UCLouvain

FACULTE DES SCIENCES

Faculté des sciences

KernelICA for Identification of Structural Innovations in Multivariate Time Series Models

Author: **Victor DUJARDIN**
Supervisor: **Christian HAFNER**
Reader: **Rainer VON SACHS**
Academic year 2024
Master [120] in Data Science : Statistic

Abstract

This master's thesis explores how Kernel Independent Component Analysis (KernelICA) can solve the identification problem in Structural Vector Autoregressive models. Identifying structural shocks is essential for accurately interpreting and forecasting multivariate time series data. We start by discussing the basics of SVAR models and the difficulties in identifying these shocks. Then, we introduce KernelICA as a new method of identification and compare it to traditional Independent Component Analysis techniques. Using Monte Carlo simulations, we test how well KernelICA identifies independent shocks and compare its performance with existing methods. An empirical analysis using data from Blanchard and Perotti shows KernelICA's practical use. Our results indicate that KernelICA is a strong alternative for identifying structural innovations in SVAR models.

Acknowledgments

I would like to express my deepest gratitude to my supervisor, Professor Christian Hafner, for his guidance throughout the research and writing of this thesis. My sincere thanks also go to Professor Rainer Von Sachs for accepting being a reader of this master's thesis.

Contents

List of Abbreviations

SVAR	Structural Vector Autoregression
VAR	Vector Autoregression
VMA	Vector Moving Average
IRF	Impulse Response Function
FEVD	Forecast Error Variance Decomposition
MSPE	Mean Squared Prediction Error
ICA	Independent Component Analysis
PCA	Principal Component Analysis
KL	Kullback-Leibler
KernelICA	Kernel Independent Component Analysis
CCA	Canonical Correlation Analysis
RKHS	Reproducing Kernel Hilbert Space
MI	Mutual Information
KGV	Kernel Generalized Variance
DCov	Distance Covariance
MDI	Minimum Distance Index

List of Figures

List of Tables

Chapter 1

Introduction

Every day, large amounts of data are created, creating a growing need for analysis to better understand the world around us. In econometrics, which uses statistical methods to analyze economic data, data-driven approaches are becoming increasingly important to fully utilize the available data.

Models that handle multiple variables are essential because the interactions between different factors are often complex. Among these models, Structural Vector Autoregressive (SVAR) models are useful tools for analyzing how different time series variables affect each other over time.

A key challenge in using SVAR models is identifying the structural innovations—unexpected changes in the data that drive the system. While there are many methods to identify these shocks, Independent Component Analysis (ICA) has become a popular approach, assuming these structural innovations are independent.

This thesis will explore different methods for identifying independent structural innovations, with a focus on Kernel Independent Component Analysis (KernelICA). We will evaluate how well KernelICA performs in identifying these innovations and compare it with other ICA methods using Monte Carlo simulations.

Finally, the empirical part of this study will demonstrate how KernelICA can be applied to real-world economic data, showing its potential as a useful tool in econometric analysis.

The code used for this master's thesis is available on GitHub. (https://github.com/victordujardin/KernelICA-for-SVAR-identification)

Chapter 2

Structural VAR

Econometrics often tries to understand and measure relationships between different economic variables. The Vector Autoregression (VAR) model, introduced by Sims (1980), does precisely so by allowing for the joint modeling of multiple time series, such as GDP, inflation rates, and unemployment rates. In this model, each variable is represented as a linear combination of its own lagged values and the lagged values of the other variables. Even though VAR models are powerful tools in econometrics, this model does identify interpretable movements in the data, also named structural shocks or innovations.

SVAR models improve on traditional methods by adding restrictions to the VAR framework. These restrictions help identify structural shocks. By isolating these shocks, SVAR models clarify the cause-and-effect relationships between economic variables.

Consider a vector Y_t comprising economic indicators observed at regular intervals. The SVAR model posits that the current value of Y_t can be explained by its past values, embodying the autoregressive nature. Specifically, the value of Y_t is influenced by its own previous values (e.g., $Y_{t-1}, Y_{t-2}, \ldots$) as well as by the lagged values of other variables within the vector. Moreover, SVAR captures the contemporaneous interactions among variables, allowing for the modeling of immediate effects one variable may have on another within the same time period.

Through this framework, the SVAR model offers a nuanced understanding of how economic indicators interact, accounting for both historical data and the immediate relationships among variables.

Formally, the SVAR model is expressed by Kilian and Lütkepohl (2017, p. 109) as follows:

$$B_0 y_t = B_1 y_{t-1} + B_2 y_{t-2} + \cdots + B_p y_{t-p} + w_t \tag{2.1}$$

where the $K \times 1$ vector y_t is presumed to have zero mean. The dimension of B_i for $i = 0, \ldots, p$ is $K \times K$. In this thesis, the components $w_{1t}, \ldots, w_{Kt}$ of the $K \times 1$ vector w_t are assumed to be mutually independent and to have a non-Gaussian distribution except possibly for one component, with reasons for this choice discussed in the next chapter. Additionally, we assume B_0 to be invertible.

As stated by Moneta and Pallante (2022), this model is structural because it tracks the effect of statistically independent shocks on endogenous variables.

This model can be expressed in reduced-form as per Kilian and Lütkepohl (2017, p. 109):

$$y_t = B_0^{-1} B_1 y_{t-1} + \cdots + B_0^{-1} B_p y_{t-p} + B_0^{-1} w_t \tag{2.2}$$

$$= A_1 y_{t-1} + \cdots + A_p y_{t-p} + u_t \tag{2.3}$$

This reduced-form is called a VAR process. The parameter matrices of a VAR process can be estimated using Maximum Likelihood and these estimates are asymptotically normal. It is important to note that the VAR process does not account for contemporaneous relationships between the variables, which is a limitation.

Please note that the reduced-form residuals u_t are a linear mixture of the structural shocks w_t, namely:

$$w_t = B_0 u_t \tag{2.4}$$

This is the B-model described by Lütkepohl (2005).

A VAR(p) process is termed stable if

$$\det(I_K - A_1 z - \cdots - A_p z^p) \neq 0 \text{ for } |z| \leq 1.$$

This condition ensures that the process does not "explode," meaning it prevents the values generated by the model from growing uncontrollably over time. A stable VAR(p) process can be expressed as the weighted sum of past and present innovations (Kilian & Lütkepohl, 2017, p. 25).

For a VAR(1) process, we have

$$y_t = A_1 y_{t-1} + u_t.$$

Successive substitution implies

$$y_t = \sum_{i=0}^{\infty} A_1^i u_{t-i}.$$

The sum on the right-hand side of this infinite-order representation exists if the eigenvalues of A_1 are all less than 1 in modulus. Similarly, a representation in terms of past and present innovations of a VAR(p) model can be obtained via the corresponding VAR(1) representation, resulting in

$$\begin{aligned} y_t &= A(L)^{-1} u_t \\ &= \sum_{i=0}^{\infty} J\mathbf{A}^i J' J U_{t-i} \\ &= \sum_{i=0}^{\infty} \Phi_i u_{t-i} \end{aligned} \tag{2.5}$$

where $J \equiv [I_K, 0_{K \times K(p-1)}]$ is a $K \times Kp$ matrix, and the $K \times K$ coefficient matrices of the inverse VAR operator $A(L)^{-1} = \sum_{i=0}^{\infty} \Phi_i L^i$ are given by $\Phi_i = J\mathbf{A}^i J'$, for $i = 0, 1, \ldots$.

The existence of the inverse VAR operator is ensured by the stability of the process. This representation is known as the vector moving average (VMA) representation.

The idea of SVAR, in contrast to its reduced-form, is to study the impact of a shock in one of the variables on the system. However, in a basic VAR(p), the different residuals may be correlated with each other, meaning a shock in a particular variable can affect another one. That is, the variance-covariance matrix Σ_u is not diagonal.

By performing a linear transformation, we can impose orthogonality between the variables so that the variance-covariance matrix of the structural shocks Σ_w is diagonal.

With this restriction, it becomes easier to study the impact of one particular structural shock on the system.

2.1 The Identification Problem

To achieve full identification of the SVAR model described in Equation 2.1, it is necessary to estimate the B_0 matrix, which contains $K \times K$ parameters.

Sims (1980) argues that a model's identification occurs when different sets of parameters within its framework lead to uniquely observable behaviors or outcomes. Essentially, this means that by changing the values of the parameters, we can see different patterns or reactions in the model's variables, highlighting a direct relationship between parameter values and the model's observable behavior.

The identification problem arises because the matrix B_0, which contains $K \times K$ elements, where K is the number of variables in the system, is not directly observable. As a result, there exists an infinite set of different values of B_0 which all imply the same probability distribution for the observed data (Gottschalk, 2001). To estimate B_0, we need additional restrictions or assumptions based on economic theory and statistical properties.

To achieve identification of the model, a lot of strategies can be employed, some are explained below by Kilian and Lütkepohl (2017):

Identification by Short-Run Restrictions

Consider the model 2.1 where w_t represents the structural innovations with mean zero and serially uncorrelated errors. The variance-covariance matrix of w_t is normalized as:

$$\mathbb{E}(w_t w_t') = \Sigma_w = I_K,$$

implying that structural shocks are uncorrelated and have unit variance. This normalization is only about the scaling of the system and helps simplify the estimation process without altering anything of substance, as explained by Gottschalk (2001).

To estimate the structural model, its reduced-form representation can be used:

$$A(L)y_t = u_t,$$

where $A(L) = I - A_1 L - \cdots - A_p L^p$.

From standard estimation methods, we can consistently estimate the parameters A_i, the reduced-form errors u_t, and their covariance matrix:

$$\mathbb{E}(u_t u_t') = \Sigma_u.$$

Equation 2.4 leads to:

$$\mathbb{E}(u_t u_t') = B_0^{-1} \Sigma_w (B_0^{-1})'.$$

Given $\Sigma_u = I_K$, identifying B_0^{-1} requires solving the equation :

$$\Sigma_u = B_0^{-1} (B_0^{-1})'. \tag{2.6}$$

This involves imposing additional restrictions on B_0^{-1}, such as zero restrictions on specific elements. Based on economic theory, certain variables are assumed not to have a direct immediate effect on others. These assumptions introduce zeros in the B_0 matrix, providing necessary restrictions for identification.

This symmetry of the covariance matrix specifies $K(K+1)/2$ different equations, and we still need $K(K-1)/2$ further relations to identify all K^2 elements of B_0 (Lütkepohl, 2005).

A common method is to use recursive identification via Cholesky decomposition of Σ_u. Define a lower-triangular matrix P such that:

$$PP' = \Sigma_u.$$

This decomposition imposes a recursive ordering on the contemporaneous relationships among variables, essentially assuming a hierarchical structure in which the order of variables matters. This can sometimes be a limitation, as the imposed causality may not always be justifiable or evident in real-world data, as critiqued by Moneta and Pallante (2022).

Identification by Long-Run Restrictions

Another idea is to place limits on how variables respond to shocks in the long term. This can help identify some shocks, especially when some variables have unit roots, and others

do not. This approach can simplify things by focusing on long-term aspects of models, which economists tend to agree on, rather than the more debated short-term restrictions. (Kilian, 2011)

We consider the structural VAR model, denoted by $B(L)y_t = w_t$, and its structural VMA representation, $y_t = \Theta(L)w_t$. Similarly, for the reduced-form VAR model, represented as $A(L)y_t = u_t$, we have the corresponding VMA representation, $y_t = \Phi(L)u_t$.

The relationship between the structural and reduced-form models can be established using $A(L) = B_0^{-1}B(L)$. For $L = 1$, we derive $B_0^{-1} = A(1)B(1)^{-1}$.

By manipulating these equations and with the help of equations 2.4 and 2.6, we derive the relation $\Phi(1)\Sigma_u\Phi(1)^T = \Theta(1)\Theta(1)^T$. With sufficient restrictions on $\Theta(1)$, specifically $\frac{K(K-1)}{2}$ restrictions, we can uniquely identify the elements of $\Theta(1)$.

To ensure these restrictions are met, we often assume a recursive structure, allowing the use of a Cholesky decomposition on the matrix $\Phi(1)\Sigma_u\Phi(1)^T$. These restrictions enable us to identify $\Theta(1) = B(1)^{-1}$, representing the long-run effects of structural shocks.

Ultimately, with $\Theta(1)$ determined, we can estimate B_0^{-1} as $A(1)\Theta(1)$, combining short-run and long-run restrictions in the estimation process.

Identification by Sign Restrictions

Kilian (2011) explains that an alternative approach is to identify structural shocks by setting restrictions on the sign of the responses of certain variables to these shocks.

The key to sign-identified models is associating each identified shock with a unique pattern of sign changes. These sign restrictions can be based on economic theory and applied to the coefficients in the model.

To implement this approach, we follow these steps:

1. Start with a reduced-form VAR model and compute its variance-covariance matrix.
2. Use the Cholesky decomposition to obtain an initial estimate of the structural matrix.
3. Randomly draw orthogonal matrices and combine them with the initial estimate to generate candidate models.

4. Among the candidate models, retain only those that satisfy the predefined sign restrictions on the impulse response functions.

Identification by Heteroskedasticity

Rigobon (2003) introduces a method for solving the VAR identification problem by using the heteroskedasticity of structural shocks. Heteroskedasticity, which can occur during events like financial crises, allows the identification of structural parameters when it can be described as a two-regime process. Rigobon demonstrates that under this condition, the structural parameters can be identified. He also extends this method to more complex situations, such as having more than two regimes or unobservable common shocks. However, the method has limitations. It can be difficult to determine the existence, number, and timing of variance regimes.

Identification by Non-Gaussianity

Going beyond the assumption of uncorrelated residuals, we can identify the model by assuming that the residuals are statistically independent. This stronger assumption facilitates the identification of the B_0 matrix by exploiting the non-Gaussian nature of the residuals. This method of identification is further detailed in Section 3.

Alternative Structural VAR Approaches

Kilian and Lütkepohl (2017) explain that other identification strategies exist, such as identification based on extraneous data.

We can also achieve identification by the spectral decomposition of Σ_u. Spectral decomposition offers a different perspective on matrix decomposition, focusing on the eigenvalues and eigenvectors of the variance-covariance matrix. For Σ_u, the spectral decomposition can be expressed as $\Sigma_u = B_0 \Lambda B_0'$, where B_0 is a matrix composed of the eigenvectors of Σ_u, and Λ is a diagonal matrix containing the corresponding eigenvalues. Each column $\mathbf{B}_{0,i}$ of B_0 represents an eigenvector, and each diagonal element Λ_{ii} of Λ corresponds to an eigenvalue λ_i. By using the Spectral decomposition we impose a symmetry between the effects of one variable on the other.

The choice among these strategies depends on the specific characteristics of the data and the underlying economic theory. Each method has its advantages and limitations, so selecting the right approach is essential for accurate SVAR model interpretation. In our case, we focus on the use of independent component analysis, which relies on Non-Gaussianity, as discussed in more detail in Chapter 3.

2.2 Structural Impulse Response Function Analysis

In SVAR models, analyzing the effects of shocks within the system is fundamental. Consider the Equation 2.4. The focus is on how each element of y_t, our vector of variables, responds to an impulse in w_t. This response is described by the structural impulse response function (IRF), initially introduced by Sims (1980), and is mathematically expressed as:

$$\frac{\partial y_{t+i}}{\partial w_t'} = \Theta_i, \text{ for } i = 0, 1, 2, \ldots, H,$$

where Θ_i is a $K \times K$ matrix that captures the response of y_t to the shock in w_t after i periods (Kilian & Lütkepohl, 2017, pp. 110).

To derive the structural impulse responses ($\theta_{jk,i}$), we start by considering the responses of y_{t+i} to the reduced-form errors u_t. These can be obtained from the VAR(1) representation of the VAR(p) process:

$$Y_t = \mathbf{A}Y_{t-1} + U_t,$$

where:

$$Y_t \equiv \begin{pmatrix} y_t \\ \vdots \\ y_{t-p+1} \end{pmatrix}, \quad \mathbf{A} \equiv \begin{pmatrix} A_1 & A_2 & \cdots & A_{p-1} & A_p \\ I_K & 0 & \cdots & 0 & 0 \\ 0 & I_K & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & I_K & 0 \end{pmatrix}, \quad U_t \equiv \begin{pmatrix} u_t \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

Through successive substitutions for Y_{t-i}, we can express it as:

$$Y_{t+i} = \mathbf{A}^{i+1}Y_{t-1} + \sum_{j=0}^{i} \mathbf{A}^j U_{t+i-j}.$$

By left-multiplying by $J \equiv [I_K, 0_{K\times K(p-1)}]$, we derive:

$$\begin{aligned} y_{t+i} &= J\mathbf{A}^{i+1}Y_{t-1} + \sum_{j=0}^{i} J\mathbf{A}^{j}U_{t+i-j} \\ &= J\mathbf{A}^{i+1}Y_{t-1} + \sum_{j=0}^{i} J\mathbf{A}^{j}J'JU_{t+i-j} \\ &= J\mathbf{A}^{i+1}Y_{t-1} + \sum_{j=0}^{i} J\mathbf{A}^{j}J'u_{t+i-j}. \end{aligned}$$

The response of variable $j = 1, \ldots, K$ to a unit shock $u_{kt}, k = 1, \ldots, K$ i periods ago, is given by:

$$\Phi_i = [\phi_{jk,i}] \equiv J\mathbf{A}^{i}J^{'}.$$

These responses are known as dynamic multipliers or reduced-form impulse responses (Kilian & Lütkepohl, 2017, p. 111).

