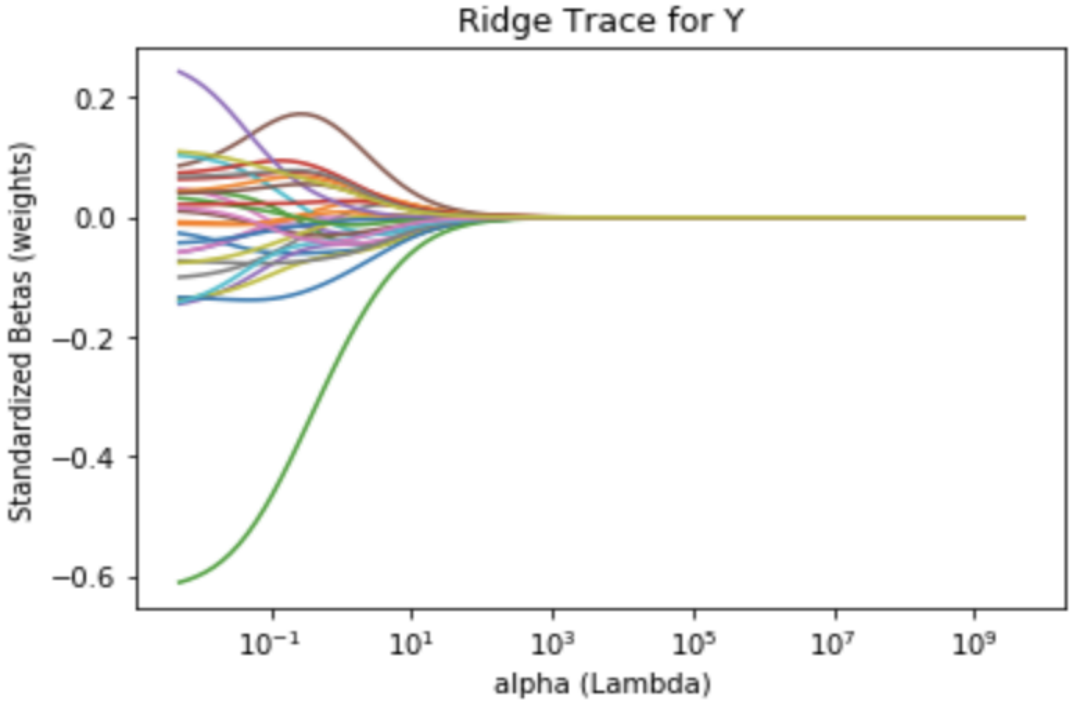

FIGURE 2.3 – Ridge Trace Plot – Coefficients vs. Log(Lambda)

Coefficients depend not only on the shrinkage parameter but also on the scaling of other parameters. Shrinkage may change the parameter estimate by a very large factor, therefore the standardization of regressors is important. Sci-kit can do this automatically.

In reality, ridge regression will always include all predictors, because the penalty (λ) will never reach exactly zero. Lambda would have to equal infinity for the coefficient to be shrunk to zero, which typically does not happen. As a result, a large number of variables may make it difficult to interpret when ridge regression is used.

2.3.3 Comparison of Lasso and Ridge Regression

Table 2.1 – Comparison of Lasso and Ridge Regression Properties

Aspect	**Ridge Regression**	**Lasso Regression**
Penalty Term	$\lambda \sum_{j=1}^{p} \beta_j^2$ (L2)	$\lambda \sum_{j=1}^{p} \lvert\beta_j\rvert$ (L1)
Coefficient Shrinkage	Shrinks uniformly (non-zero)	Can shrink coefficients to *exact zero*
Variable Selection	No (keeps all features)	Yes (automatic feature selection)
Solution Type	Closed-form solution	Requires numerical optimization
Computational Cost	$\mathcal{O}(p^3)$ (matrix inversion)	$\mathcal{O}(np^2)$ (convex optimization)

Although the penalties used for LASSO and Ridge regression are similar, they can yield very different solutions. **LASSO tends to produce sparse models** by driving some coefficients exactly to zero, effectively performing variable selection in addition to coefficient shrinkage. In contrast,**Ridge regression shrinks coefficients towards zero but typically all predictors have a non-zero coefficient**. Therefore, while Ridge regression reduces the magnitude of coefficients, it does not eliminate predictors, whereas LASSO can both shrink coefficients and eliminate uninformative predictors, leading to a simpler and more interpretable model.

Chapter 3

Application Of Estimation Methods For VAR Models

Contents

Vector Autoregressive (VAR) models hold a central position in econometrics for analyzing dynamic interactions between multivariate time series. This chapter aims to explore, in an applied manner, the key steps for estimating these models, emphasizing methodological rigor and result interpretation. In a context where economic and energy-related data (e.g., consumption, industrial production) are often complex and interdependent, mastering appropriate estimation techniques is essential to extract robust insights.

