Doctoral Program in Economics and Management

Identification and Inference in High-Dimensional Econometric Models

A Thesis Presented in Partial Fulfillment
of the Requirements for the PhD Degree in
Economics and Management Sciences by

Mathieu Sauvenier

The 21th of April 2023, at 2pm
Université Catholique de Louvain, Louvain-la-Neuve,
Faculté des sciences économiques, sociales, politiques et de la communication.

Jury
Prof. Sébastien Van Bellegem (UCLouvain), Supervisor
Prof. Francesca Monti (UCLouvain), Chairman of the Jury
Prof. Christian Hafner (UCLouvain), Secretary of the Jury
Prof. Anna Simoni (ENSAE)
Prof. Germain Van Bever (UNamur)

UNIVERSITE CATHOLIQUE DE LOUVAIN

Acknowledgements

I would like to express my deep gratitude to Professor Sébastien Van Bellegem for guiding me through this thesis, for his willingness to teach me cool mathematics, and for the energy he put into advising my work. I also thank the readers of this thesis, and in particular the members of the jury for their comments and remarks. They helped me to improve the presentation of this work. I would like to give special thanks to Professor Ernesto San Martin for sharing his insights on the foundations of statistics. I would also like to thank all the academic and administrative staff, as well as all the workers of the university.

If, to paraphrase a philosopher, humans are often unaware of the causes that determine them, I am glad to be aware of those that some of my friends have had on my intellectual journey. I would like to thank them in chronological order. Thank you Aline, you opened the way to research. Thank you Jean for having introduced me to philosophy. Thank you Valérian, you show me the power (and the limits) of empiricism. Thank you, K, you reminded me that reading is the key, and that, sadly, at this point in history, reason is not universal, but divided. Thank you, Benjamin, for many things, but here, mainly, for making me understand the fundamental importance of history. Thank you Yannick for always discussing mathematics and physics with me and especially for helping me to humbly realize the beauty of geometry. Thank you Jean-Christophe for teaching me history. Thank you M for introducing me to literature, it helps a lot to struggle against the bad side of abstractions. Thank you to all the workers for producing all the goods that have allowed me to live a comfortable life, study, and do research, we owe our existence to your work.

I would like to extend special thanks to Simon and Suzanne, as well as express my gratitude to all those with whom I have shared a home over the past years.

Je tiens également à exprimer ma profonde gratitude à l'égard de ma famille : Sophie Cavigioli, Bernard Sauvenier, Simone Heine, Lola et Jeanne.

Et bien sûr, je veux remercier ma chère Leïla, de partager cette vie, pour toute l'énergie que tu as déployées pour m'aider à terminer cette thèse, et, pour tant d'autres choses encore.

Merci à tous, et courage, tout indique que les temps qui viennent ne seront pas facile.

Contents

Contents 3

Acronyms 7

List of Figures 8

List of Tables 10

List of Main Symbols 11

1 Introduction 15
- 1.1 General introduction . . . 15
- 1.2 Topics of the thesis . . . 16
  - 1.2.1 High-dimensional cross-sections and high-dimensional non-stationary time series . . . 16
  - 1.2.2 The challenges of inference in high dimensions . . . 17
  - 1.2.3 Identification of parameters . . . 19
- 1.3 Contributions and structure of the thesis . . . 21

2 Identification and estimation by GEP 25
- 2.1 Introduction . . . 25
- 2.2 Structural Model . . . 29
- 2.3 Identification of the direction of interest . . . 30
- 2.4 Estimation under sparsity . . . 31
  - 2.4.1 Directional Best Subset Selection . . . 31
  - 2.4.2 Algorithms for Directional Best Subset Selection . . . 33
  - 2.4.3 Consistency of the estimation and the rate of convergence 35
  - 2.4.4 Central Limit Theorem . . . 39
- 2.5 Monte Carlo experiments . . . 42
  - 2.5.1 Implementation . . . 42
  - 2.5.2 Results . . . 44
- 2.6 Conclusion . . . 45

2.7 Appendix . . . 48
2.7.1 Proof of Theorem 1 . . . 48
2.7.2 Proof of Proposition 1 . . . 49
2.7.3 Proof of Proposition 2 . . . 50
2.7.4 Proof of Proposition 4 . . . 50
2.7.5 Proof of Proposition 5 . . . 53
2.7.6 Proof of Proposition 6 . . . 55
2.7.7 Proof of Theorem 3 . . . 57
2.8 Supplementary Materials . . . 59
2.8.1 Proof of Theorems 2 . . . 59
2.8.2 Proof of Proposition 3 . . . 66
2.8.3 Proof of Lemma 1 . . . 69
2.8.4 Proof of Lemma 3 . . . 70
2.8.5 Proof of Lemma 5 . . . 71
2.8.6 Proof of Lemma 6 . . . 71
2.8.7 Proof of Lemma 7 . . . 71

3 Goodness-of-fit test for sparse model 73
3.1 Introduction . . . 73
3.2 Necessary and sufficient conditions for sparsity . . . 75
3.3 Results for inference . . . 76
3.3.1 Multiple-Splitting Projection Test . . . 79
3.4 Experiments and applications . . . 84
3.4.1 Monte Carlo experiment: assessing the power of the test 84
3.4.2 Application to synthetic datasets: comparing variable selections . . . 85
3.4.3 Application: assessing the quality of the selection prior to nowcasting macroeconomic time series . . . 86
3.5 Conclusion . . . 92
3.6 Appendix . . . 93
3.6.1 Preliminary results . . . 93
3.6.2 Proof of Theorem 4 . . . 94
3.6.3 Proof of Proposition 8 . . . 95

4 Multivariate Locally Stationary Wavelet Processes 97
4.1 Introduction . . . 97
4.2 Wavelet Analysis . . . 99
4.2.1 Multiresolution analysis of $L^2(\mathbb{R})$ . . . 100
4.2.2 Construction of wavelets from a multiresolution analysis 100
4.2.3 The discrete wavelet transform . . . 103
4.2.4 Nondecimated discrete wavelets . . . 105
4.3 The Cross-Correlation of Wavelet Functions . . . 106
4.3.1 Linear independence of the CCWF . . . 107
4.4 Multivariate Locally Stationary Wavelet Process . . . 111

4.4.1 Covariance representation . . . 113
4.4.2 Asymptotic identification of the covariance representation 116
4.5 Appendix . . . 118
4.5.1 Proof of Lemma 13 . . . 118
4.5.2 Proof of Lemma 14 . . . 118
4.5.3 Proof of Lemma 15 . . . 120
4.5.4 Proof of Lemma 16 . . . 122
4.5.5 Proof of Proposition 9 . . . 123

5 Inference in the MLSW and application 125
5.1 Estimation of the EWS in the $(J, q)-$MLSW model . . . 126
5.1.1 Raw wavelet periodogram . . . 127
5.1.2 Smoothed wavelet periodogram . . . 128
5.1.3 Smoothed and corrected wavelet periodogram . . . 130
5.1.4 Estimation in High Dimensions . . . 132
5.1.5 Simulations . . . 133
5.2 Application to real data: oil market . . . 134
5.2.1 Motivations . . . 134
5.2.2 The model . . . 138
5.2.3 Results . . . 140
5.2.4 Robustness analysis . . . 144
5.3 Appendix . . . 150
5.3.1 Proof of Theorem 7 . . . 150
5.3.2 Proof of Theorem 8 . . . 151
5.3.3 Proof of Theorem 9 . . . 151
5.3.4 Proof of Theorem 10 . . . 152
5.3.5 Proof of Theorem 11 . . . 160
5.3.6 Proof of Theorem 12 . . . 161

6 From double asymptotics to functional regression 163
6.1 Introduction . . . 163
6.2 From double asymptotic to functional regression . . . 167
6.2.1 A simple theory for approximation of square-integrable functions . . . 167
6.2.2 Estimating a functional regression via HDLM . . . 170
6.3 Consistency in two types of sequences of statistical experiments 174
6.3.1 Simple asymptotic, independent and identically distributed data . . . 174
6.3.2 Double-asymptotic, independent and identically distributed data . . . 178

7 Conclusion 181
7.1 Overview of the contribution . . . 181
7.2 Possible directions for future research . . . 182

Bibliography **185**

Acronyms

CCWF Cross-Correlation of Wavelet Functions. 99, 107–109, 113, 115–117, 120, 125, 150

CLT Central Limit Theorem. 28, 95

EWS Evolutionary Wavelet Spectrum. 112, 113, 116, 123, 125–127, 130, 139, 152, 156

GDP Gross Domestic Product. 74, 86, 87, 92

GEP Generalized Eigenvalue Problem. 25–27, 29, 31–33, 35, 36, 47, 59

HD High dimensions. 26, 78

IHT Iterative Hard Tresholding. 28, 42, 43, 85, 86, 88, 92

LASSO Least Absolute Shrinkage and Selection Operator. 28, 39, 42–45, 73, 74, 92

LR Loaded-truncated Rayleigh flow method. 28, 29, 34, 40, 42–45, 85, 86, 88

MLSW Multivariate Locally Stationary Wavelet Process. 99, 113, 125–127, 132, 134, 137, 138

MPT Multiple-splitting Projection Test. 74, 78, 92

MRA Multiresolution analysis. 100, 101, 103

OPEC The Organization of the Petroleum Exporting Countries. 23, 125, 126, 134–138, 140–144

RIFLE Truncated Rayleigh flow method. 27–29, 33, 35, 47, 73, 85

SDAR Support Detection And Root finding. 28, 42, 43, 73, 85, 86, 88, 92

List of Figures

1.1 Oil production of Saudi Arabia and of the remaining of OPEP . . 18
1.2 Variation in oil production of Saudi Arabia and of the remaining of OPEP . . . 18
1.3 Diagram of the thesis . . . 21

2.1 Sketch of β, Φ and $\phi^\star$ . . . 30
2.2 Relative error of estimation of the direction . . . 44

3.1 Probabilities computed by numerical integration . . . 82
3.2 Scenario 1 . . . 87
3.3 Scenario 2 . . . 88
3.4 Scenario 3 . . . 89
3.5 Illustration of the control on the error of Type I (scenario 5) . . . 89
3.6 Illustration of the control on the error of Type I (scenario 6) . . . 90
3.7 Illustration of the control on the error of Type I (scenario 7) . . . 90
3.8 Results of the goodness-of-fit tests on real data . . . 91

4.1 Illustration of the Haar system . . . 103

5.1 Estimation of the EWS for $n = 1000$ . . . 134
5.2 Estimation of the EWS for $n = 10000$ . . . 135
5.3 Cosine similarity . . . 141
5.4 Local Coherency . . . 145
5.5 Instant correlation estimated using rolling window . . . 146
5.6 Cosine similarity estimated with a shorter smoothing parameter . 147
5.7 Local coherency at different scales estimated with a shorted smoothing parameter . . . 147
5.8 Instantaneous correlation estimated using rolling window and a shorter smoothing parameter . . . 148
5.9 Cosine similarity estimated with monthly data . . . 148
5.10 Local coherency at different scales estimated with monthly data . 149

5.11 Instantaneous correlation estimated using rolling window and monthly data 149

List of Tables

2.1 Summary of our comparison of four algorithms 28
2.2 Signal-to-noise ratios considered in the four scenarios. 43
2.3 Frequency of support recovery 45
2.4 Relative risk of estimation 46

3.1 Statistical table 84

List of Main Symbols

$\in$	Belong to	16
$\mathbb{R}^{\cdot}$	Euclidean space	16
$\times$	Cross product	16
$\{\cdot\}$	Set	16
$\mathbb{N}$	Set of natural number	16
$=$	Equal to	16
$O(a)$	Bounded by a constant times a.	16
$>$	Greater than	16
$-$	Subtracted by, *also*, opposite	17
$\mathbb{Z}$	Set of integers	17
$\mathbb{P}_{\cdot}$	Probability measure	19
$\cdot\vert\cdot$	Given that	19
$\neq$	Not equal to	19
$\Rightarrow$	Only if, *or equivalently*, implies	19
$\forall$	For all	20
$\sigma(\cdot)$	$\sigma-$algebra generated by argument	20
$\int$	Integral (most often, Lebesgue)	20
d	Differential	20
$\frac{\cdot}{\cdot}$	Division	23
argmin	Set of arguments that minimize	25
$\Vert\cdot\Vert_2$	Canonical l_2 norm	25
$\cdot^a$	Power of a	25
$\Vert\cdot\Vert_0$	l_0-pseudonorm	25
$\leq$	Lower or equal to	25
$\sum_{\cdot}$	Summation	25
$\vert\cdot\vert$	Absolute value	25
$\mathbf{1}_{\cdot}$	Indicator function	25
$+$	Added to	26
$\mathbb{E}(\cdot\vert\cdot)$	Conditional expectation	26
$\cdot^\top$	Transpose	26
Φ	Direction of the parameter of interest	26

$\subset$	Subset of	26
$<$	Lower than	26
argmax	Set of arguments that maximize	27
$:=$	Defined as	27
log	Logarithm function	27
$/$	Division	27
$O(a)$	Bounded by a constant multiplied by a	29
$\mathbb{E}$	Mathematical expectation	30
$\langle \cdot, \cdot \rangle_2$	Canonical l_2 inner product	30
$\mathcal{D}$	Direction space	30
$\phi^\star$	Normalized parameter of interest	30
$\langle \cdot, \cdot \rangle_\Sigma$	Inner product induced by Σ	30
ker	Kernel of an operator	30
$\| \cdot \|_\Sigma$	Norm assocciated to $\langle \cdot, \cdot \rangle_\Sigma$	31
$\geq$	Greater or equal to	32
max	Maximum	32
$o(a)$	Converges two zero when divided by a	32
$O_p(a)$	Bounded in probability by a constant multiplied by a	32
$\rightarrow$	Tends to *or* (when between sets) application between	35
sup	Supremum	35
$\mathcal{F}(\cdot, \cdot)$	Set containing different integers	36
$\mathcal{P}(\cdot)$	Set containing orthogonal projectors	36
$I_\cdot$	Identity operator	36
$\| \cdot \|_{\text{spec}}$	Spectral norm	36
$Cr(\cdot, \cdot)$	Crowford number	36
$\sqrt{\cdot}$	Square root	37
$\prod_\cdot$	Multiplication	40
$\mathcal{AN}$	Asymptotically normally distributed	40
$\mathbb{V}\text{ar}$	Variance	41
$o_p(1)$	Converges in probability to zero	41
$\lim_\cdot$	Limit of	42
$\sim$	Distributed as	42
$\mathbb{E}[\cdot]^2$	Mathematical expectation of the squared value	44
$\sqcup$	Disjoint union	48
$\mathcal{L}$	Lagrangian	49
$\nabla_\cdot$	Gradient	49
$\square$	End of proof	49
$\mathbb{P}(\cdot \vert \cdot)$	Conditional Probability	53
$\cap$	Intersection	54
$\xrightarrow{p}$	Converges in probability to	55
$\mathcal{N}_\epsilon$	ϵ-net	66
range$(\cdot)$	Range of an operator	75
rank$(\cdot)$	Rank of an operator	75
$\neq$	Not equal to	76

$\Leftrightarrow$	Equivalent to	76
$\setminus$	Subtracted of the elements of the set	77
H_0	Null-hypothesis	78
$\mathbb{C}\mathrm{ov}$	Covariance	78
Φ_G	Cumulative distribution function of a standard normal	80
H_1	Alternative hypothesis	80
$\mathcal{N}$	Normal distribution	81
R^2	Coefficient of determination	88
$\varepsilon_{\cdot,\cdot}$	Standard normal increment	98
$\psi_{\cdot,\cdot}(\cdot)$	Wavelet function	98
$h_{\cdot}$	Scaling coefficient	100
$g_{\cdot}$	Element of the wavelet filter	102
$L[h]$	Length of the scaling filter	105
$L_{\cdot}$	Length of the support of wavelets	106
$\Psi_{\cdot,\cdot}$	Cross-correlation of wavelet functions (CCWF)	107
$\Psi_{\cdot}$	Collection of CCWF	107
$\psi_{\cdot,\cdot}$	Vector of wavelet functions	113
$S^{(\cdot,\cdot)}(\cdot)$	Evolutionary Wavelet Spectrum (EWS)	113
$C_{\cdot,\cdot}^{(\cdot,\cdot)}$	Lipschitz constant of an element of the EWS	113
$C_n^{(\cdot,\cdot)}(\cdot,\cdot)$	Covariance between two times series for a given n	114
$R^{(\cdot,\cdot,\cdot,\cdot)}(\cdot,\cdot)$	A rest studied in Chapter 5	114
$\Phi_{\cdot,\cdot}$	Vector of CCWF	115
$S_{\cdot,\cdot}^{(\cdot,\cdot)}$	Vector of elements of the EWS	115
$C^{(\cdot,\cdot)}(\cdot,\cdot)$	Local covariance function between two time series	116
$A_{\cdot,\cdot}$	Some Gram matrix	116
$d_{\cdot,\cdot}^{(\cdot)}$	Wavelet coefficient	126
$I_{\cdot,\cdot}^{(\cdot,\cdot)}$	Raw wavelet periodogram (vector)	127
$I_{\cdot,\cdot}^{(\cdot,\cdot)}$	Smoothed wavelet periodogram (vector)	128
$\tilde{I}_{\cdot,\cdot}^{(\cdot,\cdot)}$	Smoothed and corrected wavelet periodogram (vector)	130
$\rho^{(\cdot,\cdot)}(\cdot)$	Instantaneous correlation	140
$\kappa_{(\cdot,\cdot)(j,\cdot)}$	Local coherency at scale j	140
$\mathcal{E}_{\cdot}$	A statistical experiment	164
$[\cdot,\cdot)$	Left closed and right open interval on $\mathbb{R}$	164
$L^2\cdot$	Set of square-integrable functions on $\cdot$	164
$\langle\cdot,\cdot\rangle_{L^2}$	Inner product on $L^2[0,1)$ with Lebesgue measure	164
$D_{\cdot}$	Continuation operator	168
$\langle\cdot,\cdot\rangle_{l_p}$	Normalized canonical inner product on $\mathbb{R}^p$	168
$D_{\cdot}^{\star}$	Discretization operator	169

One

Introduction

1.1 General introduction

The amount of information provided to communities and individuals each day is enormous. Our systems designed to learn from this information are overloaded to the point where they often cannot process conflicting reports from multiple sources. To function, we collectively rely more and more on some form of auditing services that we trust to help us distinguish between valid and invalid information, between truth and falsehood. All of this inhibits our ability to learn from the entire stream of uncertified information.

This thesis provides theories and methods that enable learning from the maelstrom of available facts known as big data.

Big data sets are often called high-dimensional data sets in statistical literature. In a linear regression model, for instance, high dimensionality means that the number of covariates is greater than the number of observations. In a multivariate time series framework, high dimensionality refers to situations where many time series are studied jointly.

In today's practice, the empirical researcher is often confronted with high-dimensional data. This is the case, for instance, when central banks seek to forecast key economic variables (*e.g.* Bok, Caratelli, Giannone, Sbordone and Tambalotti (2018)), when biologist seeks to determine which genes are causing a particular disease (*e.g.* Carvalho, Chang, Lucas, Nevins, Wang and West (2008)), and even in the training of cutting-edge machine learning models (e.g. Brown, Mann, Ryder, Subbiah, Kaplan, Dhariwal, Neelakantan, Shyam, Sastry, Askell et al. (2020)).

Many successful applications have demonstrated that there is much to be gained by jointly exploiting a very large data set to answer empirical questions. However, modern applications push the dimensionality of the data to such an extent that it has become necessary to consider that the dimensionality of the data can grow very fast with the sample size or even that it is infinite.

Several methodological obstacles hinder the application of standard statistical methods to the high-dimensionality framework. For instance, for cross-

sectional data sets, they are often not consistent or even exist when the ratio between the number of variables and the number of observations diverges.

Moreover, in the treatment of time series, much of the existing literature assumes that the data have characteristics that are convenient for simplifying the mathematical treatment of the problem but that are rarely encountered in reality. One of these unrealistic assumptions is the stationarity of time series which assumes that the statistical properties of time series, such as the mean and variance, remain constant over time. When the time series under consideration are not stationary, common statistical methods often become inconsistent.

In this thesis, we study novel ways to conduct inferences on high-dimensional cross-sectional data and high-dimensional non-stationary time series.

1.2 Topics of the thesis

1.2.1 High-dimensional cross-sections and high-dimensional non-stationary time series

This work is concerned with *high-dimensional* data in the forms of random matrices $X_n \in \mathbb{R}^{n\times p}$ where n, is the *number of observations* and p is the *number of covariates*.

Those datasets may be called *cross-sections* when the observations of the p covariates represent records of properties (the covariates) of different *items* (persons, groups, units, firms, countries,...) *at some unique* point *in time*. Where this thesis considers high-dimensional cross-sections, it is assumed that, in addition to X_n, a random vector $Y_n \in \mathbb{R}^n$ is observed. The inference considered in the thesis concentrates on the linear regression of Y_n, the *explained variable*, on X_n, the *explanatory variables*. There, high-dimensionality characterizes a situation where the number of covariates p is larger than n. Formally, it is assumed that X_n is an element of a sequence of random matrices of growing dimension, that is $\{X_n \in \mathbb{R}^{n\times p(n)}\}_{n\in\mathbb{N}}$ $p(n) = O(n^\alpha)$ where $\alpha > 1$.

The observations of the p covariates may, alternatively, be supposed to have happened sequentially in time. In this case each line of X_n is written in vector form as $X_n(t) \in \mathbb{R}^p$ for $(t = 0, ..., n-1)$, that is $X_n = (X_n^\top(0), ..., X_n^\top(n-1))^\top$, and these vectors are called *multivariate time series*. Where this thesis considers high-dimensional multivariates times series, it concentrates on *zero-mean* vector processes that arise, for instance, when the overall trend of each covariate has been removed. In this setting, the inference considered in this thesis concentrates on the estimation of the covariance between two vectors of time series, say $X_n(t)$ and $X_n(t-\tau)$ where $\tau \in \mathbb{Z}$ is a *lag*. In this context, high dimensionality refers to the high number of time series that are considered jointly. Formally, as for cross-sectional data, it is assumed that X_n is an

element of a sequence of random matrices of growing dimension, that is $\{X_n \in \mathbb{R}^{n\times p(n)}\}_{n\in\mathbb{N}}$, however, for time series, we only consider such sequences with $p(n) = O(n^\alpha)$ where $\alpha < 1$. Hence, for time series, we do not treat cases where $p > n$ and always suppose the opposite. This simpler form of high-dimensionality renders possible to conduct the type of inference considered in this thesis, without relying on a simplifying assumption that is common in the analysis of zero-mean multivariates time series, namely the *stationarity of the second order structure*, a notion that we explain now.

Most existing models for multivariates time series analysis assume that the zero mean $X_n(t)$ is *covariance stationary*, meaning that the covariance between $X_n(t)$ and $X_n(t-\tau)$ only depends on the lag τ and not on the time t. This assumption is very useful to simplify the mathematical treatment of estimators of the covariance between two multivariate time series. However, in applied sciences, many time series are *not* covariance stationary and, rather display *time-varying second-order structure* (variance and covariance change over time). An example of such a multivariate time series is given in Figures 1.1 and 1.2. The former displays the quarterly oil production of Saudi Arabia and the one of the rest of OPEC in million barrels per day between the first quarter of 1973 and the third quarter of 2022. From Figure 1.1, it is clear that the processes are not zero means. However, it is possible to remove their global trend by computing their first log difference. The resulting time series are displayed in Figure 1.2 where it is clear that, first, the time series may be modeled as zero-mean processes, and second, that the second order structures of the two time series change over time.

The high-dimensional time series considered in this thesis are *not* supposed to be covariance stationary.

1.2.2 The challenges of inference in high dimensions

With high-dimensional cross-sections, the thesis concentrates on the *High-Dimensional Sparse Linear model*. That is a linear regression model with more covariates than observations and where only a small subset of the covariates are significant. In such a setting, the three main challenges are (1) recovering the unknown set of significant covariates (the selection problem), (2) estimating the marginal effect of the selected covariates (the estimation problem) (3) statistically testing the validity of the selections and estimations procedures. Performing these tasks is difficult because standard estimation techniques (such as ordinary least squares), as well as standard statistical tests (such as the Hoteling T^2 test (Hotelling, 1931)) fail in high dimensions because the empirical second moment of the covariates is not invertible.

With high-dimensional time series, the thesis concentrates on a model of *locally stationary wavelet processes*. In these kinds of models, times series are represented as a linear combination of some analytically tractable functions, called *wavelets*, and of random increments. Those types of models have already

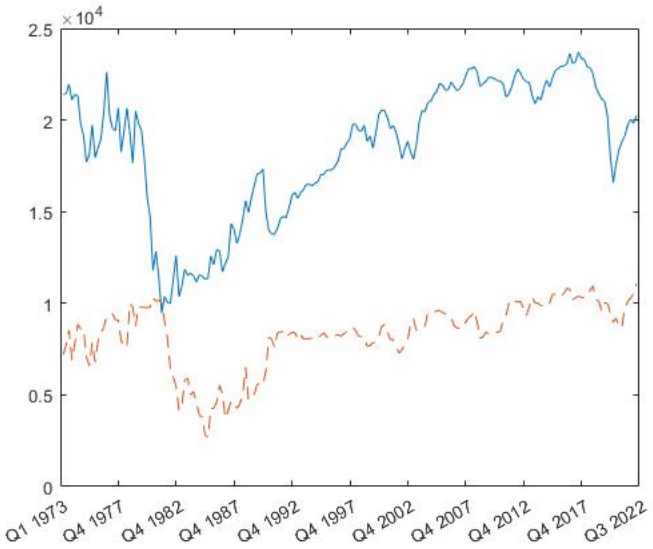

Figure 1.1: Oil production of Saudi Arabia and of the remaining of OPEP

Quarterly oil production of Saudi Arabia (dashed red line) and of the rest of OPEC (plain blue line) in million barrels per day between the first quarter of 1973 and the third quarter of 2022.

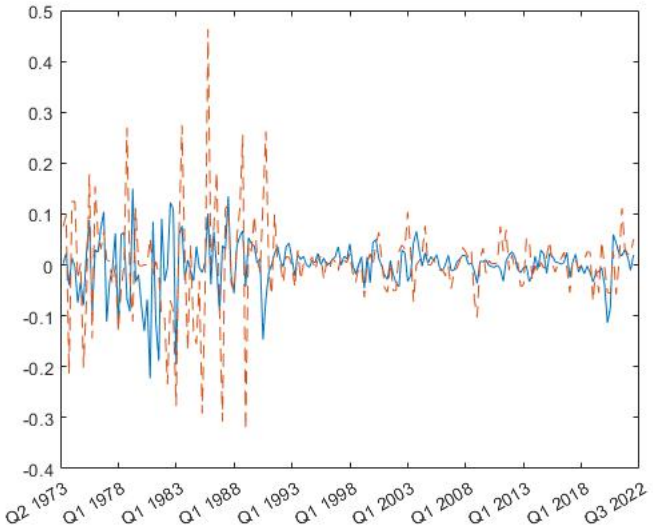

Figure 1.2: Variation in oil production of Saudi Arabia and of the remaining of OPEP

First log difference of the quarterly oil production of Saudi Arabia (dashed red line) and the rest of OPEC (plain blue line) between the second quarter of 1973 and the third quarter of 2022.

been studied for univariate non-stationary time series in Nason, Von Sachs and Kroisandt (2000) and Van Bellegem and von Sachs (2008). They are based on the concept of *local stationarity* introduced by Dahlhaus (1997) that allows the modeling and the estimation of time-varying autocovariance. The main challenge here is to generalize the approach of Nason et al. (2000) and Van Bellegem and von Sachs (2008) to a multivariate setting imposing the least stringent constraint on the time-varying covariance of the time series.

1.2.3 Identification of parameters

Before conducting inference, one wants to ensure that parameters of interest are *identified*. This section briefly discussed the concept of identification for unconditional and conditional models.

Consider a statistical experiment $\{S, \mathcal{S}, \mathbb{P}_\theta | \theta \in \Theta\}$ where S is the *sample space*, $\mathcal{S}$ a $\sigma-$algebra on S, $\mathbb{P}_\theta$ is a (joint) *probability measure* and Θ is set called *parameter space*.

In this thesis, a parameter of interest, say $g(\theta)$ (where g is a function defined on the parameter space), is said *identified*, if $g(\theta) \neq g(\theta') \Rightarrow \mathbb{P}_\theta \neq \mathbb{P}_{\theta'}$.

The relevance of this property is not obvious and related to the *interpretation* of the model. Suppose that our parameter of interest is not identified, hence this means that there might exist $\theta, \theta' \in \Theta$ such that $\mathbb{P}_\theta = \mathbb{P}_{\theta'}$ with $g(\theta) \neq g(\theta')$. In this case, we are left with a quantitatively equivalent, but qualitatively different, representation of the probability we are interested in. In other words, we are left with different interpretations of the phenomena we are interested in, without the ability to prefer one interpretation over another. Consequently, when the model is not identified, the statistical analysis does not permit to gain information concerning the parameter of interest that cannot be uniquely represented as a function of the probability measure of the data.

An identification result demonstrates the above injection and, by doing so, defining uniquely the parameter of interest as a function of the probability measure of our observables. In Chapter 5 we present a new identification result for an unconditional model.

The above notion of identification is given in the context of unconditional models where $\mathbb{P}_\theta$ represents a joint law of probability. However, it is often the case in econometrics that we are only interested in some particular exogenous (in the sense of Engle, Hendry and Richard (1983)) unknown parameter of a *conditional* model. For instance, in a linear regression problem, one is often interested in the parameter of the linear function of the covariates (often denoted by β). In the following, we give a (simplified) notion of identification for conditional models that is due to Mouchar and Oulhaj (2006) (to which we refer the reader for a deeper discussion on the matter).