If y_t is covariance stationary, meaning that its covariance structure is independent of time, it can be represented as a weighted average of current and past shocks with diminishing weights Φ_i. This VMA representation is:

$$y_t = \sum_{i=0}^{\infty} \Phi_i u_{t-i} = \sum_{i=0}^{\infty} \Phi_i B_0^{-1} B_0 u_{t-i} = \sum_{i=0}^{\infty} \Theta_i w_{t-i},$$

where $w_{t-1} = B_0 u_{t-i}$ and $\Theta_i \equiv \Phi_i B_0^{-1}$. Therefore,

$$\frac{\partial y_t}{\partial w_{t-1}'} = \frac{\partial y_{t+i}}{\partial w_t'} = \Theta_i.$$

These responses can be calculated by post-multiplying Φ_i, $i = 0, \ldots, H$, by B_0^{-1}:

$$\Theta_i = \Phi_i B_0^{-1}.$$

In cases where the VAR is not stable, this method still applies, though the impulse responses may not converge to zero as $i \to \infty$, and they will not reflect the coefficients of the structural MA representation (Kilian & Lütkepohl, 2017, p. 111).

As noted by Kilian and Lütkepohl (2017), it is customary to scale B_0^{-1} so that the structural shocks represent one standard deviation, viewing such a shock as typical in

magnitude. Structural shocks are generally unit-free and cannot be expressed in the units of the model variables (Kilian & Lütkepohl, 2017, p. 112).

To estimate impulse responses in practice, the unknown parameters in the reduced-form VAR(p) model are replaced with consistent estimates. With estimates of the VAR parameters $\hat{A}_j$, $j = 1, \ldots, p$, and $\hat{\Sigma}_u$, and the implied estimate of B_0^{-1}, $\hat{\Phi}_i$ (and thus $\hat{\Theta}_i$) can be constructed recursively for $i = 0, \ldots, H$ (Kilian & Lütkepohl, 2017, p. 112).

To illustrate the application of structural impulse response functions, consider the VAR model by Kilian and Park (2009) which explores the global crude oil market and the U.S. stock market. This model includes the growth rate in global crude oil production ($\Delta prod_t$), a measure of the global business cycle in industrial commodity markets (rea_t), the real price of crude oil ($rpoil_t$), and U.S. real dividend growth (Δrd_t). Let $y_t = (\Delta prod_t, rea_t, rpoil_t, \Delta rd_t)'$. The objective is to decompose the reduced-form innovations into structural shocks: oil supply shocks ($w_{1t}^{oil\ supply}$), aggregate demand shocks for industrial commodities ($w_{2t}^{aggregate\ demand}$), oil-specific demand shocks ($w_{3t}^{oil-specific\ demand}$), and a residual shock capturing other determinants of U.S. real dividends (w_{4t}^{other}). By imposing a recursive ordering on B_0^{-1} so that the elements above the diagonal are zero, the remaining elements can be uniquely identified from Σ_u. With estimates of A_j and thus Φ_i, $i = 0, \ldots, H$, and knowledge of B_0^{-1}, the corresponding structural impulse response matrices Θ_i can be estimated (Kilian & Lütkepohl, 2017, p. 112).

To quantify the effects of oil demand and supply shocks on U.S. real dividends (measured in percent deviations from the baseline), we compute the cumulative effects of these shocks on the fourth variable in the VAR model by summing the estimates of $\theta_{4,1,i}$, $\theta_{4,2,i}$, and $\theta_{4,3,i}$ for $i = 0, \ldots, H$. The figure 2.1 illustrates that a positive global aggregate demand shock increases U.S. real dividends by about 1% one year later, supporting the notion that a global demand boom positively impacts the U.S. economy. Conversely, positive oil-specific demand shocks and negative oil supply shocks reduce real dividends by approximately 0.5% and 1%, respectively, indicating that supply disruptions and other adverse events in the global oil market negatively affect the U.S. economy (Kilian & Lütkepohl, 2017, p. 113).

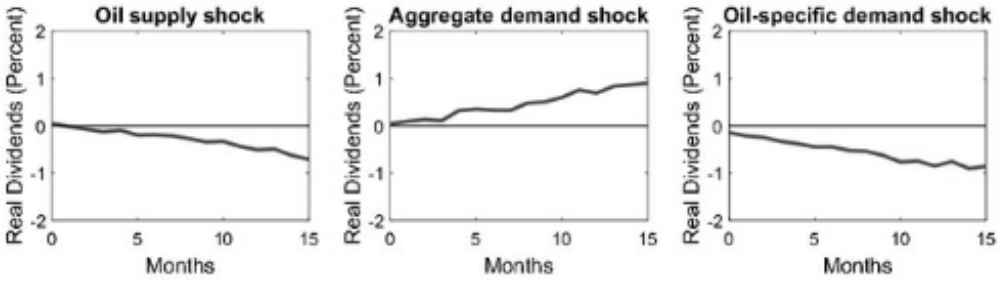

Figure 2.1: Point estimates of responses of U.S. real dividends to selected structural shocks, Source : Kilian and Park (2009)

2.3 Forecast Error Variance Decomposition

Forecast Error Variance Decomposition (FEVD) is a technique used to understand the impact of various structural shocks on the forecast error when predicting future values of a variable within a SVAR model. FEVD breaks down the contribution of each shock to the forecast error variance for a specific variable (Kilian & Lütkepohl, 2017, p. 113).

To compute FEVD, we consider the proportion of the Mean Squared Prediction Error (MSPE) of y_{t+h} attributable to each shock. For a VAR process, the h-step ahead forecast error can be written as:

$$y_{t+h} - y_{t+h|t} = \sum_{i=0}^{h-1} \Phi_i u_{t+h-i} = \sum_{i=0}^{h-1} \Theta_i w_{t+h-i}$$

where $u_t = B_0^{-1} w_t$, allowing the substitution of $\Phi_i u_{t+h-i}$ with $\Theta_i w_{t+h-1}$ (Kilian & Lütkepohl, 2017, p. 114).

The MSPE at horizon h is then given by:

$$
\begin{aligned}
MSPE(h) &\equiv \mathbb{E}[(y_{t+h} - y_{t+h|t})(y_{t+h} - y_{t+h|t})'] \\
&= \sum_{i=0}^{h-1} \Phi_i \Sigma_u \Phi_i' \\
&= \sum_{i=0}^{h-1} \Theta_i \Sigma_w \Theta_i' \\
&= \sum_{i=0}^{h-1} \Theta_i I_K \Theta_i' \\
&= \sum_{i=0}^{h-1} \Theta_i \Theta_i'
\end{aligned}
$$

Denote $\theta_{kj,h}$ as the kj^{th} element of Θ_h. The contribution of shock j to the MSPE of y_{kt}, for $k = 1, \ldots, K$, at horizon h is:

$$MSPE_j^k(h) = \theta_{kj,0}^2 + \ldots + \theta_{kj,h-1}^2$$

The total MSPE of y_{kt}, for $k = 1, \ldots, K$, at horizon h is:

$$MSPE^k(h) = \sum_{j=1}^{K} MSPE_j^k(h) = \sum_{j=1}^{K} (\theta_{kj,0}^2 + \ldots + \theta_{kj,h-1}^2)$$

By dividing:

$$MSPE^k(h) = \sum_{j=1}^{K} MSPE_j^k(h)$$

by $MSPE^k(h)$, we obtain the following decomposition for a given h and k:

$$1 = \frac{MSPE_1^k(h)}{MSPE^k(h)} + \frac{MSPE_2^k(h)}{MSPE^k(h)} + \ldots + \frac{MSPE_K^k(h)}{MSPE^k(h)}$$

Each ratio indicates the fraction of the contribution of the j^{th} shock to the MSPE(h) of variable k for $j = 1, \ldots, K$. Therefore, $FEVD = \frac{MSPE_j^k(h)}{MSPE^k(h)}$ measures the fraction of the contribution of shock j to the forecast error variance of variable k (Kilian & Lütkepohl, 2017, p. 115).

The application of FEVD is best illustrated through practical examples. For instance, in their model explained in Section 2.2 of this thesis, Kilian and Park (2009) explore how

much variability in U.S. real dividend growth is explained by oil demand and oil supply shocks. This can be evaluated using FEVD for U.S. real dividend growth. Kilian and Park (2009) analyze horizons of 1, 2, 3, and 12 months, and also consider an infinite horizon ($h = \infty$) for stationary real dividend growth (Δrd_t). The results are shown in the following Table 2.2.

	Percent of h-Step Ahead Forecast Error Variance Explained by:			
Horizon	Oil supply shock	Aggregate demand shock	Oil-specific demand shock	Residual shock
1	0.2	0.2	1.7	98.0
2	0.6	0.4	2.1	97.0
3	0.8	0.5	2.1	96.6
12	2.8	6.8	4.5	85.8
∞	6.6	8.4	7.9	77.1

Figure 2.2: Forecast error variance decomposition for U.S. real dividend growth, Source : Kilian and Park (2009)

Ignoring rounding errors, the entries in each row of the table sum to 100 % by design. The entries for horizon ∞ represent the variance decomposition of U.S. real dividend growth. Practically, ∞ can be approximated by a sufficiently large number, showing that further increases in the horizon do not significantly change the results (Kilian & Lütkepohl, 2017, p. 115).

Analyzing FEVD across different horizons reveals patterns of interest. In this example, oil supply and oil demand shocks combined account for only 2 % of the MSPE of U.S. real dividend growth at the one-month horizon, but their explanatory power rises to 22.9 % in the long run. This suggests a weak relationship between the global oil market and the U.S. stock market. Additionally, the relative contributions of different shocks at specific horizons are notable. At the one-month horizon, oil-specific demand shocks are more significant than oil supply or aggregate demand shocks in explaining the forecast error variance of real dividend growth. However, in the long run, each type of shock contributes approximately equally to the unconditional variance (Kilian & Lütkepohl, 2017, p. 115).

2.4 Linking Structural VAR and ICA for Identification

As explained in Section 2.1, achieving identification in an SVAR model requires imposing certain restrictions. One way to do this is by using Independent Component Analysis, which assumes that the model's residuals are statistically independent. This assumption is reasonable if we consider the structural innovations as unrelated shocks that independently affect the system. In the next section, we will present three different ICA algorithms, including Kernel Independent Component Analysis, which is the primary focus of this thesis. These algorithms aim to recover independent time series from dependent ones by estimating an unmixing matrix, denoted as W. In the context of the SVAR model, the matrix B_0 in Equation 2.4 corresponds to this unmixing matrix W. Therefore, these different algorithms are valuable for recovering the structural innovations within the SVAR model.

Chapter 3

Independent Component Analysis

To address the issue of identification of structural innovations within the previously discussed structural models, we will employ a method known as independent component analysis or ICA for short. This approach aims to discover new, independent features by decomposing a multivariate signal into its additive subcomponents. It is an effective strategy for revealing the fundamental structure of a model. The identification will be based on the hypothesis of the independence of these residuals.

There are various methods within ICA, but the fundamental principle can be summarized as follows: the relationship between observed signals x and sources s is represented by the equation $x = As$, where s is a latent random vector with m independent components, A is an $m \times m$ matrix of parameters, assumed invertible, and x is an observed vector of m components. The sources s can be retrieved by multiplying the observed signals x with the inverse of the mixing matrix, $W = A^{-1}$, also referred to as the unmixing matrix.

We want to find an estimate of A and recover the values of s by solving the linear system of equations. The distribution of s is assumed to be unknown, and ICA is formulated as a semiparametric model (Bickel et al., 1998). Bach and Jordan (2003) explain that the objective of ICA is to estimate A using maximum likelihood estimation. First, consider the population version of Independent Component Analysis, where $p^*(x)$ represents the true distribution of x, and $p(x)$ represents the model distribution. We aim to minimize the Kullback-Leibler (KL) divergence between p^* and p, denoted as $D(p^*(x)\|p(x))$. The KL divergence is defined as

$$D(p^*\|p) = \int p^*(x) \log\left(\frac{p^*(x)}{p(x)}\right) dx \tag{3.1}$$

and measures the difference between two probability distributions. Given that the KL divergence remains invariant under invertible transformations, we can apply W to x in both arguments of the KL divergence. Consequently, the problem reduces to minimizing $D(p^*(s)\|p(s))$.

Let $\tilde{p}(s)$ denote the joint probability distribution obtained by taking the product of the marginals of $p^*(s)$. According to Cover and Thomas (1991), for any distribution $p(s)$ with independent components, we can decompose the KL divergence as follows:

$$D(p^*(s)\|p(s)) = D(p^*(s)\|\tilde{p}(s)) + D(\tilde{p}(s)\|p(s)),$$

Therefore, for a given A, the optimal distribution $p(s)$ that minimizes the objective function is achieved when $p(s) = \tilde{p}(s)$, with the minimum value being $D(p^*(s)\|\tilde{p}(s))$. This value represents the mutual information between the components of $s = Wx$. Therefore, maximizing the likelihood with respect to W is equivalent to minimizing the mutual information between the components of $s = Wx$.

In practice, since the true distribution $p^*(y)$ is unknown, we need to replace mutual information or KL divergence with empirical estimates. Although empirical mutual information or likelihood can be computed and optimized with respect to W, a more typical approach in ICA is to use approximations of mutual information Amari et al., 1996; Comon, 1994; Hyvärinen, 1999 or alternative contrast functions Jutten and Herault, 1991.

The ability to identify independent components is constrained to potential permutations and scaling of the sources, as explained from the following theorem, derived in Eriksson and Koivunen (2004) Th. 3 :

Theorem 1 *Consider the model: $x = As$. Under the following conditions:*

(i) A is invertible,

(ii) The components $s_1, \ldots, s_K$ are independent, with at most one Gaussian distribution,

then matrix A is identifiable up to the post multiplication by $D^{-1}P^T$, where P is a permutation matrix and D a diagonal matrix with non-zero diagonal elements.

In other words A is identifiable up to a permutation of indexes and to signed scaling.

Kilian and Lütkepohl (2017) explain that this follows by noting that any other K-dimensional random vector, with independent components, that is obtained by a linear transformation of w_t must be just a reordering of the components of w_t, possibly with a reversed sign. Hence, the only linear transformations that preserve the independence of the components are of the form PB_0, where P is a permutation matrix that permutes the rows of B_0. Hence P^{-1} is also a permutation matrix and $B_0^{-1}P^{-1}$ is the matrix B_0^{-1} with permuted columns. In other words, the matrix of impact effects is unique apart from column permutations and column sign changes.

Independent Component Analysis can be seen as an extension of Principal Components Analysis (PCA). While PCA produces uncorrelated components based only on second-order moments, ICA goes further by generating independent components. Consequently, every ICA solution also satisfies the conditions of a PCA solution, although the reverse does not hold. In practice, ICA algorithms often leverage this relationship by using PCA as a preprocessing step. This involves whitening the random variable $\mathbf{y}$, which is achieved by multiplying $\mathbf{y}$ by a matrix $\mathbf{P}$, resulting in $\tilde{\mathbf{y}} = \mathbf{P}\mathbf{y}$ with an identity covariance matrix. The matrix $\mathbf{P}$ can be selected as the inverse of the square root of the covariance matrix of $\mathbf{y}$. This approach offers computational benefits: once the data are whitened, the resulting matrix $\mathbf{W}$ must be orthogonal (Hyvärinen et al., 2001).

A criterion for independence is required to determine the Independent Components. Independence is often defined through either the maximization of non-Gaussianity or the minimization of mutual information.

Maximization of Non-Gaussianity

Within the domain of blind source separation, we posit that the sources exhibit non-Gaussian characteristics. When we combine independent signals, the resultant signal becomes increasingly Gaussian, a phenomenon attributed to the Central Limit Theorem. To elaborate, if a set of signals $\mathbf{s} = (s_1, s_2, \ldots, s_M)$ are independent, with means $(\mu_1, \mu_2, \ldots, \mu_M)$ and variances $(\sigma_1^2, \sigma_2^2, \ldots, \sigma_M^2)$, then, for a large number M of signals $\mathbf{s}$, the signal

$$x = \sum_{j=1}^{M} s_j$$

has a probability density function (pdf) which is approximately Gaussian, with mean $\sum_j \mu_j$ and variance $\sum_j \sigma_j^2$.

From this perspective, by reverse logic, as explained in Stone (2004), it is plausible to perceive the sources as those signals exhibiting the highest deviation from Gaussianity. To extract the sources, it is necessary to assume that they follow a non-Gaussian distribution, with the possibility of at most one exception.

Minimization of Mutual Information

To assess independence, Comon (1994) explains that the Kullback-Leibler divergence can be used as follows :

Let Y be a random vector with values in $\mathbb{R}^m$, and let $p(y)$ denote its joint probability density function, with $p_j(y_j)$ representing the marginal density of the j-th component Y_j. The components of Y are mutually independent if and only if

$$p(y) = \prod_{i=1}^{m} p_i(y_i).$$

The mutual information, first introduced by Shannon (1948), between the m components $Y_1, \ldots, Y_m$ of Y is the KL divergence between the joint distribution of Y and the product of the marginal distributions of its components, as detailed in Equation (3.1) (Izenman, 2008). Specifically, the mutual information is given by:

$$\mathrm{MI}(Y) = \mathrm{KL}\left(p \parallel \prod_{j=1}^{m} p_j\right),$$

where $\prod_{j=1}^{m} p_j$ denotes the product of the marginal densities of Y_j.

The Kullback-Leibler divergence is nonnegative. Formally,

$$\begin{aligned}
\mathrm{KL}(p \parallel q) &= \mathbb{E}_p\left[\log \frac{p(y)}{q(y)}\right] \\
&\geq -\log \mathbb{E}_p\left[\frac{q(y)}{p(y)}\right] \\
&= -\log\left(\int q(y)\, dy\right) = 0,
\end{aligned}$$

following from Jensen's inequality $\mathbb{E}\{f(x)\} \geq f(\mathbb{E}\{x\})$ applied to the convex function $f(x) = -\log(x)$. The KL divergence is zero if and only if $p = q$. Here, $\mathbb{E}_p$ denotes the expectation with respect to the density p (Izenman, 2008).