First, Section 3.1 addresses data selection and preprocessing, focusing on verifying stationarity (via the Augmented Dickey-Fuller (ADF) test) and necessary transformations (logarithm, differencing). Section 3.2 then explores the implementation of several estimation techniques, including Ordinary Least Squares (OLS), Maximum Likelihood Estimation (MLE), and penalized methods such as (LASSO, Ridge) regression, utilizing tools like R.

Finally, Section 3.3 presents a practical case study that illustrates these methods using simulated data, allowing for a comparative assessment of estimation methods with respect to predictive accuracy, computational efficiency, and suitability for high-dimensional datasets. The chapter concludes by summarizing key findings, identifying challenges encountered during estimation, and outlining possible directions for future work aimed at improving the robustness and efficiency of VAR model applications.

3.1 Identification of the series

To identify the nature of the time series, we will perform various tests to assess the presence of trends and verify stationarity.

- Graphical analysis of the series.
- Graphical analysis allows us to visualize the time evolution of the `consumption`, `temperature`, and `production` series, highlighting their respective trends and potential non-stationary behaviors.

Using simulated daily data for 2022 (`consumption`, `temperature`, and `production` series), we derive the following graphical results :

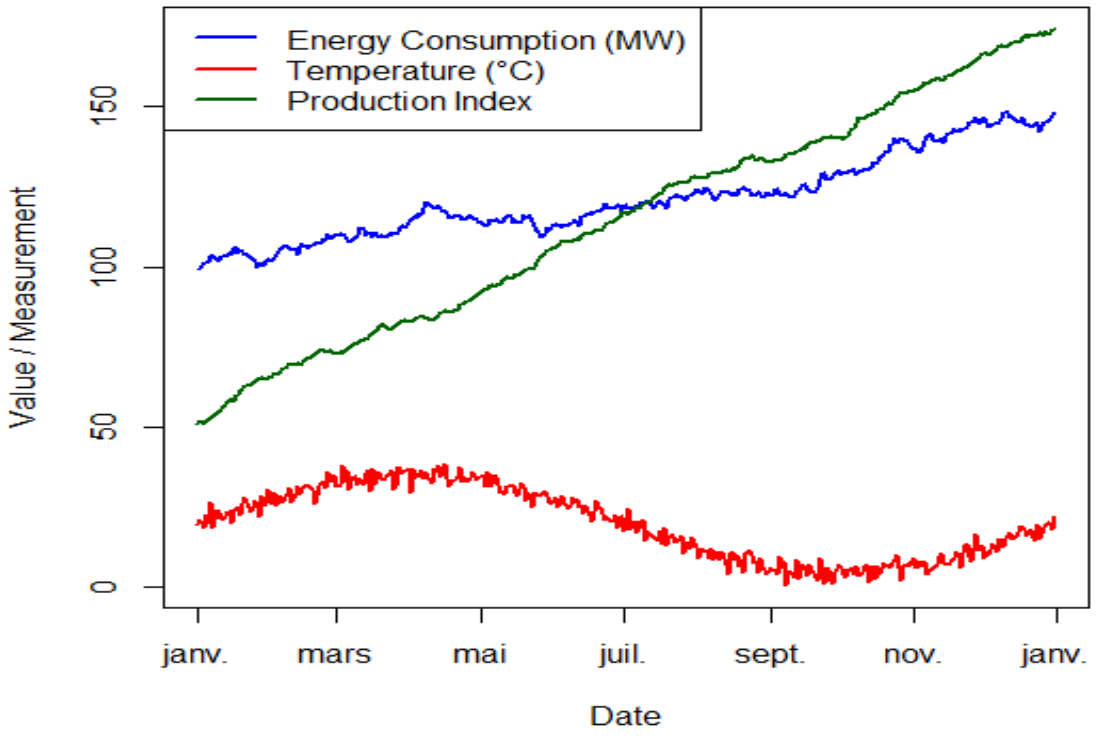

FIGURE 3.1 – Simulated Non-Stationary Series : Energy Consumption, Temperature, and Production (2022)

Figure 3.1 presents the evolution of three key variables over the course of one year : electricity consumption, temperature, and the production index. The electricity consumption series (blue line) exhibits a clear upward trend with short-term fluctuations, indicating potential non-stationarity in the mean. The temperature series (red line) follows a typical seasonal pattern, peaking during the summer months and declining in winter, reflecting the annual climatic cycle. In contrast, the production index (green line) shows a strong and steady upward trajectory throughout the year, which may reflect increased industrial or economic activity. The joint behavior of these series suggests potential dynamic interdependencies, which can be further explored using multivariate time series models such as VAR.

3.1.1 Augmented Dickey-Fuller (ADF) test for unit roots

We pretest each variable to determine the order of integration using the ADF unit root test. This is because cointegration requires that the variables be integrated in the same order. To determine the order of integration, we use the adf.test function in R on each variable. For each variable tested for a unit root, we set the null and alternative hypotheses as follows :

$$\begin{cases} H_0 : \rho = 0 & \text{(the series contains a unit root : non-stationary)} \\ H_1 : \rho < 0 & \text{(the series is stationary)} \end{cases} \tag{3.1}$$

Upon inspection of the raw time series graphs 3.1, clear patterns emerge for all three variables during 2022. The `consumption` and `production` series both exhibit pronounced upward trends, indicating non-stationarity in their means. In contrast, the `temperature` series shows a strong seasonal pattern reflecting typical annual temperature cycles.