Suppose that $S = S_Y \times S_X$ where S_Y is the sample space of the explained variable and S_X is the sample space of the explanatory variable, moreover write $\mathcal{S} = \sigma(\mathcal{S}_Y \cup \mathcal{S}_X)$ where $\mathcal{S}_Y$ and $\mathcal{S}_Z$ are the sigma-algebras generated[1] respectively by S_Y and S_X, and $\sigma(\mathcal{S}_Y \cup \mathcal{S}_X)$ is the $\sigma-$algebra generated by $\mathcal{S}_Y \cup \mathcal{S}_X$. Moreover, suppose that $\Theta = \Theta_1 \times \Theta_2$ where Θ_1 is the space in which live the (exogenous) parameters of the conditional model and Θ_2 is the space in which live the parameters of the marginal distribution of the explanatory variables. Finally suppose that for any $\theta \in \Theta$ the joint probability law $\mathbb{P}_\theta$ is (uniquely) defined by

$$\forall(\mathcal{Y}, \mathcal{X}) \in \mathcal{S}_Y \times \mathcal{S}_X, \quad \mathbb{P}_\theta(\mathcal{Y} \times \mathcal{X}) = \int_{\mathcal{X}} \mathbb{P}_{(Y|x);\theta_1}(\mathcal{Y}) \mathbb{P}_{X;\theta_2}(dx)$$

where for $\theta_1 \in \Theta$, and $x \in S_X$, $\mathbb{P}_{(Y|x);\theta_1}(y)$ is the conditional probability measure of the explained variable Y given the realisation x the explanatory variable and $\mathbb{P}_{X;\theta_2}$ is the marginal probability measure of the explanatory variable. We then have defined the (simplified) *conditional model*

$$\{S_Y \times S_X,\ \sigma(\mathcal{S}_Y \cup \mathcal{S}_X),\ \mathbb{P}_{(\theta_1,\theta_2)} | (\theta_1, \theta_2) \in \Theta_1 \times \Theta_2\}.$$

In such a model our *parameter of interest* is a function of the exogenous parameter of the conditional distribution, say $g(\theta_1)$ where $\theta_1 \in \Theta_1$. We said that our parameter of interest is *(weakly) identified* if for all $\theta_2 \in \Theta_2$, $g(\theta_1) \neq g(\theta_1') \Rightarrow \mathbb{P}_{(\theta_1,\theta_2)} \neq \mathbb{P}_{(\theta_1',\theta_2)}$.

Often, proving the weak identification of the parameter of interest in a conditional model comes to impose restrictions on the marginal distribution of the explanatory variable. For instance in linear regression, one usually assumes that the kernel of the second moment of the explanatory variable contains only 0. In Chapter 2 and 3, we provide two new identification results in the context of conditional models.

[1] Here we follow the phrasing in Florens, Mouchart and Rolin (1990) where, given a collection of subsets of a set M, say $\mathcal{C}$, the sigma-algebra generated by $\mathcal{C}$ is constructed as the intersection of all the $\sigma-$algebra of $\sigma(M)$ (a $\sigma-$algebra on M) containing C, which is the smallest $\sigma-$algebra containing C.

1.3 Contributions and structure of the thesis

The thesis contributes to the development of statistical analysis of both high-dimensional cross-sectional data and high-dimensional time series. Our contributions relate both to the theoretical and to the empirical analysis of high-dimensional data. For high-dimensional cross-sections, we provide new results concerning the modeling, the selection, the estimation, and the statistical testing of hypothesis in a high-dimensional linear sparse model. For high-dimensional time series data, we provide new results for modeling and estimating the time-varying covariance of high-dimensional zero-means processes through the definition, the identification, and the estimation of a novel class of locally stationary wavelet process model.

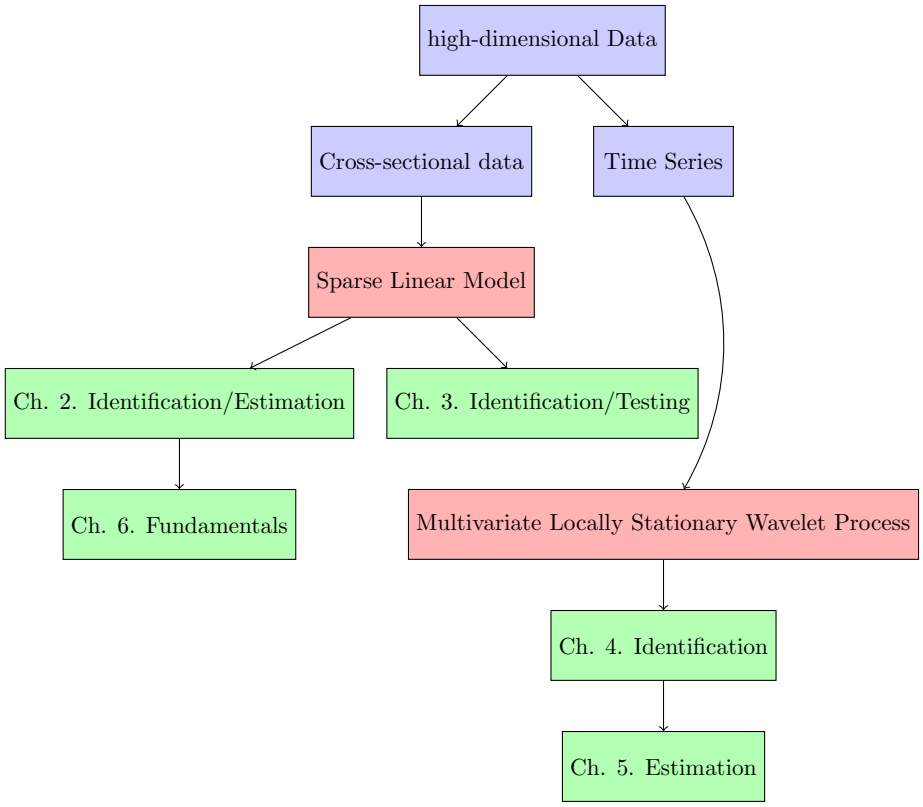

Figure 1.3: Diagram of the thesis

Chapter 2 establishes the connection between two well-studied statistical problems: the *Best Subset Selection* and the *Generalized Eigenvalue Problem*. This connection is based on the idea that in linear regression, the parameter of interest, that is the parameter of the linear functional of the covariates, is completely characterized by two objects. First, its direction thereafter called the *direction of interest*, defined as the vector subspace of all the vectors

collinear to the parameter of interest. Second its magnitude, which is its norm under a suitable choice of basis for the direction of interest. In this chapter, the direction of interest is identified with the *generalized eigenspace* associated with the only nonnull *generalized eigenvalue* of a pair of measurable matrices. Because our identification result connects the linear regression problem and the generalized eigenvalue problem, it extends the class of estimators available for the former to the one available for the latter. This is especially useful in the context of high dimensions, where standard estimation techniques fail. Consequently, we suggest addressing the best subset selection problem from a new perspective by solving an empirical generalized eigenvalue problem to estimate the direction of the high-dimensional vector of parameters. We then study a new estimator based on the well-established RIFLE algorithm (see Tan, Wang, Liu and Zhang (2018)) and demonstrate a nonasymptotic bound of the l_2 risk, the minimax convergence of the estimator and a central limit theorem. Simulations show the superiority of the proposed inference over some direct competitors.

Chapter 3 introduces a goodness-of-fit test for the outcome of variable selection in a high-dimensional sparse linear model. The test aims to minimize a moment condition that is proven to reflect the sparsity constraint. Testing this constraint statistically is possible thanks to a high-dimensional Central Limit Theorem proven under heteroskedasticity. To implement the test in high dimensions a Multiple-splitting Projection Test procedure that has been recently proposed in the literature is employed (Liu, Yu and Li, 2022). An upstream result addresses the question of collinearity in the set of non-significant covariates. This result provides a new identification condition for the parameters of a sparse linear model that allows collinearity in the set of covariates that do not affect the conditional expectation of the explained variables. Monte Carlo experiments demonstrate the power of the test. A real data application considers the problem of selecting predictors to nowcast the quarterly Belgian GDP via monthly time series. The empirical results show that it is possible to select a minimal number of variables such that every larger set of variables would pass the goodness-of-fit test.

Chapter 4 aims three goals. Firstly, introducing the reader to the fundamental concepts of wavelet analysis. This is done by reviewing the construction of wavelet functions via multiresolution analysis, reviewing the construction of *non-decimated discrete* wavelet functions from continuous wavelet functions, and presenting some of the properties of the former. Secondly, the chapter defined the notion of *cross-correlation of wavelet functions* that is fundamental to the present extension of the univariate framework of Nason et al. (2000) to a multivariate framework. Proof that a collection of cross-correlation of wavelet functions forms a basis is given. This result is essential to be able to uniquely represent a function in the space of cross-correlation of wavelet functions. Thirdly, the chapter defines the class of *multivariate lo-*

cally stationary wavelet process and asymptotically identifies a representation of the time-varying covariance function of the vector time series, in the space of the cross-correlation of wavelet functions.

Chapter 5 studies the estimation of the representation of the covariance structure of the vector of time series in the space of the cross-correlation of wavelet functions. The chapter demonstrates the l_2 convergence in probability of our estimator when $p = o(n^\alpha)$ when $\alpha < 1/4$. The last section of the chapter presents an economic application of the method. The application considers measuring the (lack of) coordination between Saudi Arabia and the rest of the OPEC organization in oil production. This application is a contribution to a large field of applied econometrics that has been puzzled since the eighties by the contradictory reports of (lack) of statistical measurement of the said coordination (e.g. Griffin (1985), Smith (2005), Alkhathlan, Gately and Javid (2014), Al Rousan, Sbia and Tas (2018)). The chapter argues that the issues encountered with the statistical analysis of the behavior of Saudi Arabia and one of the rest of OPEC can be tracked back to the assumption of stationarity upon which have relied on the econometric methods employed to tackle the measurements so far. New measurements, that are based on the model presented in Chapter 5, and conducted via the estimator studied in Chapter 6, are presented. To the best of our knowledge, those measurements of the coordination of Saudi Arabia with its partners in OPEC are the first ones not to rely on the assumption of stationarity. The results of the measurements are consistent with the historical trends of the last half-century in the oil market.

Chapter 6 discusses the present duality there is in the modelization of high-dimensional data. In the first approach, which is the one considered in this thesis and may be called the double asymptotic setting, the number of covariates is an increasing function of the number of observations. Asymptotic reasoning conducted in this setting is different than when only the number of observations differs. One consequence of this fact is that, in this context, it is not clear how to characterize the consistency of an estimator. The last section of the chapter provides an analysis of two abstract statistical experiments in order to provide definitions of consistency that are applicable to the double asymptotic setting. A second fruitful path to model high dimensionality works in an infinite dimensional space. This is for instance the case in functional regression analysis (e.g. Florens and Van Bellegem (2015)) that is becoming popular in applied studies (see for instance the recent work of Meeks and Monti (2022)). At this point in the development of the literature, the links between the two approaches remain unclear. Nevertheless, the second one may/should be considered as the limit of the first one. However, connecting the two approaches raises some technical challenges that can only be addressed in an appropriate and general theoretical framework that is outside the scope of this thesis. Nevertheless, after discussing the issue the chapter provides

early results derived to answer a simple question concerning the estimation of the functional parameters in a functional regression defined in L^2 that is interestingly linked to the problem of the connection between the double asymptotic setting and the infinite-dimensional setting.

Two

Direction Identification and Minimax Estimation by Generalized Eigenvalue Problem in High-Dimensional Sparse Regression

2.1 Introduction

This chapter connects two well-known statistical problems. First, the *best subset selection* problem that is a solution of

$$\tilde{\beta} \in \underset{\hat{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} \frac{1}{n}\|Y - X\hat{\beta}\|_2^2 \text{ s.t. } \|\hat{\beta}\|_0 \leq k, \tag{2.1.1}$$

for $k \in \mathbb{N}$ (Miller, 2002), where $Y \in \mathbb{R}^n$ and $X \in \mathbb{R}^{n\times p}$ and where for any $v \in \mathbb{R}^p$, $\|v\|_2 = \sum_{i=1}^p |v_i|^2$ is the l_2-norm of v and $\|v\|_0 = \sum_{i=1}^m \mathbf{1}_{|v_i|>0}$ is the $l_0-pseudonorm$ of v. Second, the *Generalized Eigenvalue Problem* (GEP), that is, given two square matrices $A, B \in \mathbb{R}^{p\times p}$, solving

$$Av_i = \lambda_i Bv_i \text{ for } i = 1, ..., p \tag{2.1.2}$$

for λ_i and v_i, which are called *generalized eigenvalues* and *generalized eigenvectors*, respectively. We observe that if B is the identity matrix, then the GEP is the standard eigenvalue problem (for an introduction to GEP, see *e.g.,* Ghojogh, Karray and Crowley (2019)).

The connection between the two problems is based on the idea that in the linear regression $Y = X\beta + U$, where $Y \in \mathbb{R}^n$ and $X \in \mathbb{R}^{n\times p}$ are random and *iid* across rows and $U = (Y_1 - \mathbb{E}(Y_1|X_1), ..., Y_n - \mathbb{E}(Y_n|X_n))^\top$, the parameter β is completely characterized by two objects. First, its *direction*, $\Phi \subset \mathbb{R}^p$, defined as the vector subspace of $\mathbb{R}^p$ of all the vectors colinear to β, and second, its *magnitude* that its norm in $\mathbb{R}^p$ under a suitable choice of basis for Φ. In Theorem 1 below, we identify the direction Φ with the *generalized*

eigenspace associated with the only nonnull *generalized eigenvalue* of a pair of measurable matrices.

One important consequence of our identification result is that by exposing the connection between the linear regression problem and the GEP, we extend the class of estimators available for the former to the one available for the latter. This is especially useful in the context of *high dimensions* (HD).

When the dimension of the target parameter β, namely, p, is greater than the number of observations, n, for instance, because the former is an increasing function of the latter, we say that the model is *high-dimensional*. Estimating the parameter β in HD is a problem of great interest from both theoretical and practical points of view. Since the high dimensionality of the model implies that the empirical second moment of the covariates is not invertible, the parameter β cannot be consistently estimated via ordinary least squares. However, situations where $p > n$ arise in practice, for instance, in biology, health studies, economics, or machine learning (Fan and Li, 2006), and call for efficient estimation procedures.

There is no unique way to tackle the problem of consistently estimating the parameters of a high-dimensional linear model. One line of research is based on the notion of *sparsity*. According to Stodden (2006), Box and Meyer (1986) coined the term *factor sparsity* to describe a model where the majority of covariates have zero effect. Since then, a wide and active area of research has specialized in the estimation of the parameters of high-dimensional *sparse* linear models *i.e.* high-dimensional linear models where it is assumed that only a subset of the elements of β differs from zeros (*e.g.* Belloni, Chernozhukov and Hansen (2013), van de Geer (2016)). Sparsity assumes that

$$\|\beta\|_0 = s \leq n < p$$

where s is the *sparsity level*.

The best subset selection problem (2.1.1) originates from the combination of the sparsity assumption with the least squares principle. Due to the l_0 constraint, this program is NP-hard (Natarajan, 1995). Hence, solving (2.1.1) in polynomial time is impossible unless $P = NP$. Despite this computational difficulty, a set of theoretical results indicates that the best subset selection is a desirable estimator for sparse regression (*e.g.* Greenshtein (2006), Raskutti, Wainwright and Yu (2011), Zhang and Zhang (2012), Shen et al. (2013), Zhang et al. (2014), Bertsimas, King and Mazumder (2016)). Recently, optimal or near-optimal solutions have been proposed in Bertsimas et al. (2016) via a reformulation of problem (2.1.1) as a mixed integer optimization (MIO) problem. The implementation of this method is performed via the licensed GUROBI solver and is demanding in computing time as the dimension of the model increases. For a critical review, see *e.g.* Hastie, Tibshirani and Tibshirani (2020).

By following the same line of thought that leads to the formulation of

the best subset selection, we formulate an estimator of Φ over a sparsity constraint. To do so, we observe that the generalized eigenvector associated with the maximum generalized eigenvalue of the GEP (2.1.2) is the solution of the maximization of the following Rayleigh quotient

$$\underset{v \in \mathbb{R}^p}{\operatorname{argmax}} \frac{v^\top A v}{v^\top B v},$$

see *e.g.* Golub and Van Loan (2013). Our identification result in Theorem 1 allows combining this fact with the sparsity assumption to obtain the sparse estimator of Φ, that is,

$$\hat{\Phi} := \underset{\phi \in \mathbb{R}^p}{\operatorname{argmax}} \frac{\phi^\top X^\top Y Y^\top X \phi}{n \phi^\top X^\top X \phi} \text{ s.t. } \|\phi\|_0 \leq k \tag{2.1.3}$$

for $k \in \mathbb{N}$. We call (2.1.3) the *directional best subset selection* problem.

Below, we demonstrate that for any solution of problem (2.1.3), say v, the product of v with the orthogonal projection of Y on Xv is a solution of the best subset selection (2.1.1). Hence, solving problem (2.1.3) allows solving (2.1.1). We also demonstrate that the two problems have the same minimax rate of convergence in the l_2 norm, that is, $(s \log(p/s)/n)^{1/2}$ in the sub-Gaussian setting (Raskutti et al., 2011).

Next, we show how these ideas may be fruitful for the estimation of high-dimensional sparse linear models (HDSLM) by proposing an efficient estimator of Φ, which gives as a byproduct a novel estimator of β. This estimator is a modification of a well-established algorithm called RIFLE that was proposed by Tan, Wang, Liu and Zhang (2018) to approximate the solution of GEP over a l_0 constraint (which are also NP-Hard (Moghaddam, Weiss and Avidan, 2006)) when $p > n$. We demonstrate that our estimator converges at the minimax rate of convergence with high probability in the sub-Gaussian setting and that it possesses the sure screening property.

Finally, we provide a central limit theorem for the estimator, which is a nontrivial task given its nonlinearity. To this aim, we prove a theorem that gives sufficient conditions for the convergence in probability of a *Cauchy product*. This theorem may be of independent interest.

Generalized Eigenvalue Problems have already been employed to tackle models with many covariates. In particular, the *sufficient dimension reduction* approach introduced in Cook (1994) employs moment-based methods that can be formulated as GEP to replace the original covariate with a minimal set of their linear combinations without loss of information (Li, 2007). Those methods are not designed specifically to tackle high-dimensional sparse linear models and are not an answer to the best subset selection problem. We still mention the research of Chen, Zou and Cook (2010) that introduced a method

to induce sparsity in the estimated linear combination of the covariate and thus to *de facto* select variables.

Like RIFLE, our estimator requires an initial solution that is then iteratively improved toward convergence. The article concludes with a set of Monte Carlo experiments that show the superiority of the proposed inference when our estimator is initialized by using LASSO, that is, the Lagrangian version of the l_1-relaxation of problem (2.1.1) (Tibshirani (1996), Chen, Donoho and Saunders (1998)). LASSO is known to select too many random variables when tuned for prediction accuracy (Bertsimas et al. (2016), Hastie et al. (2020)). Our experiment shows that for signal-to-noise ratios higher than two, our estimator improves on the initial solution provided by LASSO, *while* reducing the number of variables associated with nonnull coefficients.

The performances of the proposed inference are compared with those of three other l_0 constrained/penalized estimators. First and second are the Iterative Hard Thresholding (IHT) and the two-stage IHT (Jain, Tewari and Kar (2014)) that are designed to approximate problem (2.1.1). The third competitor is the support detection and root finding (SDAR), proposed by Huang, Jiao, Liu and Lu (2018) to approximate the Lagrangian version of the best subset selection (which is not equivalent to problem (2.1.1) because of the l_0-constraint) but still allows fixing the sparsity level of the estimate.

We summarize our comparison of the four algorithms in Table 2.1 where *Loaded-RIFLE* (LR) is the name of our estimator. Minimax convergence rates and the sure screening property have been established for the IHT and two-stage IHT (Jain et al. (2014) and Guo, Zhu and Fan (2020)). Minimax convergence rates have been established for SDAR (Huang et al., 2018). To the best of our knowledge, no CLT has been established for these algorithms. The numerical performances are compared via Monte Carlo experiments in Section 5. Accuracy in estimation, variable selection, and prediction are compared. The computing time given in Table 2.1 is the rounded mean computing time of 100 replications in the second setup described in Section 5. The computational time of the LR *includes* the time for computing the initial vector via LASSO. Our implementation of the IHT optimizes the step size at each iteration (as in Blumensath and Davies (2009*a*)), which greatly improves estimation but is costly in computational time.

Estimator	Efficiency	CLT	Num. perf.	Time (s)
IHT	near minimax			$\approx 5e^{-1}$
2 steps-IHT	near minimax			$\approx \mathbf{3e^{-2}}$
SDAR	**minimax**			$\approx 1e^{-1}$
Loaded-RIFLE	**minimax**	**yes**	**best overall**	$\approx 2e^{-1}$

Table 2.1: Summary of our comparison of four algorithms

The organization of the chapter is as follows. Section 2 provides the struc-

tural model and the main objects with which we are working. Section 3 addresses the identification of Φ with the solution of a particular GEP. In Section 4, we use the result of the preceding section to treat the estimation of Φ under sparsity. First, we introduce the directional best subset selection (2.1.3) and establish a theory connecting it with the best subset selection (2.1.1). Second, we show how these results may be applied by presenting RIFLE and LR, which are two algorithms that produce solutions for the directional best subset selection problem. Third, we provide nonasymptotic bounds as well as the convergence rate in the sub-Gaussian setting for the relative error of these two estimators. Fourth, we present a central limit theorem for these estimators and discuss the theorem. This discussion allows us to present intermediate results that provide deep insights into the functioning of RIFLE. In particular, one of these results gives sufficient conditions for the convergence in the probability of the stochastic Cauchy product and may be of independent interest. In Section 5, we present numerical experiments to compare the performance of the proposed algorithms that demonstrate the competitiveness of LR. Section 6 contains concluding remarks.

2.2 Structural Model

We consider a sequence of statistical experiments indexed by $n \in \mathbb{N}$. *For each* n we assume that we observe a random vector of *explanatory variables*, $Y^{(n)} \in \mathbb{R}^n$ and a random matrix of *covariates* $X^{(n)} \in \mathbb{R}^{n \times p}$. We also assume that *for each* $n \in \mathbb{N}$ the rows of $Y^{(n)}$ and $X^{(n)}$ are independent and identically distributed. To model high dimensionality, we assume $p = p(n)$, that is, p, the *number of explanatory variables*, is a function of n, *the number of observations* and then consider situations where $p(n)$ may increase faster than n (for instance, below, we consider situations where the dimension p is of order $O(n^\alpha)$ with $\alpha > 1$).

We assume that for each $n \in \mathbb{N}$ there exists $\beta^{(n)} \in \mathbb{R}^p$ such that we have $\mathbb{E}^{(n)}(Y_1^{(n)}|X_1^{(n)}) = (X_1^{(n)})^\top \beta^{(n)}$ where $Y_1^{(n)}$ and $X_1^{(n)}$ represent the transpose of the first rows of $Y^{(n)}$ and $X^{(n)}$.

This chapter is only indirectly interested in $\beta^{(n)}$ and focuses instead on the identification and estimation of the *direction of* $\beta^{(n)}$, which is the one-dimensional subspace of $\mathbb{R}^p$ containing all the vectors collinear to β.

Definition 1 *(Direction space) For any non-null vector* $v \in \mathcal{V}$*, we call the* direction *of* v *the one-dimensional subspace of* $\mathcal{V}$ *defined by*

$$\mathcal{D}(v) := \{mv | m \in \mathbb{R}\}.$$

For the remainder of this chapter, we let $\Phi^{(n)} := \mathcal{D}(\beta^{(n)})$ and choose $\phi^{(n)\star} := \beta^{(n)}/\|\beta^{(n)}\|_2$ as the basis for $\Phi^{(n)}$, where for any and $u, v \in \mathbb{R}^p$, the l_2-norm of v, $\|v\|_2$, is induced by the *canonical inner product* on $\mathbb{R}^p$, namely, $\langle u, v\rangle_2 :=$

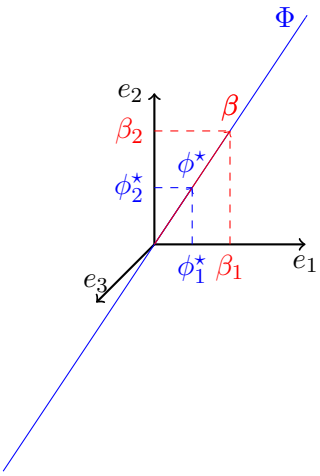

Figure 2.1: Sketch of β, Φ and $\phi^\star$

Graphical representation of β, Φ and $\phi^\star$ where $\{e_1, e_2, e_3\}$ is an orthogonal basis of $\mathbb{R}^3$.

$\sum_{i=1}^{p} u_i v_i$. We observe that this choice of basis for $\Phi^{(n)}$ implies that the magnitude of $\beta^{(n)}$ is its l_2-norm.

Since p may increase faster than n, we may encounter situations where the dimension of the parameter space is larger than the sample size. Hence, we require our high-dimensional linear model to be sparse; that is, for each $n \in \mathbb{N}$, we assume that $\|\phi^{\star(n)}\|_0 = s(n) < p(n)$.

In the remainder of the chapter, we drop the upper script (n) to simplify the notation.

2.3 Identification of the direction of interest

This section provides a theorem that identifies Φ to some functions of moments of Y_1 and X_1. The result suggests a natural estimator that is exploited in Section 2.4. Before stating the theorem, we introduce some notation. Let $\Sigma_X := \mathbb{E}(X_1 X_1^\top)$ be the second moment of X_1; moreover, let $\Sigma_{XY} := \mathbb{E}(X_1 Y_1)$ and $\Sigma_{YX} := (\Sigma_{XY})^\top$. For all $u, v \in \mathbb{R}^p$, let $\langle u, v \rangle_{\Sigma_X} := \langle u, \Sigma_X v \rangle_2$ be the inner product induced by Σ_X on $\mathbb{R}^p$. For any $m \in \mathbb{N}$ and $Z \in \mathbb{R}^{m \times m}$, let $\mathrm{Ker}(Z) = \{v \in \mathbb{R}^m; \|Zv\|_2 = 0\}$ be the *kernel* of Z.

Theorem 1 *Let $Y_1 \in \mathbb{R}$ and $X_1 \in \mathbb{R}^p$ be random. If there exists $\beta \in \mathbb{R}^p$ s.t.*

$\mathbb{E}(Y_1|X_1) = X_1^\top \beta$ *and if* Σ_X *is positive definite then*

$$\Phi = \underset{\phi \in \mathbb{R}^p}{\operatorname{argmax}} \frac{\phi^\top \Sigma_{XY} \Sigma_{YX} \phi}{\phi^\top \Sigma_X \phi} \tag{2.3.1}$$

$$= \underset{\phi \in \mathbb{R}^p}{\operatorname{argmax}} \left(\left(\frac{\phi}{\|\phi\|_{\Sigma_X}} \right)^\top \Sigma_{XY} \right)^2 \tag{2.3.2}$$

$$= \operatorname{Ker}(\Sigma_{XY}\Sigma_{YX} - \Sigma_X \lambda_{\max}) \tag{2.3.3}$$

where $\lambda_{\max} = \max_{\phi \in \mathbb{R}^p} \frac{\phi^\top \Sigma_{XY} \Sigma_{YX} \phi}{\phi^\top \Sigma_X \phi}$ *and* $\Phi = \mathcal{D}(\beta)$.

The proof is given in Appendix. Theorem 1 identifies Φ as the solution set of a generalized Rayleigh quotient (equations (2.3.1) and (2.3.2)) or equivalently as the generalized eigenspace associated with the maximum generalized eigenvalue of the matrix pair $(\Sigma_{XY}\Sigma_{YX}, \Sigma_X)$ (equation (2.3.3)).

This result is a general property of the geometry of linear models that, to the best of our knowledge, has not yet been pointed out as such in the literature. We still mention that a somewhat similar equation is given in the context of partial least squares in Sun, Ji and Ye (2009).

The literature sometimes suggests solving specific GEP via various least squares programs (*e.g.* Moghaddam et al. (2006), Sun et al. (2009), Golub and Van Loan (2013)). One original aspect of the present chapter lies in the fact that we propose to estimate the parameters of linear models, a task that is commonly achieved via least squares, through a Generalized Eigenvalue Problem.

In addition to the linearity of the conditional expectation, the only assumption required by Theorem 1 is the positive definitiveness of Σ_X. This assumption is well known to be an identification condition for β (*e.g.,* Florens, Marimoutou, Peguin-Feissolle, Perktold and Carrasco (2007)).

2.4 Estimation under sparsity

2.4.1 Directional Best Subset Selection

We introduce the *directional best subset selection*, prove that its solution has the same direction as the solution of the *best subset selection*, and prove that the latter can be recovered by multiplying the former with an estimation of the magnitude, ν, and connect the rate of convergence of the two estimators. Let $\hat{\Sigma}_X := X^\top X/n$, $\hat{\Sigma}_{XY} := X^\top Y/n$ and $\hat{\Sigma}_{YX} := (\hat{\Sigma}_{XY})^\top$ be estimators of Σ_X, Σ_{XY} and Σ_{YX}, respectively.

We consider the empirical version of problem (2.3.1) to estimate Φ over a sparsity constraint, that is, problem (2.1.3). The fact that its solution set, $\hat{\Phi}$, finds the direction of the solution of the best subset selection is established in the following proposition.