This property motivates using mutual information to characterize the dependence between random variables.

3.1 FastICA

As described by Hyvärinen and Oja (2000), The FastICA algorithm has become a widely acknowledged method for blind source separation and serves as a benchmark in the field of Independent Component Analysis. Its core principle lies in the maximization of negentropy, which is an indicator of non-Gaussianity.

Negentropy is mathematically represented as:

$$J(y) = H(y_{\text{gauss}}) - H(y)$$

where y_{gauss} is a Gaussian random variable with the same covariance matrix as y, and H symbolizes entropy. Entropy, first introduced by Shannon (1948), is defined as:

$$H(y) = -\int f(y) \log f(y)\, dy$$

for continuous variables, and

$$H(Y) = -\sum_i P(Y = a_i) \log P(Y = a_i)$$

for discrete variables.

Calculating negentropy directly is challenging in practice. By employing the maximum entropy principle, an approximation can be derived as:

$$J(y) \approx [\mathbb{E}\{G(y)\} - \mathbb{E}\{G(\nu)\}]^2 \tag{3.2}$$

where ν is a standard normal variable and G is a non-quadratic function used to approximate negentropy.

3.1.1 FastICA for one Unit

We can start with the one-unit version of FastICA. By a "unit", we refer to a computational unit, possibly an artificial neuron, having a weight vector w that a neuron can update by a learning rule. The FastICA learning rule finds a direction, i.e. a unit vector w such that the projection $w^T x$ maximizes non-Gaussianity. Non-Gaussianity is here measured by the approximation of negentropy $J(w^T x)$. The variance of $w^T x$ must here be constrained to unity; for whitened data, this is equivalent to constraining the norm of w to be unity.

The basic form of the FastICA algorithm is as follows :

---

1. Choose an initial (e.g., random) weight vector $\mathbf{w}$.
2. Let $\mathbf{w}^+ = \mathbb{E}\{\mathbf{x}g(\mathbf{w}^{\mathrm{T}}\mathbf{x})\} - \mathbb{E}\{g'(\mathbf{w}^{\mathrm{T}}\mathbf{x})\}\mathbf{w}$.
3. Normalize the new weight vector: $\mathbf{w} = \frac{\mathbf{w}^+}{\|\mathbf{w}^+\|}$.
4. If not converged, go back to step 2.

Origin of the Nonlinearity Functions

The nonlinearity functions g, which are the derivatives of the nonquadratic function G used in Equation 3.2, are selected for their effectiveness in approximating negentropy (Hyvärinen, 1999).

- $\boldsymbol{g_1(u) = \tanh(a_1 u)}$ where $1 \leq a_1 \leq 2$ is a constant, often taken as $a_1 = 1$: This function is derived from the logistic function and is used because it approximates the cumulative distribution function of a super-Gaussian distribution. g_1 emphasize larger values of u, helping to identify the non-Gaussian features. The parameter a_1 adjusts the steepness of the function, ensuring flexibility in capturing different levels of non-Gaussianity.
- $\boldsymbol{g_2(u) = u\exp(-u^2/2)}$: This function is inspired by the Gaussian function and is also useful for identifying super-Gaussian sources. It behaves similarly to a Gaussian distribution but emphasizes the tails, making it effective for handling data with outliers. The form of g_2 allows it to act like a sigmoid function for small values of

u while attenuating larger values. Thus, this function provides robustness against extreme values and ensures stable convergence of the algorithm.

These specific forms of g are chosen because they are effective in practical scenarios for maximizing non-Gaussianity.

3.1.2 Relationship with Mutual Information

Negentropy has a direct relationship with mutual information. Assuming that the components y_i are uncorrelated and each has unit variance, the mutual information $I(y_1, y_2, \ldots, y_n)$ can be expressed as:

$$I(y_1, y_2, \ldots, y_n) = C - \sum_i J(y_i)$$

where C is a constant.

3.1.3 FastICA for Multiple Components

To find multiple independent components, the FastICA algorithm can be extended using two main approaches: deflation and symmetric decorrelation.

Deflation Approach

In the deflation approach, independent components are estimated one by one. After estimating each component, the algorithm removes its effect from the data and normalizes the resulting weight vector.

1. Let $\mathbf{w}_{p+1} \leftarrow \mathbf{w}_{p+1} - \sum_{j=1}^{p} (\mathbf{w}_{p+1}^{\mathrm{T}} \mathbf{w}_j) \mathbf{w}_j$.
2. Normalize $\mathbf{w}_{p+1} \leftarrow \frac{\mathbf{w}_{p+1}}{\|\mathbf{w}_{p+1}\|}$.

Symmetric Decorrelation

In the symmetric decorrelation approach, all components are estimated simultaneously, and the weight vectors are orthogonalized in each iteration.

1. Normalize the weight matrix: $\mathbf{W} \leftarrow \frac{\mathbf{W}}{\|\mathbf{W}\|}$.
2. Update the weight matrix: $\mathbf{W} \leftarrow \frac{3}{2}\mathbf{W} - \frac{1}{2}\mathbf{W}\mathbf{W}^{\mathrm{T}}\mathbf{W}$.

3.2 Kernel Independent Component Analysis

This section presents the Kernel Independent Component Analysis (KernelICA) algorithm as detailed in the paper by Bach and Jordan (2003). The explanation follows closely from their work. KernelICA, like other Independent Component Analysis techniques, is a blind source separation method aimed at extracting underlying latent vectors.

KernelICA does commit to a predefined distribution for x and uses a semiparametric approach instead. This flexibility makes KernelICA different from other methods, allowing it to adapt to various distributions.

KernelICA, as with other ICA algorithms, is related to minimizing mutual information among independent components. The latter indirectly maximizes the non-Gaussianity of the data distribution. However, calculating mutual information is hard when dealing with finite sample size, so Bach and Jordan (2003) uses an approximation of mutual information as the objective function.

In their approach, Bach and Jordan (2003) do not rely on a single nonlinear function whose expectations could robustly approximate mutual information. Instead, they opt for an entire spectrum of candidate nonlinear functions. Specifically, they utilize functions from a reproducing kernel Hilbert space, leveraging the kernel trick for efficient exploration of this functional space. This strategy enables the KernelICA algorithm to adjust to various sources, enhancing its robustness.

3.2.1 Canonical Correlation

To go further in the explanation of KernelICA, it is necessary to understand the concept of canonical correlation. Bach and Jordan (2003) detail in their paper how KernelICA makes use of this concept.

Canonical correlation analysis (CCA) (Hotelling, 1936) is a technique used to explore the linear relationships between two vector variables by identifying a pair of linear transformations of the vectors x_1 and x_2, which have dimensions p_1 and p_2, respectively. These transformations are designed so that one component from each transformed set is correlated with a corresponding component in the other set (Bach & Jordan, 2003). As a result, the correlation matrix between x_1 and x_2 is simplified to a block diagonal matrix

composed of two-by-two blocks, where each block takes the form $\begin{pmatrix} 1 & \rho_i \\ \rho_i & 1 \end{pmatrix}$. The values ρ_i are known as canonical correlations.

CCA can be described in a recursive manner, addressing one component at a time. Specifically, the first canonical correlation is defined as the maximum achievable correlation between the projections $\xi_1^T \mathbf{x}_1$ and $\xi_2^T \mathbf{x}_2$ of $\mathbf{x}_1$ and $\mathbf{x}_2$.

$$\begin{aligned} \rho(\mathbf{x}_1, \mathbf{x}_2) &= \max_{\xi_1, \xi_2} \operatorname{corr}(\xi_1^T \mathbf{x}_1, \xi_2^T \mathbf{x}_2) \\ &= \max_{\xi_1, \xi_2} \frac{\operatorname{cov}(\xi_1^T \mathbf{x}_1, \xi_2^T \mathbf{x}_2)}{\sqrt{\operatorname{var}(\xi_1^T \mathbf{x}_1) \operatorname{var}(\xi_2^T \mathbf{x}_2)}} \\ &= \max_{\xi_1, \xi_2} \frac{\xi_1^T \mathbf{C}_{12} \xi_2}{\sqrt{\xi_1^T \mathbf{C}_{11} \xi_1 \, \xi_2^T \mathbf{C}_{22} \xi_2}}. \end{aligned}$$

where $\mathbf{C} = \begin{pmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{21} & \mathbf{C}_{22} \end{pmatrix}$ is the covariance matrix of $(\mathbf{x}_1, \mathbf{x}_2)$. By differentiating with respect to ξ_1 and ξ_2, we derive the following:

$$\mathbf{C}_{12} \xi_2 = \frac{\xi_1^T \mathbf{C}_{12} \xi_2}{\xi_1^T \mathbf{C}_{11} \xi_1} \mathbf{C}_{11} \xi_1 \quad \text{and} \quad \mathbf{C}_{21} \xi_1 = \frac{\xi_1^T \mathbf{C}_{12} \xi_2}{\xi_2^T \mathbf{C}_{22} \xi_2} \mathbf{C}_{22} \xi_2.$$

By normalizing the vectors ξ_1 and ξ_2 such that $\xi_1^T \mathbf{C}_{11} \xi_1 = 1$ and $\xi_2^T \mathbf{C}_{22} \xi_2 = 1$, the CCA formulation simplifies to the following generalized eigenvalue problem:

$$\begin{pmatrix} 0 & \mathbf{C}_{12} \\ \mathbf{C}_{21} & 0 \end{pmatrix} \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix} = \rho \begin{pmatrix} \mathbf{C}_{11} & 0 \\ 0 & \mathbf{C}_{22} \end{pmatrix} \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix}.$$

The problem yields $p_1 + p_2$ eigenvalues: $\{\rho_1, -\rho_1, \ldots, \rho_p, -\rho_p, 0, \ldots, 0\}$. We can note that the generalized eigenvector problem can also be expressed as follows (Bach & Jordan, 2003):

$$\begin{pmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{21} & \mathbf{C}_{22} \end{pmatrix} \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix} = (1 + \rho) \begin{pmatrix} \mathbf{C}_{11} & 0 \\ 0 & \mathbf{C}_{22} \end{pmatrix} \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix},$$

The eigenvalues are given by $\{1 + \rho_1, 1 - \rho_1, \ldots, 1 + \rho_p, 1 - \rho_p, 1, \ldots, 1\}$. It is worth noting that determining the maximum generalized eigenvalue, $\lambda_{\max} = 1 + \rho_{\max}$, where $\rho_{\max}$ represents the highest canonical correlation, is equivalent to finding the minimum

generalized eigenvalue, $\lambda_{\min} = 1 - \rho_{\max}$. This latter value lies between zero and one, offering a more natural progression when extending the analysis to more than two variables. Thus, the primary objective will be to compute the minimum generalized eigenvalues.

Generalizing to More Than Two Variables

The extension to multiple variables can be described as follows. Given m multivariate random variables $\mathbf{x}_1, \ldots, \mathbf{x}_m$, the goal is to determine the smallest generalized eigenvalue $\lambda(\mathbf{x}_1, \ldots, \mathbf{x}_m)$ in the following problem:

$$\begin{pmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} & \cdots & \mathbf{C}_{1m} \\ \mathbf{C}_{21} & \mathbf{C}_{22} & \cdots & \mathbf{C}_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{C}_{m1} & \mathbf{C}_{m2} & \cdots & \mathbf{C}_{mm} \end{pmatrix} \begin{pmatrix} \xi_1 \\ \xi_2 \\ \vdots \\ \xi_m \end{pmatrix} = \lambda \begin{pmatrix} \mathbf{C}_{11} & 0 & \cdots & 0 \\ 0 & \mathbf{C}_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathbf{C}_{mm} \end{pmatrix} \begin{pmatrix} \xi_1 \\ \xi_2 \\ \vdots \\ \xi_m \end{pmatrix},$$

or, in short, $\mathbf{C}\xi = \lambda \mathbf{D}\xi$, where $\mathbf{C}$ represents the covariance matrix of $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m)$ and $\mathbf{D}$ is a block-diagonal matrix containing the covariances of the individual vectors $\mathbf{x}_i$.

The minimum generalized eigenvalue is confined to the range $[0, 1]$, while the maximum generalized eigenvalue varies depending on the dimensions of the variables. Therefore, focusing on the minimum generalized eigenvalue is more convenient.

3.2.2 Reproducing Kernel Hilbert Space

To explore the functional space efficiently, Bach and Jordan (2003) use a reproducing kernel Hilbert space and make use of the "kernel trick".

Bach and Jordan (2003) explain the kernel trick as follows:

> Let $K(x, y)$ be a Mercer kernel (Saitoh, 1988) on $X \in \mathbb{R}^p$, that is, a function for which the Gram matrix $K_{ij} = K(x_i, x_j)$ is positive semidefinite for any collection $\{x_i\}_{i=1,\ldots,N}$ in X. Corresponding to any such kernel K there is a map Φ from X to a *feature space* $\mathcal{F}$, such that:
>
> $$K(x, y) = \langle \Phi(x), \Phi(y) \rangle.$$
>
> That is, the kernel can be used to evaluate an inner product in the feature space. This is often referred to as the “kernel trick.”

A useful feature space is the *reproducing kernel Hilbert space* (RKHS) associated with a kernel K. Consider the set of functions $\{K(\cdot, x) : x \in X\}$, where $K(\cdot, x)$ represents a function indexed by x in the input space X. The span of these functions forms a linear space, which can always be completed into a Hilbert space, as shown by Saitoh (1988).

A key property of these Hilbert spaces is the *reproducing property*, which states that for any function f in the space:

$$f(x) = \langle K(\cdot, x), f \rangle \quad \forall f \in \mathcal{F}.$$

If we define the mapping $\Phi(x) = K(\cdot, x)$ from the input space to the RKHS, then:

$$\langle \Phi(x), \Phi(y) \rangle = \langle K(\cdot, x), K(\cdot, y) \rangle = K(x, y),$$

showing that $\Phi(x) = K(\cdot, x)$ is an instantiation of the "kernel trick."

To focus on translation-invariant kernels, consider kernels of the form $K(x, y) = k(x-y)$, where k is a function from $\mathbb{R}^p$ to $\mathbb{R}$. For these kernels, the feature space $\mathcal{F}$ is infinite-dimensional. The RKHS can be described using Fourier theory (Girosi et al., 1995). Specifically, for a given function k, the space $\mathcal{F}$ consists of functions $f \in L^2(\mathbb{R}^p)$ such that:

$$\int_{\mathbb{R}^p} \frac{|\hat{f}(\omega)|^2}{\nu(\omega)} \, d\omega < \infty,$$

where $\hat{f}(\omega)$ is the Fourier transform of f and $\nu(\omega)$ is the Fourier transform of k. For k to be a Mercer kernel, $\nu(\omega)$ must be real and positive. This implies that functions in $\mathcal{F}$ have rapidly decaying Fourier transforms, indicating that $\mathcal{F}$ consists of smooth functions.

Consider an isotropic Gaussian kernel in p dimensions:

$$K(x, y) = G_\sigma(x - y) = \exp\left(-\frac{1}{2\sigma^2}\|x - y\|^2\right).$$

The Fourier transform of this kernel is $\nu(\omega) = (2\pi\sigma^2)^{p/2} \exp\left(-\frac{\sigma^2}{2}\|\omega\|^2\right)$, and the feature space $\mathcal{F}_\sigma$ includes functions with rapidly decaying Fourier transforms. Alternatively, functions in $\mathcal{F}_\sigma$ can be interpreted as convolutions of functions from $L^2(\mathbb{R}^p)$ with a Gaussian kernel $G_{\sigma/\sqrt{2}}(x) = \exp\left(-\frac{1}{\sigma^2}\|x\|^2\right)$.

As σ increases from 0 to ∞, the Gaussian kernel $G_{\sigma/\sqrt{2}}$ transitions from an impulse to

a constant function, and the corresponding function spaces $\mathcal{F}_\sigma$ change from $L^2(\mathbb{R}^p)$ to an empty set.

3.2.3 $\mathcal{F}$-Correlation

The foundation of the KernelICA algorithm is the concept of $\mathcal{F}$-correlation :

Theorem 2 *Let x_1 and x_2 be random variables in $\mathcal{X} = \mathbb{R}^p$. Let K_1 and K_2 be Mercer kernels with feature maps Φ_1 and Φ_2, and feature spaces $\mathcal{F}_1$ and $\mathcal{F}_2 \subseteq \mathbb{R}^{\mathcal{X}}$. Then the canonical correlation $\rho_{\mathcal{F}}$ between $\Phi_1(x_1)$ and $\Phi_2(x_2)$, which is defined as*

$$\rho_{\mathcal{F}} = \max_{(f_1,f_2)\in\mathcal{F}_1\times\mathcal{F}_2} corr\left(\langle\Phi_1(x_1), f_1\rangle, \langle\Phi_2(x_2), f_2\rangle\right),$$

is equal to

$$\rho_{\mathcal{F}} = \max_{(f_1,f_2)\in\mathcal{F}_1\times\mathcal{F}_2} corr(f_1(x_1), f_2(x_2)).$$

This comes from the reproducing property explained in Section 3.2.2.

The $\mathcal{F}$-correlation represents the highest correlation achievable between one-dimensional linear projections of $\Phi(x_1)$ and $\Phi(x_2)$, which aligns with the concept of the first canonical correlation between $\Phi(x_1)$ and $\Phi(x_2)$.

If the variables are independent, the $\mathcal{F}$-correlation will be zero. Moreover, if the space $\mathcal{F}$ is large enough, the converse holds true as well. We will take the case of a gaussian kernel.

Theorem 3 *If $\mathcal{F}$ is the RKHS corresponding to a Gaussian kernel on $X = \mathbb{R}$, then $\rho_{\mathcal{F}} = 0$ if and only if the variables y_1 and y_2 are independent.*

This theorem does not hold if the function space has finite dimension.