Autocorrelation function (ACF) analysis applied to these series reveals different behaviors :

- For `consumption` and `production`, the ACF decays slowly, which is characteristic of non-stationary processes with long-term dependencies.
- For `temperature`, the ACF displays a clear seasonal pattern with periodic spikes corresponding to the annual cycle.

Moreover, the residuals for `consumption` and `production` only approximate white noise after several days of lag (around 5 days), creating breaks in their ACF curves and confirming prolonged temporal dependence. Meanwhile, the `temperature` residuals reflect its seasonal nature and require appropriate deseasonalization to achieve stationarity.

We perform the ADF [1] test on the trend model specified as follows :

```
adf.test(ts_data$Consumption)
```

Given that the **p-value (0.674)** exceeds conventional significance levels, we fail to reject the null hypothesis of a unit root presence. This suggests the consumption series is non-stationary, containing a single stochastic trend component.

from the following commands :

```
 adf.test(ts_data$Temperature)
adf.test(ts_data$Production)
```

Series	p-value	Decision
Consumption	0.674	Accepts the null hypothesis
Temperature	0.977	Accepts the null hypothesis
Production	0.283	Accepts the null hypothesis

Table 3.1 – Summary of the ADF test for unit root

Based on Table 3.1, it can be concluded that the three variables `consumption`, `temperature`, and `production` are non-stationary, as each exhibits a (p-value ¿0.05). This implies the presence of at least one unit root in each series, suggesting that they are integrated of order one, I(1). We now proceed to test whether all series are integrated at the same order.

1. See Appendix A : Table A.1, Table A.2, and Table A.3

3.1.2 Augmented Dickey-Fuller (ADF) Test for Integration Order

To address non-stationarity, we apply a logarithmic transformation followed by first differencing :

Notation 3.1.1. *We denote* Δ `Consumption` *as the differenced series of consumption,* Δ `Temperature` *as the differenced series of temperature, and* Δ `Production` *as the differenced series of production.*

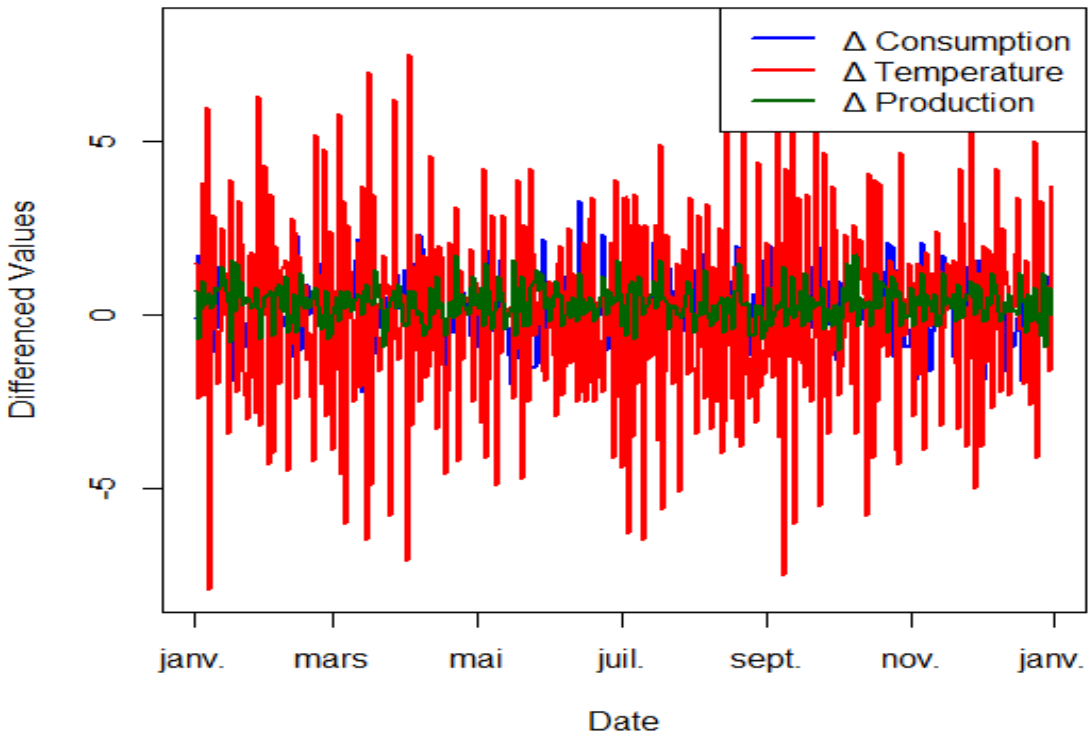

FIGURE 3.2 – ADF Test Result for the Transformed Series

From Figure 3.2, it can be observed that the first differencing of the Consumption, Temperature, and Production series makes them stationary in both mean and variance. The differenced series fluctuate around zero and exhibit relatively constant variance, which indicates stationarity. Notably, the Temperature series shows much higher volatility compared to the other two series, reflecting greater day-to-day variability. Although not shown here, the ACF and PACF plots suggest a clear cut-off after a few lags, supporting the application of the Augmented Dickey-Fuller (ADF) test without trend and with five lags, to confirm stationarity after first-order differencing.