Proposition 1 *Let $Y \in \mathbb{R}^n$ and $X \in \mathbb{R}^{n\times p}$ be random functions. For a given $k \in \mathbb{N}$, let $\tilde{\beta}$ and $\hat{\Phi}$ be as in problems (2.1.1) and (2.1.3) respectively. If $\tilde{\beta}$ is unique almost surely, then conditionally on the set of events where $\min_{v\in\mathbb{R}^p,\|v\|_0\leq k} v^\top \hat{\Sigma}_X v > 0$ we have $\mathcal{D}(\tilde{\beta}) = \hat{\Phi}$ almost surely.*

The proof is given in the Appendix and leads to the following Corollary.

Corollary 1 *Suppose the assumptions of Proposition 1 hold true. Let $\tilde{\phi} \in \hat{\Phi}$ and $\tilde{v} = \operatorname{argmin}_{m\in\mathbb{R}} n^{-1}\|Y - X\tilde{\phi}m\|_2^2$ then conditionally on the same set of events as in Proposition 1 we have $\tilde{\beta} = \tilde{\phi}\tilde{v}$ almost surely.*

Proposition 1 and Corollary 1 exemplify how Theorem 1 can be employed to extend the class of estimators of the high-dimensional sparse linear model (HDSLM) defined in the preceding section to those of high-dimensional sparse GEP. Here, we focus on l_0-constrained estimators, but the proofs and results can be straightforwardly generalized.

The strategy of estimation relies on two steps: First, we estimate Φ via GEP, and second, if the parameter of interest is β, we choose any element in $\hat{\Phi}$, say $\tilde{\phi}$, and then regress Y on the span of $X\tilde{\phi}$ to estimate the magnitude of β. Corollary 1 ensures that if in the first step, the directional best subset selection is employed, this procedure gives an estimator of β (the direction times the magnitude) that is actually the solution of the best subset selection.

The next proposition expresses the error of estimation of β as a function of the error of the estimation of $\phi^\star$ to connect the minimax rate of convergence of the former with that of the latter.

Proposition 2 *For all $n \in \mathbb{N}$, let $Y \in \mathbb{R}^n$ and $X \in \mathbb{R}^{n\times p}$ be random functions* iid *across rows. For all $n \in \mathbb{N}$, assume $\Sigma_X := \mathbb{E}(X_1 X_1^\top)$ is positive definite, that there exists $\beta \in \mathbb{R}^p$ s.t. $\mathbb{E}(Y_1|X_1) = X_1^\top\beta$, $\|\beta\|_0 = s$. Let $\phi^\star := \beta/\|\beta\|_2$ and assume $\hat{\phi} \in \mathbb{R}^p$ is an estimator of $\phi^\star$ for which $\langle\hat{\phi},\phi^\star\rangle_2 \geq 0$, $\|\hat{\phi}\|_0 = s$ and $\|\hat{\phi}\|_2 = 1$ hold almost surely and whose convergence is such that $\|\hat{\phi} - \phi^\star\|_2 = O_p(\alpha_n)$ where $\alpha_n = o(1)$ satisfies $(2s/n)^{1/2}/\alpha_n = o(1)$. Finally, let $\hat{v} = \operatorname{argmin}_{m\in\mathbb{R}} \|Y - X\hat{\phi}m\|_2^2$ be an estimator of the magnitude of β.*

Conditionally on the set of events where there are $0 < \underline{\kappa} \leq \overline{\kappa} < \infty$ such that for all $u \in \mathbb{R}^p$ with $\|u\|_2 = 1$, $\|u\|_0 = \max\{2s, p\}$ we have

$$\underline{\kappa} \leq u^\top \hat{\Sigma}_X u \leq \overline{\kappa}, \tag{2.4.1}$$

if $\hat{\beta} \in \mathbb{R}^p$ is an estimator of β that satisfies $\hat{\beta} = \hat{\phi}\hat{v}$ almost surely, we have

$$\|\hat{\beta} - \beta\|_2^2 = c\|\hat{\phi} - \phi^\star\|_2^2 + O_p(\alpha_n^2).$$

where $c = O_p(1)$.

The proof is given in Appendix. The proposition provides sufficient conditions to ensure that the l_2 convergence rate of $\hat{\beta}$ is one of the l_2 convergences of the direction estimator $\hat{\phi}$.

This result is important since it implies that the minimax rate of convergence in the l_2 loss of the directional best subset selection (2.1.3) is the same as that of the best subset selection (2.1.1). Raskutti et al. (2011) have proven that in a Gaussian and fixed design (where the matrix X is assumed to be nonrandom), if a version of assumption (2.4.1) holds, the minimax rate of convergence of $\|\tilde{\beta} - \beta\|_2$ is $(s \log(p/s)/n)^{1/2}$. In the next section, we propose algorithms that *aim* to solve problem (2.1.3) and to achieve this statistical rate of convergence with high probability.

We note that we assumed $\|\hat{\phi}\|_0 = s$, the true sparsity level of $\theta^\star$, in order to parallel the setting of Raskutti et al. (2011). By a slight adjustment in the statement of assumption (2.4.1) and in the proof, we can replace this assumption by $\|\hat{\phi}\|_0 = k$ for any $k \in \{1, ..., p\}$.

2.4.2 Algorithms for Directional Best Subset Selection

We aim to solve problem (2.1.3). However, this problem is NP-hard and hence cannot be exactly solved in polynomial time unless $P = NP$. Nevertheless, Tan et al. (2018) propose an algorithm that approximates the solution of GEP over a sparsity constraint when $p > n$. In essence, their proposition consists of iteratively updating the estimated vector in its ascent direction, normalizing it, truncating it such that only the k coefficients with the highest absolute magnitude are nonzero and finally normalizing it again. The resulting algorithm applied to the problem (2.1.3) is detailed here.

Notation 1 *When we write Truncate$(v, \mathcal{K})$ for $v \in \mathbb{R}^p$ and $\mathcal{K} \subset \{1, ..., p\}$ we refer to the vector of $\mathbb{R}^p$ defined by*

$$\big(\mathit{Truncate}(v, \mathcal{K})\big)_i = \begin{cases} v_i & \text{if } i \in \mathcal{K} \\ 0 & \text{else} \end{cases}$$

for all $i = 1, ..., p$.

Algorithm 1 Truncated Rayleight flow method (RIFLE)

Input*: matrices $\hat{\Sigma}_{XY}\hat{\Sigma}_{YX}$ and $\hat{\Sigma}_X$, initial vector ϕ_0, sparsity level $k \in \{1, ..., p\}$ and step size η.*

START:

1. *$\mathcal{F}_0 =$ the set of index of the k highest elements of $|\phi_0|$;*
2. *$\phi_0 =$ Truncate$(\phi_0, \mathcal{F}_0)$;*
3. *Let $t = 1$ and repeat the following steps until convergence:*

a) $\rho_{t-1} = (\phi_{t-1}^{\top}\hat{\Sigma}_{XY}\hat{\Sigma}_{YX}\phi_{t-1})/(\phi_{t-1}^{\top}\hat{\Sigma}_{X}\phi_{t-1})$;

b) $C_{t-1} = I_{p\times p} + (\eta/\rho_{t-1})(\hat{\Sigma}_{XY}\hat{\Sigma}_{YX} - \rho_{t-1}\hat{\Sigma}_{X})$;

c) $\phi'_t = C_{t-1}\phi_{t-1}/\|C_{t-1}\phi_{t-1}\|_2$;

d) $\mathcal{F}_t =$ *the set of index of the* k *highest elements of* $|\phi'_t|$;

e) $\hat{\phi}_t =$ *Truncate*$(\phi'_t, \mathcal{F}_t)$;

f) $\phi_t = \hat{\phi}_t/\|\hat{\phi}_t\|_2$;

g) $t = t + 1$;

END

output:ϕ_t

Tan et al. (2018) provides the computational complexity of the algorithm, that is, the sum between two terms: First, $O(p)$ for selecting the k largest elements of a p dimensional vector and second, $O(kp)$ for taking the product between a truncated vector and a matrix with columns restricted to the set F_t and for computing the difference between two matrices with columns restricted to set F_t (Tan et al., 2018).

The parameter k is understood as a regularization (or tuning) parameter that is chosen.

Our experience with Algorithm 1 leads us to believe that when η is small (as suggested by the theory in Tan et al. (2018)), the variable selection operated by the algorithm is highly dependent on the ordering of the absolute value of the element of ϕ_0. In many situations, this constrains the quality of the estimation. To circumvent this feature, we introduce a variation of Algorithm 1 that we call Loaded-RIFLE (LR), and that is described here.

Algorithm 2 Loaded truncated Rayleight flow method (LR)

Input*: matrices* $\hat{\Sigma}_{XY}\hat{\Sigma}_{YX}$ *and* $\hat{\Sigma}_X$*, initial vector* ϕ_0*, initial and final sparsity level* $\overline{k}, \underline{k} \in \{1, ..., p\}$ *with* $\overline{k} > \underline{k}$*, step size* η *and number of iteration* T*.*

START

1. $\mathcal{F}_{0,\overline{k}} =$ *the set of index of the* $\underline{k}$ *highest elements of* $|\phi_0|$;
2. $\phi_{0,\overline{k}} =$ *Truncate*$(\phi_0, \mathcal{F}_{0,\overline{k}})$;
3. *Let* $k = \overline{k}$;
4. *WHILE* $k > \underline{k}$ *repeat the following steps:*
   a) Let $t = 1$
   b) WHILE $t \leq T$ *repeat the followings steps*
      i. $\rho_{t-1,k} = (\phi_{t-1,k}^{\top}\hat{\Sigma}_{XY}\hat{\Sigma}_{YX}\phi_{t-1,k})/(\phi_{t-1,k}^{\top}\hat{\Sigma}_{X}\phi_{t-1,k})$;

ii. $C_{t-1,k} = I_{p\times p} + (\eta/\rho_{t-1,k})(\hat{\Sigma}_{XY}\hat{\Sigma}_{YX} - \rho_{t-1,k}\hat{\Sigma}_X)$;
iii. $\phi'_{t,k} = C_{t-1,k}\phi_{t-1,k}/\|C_{t-1,k}\phi_{t-1,k}, k\|_2$;
iv. $\mathcal{F}_{t,k}$ = *the set of index of the* k *highest elements of* $|\phi'_t|$;
v. IF $Card\big(\mathcal{F}_{t,k} \cup \mathcal{F}_{t-1,k}\big) > \overline{k}$
  A. $\phi_{T,k} = \phi_{0,k}$
  B. set $t = T + 1$ *(BREAK the while loop on* t*);*
vi. ELSE
  A. $\hat{\phi}_{t,k} = Truncate(\phi'_{t,k}, \mathcal{F}_{t,k})$;
  B. $\phi_{t,k} = \hat{\phi}_{t,k}/\|\hat{\phi}_{t,k}\|_2$;
  C. $t = t + 1$;

c) $\phi_{0,k-1} = \phi_{T,k}$;

d) $k = k - 1$;

5. *Apply Algorithm 1 with* $\phi_0 = \phi_{t,\underline{k}+1}$, $k = \underline{k}$, η;

END

Output ϕ_t.

Loaded-RIFLE can be seen as a kind of backward stepwise rifle. It basically runs RIFLE for $k = \overline{k}$ and then uses the output as input of a new run of RIFLE with $k = \overline{k} - 1$. The algorithm proceeds as such until $k = \underline{k}$. The only difference lies in the IF condition in step 4.b)v., which ensures a somewhat smooth update of the set of selected variables. This step is important to extend the theory of RIFLE developed by Tan et al. (2018) to Loaded-RIFLE.

2.4.3 Consistency of the estimation and the rate of convergence

By using the techniques employed in Tan et al. (2018) to study the convergence of Algorithm 1, we derive a nonasymptotic bound on the l_2 distance between the output of Algorithm 2 and the population target vector (which is $\phi^\star$ in this chapter). Based on this result, we derive the convergence rate of the estimator in the sub-Gaussian setting. Before stating the theorems, we introduce more notation where some are specific to GEP and others are nonstandard but useful in the context of this chapter.

For any $Z \in \mathbb{R}^{p\times p}$ and $k \in \mathbb{N}$ s.t. $k \leq p$, let

$$\rho(Z, k) := \sup_{\|v\|_{\mathbb{R}^p}, \|v\|_0 \leq k} |v^\top Z v|.$$

For fixed k, $\rho(\cdot, k) : \mathbb{R}^{p\times p} \to \mathbb{R}$ is a norm on $\mathbb{R}^{p\times p}$ over $\mathbb{R}$, if Z is symmetric then $\rho(Z, k)$ is the supremum on a set of maximum eigenvalues of sub-matrices of Z.

For any $k \in \mathbb{N}$ such that $k \leq p$ let $\mathcal{F}(p,k) := \{F \subset \{1,...,p\}; \text{Card}(F) = k\}$ be the set of sets containing k different integers between 1 and p. Moreover let $\{e_i\}_{i=1}^p$ be the canonical basis in $\mathbb{R}^p$ (i.e. e_i has zero's on all its elements except on the i^{th} position which contains a one) and let

$$\mathcal{P}(\mathcal{F}(p,k)) := \{P_{(F)} := (e_{\alpha_1},...,e_{\alpha_k}) \in \mathbb{R}^{p\times k} | \\ F \in \mathcal{F}(p,k); \alpha_i \in F; \alpha_1 < \alpha_2 < ... < \alpha_k\}$$

be a set of orthogonal projectors from $\mathbb{R}^p$ to its k dimensional subspaces that are spanned by the subsets of $\{e_1,...,e_p\}$ whose cardinalities equal k and that preserve the order of the indexes. This set is useful since if $v \in \mathbb{R}^p$ such that $\|v\|_0 = k$ then $\mathcal{P}(\mathcal{F}(p,k))$ contains a matrix P such that $w = P^\top v$, $w \in \mathbb{R}^k$ and w is equal to v if the former is embodied in $\mathbb{R}^p$. Observe we have that $P^\top P = I_{k\times k}$, $\|PP^\top\|_{\text{spec}} = 1$ and $\|P\|_{\text{spec}} = 1$, where for any $Z \in \mathbb{R}^{p\times p}$, $\|Z\|_{\text{spec}} := \max_{\|v\|_2=1} \|Zv\|_2$ is the *spectral norm* of Z.

For any $m \in \mathbb{N}$ and $Z \in \mathbb{R}^{m\times m}$, let $\lambda_{\max}(Z), \lambda_{\min}(Z)$ and $\lambda_i(Z)$ be respectively the maximum, minimum and i^{th} (in descending order) *eigenvalues* of Z. If Z is positive definite, let $\kappa(Z) = \lambda_{\max}(Z)/\lambda_{\min}(Z)$ be the *condition number* of Z (*e.g.* Golub and Van Loan (2013)).

For any $k \in \mathbb{N}$ s.t. $k \leq p$ and any $F \in \mathcal{F}(k,p)$, let λ_i, $\lambda_i(F)$, $\hat{\lambda}_i$ and $\hat{\lambda}_i(F)$ be the i^{th} (in descending order) *generalized* eigenvalues of respectively the pairs of matrices $(\Sigma_{XY}\Sigma_{YX}, \Sigma_X)$, $(P_{(F)}^\top \Sigma_{XY}\Sigma_{YX}P_{(F)}, P_{(F)}^\top \Sigma_X P_{(F)})$, $(\hat{\Sigma}_{XY}\hat{\Sigma}_{YX}, \hat{\Sigma}_X)$ and finally $\left(P_{(F)}^\top \hat{\Sigma}_{XY}\hat{\Sigma}_{YX}P_{(F)}, P_{(F)}^\top \hat{\Sigma}_X P_{(F)}\right)$.

For any $p \in \mathbb{N}$ and any symmetric matrices $A, B \in \mathbb{R}^{p\times p}$, let

$$Cr(A,B) = \min_{\|v\|_2=1} \left(\left(v^\top Av\right)^2 + \left(v^\top Bv\right)^2 \right)^{1/2}$$

be the *Crawford number* of (A,B). Stewart (1979) shows that if $Cr(A,B) > 0$ the GEP associated with (A,B) is *definite*, namely it has a complete system of generalized eigenvectors and all its generalized eigenvalues are real.

For any $k \in \mathbb{N}$ such that $k \leq p$ define

$$\Delta(k) := \inf_{P \in \{\cup_{i=1}^k \mathcal{P}(\mathcal{F}(p,i))\}} Cr(P^\top \Sigma_{XY}\Sigma_{YX}P, P^\top \Sigma_X P).$$

and

$$\varepsilon(k) := \left(\rho(\hat{\Sigma}_{XY}\hat{\Sigma}_{YX} - \Sigma_{XY}\Sigma_{YX}, k)^2 + \rho(\hat{\Sigma}_X - \Sigma_X, k)^2 \right)^{1/2}.$$

In the next results, we have employed the techniques developed in Tan et al. (2018) to provide a nonasymptotic bound on the l_2 error of Algorithm 2.

Theorem 2 *For all $n \in \mathbb{N}$, let $Y \in \mathbb{R}^n$ and $X \in \mathbb{R}^{n\times p}$ be random functions* iid *across rows. For all $n \in \mathbb{N}$ assume $\Sigma_X := \mathbb{E}(X_1 X_1^\top)$ is positive definite and there exists $\beta \in \mathbb{R}^p$ s.t. $\mathbb{E}(Y_1|X_1) = X_1^\top \beta$ and $\|\beta\|_0 = s$. Let $\phi^\star = \beta/\|\beta\|_2$ and $\Lambda := \lambda_1/(1+\lambda_1^2)$.*

Let $\underline{k}$ in Algorithm 2 be such that $\underline{k} = C's$ for $C' > 1$ and let $\overline{k} = C''\underline{k}$ for $C'' > 1$ and $k' = \overline{k} + s$.

Let $\phi_{t,\underline{k}}$ be the output of Algorithm 2 (i.e. after t iterations of the last step of the algorithm). Suppose that $t \geq T$ where T is like in Algorithm 2. Let ϕ_0 be the initial vector in Algorithm 2 after truncation (as in step 2.).

Choose a constant $c > 0$ and η in Algorithm 2 such that

$$\eta\lambda_{\max}(\Sigma_X) < 1/(1+c), \tag{2.4.2}$$

and

$$\nu = (1+\sqrt{\frac{s}{\underline{k}}})\left(1 - \frac{1+c}{8}\eta\lambda_{\min}(\Sigma_X)\frac{1}{C_\diamond\kappa(\Sigma_X)}\right)^{1/2} < 1 \tag{2.4.3}$$

hold true, where $C_\diamond = (1+c)/(1-c)$ and $\kappa(\Sigma_X)$ is the condition number of Σ_X.

Then, conditionally on the set of events where

$$\frac{\varepsilon(k')}{\Delta(k')} < \frac{\Lambda(1-\nu)\sqrt{\xi}}{4\sqrt{10} + \Lambda(1-\nu)\sqrt{\xi}}, \tag{2.4.4}$$

$$\rho(\hat{\Sigma}_X - \Sigma_X, k') \leq c\lambda_{\min}(\Sigma_X) \tag{2.4.5}$$

$$|\langle \phi^\star, \phi_0\rangle_2| \leq 1 - \nu^2\xi, \tag{2.4.6}$$

hold almost surely with ξ defined as

$$\xi := \min\left\{\frac{1}{8C_\diamond\kappa(\Sigma_X)}, \frac{1}{30(1+c)C_\diamond^3\eta\lambda_{\max}\kappa^3(\Sigma_X)}\right\},$$

we have

$$\sqrt{1 - |(\phi^\star)^\top\phi_{t,\underline{k}}|} \leq \nu^t\sqrt{\xi} + \frac{1}{(1-\nu)^2}\sqrt{10}\frac{\sqrt{2}}{\Lambda\big(\Delta(k') - \varepsilon(k')\big)}\varepsilon(k')$$

almost surely.

Since the theorem is an extension to Algorithm 2 of a known result given in Tan et al. (2018) for Algorithm 1, we give a complete proof in Supplementary Material. We note that this proof provides as a byproduct a version of the theorem for Algorithm 1 that fixes the one given in Tan et al. (2018).

Theorem 2 indeed allows one to bound the l_2 norm between $\phi^\star$ and ϕ_t since if $\langle \phi^\star, \phi_t\rangle_2 \geq 0$, then $\sqrt{1 - |(\phi^\star)^\top\phi_t|} = 1/\sqrt{2}\|\phi^\star - \phi_t\|_2$.

The first term on the right-hand side, namely, $\nu^t\sqrt{\xi}$, is the optimization error. The second term on the right-hand side is the statistical error (Tan et al., 2018).

Assumptions (2.4.4) and (2.4.5) control for the error of estimation of the (generalized) spectrum of the (pair of) matrices at hand. In particular, they provide upper and lower bounds on the condition number of Σ_X, on the spectrum of submatrices of $\hat{\Sigma}_X$ and on $\hat{\lambda}_i(F)$ for any $F \in \mathcal{F}(p, k')$ and $i = 1, ..., k'$; Tan et al. (2018) contains the details. As we show later, if Y_1 and X_1 are sub-Gaussian, there exist ratios between n, $p(n)$ and $s(n)$ such that the probability of the events where these assumptions do not hold converges to zero as $n \to \infty$.

Theorem 2 also requires theoretical conditions on the input parameters of Algorithm 2. Inequality (2.4.2) and (2.4.3) are technical constraints on the value of the *step size* η. Equation (2.4.6) requires the cosine between the (population) target vector $\phi^\star$ and the input vector ϕ_0 to be sufficiently high. We discuss this assumption in the last paragraph of this section.

To obtain a convergence rate and to give explicit constraints on the asymptotic behavior of $p(n), s(n)$ and $k(n)$, we provide the following result.

Proposition 3 *For every $n \in \mathbb{N}$, let $Y \in \mathbb{R}^n$ and $X \in \mathbb{R}^{n\times p}$ be random functions* iid *across rows where $Y_1, X_{1,1}, ..., X_{1,p(n)}$ are sub-gaussian random variables. For all $n \in \mathbb{N}$ assume $\Sigma_X := \mathbb{E}(X_1 X_1^\top)$ is positive definite and there exists $\beta \in \mathbb{R}^p$ s.t. $\mathbb{E}(Y_1|X_1) = X_1^\top\beta$. Let $\phi^\star = \beta/\|\beta\|_{\mathbb{R}^p}$, $s = \|\beta\|_0$, $\hat{\Sigma}_X = X^\top X/n$ and $k' = Cs$ for $C > 1$.*

Assume $s\log(p/s)/n = o(1)$ and $s/p = o(1)$ then

$$\varepsilon(k') = O_p(\sqrt{\frac{s\log(p/s)}{n}})$$

and

$$\rho(\hat{\Sigma}_X - \Sigma_X, k') = O_p(\sqrt{\frac{s\log(p/s)}{n}})$$

Proposition 3 implies that if ϕ_t is the output of Algorithm 2 after t iterations of step 5, then for an appropriate choice of η and ϕ_0, $|(1-|(\phi^\star)^\top\phi_t|)^{1/2} - \nu^t\sqrt{\xi}| = O_p((s\log(p/s)/n)^{1/2})$. Moreover, by assuming that t is a *fast enough* (we give a precise sense to this notion in the next result) function of n and that $(\phi^\star)^\top\phi_t \geq 0$, we have $\|\phi^\star - \phi_t\|_2 = O_p((s\log(p/s)/n)^{1/2})$. As we mention above, by combining Theorem 1 in Raskutti et al. (2011) with Proposition 2, we know that in the Gaussian fixed design, $(s\log(p/s)/n)^{1/2}$ is the minimax rate of convergence for the l_0 constrained estimation of $\phi^\star$.

One can obtain from ϕ_t an estimate of the magnitude of β by regressing Y on the span of $X\phi_t$. Let $\hat{\upsilon}$ be this estimator and $\hat{\beta}$ be the estimator of β obtained as the product of ϕ_t and $\hat{\upsilon}$. Proposition 3 tells us that for $t = t(n)$

fast enough $\|\hat{\beta} - \beta\|_2 = O_p\big((s\log(p/s)/n)^{1/2}\big)$, which is minimax (Raskutti et al., 2011).

Theorem 2 relies heavily on the quality of the initial vector ϕ_0. To obtain a good initial vector, we follow the same heuristic as in Tan et al. (2018), that is, by using a solution of an l_1 relaxation of the l_0-constrained problem as the initial vector, which is the LASSO for problem (2.1.1) (Bertsimas et al. (2016), Hastie et al. (2020)). This choice of initialization makes particular sense since the LASSO is known to select too many covariates when tuned for predictive accuracy. Using its (normalized) solution as the initial vector for Algorithm 2 allows us to reduce the size of the selected support while improving the accuracy of the estimation. As the Monte Carlo experiment in Section 5 demonstrates, this procedure outperforms the other l_0-constrained estimators considered in this chapter. This initialization procedure is theoretically justified by the fact that it has been established in Chetverikov et al. (2021) that the l_2 error of the cross-validated LASSO is $O_p\big((s\log(p/s)/n)^{1/2} \times \log(pn)\big)$. This error ensures that the events on which the assumption on ϕ_0 in Theorems 2 is not verified have a probability shrinking to zero.

Unfortunately, we are not aware of any method to select η that has an analog quality. Experience and Assumption (2.4.2) Tan et al. (2018) suggest selecting η small enough such that $\eta\lambda(\hat{\Sigma}_X) < 1$.

2.4.4 Central Limit Theorem

In this section, we provide a central limit theorem for the output of Algorithm 2. In the discussion that follows the results, we state three propositions that analyze in depth the behavior of the last step of Algorithm 2, that is, Algorithm 1. One of these propositions gives sufficient conditions for the convergence of a stochastic Cauchy product and may be of independent interest.

Before stating the results, we give the following definition that is necessary for the statement and proof of the theorem.

Definition 2 *Given $\phi^\star$ and p, let*

$$\mathcal{Q}^\star := \{Q = PP^\top | P \in \cup_{i=s}^{k} \mathcal{P}\big(\mathcal{F}(p,i)\big), Q\phi^\star = \phi^\star\}.$$

Moreover, for any ϕ_t that is the output of Algorithm 1 or 2, let

$$Q(\phi_t) = Q_t = P_t P_t^\top$$

where P_t is the element of $\mathcal{P}(\mathcal{F}(p, \|\phi_t\|_0))$ that is such that if $w = P_t^\top \phi_t$ then w is equal to ϕ_t if the former is embodied in $\mathbb{R}^p$.

We note that all ϕ_t s.t. $Q(\phi_t) \in \mathcal{Q}^\star$ have a support containing the support of $\phi^\star$. From Lemma 5 in the Appendix, one can show that in the sub-Gaussian setting, for $t = t(n)$ fast enough, $Q(\phi_t) \in \mathcal{Q}^\star$ with probability tending to one,

i.e., Algorithm 2 possesses the *sure screening property* (Fan and Lv (2008), Guo et al. (2020)).

In the remainder of this section, we make use of the following convention.

Notation 2 *Let $\{a_\alpha\}_{\alpha\in\mathbb{N}}$ be a family of operators on a vector space, if $j < i$ then $\prod_{\alpha=i}^{j} a_\alpha = 1$ where 1 symbolize the unitary operator on the vector space. If $j \geq i$ the symbol $\prod$ is used conventionally.*

We now state the central limit theorem for the output of Algorithm 1. Since the last step of Algorithm 2 is an application of Algorithm 1, the extension of the result to the output of LR is direct.

Theorem 3 *For all $n \in \mathbb{N}$, let $Y \in \mathbb{R}^n$ and $X \in \mathbb{R}^{n\times p}$ be random functions* iid *across rows. For all $n \in \mathbb{N}$ assume $\Sigma_X := \mathbb{E}(X_1 X_1^\top)$ is finite and positive definite and there exists $\beta \in \mathbb{R}^p$ s.t. $\mathbb{E}(Y_1|X_1) = X_1^\top\beta$. Let $U = Y - X\beta$, $\sigma^2 = \mathbb{V}ar(U_1)$, $\phi^\star = \beta/\|\beta\|_2$ and $s = \|\beta\|_0$. Let t, η, ϕ_t and C_{t-1} be as in Algorithm 1. Let $\tilde{\phi}_t = C_{t-1}\phi_{t-1}$, let t be an increasing function of n and assume that $\langle\phi^\star, \phi_t\rangle_2 \geq 0$ almost surely.*

If for some constants $\xi > 0$ and $\nu < 1$

$$\|\phi^\star - \phi_t\|_2 \leq \nu^t\sqrt{\xi} + O_p\Big(\sqrt{\frac{s\log(p/s)}{n}}\Big) \tag{2.4.7}$$

and if $\inf_{\|v\|_2=1,\|v\|_0=k} |v^\top\Sigma_{XY}| > 0$, and $\log(n)/t(n) = o(1)$, then conditionally on X, there exists $Q^\star \in \mathcal{Q}^\star$ and $K > 0$ such that if $\eta < K$ then for all $w \in \mathbb{R}^p$ such that $\|w\|_2 = 1$ and $\|w\|_0 \leq n$ we have that

$$\sqrt{n}\langle w, \phi_t - \phi^\star + \gamma_n\rangle_2$$

is

$$\mathcal{AN}\Big(0, w^\top[(I - Q^\star) + \eta\hat{\Sigma}_X Q^\star]^{-1}\eta^2\hat{\Sigma}_X[(I - Q^\star) + \eta\hat{\Sigma}_X Q^\star]^{-1}w\frac{\sigma^2}{\|\beta\|_2^2}\Big|X\Big)$$

and $\|\gamma_n\|_2 = o_p(1)$, where $\gamma_n := \kappa_n + \tau_n$ with

$$\kappa_n = \sum_{m=0}^{t-1}\Big(\prod_{l=1}^{m}(I_{p\times p} - \eta\hat{\Sigma}_X)Q_{t-l}\Big)(\phi_{t-l} - \tilde{\phi}_{t-l}),$$

and

$$\tau_n = \Big[\Big(\sum_{m=0}^{t-1}\Big(\prod_{l=1}^{m}(I_{p\times p} - \eta\hat{\Sigma}_X)Q_{t-l}g_{t-1-m}\Big)\big((I - Q^\star) + \eta\hat{\Sigma}_X Q^\star\big) - I_{p\times p}\Big]\phi^\star$$

where for all $m \in \mathbb{N}$

$$g_{t-1-m} = \frac{\phi_{t-1-m}^\top\hat{\Sigma}_X\phi_{t-1-m}}{\phi_{t-1-m}^\top\hat{\Sigma}_X\phi^\star + \phi_{t-1-m}^\top\frac{X^\top U}{\|\beta\|_2 n}}.$$

The proof of the result is given in Appendix. The bias γ_n is the sum of κ_n, which is observable (and hence can be corrected), and τ_n, which cannot be observed. In the proof, we show that both terms converge to zero but without being able to pin down their rates of convergence. From Assumption (2.4.7) (which can be shown to hold in the sub-Gaussian setting by combining Proposition 3 with Proposition 7 in the Supplementary Material), we conclude that γ_n must converge to zero at a slower rate than $n^{-1/2}$. The term κ_n accounts for the effect of the normalization in step 3.c) and from the truncation in step 3.e) of Algorithm 1. The term τ_n is mainly driven by the difference between the supports of ϕ_t and $\phi^\star$ and can be understood as a bias introduced by the regularization of the inverse problem posed by the ill-conditioned matrix $\hat{\Sigma}_X$ in regard to estimating $\phi^\star$ in high dimensions.