In order to compute the canonical correlations in an RKHS, we now need to to develop an empirical estimate of the $\mathcal{F}$-correlation. The approach of Bach and Jordan (2003) is to develop a kernelized version of canonical correlations.

Kernelization of CCA

For two variables, the goal is to maximize the correlation between their projections in the feature space. Directly mapping data points to this high-dimensional space and using CCA would be computationally inefficient or even impossible. Therefore, we perform all calculations in the input space instead.

Consider the sets $\{x_1^1, \ldots, x_1^N\}$ and $\{x_2^1, \ldots, x_2^N\}$, which represent N empirical observations of x_1 and x_2, respectively. Let the corresponding feature space representations be $\{\Phi(x_1^1), \ldots, \Phi(x_1^N)\}$ and $\{\Phi(x_2^1), \ldots, \Phi(x_2^N)\}$. Assume, for the moment, that the data are centered in the feature space, meaning $\sum_{k=1}^N \Phi(x_1^k) = \sum_{k=1}^N \Phi(x_2^k) = 0$. The empirical canonical correlation, denoted as $\hat{\rho}_{\mathcal{F}}(x_1, x_2)$, is based on empirical covariances rather than population covariances. As we will see, this empirical canonical correlation $\hat{\rho}_{\mathcal{F}}(x_1, x_2)$ is determined solely by the Gram matrices K_1 and K_2 corresponding to these observations. Therefore, we will also use the notation $\hat{\rho}_{\mathcal{F}}(K_1, K_2)$ to refer to this canonical correlation.

We need to focus solely on the subspace of $\mathcal{F}$ that encompasses the span of the data. Given fixed f_1 and f_2, the empirical covariance of the projections in feature space can be expressed as:

$$\hat{\text{cov}}(\langle \Phi(x_1), f_1 \rangle, \langle \Phi(x_2), f_2 \rangle) = \frac{1}{N} \sum_{k=1}^N \langle \Phi(x_1^k), f_1 \rangle \langle \Phi(x_2^k), f_2 \rangle.$$

Let $\mathcal{S}_1$ and $\mathcal{S}_2$ denote the linear spaces spanned by the Φ-images of the data points. Consequently, we can express f_1 and f_2 as $f_1 = \sum_{k=1}^N \alpha_k^1 \Phi(x_1^k) + f_1^\perp$ and $f_2 = \sum_{k=1}^N \alpha_k^2 \Phi(x_2^k) + f_2^\perp$, where $f_1^\perp$ and $f_2^\perp$ are components orthogonal to $\mathcal{S}_1$ and $\mathcal{S}_2$, respectively. Then, we have:

$$\hat{\text{cov}}(\langle \Phi(x_1), f_1 \rangle, \langle \Phi(x_2), f_2 \rangle) = \frac{1}{N} \sum_{k=1}^N \left\langle \Phi(x_1^k), \sum_{i=1}^N \alpha_i^1 \Phi(x_1^i) \right\rangle \left\langle \Phi(x_2^k), \sum_{j=1}^N \alpha_j^2 \Phi(x_2^j) \right\rangle.$$

$$= \frac{1}{N} \sum_{k=1}^N \sum_{i=1}^N \sum_{j=1}^N \alpha_i^1 K_1(x_1^i, x_1^k) K_2(x_2^j, x_2^k) \alpha_j^2 = \frac{1}{N} (\alpha_1^T K_1 K_2 \alpha_2),$$

where K_1 and K_2 represent the Gram matrices corresponding to the data sets $\{x_1^i\}$ and $\{x_2^i\}$, respectively. Additionally, we derive:

$$\hat{\text{var}}(\langle \Phi(x_1), f_1 \rangle) = \frac{1}{N}(\alpha_1^T K_1^2 \alpha_1) \quad \text{and} \quad \hat{\text{var}}(\langle \Phi(x_2), f_2 \rangle) = \frac{1}{N}(\alpha_2^T K_2^2 \alpha_2).$$

Combining these results, the kernelized CCA problem reduces to the following maximization problem:

$$\hat{\rho}_{\mathcal{F}}(K_1, K_2) = \max_{\alpha_1, \alpha_2 \in \mathbb{R}^N} \frac{\alpha_1^T K_1 K_2 \alpha_2}{\sqrt{\alpha_1^T K_1^2 \alpha_1}\sqrt{\alpha_2^T K_2^2 \alpha_2}}. \tag{3.3}$$

Equation 3.3 is equivalent to conducting CCA on two vectors of dimension N, where the covariance matrix is given by:

$$\begin{pmatrix} K_1^2 & K_1 K_2 \\ K_2 K_1 & K_2^2 \end{pmatrix}.$$

Therefore, we can perform a kernelized version of CCA by solving the following generalized eigenvalue problem:

$$\begin{pmatrix} 0 & K_1 K_2 \\ K_2 K_1 & 0 \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} = \rho \begin{pmatrix} K_1^2 & 0 \\ 0 & K_2^2 \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix}. \tag{3.4}$$

If the points $\Phi(x_i^k)$ are not centered, we can still compute the Gram matrix for the centered data points using the following method: Given the Gram matrix K for the non-centered data, the centered Gram matrix $\tilde{K}$ is obtained as $\tilde{K} = N_0 K N_0$, where $N_0 = I - \frac{1}{N}\mathbf{1}\mathbf{1}^T$ is the centering matrix (Schölkopf et al., 1998). We assume that any Gram matrix used has been centered in this way.

Regularization

The issue with the eigenequation 3.4 is that $\begin{pmatrix} K_1^2 & 0 \\ 0 & K_2^2 \end{pmatrix}$ will be singular because centering makes both Gram matrices, $\mathbf{K}_1$ and $\mathbf{K}_2$, singular. Additionally, all pairs of "kernel canonical variates" in feature space will be perfectly correlated, even if the non-centered $\mathbf{K}_1$ and $\mathbf{K}_2$ are invertible. (Izenman, 2008)

Therefore, this "naive" kernelization approach does not yield a practical estimator for general kernels. However, it serves as a foundation for developing a useful regularized estimator, which we will now explore. Our regularization strategy involves penalizing

the RKHS norms of f_1 and f_2, thereby offering control over the statistical properties of KernelICA.

In particular, we define the regularized F-correlation $\rho_{\kappa,\mathcal{F}}$ as:

$$\rho_{\kappa,\mathcal{F}} = \max_{f_1,f_2\in\mathcal{F}} \frac{\operatorname{cov}(f_1(x_1), f_2(x_2))}{\sqrt{\operatorname{var}(f_1(x_1)) + \kappa\|f_1\|_{\mathcal{F}}^2}\sqrt{\operatorname{var}(f_2(x_2)) + \kappa\|f_2\|_{\mathcal{F}}^2}},$$

where κ is a small positive constant. The choice of κ is detailed when we discuss the choice of free parameters. Note that the regularized $\mathcal{F}$-correlation inherits the independence characterization property of the $\mathcal{F}$-correlation. In order to estimate it from a finite sample, we expand $\operatorname{var}(f_1(x_1)) + \kappa\|f_1\|_{\mathcal{F}}^2$ up to second order in κ, to obtain:

$$\operatorname{var}(f_1(x_1)) + \kappa\|f_1\|_{\mathcal{F}}^2 = \frac{1}{N}\alpha_1^T K_1^2\alpha_1 + \kappa\alpha_1^T K_1\alpha_1 \approx \frac{1}{N}\alpha_1^T(K_1 + \frac{N\kappa}{2}I)^2\alpha_1.$$

Thus, the regularized kernel CCA problem becomes:

$$\hat{\rho}_{\kappa,\mathcal{F}}(K_1, K_2) = \max_{\alpha_1,\alpha_2\in\mathbb{R}^N} \frac{\alpha_1^T K_1 K_2 \alpha_2}{\sqrt{\alpha_1^T(K_1 + \frac{N\kappa}{2}I)^2\alpha_1}\sqrt{\alpha_2^T(K_2 + \frac{N\kappa}{2}I)^2\alpha_2}}, \tag{3.5}$$

The value of κ determines the balance between the penalty terms and the variance terms. As κ approaches zero, the variance term becomes more dominant. However, as κ increases, the influence of the penalty term grows. Therefore, careful consideration is needed when choosing the value of κ. (Izenman, 2008)

Differentiating Equation 3.5 with respect to α_1 and α_2 and then setting the result equal to zero yields the following generalized eigenvalue problem :

$$\begin{pmatrix} 0 & K_1K_2 \\ K_2K_1 & 0 \end{pmatrix}\begin{pmatrix}\alpha_1 \\ \alpha_2\end{pmatrix} = \rho\begin{pmatrix}(K_1 + \frac{N\kappa}{2}I)^2 & 0 \\ 0 & (K_2 + \frac{N\kappa}{2}I)^2\end{pmatrix}\begin{pmatrix}\alpha_1 \\ \alpha_2\end{pmatrix}.$$

The regularized kernel CCA problem can be reformulated into a more computationally stable eigenvalue problem. Additionally, the regularized first canonical correlation offers a key statistical benefit compared to the unregularized version—it acts as a consistent estimator of the regularized $\mathcal{F}$-correlation. This implies that as the sample size N approaches infinity, the estimate converges in probability to the true population value.

Generalizing to More than Two Variables

The extension of regularized kernelized canonical correlation analysis (CCA) to more than two sets of variables follows naturally from the generalization of CCA itself to multiple variables. We define $\mathcal{K}_\kappa$ as an $mN \times mN$ matrix, where its blocks are $(\mathcal{K}_\kappa)_{ij} = K_i K_j$ for $i \neq j$, and $(\mathcal{K}_\kappa)_{ii} = (K_i + \frac{N\kappa}{2} I)^2$. Additionally, let $\mathcal{D}_\kappa$ be the $mN \times mN$ block-diagonal matrix with blocks $(K_i + \frac{N\kappa}{2} I)^2$. This leads us to the following generalized eigenvalue problem:

$$\begin{pmatrix} (K_1 + \frac{N\kappa}{2} I)^2 & K_1 K_2 & \cdots & K_1 K_m \\ K_2 K_1 & (K_2 + \frac{N\kappa}{2} I)^2 & \cdots & K_2 K_m \\ \vdots & \vdots & \ddots & \vdots \\ K_m K_1 & K_m K_2 & \cdots & (K_m + \frac{N\kappa}{2} I)^2 \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_m \end{pmatrix}$$
$$= \lambda \begin{pmatrix} (K_1 + \frac{N\kappa}{2} I)^2 & 0 & \cdots & 0 \\ 0 & (K_2 + \frac{N\kappa}{2} I)^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & (K_m + \frac{N\kappa}{2} I)^2 \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_m \end{pmatrix}$$

This can be compactly written as $\mathcal{K}_\kappa \alpha = \lambda \mathcal{D}_\kappa \alpha$. The smallest eigenvalue from this problem is denoted by $\hat{\lambda}_{\kappa,\mathcal{F}}(K_1, \ldots, K_m)$ and is termed the first kernel canonical correlation, also referred to as an (empirical) $\mathcal{F}$-correlation.

In the case of two variables, we defined a function $\rho_{\mathcal{F}}(x_1, x_2)$ that is based on the covariances of the random variables $\Phi(x_1)$ and $\Phi(x_2)$. From this, we derived an empirical contrast function $\hat{\rho}_{\kappa,\mathcal{F}}(x_1, x_2)$ by replacing population covariances with empirical covariances and introducing regularization. For the case involving m variables, we have directly defined the empirical function $\hat{\lambda}_{\kappa,\mathcal{F}}(K_1, \ldots, K_m)$.

To define our contrast functions, we use the negative logarithm of canonical correlations due to the relationship between canonical correlations and mutual information, as discussed in the section 3.2.4. Specifically, we define a contrast function $I_{\lambda_{\mathcal{F}}}(x_1, \ldots, x_m) = -\frac{1}{2} \log \lambda_{\mathcal{F}}(x_1, \ldots, x_m)$ and aim to minimize this function.

For the empirical contrast function, we denote it as $\hat{I}_{\lambda_{\mathcal{F}}}(K_1, \ldots, K_m) = -\frac{1}{2} \log \hat{\lambda}_{\kappa,\mathcal{F}}(K_1, \ldots, K_m)$, highlighting that this contrast function is determined by the data solely through the Gram matrices.

3.2.4 KernelICA Algorithms

Now that the contrast function has been defined, one can understand the KernelICA-KCCA algorithm from Bach and Jordan (2003). The resulting algorithm is as follows:

Input: Data vectors $y^1, y^2, \ldots, y^N$
Kernel $K(x, y)$

1. Whiten the data
2. Minimize (with respect to W) the contrast function $C(W)$ defined as:
   (a) Compute the centered Gram matrices $K_1, K_2, \ldots, K_m$ of the estimated sources $\{x_1, x_2, \ldots, x_N\}$, where $x^i = Wy^i$
   (b) Define $\hat{\lambda}^{\kappa}_{\mathcal{F}}(K_1, \ldots, K_m)$ as the minimal eigenvalue of the generalized eigenvector equation $\mathcal{K}_\kappa \alpha = \lambda \mathcal{D}_\kappa \alpha$
   (c) Define $C(W) = \hat{I}_{\lambda_\mathcal{F}}(K_1, \ldots, K_m) = -\frac{1}{2} \log \hat{\lambda}^{\kappa}_{\mathcal{F}}(K_1, \ldots, K_m)$

Output: W

Kernel Generalized Variance

The concept of $\mathcal{F}$-correlation can be expanded to more closely approximate mutual information.

For jointly Gaussian variables x_1 and x_2, the mutual information is given by:

$$I(x_1, x_2) = -\frac{1}{2} \sum_{i=1}^{p} \log(1 - \rho_i^2)$$

where ρ_i represents the canonical correlations.

This formulation can be adapted for m Gaussian variables, leading to:

$$I(x_1, x_2, \ldots, x_m) = -\frac{1}{2} \log \frac{\det C}{\det D} = -\frac{1}{2} \sum_{i=1}^{P} \log(\lambda_i)$$

Here, λ_i are the eigenvalues from solving the equation $C\xi = \lambda D\xi$.

The kernel generalized variance (KGV) is then defined as:

$$\hat{\delta}_{\mathcal{F}}^{\kappa}(K_1, \ldots, K_m) = \frac{\det \mathcal{K}_\kappa}{\det \mathcal{D}_\kappa}$$

And the contrast function is formulated as:

$$\hat{I}_{\delta_{\mathcal{F}}} = -\frac{1}{2} \log \hat{\delta}_{\mathcal{F}}^{\kappa}(K_1, \ldots, K_m)$$

This motivates the following algorithm known as KernelICA-KGV.

Input: Data vectors $y_1, y_2, \ldots, y_N$
Kernel $K(x, y)$

1. Whiten the data.
2. Minimize (with respect to W) the contrast function $C(W)$ defined as:
   (a) Compute the centered Gram matrices $K_1, K_2, \ldots, K_m$ of the estimated sources $\{x^1, x^2, \ldots, x^N\}$, where $x^i = Wy^i$.
   (b) Define $\hat{\delta}_{\mathcal{F}}^{\kappa}(K_1, \ldots, K_m) = \det \mathcal{K}_\kappa / \det \mathcal{D}_\kappa$.
   (c) Define $C(W) = \hat{I}_{\delta_{\mathcal{F}}}(K_1, \ldots, K_m) = -\frac{1}{2} \log \hat{\delta}_{\mathcal{F}}^{\kappa}(K_1, \ldots, K_m)$.

Output: W

Free Parameters

The KernelICA algorithms involve two adjustable parameters: the regularization parameter κ and the kernel width σ, assuming identical Gaussian kernels are used for each source. In their experimental work, Bach and Jordan (2003) found that the KernelICA algorithms were reasonably robust to the settings of these parameters. Their choices were to set $\kappa = 2 \times 10^{-3}$, $\sigma = 1/2$ for large samples ($N > 1000$) and $\kappa = 2 \times 10^{-2}$, $\sigma = 1$ for smaller samples ($N \leq 1000$).

Leurgans et al. (1993) suggest using cross-validation to choose a good value for κ. However, they found that cross-validation is more effective for the leading canonical variate than for the subsequent ones.

For a finite N, selecting a very small σ results in diagonal Gram matrices, making the criteria trivial. Conversely, as N increases, the Kernel Generalized Variance (KGV) converges to the mutual information as σ approaches zero, indicating that σ should be as small as possible. However, computational considerations must also be taken into account—when σ is too small, the eigenvalues of the Gram matrices decay more slowly, leading to increased computational complexity. This issue can be addressed by choosing an appropriate κ; specifically, the algorithm could adjust κ to maintain a constant number of retained eigenvalues for each Gram matrix (Bach & Jordan, 2003).

3.3 Distance Covariance (DCov)

To estimate the Independent Component (IC) model, Matteson and Tsay (2017) propose leveraging distance covariance, a statistical measure introduced by Székely et al. (2007).