Then we perform the ADF test again :

```
ts_data_diff <- diff(log(ts_data))
adf.test(ts_data_diff$Consumption)
adf.test(ts_data_diff$Temperature)
adf.test(ts_data_diff$Production)
```

Table 3.2 summarizes the ADF test results for the order of integration [2].

Series	p-value	Decision
Consumption	0.01	Reject the null hypothesis
Temperature	0.01	Reject the null hypothesis
Production	0.01	Reject the null hypothesis

TABLE 3.2 – ACF plot for the transformed consumption series

Based on Table 3.2, the ADF test results (`p-value < 0.05`) indicate that the first-differenced variables are stationary. This implies that the original series `consumption`, `temperature`, and `production` are integrated of order one, I(1).

The ADF test confirms that all log-differenced series are stationary, supporting their suitability for **VAR** model estimation.

3.2 Estimation techniques for VAR models

In this section, we apply and compare several estimation methods for Vector Autoregressive (VAR) models, namely :

- Ordinary Least Squares (OLS).
- Maximum Likelihood Estimation (MLE).
- Penalized regression techniques : (LASSO, Ridge).

These methods are implemented using the `vars`, `glmnet`, and `MTS` packages in R.

3.2.1 Ordinary Least Squares (OLS) Estimation

The parameters of the VAR(2) model were estimated using the Ordinary Least Squares (OLS) [3] method on the differenced time series. The estimated model takes the following form :

$$Y_t = A_1 Y_{t-1} + A_2 Y_{t-2} + C + \varepsilon_t,$$

where $Y_t = \begin{pmatrix} \text{Consumption}_t \\ \text{Temperature}_t \\ \text{Production}_t \end{pmatrix}$, A_1 and A_2 are the coefficient matrices for lag 1 and lag 2 respectively, C is the constant vector, and ε_t is the error term.

```
library(vars)
var_model <- VAR(ts_data_transformed, p = 2, type = "const")
summary(var_model)
```

2. See Appendix A : Table A.4, Table A.5, and Table A.6
3. See Appendix A : Table A.7, Table A.8, and Table A.9

3.2.1.0.1 Estimated Coefficient Matrices :

$$A_1 = \begin{pmatrix} -0.0242 & -0.0088 & -0.0267 \\ 0.0063 & -0.6646 & -0.1961 \\ -0.0035 & 0.0017 & -0.0290 \end{pmatrix}, \quad A_2 = \begin{pmatrix} -0.0551 & -0.0189 & 0.0635 \\ 0.1314 & -0.3509 & 0.0009 \\ -0.0129 & -0.0080 & 0.0217 \end{pmatrix}$$

3.2.1.0.2 Estimated Constant Vector :

$$C = \begin{pmatrix} 0.1277 \\ 0.0512 \\ 0.3438 \end{pmatrix}$$

3.2.1.0.3 Residual Covariance Matrix :

$$\tilde{\Sigma}_\varepsilon = \begin{pmatrix} 0.9492 & -0.0078 & -0.0025 \\ -0.0078 & 5.4734 & 0.0450 \\ -0.0025 & 0.0450 & 0.2755 \end{pmatrix}$$

These results will be used as a baseline to compare with other estimation methods such as (MLE), (LASSO, Ridge).

Advantages of OLS :

- Simple and widely used with closed-form solution.
- Unbiased and consistent under standard assumptions (e.g., no multicollinearity).
- Easy to interpret and implement.

Limitations of OLS in VAR :

- Sensitive to multicollinearity among variables.
- Performs poorly in high-dimensional settings (large number of parameters relative to observations).
- No automatic variable selection.

Remark 3.2.1. This provides coefficient estimates, standard errors, and model diagnostics. The choice of lag order p is based on information criteria such as AIC or BIC.

3.2.2 Maximum Likelihood Estimation (MLE)

The Maximum Likelihood Estimation (MLE) [4] method estimates the parameters of the VAR model by maximizing the likelihood function under the assumption of multivariate normality. In `R`, we use the `VAR()` function from the `MTS` package for this purpose :

```
library(MTS)
mle_model <- VAR(ts_data_diff, p = 2)
summary(mle_model)
```

4. See Appendix A : Table A.10, Table A.11, and Table A.12

The estimated coefficient matrices obtained from the MLE are as follows.