The keys to the proof of Theorem 3 are three propositions that we now state and comment on. The proofs of the three propositions are given in the Appendix.

The first proposition decomposes the estimator ϕ_t to clarify its behavior.

Proposition 4 *For all $n \in \mathbb{N}$, let $Y \in \mathbb{R}^n$ and $X \in \mathbb{R}^{n\times p}$ be random functions* iid *across rows. For all $n \in \mathbb{N}$ assume $\Sigma_X := \mathbb{E}(X_1 X_1^\top)$ is positive definite and there exists $\beta \in \mathbb{R}^p$ s.t. $\mathbb{E}(Y_1|X_1) = X_1^\top \beta$. Let $\phi^\star = \beta/\|\beta\|_2$ and $s = \|\beta\|_0$. Let t, η, ϕ_t, C_{t-1} and ϕ_0 be as in Algorithm 1. Finally, let $\tilde{\phi}_t = C_{t-1}\phi_{t-1}$. Then for every $t \geq 2$*

$$\begin{aligned}\phi_t =& \left(\prod_{m=1}^{t} \Gamma_{t-m}\right)\phi_0 + \sum_{m=0}^{t-1}\left(\prod_{l=1}^{m}\Gamma_{t-l}\right)(\phi_{t-l} - \tilde{\phi}_{t-l}) \\ &+ \eta \sum_{m=0}^{t-1}\left(\prod_{l=1}^{m}\Gamma_{t-l}\right)\frac{\phi_{t-1-m}^\top \hat{\Sigma}_X \phi_{t-1-m}}{\phi_{t-1-m}^\top \hat{\Sigma}_X \phi^\star + \phi_{t-1-m}^\top \frac{X^\top U}{\|\beta\|_{\mathbb{R}^p} n}}\left(\hat{\Sigma}_X \phi^\star + \frac{X^\top U}{\|\beta\|_2 n}\right)\end{aligned}$$

almost surely, where $\Gamma_i = (I_{p\times p} - \eta\hat{\Sigma}_X)Q_i$ for any $i \in \mathbb{N}$.

The first term explains the numerical dependence of ϕ_t on ϕ_0. From Lemma 1 in the Appendix, one can show that there exists η, making this dependence go to zero in probability as $t \to \infty$. The second term, which is a Cauchy product, explains the impact of the normalization in step 3.c) and of the truncation in step 3.e) of Algorithm 1. Using Proposition 5 below to prove that this term is $o_p(1)$ is a key step in the proof of Theorem 3. The third term, which is also a Cauchy product, is shown to converge to $\phi^\star$ by using Proposition 6.

The next proposition, which might be of independent interest, gives conditions for the convergence in probability of a stochastic Cauchy product.

Proposition 5 *Let $\{a_i\}_{i=0}^m$ and $\{b_i\}_{i=0}^m$ be two sequences of* positive *random variables such that $m = m(n)$ where $m \to \infty$ as $n \to \infty$. If $\{a_i\}_{i=0}^m =$*

$o_p(1)$, $\sup_{i\in\{0,\dots,m\}} |a_i| = O_p(1)$ *and it exists* $M \in \mathbb{R}$ *such that we have* $\lim_{n\to\infty} \sum_{i=0}^{m(n)} b_i = M$ *almost surely.*

Then $\sum_{i=0}^{m} a_{m-i} b_i = o_p(1)$.

The following and last proposition studies the convergence of the Cauchy product in the last term of the decomposition of Proposition 4.

Proposition 6 *Under the same assumptions than in Theorem 3, conditionally on* X*, there exists* $\eta > 0$ *and* $Q \in \mathcal{Q}^\star$ *such that*

$$\sum_{m=0}^{t-1} \left(\prod_{l=1}^{m} \Gamma_{t-l} \right) \frac{\phi_{t-1-m}^\top \hat{\Sigma}_X \phi_{t-1-m}}{\phi_{t-1-m}^\top \hat{\Sigma}_X \phi^\star + \phi_{t-1-m}^\top \frac{X^\top U}{\|\beta\|_{\mathbb{R}^p} n}} = [(I - Q^\star) + \eta \hat{\Sigma}_X Q^\star]^{-1} + o_p(1)$$

where $\Gamma_m = (I_{p\times p} - \eta \hat{\Sigma}_X) Q_m$ *for any* $m \in \mathbb{N}$.

This proposition is the key to treating the third term of Proposition 4 and thus obtaining the exact asymptotic variance given in Theorem 3.

2.5 Monte Carlo experiments

In our experiments, we use Algorithm 2 initialized by a solution of LASSO to estimate $\phi^\star$. From this estimator, we calculate an estimate of β by using the strategy described in Section 4.1.

We compare the performances of LR against those of three well-studied l_0-constrained estimators of β that we normalized to find estimators of $\phi^\star$. The competitors are the IHT (Blumensath and Davies (2008), Blumensath and Davies (2009*b*), Jain et al. (2014)) that we implement as in (Blumensath and Davies, 2009*a*), the two-stage IHT Jain et al. (2014) and the SDAR Huang et al. (2018).

2.5.1 Implementation

In each experiment we let $n = 100, p = 1000$ and generate $X \in \mathbb{R}^{n\times p}$ *iid* across rows such that $X_1 \sim \mathcal{N}_p(0, \Sigma)$ and $Y|X \in \mathbb{R}^n$ such that $Y \sim \mathcal{N}_n(X\beta, \beta^\top \Sigma_X \beta / m)$ where m is the *signal-to-noise ratio* (SNR) and β is generated such that $\|\beta\|_0 = s = 10$ with every element of β having an equal probability to be assigned a value of 1. We consider four scenarios that go from the least correlated design to a highly correlated design.

Scenario 1 *We took* $\Sigma = I_{p\times p}$.

Scenario 2 *We took* $\Sigma_{i,j} = 0.5^{|i-j|}$.

Scenario 3 *We let* $W \in \mathbb{R}^{p\times p}$ *such that* $W_{i,j} = 0.5^{|i-j|}$ *and took* $\Sigma = U^\top DU$ *where* U *is the matrix of eigenvectors of* W*,* $D_{j,j} = 10/\sqrt{j}$ *and* $D_{i,j} = 0$ *when* $i \neq j$*.*

Scenario 4 *We let* $W \in \mathbb{R}^{p\times p}$ *such that* $W_{i,j} = 0.5^{|i-j|}$ *and took* $\Sigma = U^\top DU$ *where* U *is the matrix of eigenvectors of* W*,* $D_{j,j} = 10/j$ *and* $D_{i,j} = 0$ *when* $i \neq j$*.*

As pointed out in Hastie et al. (2020), the performances of estimators vary greatly with the SNR, and it may happen that one estimator outperforms another for a high SNR while underperforming in low SNR settings. Hence, we consider a range of eleven signal-to-noise ratios for the four scenarios. Since the relative performances of the algorithms often change before and after an SNR level of 1, we choose 5 equally spaced values between 0.05 and 1 and between 1 and 6.

0.05	0.24	0.43	0.62	0.81	1.00	2.00	3.00	4.00	5.00	6.00

Table 2.2: Signal-to-noise ratios considered in the four scenarios.

We fix the sparsity level of the estimated vectors at $k = 12$ for all the estimators. Hence, this experiment assumes that a (good) prior on the sparsity level s is available. This means we do not problematize here the selection of the parameter k. There exist other experimental paradigms where the tuning parameter is selected by cross-validation (as in Hastie et al. (2020) or via BIC-type criteria (as in Huang et al. (2018)).

SDAR and pseudo-LASSO have been initialized with the solution of a ridge regression with tuning parameter $2s\log(p)/n$. LR has been initialized with a solution from LASSO that has approximately $3k = 36$ nonzero components. In addition to k, LR and two-stage IHT both depend on $\overline{k}$, a parameter that plays different technical functions in the two algorithms but that encompasses the same heuristic, namely, preoptimizing on a support of dimension $\overline{k} > k$ before selecting a support of k on which the final optimization takes place (Jain et al., 2014). For both algorithms, we set $\overline{k} = 4k = 48$. For LR, we set the *step-size* $\eta = 0.5/\|\hat{\Sigma}_X\|_{spec}$. For IHT, the step size changes at each iteration following the rules proposed in Blumensath and Davies (2009*a*). For LR, we set $T = \log(n) \approx 5$. Following the theory and practice in Huang et al. (2018), we fed SDAR with X and Y, after dividing them by $\sqrt{n}$.

Let $\hat{\beta}$ be an estimate of β. For each experiment, we record three accuracy metrics.

Estimation metric: measuring the (relative) l_2 error of estimation of $\phi^\star$, namely $\hat{\beta}/\|\hat{\beta}\|_2 - \beta/\|\beta\|_2$.

Selection metric: measuring the recovery of the true support, that is a binary variable taking value 1 if the support of β is contained in the support of $\hat{\beta}$, and 0 otherwise.

Prediction metric: The relative risk (e.g. Bertsimas et al. (2016), Hastie et al. (2020)) that is given by, $\mathbb{E}[X_1^\top \hat{\beta} - X_1^\top \beta]^2 / \mathbb{E}[X_1^\top \beta]^2 = (\hat{\beta} - \beta)^\top \Sigma (\hat{\beta} - \beta) / \beta^\top \Sigma \beta$.

2.5.2 Results

For each scenario and each signal-to-noise ratio, we repeated the experiment 100 times. The results of the *estimation metric* are displayed in Figure 2.2. The results of the *selection metric* are displayed in Table 2.3 for SNRs higher than 2 (for SNRs strictly lower than two, no algorithm was able to recover complete true support at least once). The results for the *prediction metric* are displayed in Table 2.4 for three SNRs.

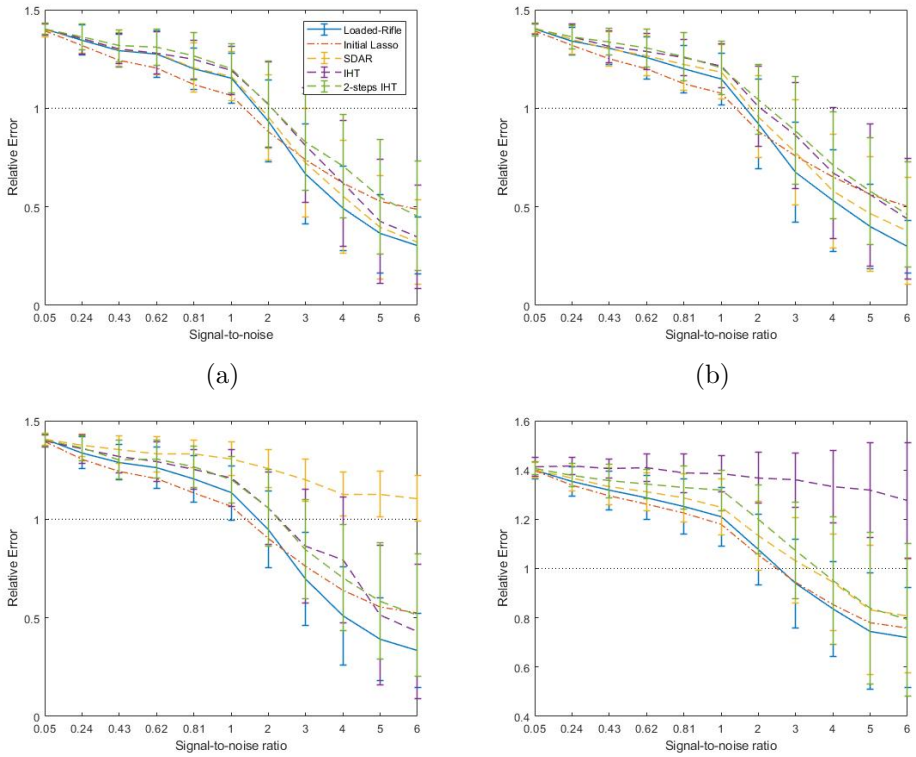

Figure 2.2: Relative error of estimation of the direction

Mean and standard deviation of the relative error of estimation of $\phi^\star = \beta / \|\beta\|_2$ across various signal-to-noise ratios in (a) Scenario 1, (b) Scenario 2, (c) Scenario 3 and (d) Scenario 4. In all the figures, the horizontal dot line represents the score of the null vector.

We draw four main conclusions from the experiment. First, LR initialized with LASSO outperforms its competitors both in terms of variable selection

	Signal-to-noise ratio	2	3	4	5	6
Scenario 1	Loaded-RIFLE	**0.02**	**0.18**	0.4	**0.72**	**0.82**
	SDAR	0.01	0.16	**0.41**	0.68	0.81
	IHT	**0.02**	0.09	0.33	**0.72**	**0.82**
	Two-stages IHT	*	0.01	0.09	0.28	0.43
Scenario 2	Loaded-RIFLE	**0.03**	**0.19**	**0.41**	**0.61**	**0.84**
	SDAR	*	0.11	0.36	0.54	0.69
	IHT	0.01	0.08	0.3	0.47	0.63
	Two-stages IHT	*	0.05	0.1	0.22	0.41
Scenario 3	Loaded-RIFLE	**0.01**	**0.14**	**0.43**	**0.65**	**0.75**
	SDAR	*	*	*	*	*
	IHT	*	0.08	0.16	0.58	0.69
	Two-stages IHT	*	0.05	0.1	0.23	0.36
Scenario 4	Loaded-RIFLE	*	*	0.02	**0.08**	**0.1**
	SDAR	*	*	0.02	0.07	0.03
	IHT	*	*	*	0.01	0.01
	Two-stages IHT	*	*	**0.03**	0.05	**0.1**

Table 2.3: Frequency of support recovery

Frequency at which the support selected by the algorithms contained the true support of β for signal-to-noise ratio greater or equal to 2. The asterisk are zeros. The best result for each scenario are written in bold.

and estimation as well as for prediction in most scenarios. This conclusion is particularly observable for SNRs higher than two and in the more correlated designs (scenarios 3 and 4). Second, for SNRs higher than 2, LR improves on the initial LASSO solution *while* significantly reducing the size of the estimated support. Third, for SNRs lower than 2, no estimator was able to recover the true support at least once. Fourth, for SNRs lower than 2, all the estimators have a relative error higher than one, that is, the score of a null vector.

2.6 Conclusion

We propose to interpret the parameter vector of the high-dimensional sparse linear model as the product between a basis in *the direction of interest*, which is identified as the generalized eigenspace of a pair of measurable matrices, and the magnitude relative to this basis.

On the one hand, this approach sheds new light on the well-studied linear regression while inducing a parameterization that emerges naturally in economic theory. To exemplify this, let $Y \in \mathbb{R}$ and $X \in \mathbb{R}^p$ be random variables such that $\mathbb{E}(Y|X) = X^\top \phi\nu$ where ϕ is some basis in Φ and $\nu \in \mathbb{R}$ is the *magnitude relative to the basis*. As a first example, let Y be the random log-

Signal-to-noise ratio		0.05		1		6	
		mean	sd	mean	sd	mean	sd
Scenario 1	Loaded-RIFLE	15.878	**2.8**	**1.632**	**0.3**	**0.109**	**0.1**
	SDAR	**14.627**	4.5	1.666	0.3	0.145	0.2
	IHT	18.509	3.8	1.895	0.4	0.181	0.2
	Two-stages IHT	15.848	2.9	1.737	0.3	0.255	0.2
Scenario 2	Loaded-RIFLE	16.013	**2.8**	**1.546**	**0.3**	**0.096**	**0.1**
	SDAR	**15.196**	4.6	1.644	0.4	0.181	0.2
	IHT	18.136	4.6	1.938	0.3	0.248	0.2
	Two-stages IHT	16.586	3.7	1.743	0.4	0.245	0.2
Scenario 3	Loaded-RIFLE	15.734	2.8	1.477	0.3	**0.131**	0.1
	SDAR	**4.679**	**0.9**	**1.175**	**0.1**	0.856	**0.1**
	IHT	16.337	4.9	1.830	0.5	0.267	0.3
	Two-stages IHT	16.354	3.1	1.710	0.3	0.312	0.3
Scenario 4	Loaded-RIFLE	**11.691**	**2.2**	**0.962**	**0.2**	**0.258**	**0.1**
	SDAR	14.912	3.3	1.239	0.3	0.336	0.2
	IHT	20.168	10.3	2.895	1.1	1.908	1.0
	Two-stages IHT	20.707	5.3	1.831	0.3	0.381	0.2

Table 2.4: Relative risk of estimation

Mean and standard deviation (sd) of the Relative Risk of the four estimators in the four scenarios for three signal-to-noise ratio. The best results for each scenario are written in bold.

output of a Cobb-Douglass production function, X be the random vector of the quantity of factors of production and ϕ such that $\sum_{j=1}^{p} \phi_j = 1$. Then, the return to scale of the production function is decreasing, constant or increasing when ν is lower, equal to, or higher than 1, respectively. A second example is the *Cambridge Equation* (Pigou (1917)) where Y is the random money demand, X is the random vector of good to be exchanged, ϕ is the vector of prices, and ν the proportion of money detained for convenience by individuals or if equilibrium on the money market is assumed, the inverse of the velocity of money (Fisher and Brown (1911)). Third, we observe that the parametrization is induced by the relativity of prices. In the easiest example of this, let Y be a random quantity of dollars, X be a random vector of the quantity of goods and ϕ be the vector of price expressed in terms of Euro; then, ν is the exchange rate between the two currencies. As it appears from these examples, the choice of the basis $\phi \in \Phi$ should depend on the context to study; for instance, one could choose ϕ such that $\|\phi\| = 1$ for some norm; nevertheless, infinitely many other specifications are available.

On the other hand, our approach connects linear regression with the Generalized Eigenvalue Problem and extends the class of estimators available for the former to the one of the latter. This is especially fruitful in the context

of high dimensions where standard optimal procedures of estimation are not available.

We show how to apply this last idea by tackling the best subset selection problem via a novel estimator called directional best subset selection. We provide a theory connecting the solutions of the two problems as well as the minimax rate of convergence of their l_2 error of estimation. This leads us to approximate the solution of the directional best subset selection via a novel estimator based on RIFLE, which is a well-established algorithm developed by Tan et al. (2018) and designed to solve sparse Generalized Eigenvalue Problems in a high-dimensional context. We demonstrate that this proposition leads to an estimator that has the minimax rate of convergence in l_2 for the sub-Gaussian setting. Finally, we provide a central limit theorem for the estimator. To do so, we demonstrate a theorem that gives sufficient conditions for the convergence of a stochastic Cauchy product, which might be of independent interest. Our simulations show that our method is competitive and can outperform existing competitors.

Our estimator is an answer to the important best-subset selection problem of selecting k out of p covariates given n observations. While our estimator does not equal the best subset selection, it still allows an l_0-*constrained* estimation while attaining the minimax rate of convergence of the best subset selection and providing solutions in less than a second, even for large problems. This is one of the rare propositions in the literature to possess these qualities altogether.

Although this chapter focuses on the l_0 constrained regression and concentrates on the estimator derived from RIFLE, the result of the second section applies or may be generalized more broadly. In particular, other estimators of GEP exist (*e.g.,* Safo, Ahn, Jeon and Jung (2018), Jung, Ahn and Jeon (2019) and Yuan, Shen and Zheng (2019)) and might be of interest, for instance, to solve least squares problems constrained by norms other than l_0. More generally, the connection between least squares programs and GEP opens the possibilities of bridging new numerical techniques into high-dimensional estimation theory.

Finally, in this chapter, we have assumed the exogeneity of the parameters. However, the theoretical study of endogeneity, as well as of instrumental variables (IV) regressions, in high dimensions are questions at the edge of the research frontier in econometrics (*e.g.* Belloni, Chen, Chernozhukov and Hansen (2012), Fan and Liao (2014), Chernozhukov, Hansen and Spindler (2015*b*), Chernozhukov, Hansen and Spindler (2015*a*), Gold, Lederer and Tao (2020)). Further research might explore the possibility of providing an extension of Theorem 1 to the case of an IV regression and study the opportunity it could open for inference. It might also be fruitful to explore the possibility of estimating simultaneous linear equations by using the GEP approach. Since the general form of the GEP of a pair of matrices (A, B) allows the matrix A

to have a rank higher than 1, it might be of interest to explore an extension of Theorem 1 to the case where the explained variable is not a scalar but a vector.

2.7 Appendix

The proof of Theorem 2, while correcting and extending results given in Tan et al. (2018), employ the same techniques as in the aforementioned paper and hence are given in the Supplementary Material. The proof of Proposition 3 employs known techniques and hence is also given in the Supplementary Material.

2.7.1 Proof of Theorem 1

Let $\upsilon = \|\beta\|_2$. The linearity in X_1 of the conditional expectation implies $\Sigma_{XY} = \Sigma_X \phi^\star \upsilon$. Hence we have

$$\frac{(\phi^\star)^\top \Sigma_{XY}}{\|\phi^\star\|^2_{\Sigma_X}} = \upsilon. \tag{2.7.1}$$

Substituting equation (2.7.1) into the functional form of the conditional expectation we obtain

$$\mathbb{E}(Y_1|X_1) = X_1^\top \frac{\phi^\star (\phi^\star)^\top \Sigma_{XY}}{\|\phi^\star\|^2_{\Sigma_X}} \tag{2.7.2}$$

Since $\mathbb{E}(Y_1|X_1)$ is a projection onto the vector space of real-valued linear functions of X that are square integrable, we have that

$$\begin{aligned}
\phi^\star &\in \operatorname*{argmin}_{\phi\in\mathbb{R}^p} \mathbb{E}\left[Y_1 - X_1^T \frac{\phi\phi^\top \Sigma_{XY}}{\|\phi\|^2_{\Sigma_X}}\right]^2 \\
&\in \operatorname*{argmin}_{\phi\in\mathbb{R}^p} \mathbb{E}\left(Y_1^2 - 2Y_1 X_1^T \frac{\phi\phi^\top \Sigma_{XY}}{\|\phi\|^2_{\Sigma_X}} + \frac{\Sigma_{YX}\phi}{\|\phi\|^2_{\Sigma_X}} \phi^\top X_1 X_1^\top \phi \frac{\phi^\top \Sigma_{XY}}{\|\phi\|^2_{\Sigma_X}}\right) \\
&\in \operatorname*{argmin}_{\phi\in\mathbb{R}^p} \mathbb{E}(Y_1^2) - 2\frac{\Sigma_{YX}\phi\phi^\top \Sigma_{XY}}{\|\phi\|^2_{\Sigma_X}} + \frac{\Sigma_{YX}\phi}{\|\phi\|^2_{\Sigma_X}} \|\phi\|^2_{\Sigma_X} \frac{\phi^\top \Sigma_{XY}}{\|\phi\|^2_{\Sigma_X}} \\
&\in \operatorname*{argmax}_{\phi\in\mathbb{R}^p} \left(\frac{\phi}{\|\phi\|_{\Sigma_X}}\right)^\top \Sigma_{XY}\Sigma_{YX} \frac{\phi}{\|\phi\|_{\Sigma_X}}.
\end{aligned}$$

Let $\mathcal{S}$ be the solution set of the previous problem and observe that it is closed under scalar multiplication. Let $\mathcal{S}_1 := \{\phi \in \mathcal{S} \text{ s.t. } \|\phi\|_{\Sigma_X} = 1\}$ and $\mathcal{S}_2 := \{\phi \in \mathcal{S} \text{ s.t. } \|\phi\|_{\Sigma_X} \neq 1\}$, hence we have

$$\mathcal{S} = \mathcal{S}_1 \sqcup \mathcal{S}_2 \text{ (disjoint union)},$$
$$\text{if } \phi \in \mathcal{S}_2 \text{ then } \frac{\phi}{\|\phi\|_{\Sigma_X}} \in \mathcal{S}_1, \text{ and}$$
$$\text{if } \phi \in \mathcal{S}_1 \text{ then } \alpha\phi \in \mathcal{S}_2 \text{ for all } \alpha \neq 1.$$

Consider now the following problem whose solutions set is equal to $\mathcal{S}_1$

$$\operatorname*{argmax}_{\phi \in \mathbb{R}^p} \phi^\top \Sigma_{XY}\Sigma_{YX}\phi \qquad \text{such that } \|\phi\|^2_{\Sigma_X} = 1. \tag{2.7.3}$$

By definition of $\|.\|_{\Sigma_X}$ the Lagrangian of problem (2.7.3) is

$$\mathcal{L} = \phi^\top \Sigma_{XY}\Sigma_{YX}\phi - \lambda(\phi^\top \Sigma_X \phi - 1)$$

computing $\nabla_\phi \mathcal{L} = 0$ leads to

$$\Sigma_{XY}\Sigma_{YX}\phi = \Sigma_X \phi\lambda \tag{2.7.4}$$

which is a scalar equation of the Generalized Eigenvalue Problem of the matrix pair $\big(\Sigma_{XY}\Sigma_{YX}, \Sigma_X\big)$. Pre-multiplying each side of this equation by $\phi^\top$ identifies λ as the maximum generalized eigenvalue which is why we symbolize it by $\lambda_{\max}$. Remark that if ϕ satisfies (2.7.4) then $\alpha\phi$ also satisfies this equation for any $\alpha \in \mathbb{R}$. This last remark allows us to conclude that

$$\mathcal{D}(\beta) = \operatorname{Ker}\big(\Sigma_{XY}\Sigma_{YX} - \lambda_{\max}\Sigma_X\big) = \mathcal{S}_1 \sqcup \mathcal{S}_2 = \mathcal{S},$$

which is the result to be proven. □

2.7.2 Proof of Proposition 1

Let $\mathcal{O}(p,k) := \{v \in \mathbb{R}^p; \|v\|_0 = k\}$. Observe that for any $v \in \mathcal{O}(p,k)$ there exists an infinity of $(\phi, m) \in \mathcal{O}(p,k) \times \mathbb{R}$ such that $\phi m = v$. Hence note that if $(\hat{\phi}, \hat{m}) \in \operatorname{argmin}_{(\phi,m)\in\mathcal{O}(p,s)\times\mathbb{R}} n^{-1}\|Y - X\phi m\|_2^2$ then $\tilde{\beta} = \hat{\phi}\hat{m}$ almost surely. Now observe that *for any* $\phi \in \mathcal{O}(p,k)$ we have

$$\operatorname*{argmin}_{m\in\mathbb{R}} \frac{1}{n}\|Y - X\phi m\|_2^2 = \frac{\phi^\top X^\top Y}{\phi^\top X^\top X \phi}.$$

Then in particular if $\hat{\phi} \in \operatorname{argmin}_{\phi\in\mathcal{O}(p,k)} n^{-1}\|Y - (X\phi\phi^\top XY)/\phi^\top X^\top X\phi)\|_2^2$ we have $\mathcal{D}(\tilde{\beta}) = \mathcal{D}(\hat{\phi})$ almost surely. Finally, computations similar to ones in the proof of Theorem 1 lead to

$$\operatorname*{argmin}_{\hat{\phi}\in\mathcal{O}(p,k)} \frac{1}{n}\|Y - \frac{X\hat{\phi}\hat{\phi}^\top XY}{\hat{\phi}^\top X^\top X\hat{\phi}}\|_2^2 = \operatorname*{argmax}_{\hat{\phi}\in\mathcal{O}(p,k)} \frac{\hat{\phi}^\top X^\top YY^\top X\hat{\phi}}{\hat{\phi}^\top X^\top X\hat{\phi} n},$$

that gives the result. □

2.7.3 Proof of Proposition 2

Let $\upsilon = \|\beta\|_2$ and $U := Y - X\beta$, straightforward computations give

$$\hat{\upsilon} = \frac{\hat{\phi}^\top X^\top Y}{\hat{\phi}^\top X^\top X \hat{\phi}} = \upsilon + \frac{\upsilon}{\hat{\phi}^\top \hat{\Sigma}_X \hat{\phi}} \big(\hat{\phi}^\top \hat{\Sigma}_X (\phi^\star - \hat{\phi}) + \frac{\hat{\phi}^\top X^\top U}{n\upsilon} \big)$$

where the second equality is obtained by adding zero and simplifying. Hence we have an equality for $\hat{\upsilon} - \upsilon$ that we now use to obtain an inequality. Let $P \in \mathcal{P}(F(p,l))$ be defined as in section 4.3, such that $l \leq 2s$ and $PP^\top \hat{\phi} = \hat{\phi}$ and $PP^\top(\phi^\star - \hat{\phi}) = (\phi^\star - \hat{\phi})$.