Distance covariance $I(\mathbf{X}^{(1)}, \mathbf{X}^{(2)})$ serves as a multivariate measure of independence between random vectors $\mathbf{X}^{(1)} \in \mathbb{R}^{d_1}$ and $\mathbf{X}^{(2)} \in \mathbb{R}^{d_2}$, where d_1 and d_2 are arbitrary dimensions. It applies to distributions with finite first absolute moments. In this context, $|\cdot|$ denotes the Euclidean distance. Let $(\mathbf{X}^{(1)\prime}, \mathbf{X}^{(2)\prime})$ and $(\mathbf{X}^{(1)\prime\prime}, \mathbf{X}^{(2)\prime\prime})$ denote independent and identically distributed (i.i.d.) copies of $(\mathbf{X}^{(1)}, \mathbf{X}^{(2)})$. Distance covariance is defined as follows:

$$\begin{aligned} I(\mathbf{X}^{(1)}, \mathbf{X}^{(2)}) =& \mathbb{E}\left[|\mathbf{X}^{(1)} - \mathbf{X}^{(1)\prime}||\mathbf{X}^{(2)} - \mathbf{X}^{(2)\prime}|\right] \\ &+ \mathbb{E}\left[|\mathbf{X}^{(1)} - \mathbf{X}^{(1)\prime}|\right]\mathbb{E}\left[|\mathbf{X}^{(2)} - \mathbf{X}^{(2)\prime}|\right] \\ &- \mathbb{E}\left[|\mathbf{X}^{(1)} - \mathbf{X}^{(1)\prime}||\mathbf{X}^{(2)} - \mathbf{X}^{(2)\prime\prime}|\right] \\ &- \mathbb{E}\left[|\mathbf{X}^{(1)} - \mathbf{X}^{(1)\prime\prime}||\mathbf{X}^{(2)} - \mathbf{X}^{(2)\prime}|\right]. \end{aligned}$$

The distance covariance measure has several important properties. It is always non-negative, meaning $0 \leq I(\mathbf{X}^{(1)}, \mathbf{X}^{(2)})$. Additionally, it is invariant under orthogonal transformations: for any constant vectors a_1 and a_2, non-zero scalars b_1 and b_2, and orthogonal matrices C_1 and C_2 of conforming dimensions, the measure satisfies $I(a_1 + b_1 C_1 \mathbf{X}^{(1)}, a_2 + b_2 C_2 \mathbf{X}^{(2)}) = |b_1||b_2|^{1/2} I(\mathbf{X}^{(1)}, \mathbf{X}^{(2)})$. Finally, the distance covariance $I(\mathbf{X}^{(1)}, \mathbf{X}^{(2)})$ equals zero if and only if $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ are independent.

For a sample $\{(X_i^{(1)}, X_i^{(2)}) : i = 1, \ldots, n\}$, the empirical distance covariance $I_n(X^{(1)}, X^{(2)})$ is approximated using U-statistics. The approximation involves:

$$T_{1,n}(X^{(1)}, X^{(2)}) = \frac{1}{\binom{n}{2}} \sum_{i<j} \left|X_i^{(1)} - X_j^{(1)}\right| \left|X_i^{(2)} - X_j^{(2)}\right|,$$

$$T_{2,n}(X^{(1)}, X^{(2)}) = \left(\frac{1}{\binom{n}{2}} \sum_{i<j} \left|X_i^{(1)} - X_j^{(1)}\right|\right) \left(\frac{1}{\binom{n}{2}} \sum_{i<j} \left|X_i^{(2)} - X_j^{(2)}\right|\right),$$

$$\begin{aligned} T_{3,n}(X^{(1)}, X^{(2)}) = \frac{1}{\binom{n}{3}} \sum_{i<j<k} \frac{1}{3} \Bigg[& \left|X_i^{(1)} - X_j^{(1)}\right| \left|X_i^{(2)} - X_k^{(2)}\right| \\ & + \left|X_i^{(1)} - X_k^{(1)}\right| \left|X_i^{(2)} - X_j^{(2)}\right| + \left|X_j^{(1)} - X_k^{(1)}\right| \left|X_j^{(2)} - X_k^{(2)}\right| \\ & + \left|X_i^{(1)} - X_j^{(1)}\right| \left|X_j^{(2)} - X_k^{(2)}\right| + \left|X_i^{(1)} - X_k^{(1)}\right| \left|X_k^{(2)} - X_j^{(2)}\right| \\ & + \left|X_j^{(1)} - X_k^{(1)}\right| \left|X_i^{(2)} - X_k^{(2)}\right| \Bigg]. \end{aligned}$$

Therefore, the empirical distance covariance is given by:

$$I_n(X^{(1)}, X^{(2)}) = T_{1,n}(X^{(1)}, X^{(2)}) + T_{2,n}(X^{(1)}, X^{(2)}) - T_{3,n}(X^{(1)}, X^{(2)}).$$

As the sample size n approaches infinity, $I_n(X^{(1)}, X^{(2)})$ converges to $I(X^{(1)}, X^{(2)})$.

In the context of ICA, the objective is to estimate the unmixing matrix W_θ such that the components of $S(\theta) = W_\theta Z$ are as independent as possible. Z is the uncorrelated observed data. To achieve this, Matteson and Tsay (2017) propose defining the objective function $J_n(\theta)$ as follows:

$$J_n(\theta) = \sum_{k=1}^{d-1} I_n(S_k(\theta), S_{k^+}(\theta)).$$

where $S_k(\theta)$ is the k-th component of $S(\theta)$, and $S_{k+}(\theta)$ refers to the subsequent components, i.e., the components that follow $S_k(\theta)$ in the sequence.

The task then becomes finding θ that minimizes $J_n(\theta)$, which estimates the unmixing matrix W_θ. The objective function has $d(d-1)/2$ parameters, which can be estimated jointly.

In the two-dimensional case, where there is only one angle to estimate, the mixing matrix is:

$$\mathbf{G}(\theta) = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}.$$

In the three-dimensional case, where there are three angles to estimate, the mixing matrix is expressed as:

$$\mathbf{G}(\theta) = \begin{bmatrix} \cos\theta_1 & -\sin\theta_1 & 0 \\ \sin\theta_1 & \cos\theta_1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \cos\theta_2 & 0 & -\sin\theta_2 \\ 0 & 1 & 0 \\ \sin\theta_2 & 0 & \cos\theta_2 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\theta_3 & -\sin\theta_3 \\ 0 & \sin\theta_3 & \cos\theta_3 \end{bmatrix}.$$

In summary, the ICA problem involves finding $\hat{\theta}$ such that the dependence among the components resulting from $\mathbf{G}(\hat{\theta})^{-1}u_t$ is minimized. The resulting mixing matrix B_0 is set to $\mathbf{G}(\hat{\theta})$.

Chapter 4

Monte Carlo Simulation

To assess the efficacy of KernelICA in identifying structural innovations and comparing its performance with other identification methodologies, we will adopt the comprehensive evaluation framework proposed by Moneta and Pallante (2022). This approach is inspired by the methodology outlined by Matteson and Tsay (2017). Our analysis will evaluate the performance of three distinct techniques: KernelICA, FastICA, and Distance-Covariance.

4.1 Data Generating Process

In our simulation, we utilize data generated from a p-generalized normal distribution. This choice is motivated by the distribution's adjustable Gaussianity parameter, p, offering a versatile framework for our analysis. As detailed by Kalke and Richter (2013), this distribution is defined by the probability density function :

$$f(x,p) = \frac{p^{1-\frac{1}{p}}}{2\Gamma\left(\frac{1}{p}\right)} \exp\left(-\frac{|x|^p}{p}\right), \quad x \in \mathbb{R},\, p > 0$$

Here, Γ represents the gamma function, and p serves as a shape parameter that adjusts the distribution's gaussianity. Specifically, the distribution is gaussian at $p = 2$. For $p > 2$, it becomes sub-gaussian, while values of $p < 2$ lead to a super-gaussian distribution.

This distribution is useful for our study because it allows us to systematically examine how the gaussianity of data, controlled by p, affects the performance of the methods we are investigating. Figure 4.1, illustrates the density curves for various p values, showcasing how the distribution deviates from gaussianity.

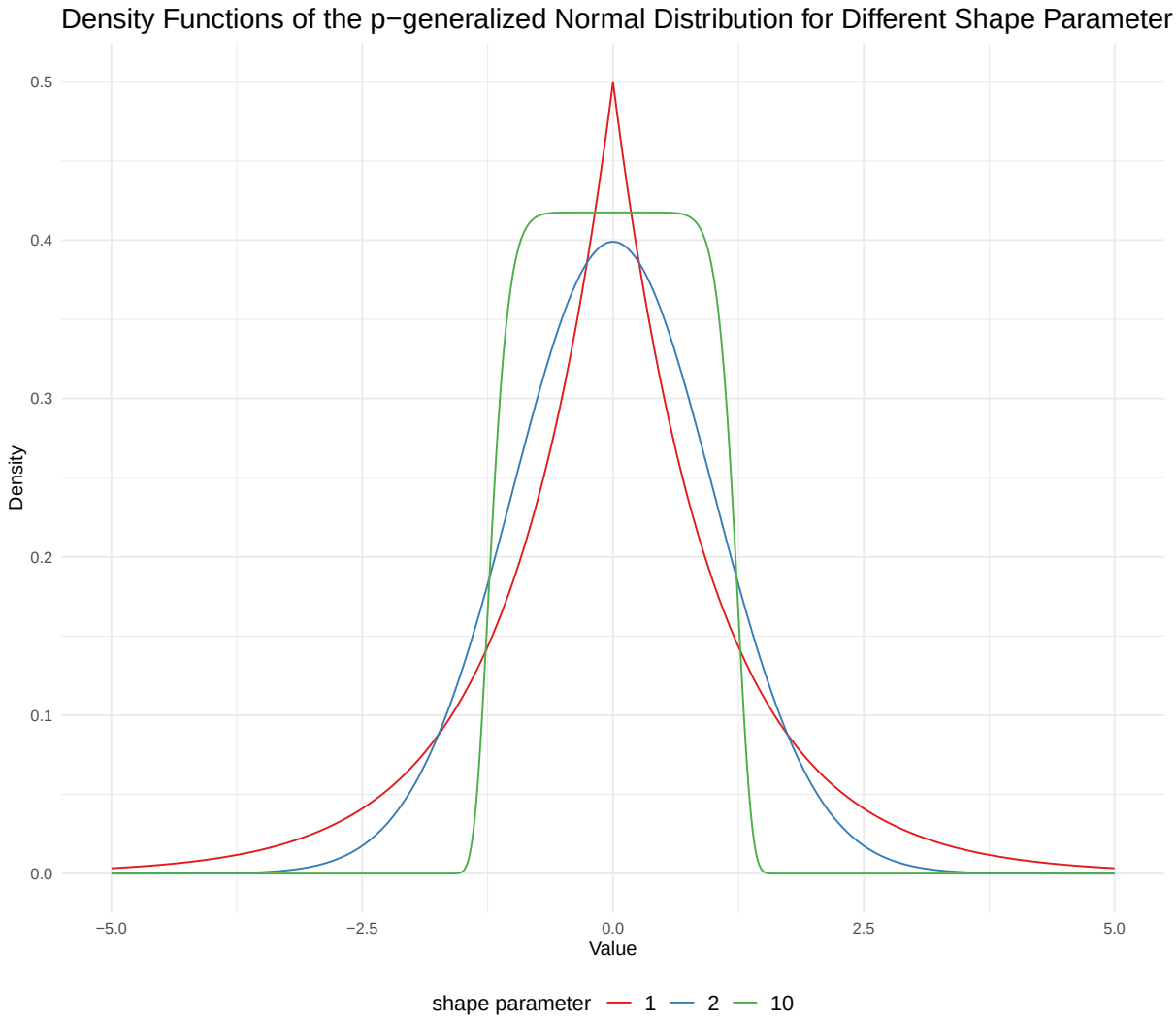

Figure 4.1: Density functions for different p values in the p-generalized normal distribution

4.2 Evaluating the Performance

4.2.1 Simulation Setup

Our simulation setup is designed to evaluate the performance of ICA methods by assessing how accurately they estimate a randomly generated mixing matrix in each experiment. The random mixing matrices are generated using the `mixmat` function from the `ProDenICA` R package (Hastie & Tibshirani, 2010). The matrices are constructed by performing singular value decomposition, with singular values drawn uniformly from the interval [1,

2], ensuring a condition number between 1 and 2.

We opted for a sample size of 200 to maintain a balance between computational efficiency and adequate sample size. Our goal is to assess the performance of ICA methods across datasets with different sizes and complexities. To achieve this, we designed our evaluation to cover three key scenarios: the first involving datasets with 2 variables, the second with 3 variables, and the third with 4 variables.

To study how ICA methods handle different non-Gaussian levels, we changed the shape parameter (p) across a range of values. This approach helps us assess how well ICA methods adapt to varying non-Gaussian conditions.

Each method is initialized five times for each Monte Carlo iteration, with the selection of the best-performing initialization for further analysis. This comparative analysis is facilitated through 1000 Monte Carlo iterations to ensure the generation of substantial statistical data.

Regarding the implementation of specific ICA methods, KernelICA is applied using the `kernel_ica` function from the `KernelICA` R-package. The FastICA algorithm is made available through the `fastICA` R-package, and the Distance-Covariance method is implemented using the `steadyICA` package.

Choice of Parameters

As advised by Bach and Jordan (2003), for a Gaussian kernel, we have chosen $\kappa = 2 \times 10^{-2}$ and $\sigma = 1$ since our sample size is less than 1000.

For FastICA, we have chosen the function $g(x) = tanh(ax)$, where $a = 1$ and the independent components are extracted in parallel.

4.2.2 Evaluation Metric

After generating independent vectors, we combine them using a randomly generated mixing matrix. The accuracy of the ICA methods' estimated matrices is then assessed by comparing them to the original mixing matrix. For this comparison, we employ a metric proposed by Ilmonen et al. (2010) and recommended by Matteson and Tsay (2017). This metric, called the Minimum Distance Index (MDI), is specifically designed to address the

intrinsic indeterminacies of the ICA model, such as the limitations in identifying the scale, sign, and permutation of the columns of B_0.

The metric is defined as follows:

$$D\left(B_0^{(m)}, \hat{B}_0^{(m)}\right) = \frac{1}{\sqrt{k-1}} \inf_C \left\| C\hat{B}_0^{(m)} B_0^{(m)-1} - I_k \right\|_F$$

In this equation, $B_0^{(m)}$ represents the original randomly generated mixing matrix for the m-th replication, and $\hat{B}_0^{(m)}$ denotes its estimated counterpart. The term C stands for $P_{\pm}D_+$, where $P_{\pm}$ is any signed permutation matrix of size $k \times k$, and D_+ is any $k \times k$ diagonal matrix with strictly positive diagonal elements. The Frobenius norm is represented by $\|\cdot\|_F$. The closer the Minimum Distance Index is to zero, the more accurate the estimate $\hat{B}_0^{(m)}$ is in relation to the true mixing matrix $B_0^{(m)}$, providing a quantitative measure of the ICA method's performance.

Other metrics can be used; for example, Bach and Jordan (2003) selected the Amari errors, defined as

$$d(V,W) = \frac{1}{2m}\sum_{i=1}^{m}\left[\frac{\sum_{j=1}^{m}|a_{ij}|}{\max_j |a_{ij}|} - 1\right] + \frac{1}{2m}\sum_{j=1}^{m}\left[\frac{\sum_{i=1}^{m}|a_{ij}|}{\max_i |a_{ij}|} - 1\right],$$

where $a_{ij} = (VW^{-1})_{ij}$. Amari errors use the ℓ_1 norm, while the Minimum Distance Index (MDI) uses the Frobenius norm, related to the ℓ_2 norm. The choice of MDI over Amari errors is motivated by the fact that MDI does not depend on the model formulation, since the Amari errors are not affine invariant. Therefore, there might be pitfalls when different algorithms are compared using the Amari index. (Ilmonen et al., 2010)

4.2.3 Implementation Details

The simulation was automated through a series of R scripts, facilitating the execution of multiple Monte Carlo iterations across the defined values of p. Each iteration involved:

- Generating synthetic data using the p-generalized normal distribution.
- Mixing the generated data using a random matrix.
- Applying various ICA methods to estimate the mixing and unmixing matrices.
- Calculating the MDI for each method to evaluate performance.

The process was parallelized using the `parallel` and `doParallel` packages in R to enhance computational efficiency.

4.3 Results

This section presents our simulation results, shown in figures displaying the mean evaluation metric for each method and value of p. These results reveal how different ICA methods perform with varying shape parameters (p) of the p-generalized normal distribution used for generating synthetic data.

Figure 4.2 displays the results for a model with two variables.

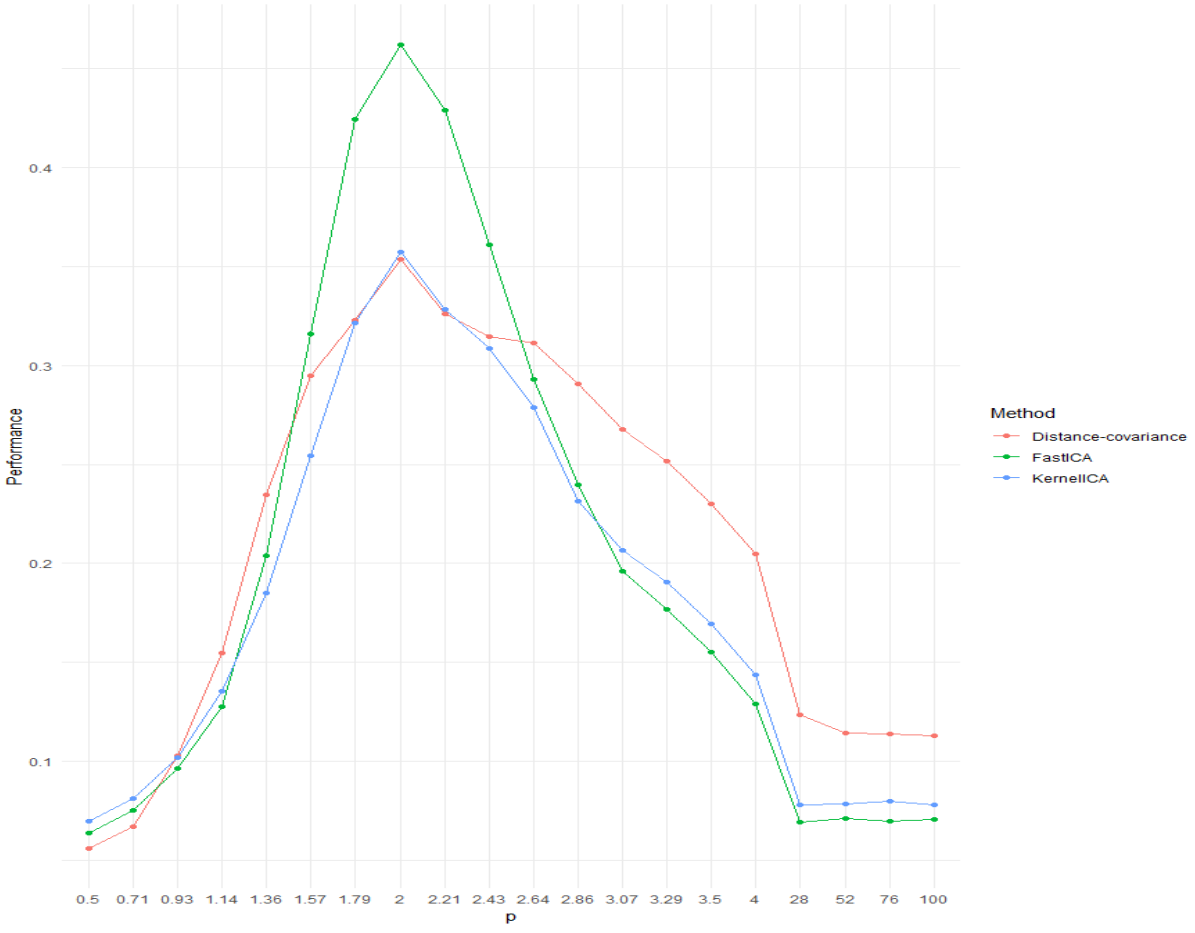

Figure 4.2: General assessment with 2 variables

For the two-variable simulation, a distinct peak in MDI occurs at $p = 2$, indicating

normally distributed data. The KernelICA method performs best near $p = 2$, except precisely at $p = 2$, where the Distance-Covariance method excels. FastICA outperforms the others for p values above 3.07 but lags behind KernelICA in the intermediate range of 1.36 to 2.86.