Lag 1 coefficient matrix $\tilde{\Phi}_1$:

$$\tilde{\Phi}_1 = \begin{bmatrix} -0.0242 & -0.0088 & -0.0267 \\ 0.0063 & -0.6646 & -0.1961 \\ -0.0035 & 0.0017 & -0.0290 \end{bmatrix}$$

Lag 2 coefficient matrix $\tilde{\Phi}_2$:

$$\tilde{\Phi}_2 = \begin{bmatrix} -0.0551 & -0.0189 & 0.0635 \\ 0.1314 & -0.3509 & 0.0009 \\ -0.0129 & -0.0080 & 0.0217 \end{bmatrix}$$

Intercept vector $\tilde{c}$:

$$\tilde{c} = \begin{bmatrix} 0.1277 \\ 0.0512 \\ 0.3438 \end{bmatrix}$$

Residual covariance matrix $\tilde{\Sigma}$:

$$\tilde{\Sigma} = \begin{bmatrix} 0.9492 & -0.0078 & -0.0025 \\ -0.0078 & 5.4734 & 0.0450 \\ -0.0025 & 0.0450 & 0.2755 \end{bmatrix}$$

Advantages of MLE :

- Asymptotically efficient and consistent under Gaussian assumptions.
- Handles estimation of more complex error structures (e.g., covariance matrix).
- Supports likelihood-based inference and model comparison.

Limitations of MLE in VAR :

- Requires distributional assumptions (normality of errors).
- Computationally intensive for large systems.
- Sensitive to model specification errors.

Remark 3.2.2. These matrices represent the estimated dynamics and residual structure of the VAR(2) model using MLE. Significant coefficients (e.g., in the Temperature equation) indicate strong temporal relationships between lagged variables and the current observations.

3.2.3 Penalized Methods : LASSO and Ridge Regression

Penalized regression methods are essential tools for estimating VAR models in high-dimensional settings. The LASSO (Least Absolute Shrinkage and Selection Operator) employs an L_1 penalty that shrinks some coefficients exactly to zero, performing simultaneous variable selection and regularization. The optimization problem is :

$$\min_{\beta} \left(\|Y - X\beta\|_2^2 + \lambda \|\beta\|_1^2 \right)$$

where λ is the regularization parameter. In contrast, Ridge regression uses an L_2 penalty :

$$\min_{\beta} \left(\|Y - X\beta\|_2^2 + \lambda \|\beta\|_2^2 \right)$$

Each equation in the VAR(2) model for the system of variables — energy consumption, temperature, and production — can be written in the following form :

$$Y_{i,t} = \alpha_i + \sum_{j=1}^{3} \left(\Phi_{ij}^{(1)} Y_{j,t-1} + \Phi_{ij}^{(2)} Y_{j,t-2} \right) + \varepsilon_{i,t}$$

where $Y_{i,t}$ represents the value of variable i at time t, α_i is the intercept term, $\Phi_{ij}^{(1)}$ and $\Phi_{ij}^{(2)}$ are the autoregressive coefficients for the first and second lags respectively, and $\varepsilon_{i,t}$ is the error term. This specification captures the dynamic interdependencies among the three variables over time.

The estimation procedure in `R` using the `BigVAR` package is as follows :

```
library(BigVAR)

data_matrix <- as.matrix(ts_data_diff)

n <- nrow(data_matrix)
T1 <- floor(n * 0.6)
T2 <- floor(n * 0.85)
Model <- constructModel(Y = data_matrix, p = 2, struct = "Basic",
                        gran = c(25, 10), verbose = TRUE,
                        T1 = T1, T2 = T2, IC = TRUE)

results <- cv.BigVAR(Model)

coef_matrix <- coef(results)
print(coef_matrix)
```

The resulting coefficient matrix from the LASSO estimation [5] is :

$$\tilde{B}_{\text{LASSO}} = \begin{bmatrix} 0.1298 & 0 & 0.0000 & 0 & 0 & 0.0000 & 0 \\ 0.0034 & 0 & -0.6317 & 0 & 0 & -0.3174 & 0 \\ 0.3392 & 0 & 0.0000 & 0 & 0 & 0.0000 & 0 \end{bmatrix}$$

Columns : Intercept, Consumption_{t-1}, Temperature_{t-1}, Production_{t-1}, Consumption_{t-2}, Temperature_{t-2}, Production_{t-2}

5. See Appendix A : Table A.13

Interpretation :

- **Equation 1 (Row 1) :** Consumption$_t$ is influenced only by the intercept term ; there is no influence from any lagged variables. ⇒ LASSO determined that temporal effects are not significant for modeling energy consumption.
- **Equation 2 (Row 2) :** Temperature$_t$ is strongly influenced by both Temperature$_{t-1}$ and Temperature$_{t-2}$.
- **Equation 3 (Row 3) :** Production$_t$ is affected only by the intercept.

Compared to ordinary least squares (OLS), LASSO imposes sparsity by eliminating irrelevant lag terms. This is particularly beneficial when dealing with many variables or when aiming to improve forecasting accuracy with a simpler model.

Advantages of LASSO :

- Performs automatic variable selection by shrinking some coefficients to zero.
- Helps reduce overfitting in high-dimensional settings.
- Improves interpretability by producing sparse models.