Observe that

$$\hat{\phi}^\top \hat{\Sigma}_X (\phi^\star - \hat{\phi}) = \hat{\phi}^\top PP^\top \hat{\Sigma}_X PP^\top (\phi^\star - \hat{\phi}) \leq \overline{\kappa} \|\phi^\star - \hat{\phi}\|_2$$

where the second equation is obtained by Cauchy–Schwarz inequality, the sub-multiplicativity of the spectral norm and $\|P\|_{\text{spec}} = 1$. Hence

$$|\hat{\upsilon} - \upsilon| \leq \frac{\upsilon}{\underline{\kappa}} \bigg(\overline{\kappa} \|\phi^\star - \hat{\phi}\|_2 + \frac{1}{\upsilon} \| \frac{PX^\top U}{n} \|_2 \bigg),$$

which implies $|\hat{\upsilon} - \upsilon| = O_p\big(\alpha_n\big) + O_p\big(\sqrt{l/n}\big) = O_p\big(\alpha_n\big)$. And since

$$\begin{aligned}
\|\hat{\beta} - \beta\|_2^2 &= \|\hat{\upsilon}\hat{\phi} - \upsilon\phi^\star\|_2^2 \\
&= \hat{\upsilon}^2 - 2\hat{\upsilon}\upsilon\langle\hat{\phi}, \phi^\star\rangle_2 + \upsilon^2 && \text{(Since } \|\hat{\phi}\|_2 = \|\phi^\star\|_2 = 1\text{)} \\
&= (\hat{\upsilon} - \upsilon)^2 + 2\hat{\upsilon}\upsilon(1 - \langle\hat{\phi}, \phi^\star\rangle_2) \\
&= (\hat{\upsilon} - \upsilon)^2 + \hat{\upsilon}\upsilon\|\hat{\phi} - \phi^\star\|_2^2 && \text{(Since } \langle\hat{\phi}, \phi^\star\rangle_2 \geq 0\text{)} \\
&= (\hat{\upsilon} - \upsilon)^2 + \big((\hat{\upsilon} - \upsilon)\upsilon + \upsilon^2\big)\|\hat{\phi} - \phi^\star\|_2^2,
\end{aligned}$$

since $|\hat{\upsilon} - \upsilon| = O_p(\alpha_n)$, $(\hat{\upsilon} - \upsilon)^2 = O_p(\alpha^2)$ and $(\hat{\upsilon} - \upsilon)\upsilon + \upsilon^2 = O_p(1)$, the result follows. □

2.7.4 Proof of Proposition 4

Let $f(\phi_t) := \rho_t = (\phi_t^\top \hat{\Sigma}_{XY} \hat{\Sigma}_{YX} \phi_t)/(\phi_t^\top \hat{\Sigma}_X \phi_t)$. Consequently, we have $\nabla_{\phi_t} f = (2/\phi_t^\top \hat{\Sigma}_X \phi_t)\big(\hat{\Sigma}_{XY}\hat{\Sigma}_{YX}\phi_t - f(\phi_t)\hat{\Sigma}_X\phi_t\big)$. Moreover, note that

$$\begin{aligned}
\phi_t^\top \hat{\Sigma}_{XY}\hat{\Sigma}_{YX}\phi_t &= \big(\frac{\phi_t^\top X^\top Y}{n}\big)^2 \\
&= \big(\frac{\phi_t^\top X^\top X\beta}{n} + \frac{\phi^\top X^\top U}{n}\big)^2 \\
&= \big(\frac{\phi_t^\top X^\top X\phi_t\|\beta\|_2}{n} + \frac{\phi_t^\top X^\top X(\phi^\star - \phi_t)\|\beta\|_2}{n} + \frac{\phi_t^\top X^\top U}{n}\big)^2 \\
&= \bigg(\frac{\|X\phi_t\|_2\|\beta\|_2}{n}\big\{1 + \frac{\phi_t^\top X^\top X(\phi^\star - \phi_t)}{\|X\phi_t\|_2^2} + \frac{\phi_t^\top X^\top U}{\|X\phi_t\|_2^2\|\beta\|_2}\big\}\bigg)^2.
\end{aligned}$$

Hence we have

$$\begin{aligned}
\tilde{\phi}_t &= \phi_{t-1} + \eta \frac{1}{f(\phi_{t-1})}\big(\hat{\Sigma}_{XY}\hat{\Sigma}_{YX}\phi_{t-1} - f(\phi_{t-1})\hat{\Sigma}_X\phi_{t-1}\big) \\
&= \phi_{t-1} + \frac{\eta}{2}\big(\phi_{t-1}^\top\hat{\Sigma}_X\phi_{t-1}\big)\frac{\nabla_{\phi_{t-1}} f}{f(\phi_{t-1})} \\
&= \phi_{t-1} + \frac{\eta}{2}\big(\phi_{t-1}^\top\hat{\Sigma}_X\phi_{t-1}\big)\nabla_{\phi_{t-1}}\log(f) \\
&= \phi_{t-1} + \\
&\quad \frac{\eta}{2}\big(\phi_{t-1}^\top\hat{\Sigma}_X\phi_{t-1}\big)\nabla_{\phi_{t-1}}\big\{\log(\phi_{t-1}^\top\hat{\Sigma}_{XY}\hat{\Sigma}_{YX}\phi_{t-1}) - \log(\phi_{t-1}^\top\hat{\Sigma}_X\phi_{t-1})\big\} \\
&= \phi_{t-1} + \frac{\eta}{2}\frac{\|X\phi_{t-1}\|_2^2}{n}\nabla_{\phi_{t-1}}\big\{\log(\phi_{t-1}^\top\hat{\Sigma}_{XY}\hat{\Sigma}_{YX}\phi_{t-1}) - \log(\frac{\|X\phi_{t-1}\|_2^2}{n})\big\}
\end{aligned}$$

which is equivalent to

$$\begin{aligned}
\phi_t' = \phi_{t-1} + \frac{\eta}{2}\frac{\|X\phi_{t-1}\|_2^2}{n}\nabla_{\phi_{t-1}}\big\{&2\log(\|\beta\|_2) + 2\log(\frac{\|X\phi_{t-1}\|_2^2}{n}) + \\
&2\log(1 + \frac{\phi_{t-1}^\top X^\top X(\phi^\star - \phi_{t-1})}{\|X\phi_{t-1}\|_2^2} + \frac{\phi_{t-1}^\top X^\top U}{\|X\phi_{t-1}\|_2^2\|\beta\|_2}) - \log(\frac{\|X\phi_{t-1}\|_2^2}{n})\big\}.
\end{aligned}$$

Simplifying we get

$$\begin{aligned}
\tilde{\phi}_t = \phi_{t-1} + \frac{\eta}{2}\frac{\|X\phi_{t-1}\|_2^2}{n}\nabla_{\phi_{t-1}}\big\{&2\log(\|\beta\|_2) + \log(\frac{\|X\phi_{t-1}\|_2^2}{n}) + \\
&2\log(1 + \frac{\phi_{t-1}^\top X^\top X(\phi^\star - \phi_{t-1})}{\|X\phi_{t-1}\|_2^2} + \frac{\phi_{t-1}^\top X^\top U}{\|X\phi_{t-1}\|_2^2\|\beta\|_2})\big\}.
\end{aligned}$$

Computing the gradient for the first two terms inside the brackets and simplifying further leads to

$$\begin{aligned}
\tilde{\phi}_t = \phi_{t-1} + \eta\hat{\Sigma}_X\phi_{t-1} + \\
\eta\frac{\|X\phi_{t-1}\|_2^2}{n}\nabla_{\phi_{t-1}}\log(1 + \frac{\phi_{t-1}^\top X^\top X(\phi^\star - \phi_{t-1})}{\|X\phi_{t-1}\|_2^2} + \frac{\phi_{t-1}^\top X^\top U}{\|X\phi_{t-1}\|_2^2\|\beta\|_2}).
\end{aligned}$$

Let $g_{t-1} := 1 + \frac{\phi_{t-1}^\top X^\top X(\phi^\star - \phi_{t-1})}{\|X\phi_{t-1}\|_2^2} + \frac{\phi_{t-1}^\top X^\top U}{\|X\phi_{t-1}\|_2^2\|\beta\|_2}$, we compute gradient in the last term

$$\begin{aligned}
&\nabla_{\phi_{t-1}}\log(1 + \frac{\phi_{t-1}^\top X^\top X(\phi^\star - \phi_{t-1})}{\|X\phi_{t-1}\|_2^2} + \frac{\phi_{t-1}^\top X^\top U}{\|X\phi_{t-1}\|_2^2\|\beta\|_2}) \\
&= \frac{1}{g_{t-1}}\left((2\hat{\Sigma}_X\phi_{t-1} + \hat{\Sigma}_X\phi^\star - 2\hat{\Sigma}_X\phi_{t-1} + \frac{X^\top U}{\|\beta\|_2})\frac{1}{\|X\phi_{t-1}\|_2^2} - g_{t-1}\frac{2\hat{\Sigma}_X\phi_{t-1}}{\|X\phi_{t-1}\|_2^2}\right) \\
&= \frac{1}{g_{t-1}}(\hat{\Sigma}_X\phi^\star + \frac{X^\top U}{\|\beta\|_2})\frac{1}{\|X\phi_{t-1}\|_2^2} - \frac{2\hat{\Sigma}_X\phi_{t-1}}{\|X\phi_{t-1}\|_2^2}.
\end{aligned}$$

And since for all t, $\phi_t = Q_t\phi_t$ we have

$$\begin{aligned}\tilde{\phi}_t &= \phi_{t-1} - \eta\hat{\Sigma}_X\phi_{t-1} + \eta\frac{1}{g_{t-1}}\hat{\Sigma}_X\phi^\star + \eta\frac{1}{g_{t-1}}\frac{X^\top U}{\|\beta\|_{\mathbb{R}^p}n} \\ &= (I_{p\times p} - \eta\hat{\Sigma}_X)\phi_{t-1} + \eta\frac{1}{g_{t-1}}(\hat{\Sigma}_X\phi^\star + \frac{X^\top U}{\|\beta\|_{\mathbb{R}^p}n}) \\ &= (I_{p\times p} - \eta\hat{\Sigma}_X)Q_{t-1}\phi_{t-1} + \eta\frac{1}{g_{t-1}}(\hat{\Sigma}_X\phi^\star + \frac{X^\top U}{\|\beta\|_{\mathbb{R}^p}n}),\end{aligned} \tag{2.7.5}$$

which implies that for all $t \in \mathbb{N}$

$$\begin{aligned}\tilde{\phi}_t = (I_{p\times p} - \eta\hat{\Sigma}_X)Q_{t-1}\tilde{\phi}_{t-1} \\ + (I_{p\times p} - \eta\hat{\Sigma}_X)Q_{t-1}(\phi_{t-1} - \tilde{\phi}_{t-1}) \\ + \frac{\eta}{g_{t-1}}(\hat{\Sigma}_X\phi^\star + \frac{X^\top U}{\|\beta\|_{\mathbb{R}^p}n})\end{aligned} \tag{2.7.6}$$

by letting $\tilde{\phi}_0 = \phi_0$.

Then by recursively applying equation (2.7.6) we have for all $t \geq 2$

$$\begin{aligned}\tilde{\phi}_t =& \left(\prod_{k=1}^{t}(I_{p\times p} - \eta\hat{\Sigma}_X)Q_{t-k}\right)\phi_0 \\ &+ \sum_{m=1}^{t-1}\left(\prod_{l=1}^{m}(I_{p\times p} - \eta\hat{\Sigma}_X)Q_{t-l}\right)(\phi_{t-m} - \tilde{\phi}_{t-m}) \\ &+ \eta\sum_{m=0}^{t-1}\left(\prod_{l=1}^{m}(I_{p\times p} - \eta\hat{\Sigma}_X)Q_{t-l}\right)\frac{1}{g_{t-1-m}}\left(\hat{\Sigma}_X\phi^\star + \frac{X^\top U}{\|\beta\|_2 n}\right).\end{aligned}$$

By adding zero, on the left-hand side and rearranging the terms and noting that

$$g_{t-1} = \frac{\phi_{t-1}^\top\hat{\Sigma}_X\phi^\star + \frac{\phi_{t-1}^\top X^\top U}{\|\beta\|_{\mathbb{R}^p}n}}{\phi_{t-1}^\top\hat{\Sigma}_X\phi_{t-1}}.$$

we obtain the thesis. □

2.7.5 Proof of Proposition 5

Since $\sup_{i\in\{0,\dots,m\}} a_i = O_p(1)$, for all $\epsilon > 0$, there exists $N_1 \in \mathbb{N}$ and $C > 0$ such that for all $m > m(N_1)$ and all $\gamma > 0$ we have

$$\begin{aligned}
&\mathbb{P}\Big(\sum_{i=0}^{m} a_{m-i}b_i > \gamma\Big) \\
&\leq \mathbb{P}\Big(\sum_{i=0}^{m} a_{m-i}b_i > \gamma \Big| \sup_{i\in\{0,\dots,m\}} a_i \leq C\Big) + \mathbb{P}\Big(\sup_{i\in\{0,\dots,m\}} a_i > C\Big) \\
&= \mathbb{P}\Big(\sum_{i=0}^{m} a_{m-i}b_i > \gamma \Big| \sup_{i\in\{0,\dots,m\}} a_i \leq C\Big) + \epsilon/2.
\end{aligned}$$

Now since $\lim_{n\to\infty}\sum_{i=0}^{m} b_i = M$ a.s., for all $\alpha > 0$, there exists $N_2 \in \mathbb{N}$ such that for all $m \geq m(N_2)$ we have $\sum_{m\geq m(N_2)} b_i \leq \alpha/(2C)$ almost surely. Consequently we have for all $\gamma > 0$ and all $m > M(N_2)$ that

$$\begin{aligned}
&\mathbb{P}\Big(\sum_{i=0}^{m} a_{m-i}b_i > \gamma \Big| \sup_{i\in\{0,\dots,m\}} a_i \leq C\Big) \\
&\qquad\qquad \leq \mathbb{P}\Big(\sum_{i=0}^{m(N_2)} a_{m-i}b_i + \frac{\alpha}{2} > \gamma \Big| \sup_{i\in\{0,\dots,m\}} a_i \leq C\Big).
\end{aligned}$$

Note now that since $\{a_i\}_{i=0}^{m} = o_p(1)$, for any $\alpha, \tilde{\epsilon} > 0$ there exists $N_3 \in \mathbb{N}$ such that for all $m > m(N_3)$ we have

$$\mathbb{P}\Big(a_m > \frac{\alpha}{2m(N_2)M}\Big) < \frac{\tilde{\epsilon}}{(m(N_2)+1)}$$

which implies by a union bound that for all $m > m(N_2) + m(N_3)$

$$\mathbb{P}\Big(\sup_{i\in\{0,\dots,m(N_2)\}} a_{m-i} > \frac{\alpha}{2m(N_2)M}\Big) \leq \tilde{\epsilon}.$$

Hence we have for all $m > m(N_2) + m(N_3)$ that

$$\mathbb{P}\Big(\sum_{i=0}^{m(N_2)} a_{m-i}b_i + \frac{\alpha}{2} > \gamma \Big| \sup_{i\in\{0,\dots,m\}} a_i \leq C\Big)$$

is bounded above by the *sum* of

$$\begin{aligned}
&\leq \mathbb{P}\Big(\sum_{i=0}^{m(N_2)} a_{m-i}b_i + \frac{\alpha}{2} > \gamma \Big| \\
&\qquad \{\sup_{i\in\{0,\dots,m\}} a_i \leq C\} \cap \{\sup_{i\in\{0,\dots,m(N_2)\}} a_{m(N_2)-i} < \frac{\alpha}{2m(N_2)M}\}\Big)
\end{aligned}$$

and

$$\mathbb{P}\Big(\sup_{i\in\{0,\dots,m(N_2)\}} a_{m(N_2)-i} > \frac{\alpha}{2m(N_2)M} \,\Big|\, \{\sup_{i\in\{0,\dots,m\}} a_i \le C\}\Big).$$

And since for any events A and B such that $\mathbb{P}(B) > 0$, the Law of Total Probability implies that

$$\frac{\mathbb{P}(A)}{1-\mathbb{P}(\bar{B})} \ge \mathbb{P}(A|B)$$

by choosing the events $A := \{\sup_{i\in\{0,\dots,m(N_2)\}} a_{m(N_2)-i} < \alpha/(2m(N_2)M\}$ and $B = \{\sup_{i\in\{0,\dots,m\}} a_i \le C\}$, we have for every $\epsilon, \tilde{\epsilon}$

$$\mathbb{P}\Big(\sup_{i\in\{0,\dots,m(N_2)\}} a_{m(N_2)-i} > \frac{\alpha}{2m(N_2)M} \,\Big|\, \{\sup_{i\in\{0,\dots,m\}} a_i \le C\}\Big) \le \frac{\tilde{\epsilon}}{1-\epsilon/2}.$$

Hence for every ϵ we choose $\tilde{\epsilon}$ such that $\tilde{\epsilon}/(1-\epsilon/2) < \epsilon/2$. Remark that this choice only affects N_3 . As a consequence we have for all $m > m(N_2)+m(N_3)$

$$\begin{aligned}\mathbb{P}\Big(\sum_{i=0}^{m(N_2)} a_{m-i}b_i + \frac{\alpha}{2} > \gamma \,\Big|\, \sup_{i\in\{0,\dots,m\}} a_i \le C\Big) \\ \le \mathbb{P}\Big(\sum_{i=0}^{m(N_2)} a_{m-i}b_i + \frac{\alpha}{2} > \gamma \,\Big| \\ \{\sup_{i\in\{0,\dots,m\}} a_i \le C\} \cap \{\sup_{i\in\{0,\dots,m(N_2)\}} a_{m(N_2)-i} < \frac{\alpha}{2m(N_2)M}\}\Big) + \epsilon/2.\end{aligned}$$

where

$$\begin{aligned}\mathbb{P}\Big(\sum_{i=0}^{m(N_2)} a_{m-i}b_i + \frac{\alpha}{2} > \gamma \,\Big| \\ \{\sup_{i\in\{0,\dots,m\}} a_i \le C\} \cap \{\sup_{i\in\{0,\dots,m(N_2)\}} a_{m(N_2)-i} < \frac{\alpha}{2m(N_2)M}\}\Big)\end{aligned}$$

is bounded above by

$$\begin{aligned}\mathbb{P}\Big(\frac{\alpha}{2} + \frac{\alpha}{2} > \gamma \,\Big|\, \{\sup_{i\in\{0,\dots,m\}} a_i \le C\} \cap \{\sup_{i\in\{0,\dots,m(N_2)\}} a_{m(N_2)-i} < \frac{\alpha}{2m(N_2)M}\}\Big) \\ \le \mathbb{P}\Big(\alpha > \gamma \,\Big|\, \{\sup_{i\in\{0,\dots,m\}} a_i \le C\} \cap \{\sup_{i\in\{0,\dots,m(N_2)\}} a_{m(N_2)-i} < \frac{\alpha}{2m(N_2)M}\}\Big)\end{aligned}$$

and since we can pick $\alpha > 0$ as small as we want, for any fixed $\gamma > 0$ we can make the probability in the last inequality equal to zero. Hence, for all $\gamma, \epsilon > 0$, there exists $N_1, N_2, N_3 \in \mathbb{N}$ such that for all $m > m(N_1) + m(N_2) + m(N_3)$ we have $\mathbb{P}\Big(\sum_{i=0}^{m} a_{m-i}b_i > \gamma\Big) < \epsilon$ which is what needed to be proven. □

2.7.6 Proof of Proposition 6

Lemmas

Lemma 1 *For all $n \in \mathbb{N}$ let $X \in \mathbb{R}^{n\times p}$ be a random matrix* iid *across and such that $\mathbb{E}(\|X_1\|^2) < \infty$. Let $\hat{\Sigma}_X = X^\top X/n$, and $P \in \mathcal{P}(\mathcal{F}(p,k))$ for some $k \in \mathbb{N}$, $k < p$.*

Then conditionally on X, if $P^\top \hat{\Sigma}_X P$ is positive definite a.s. then there exists $\eta > 0$ such that

$$\|(I_{p\times p} - \eta\hat{\Sigma}_X)P\|_{\text{spec}} < 1 \qquad a.s.$$

Proof. The proof is given in the Supplementary Material

Lemma 2 *Under the same set of assumptions as in Theorem 3 we have*

$$\left|\frac{\phi_t^\top \hat{\Sigma}_X \phi_t}{\phi_t^\top \hat{\Sigma}_X \phi^\star + \phi_t^\top \frac{X^\top U}{\|\beta\|_2 n}} - 1\right| = o_p(1)$$

Proof. The result follows from the facts that $\phi_t \xrightarrow{p} \phi^\star$ and that $\frac{X^\top U}{\|\beta\|_2 n} \xrightarrow{p} 0$. □

Lemma 3 *Under the same set of assumptions as in Theorem 3 we have (and in particular if $\inf_{\|v\|_2=1,\|v\|_0=k} |v^\top \Sigma_{XY}| > 0$),*

$$\sup_{\tilde{t}\in\{1,\dots,t(n)\}} \left|\frac{\phi_{\tilde{t}}^\top \hat{\Sigma}_X \phi_{\tilde{t}}}{\phi_{\tilde{t}}^\top \hat{\Sigma}_X \phi^\star + \phi_{\tilde{t}}^\top \frac{X^\top U}{\|\beta\|_2 n}} - 1\right| = O_p(1)$$

Proof. The proof is given in the Supplementary Material.

Lemma 4 *For all $n \in \mathbb{N}$, let $Y \in \mathbb{R}^n$ and $X \in \mathbb{R}^{n\times p}$ be random functions* iid *across rows. For all $n \in \mathbb{N}$ assume $\Sigma_X := \mathbb{E}(X_1 X_1^\top)$ is finite and positive definite and there exists $\beta \in \mathbb{R}^p$ s.t. $\mathbb{E}(Y_1|X_1) = X_1^\top \beta$ and $\|\beta\|_0 = s$. Let $\phi^\star = \beta/\|\beta\|_2$. Let t be as in Algorithm 1 and assume that $t \to \infty$ as $n \to \infty$.*

If conditionally X, for given k and ϕ_0, for any η, for all $\phi_t \in \Phi(\phi_0, k, \eta)$ we have that $P_t^\top \hat{\Sigma}_X P_t$ is positive definite a.s. then there exists η and $M_\eta < \infty$ such that, conditionally on X,

$$\lim_{n\to\infty} \|\sum_{m=0}^{t-1}\prod_{l=1}^{m}(I_{p\times p} - \eta\hat{\Sigma}_X)Q_{t-l}\|_{\text{spec}} = M_\eta \qquad a.s.$$

Proof. Note that

$$\|\sum_{m=0}^{t-1}\prod_{l=1}^{m}(I_{p\times p} - \eta\hat{\Sigma}_X)Q_{t-l}\|_{\text{spec}} \leq \sum_{m=0}^{t-1}\sup_{l\in\{1,\dots,t-1\}}\|(I_{p\times p} - \eta\hat{\Sigma}_X)P_{t-l}\|_{\text{spec}}^m \|P_{t-l}\|_{\text{spec}}^m.$$

and since for all $l \in \{1, ..., t-1\}$, by construction we have $\|P_{t-l}\|_{\text{spec}} = 1$ and by Lemma 1, there exists η such that $\|(I_{p\times p} - \eta\hat{\Sigma}_X)P_{t-l}\|_{\text{spec}} < 1$ a.s., conditionally on X. Hence for such η the series on the right-hand side is a convergent geometric series. This proves the thesis since $t \to \infty$ as $n \to \infty$. □

Lemma 5 *Under the same assumptions than in Theorem 3, there exists $\alpha > 0$ such that we have $\mathbb{P}\big(Q(\phi_t) \in \mathcal{Q}^\star \big| \|\phi_{\tilde{t}} - \phi^\star\|_2 \leq \alpha\big) = 1$ for all $t \geq \tilde{t}$, where Q and $\mathcal{Q}^\star$ are defined in Equation (2) and thereafter.*

Proof. The proof is given in the Supplementary Material

Proof of Proposition 6

The existence of η follows from Lemma 1. For the remaining of the proof, we consider that η is such that $\|(I_{p\times p} - \eta\hat{\Sigma}_X)P_t\|_{\text{spec}} < 1$ a.s. conditionally on X.

Let $\Gamma_m := \prod_{l=1}^{m}(I_{p\times p} - \eta\hat{\Sigma}_X)Q_{t-l}$ (recall that our convention on $\prod$ implies that $\Gamma_0 = I_{p\times p}$) and let

$$g_{t-1-m} := \frac{\phi_{t-1-m}^\top \hat{\Sigma}_X \phi_{t-1-m}}{\phi_{t-1-m}^\top \hat{\Sigma}_X \phi^\star + \phi_{t-1-m}^\top \frac{X^\top U}{\|\beta\|_2 n}}.$$

Define $L := g_{t-1}I_{p\times p} + g_{t-2}\Gamma_1 + ... + g_0\Gamma_{t-1}$ and $M := I_{p\times p} + \Gamma_1 + ... + \Gamma_{t-1}$ as a consequence $L - M = (g_{t-1}-1)I_{p\times p} + (g_{t-2}-1)\Gamma_1 + ... + (g_0-1)\Gamma_{t-1}$. Lemmas 2, 3 and 4 ensures that by Proposition 5 we have $\|L - M\|_{\text{spec}} = o_p(1)$. Hence we have shown that $L = M + o_p(1)$. Now consider M. For any $\alpha > 0$ and $t' \in \mathbb{N}$, let $\mathcal{E}_{\alpha t'} := \{\|\phi_{t'} - \phi^\star\|_2 \leq \alpha\}$, then we can write $M = M \times \mathbf{1}_{\mathcal{E}_{\alpha\tilde{t}}} + M \times \mathbf{1}_{\mathcal{E}^c_{\alpha t'}}$. Given our assumptions about the convergence of ϕ_t, we know that for any $\alpha > 0$ and some t' high enough we have $M \times \mathbf{1}_{\mathcal{E}^c_{\alpha t'}} = o_p(1)$. Furthermore by Lemma 5, by choosing α low enough we have that there exists $Q^\star \in \mathcal{Q}^\star$ such that

$$M \times \mathbf{1}_{\mathcal{E}_{\alpha t'}} = \left[\sum_{m=0}^{t-t'}\Big((I - \eta\hat{\Sigma}_X)Q^\star\Big)^m + \sum_{m=t-t'+1}^{t-1}\prod_{l=1}^{m}(I - \eta\hat{\Sigma}_X)Q_{t-l}\right] \times \mathbf{1}_{\mathcal{E}_{\alpha t'}}$$

By Lemma 5 and dealing with indices, the right-hand-side can be rewritten as

$$\left[\sum_{m=0}^{t-t'}\Big((I - \eta\hat{\Sigma}_X)Q^\star\Big)^m + \sum_{m=1}^{t'-1}\Big((I - \eta\hat{\Sigma}_X)Q^\star\Big)^{(t-t')}\prod_{l=1}^{m}(I - \eta\hat{\Sigma}_X)Q_{t'-l}\right] \times \mathbf{1}_{\mathcal{E}_{\alpha\tilde{t}}}.$$

Since by submutiplicativity of the spectral norm and Lemma 1 we obtain that $\|(I-\eta\hat{\Sigma}_X)Q^\star\|_{\text{spec}} \leq \|(I-\eta\hat{\Sigma}_X)P^\star\|_{\text{spec}} < 1$, we have (geometric series) that

$$\sum_{l=0}^{t-t'}\Big((I-\eta\hat{\Sigma}_X)Q^\star\Big)^m = [(I-Q^\star)+\eta\hat{\Sigma}_X Q^\star]^{-1}[(I-\Big((I-\eta\hat{\Sigma}_X)Q^\star\Big)^{t-t'}].$$

Moreover since $\Big((I-\eta\hat{\Sigma}_X)Q^\star\Big)^{t-t'} = o_p(1)$ and since

$$\sum_{m=0}^{t'}\Big((I-\eta\hat{\Sigma}_X)Q^\star\Big)^{t-t'}\prod_{l=0}^{m}(I-\eta\hat{\Sigma}_X)Q_{t'-l} = o_p(1)$$

because $t-t' \to \infty$ as $n \to \infty$, we have shown that

$$L = M + o_p(1) = [(I-Q^\star)+\eta\hat{\Sigma}_X Q^\star]^{-1} + o_p(1)$$

which is what we needed to prove. □

2.7.7 Proof of Theorem 3

Lemmas

Lemma 6 *Under the same set of assumptions as in Theorem, 3 we have* $\|\phi_t - \tilde{\phi}_t\|_2 = o_p(1)$

Proof. The proof is given in the Supplementary Material.

Lemma 7 *Under the same set of assumptions as in Theorem 3 we have (and in particular if* $\inf_{\|v\|_2=1,\|v\|_0=k}|v^\top\Sigma_{XY}| > 0$ *and given the definition of* $\tilde{\phi}$*), then* $\sup_{t\in\{1,\dots,t(n)\}}\|\phi_t - \tilde{\phi}_t\|_2 = O_p(1)$.

Proof. The proof is given in the Supplementary Material

Proof of Theorem 3

The existence of η follows from Lemma 1. For the remaining of the proof, we consider that η is such that $\|(I_{p\times p}-\eta\hat{\Sigma}_X)P_t\|_{\text{spec}} < 1$ a.s. conditionally on X.