Figure 4.3 shows the results for a model with three variables.

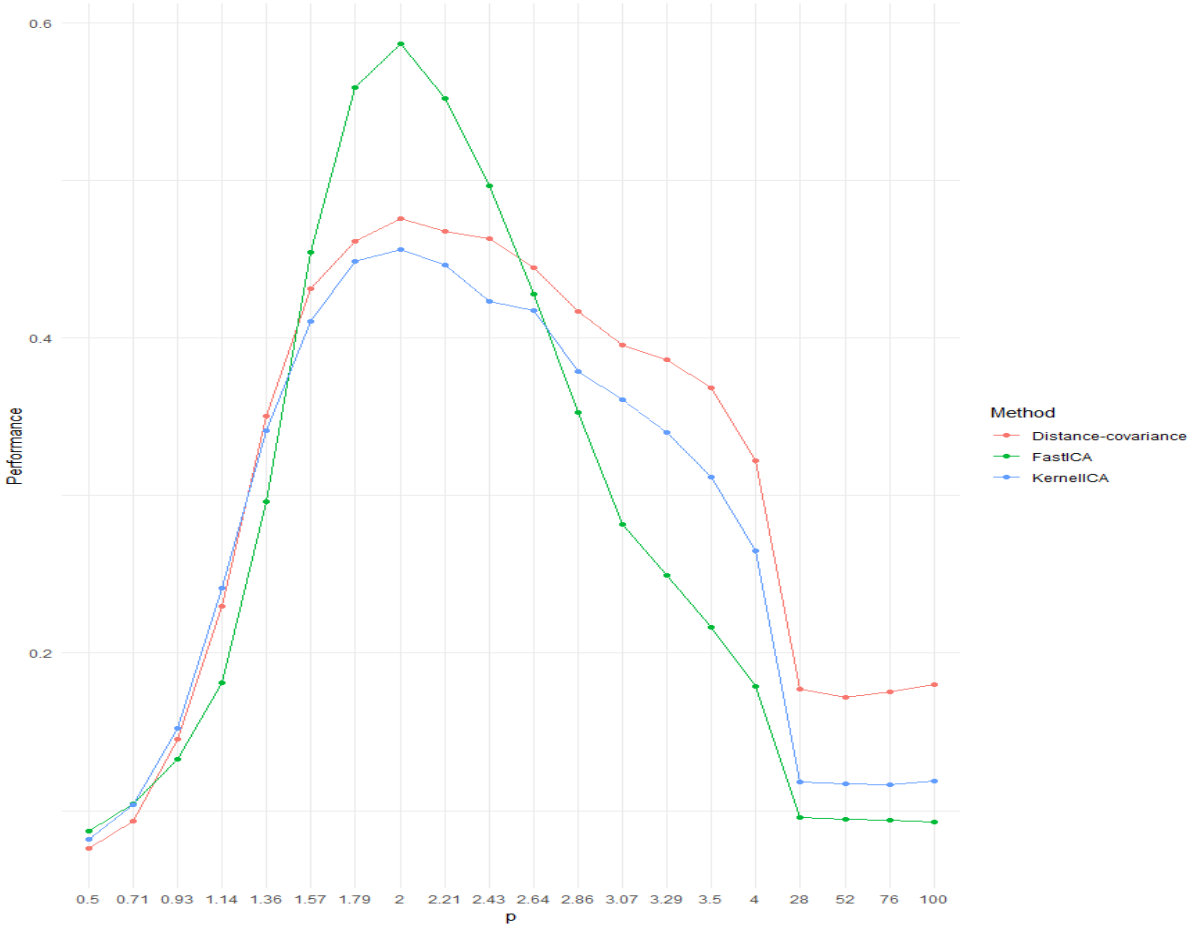

Figure 4.3: General assessment with 3 variable

The 3-variable simulation also show a performance minimum at $p = 2$. Unlike the 2-variable simulation, the KernelICA method outperforms Distance-Covariance for nearly all p values, except at low values where their performances are almost similar. KernelICA is the best method around $p = 2$. FastICA significantly outperforms KernelICA at high p values, a difference from the 2-variable case.

Figure 4.4 illustrates the results for a model with four variables.

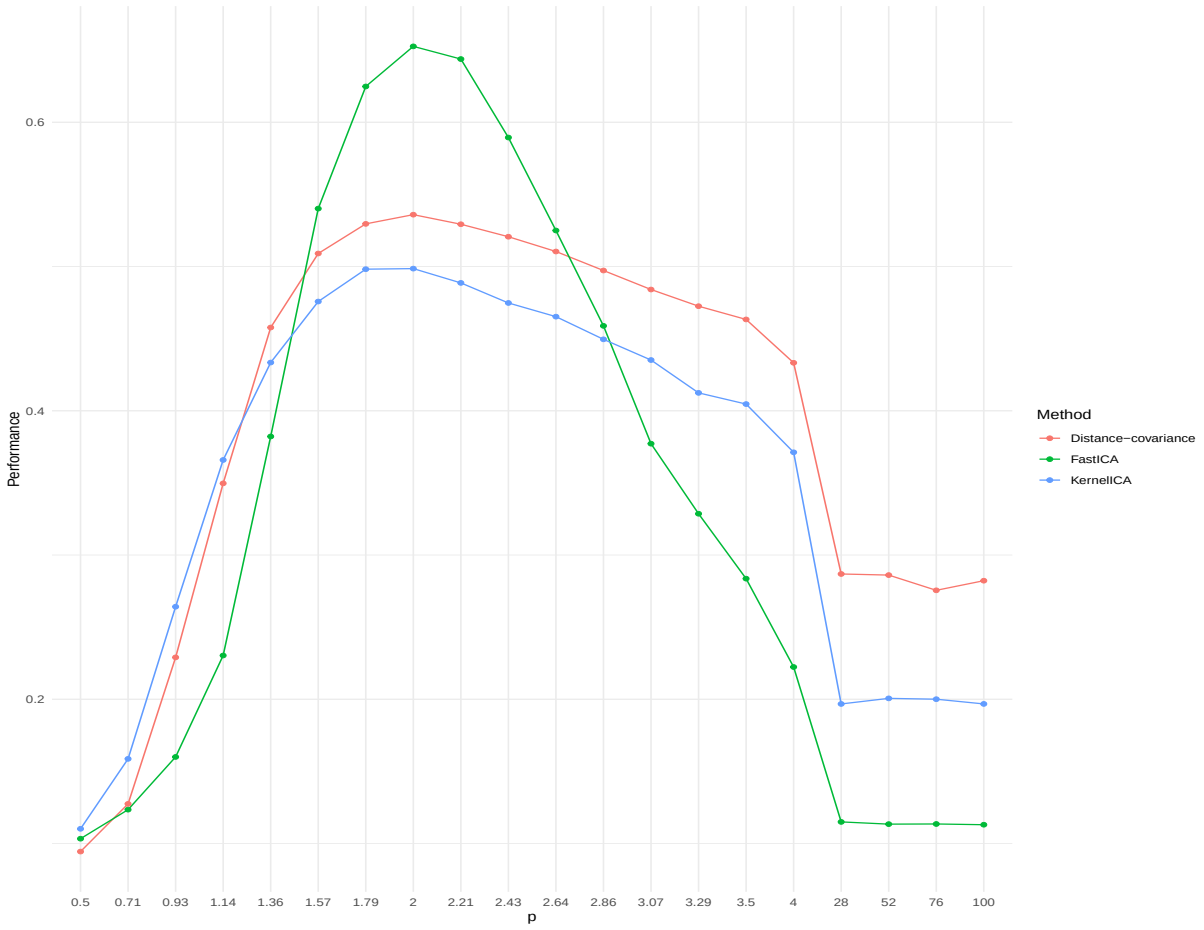

Figure 4.4: General assessment with 4 variables

With four variables, the results are similar to the three-variable scenario: KernelICA performs best for p values near 2, while FastICA shows significantly superior performance at high p values.

The performances of the various methods appear to converge as the number of variables in the model increases. As noted by Moneta and Pallante (2022), FastICA tends to demonstrate better results as the dimensionality increases. KernelICA excels with distributions close to Gaussian but is outperformed by FastICA for more extreme values of p. This observation underscores the importance of conducting normality tests before choosing the identification method.

Chapter 5

Empirical Exercice

In this chapter, we undertake an empirical investigation using real data to analyze economic dynamics. Our focus will be on data from Blanchard and Perotti (2002), which examines the impact of government spending and taxes on U.S. activity in the postwar period. The dynamic relationships and effects of these fiscal policies are studied through the estimation of a SVAR model.

5.1 The Blanchard-Perotti Model

Government spending is defined by Blanchard and Perotti (2002) as the total purchases of goods and services, i.e., government consumption plus government investment. The revenue is defined as the total tax revenues minus transfers (including interest payments). We will call it "taxes" for short to be consistent with the original paper.

In the formulation provided by Kilian and Lütkepohl (2017) on page 236, the Blanchard-Perotti model for $y_t = [tax_t, gov_t, gdp_t]$ is expressed without loss of generality as follows:

$$\begin{pmatrix} 1 & 0 & b_{13,0} \\ 0 & 1 & b_{23,0} \\ b_{31,0} & b_{32,0} & 1 \end{pmatrix} \begin{pmatrix} u_t^{tax} \\ u_t^{gov} \\ u_t^{gdp} \end{pmatrix} = \begin{pmatrix} 1 & c_{12} & 0 \\ c_{21} & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} w_t^{tax} \\ w_t^{gov} \\ w_t^{gdp} \end{pmatrix}$$

such that

$$\begin{cases} u_t^{tax} = -b_{13,0} u_t^{gdp} + c_{12} w_t^{gov} + w_t^{tax}, \\ u_t^{gov} = -b_{23,0} u_t^{gdp} + c_{21} w_t^{tax} + w_t^{gov}, \\ u_t^{gdp} = -b_{31,0} u_t^{tax} - b_{32,0} u_t^{gov} + w_t^{gdp}. \end{cases} \tag{5.1}$$

Here, $y_t = [tax_t, gov_t, gdp_t]'$ is a $n = 3$ dimensional vector in the logarithms of quarterly taxes, primary expenditure, and GDP - all measured in real, per capita terms.

In the system of equations 5.1, the first equation shows that unexpected movements in taxes within a quarter, u_t^{tax}, can come from three sources: the response to unexpected movements in GDP, represented by $-b_{13,0} u_t^{\text{gdp}}$, the response to structural shocks to government spending, represented by $c_{12} w_t^{\text{gov}}$, or structural shocks to taxes, represented by w_t^{tax}. The second equation can be understood similarly for unexpected movements in government spending. The third equation suggests that unexpected movements in output can be caused by unexpected movements in taxes, government spending, or other unexpected shocks, w_t^{gdp}.

Kilian and Lütkepohl (2017) discuss the work of Blanchard and Perotti (2002), who initially provides institutional reasoning for the delay restriction $b_{23,0} = 0$, which prevents immediate feedback from economic activity to government spending within the same quarter. They demonstrate that the within-quarter response of taxes to economic activity, represented by $-b_{13,0}$, can be calculated using external estimates of tax elasticity, yielding $b_{13,0} = -2.08$. The parameters $b_{31,0}$ and $b_{32,0}$ remain unrestricted. To address the potential endogeneity between taxes and spending, either $c_{21} = 0$ or $c_{12} = 0$ is imposed. In the latter scenario, for example, the following result is obtained.

$$\begin{cases} u_t^{tax} = 2.08 u_t^{gdp} + w_t^{tax}, \\ u_t^{gov} = c_{21} w_t^{tax} + w_t^{gov}, \\ u_t^{gdp} = -b_{31,0} u_t^{tax} - b_{32,0} u_t^{gov} + w_t^{gdp}. \end{cases}$$

This system has enough restrictions and can be estimated numerically.

5.2 Model Estimation

We will estimate a SVAR model for the Blanchard-Perotti data using FastICA, KernelICA, and Distance Covariance methods. This empirical exercise aims to compare in practice the different ICA methods presented in this thesis.

In the estimation of our model, we obtained the following mixing matrices for the residuals. These matrix elements have been multiplied by 100 and rounded to facilitate interpretation.

	$B^{-1}_{0,\text{FastICA}}$			$B^{-1}_{0,\text{KernelICA}}$			$B^{-1}_{0,\text{DC}}$		
	$w_{t,1}$	$w_{t,2}$	$w_{t,3}$	$w_{t,1}$	$w_{t,2}$	$w_{t,3}$	$w_{t,1}$	$w_{t,2}$	$w_{t,3}$
G_t	0.67	-0.37	-0.15	0.73	-0.26	0.08	0.72	-0.19	-0.24
Tax_t	1.09	1.41	0.68	0.62	1.59	0.86	0.78	1.61	0.65
GDP_t	0.22	-0.13	0.59	0.05	-0.07	0.64	0.31	-0.09	0.56

Table 5.1: Comparison of mixing matrices using FastICA, KernelICA, and Direct Covariance Methods

These matrices capture the contemporaneous relationships between the model variables, enabling the recovery of the system's structural innovations. As discussed in Chapter 3, the identification of independent components is determined up to a permutation of the sources.

Following the approach of Moneta and Pallante (2022), we select the permutation matrix according to the LiNGAM criterion (Shimizu et al., 2006). The LiNGAM criterion explores the space of all possible permutation matrices to find the matrix $\tilde{P}$ that minimizes a cost function. This function specifically penalizes small absolute values in the main diagonal of the permuted mixing matrix. Based on this criterion, we identify two independent shocks that primarily influence taxes and government spending, and a third shock that significantly impacts both output and taxes. This configuration aligns with the findings of Blanchard and Perotti. Consequently, similar to the labeling in Blanchard and Perotti (2002), we can designate the first (w_1) and second (w_2) shocks as a spending shock and a tax shock, respectively.

The main difference between the mixing matrix estimated using KernelICA and the other two is the entry representing the relationship between the spending and the shock

w_3, which is positive whereas this entry is negative in the other matrices. This observation can be balanced by the fact that the magnitude of this entry is small in absolute value for the three matrices.

Another observation is that the entry representing the relationship between the taxes and w_1 is greater with the FastICA method than with the other two methods.

The next step is to include Cholesky and Spectral decomposition to enrich our analysis and to compare them to the ICA methods.

5.3 Structural Impulse Response Function Analysis

As discussed previously, we can leverage impulse response functions to observe the relationships between variables. In this empirical exercise, we will examine how the impulse response functions estimated by KernelICA compare to other methods and assess their consistency with economic theory.

We begin by studying the impulse response functions of a tax shock and employ a bootstrapping procedure to compute confidence intervals for FastICA, Distance-Covariance, and KernelICA. The upper and lower dashed lines represent the 84th and 16th percent quantiles, respectively, of the bootstrap estimates, following the approach of Moneta and Pallante (2022). The mixing matrices have been normalized so that each structural innovation has a unit contemporaneous impact on the logarithm of the corresponding variable.

The exact estimates of the impulse response function are in Appendix A.1, A.2 and A.3.

Figure 5.1 illustrates the normalized impulse response functions to a tax shock, with confidence intervals obtained through bootstrapping.