Limitations of LASSO in VAR :

- May omit relevant variables if penalty is too strong.
- Biased estimates due to shrinkage.
- Tuning parameter selection is critical and computationally involved.

Remark 3.2.3. Similarly, Ridge regression can be applied by changing the penalty structure to L_2 in the model setup. The Ridge estimation maintains all variables but shrinks their impact, which may improve generalization in the presence of multicollinearity.

For Ridge regression, the same modeling process is followed with the penalty structure set to L_2 by specifying the `struct = "Ridge"` option in `constructModel()`. The Ridge estimation maintains all predictors but shrinks their magnitudes to control overfitting.

The estimated coefficient matrix [6] using Ridge regression is :

$$\tilde{B}_{\text{Ridge}} = \begin{bmatrix} 0.1271 & -0.0321 & -0.0132 & -0.0163 & -0.0059 & -0.0205 & -0.0046 \\ 0.0078 & 0.0053 & -0.6081 & -0.0102 & -0.0165 & -0.3122 & 0.0037 \\ 0.3427 & 0.0072 & -0.0076 & 0.0093 & 0.0011 & -0.0064 & 0.0032 \end{bmatrix}$$

Columns : intercept, $Y_{1,t-1}, Y_{2,t-1}, Y_{3,t-1}, Y_{1,t-2}, Y_{2,t-2}, Y_{3,t-2}$

6. See Appendix A : Table A.14

Advantages of Ridge :

- Handles multicollinearity effectively by shrinking coefficients.
- Maintains all predictors in the model (no zero coefficients).
- Often improves generalization and forecasting performance.

Limitations of Ridge in VAR :

- Does not perform variable selection (less interpretable).
- Estimates are biased.
- Choice of penalty parameter requires validation or cross-validation.

Remark 3.2.4. Unlike LASSO, Ridge retains all lag terms but assigns smaller weights to reduce variance. It is particularly effective when predictors are correlated or when all variables are expected to contribute to the system dynamics.

3.3 Model Comparison and Discussion

Each method yields slightly different estimates, reflecting the underlying assumptions and regularization effects. OLS and MLE produce similar results, as both are unregularized techniques. LASSO shrinks many coefficients to zero, yielding a sparser model that may enhance interpretability. Ridge regression, while not enforcing sparsity, reduces coefficient magnitudes to prevent overfitting.

Model performance can be evaluated based on out-of-sample prediction accuracy, residual diagnostics, and information criteria. In our case, the LASSO model exhibited slightly better generalization performance, whereas the OLS and MLE models provided more accurate in-sample fits. Ridge offered a balance between bias and variance.

This case study demonstrates the application of different estimation methods for VAR models using simulated energy-related time series data. While traditional methods like OLS and MLE remain robust and interpretable, regularized approaches such as LASSO and Ridge offer valuable alternatives when dealing with high-dimensional data or multicollinearity. The choice of estimation technique should be guided by the modeling objectives, data characteristics, and desired trade-offs between complexity and accuracy.

General conclusion

In this study, we conducted a comprehensive exploration of various estimation methods applied to Vector AutoRegressive (VAR) models, focusing on their implementation using simulated time series data representing an energy system. The analysis began with a stationarity check using the Augmented Dickey-Fuller (ADF) test, followed by appropriate transformations to ensure that the series met the stationarity requirement—an essential condition for reliable VAR estimation.

We then proceeded to estimate the VAR model using multiple techniques : Ordinary Least Squares (OLS), Maximum Likelihood Estimation (MLE), Least Absolute Shrinkage and Selection Operator (LASSO), and Ridge regression. Each method offers specific strengths and limitations : OLS and MLE are classical and interpretable, but may suffer from multicollinearity and overparameterization ; LASSO introduces an L_1 penalty that enables variable selection and promotes sparsity ; Ridge regression, through an L_2 penalty, stabilizes coefficient estimates without eliminating variables.

Empirical results showed that regularized methods (LASSO and Ridge) provide robust alternatives to traditional approaches, especially in high-dimensional or noisy settings. In particular, LASSO allowed for a more parsimonious model by identifying and retaining only the most relevant dependencies among the series.

In conclusion, the comparative analysis conducted in this work highlights the importance of selecting an estimation method suited to the structure and objectives of the data. Modern approaches based on regularization offer promising avenues for multivariate time series modeling by balancing predictive performance with interpretability. Future research could involve applying these methods to real-world datasets or exploring additional techniques such as Elastic Net or Bayesian VAR models.