By Proposition 4 we have for all $t \geq 2$

$$\phi_t - \phi^\star = \left(\prod_{m=1}^{t} (I_{p\times p} - \eta\hat{\Sigma}_X) Q_{t-m} \right) \phi_0 \tag{2.7.7}$$

$$+ \sum_{m=0}^{t-1} \left(\prod_{l=1}^{m} (I_{p\times p} - \eta\hat{\Sigma}_X) Q_{t-l} \right) (\phi_{t-l} - \tilde{\phi}_{t-l}) \tag{2.7.8}$$

$$+ \eta \sum_{m=0}^{t-1} \left(\prod_{l=1}^{m} (I_{p\times p} - \eta\hat{\Sigma}_X) Q_{t-l} \right) g_{t-1-m} \left(\hat{\Sigma}_X \phi^\star + \frac{X^\top U}{\|\beta\|_2 n} \right) - \phi^\star \tag{2.7.9}$$

where

$$g_{t-1-m} = \frac{\phi_{t-1-m}^\top \hat{\Sigma}_X \phi_{t-1-m}}{\phi_{t-1-m}^\top \hat{\Sigma}_X \phi^\star + \phi_{t-1-m}^\top \frac{X^\top U}{\|\beta\|_2 n}}.$$

We treat the term on the right-hand side of (2.7.7). By Cauchy–Schwarz inequality we have

$$\begin{aligned}
&\sqrt{n} w^\top \left(\prod_{k=1}^{t} (I_{p\times p} - \eta\hat{\Sigma}_X) Q_{t-k} \right) \phi_0 \\
&\leq \sqrt{n} \| \left(\prod_{k=1}^{t} (I_{p\times p} - \eta\hat{\Sigma}_X) Q_{t-k} \right) \phi_0 \|_2 \\
&\leq \sqrt{n} \prod_{m=1}^{t} \| (I_{p\times p} - \eta\hat{\Sigma}_X) Q_{t-m} \|_{\text{spec}} \\
&\leq \sqrt{n} \prod_{m=1}^{t} \| (I_{p\times p} - \eta\hat{\Sigma}_X) P_{t-m} \|_{\text{spec}} \| P_{t-m} \|_{\text{spec}} \\
&\leq \sqrt{n} \sup_{m\in\{1,\dots,t\}} \| (I_{p\times p} - \eta\hat{\Sigma}_X) P_{t-m} \|_{\text{spec}}^t
\end{aligned}$$

and by Lemma 1

$$\mathbb{P}\left(\lim_{n\to\infty} \sqrt{n} \sup_{m\in\{1,\dots,t\}} \| (I_{p\times p} - \eta\hat{\Sigma}_X) P_{t-m} \|_{\text{spec}}^t = 0 \middle| X \right) = 1.$$

Now we treat terms (2.7.8). Lemmas 4, 6 and 7 ensure that by Proposition 5, conditionally on X we have

$$\sum_{m=0}^{t-1} \left(\prod_{l=1}^{m} (I_{p\times p} - \eta\hat{\Sigma}_X) Q_{t-l} \right) (\phi_{t-l} - \tilde{\phi}_{t-l}) = o_p(1).$$

We now treat the term (2.7.9). First consider

$$\eta \sum_{m=0}^{t-1} \left(\prod_{l=1}^{m} (I_{p\times p} - \eta\hat{\Sigma}_X) Q_{t-l} \right) g_{t-1-m} \hat{\Sigma}_X \phi^\star - \phi^\star.$$

Since $(I - Q^\star)\phi^\star = 0$ we can write

$$\left[\left(\sum_{m=0}^{t-1} \left(\prod_{l=1}^{m} (I_{p\times p} - \eta\hat{\Sigma}_X) Q_{t-l} \right) g_{t-1-m} \right) \left((I - Q^\star) + \eta\hat{\Sigma}_X Q^\star \right) - I_{p\times p} \right] \phi^\star.$$

Then by Proposition 6, we have that this quantity converges to zero in probability conditionally on X. Finally, Proposition 6, Slutsky's Lemma and the Central Limit Theorem imply that

$$w^\top \eta \sum_{m=0}^{t-1} \left(\prod_{l=1}^{m} (I_{p\times p} - \eta\hat{\Sigma}_X) Q_{t-l} \right) g_{t-1-m} \frac{X^\top U}{\|\beta\|_2 \sqrt{n}}$$

is

$$\mathcal{AN}\left(0, w^\top [(I - Q^\star) + \eta\hat{\Sigma}_X Q^\star]^{-1} \eta^2 \hat{\Sigma}_X [(I - Q^\star) + \eta\hat{\Sigma}_X Q^\star]^{-1} w \frac{\mathbb{V}\mathrm{ar}(U_1)}{\|\beta\|_2^2} \middle| X \right)$$

hence we have proven the thesis. □

2.8 Supplementary Materials

2.8.1 Proof of Theorems 2

Remark 1 *In the proofs of this section, it will always be understood that the fact that $\phi^\star$ is the leading eigenvector of the pair of matrices $(\Sigma_X, \Sigma_{XY}\Sigma_{YX})$ is proven in Theorem 1.*

Remark 2 *Note that $\Sigma_{XY}\Sigma_{YX}$ is symmetric, hence by Corollary 8.7.2 in Golub and Van Loan (2013) there exists a non-singular $W \in \mathbb{R}^{p\times p}$ such that*

$$W^\top \Sigma_{XY}\Sigma_{YX} W = diag(a_1, ..., a_p) \quad \text{and} \quad W^\top \mathbb{E}(X_1 X_1^\top) W = diag(b_1, ..., b_p)$$

and such that, for all $i \in \{1, ..., p\}$

$$\Sigma_{XY}\Sigma_{YX} W_i = \lambda_i \Sigma_X W_i \qquad \text{where } \lambda_i = a_i/b_i.$$

But since $\Sigma_{XY}\Sigma_{YX}$ is of rank 1, only one a_i is different from zero. Hence only one λ_i is different from zero. This implies that in Tan et al. (2018) $\gamma = 0$. We note that Lemma 3 in Tan et al. (2018) remains valid by interpreting the divisions by zero implied by $\gamma = 0$ as infinities.

Note that the same reasoning holds (when the associated GEP are definite) for $\hat{\Sigma}_{XY}\hat{\Sigma}_{YX}$ and every sub-matrices one can construct by pre- and post-multiplying $\Sigma_{XY}\Sigma_{YX}$ or $\hat{\Sigma}_{XY}\hat{\Sigma}_{YX}$ by any $P \in \mathcal{P}(\mathcal{F}(p, k)$ for any $k \leq p$.

Remark 3 *In order to prove the results we need to define the following quantity. For any vector* $v \in \mathbb{R}^p$*, let* $\text{supp}(v)$ *be the set of index of non-zero elements of* v*. For any set* $F \subset \{1, ..., p\}$ *let* $\phi(F)$ *be defined as*

$$\phi(F) := \underset{\phi \in \mathbb{R}^p}{\text{argmax}} \frac{\phi^\top X^\top Y Y^\top X \phi}{n \phi^\top X^\top X \phi} \qquad s.t.\ \phi^\top \hat{\Sigma}_X \phi = 1,\ \text{supp}(\phi) \subseteq F.$$

Lemmas

The next lemma specializes the result of Lemma 3 in Tan et al. (2018) to the setup of this chapter.

Lemma 8 *For all* $n \in \mathbb{N}$*, let* $Y \in \mathbb{R}^n$ *and* $X \in \mathbb{R}^{n \times p}$ *be random functions* iid *across rows. For all* $n \in \mathbb{N}$ *assume* $\Sigma_X := \mathbb{E}(X_1 X_1^\top)$ *is positive definite and there exists* $\beta \in \mathbb{R}^p$ *s.t.* $\mathbb{E}(Y_1 | X_1) = X_1^\top \beta$ *and* $\|\beta\|_0 = s$*.*

For $k' \in \mathbb{N}$ *s.t.* $k' > s$*, let* $F \in \mathcal{F}(p, k')$*. Given any* $\tilde{\phi} \in \mathbb{R}^k$ *such that* $\|\tilde{\phi}\|_2 = 1$ *and* $\tilde{\phi}^\top P_{(F)}^\top \phi(F) > 0$ *a.s, let*

$$\rho = \frac{\tilde{\phi}^\top P_{(F)}^\top \hat{\Sigma}_{XY} \hat{\Sigma}_{YX} P_{(F)} \tilde{\phi}}{\tilde{\phi}^\top P_{(F)}^\top \hat{\Sigma}_X P_{(F)} \tilde{\phi}}$$

where $P_{(F)}$ *is defined as in section 4.3. Moreover let* $\phi' = C_F \tilde{\phi} / \|C_F \tilde{\phi}\|_2$ *where*

$$C_F = P_{(F)}^\top P_{(F)} + (\eta/\rho) P_{(F)}^\top (\Sigma_{XY} \Sigma_{YX} - \rho \Sigma_X) P_{(F)}$$

and $\eta > 0$ *is some positive constant. Let* $\delta = 1 - \tilde{\phi}^\top P_{(F)}^\top \phi(F)$*. Pick* $c > 0$ *and* η *(sufficiently small) such that*

$$\eta \lambda_{\max}(\Sigma) < 1/(1 + c),$$

Then, conditionally on the set of events where for some $b > 0$

$$\frac{\varepsilon(k')}{\Delta(k')} < b,$$
$$\rho(\hat{\Sigma}_X - \Sigma_X, k') \leq c \lambda_{\min}(\Sigma_X),$$
$$1 - \delta \geq 1 - \xi$$

hold almost surely with ξ *defined as*

$$\xi := \min \left\{ \frac{1}{8 C_\diamond \kappa(\Sigma_X)}, \frac{1}{30(1+c) C_\diamond^3 \eta \lambda_{\max} \kappa^3(\Sigma_X)} \right\},$$

where $C_\diamond = (1 + c)/(1 - c)$*, we have*

$$\phi(F)^\top P_{(F)} \phi' \geq \phi(F)^\top P_{(F)} \tilde{\phi} + \frac{1 + c}{8} \eta \lambda_{\min}(\Sigma_X)(1 - \phi(F)^\top P_{(F)} \tilde{\phi}) \frac{1}{C_\diamond \kappa(\Sigma_X)}$$

almost surely,

Proof. The result is a direct consequence of Lemma 3 in Tan et al. (2018) and of Remark 2. □

The following lemma follows directly from Lemma 4 in Tan et al. (2018).

Lemma 9 [Lemma 4 in Tan et al. (2018)] *Consider a vector v' with $F' = \text{supp}(v')$ and $\text{Card}(F) = \bar{k}$. Let F be the indices the element of an other vector v that have the largest k absolutes values, with $\text{Card}(F) = k$. If $\|v'\|_2 = \|v\|_2 = 1$ then*

$$\left|\text{Truncate}(v, F)^\top v'\right| \geq |v^\top v'| - (\bar{k}/k)^{1/2} \min\{\sqrt{1 - (v^\top v')^2}, (1 + (\bar{k}/k)^{1/2})(1 - (v^\top v')^2)\}$$

The next lemma specializes the result of Lemma 5 in Tan et al. (2018) to the setup of this chapter.

Lemma 10 *For all $n \in \mathbb{N}$, let $Y \in \mathbb{R}^n$ and $X \in \mathbb{R}^{n \times p}$ be random functions* iid *across rows. For all $n \in \mathbb{N}$ assume $\Sigma_X := \mathbb{E}(X_1 X_1^\top)$ is positive definite and there exists $\beta \in \mathbb{R}^p$ s.t. $\mathbb{E}(Y_1|X_1) = X_1^\top \beta$ and $\|\beta\|_0 = s$. Let $\phi^\star = \beta/\|\beta\|_2$, $V = \text{supp}(\phi^\star)$ and $F \subset \{1, ..., p\}$ such that $V \subset F$ with $\text{Card}(F) = k' > s$.*

Conditionally on the set of events where $\Lambda := \lambda_{\max}/(1+\lambda_{\max}^2) > \varepsilon(k')/\Delta(k')$ a.s., we have that

$$\sqrt{1 - |\phi(F)^\top \phi^\star|} \leq \frac{\sqrt{2}}{\Lambda\big(\Delta(k') - \varepsilon(k')\big)} \varepsilon(k').$$

holds almost surely, where $\Delta(k')$ and $\varepsilon(k')$ are defined as in section 4.3.

Proof. The result follows directly from Lemma 5 in Tan et al. (2018) and from Remark 2. □

The following proposition originates from an application of Corollary 1 in Tan et al. (2018) in order to demonstrate a non-asymptotic bound for Algorithm 2. We provide our own proof since the hypothesis given in Tan et al. (2018) are not sufficient to prove the results the authors gave in their paper.

Note that a non-asymptotic bound on the l_2 error of the output of Algorithm 1 follows directly from this proposition.

Proposition 7 *For all $n \in \mathbb{N}$, let $Y \in \mathbb{R}^n$ and $X \in \mathbb{R}^{n \times p}$ be random functions* iid *across rows. For all $n \in \mathbb{N}$ assume $\Sigma_X := \mathbb{E}(X_1 X_1^\top)$ is positive definite and there exists $\beta \in \mathbb{R}^p$ s.t. $\mathbb{E}(Y_1|X_1) = X_1^\top \beta$ and $\|\beta\|_0 = s$. Let $\phi^\star = \beta/\|\beta\|_2$ and $\Lambda := \lambda_1/(1 + \lambda_1^2)$.*

Let k in Algorithm 1 (respectively $\underline{k}$ in Algorithm 2) be such that $k = Cs$ for $C > 1$ (respectively $\underline{k} = C's$ for $C' > 1$) and let $k' = 2k + s$ (respectively let $\bar{k} = C''\underline{k}$ for $C'' > 1$ and $k' = \bar{k} + s$).

Let $\bar{\phi}$ be equal to ϕ_t in step 3.f) *of Algorithm 1 for any $t \in \mathbb{N}$ (respectively to $\phi_{t,k}$ in step* 4.b)vi.B. *in Algorithm 2 for any $t \in \mathbb{N}$ and $k \in \{\underline{k}+1, ..., \overline{k}\}$). Let $\phi_{(0)}$ be equal to ϕ_0 in step 2 of Algorithm 1 (respectively to $\phi_{0,k}$ in step* 2 *of Algorithm 2 if $k = \overline{k}$, and else to $\phi_{T,k+1}$, as in step* 4.c.*).*

Choose a constant $c > 0$ and choose η in Algorithm 1 (respectively Algorithm 2) such that

$$\eta\lambda_{\max}(\Sigma_X) < 1/(1+c), \tag{2.8.1}$$

and

$$\nu = [1 + 2(\frac{s}{k_\diamond})^{1/2} + \frac{s}{k_\diamond})]^{1/2}\left(1 - \frac{1+c}{8}\eta\lambda_{\min}(\Sigma_X)\frac{1}{C_\diamond\kappa(\Sigma_X)}\right)^{1/2} < 1 \tag{2.8.2}$$

holds true, where $C_\diamond = (1+c)/(1-c)$ and $k_\diamond$ is equal to k (respectively to $\underline{k}$).

Then, conditionally on the set of events where

$$\frac{\varepsilon(k')}{\Delta(k')} < \frac{\Lambda(1-\nu)\sqrt{\xi}}{4\sqrt{10} + \Lambda(1-\nu)\sqrt{\xi}}, \tag{2.8.3}$$

$$|\langle\phi^\star, \phi_{(0)}\rangle_2| \leq 1 - \nu^2\xi, \tag{2.8.4}$$

$$\rho(\hat{\Sigma}_X - \Sigma_X, k') \leq c\lambda_{\min}(\Sigma_X) \tag{2.8.5}$$

hold almost surely (with ξ defined as in Lemma 8), we have that

$$\sqrt{1 - |(\phi^\star)^\top\bar{\phi}|} \leq \nu^t\sqrt{1 - |(\phi^\star)^\top\phi_{(0)}|} + \frac{1}{1-\nu}\sqrt{10}\frac{\sqrt{2}}{\Lambda\big(\Delta(k') - \varepsilon(k')\big)}\varepsilon(k')$$

almost surely.

Proof: Firstly we demonstrate the proposition for $\phi_{t,k}$ in Algorithm 2.

For any $k \in \mathbb{N}$ let ν_k be defined as

$$\nu_k := [1 + 2(\frac{s}{k})^{1/2} + \frac{s}{k})]^{1/2}\left(1 - \frac{1+c}{8}\eta\lambda_{\min}(\Sigma_X)\frac{1}{C_\diamond\kappa(\Sigma_X)}\right)^{1/2}.$$

For any $k \in \{\underline{k}+1, \overline{k}\}$ let $\phi_{t,k}$ be as in step 4.*b*)*vi.B.* in Algorithm 2. Let $F_{t,k} = \mathrm{supp}(\phi_{t,k})$, $F_{t-1,k} = \mathrm{supp}(\phi_{t-1,k})$ and $V = \mathrm{supp}(\phi^\star)$. Moreover let $F^{(t,k)} = F_{t,k} \cup F_{t-1,k} \cup V$. Note that thanks to step 4.*b*)*v.* in Algorithm 2, the cardinality of $F^{(t,k)}$ is at most $\overline{k} + s$.

Now we operate like in the proof of Theorem 1 in Tan et al. (2018) in order to bound $(1 - |\langle\phi^\star, \phi_{t,k}|)^{1/2}$ for any $k \in \{\underline{k}+1, \overline{k}\}$ and $t \in \{1, ..., T\}$ (where T, $\underline{k}$ and $\overline{k}$ are as in Algorithm 2).

Let $\phi'_{t,k}$ be the as in step 4.*b*)*iii.* of Algorithm 2. Let $F^{(t,k)}$ be defined as above. Note that

$$P_{(F^{(t,k)})}P^\top_{(F^{(t,k)})}\phi_{t-1,k} = \phi_{t-1,k},$$

and that same equality holds for $\phi(F^{(t,k)})$. Moreover, in the notations of Lemma 8, observe that if $\tilde{\phi} = P_{(F^{(t,k)})}^{\top}\phi_{t-1,k}$ then $\phi' = P_{(F^{(t,k)})}^{\top}\phi'_{t,k}$. Hence we let $\tilde{\phi} = P_{(F^{(t,k)})}^{\top}\phi_{t-1,k}$ and apply Lemma 8. To this aim we need to verify the condition on the initial vector $\tilde{\phi}$. To do so we apply inductive reasoning. Suppose that the condition on the initial vector $\tilde{\phi}$ holds, then by Lemma 8 we have

$$\phi(F^{(t,k)})^{\top}\phi'_{t,k} \geq \\ \phi(F^{(t,k)})^{\top}\phi_{t-1,k} + \frac{1+c}{8}\eta\lambda_{\min}(\Sigma_X)(1-\phi(F^{(t,k)})^{\top}\phi_{t-1,k})(\frac{1}{C_{\diamond}\kappa(\Sigma_X)})$$

which implies (by subtracting 1 and rearranging the terms)

$$1-\phi(F^{(t,k)})^{\top}\phi'_{t,k} \leq \big(1-\phi(F^{(t,k)})^{\top}\phi_{t-1,k}\big)\big(1-\frac{1+c}{8}\eta\lambda_{\min}(\Sigma_X)\frac{1}{C_{\diamond}\kappa(\Sigma_X)}\big). \tag{2.8.6}$$

We prove a series of inequalities. Without loss of generality, we assume the inner products between $\phi(F^{(t,k)})$, $\phi^{\star}$ and $\phi'_{t,k}$ are positive everywhere on the set of events we are working on (because otherwise, we can simply treat the events where this inner product is negative by making appropriate sign changes in the proof). Hence inequality (2.8.6) is equivalent to

$$\|\phi(F^{(t,k)})-\phi'_{t,k}\|_2 \leq \|\phi(F^{(t,k)})-\phi_{t-1,k}\|_2\big(1-\frac{1+c}{8}\eta\lambda_{\min}(\Sigma_X)\frac{1}{C_{\diamond}\kappa(\Sigma_X)}\big)^{1/2}. \tag{2.8.7}$$

Moreover, by the triangle inequality, we have

$$\|\phi^{\star}-\phi'_{t,k}\|_2 \tag{2.8.8}$$

$$\begin{aligned}
&\leq \|\phi(F^{(t,k)})-\phi'_{t,k}\|_2 + \|\phi(F^{(t,k)})-\phi^{\star}\|_2 \\
&\leq \|\phi(F^{(t,k)})-\phi_{t-1,k}\|_2\big(1-\frac{1+c}{8}\eta\lambda_{\min}(\Sigma_X)\frac{1}{C_{\diamond}\kappa(\Sigma_X)}\big)^{1/2} + \|\phi(F^{(t,k)})-\phi^{\star}\|_2 \\
&\leq \|\phi^{\star}-\phi_{t-1,k}\|_2\big(1-\frac{1+c}{8}\eta\lambda_{\min}(\Sigma_X)\frac{1}{C_{\diamond}\kappa(\Sigma_X)}\big)^{1/2} + 2\|\phi(F^{(t,k)})-\phi^{\star}\|_2
\end{aligned} \tag{2.8.9}$$

where the second inequality comes from inequality (2.8.7). Inequality (2.8.8) is equivalent to

$$\sqrt{1-|(\phi^{\star})^{\top}\phi'_{t,k}|} \leq \sqrt{1-|(\phi^{\star})^{\top}\phi_{t-1,k}|}\left(1-\frac{1+c}{8}\eta\lambda_{\min}(\Sigma_X)\frac{1}{C_{\diamond}\kappa(\Sigma_X)}\right)^{1/2} \\ + 2\sqrt{1-|\phi(F^{(t,k)})^{\top}\phi^{\star}|} \tag{2.8.10}$$

Let $\hat{\phi}_{t,k}$ be as in step 4.*b*).*vi*.*A*. of Algorithm 2. By Lemma 9, we have

$$\begin{aligned}
&\sqrt{1-|(\phi^\star)^\top\hat{\phi}_{t,k}|} && (2.8.11)\\
&\leq \left(1-|(\phi^\star)^\top\phi'_{t,k}|+\big((s/k)^{1/2}+s/k\big)\big(1-|(\phi^\star)^\top\phi'_{t,k}|^2\big)\right)^{1/2}\\
&\leq \sqrt{1-|(\phi^\star)^\top\phi'_{t,k}|}\left(1+\big((s/k)^{1/2}+s/k\big)\big(1+|(\phi^\star)^\top\phi'_{t,k}|\big)\right)^{1/2}\\
&\leq \sqrt{1-|(\phi^\star)^\top\phi'_{t,k}|}\left(1+2\big((s/k)^{1/2}+s/k\big)\right)^{1/2}\\
&\leq \nu_k\sqrt{1-|(\phi^\star)^\top\phi_{t-1,k}|}+\sqrt{20}\sqrt{1-|\phi(F^{(t,k)})^\top\phi^\star|} && (2.8.12)
\end{aligned}$$

where the third inequality holds since $|(\phi^\star)^\top\phi'_{t,k}|\leq 1$ and the fourth inequality holds by inequality (2.8.10).

Finally, since $\phi_{t,k}=\hat{\phi}_{t,k}/\|\hat{\phi}_{t,k}\|_2$ and that by construction $\|\hat{\phi}_{t,k}\|_2\leq 1$, we have

$$\begin{aligned}
\sqrt{1-|(\phi^\star)^\top\phi_{t,k}|} &\leq \sqrt{1-|(\phi^\star)^\top\hat{\phi}_{t,k}|}\\
&\leq \nu_k\sqrt{1-|(\phi^\star)^\top\phi_{t-1,k}|}+\sqrt{20}\sqrt{1-|\phi(F^{(t,k)})^\top\phi^\star|}\\
&\leq \nu_k\sqrt{1-|(\phi^\star)^\top\phi_{t-1,k}|}+\sqrt{20}\frac{\sqrt{2}}{\Lambda\big(\Delta(k')-\varepsilon(k')\big)}\varepsilon(k') \qquad (2.8.13)
\end{aligned}$$

where the second inequality holds by inequality (2.8.11) and the third inequality holds by Lemma 10 that we can apply since inequality (2.8.3) implies that $\Lambda>\varepsilon(k')/\Delta(k')$. Starting from now, we differ from Tan et al. (2018) and verify that if $\phi_{t-1,k}$ satisfies the condition of Lemma 8, so does $\phi_{t,k}$. Let

$$\omega(k'):=\frac{\sqrt{2}}{\Lambda\big(\Delta(k')-\varepsilon(k')\big)}\varepsilon(k')$$

By the triangle inequality, we have

$$\|\phi(F^{(t,k)})-\phi_{t,k}\|_2\leq\|\phi^\star-\phi_{t,k}\|_2+\|\phi(F^{(t,k)})-\phi^\star\|_2$$

this implies that

$$\begin{aligned}
\sqrt{1-|\phi(F^{(t,k)})^\top\phi_{t,k}|} &\leq \sqrt{1-|(\phi^\star)^\top\phi_{t,k}|}+\sqrt{1-\phi(F^{(t,k)})^\top\phi^\star}\\
&\leq \nu_k\sqrt{\xi}+2\sqrt{20}\omega(k')\\
&\leq \sqrt{\xi}
\end{aligned}$$

where the second inequality holds by inequality (2.8.13) and Lemma 10, and the third inequality holds by inequality (2.8.3). This proves that if the assumption of Lemma 8 is satisfied for $t-1$ then it is for t. Since we have assumed that $|\phi_{(0)}^\top \phi^\star| \geq 1-\nu^2\xi$, namely that the initial vector satisfies the assumption of Lemma 8, it implies by induction, that the assumption is satisfied for all $t \in \mathbb{N}$. Hence, by applying inequality (2.8.13) recursively, we obtain

$$\sqrt{1-|(\phi^\star)^\top \phi_{t,k}|} \leq \nu_k^t \sqrt{1-|(\phi^\star)^\top \phi_{0,k}|} + \frac{1-\nu_k^t}{1-\nu_k}\sqrt{20}\omega(k')$$

and since $\nu \geq \nu_k$ we have

$$\sqrt{1-|(\phi^\star)^\top \phi_{t,k}|} \leq \nu^t \sqrt{1-|(\phi^\star)^\top \phi_{0,k}|} + \frac{1}{1-\nu}\sqrt{20}\omega(k')$$

which was needed to be proven.

The proof for ϕ_t in Algorithm 1 follows by replacing $\phi_{t,k}$ by ϕ_t, $\phi'_{t,k}$ by ϕ'_t, $\hat{\phi}_{t,k}$ by $\hat{\phi}_t$, $\phi_{0,k}$ by ϕ_0 (where ϕ'_t, $\hat{\phi}_t$, ϕ_0 are as in Algorithm 1) and $F_{t,k}$ by $F_t := \mathrm{supp}(\phi_t)$, $F_{t-1,k}$ by $F_{t-1} := (\phi_{t-1,k})$ and $F^{(t,k)}$ by $F^{(t)} := F_t \cup F_{t-1} \cup V$ (and noting that the cardinality of $F^{(t)}$ is at most $2k+s$).

□

Proof of Theorem 2

Let T be as in Algorithm 2 and let

$$\omega(k') := \frac{\sqrt{2}}{\Lambda\big(\Delta(k') - \varepsilon(k')\big)}\varepsilon(k').$$

We aim to apply Proposition 7 in order to bound $\sqrt{1-|(\phi^\star)^\top \phi_{t,k}|}$ for any $t \in \mathbb{N}$ and $k \in \{\underline{k}, \overline{k}\}$.

Suppose that for some $k \in \{\underline{k}+1, \overline{k}-1\}$, in the notation of Proposition 7, the condition on $\phi_{(0)}$ is satisfied (that is a condition on $\phi_{0,k}$). Hence by inequality (2.8.13) and the reasoning thereafter in the proof of Proposition (7), we know that $\phi_{T,k}$ also satisfies this condition. Hence since by Algorithm 2, $\phi_{0,k-1} = \phi_{T,k}$, Proposition 7 may be applied to bound $\sqrt{1-|(\phi^\star)^\top \phi_{T,k-1}|}$. Hence, by induction, our assumption on $\phi_{0,\overline{k}}$, ensures that we can apply Proposition 7 in order to bound $\sqrt{1-|(\phi^\star)^\top \phi_{t,k}|}$ for any $t \in \mathbb{N}$ and $k \in \{\underline{k}, \overline{k}\}$.

Finally, since $\omega(k_1) \geq \omega(k_2)$ if $k_1 > k_2$ and since $t \geq T$ by recursively applying Proposition 7, we have

$$\sqrt{1-|(\phi^\star)^\top \phi_{t,\underline{k}}|} \leq \nu^{t+[\overline{k}-\underline{k}]T} \sqrt{1-|(\phi^\star)^\top \phi_{0,\overline{k}}|} + \frac{1-\nu^{[\overline{k}-\underline{k}]T}}{1-\nu^T}\frac{1}{1-\nu}\sqrt{20}\omega(k')$$

from which the result is obtained. □

2.8.2 Proof of Proposition 3

Lemmas

Remark 4 *It is common in the literature to characterize sub-gaussian and sub-exponential random variable through the finiteness of their* Orlicz norm *defined as*

$$\|Z\|_{\Psi_\alpha} := \{C > 0; \mathbb{E}\exp(\frac{Z^\alpha}{C^\alpha}) \leq 2\}.$$

From Definitions 5.7 and 5.13 in Vershynin (2012) one easily verify that Z is sub-gaussian if and only if $\|Z\|_{\Psi_2} < \infty$, and is sub-exponential if and only if $\|Z\|_{\Psi_1} < \infty$.

In the proofs we need the following definition.