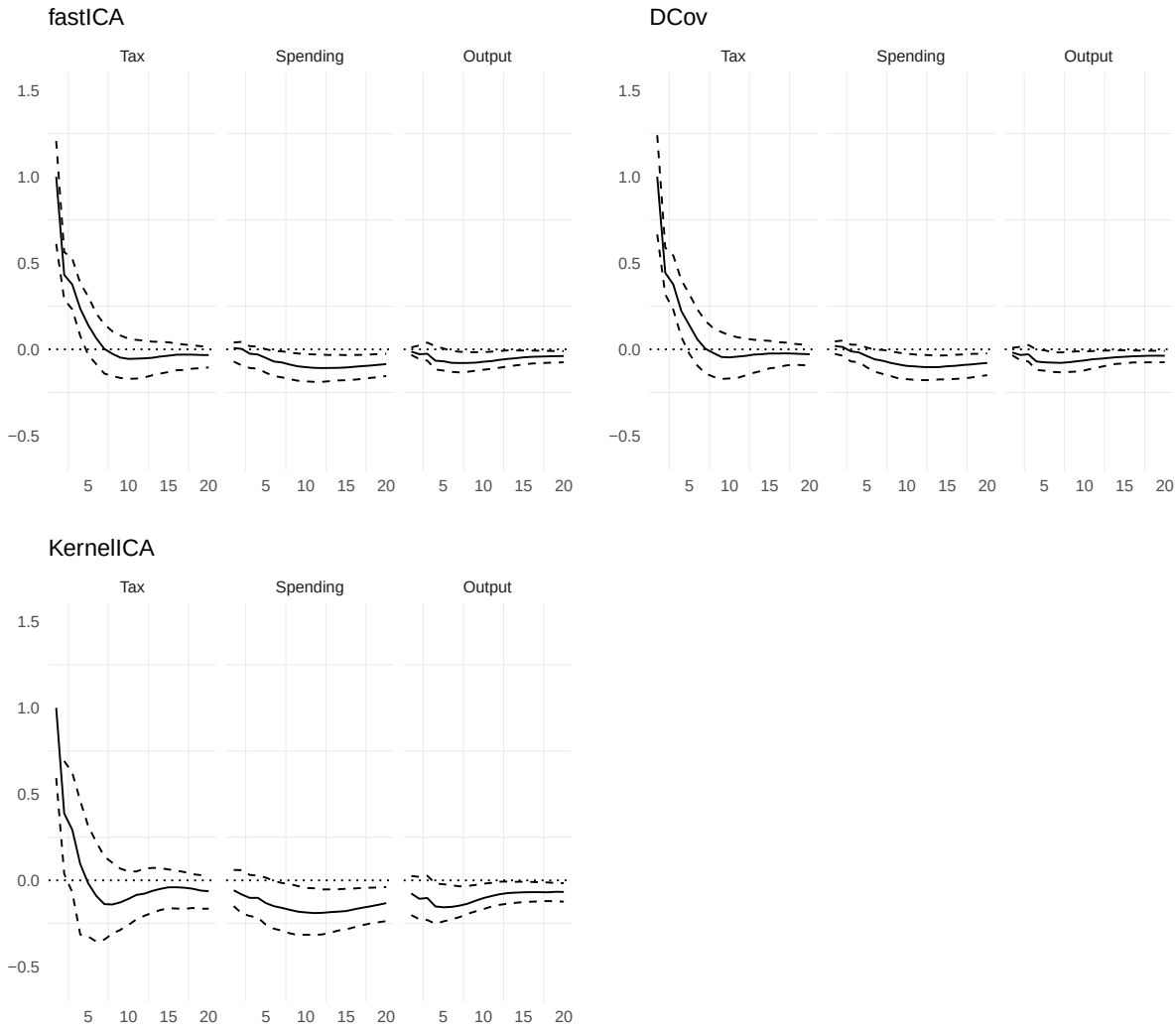

Figure 5.1: Normalized impulse response functions to a tax shock with bootstrap

FastICA, Distance-covariance, and KernelICA exhibit similar trends in the impulse response functions. The response of tax to a tax shock gradually declines to slightly below zero after 10 quarters. All methods indicate that tax shocks have a negative impact on output and that this effect decreases over time. Moreover, each method demonstrates a negative response of government spending to a tax increase, initially decreasing to a minimum between the tenth and fifteenth quarter before returning to zero. These trends align with the findings of Blanchard and Perotti (2002).

We replicate this procedure to analyze the impulse response functions to a spending shock on Figure 5.2.

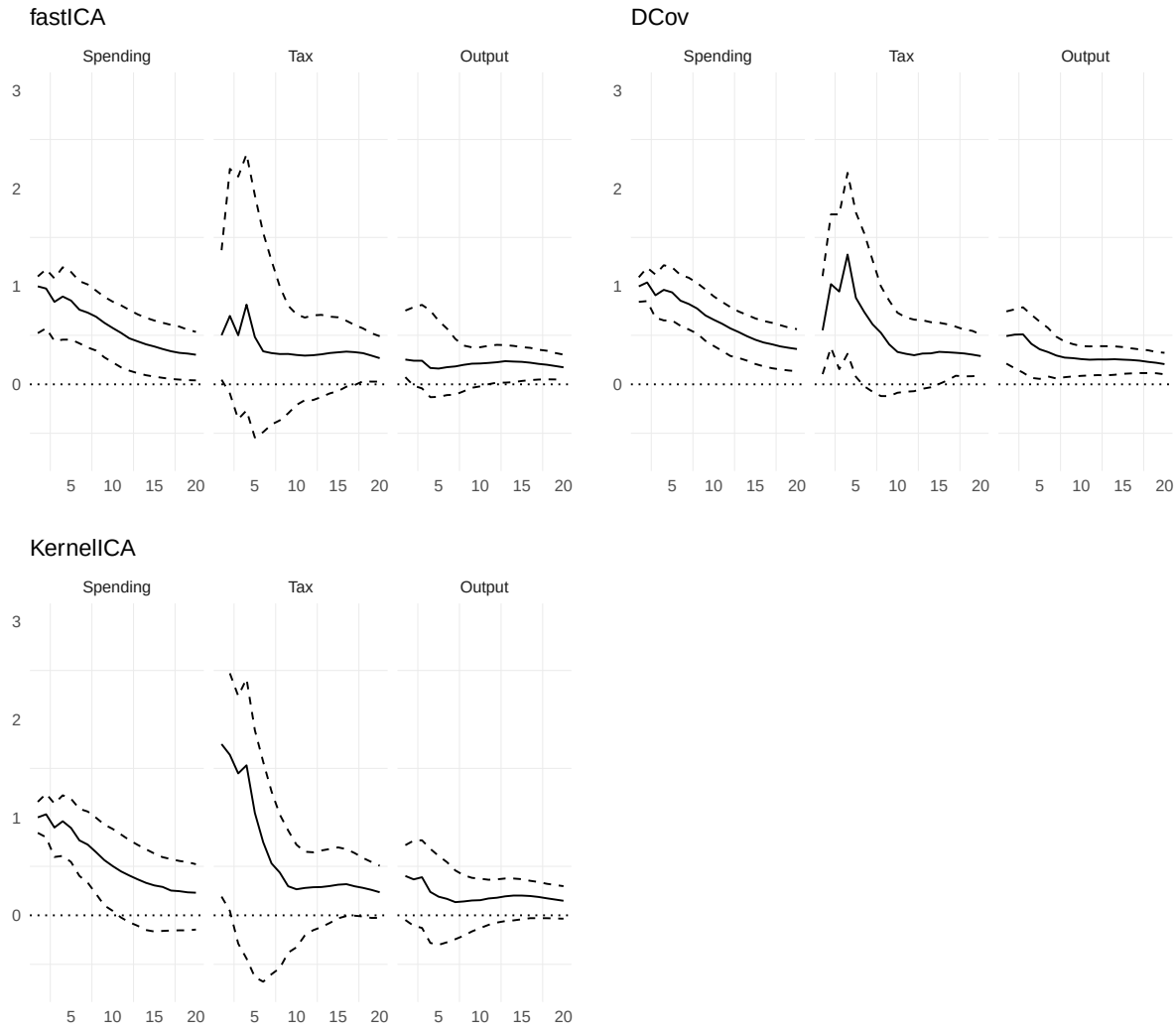

Figure 5.2: Normalized impulse response functions to a spending shock with bootstrap

The outcomes are consistent with Blanchard and Perotti (2002), where a positive government spending shock has a positive effect on output, whereas a positive tax shock negatively affects output. However, methodological differences become more apparent. The tax response to a spending shock in the first periods is more pronounced with KernelICA.

We now present a comparative analysis of the impulse response functions, incorporating both spectral decomposition and Cholesky decomposition techniques. This comparison allows us to observe the distinctive effects and patterns revealed by each method under consideration.

Figure 5.3 displays the different IRFs, the upper row shows responses to a spending

shock, while the lower row shows responses to a tax shock.

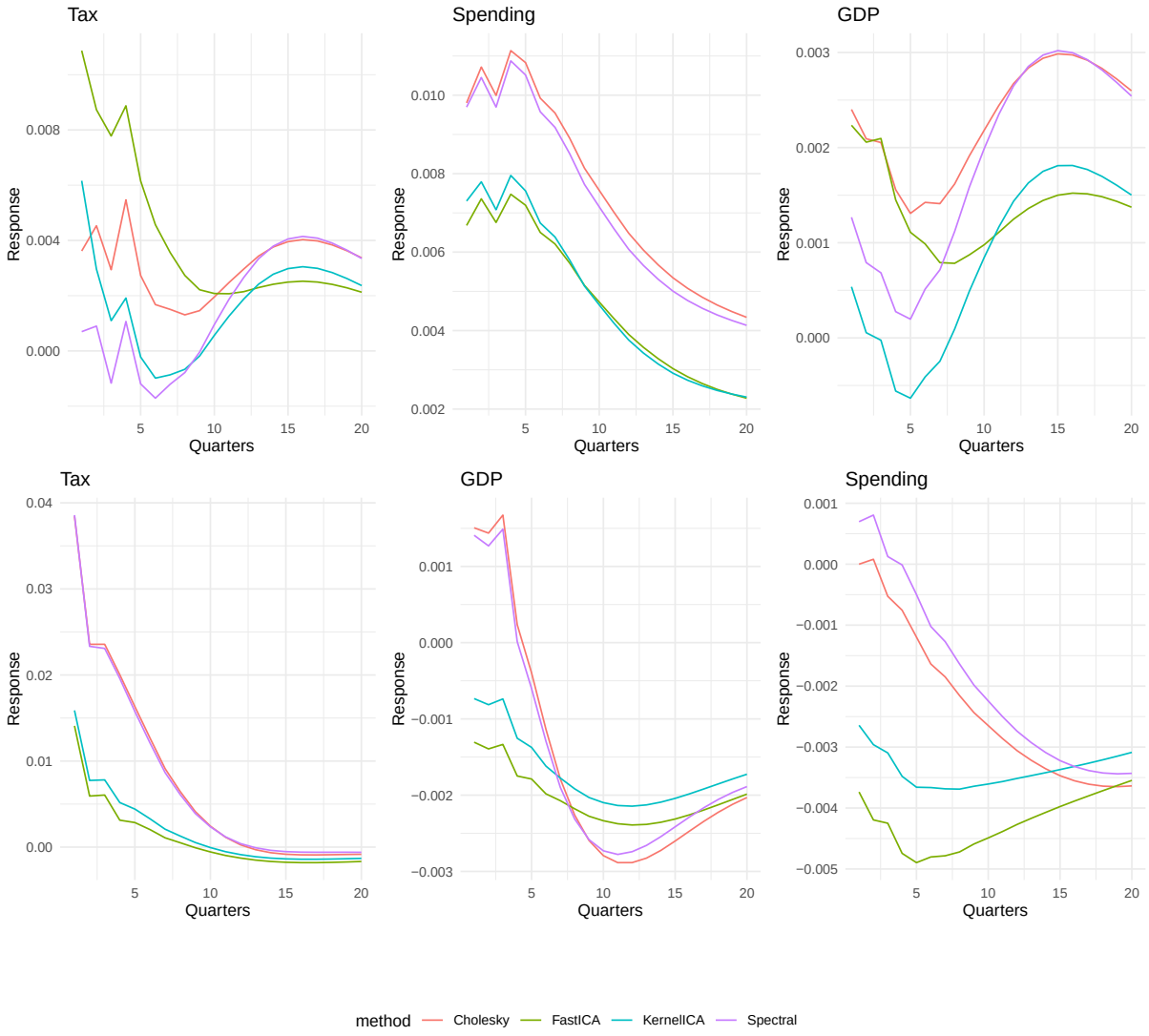

Figure 5.3: Impulse response functions including Cholesky and Spectral decomposition

Across the methods, there is a general consistency in the patterns observed. However, there are notable differences, particularly in the initial quarters'of the spending response to a tax shock. Specifically, the Cholesky and spectral decompositions suggest a positive effect in the early quarters, contrasting with the negative effect indicated by KernelICA and FastICA. This divergence highlights the variability in methodological outcomes, underscoring the importance of choosing the appropriate decomposition technique for

accurate impulse response analysis.

5.4 Forecast Error Variance Decomposition

In this section, we delve deeper into the analytical frameworks with a specific focus on the Forecast Error Variance Decomposition (FEVD) methodology that was introduced in Section 2.3. Our primary aim here is to dissect and understand the varying contributions of different economic variables to the fluctuations observed in GDP over time.

Figure 5.4 presents the Forecast Error Variance Decomposition (FEVD) of GDP attributed to spending and tax shocks.

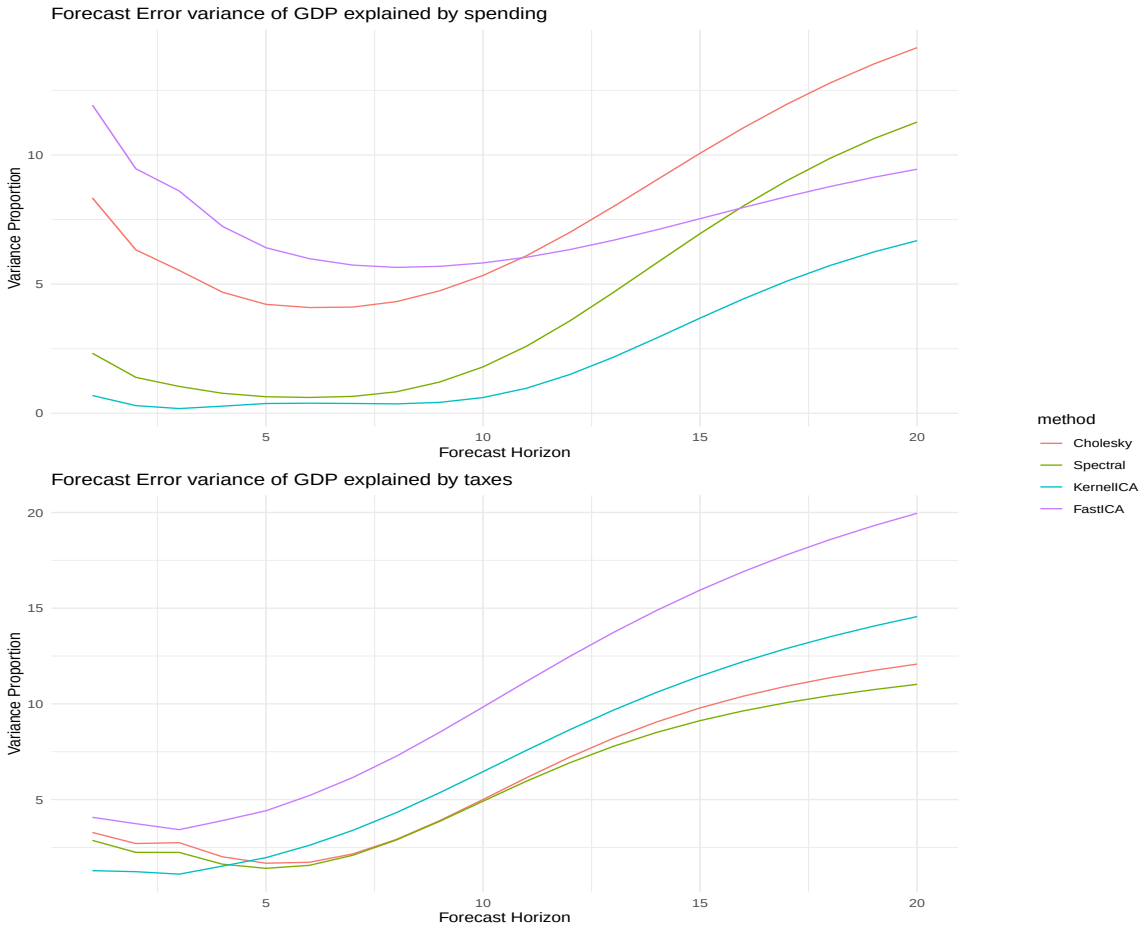

Figure 5.4: Forecast error variance decomposition of GDP for different methods

The exact estimates of the forecast error variance decomposition can be found in Appendix A.4.

The Forecast Error Variance Decomposition analysis reveals how different methods distribute the variance of GDP across various economic variables. Specifically, the FastICA method allocates a larger share of GDP variance to government spending and taxes compared to its counterparts. This suggests a perspective where public policy components, such as spending and taxation, are seen as primary influencers of GDP fluctuations.

Conversely, KernelICA initially ascribes the bulk of GDP variance to the GDP itself, indicating a strong self-referential dynamic in the early quarters. However, this self-attribution diminishes more rapidly over time, suggesting a quicker diversification in the factors influencing GDP variance.

Spectral decomposition shares a similar pattern with KernelICA, with the notable difference being a slight adjustment in the variance attributed to spending and taxes. It allocates a bit more of the variance to government spending and slightly less to taxation, pointing to a nuanced view on the relative impact of these fiscal policies on GDP.

Cholesky decomposition, meanwhile, assumes a balanced approach but gradually shifts more variance attribution to government spending over the long haul. This indicates an evolving understanding of fiscal policy's role in driving GDP changes over time, particularly government spending.

Interestingly, we can note that the results of KernelICA and FastICA are not the most similar ones, even though they rely on the same hypothesis of independence.

Chapter 6

Conclusion

In this thesis, we investigated the application of Kernel Independent Component Analysis for identifying structural innovations in Structural Vector Autoregressive models, evaluating its performance by comparing it to other independent component analysis methods and other alternative identification strategies. KernelICA uses the concepts of reproducing kernel Hilbert spaces and canonical correlation to effectively maximize non-Gaussianity and minimize mutual information among components. KernelICA was tested using a series of Monte Carlo simulations in comparison with FastICA and Distance Covariance methods. KernelICA consistently outperformed the other methods in close-to-Gaussian scenarios.