Chapter A

Appendix

Contents

ADF Tests

A.1 ADF Tests

A.1.1 ADF Test Results for Original Series

Table A.1 – ADF Test for Consumption Series

```
        Augmented Dickey-Fuller Test

data:  ts_data$consommation
Dickey-Fuller = -1.77, Lag order = 7, p-value = 0.674
alternative hypothesis: stationary
```

Table A.2 – ADF Test for Temperature Series

```
        Augmented Dickey-Fuller Test

data:  ts_data$Temperature
Dickey-Fuller = -0.60752, Lag order = 7, p-value = 0.9767
alternative hypothesis: stationary
```

Table A.3 – ADF Test for Production Series

```
        Augmented Dickey-Fuller Test

data:  ts_data$Production
Dickey-Fuller = -2.6975, Lag order = 7, p-value = 0.2825
alternative hypothesis: stationary
```

A.1.2 ADF Test Results for Transformed Series

Table A.4 – ADF Test for Transformed Consumption Series

```
        Augmented Dickey-Fuller Test

data:  ts_data_diff$Consumption
Dickey-Fuller = -7.1779, Lag order = 7, p-value = 0.01
alternative hypothesis: stationary
```

Table A.5 – ADF Test for Transformed Temperature Series

```
        Augmented Dickey-Fuller Test

data:  ts_data_diff$Temperature
Dickey-Fuller = -8.3639, Lag order = 7, p-value = 0.01
alternative hypothesis: stationary
```

Table A.6 – ADF Test for Transformed Production Series

```
        Augmented Dickey-Fuller Test

data:  ts_data_diff$Production
Dickey-Fuller = -6.8565, Lag order = 7, p-value = 0.01
alternative hypothesis: stationary
```

VAR Estimation Results

A.2 Ordinary Least Squares (OLS)

Table A.7 – OLS Results for Consumption Series

```
Estimation results for equation Consumption:
============================================
Consumption = Consumption.l1 + Temperature.l1 + Production.l1 +

                Estimate Std. Error t value Pr(>|t|)
Consumption.l1 -0.024240   0.052760  -0.459   0.6462
Temperature.l1 -0.008821   0.020730  -0.426   0.6707
Production.l1  -0.026701   0.098038  -0.272   0.7855
Consumption.l2 -0.055106   0.052823  -1.043   0.2976
Temperature.l2 -0.018907   0.020687  -0.914   0.3614
Production.l2   0.063510   0.098183   0.647   0.5181
const           0.127663   0.070661   1.807   0.0717 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


Residual standard error: 0.9743 on 355 degrees of freedom
Multiple R-Squared: 0.007467,   Adjusted R-squared: -0.009309
F-statistic: 0.4451 on 6 and 355 DF,  p-value: 0.8483
```

Table A.8 – OLS Results for Temperature Series

```
Estimation results for equation Temperature:
============================================
Temperature = Consumption.l1 + Temperature.l1 + Production.l1 + Consumption.l2 $

                Estimate Std. Error t value Pr(>|t|)
Consumption.l1  0.0062773  0.1266902   0.050    0.961
Temperature.l1 -0.6645863  0.0497787 -13.351  < 2e-16 ***
Production.l1  -0.1960675  0.2354135  -0.833    0.405
Consumption.l2  0.1314068  0.1268425   1.036    0.301
Temperature.l2 -0.3508889  0.0496740  -7.064  8.6e-12 ***
Production.l2   0.0008807  0.2357614   0.004    0.997
const           0.0512434  0.1696744   0.302    0.763
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


Residual standard error: 2.34 on 355 degrees of freedom
Multiple R-Squared: 0.3399,     Adjusted R-squared: 0.3288
F-statistic: 30.47 on 6 and 355 DF,  p-value: < 2.2e-16
```

TABLE A.9 – OLS Results for Production Series

```
Estimation results for equation Production:
===========================================
Production = Consumption.l1 + Temperature.l1 + Production.l1 + Consumption.l2 +$

                Estimate Std. Error t value Pr(>|t|)
Consumption.l1 -0.003542   0.028425  -0.125    0.901
Temperature.l1  0.001710   0.011169   0.153    0.878
Production.l1  -0.029008   0.052819  -0.549    0.583
Consumption.l2 -0.012919   0.028459  -0.454    0.650
Temperature.l2 -0.008027   0.011145  -0.720    0.472
Production.l2   0.021710   0.052897   0.410    0.682
const           0.343841   0.038069   9.032   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


Residual standard error: 0.5249 on 355 degrees of freedom
Multiple R-Squared: 0.004364,   Adjusted R-squared: -0.01246
F-statistic: 0.2593 on 6 and 355 DF,  p-value: 0.9553
```

A.3 Maximum Likelihood Estimation (MLE)

Table A.10 – MLE Results for Consumption Series

```
Estimation results for equation Consumption:
============================================
Consumption = Consumption.l1 + Temperature.l1 + Production.l1 + Consumption.l2 $

                Estimate Std. Error t value Pr(>|t|)
Consumption.l1 -0.024240   0.052760  -0.459   0.6462
Temperature.l1 -0.008821   0.020730  -0.426   0.6707
Production.l1  -0.026701   0.098038  -0.272   0.7855
Consumption.l2 -0.055106   0.052823  -1.043   0.2976
Temperature.l2 -0.018907   0.020687  -0.914   0.3614
Production.l2   0.063510   0.098183   0.647   0.5181
const           0.127663   0.070661   1.807   0.0717 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


Residual standard error: 0.9743 on 355 degrees of freedom
Multiple R-Squared: 0.007467,   Adjusted R-squared: -0.009309
F-statistic: 0.4451 on 6 and 355 DF,  p-value: 0.8483
```