Definition 3 (Definition 5.1 in Vershynin (2012)) (Nets, covering numbers)

Let (U, d) be a metric space and let $\epsilon > 0$. A subset $\mathcal{N}_\epsilon$ of U is called a ϵ-net of U if every point $u \in U$ can be approximated to within ϵ by some point $v \in \mathcal{N}_\epsilon$, i.e. so that $d(u, v) \leq \epsilon$. The minimal cardinality of a ϵ-net of U, if finite, is denoted $\mathcal{N}(U, \epsilon)$ and is called the covering number of U (at scale ϵ).

Lemma 11 *Assume that $X_{11}, ..., X_{1p}$ are sub-gaussian random variables and that $p \to \infty$ as $n \to \infty$. Assume that $X = (X_1, ..., X_p) \in \mathbb{R}^{n \times p}$ where each row are independent and identically to $(X_{11}, ..., X_{1p})$. Let $s \in \mathbb{N}$, $k = Cs$ for $C > 0$, $\Sigma_X = \mathbb{E}(X_1 X_1^\top)$ and $\hat{\Sigma}_X = X^\top X/n$.*

Then if $s\log(p/s)/n = o(1)$ and if $s/p = o(1)$ then

$$\rho(\hat{\Sigma}_X - \Sigma, k) = O_p(\sqrt{\frac{s\log(p/s)}{n}})$$

where $\rho(\cdot, \cdot)$ is defined as in section 4.3..

Proof.

$$
\begin{aligned}
&\mathbb{P}\Big(\sup_{\|v\|_0\leq k,\|v\|_2=1} |\langle(\hat{\Sigma}_X-\Sigma_X)v,v\rangle_2|>t\Big)\\
&=\mathbb{P}\Big(\sup_{P\in\mathcal{P}(\mathcal{F}(p,k)),\|v\|_2=1} |\langle(\hat{\Sigma}_X-\Sigma_X)Pv,Pv\rangle_2|>t\Big)\\
&\leq \sum_{P\in\mathcal{P}(\mathcal{F}(p,k))}\mathbb{P}\Big(\sup_{\|v\|_2=1} |\langle(\hat{\Sigma}_X-\Sigma_X)Pv,Pv\rangle_2|>t\Big) \qquad (2.8.14)\\
&\leq \sum_{P\in\mathcal{P}(\mathcal{F}(p,k))}\mathbb{P}\Big(2\sup_{v\in\mathcal{N}_{\frac{1}{4}}} |\langle(\hat{\Sigma}_X-\Sigma_X)Pv,Pv\rangle_2|>t\Big) \qquad (2.8.15)\\
&\leq \sum_{P\in\mathcal{P}(\mathcal{F}(p,k))}\sum_{v\in\mathcal{N}_{\frac{1}{4}}}\mathbb{P}\big(2|\langle(\hat{\Sigma}_X-\Sigma_X)Pv,Pv\rangle_2|>t\big) \qquad (2.8.16)\\
&= \sum_{P\in\mathcal{P}(\mathcal{F}(p,k))}\sum_{v\in\mathcal{N}_{\frac{1}{4}}}\mathbb{P}\big(2|\frac{1}{n}\langle XPv,XPv\rangle_2-\langle\Sigma_XPv,Pv\rangle_2|>t\big)\\
&= \sum_{P\in\mathcal{P}(\mathcal{F}(p,k))}\sum_{v\in\mathcal{N}_{\frac{1}{4}}}\mathbb{P}\big(2|\frac{1}{n}\|XPv\|_2^2-\langle\Sigma_XPv,Pv\rangle_2|>t\big). \qquad (2.8.17)
\end{aligned}
$$

Where inequalities (2.8.14) and (2.8.16) follow by union bounds, and where inequality (2.8.15) follows by Lemma 5.2 in Vershynin (2012) with scale $\epsilon = 1/4$. Now we treat the terms inside the probability on the right-hand side of inequality (2.8.17). Remark that for all $P\in\mathcal{P}(\mathcal{F}(p,k))$ and all $v\in\mathcal{N}_{\frac{1}{4}}$ we have $\|XPv\|_2^2=\sum_{i=1}^n\langle P^\top X_i,v\rangle_2^2$ (by definition of the canonical norm on $\mathbb{R}^n$), that

$$
\begin{aligned}
\mathbb{E}\langle P^\top X_i,v\rangle_2^2 &= \mathbb{E}(\langle P^\top X_i,v\rangle_2\langle P^\top X_i,v\rangle_2)\\
&= \mathbb{E}(\langle P^\top X_i\langle P^\top X_i,v\rangle_2,v\rangle_2)\\
&= \langle\mathbb{E}(P^\top X_i\langle P^\top X_i,v\rangle_2),v\rangle_2\\
&= \langle P^\top\Sigma_XPv,v\rangle_2
\end{aligned}
$$

and that $\|\langle P^\top X_i,v\rangle_2\|_{\Psi_2}\leq\sum_{j=1}^k\|X_{i,\alpha_j}\|_{\Psi_2}<\infty$ since we have assumed that the components of X_i are sub-gaussian. Hence we have that

$$
\frac{1}{n}\sum_{i=n}^n(\langle P^\top X_i,v\rangle_2^2-\langle P^\top\Sigma_XPv,v\rangle_2)
$$

is an iid sum of centered sub-exponential. Then by Corollary 5.17 in Vershynin (2012), for all $P\in\mathcal{P}(\mathcal{F}(p,k))$ and for all $v\in\mathcal{N}_{1/4}$ there exists an absolute constant $c>0$ and $K_{P,v}=\sup_{l\leq1}l^{-1}(\mathbb{E}(\langle P^\top X_i,v\rangle_2^2))^{1/l}$ such that

$$
\mathbb{P}\big(|\frac{1}{n}\sum_{i=n}^n(\langle P^\top X_i,v\rangle_2^2-\langle P^\top\Sigma Pv,v\rangle_2)|>t/2\big)\leq 2e^{-\frac{nc}{2}\min\{\frac{t}{K},\frac{t^2}{2K^2}\}}.
$$

Let $K = \min_{P\in\mathcal{P}(\mathcal{F}(p,k)),v\in\mathcal{N}_{1/4}} K_{P,v}$, then by inequality (2.8.17), we conclude that there exists an absolute constant $c > 0$ such that

$$\mathbb{P}(\sup_{\|v\|=1,\|v\|_0\leq 0} |\langle(\hat{\Sigma}_X - \Sigma_X)v, v\rangle_2| > t) \leq 9^k C_p^k 2e^{-\frac{nc}{2}\min\{\frac{t}{K},\frac{t^2}{2K^2}\}} \tag{2.8.18}$$

$$\leq (\frac{9ep}{k})^k 2e^{-\frac{nc}{2}\min\{\frac{t}{K},\frac{t^2}{2K^2}\}} \tag{2.8.19}$$

$$= 2e^{k\log(\frac{9ep}{k})-\frac{nc}{2}\min\{\frac{t}{K},\frac{t^2}{2K^2}\}}.$$

Where inequality (2.8.18) follows by Lemma 5.4 in Vershynin (2012) and the fact that $\text{Card}(\mathcal{P}(\mathcal{F}(p,k))) = C_p^k$, and inequality (2.8.19) follows since $C_p^k \leq (\frac{ep}{k})^k$.

Now we show the bound in probability at the desired rate. For $C > 0$, we have

$$\mathbb{P}(\sup_{\|v\|=1,\|v\|_0\leq 0} |\langle(\hat{\Sigma}_X - \Sigma_X)v, v\rangle_2| > C\sqrt{\frac{k\log(p/k)}{n}})$$

$$\leq 2e^{k\log(\frac{9ep}{k})-\frac{nc}{2}\min\{\frac{C}{K}\sqrt{\frac{k\log(p/k)}{n}},\frac{C^2}{2K^2}\frac{k\log(p/k)}{n}\}}$$

and it is easy to check that for all $C > 0$ if $\frac{4K^2}{C^2} > \frac{k\log(p/k)}{n}$ then

$$\min\{\frac{C}{K}\sqrt{\frac{k\log(p/k)}{n}}, \frac{C^2}{2K^2}\frac{k\log(p/k)}{n}\} = \frac{C^2}{2K^2}\frac{k\log(p/k)}{n}$$

and the bound becomes

$$\mathbb{P}(\sup_{\|v\|=1,\|v\|_0\leq 0} |\langle(\hat{\Sigma}_X - \Sigma_X)v, v\rangle_2| > C\sqrt{\frac{\log(p/k)}{n}})$$

$$\leq 2e^{k\left(\log(\frac{9ep}{k})-\frac{c}{4}\frac{C^2}{K^2}\log(p/k)\right)}$$

$$\leq 2e^{k\left(\log(9e)+\left(1-\frac{c}{4}\frac{C^2}{K^2}\right)\log(p/k)\right)}$$

And for C such that $1 < \frac{c}{4}\frac{C^2}{K^2}$ the right hand-side can be made arbitrarily small by increasing p/k. Hence since $p \to \infty$ as $n \to \infty$, $k/p = o(1)$, $\frac{k\log(p/k)}{n} = o(1)$ and $k = C's$ for some $C' > 0$, the thesis is proven. □

Lemma 12 *Assume that $Y_1, X_{11}, ..., X_{1p}$ are sub-gaussian random variables. Assume that $(Y, X) = (Y, X_1, ..., X_p) \in \mathbb{R}^{n\times 1+p}$ where each row are independent and identically distributed to $(Y_1, X_{11}, ..., X_{1p})$. Let $s \in \mathbb{N}$, $k = Cs$ for $C > 0$, $\Sigma_{XY} = \mathbb{E}(X_1Y_1)$, $\Sigma_{YX} = \Sigma_{XY}^\top$, $\hat{\Sigma}_{XY} = X^\top Y/n$ and $\hat{\Sigma}_{YX} = \hat{\Sigma}_{XY}^\top$.*

Then if $s\log(p/s)/n = o(1)$ and if $s/p = o(1)$ then

$$\rho(\hat{\Sigma}_{XY}\hat{\Sigma}_{YX} - \Sigma_{XY}\Sigma_{YX}, k) = O_p(\sqrt{\frac{k\log(p/k)}{n}}).$$

where $\rho(\cdot,\cdot)$ is defined as in section 4.3..

Proof. We have

$$\begin{aligned}
&\sup_{\|v\|_2=1;\|v\|_0=k} |v^\top \hat{\Sigma}_{XY}\hat{\Sigma}_{YX}v - v^\top \Sigma_{XY}\Sigma_{YX}^\top v| \\
&= \sup_{\|v\|_2=1;\|v\|_0=k} |(v^\top \hat{\Sigma}_{XY})^2 - (v^\top \Sigma_{XY})^2| \\
&= \sup_{\|v\|_2=1;\|v\|_0=k} |v^\top \hat{\Sigma}_{XY} - v^\top \Sigma_{XY}||v^\top \hat{\Sigma}_{XY} + v^\top \Sigma_{XY}| \\
&= \sup_{\|v\|_2=1;\|v\|_0=k} |v^\top \hat{\Sigma}_{XY} - v^\top \Sigma_{XY}| \sup_{\|v\|_2=1;\|v\|_0=k} |v^\top \hat{\Sigma}_{XY} + v^\top \Sigma_{XY}|
\end{aligned}$$

And it is easy to check (using the same techniques that in the proof of Lemma 11) that if $p \to \infty$ as $n \to \infty$, $k/p = o(1)$, $\frac{k\log(p/k)}{n} = o(1)$ and $k = Cs$ for some $C > 0$

$$\sup_{\|v\|_2=1;\|v\|_0=k} |v^\top \hat{\Sigma}_{XY} - v^\top \Sigma_{XY}| = O_p(\sqrt{\frac{s\log(p/s)}{n}})$$

and that

$$\sup_{\|v\|_2=1;\|v\|_0=k} |v^\top \hat{\Sigma}_{XY} + v^\top \Sigma_{XY}| = O_p(1).$$

then the proof follows from the fact that

$$O_p(R_n)O_p(1) = O_p(R_n).$$

□

Proof of Proposition 3

The result follows directly from Lemmas 11 and 12. □

2.8.3 Proof of Lemma 1

$$\|(I_{p\times p} - \eta\hat{\Sigma}_X)P\|_{\text{spec}}^2 = \sup_{\|v\|_2=1} v^\top P^\top (I_{p\times p} - \eta\hat{\Sigma}_X)^2 Pv.$$

Since for all v s.t. $\|v\|_2 = 1$ we have

$$v^\top P^\top (I_{p\times p} - \eta\hat{\Sigma}_X)^2 Pv = 1 - \eta\Big(2v^\top P^\top \hat{\Sigma}_X Pv - \eta v^\top P^\top \hat{\Sigma}_X^2 Pv\Big),$$

since $P^\top \hat{\Sigma}_X P$ is positive definite by assumption, there exists $\eta_v^\alpha > 0$ small enough such that

$$2v^\top P^\top \hat{\Sigma}_X Pv - \eta^\alpha v^\top P^\top \hat{\Sigma}_X^2 Pv > 0$$

and hence there exists $\eta_v^\beta > 0$ small enough such that

$$0 \leq 1 - \eta_v^\beta \left(2v^\top P^\top \hat{\Sigma}_X P v - \eta_v^\alpha v^\top P^\top \hat{\Sigma}_X^2 P v \right) < 1$$

the thesis follows from choosing $\eta = \inf_{\|v\|_2=1} \eta_v$ where $\eta_v = \inf\{\eta_v^\alpha, \eta_v^\beta\}$. □

2.8.4 Proof of Lemma 3

For some $M > 0$, consider

$$\mathbb{P}\left(\sup_{\tilde{t}\in\{1,\dots,t(n)\}} \left| \frac{\phi_{\tilde{t}}^\top \hat{\Sigma}_X \phi_{\tilde{t}}}{\phi_{\tilde{t}}^\top \hat{\Sigma}_X \phi^\star + \phi_{\tilde{t}}^\top \frac{X^\top U}{\|\beta\|_2 n}} - 1 \right| > M \right)$$

which is bounded above by

$$\mathbb{P}\left(\frac{\sup_{\tilde{t}\in\{1,\dots,t(n)\}} |\phi_{\tilde{t}}^\top \hat{\Sigma}_X (\phi_{\tilde{t}} - \phi^\star) - \phi_{\tilde{t}}^\top \frac{X^\top U}{\|\beta\|_2 n}|}{\inf_{\tilde{t}\in\{1,\dots,t(n)\}} |\phi_{\tilde{t}}^\top \hat{\Sigma}_{XY}|} | > \frac{M}{\|\beta\|_2} \right)$$

let $M' = M/\|\beta\|_2$, then this probability is bounded above by

$$\mathbb{P}\left(\sup_{\tilde{t}\in\{1,\dots,t(n)\}} |\phi_{\tilde{t}}^\top \hat{\Sigma}_X (\phi_{\tilde{t}} - \phi^\star)| + \| \frac{X^\top U}{\|\beta\|_2 n} \|_2 - M' \inf_{\tilde{t}\in\{1,\dots,t(n)\}} |\phi_{\tilde{t}}^\top \hat{\Sigma}_{XY}| > 0 \right) \tag{2.8.20}$$

and for all $\tilde{t}$ we have

$$\begin{aligned} |\phi_{\tilde{t}}^\top \hat{\Sigma}_{XY}| &= |\phi_{\tilde{t}}^\top \Sigma_{XY} - \phi_{\tilde{t}}^\top (\Sigma_{XY} - \hat{\Sigma}_{XY})| \\ &\geq |\phi_{\tilde{t}}^\top \Sigma_{XY}| - |\phi_{\tilde{t}}^\top (\Sigma_{XY} - \hat{\Sigma}_{XY})| \\ &\geq \inf_{\|v\|_2=1;\|v\|_0=k} |v^\top \Sigma_{XY}| - \sup_{\|v\|_2=1;\|v\|_0=k} |v^\top \hat{\Sigma}_{XY} - v^\top \Sigma_{XY}| \end{aligned}$$

hence the probability in (2.8.20) is bounded above by

$$\begin{aligned} \mathbb{P}\Bigg(& \sup_{\tilde{t}\in\{1,\dots,t(n)\}} |\phi_{\tilde{t}}^\top \hat{\Sigma}_X (\phi_{\tilde{t}} - \phi^\star)| + \| \frac{X^\top U}{\|\beta\|_2 n} \|_2 \\ & + M' \sup_{\|v\|_2=1;\|v\|_0=k} |v^\top \hat{\Sigma}_{XY} - v^\top \Sigma_{XY}| - M' \inf_{\|v\|_2;\|v\|_0=k} |v^\top \Sigma_{XY}| > 0 \Bigg). \end{aligned}$$

And for M and n high enough this probability can be bounded by any $\epsilon > 0$ since

$$\begin{aligned} \sup_{\tilde{t}\in\{1,\dots,t(n)\}} |\phi_{\tilde{t}}^\top \hat{\Sigma}_X (\phi_{\tilde{t}} - \phi^\star)| &= O_p(1), \\ \| \frac{X^\top U}{\|\beta\|_2 n} \|_2 &= o_p(1), \\ \sup_{\|v\|_2=1;\|v\|_0=k} |v^\top \hat{\Sigma}_{XY} - v^\top \Sigma_{XY}| &= o_p(1). \end{aligned}$$

Which is what we needed to prove. □

2.8.5 Proof of Lemma 5

Let x be any output of Algorithm 1 or 2 2, then following Definition 2 we have

$$Q_{i,j}(x) = \begin{cases} 0 \text{ if } i \neq j \text{ or if } x_i = 0 \\ 1 \text{ else.} \end{cases}$$

where $Q_{i,j}(x)$ is the element of the i^{th} row and j^{th} column of $Q(x)$.

Firstly, we show there exists $\epsilon_1 > 0$ s.t. for any x as above if $\|x - \phi^\star\| < \epsilon_1$ then $Q(x) \in \mathcal{Q}^\star$ almost surely.

Suppose the opposite. This implies there exists $i \in \{1, ..., p\}$ s.t. $x_i = 0$ and $\phi_i^\star \neq 0$. However $|\phi_i^\star| < \|x - \phi^\star\|_2 < \epsilon_1$. Hence by choosing ϵ_1 to be the lowest non-null element of $|\phi^\star|_2$ we have reached a contradiction.

Secondly, our assumption on the convergence of ϕ_t implies that for all $h > 0$ $\|\phi_{t+h} - \phi^\star\| < \epsilon_2$ and the same reasoning applies. Hence the thesis is proven. □

2.8.6 Proof of Lemma 6

By adding zero and the triangle inequality we have

$$\|\phi_t - \tilde{\phi}_t\|_2 \leq \|\phi_t - \phi^\star\|_2 + \|\phi^\star - \phi_{t-1}\|_2 + \|\phi_{t-1} - \tilde{\phi}_t\|_2.$$

By assumptions on the convergence of ϕ_t we have $\|\phi_t - \phi^\star\|_2 = o_p(1)$ and $\|\phi_{t-1} - \phi^\star\|_2 = o_p(1)$. It remains to show that $\|\tilde{\phi}_t - \phi_{t-1}\|_2 = o_p(1)$.

By definition of $\tilde{\phi}_t$ and Equation (2.7.5) in the proof of Proposition 4 gives us

$$\begin{aligned} \tilde{\phi}_t - \phi_{t-1} &= \eta \frac{1}{g_{t-1}} \hat{\Sigma}_X \phi^\star - \eta \hat{\Sigma}_X \phi_{t-1} + \eta \frac{1}{g_{t-1}} \frac{X^\top U}{\|\beta\|_{\mathbb{R}^p} n} \\ &= \eta \hat{\Sigma}_X ([\frac{1}{g_{t-1}} - 1]\phi^\star + \phi^\star - \phi_{t-1}) + \eta \frac{1}{g_{t-1}} \frac{X^\top U}{\|\beta\|_{\mathbb{R}^p} n}. \end{aligned}$$

Hence since we have $\phi_{t-1} \xrightarrow{p} \phi^\star$, that by Lemma 2 we have $\frac{1}{g_{t-1}} \xrightarrow{p} 1$ and that by our assumption we have $\|\frac{X^\top U}{\|\beta\|_2 n}\|_2 = o_p(1)$ then the boundedness in probability of $\hat{\Sigma}_X$ ensures the thesis follows. □.

2.8.7 Proof of Lemma 7

By the triangle inequality we have

$$\sup_{t \in \{1,...,t(n)\}} \|\phi_t - \tilde{\phi}_t\|_2 \leq 1 + \sup_{t \in \{1,...,t(n)\}} \|\tilde{\phi}_t\|_2.$$

By Equation (2.7.5) in the proof of Proposition 4 we have

$$\begin{aligned}\sup_{t\in\{1,\dots,t(n)\}}\|\tilde{\phi}_t\|_2 &\leq 1+\sup_{t\in\{1,\dots,t(n)\}}\frac{\eta}{\rho_{t-1}}\|(\hat{\Sigma}_{XY}\hat{\Sigma}_{YX}-\rho_{t-1}\hat{\Sigma}_X)\phi_{t-1}\|_2\\ &\leq 1+\sup_{t\in\{1,\dots,t(n)\}}\frac{\eta}{\rho_{t-1}}\|\hat{\Sigma}_{XY}\hat{\Sigma}_{YX}\|_{\text{spec}}+\eta\|\hat{\Sigma}_X\|_{\text{spec}}.\end{aligned}$$

Since $\hat{\Sigma}_{XY}\hat{\Sigma}_{YX}$ and $\hat{\Sigma}_X$ are bounded in probability since $\mathbb{E}\|X_1\|_2^2<\infty$ by assumption. It remains to note that

$$\begin{aligned}\sup_{t\in\{1,\dots,t(n)\}}\frac{1}{\rho_{t-1}} &\leq \sup_{t\in\{1,\dots,t(n)\}}\frac{\|\hat{\Sigma}_X\|_{\text{spec}}}{\phi_{t-1}^\top\hat{\Sigma}_{XY}\hat{\Sigma}_{YX}\phi_{t-1}}\\ &\leq \frac{\|\hat{\Sigma}_X\|_{\text{spec}}}{\inf_{t\in\{1,\dots,t(n)\}}|\phi_{t-1}^\top\hat{\Sigma}_{XY}|^2}\end{aligned}$$

which can be shown to be $O_p(1)$ by a similar strategy than in Lemma 3 since

$$\begin{aligned}\inf_{t\in\{1,\dots,t(n)\}}|\phi_{t-1}^\top\hat{\Sigma}_{XY}|^2 \geq &\inf_{\|v\|_2;\|v\|_0=k}|v\Sigma_{XY}|\\ &-2\sup_{\|v\|_2;\|v\|_0=k}|v\Sigma_{XY}|\sup_{\|v\|_2;\|v\|_0=k}|v(\Sigma_{XY}-\hat{\Sigma}_{XY})|\end{aligned}$$

and $\sup_{\|v\|_2;\|v\|_0=k}|v(\Sigma_{XY}-\hat{\Sigma}_{XY})|=o_p(1)$. Hence the thesis is proven. □

Three

Goodness-of-fit test for high-dimensional linear sparse model selection

3.1 Introduction

High-dimensional sparse models are today used in all fields of quantitative sciences. They correspond to linear models with a large number of covariates, among which only a relatively small subset is significant. Those models are used for instance in genomics (Carvalho et al., 2008), in engineering (Jing et al., 2007), and in economics (Fan et al., 2011), to name but a few trends of applications.

As a consequence of this interest in high-dimensional sparse models, a huge number of methods have been proposed to select the significant subset of variables. Among the long list of methods, we cite the LASSO (Tibshirani, 1996) and its extension the Elastic-Net (Zou and Hastie, 2005), the (two-steps) Iterative Hard Thresholding (Jain et al., 2014), the SDAR (Huang et al., 2018) and the Loaded-RIFLE studied in Chapter 2.

The purpose of this chapter is to provide a goodness-of-fit procedure in order to test whether the outcome of a variable selection procedure is statistically valid or not.

This chapter is related to a large piece of literature that considers the problem of hypothesis testing in high dimensions. Most of this literature considers testing the significance of (a fraction of) a high-dimensional estimator that is some sort or another of penalized least-squared program. For instance Bühlmann (2013) concentrates on the Ridge Estimator, van de Geer, Bühlmann, Ritov and Dezeure (2014) considers the LASSO on which they use a de-sparsification technique to obtain a trackable test statistic, Meinshausen (2015) assumes that the *true* high-dimensional parameter satisfies a l_1 penalty (that is typical of LASSO), Shah and Bühlmann (2018) consider the estimation is conducted via LASSO and utilizes some sort of bootstrap to obtain

critical values for the significance test. More recently Ning and Liu (2017) considers a more general approach based on $M-$estimators that are penalized via an abstract penalty function. Finally, we mention the work of Gold, Lederer and Tao (2020) that considers the problem of testing the estimator of a high-dimensional instrumental variable regression and that also assumes the estimation is conducted via LASSO.

In contrast, our test does not suppose any particular estimator and only supposes that (1) the true high-dimensional parameter is sparse, (2) that a subset of covariates has been selected (without assuming any particular selection procedure). Our goodness-of-fit test statistic aims, as it is usual, to minimize a moment condition that, in our context, reflects the sparsity constraint. From this point of view, our approach may be connected with the one of Meinshausen (2015), since it may be argued that our method assumes the true parameter satisfies some particular l_0 constraint. Testing this constraint statistically is possible thanks to a high-dimensional Central Limit Theorem we prove below under heteroskedasticity. However, the use of the Central Limit Theorem needs the estimation and the inversion of a high-dimensional asymptotic covariance matrix. In order to circumvent the instability of this inversion we provide a procedure based on a Multiple-splitting projection test (MPT) (Liu et al., 2022). This procedure is based on a data-driven regularization by projection scheme. Since it estimates the optimal projection space this scheme requires splitting the data in order to conserve the tractability of the asymptotic distribution of the regularized test statistic. Data splitting reduces the sample size available and hence decreases the power of the test. To circumvent this feature and improve the power of the test MPT operates multiple data-splitting and applies a technology of $p-$value combination (Liu et al., 2022). The final test is reasonably powerful and allows to control the type I error in finite sample size.

An upstream result addresses the question of collinearity in the set of non-significant covariates. We provide a new identification condition for the parameters of a sparse linear model that is of independent interest.

The new goodness-of-fit test is studied empirically both via Monte Carlo experiments and in an application in macroeconomics. This application considers the nowcasting of GDP, which is a daily prediction procedure in Central banks around the world. Our results consider the four model selection methods cited above. We question the number of variables selected by those methods. This application of our test is interesting since the recent works of Giannone, Lenza and Primiceri (2021) argues that, notably in the context of nowcasting, the sparsity assumption most often does not hold. Our empirical results interestingly show that, in our case, it is possible to select a minimal number of variables such that every larger set of variables would pass the goodness-of-fit test. This minimal number, that is between 31 and 35 in our application to Belgian GDP nowcast, may be considered to be the best variable selection in

the method considered.

The second Section of this chapter provides the theory while the third is concerned with Monte Carlo experiment and an application to a real dataset. The fourth Section concludes.

3.2 Necessary and sufficient conditions for sparsity

Let $y \in \mathbb{R}$ and $x \in \mathbb{R}^p$ be some random variables and suppose there exists non-random $\beta \in \mathbb{R}^p$ such that

$$\mathbb{E}(y|x) = x^\top \beta. \tag{3.2.1}$$

Choose k, out of p, covariates. Let $x_\circ \in \mathbb{R}^k$ be the vector containing those variables and let w be the vector containing the variables that have been left out. Hence we can rewrite Equation (3.2.1) as

$$\mathbb{E}(y|x) = \mathbb{E}(y|x_\circ, w) = x_\circ^\top \beta_z + w^\top \beta_w \tag{3.2.2}$$

where $\beta_\circ \in \mathbb{R}^k$ and β_w^{p-k} are the components of β that are associated with $x_\circ$ and w respectively.

In the following result, we give minimal conditions under which $\beta_\circ$ is identified while w has no effect on $\mathbb{E}(y|x)$.

Before stating the result, we introduce some notations. Let $A \in \mathbb{R}^{m\times n}$ and let $\text{range}(A) := \{v \in \mathbb{R}^m | v = Au, u \in \mathbb{R}^n\}$ be the *range* of A, let $\text{rank}(A) := \dim\big(\text{range}(A)\big)$ be the *rank* of A and let $\ker(A) := \{v \in \mathbb{R}^n | Av = 0\}$ be the *kernel* of A.

Theorem 4 *Let $y \in \mathbb{R}, x_\circ \in \mathbb{R}^k$ and $w \in \mathbb{R}^{p-k}$ be random variables. Assume there exists $\beta_\circ \in \mathbb{R}^k$ and $\beta_w \in \mathbb{R}^{p-k}$ such that $\mathbb{E}(y|x_\circ, w) = x_\circ^\top \beta_\circ + w^\top \beta_w$. Consider the following statements*

1. $w^\top \beta_w = 0$ *a.s.*
2. $\mathbb{E}\Big(w\big(y - x_\circ^\top \mathbb{E}(x_\circ x_\circ^\top)^{-1}\mathbb{E}(x_\circ y)\big)\Big) = 0$
3. $\beta_\circ = \mathbb{E}(x_\circ x_\circ^\top)^{-1}\mathbb{E}(x_\circ y)$.

If $\text{rank}\big(\mathbb{E}(x_\circ x_\circ^\top)\big) = k$ *and*

$$\text{range}\big(\mathbb{E}(ww^\top)\big) \cap \text{range}\big(\mathbb{E}(wx_\circ^\top)\big) = \{0\} \tag{3.2.3}$$

then

$$1 \Leftrightarrow 2 \Rightarrow 3,$$

if in addition $\text{range}\big(\mathbb{E}(wx_\circ^\top)\big) \neq \{0\}$

$$1 \Leftrightarrow 2 \Leftrightarrow 3.$$

The proof is given in the Appendix.