We illustrated KernelICA's usefulness by using it to the Blanchard-Perotti dataset of postwar US government expenditure, taxes, and GDP. This showed that KernelICA is capable of extracting significant structural shocks and yielding comprehensible impulse response functions. The results aligned with what was found using conventional techniques.

Even if KernelICA produced encouraging findings, there are a few issues that need to be looked into further. The choice of kernel parameters and regularization coefficients significantly impacts KernelICA's performance; hence, future work may concentrate on creating adaptive or data-driven techniques for parameter selection to improve resilience.

KernelICA presents a robust and innovative solution for identifying structural innovations in SVAR models, offering substantial improvements over traditional methods. Its ability to handle a wide range of data characteristics and provide clear, interpretable results makes it a valuable tool for both researchers and policymakers. Continuing to refine and extend this approach holds the potential to unlock deeper insights into the com-

plex dynamics of multivariate time series data, ultimately contributing to more informed decision-making across various fields.

Appendix A

Appendix

Table A.1: impulse response functions for FastICA

Horizon	$w[gov] \rightarrow gov$	$w[tax] \rightarrow gov$	$w[gdp] \rightarrow gov$	$w[gov] \rightarrow tax$	$w[tax] \rightarrow tax$	$w[gdp] \rightarrow tax$	$w[gov] \rightarrow gdp$	$w[tax] \rightarrow gdp$	$w[gdp] \rightarrow gdp$
1	0.0067	-0.0037	-0.0015	0.0109	0.0141	0.0068	0.0022	-0.0013	0.0059
2	0.0074	-0.0042	-0.0008	0.0087	0.0059	0.0148	0.0021	-0.0014	0.0070
3	0.0068	-0.0042	-0.0003	0.0078	0.0060	0.0180	0.0021	-0.0013	0.0074
4	0.0075	-0.0047	-0.0007	0.0089	0.0031	0.0202	0.0015	-0.0017	0.0073
5	0.0072	-0.0049	-0.0002	0.0062	0.0028	0.0189	0.0011	-0.0018	0.0065
6	0.0065	-0.0048	0.0003	0.0046	0.0020	0.0168	0.0010	-0.0020	0.0055
7	0.0062	-0.0048	0.0006	0.0036	0.0011	0.0137	0.0008	-0.0021	0.0044
8	0.0057	-0.0047	0.0010	0.0027	0.0005	0.0107	0.0008	-0.0022	0.0033
9	0.0052	-0.0046	0.0013	0.0022	-0.0001	0.0077	0.0009	-0.0023	0.0022
10	0.0047	-0.0045	0.0016	0.0021	-0.0006	0.0050	0.0010	-0.0023	0.0014
11	0.0043	-0.0044	0.0017	0.0021	-0.0010	0.0028	0.0011	-0.0024	0.0008
12	0.0039	-0.0043	0.0018	0.0021	-0.0013	0.0011	0.0012	-0.0024	0.0003
13	0.0036	-0.0042	0.0019	0.0023	-0.0015	-0.0001	0.0014	-0.0024	0.0000
14	0.0033	-0.0041	0.0018	0.0024	-0.0017	-0.0009	0.0014	-0.0024	-0.0001
15	0.0030	-0.0040	0.0018	0.0025	-0.0018	-0.0012	0.0015	-0.0023	-0.0002
16	0.0028	-0.0039	0.0017	0.0025	-0.0018	-0.0014	0.0015	-0.0023	-0.0002
17	0.0026	-0.0038	0.0016	0.0025	-0.0018	-0.0013	0.0015	-0.0022	-0.0001
18	0.0025	-0.0037	0.0014	0.0024	-0.0018	-0.0010	0.0015	-0.0021	0.0000
19	0.0024	-0.0036	0.0013	0.0023	-0.0017	-0.0007	0.0014	-0.0021	0.0002
20	0.0023	-0.0035	0.0012	0.0021	-0.0017	-0.0004	0.0014	-0.0020	0.0003

Table A.2: Impulse response functions for KernelICA

Horizon	$w[gov] \rightarrow gov$	$w[tax] \rightarrow gov$	$w[gdp] \rightarrow gov$	$w[gov] \rightarrow tax$	$w[tax] \rightarrow tax$	$w[gdp] \rightarrow tax$	$w[gov] \rightarrow gdp$	$w[tax] \rightarrow gdp$	$w[gdp] \rightarrow gdp$
1	0.0073	-0.0026	0.0008	0.0062	0.0159	0.0086	0.0005	-0.0007	0.0064
2	0.0078	-0.0030	0.0017	0.0030	0.0077	0.0162	0.0001	-0.0008	0.0074
3	0.0071	-0.0031	0.0020	0.0011	0.0078	0.0190	-0.0000	-0.0007	0.0078
4	0.0080	-0.0035	0.0019	0.0019	0.0052	0.0216	-0.0006	-0.0013	0.0076
5	0.0076	-0.0037	0.0023	-0.0002	0.0044	0.0196	-0.0006	-0.0014	0.0067
6	0.0067	-0.0037	0.0026	-0.0010	0.0033	0.0172	-0.0004	-0.0016	0.0057
7	0.0064	-0.0037	0.0027	-0.0009	0.0021	0.0140	-0.0002	-0.0018	0.0046
8	0.0058	-0.0037	0.0030	-0.0007	0.0013	0.0109	0.0001	-0.0019	0.0035
9	0.0051	-0.0036	0.0031	-0.0002	0.0005	0.0080	0.0005	-0.0020	0.0026
10	0.0047	-0.0036	0.0032	0.0006	-0.0001	0.0054	0.0008	-0.0021	0.0018
11	0.0042	-0.0036	0.0032	0.0013	-0.0005	0.0033	0.0012	-0.0021	0.0012
12	0.0038	-0.0035	0.0032	0.0019	-0.0009	0.0018	0.0014	-0.0021	0.0008
13	0.0034	-0.0035	0.0031	0.0024	-0.0011	0.0007	0.0016	-0.0021	0.0006
14	0.0031	-0.0034	0.0030	0.0028	-0.0013	0.0000	0.0018	-0.0021	0.0005
15	0.0029	-0.0034	0.0029	0.0030	-0.0014	-0.0003	0.0018	-0.0020	0.0004
16	0.0027	-0.0033	0.0027	0.0030	-0.0014	-0.0004	0.0018	-0.0020	0.0005
17	0.0026	-0.0033	0.0026	0.0030	-0.0014	-0.0003	0.0018	-0.0019	0.0005
18	0.0025	-0.0032	0.0024	0.0028	-0.0014	-0.0001	0.0017	-0.0019	0.0006
19	0.0024	-0.0031	0.0022	0.0026	-0.0014	0.0001	0.0016	-0.0018	0.0007
20	0.0023	-0.0031	0.0021	0.0024	-0.0013	0.0004	0.0015	-0.0017	0.0008

Table A.3: Impulse response functions for DCov

Horizon	$w[gov] \to gov$	$w[tax] \to gov$	$w[gdp] \to gov$	$w[gov] \to tax$	$w[tax] \to tax$	$w[gdp] \to tax$	$w[gov] \to gdp$	$w[tax] \to gdp$	$w[gdp] \to gdp$
1	0.0072	-0.0019	-0.0024	0.0078	0.0161	0.0065	0.0031	-0.0009	0.0056
2	0.0080	-0.0022	-0.0017	0.0086	0.0075	0.0142	0.0031	-0.0010	0.0067
3	0.0075	-0.0025	-0.0011	0.0080	0.0073	0.0175	0.0031	-0.0010	0.0071
4	0.0083	-0.0028	-0.0017	0.0100	0.0047	0.0194	0.0026	-0.0015	0.0071
5	0.0081	-0.0030	-0.0011	0.0073	0.0038	0.0183	0.0022	-0.0016	0.0063
6	0.0075	-0.0031	-0.0006	0.0057	0.0026	0.0164	0.0020	-0.0018	0.0053
7	0.0072	-0.0031	-0.0003	0.0046	0.0015	0.0133	0.0017	-0.0019	0.0041
8	0.0068	-0.0032	0.0002	0.0036	0.0009	0.0104	0.0016	-0.0020	0.0030
9	0.0062	-0.0032	0.0006	0.0030	0.0003	0.0074	0.0016	-0.0021	0.0020
10	0.0058	-0.0032	0.0009	0.0027	-0.0002	0.0047	0.0017	-0.0021	0.0012
11	0.0054	-0.0032	0.0011	0.0025	-0.0005	0.0025	0.0017	-0.0021	0.0005
12	0.0050	-0.0032	0.0012	0.0025	-0.0007	0.0008	0.0018	-0.0020	0.0001
13	0.0046	-0.0032	0.0013	0.0026	-0.0009	-0.0004	0.0019	-0.0020	-0.0002
14	0.0043	-0.0032	0.0013	0.0026	-0.0010	-0.0012	0.0019	-0.0019	-0.0004
15	0.0041	-0.0032	0.0013	0.0027	-0.0011	-0.0016	0.0020	-0.0019	-0.0005
16	0.0038	-0.0031	0.0012	0.0027	-0.0011	-0.0017	0.0020	-0.0018	-0.0004
17	0.0036	-0.0031	0.0011	0.0027	-0.0011	-0.0016	0.0020	-0.0018	-0.0003
18	0.0035	-0.0030	0.0010	0.0026	-0.0011	-0.0013	0.0019	-0.0017	-0.0002
19	0.0033	-0.0030	0.0009	0.0025	-0.0011	-0.0010	0.0019	-0.0016	-0.0001
20	0.0032	-0.0029	0.0008	0.0024	-0.0011	-0.0007	0.0018	-0.0016	0.0001

Table A.4: Values of forecast error variance decomposition

Horizon	FastICA			KernelICA			DCov		
	gov	tax	gdp	gov	tax	gdp	gov	tax	gdp
1	0.1193	0.0407	0.8399	0.0069	0.0129	0.9803	0.2295	0.0185	0.7520
2	0.0947	0.0374	0.8679	0.0030	0.0123	0.9847	0.1950	0.0189	0.7860
3	0.0860	0.0343	0.8797	0.0018	0.0110	0.9871	0.1819	0.0177	0.8004
4	0.0723	0.0390	0.8886	0.0028	0.0152	0.9820	0.1635	0.0238	0.8127
5	0.0641	0.0442	0.8918	0.0038	0.0197	0.9765	0.1526	0.0297	0.8177
6	0.0599	0.0521	0.8880	0.0039	0.0261	0.9700	0.1483	0.0374	0.8144
7	0.0574	0.0615	0.8811	0.0038	0.0339	0.9623	0.1465	0.0461	0.8074
8	0.0565	0.0726	0.8709	0.0037	0.0431	0.9532	0.1474	0.0559	0.7968
9	0.0569	0.0851	0.8580	0.0042	0.0536	0.9422	0.1503	0.0661	0.7835
10	0.0582	0.0983	0.8434	0.0061	0.0646	0.9293	0.1544	0.0764	0.7691
11	0.0604	0.1117	0.8279	0.0097	0.0757	0.9146	0.1594	0.0864	0.7542
12	0.0634	0.1248	0.8118	0.0150	0.0865	0.8984	0.1650	0.0956	0.7394
13	0.0670	0.1373	0.7958	0.0217	0.0967	0.8816	0.1711	0.1040	0.7249
14	0.0710	0.1488	0.7802	0.0291	0.1060	0.8648	0.1775	0.1115	0.7111
15	0.0753	0.1594	0.7652	0.0369	0.1145	0.8486	0.1840	0.1182	0.6979
16	0.0797	0.1691	0.7512	0.0443	0.1221	0.8336	0.1904	0.1241	0.6855
17	0.0839	0.1779	0.7382	0.0512	0.1290	0.8199	0.1967	0.1295	0.6739
18	0.0878	0.1859	0.7263	0.0572	0.1351	0.8077	0.2026	0.1343	0.6631
19	0.0913	0.1931	0.7156	0.0625	0.1406	0.7969	0.2081	0.1387	0.6533
20	0.0945	0.1996	0.7059	0.0668	0.1456	0.7876	0.2130	0.1427	0.6443

Bibliography

Amari, S., Cichocki, A., & Yang, H. H. (1996). A new learning algorithm for blind signal separation. In D. S. Touretzky, M. C. Mozer, & M. E. Hasselmo (Eds.), *Advances in neural information processing systems* (pp. 757–763). MIT Press.

Bach, F., & Jordan, M. (2003). Kernel independent component analysis. *Journal of Machine Learning Research*, *3*, 1–48. https://doi.org/10.1162/153244303768966085

Bickel, P. J., Klaassen, C. A. J., Ritov, Y., & Wellner, J. A. (1998). *Efficient and adaptive estimation for semiparametric models*. Springer-Verlag.

Blanchard, O., & Perotti, R. (2002). An empirical characterization of the dynamic effects of changes in government spending and taxes on output. *The Quarterly Journal of Economics*, *117*(4), 1329–1368.

Comon, P. (1994). Independent component analysis, a new concept? *Signal Processing*, *36*, 287–314. https://doi.org/10.1016/0165-1684(94)90029-9

Cover, T. M., & Thomas, J. A. (1991). *Elements of information theory*. John Wiley & Sons.

Eriksson, J., & Koivunen, V. (2004). Identifiability, separability, and uniqueness of linear ica models. *IEEE Transactions on Signal Processing*, *52*(3), 537–548.

Girosi, F., Jones, M., & Poggio, T. (1995). Regularization theory and neural networks architectures. *Neural Computation*, *7*(2), 219–269.

Gottschalk, J. (2001). *An introduction into the svar methodology: Identification, interpretation and limitations of svar models* (Kiel Working Paper No. 1072). Kiel Institute for the World Economy. Kiel, Germany.

Hastie, T., & Tibshirani, R. (2010). *Prodenica: Product density estimation for ica using tilted gaussian density estimates* [R package version, 1].

Hotelling, H. (1936). Relations between two sets of variates. *Biometrika*, *28*, 321–377.

Hyvärinen, A. (1999). Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks*, *10*(3), 626–634.

Hyvärinen, A., Karhunen, J., & Oja, E. (2001). *Independent component analysis*. John Wiley & Sons.

Hyvärinen, A., & Oja, E. (2000). Independent component analysis: Algorithms and applications. *Neural Networks*, *13*(4-5), 411–430.

Ilmonen, P., Nordhausen, K., Oja, H., & Ollila, E. (2010). A new performance index for ica: Properties, computation and asymptotic analysis. In V. Vigneron, V. Zarzoso, E. Moreau, R. Gribonval, & E. Vincent (Eds.), *Latent variable analysis and signal separation* (pp. 229–236). Springer Berlin Heidelberg.

Izenman, A. J. (2008). *Modern multivariate statistical techniques: Regression, classification, and manifold learning* (Vol. 1). Springer.

Jutten, C., & Herault, J. (1991). Blind separation of sources, part i: An adaptive algorithm based on neuromimetic architecture. *Signal Processing*, *24*(1), 1–10.

Kalke, S., & Richter, W.-D. (2013). Simulation of the p-generalized gaussian distribution. *Journal of Statistical Computation and Simulation*, *83*(4), 641–667. https://doi.org/10.1080/00949655.2011.631187

Kilian, L., & Lütkepohl, H. (2017). *Structural vector autoregressive analysis*. Cambridge University Press. https://doi.org/10.1017/9781108164818

Kilian, L. (2011). *Structural vector autoregressions* (CEPR Discussion Papers No. 8515). C.E.P.R. Discussion Papers.

Kilian, L., & Park, C. (2009). The impact of oil price shocks on the u.s. stock market. *International Economic Review*, *50*(4), 1267–1287.

Leurgans, S., Moyeed, R., & Silverman, B. W. (1993). Canonical correlation analysis when the data are curves. *Journal of the Royal Statistical Society: Series B (Methodological)*, *55*(3), 725–740.

Lütkepohl, H. (2005). *New introduction to multiple time series analysis*. Springer.

Matteson, D. S., & Tsay, R. S. (2017). Independent component analysis via distance covariance. *Journal of the American Statistical Association*, *111*(514), 623–637. https://doi.org/10.1080/01621459.2016.1150851

Moneta, A., & Pallante, G. (2022). Identification of structural var models via independent component analysis: A performance evaluation study. *Journal of Economic Dynamics and Control*, *144*, 104530. https://doi.org/10.1016/j.jedc.2022.104530

Rigobon, R. (2003). Identification through heteroskedasticity. *Review of Economics and Statistics*, *85*(4), 777–792.

Saitoh, S. (1988). *Theory of reproducing kernels and its applications*. Longman Scientific & Technical.

Schölkopf, B., Smola, A. J., & Müller, K.-R. (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, *10*(3), 1299–1319.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*(3), 379–423.

Shimizu, S., Hoyer, P. O., Hyvärinen, A., & Kerminen, A. (2006). A linear non-gaussian acyclic model for causal discovery (M. Jordan, Ed.). *Journal of Machine Learning Research*, *7*, 2003–2030. http://www.jmlr.org/papers/volume7/shimizu06a/shimizu06a.pdf

Sims, C. A. (1980). Macroeconomics and reality. *Econometrica*, *48*(1), 1–48.

Stone, J. V. (2004). *Independent component analysis: A tutorial introduction*. The MIT Press.

Székely, G. J., Rizzo, M. L., & Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, *35*(6), 2769–2794. https://doi.org/10.1214/009053607000000505

UNIVERSITÉ CATHOLIQUE DE LOUVAIN
Faculté des sciences
Place des Sciences, 2 bte L6.06.01, 1348 Louvain-la-Neuve, Belgique | www.uclouvain.be/sc