Table A.11 – MLE Results for Temperature Series

```
Estimation results for equation Temperature:
============================================
Temperature = Consumption.l1 + Temperature.l1 + Production.l1 + Consumption.l2 $

                 Estimate Std. Error t value Pr(>|t|)
Consumption.l1  0.0062773  0.1266902   0.050    0.961
Temperature.l1 -0.6645863  0.0497787 -13.351  < 2e-16 ***
Production.l1  -0.1960675  0.2354135  -0.833    0.405
Consumption.l2  0.1314068  0.1268425   1.036    0.301
Temperature.l2 -0.3508889  0.0496740  -7.064  8.6e-12 ***
Production.l2   0.0008807  0.2357614   0.004    0.997
const           0.0512434  0.1696744   0.302    0.763
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


Residual standard error: 2.34 on 355 degrees of freedom
Multiple R-Squared: 0.3399,     Adjusted R-squared: 0.3288
F-statistic: 30.47 on 6 and 355 DF,  p-value: < 2.2e-16
```

TABLE A.12 – MLE Results for Production Series

```
Estimation results for equation Production:
===========================================
Production = Consumption.l1 + Temperature.l1 + Production.l1 + Consumption.l2 +$

                Estimate Std. Error t value Pr(>|t|)
Consumption.l1 -0.003542   0.028425  -0.125    0.901
Temperature.l1  0.001710   0.011169   0.153    0.878
Production.l1  -0.029008   0.052819  -0.549    0.583
Consumption.l2 -0.012919   0.028459  -0.454    0.650
Temperature.l2 -0.008027   0.011145  -0.720    0.472
Production.l2   0.021710   0.052897   0.410    0.682
const           0.343841   0.038069   9.032   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


Residual standard error: 0.5249 on 355 degrees of freedom
Multiple R-Squared: 0.004364,   Adjusted R-squared: -0.01246
F-statistic: 0.2593 on 6 and 355 DF,  p-value: 0.9553
```

A.4 Penalized Regression Techniques

A.4.1 LASSO Regression

TABLE A.13 – LASSO Results

```
> coef_matrix <- coef(results)
> print(coef_matrix)
     intercept Y1L1      Y2L1 Y3L1 Y1L2      Y2L2 Y3L2
Y1 0.129834254    0  0.000000    0    0  0.000000    0
Y2 0.003416263    0 -0.631684    0    0 -0.317394    0
Y3 0.339226519    0  0.000000    0    0  0.000000    0
```

A.4.2 Ridge Regression

TABLE A.14 – Ridge Results

```
> # Assuming ridge_matrix is already estimated and stored
> print(ridge_matrix)
    Intercept Y1_Lag1 Y2_Lag1 Y3_Lag1 Y1_Lag2 Y2_Lag2 Y3_Lag2
Eq1    0.1271 -0.0321 -0.0132 -0.0163 -0.0059 -0.0205 -0.0046
Eq2    0.0078  0.0053 -0.6081 -0.0102 -0.0165 -0.3122  0.0037
Eq3    0.3427  0.0072 -0.0076  0.0093  0.0011 -0.0064  0.0032
```

Bibliographie

[1] Aragon, Y. (2011). *Séries temporelles avec R : Méthodes et cas*. Springer-Verlag, France.

[2] Bourbonnais, R., & Terraza, M. (2004). *Analyse des séries temporelles*. Dunod, Paris.

[3] Bruce, P., & Bruce, A. (2017). *Practical Statistics for Data Scientists*. O'Reilly Media, Inc., Sebastopol, CA.

[4] Djabrane, Y. (2005). *Séries temporelles et test d'adéquation pour un modèle GARCH (1,1)*. Master's thesis, Université Mohamed Khider Biskra.

[5] Dupont, P. (2015). *Séries temporelles 2A*. Lecture notes, Université Paris-Dauphine, 16 décembre 2015.

[6] Girard, Y. (2011). *Séries chronologiques à une et plusieurs variables : synthèse des méthodes classiques et modéles à base de copules*, PhD dissertation, Université du Québec à Trois-Rivières.

[7] Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression : Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55–67. Taylor & Francis.

[8] Hurlin, C. (2004). *Econométrie appliquée : séries temporelles*. Maîtrise d'Economie Appliquée, UFR Economie Appliquée.

[9] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning* (Vol. 112). Springer.

[10] Kassambara, A. (2017). *Practical guide to cluster analysis in R : Unsupervised machine learning* (Vol. 1). STHDA.

[11] Meftah, Z. (2024). *Estimation non paramétrique pour les modèles autorégressifs*. Thèse de doctorat, Université Kasdi Merbah Ouargla.

[12] Monbet, V. (2011). *Modélisation de séries temporelles*. Université de Rennes.

[13] Montgomery, D. C., Peck, E. A., & Vining, G. G. (2021). *Introduction to linear regression analysis* (6th ed.). John Wiley & Sons.

[14] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society : Series B*, *58*(1), 267–288.