Theorem 4 provides a moment condition equivalent to the requirement $w^\top \beta_w = 0$, that is statement (2). From a geometrical point of view, there are no difficulties with this result. Actually, the equivalence may be obtained via geometric reasoning quicker than via the proof we provided in the Appendix. Indeed statement (2) comes to say that w is perpendicular to the orthogonal complement of the orthogonal projection of y onto the space of squared integrable linear functions of x and this implies that, from a geometrical point of view, it is equivalent to project y on $(x_\circ, w)$ or on $x_\circ$ alone. If the proof we gave sort of looks like a detour from this more direct argument, it is because we also obtain a result concerning the identification of the parameters in the particular setup that are sparse linear models.

In the literature, sparsity of linear models is often introduced by imposing a constraint on the number of non-zero entries of β (*e.g.* Belloni et al. (2013), van de Geer (2016)). In the setup of Theorem, 4 this way of assuming the parsimony of the model would translate in the assumption $\beta_w = 0$. However doing so comes to assume β_w to be unique, that is to implicitly assume some identification condition such as postulating that $\mathbb{E}\big((x_\circ, w)^\top (x_\circ, w)\big)$ is invertible. We believe that this route is, from an identification point of view, too narrow. Specifically, we believe that in a sparse linear model, only the parameters associated with the important covariates are of interest (and hence should be identified) while the parameters associated with the non-important regressors may be considered as nuisances and thus might be left unidentified. Hence, in Theorem 4, we enlarge slightly the notion of parsimony by studying the statement $w^\top \beta_w = 0$ instead of the statement $\beta_w = 0$. Because the former statement does not require β_w to be identified, we are able to provide a weaker identification condition than the usual rank condition on the matrix of the second moment of the covariates. This weaker condition is contained in Equation (3.2.3) once it is coupled with the statement 1 (or equivalently statement 2) and a rank condition on $\mathbb{E}(x_\circ x_\circ^\top)$. Under those conditions, the model allows the presence of collinearity between the variables contained in w as long as there is no collinearity between $x_\circ$ and w.

In the next section, we use Theorem 4 to construct a statistical test of the hypothesis that there are no important variables that have been left out by the selection procedure.

3.3 Results for inference

In this section, we employ Theorem 4 to provide a Central Limit Theorem that allows us to test if the set of variables we selected does not left out any important variables. To this end, we introduce some observations.

Assume we access n independent observations of y and x. Let $Y = (y_1, ..., y_n)^\top$ and $X = [x_1, ..., x_n]^\top \in \mathbb{R}^{n \times p}$ be respectively the vector and

matrix that contain those observations. We introduce high dimensionality by assuming that p, the number of covariates, is a function of n, the number of observations, and that $n = o(p)$.

As before we assume that we select k out of p covariates to be candidates for being important variables. Since p is a function of n, it may also be the case that k is a function of n, hence to hold on to the sparsity assumption we suppose that $k = o(n)$. The observation of these selected covariates will be stacked in the matrix $X_\circ = [x_{\circ 1}, ..., x_{\circ n}]^\top \in \mathbb{R}^{n\times k}$. The observations of the remaining covariates will be contained in $W = [w_1, ..., w_n]^\top \in \mathbb{R}^{n\times(p-k)}$.

We suppose there exists $\beta = (\beta_\circ, \beta_w)^\top \in \mathbb{R}^p$ such that $\mathbb{E}(y_i|x_i) = x_i^\top \beta$ for all $i = 1, ..., n$. Hence we write the model as

$$Y = X_\circ \beta_\circ + W\beta_w + U$$

where $U := (y_1 - x_1^\top \beta, ..., y_n - x_n^\top \beta)^\top$.

We aim to check if the matrix $X_\circ$ contains at least all the important covariates. That is we want to check if $W\beta_w = 0$ is almost surely. For the remaining of the chapter, the proposition $W\beta_w = 0$ *holds with probability one* is our null hypothesis and will be denoted by H_0. To test H_0 we rely on Theorem 4 from which we know that $W\beta_w = 0$ a.s. is equivalent to $\mathbb{E}\Big(w_i\big(y_i - x_{\circ i}^\top \mathbb{E}(x_{\circ i}x_{\circ i}^\top)^{-1}\mathbb{E}(x_{\circ i}y_i)\big)\Big) = 0$ for $i = 1, ..., n$.

As a statistic we choose

$$T_n := \frac{1}{n} W^\top (Y - X_\circ (X_\circ^\top X_\circ)^{-1} X_\circ^\top Y)$$

and look for a way to test if it deviates significantly from zero. To this end the following Central Limit Theorem for T_n under H_0 that is given in the next proposition is useful.

Proposition 8 *Let $n \in \mathbb{N}$ and assume that $p = p(n)$ such that $p(n) \to \infty$ as $n \to \infty$. Let $Y \in \mathbb{R}^n$, $X_\circ \in \mathbb{R}^{n\times k}$ and $W \in \mathbb{R}^{n\times p-k}$ be respectively random vector and matrices that are independent across rows. Assume there exists $\beta_\circ \in \mathbb{R}^k$ and $\beta_w \in \mathbb{R}^{p-k}$ such that for all n and $i = 1, ..., n$ we have $\mathbb{E}(y_i|x_{\circ i}, w_i) = x_{\circ i}^\top \beta_\circ + w_i^\top \beta_w$. Assume also that for all n and $i = 1, ..., n$, $\operatorname{rank}(\mathbb{E}(x_{\circ i}x_{\circ i}^\top)) = k$ and that $k = o(n)$.*

Finally, let $u_i = y_i - \mathbb{E}(y_i|x_{\circ i}, w_i)$, $x_i = (x_{\circ i}, w_i)^\top$ for $i = 1, ..., n$ and $s_{n,v_n}^2 = \sum_{i=1}^n \mathbb{E}|v_n^\top x_i u_i|^2$ for $v_n \in \mathbb{R}^p$ and assume that $\sum_{i=1}^n \mathbb{E}\|x_i\|^2 = o(n^2)$ and that for all $\varepsilon > 0$

$$\lim_{n\to\infty} \sup_{v_n \in \mathbb{R}^p \setminus \{0_p\}} \frac{1}{s_{n,v_n}^2} \sum_{i=1}^n \mathbb{E}\Big(|v_n^\top x_i u_i|^2 \cdot \mathbf{1}_{\{|v_n^\top x_i u_i| > \varepsilon s_{n,v_n}\}}\Big) = 0. \tag{3.3.1}$$

Under these conditions if H_0 is true then

$$\sqrt{n}T_n \text{ is } \mathcal{AN}_p(0, \Omega_n)$$

where

$$\Omega_n = n^{-1}\sum_{i=1}^{n}\begin{pmatrix} -\Sigma_w^{(n)}(\Sigma_{x_\circ}^{(n)})^{-1} \\ I_{(p-k)\times(p-k)} \end{pmatrix}^\top \sigma_i^2 \mathbb{C}ov(x_i) \begin{pmatrix} -\Sigma_w^{(n)}(\Sigma_{x_\circ}^{(n)})^{-1} \\ I_{(p-k)\times(p-k)} \end{pmatrix}$$

with

$$\Sigma_w^{(n)} = n^{-1}\sum_{i=1}^{n}\mathbb{E}(w_i w_i^\top) \qquad \Sigma_{x_\circ}^{(n)} = n^{-1}\sum_{i=1}^{n}\mathbb{E}(x_{\circ i}x_{\circ i}^\top)$$

$$\Sigma_{wx}^{(n)} = n^{-1}\sum_{i=1}^{n}\mathbb{E}(w_i x_{\circ i}^\top) \qquad \sigma_i^2 = \mathbb{V}ar(y_1|x_i).$$

In particular, if the observations are independent and identically distributed, we have $\Omega_n = \sigma_1^2\Big(\mathbb{E}(w_1 w_1^\top) - \mathbb{E}(w_1 x_{\circ 1})(x_{\circ 1}x_{\circ 1})^{-1}(\mathbb{E}(x_{\circ 1}w_1)\Big)$.

The proof is given in the Appendix.

The previous result provides ground for testing H_0. Indeed, if one knows Ω_n, then one may use the fact that under H_0 we have that

$$nT_n^\top \Omega_n^{-1} T_n$$

converges weakly to a random variable following a Chi-squared distribution with $p-k$ degrees of freedom. In practice, however, we rarely know the covariance matrix Ω_n. In low-dimensional settings, that is when $(p-k) = o(n)$, the seminal work of Hotelling (1931) proposes to replace Ω_n by its empirical counterpart, say $\hat{\Omega}_n$. The resulting testing procedure is known as the *Hoteling* T^2 test and has been extensively studied (*e.g.* Anderson et al. (1958)). In HD, however, that is for instance when $n = o(p-k)$, the estimation of the inverse of the covariance matrix becomes a challenging task since the spectrum of the sample covariance matrix contains zeros almost surely (see e.g. Fan et al. (2016) for a review on estimating such inverses in HD). The literature proposed different variations of the Hoteling T^2 tests in HD, none that which may be considered as optimal in all situations (see Huang et al. (2022) for a recent review).

In this chapter, we follow the recent work of Liu et al. (2022) that proposes to employ a Multiple-Splitting Projection Test (MPT) to test hypotheses on high-dimensional mean vectors.

3.3.1 Multiple-Splitting Projection Test

Projection tests for H_0

Applied to our problem, the generic idea of projection tests is to project T_n onto a subspace of $\mathbb{R}^{p-k}$ of, say, $l = o(n)$ dimensions, such that the traditional Hoteling T^2 test can be applied. Let $P \in \mathbb{R}^{(p-k)\times l}$ be a full column rank matrix, the resulting test statistic is

$$nT_n^\top P(P^\top \hat{\Omega}_n P)^{-1} P T_n \tag{3.3.2}$$

that converges weakly to a random variable following a Chi-squared distribution with l degrees of freedom under H_0 conditionally on P.

Hence Projection Tests translate the problem of the inversion of $\hat{\Omega}_n$ to the problem of finding a somehow good P. To this end a recent result due to Huang (2015) informs us that since $\sqrt{n}T_n$ is asymptotically Gaussian the direction $P = \Omega_n^{-1}\mathbb{E}(T_n) \in \mathbb{R}^{(p-k)}$ maximizes the asymptotic *power of the test* (that is the probability to correctly reject H_0 asymptotically). Unfortunately, to estimate this optimal direction one needs to solve the inverse problem $\hat{\Omega}_n \hat{P} = T_n$ and once again stumble against the inversion of $\hat{\Omega}_n$.

The recent work of Liu et al. (2022) indicates how to circumvent this difficulty. They proposed to assume that P is sparse, that is that most of its components are zeros, and to solve a regularize quadratic optimization problem in order to obtain $\hat{P}$.

Observe that, in our problem, assuming that P is sparse is not a mere technical simplification but is somehow natural. Indeed, this assumption only comes to restrict the set of available alternative hypotheses to the ones compatible with the assumption of the sparsity of the conditional expectations $\mathbb{E}(y_i|x_i)$ with $i = 1, ...n$. In other words, it comes to choose as a generic alternative hypothesis that a subset of the set of important variables has been considered as non-important but that the cardinality of this subset is $o(n)$. Hence assuming the sparsity of P only comes to reject from the set of available alternative hypotheses the ones negating the parsimony of β.

Under the assumption of the sparsity of P, Liu et al. (2022) points out that one can obtain an estimator of P by solving

$$\underset{\phi\in\mathbb{R}^{p-k}}{\operatorname{argmin}}\, \phi^\top \hat{\Omega}_n \phi - T_n^\top \phi + \mathcal{P}_\lambda(\phi) \tag{3.3.3}$$

where $\mathcal{P}_\lambda$ is a sparsity inducing penalty. We note that under the assumption of Proposition 8, and if the observations are assumed to be identically distributed, we have that $\hat{\Omega}_n = n^{-1}\hat{\sigma}^2(W^\top X_\circ (X_\circ^\top X_\circ)^{-1} X_\circ W)$, where $\hat{\sigma}^2 = (n-1)^{-1}(Y - X_\circ(X_\circ^\top X_\circ)^{-1}X_\circ^\top Y)^\top (Y - X_\circ(X_\circ^\top X_\circ)^{-1}X_\circ^\top Y)$, and the problem (3.3.3) is equivalent to

$$\underset{\phi\in\mathbb{R}^{p-k}}{\operatorname{argmin}} \| \frac{Y}{\sqrt{n}\hat{\sigma}} - \frac{\hat{\sigma}(I_{n\times n} - X_\circ(X_\circ^\top X_\circ)^{-1}X_\circ^\top)W\phi}{\sqrt{n}} \|_2^2 + \mathcal{P}_\lambda(\phi). \tag{3.3.4}$$

Fortunately problem (3.3.4) is a penalized least square program and may be solved for various penalties using any of the existing consistent estimators proposed in the abundant literature on the matter (see *e.g.* the introduction of Bertsimas et al. (2016) for a brief, but synthetic, review).

Multiple data splitting

Our strategy relies on firstly estimating P and secondly conducting a projection test on T_n. However if one uses the same sample to operate the two steps, the dependence between $\hat{P}$ and T_n makes the distribution of the statistic (3.3.2) difficult to track. Hence we follow Liu et al. (2022) and split our data into two independent subsets.

Consider the set $\mathcal{D} = \{1, ..., n\}$ and split it in two such that $\mathcal{D}_1 = \{1, ..., n_1\}$ and $\mathcal{D}_2 = \{n_1 + 1, ..., n_2\}$ where n_1 is chosen and $n_2 = n - n_1$. We use the observations indexed by $i \in \mathcal{D}_1$ to estimate P and the ones indexed by $i \in \mathcal{D}_2$ to compute $T_{n_2 i} = \hat{P}^\top w_i (I_{n_2 \times n_2} - x_{\circ i}(X_\circ^\top X_\circ)^{-1} X_\circ^\top) Y$ for $i = 1, ..., n_2$. This strategy allows us to compute

$$S_{\hat{P}} = \sqrt{n_2} \frac{P^\top T_{n_2}}{\hat{\gamma}}$$

where $\hat{\gamma}^2$ is the sample variance of $\{T_{n_2 1}, ..., T_{n_2 n_2}\}$, that converges asymptotically to a standardized Gaussian random variable under H_0 by Proposition 8. Hence the $p-$value of testing H_0 via the statistics $S_{\hat{P}}$ is given by

$$p(S_{\hat{P}}) = 2\big(1 - \Phi_G(|S_{\hat{P}}|)\big)$$

where Φ_G is the cumulative distribution function of a standard normal.

Splitting the data reduces the power of the test since it is constrained by a lower sample size and by a particular split of the data. In order to increase the power of the test and to protect ourselves from the idiosyncrasies of a particular split, we keep on following Liu et al. (2022) and operate multiple splitting of the data. Formally let $\pi_1, ..., \pi_m$ be m permutations of $\mathcal{D} = \{1, ..., n\}$ and let $\mathcal{D}^{\pi_1}, ..., \mathcal{D}^{\pi_m}$ be the corresponding m permuted sets. Hence for $j = 1, ..., m$ we have that $\mathcal{D}^{\pi_j} = \{\alpha_1, ..., \alpha_n\}$ where $\alpha_1 = \pi_j(1), ..., \alpha_n = \pi_j(n)$. Finally, for each $j \in \{1, ..., m\}$ we may split $\mathcal{D}^{\pi_j}$ in the same fashion as we splitted $\mathcal{D}$ before, and obtain $\mathcal{D}_1^{\pi_j}$ and $\mathcal{D}_2^{\pi_j}$ whose cardinality are respectively n_1 and n_2. These m splits may be used to apply the same procedure as above to obtain m $p-$values denoted by $p_1, ..., p_m$.

Combining $p-$values

The $p-$values $p_1, ..., p_m$ are dependent. To combine them Liu et al. (2022) prove that there are *exchangeable* that is $(p_1, ..., p_m)$ is identically distributed to $(p_{\pi'(i)}, ..., p_{\pi'(m)})$ for any π' being a permutation of $\{1, ..., m\}$. This exchangeability property holds both on H_0 and H_1 (Liu et al., 2022).

In order to combine those $p-$values let $Z_j = \Phi_G^{-1}(p_j)$ for $j = 1, .., m$. Under H_0 $Z_1, ..., Z_m$ are marginally standard normal and jointly exchangeable with the correlation between any disjoint pair being $\rho \geq 0$ (Liu et al., 2022). If in addition, we *assume* that $Z_1, ..., Z_m$ are multivariate normal, we are in the setup of Follmann and Proschan (2012) that provides a test of location that exactly control for the type I error. In the following, we hold on to this assumption and apply the results of Follmann and Proschan (2012). Before, we mention that without the assumption of joint normality, one needs to rely on a Central Limit Theorem for exchangeable random variables (*e.g.* Klass and Teicher (1987)) and on a consistent estimator of ρ, say $\hat{\rho}$, to obtain that

$$\frac{m^{-1}\sum_{j=1}^{m} Z_m}{\sqrt{(1+(m-1)/\hat{\rho})/m}}$$

converges weakly to a standard Gaussian random variable under H_0. However since the variance of $m^{-1}\sum_{j=1}^{m} Z_m$ is not asymptotically degenerated and is exactly equal to $(1+(m-1)/\rho)/m$, the said weak convergence happens at a slow rate (Liu et al., 2022).

Finite sample test with control over type I-error

Hence we assume that $Z_1, ...Z_m$ are jointly normal. In addition, and following Follmann and Proschan (2012), we assume for all $j = 1, ..., m$ that $Z_j = U + V_j$ where $U \sim \mathcal{N}(0, \sigma_U^2)$, $V_i \sim \mathcal{N}(0, \sigma_V^2)$, that $(U, V_1, ...V_m)$ are independent and that $\sigma_U^2 + \sigma_V^2 = 1$. Under this last set of assumptions, we have that $\rho = 1 - \sigma_V^2 \geq 0$ (Follmann and Proschan, 2012). Moreover, those assumptions are useful to control the estimation of ρ and to derive the exact distribution of

$$Z_{\hat{\rho}} := \frac{\bar{Z}}{\sqrt{\big(1+(m-1)\hat{\rho}\big)/m}}$$

where $\hat{\rho}$ is an estimator of ρ. In particular Follmann and Proschan (2012) proved that for the conservative estimates of ρ that is

$$\hat{\rho}(\tau) = \max\big(0, \frac{(m-1)s_z^2}{a}\big)$$

where s_z^2 is the empirical variance of $\{Z_1, ..., Z_m\}$ and a solves $\mathbb{P}(x > a) = \tau$ for x being a Chi-square random variable, we have under H_0 that

$$\mathbb{P}(Z_{\hat{\rho}(\tau)} > c) = \int_0^{\frac{a}{1-\rho}} \bar{\Phi}\left(c\sqrt{\frac{1+(m-1)(1-\frac{(1-\rho)t}{a})}{1-(m-1)\rho}}\right) h_{m-1}(t)dt + \bar{\Phi}\left(c\sqrt{\frac{1}{1+(m-1)\rho}}\right)\bar{H}_a\big(\frac{a}{1-\rho}\big) \quad (3.3.5)$$

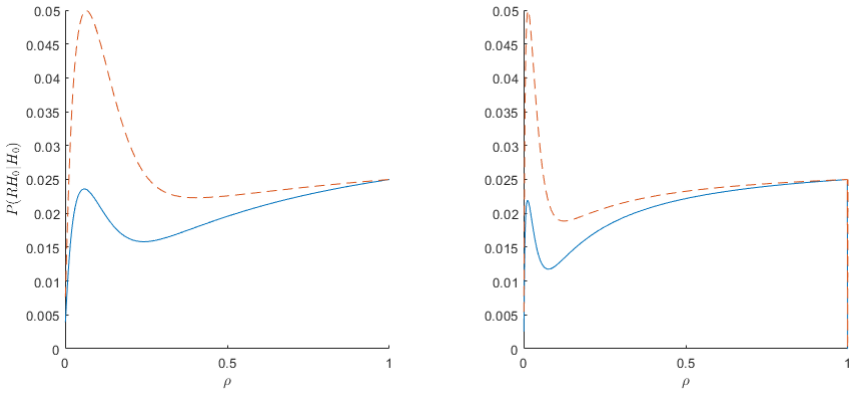

Figure 3.1: Probabilities computed by numerical integration

Behaviour of the probability in Equation (3.3.5) as a function of ρ for $m = 100$ (left) and $m = 1000$ (right), for $c^\star = 1.96$ and τ chosen such that maximum is 0.025 (solid lines) and 0.05 (dashed lines).

where $\bar{\Phi}(c)$ and $\bar{H}_{m-1}$ are $1 - \mathbb{P}(x \leq c)$ when x is respectively a standard Gaussian and a Chi-square with $m-1$ degrees of freedom and where h_{m-1} is the probability density function of a Chi-square with $m-1$ degrees of freedom.

Since the probability in Equation (3.3.5) depends on the unknown ρ, we may protect against the least favourable ρ by considering the conservative bound $\sup_{\rho\in[0,1)} \mathbb{P}(Z_{\hat{\rho}(\tau)} > c)$ (Follmann and Proschan, 2012). This ensures that the type I error is controlled for *any* ρ. In practice solving this maximization problem can be achieved by solving Equation (3.3.5) via numeric integration for various ρ.

To control the type I error in practice we fix m and choose $c = c^\star$ and find τ such that $\sup_{\rho\in[0,1)} \mathbb{P}(Z_{\hat{\rho}(\tau)} > c^\star) = \alpha$ where α is the probability of type I error. Table 3.1, gives values for τ when $c^\star = 1.96$ and $\alpha = 0.05, 0.1$ for a two-sided test.

We summarize the testing procedure in Algorithm 3.

Algorithm 3 *Inputs:* $m \in \mathbb{N}$, $c^\star > 0$, $\tau_\alpha \in (0,1)$ *(they may be obtained from Table 3.1),* $Y \in \mathbb{R}$, $X_\circ \in \mathbb{R}^{n\times k}$, $W \in \mathbb{R}^{n\times(p-k)}$, $n_1 \in \mathbb{N}$ *such that* $n_1 < n$ *and* $\mathcal{P}_\lambda(\cdot)$ *being a penalty function.*

1. *For* $j = 1, ..., m$
   a) *Randomly permute the* $\mathcal{D} = 1, ..., n$ *to obtain* $\mathcal{D}^{\pi_j}$.
   b) *Split the data such that* $\mathcal{D}_1^{\pi_j} = \{\pi_j(1), ..., \pi_j(n_1)\}$ *and* $\mathcal{D}_2^{\pi_j} = \{\pi_j(n_1+1), ..., \pi_j(n)\}$.

c) *Let* $Y^{(1)} \in \mathbb{R}^{n_1}$, $X_\circ^{(1)} \in \mathbb{R}^{n_1 \times k}$, $W^{(1)} \in \mathbb{R}^{n_1 \times (p-k)}$ *be respectively the vector and matrices that contain the row of respectively* Y, $X_\circ$ *and* W, *that are indexed by the elements of* $\mathcal{D}_1^{\pi_j}$.

d) *Let* $P_{X_\circ^{(1)}} = X_\circ^{(1)}((X_\circ^{(1)})^\top X_\circ^{(1)})^{-1}(X_\circ^{(1)})^\top$ *and solve*

$$\underset{\phi^{p-k}}{\operatorname{argmin}} \| \frac{Y^{(1)}}{\sqrt{n_1}\hat{\sigma}} - \frac{\hat{\sigma}(I_{n_1 \times n_1} - P_{X_\circ^{(1)}} W^{(1)}\phi}{\sqrt{n_1}} \|_2^2 + \mathcal{P}_\lambda(\phi)$$

where $\hat{\sigma}^2 = (n_1 - 1)^{-1}\|(Y^{(1)} - P_{X_\circ^{(1)}} Y)\|_2^2$.

e) *Let* $n_2 = n - n_1$ *and let* $Y^{(2)} \in \mathbb{R}^{n_2}$, $X_\circ^{(2)} \in \mathbb{R}^{n_2 \times k}$, $W^{(2)} \in \mathbb{R}^{n_2 \times (p-k)}$ *be respectively the vector and matrices that contain the row of respectively* Y, $X_\circ$ *and* W, *that are indexed by the elements of* $\mathcal{D}_2^{\pi_j}$.

f) *Compute* $T_{n_2 i} = \hat{P}^\top w_i (I_{n_2 \times n_2} - x_{\circ i}((X_\circ^{(2)})^\top X_\circ^{(2)})^{-1}(X_\circ^{(2)})^\top) Y^{(2)}$ *for* $i = 1, ..., n_2$.

g) *Compute*

$$S_{\hat{P}} = \sqrt{n_2} \frac{P^\top T_{n_2}}{\hat{\gamma}}$$

where $\hat{\gamma}^2$ *is the sample variance of* $\{T_{n_2 1}, ..., T_{n_2 n_2}\}$.

h) *Compute* $p_j = 2\big(1 - \Phi_G(|S_{\hat{P}}|)\big)$

i) *Compute and save* $Z_j = \Phi_G^{-1}(p_j)$.

2. *Compute*

$$\hat{\rho}(\tau) = \max\big(0, \frac{(m-1)s_z^2}{a}\big)$$

where s_z^2 *is the empirical variance of* $\{Z_1, ..., Z_m\}$ *and* a *solves* $\mathbb{P}(x > a) = \tau$ *for* x *being a Chi-square random variable.*

3. *Compute*

$$Z_{\hat{\rho}} := \frac{m^{-1}\sum_{j=1}^m Z_j}{\sqrt{\big(1 + (m-1)\hat{\rho}\big)/m}}$$

4. *If* $|Z_{\hat{\rho}}| > c^\star$ *reject* H_0.

5. *(Optional) To obtain the bound on the p−value, numerically solve*

$$p = 2 \sup_{\rho \in (0,1)} \int_0^{\frac{a}{1-\rho}} \bar{\Phi}\left(|Z_{\hat{\rho}}| \sqrt{\frac{1 + (m-1)(1 - \frac{(1-\rho)t}{a})}{1 - (m-1)\rho}}\right) h_{m-1}(t)dt$$
$$+ \bar{\bar{\Phi}}\left(|Z_{\hat{\rho}}| \sqrt{\frac{1}{1 + (m-1)\rho}}\right) \bar{H}_a\big(\frac{a}{1-\rho}\big)$$

m	2	5	10	20	40	100	1000	10000
$\tau_{0.05}$	.25	.25	.20	.20	.15	.15	.10	.05
$\tau_{0.1}$	.74	.535	.45	.39	.345	.295	.215	.17

Table 3.1: Statistical table

Value of τ that, for the given m, ensures the probability of type I error is bounded by $\alpha = 0.05, 0.1$ for a two-sided test with critical value $c^\star = 1.96$.

3.4 Experiments and applications

3.4.1 Monte Carlo experiment: assessing the power of the test

To assess the power of the test in various settings we use the following experiment.

We let $n = 100, p = 200$ and generate $X \in \mathbb{R}^{n\times p}$ *iid* across row such that $X_1 \sim \mathcal{N}_p(0, \Omega)$ and $Y|X \in \mathbb{R}^n$ such that $Y \sim \mathcal{N}_n(X\beta, \beta^\top \Sigma_X \beta/\xi)$ where ξ is the *signal-to-noise ratio* (SNR) and β is generated such that there are 10 randomly selected elements of beta that are assigned a value of 1 while the other elements of β are assigned the value zero. Next, for $k = 1, ..., 20$ we construct $X_\circ \in \mathbb{R}^{n\times k}$ and $W \in \mathbb{R}^{n\times(p-k)}$ through the following rules.

If $k \leq 10$ let $X_\circ$ contain k columns of X that are associated with a non-null element of β, and let W contain the other columns of X. If $k \geq 10$, let $X_\circ$ contain all the columns of X associated with a non-null element of β plus $k - 10$ others randomly selected columns of X, let W contain the remaining columns of X. Below, for any given k, we refer to this selection of covariates as the *Oracle's selection*.

Finally, we proceed by testing $H_0 : W^\top \beta = 0$ through the procedure describe in Algorithm 3, for $m = 100$, $\alpha = 0.1$ and $c^\star = 1.96$. We run the experiment in three different scenarios.

Scenario 5 *We took* $\Sigma_{i,j} = 0.5^{|i-j|}$. *We let the SNR be such that* $\xi = 6$.

Scenario 6 *We took the same* Σ *as in Scenario 1. We let the SNR be such that* $\xi = 2$.

Scenario 7 *We let* $W \in \mathbb{R}^{p\times p}$ *such that* $W_{i,j} = 0.5^{|i-j|}$ *and took* $\Sigma = U^\top DU$ *where* U *is the matrix of eigenvectors of* W, $D_{j,j} = 10/\sqrt{j}$ *and* $D_{i,j} = 0$ *when* $i \neq j$. *We let the SNR be such that* $\xi = 2$.

The results are displayed in Figures 3.2, 3.3 and 3.4. We draw four conclusions from the experiment. Firstly, we observe the test is powerful in all scenarios. Secondly, decreasing the signal-to-noise ratio inflates the Type II error when the selection contains most of the important variables. Thirdly,

when the covariance matrix becomes more complicated to estimate (Scenario 3) the Type II error also inflates when the selection contains most of the important variables. In both cases these losses of power may be attributed to the increased difficulty of estimating the optimal direction in step 1.(d) of Algorithm 3. Fourthly, the test procedure adequately controls for the Type I error. Finally, we also note the presence of non-marginal type II errors when $k = 1, 2$.

3.4.2 Application to synthetic datasets: comparing variable selections

Here we apply our test to selections of variables produced by three different sparsity-inducing estimators. Specifically, we consider the following methods of selection. The SDAR, for Support Detection And Root finding that is proposed by Huang et al. (2018) to approximate an l_0-penalized least squares; the two-stages IHT (Jain et al., 2014) that is designed to approximate l_0 constrained least squares programs; and finally the Loaded-RIFLE (LR) a directional l_0 constrained estimator (see Chapter 2).

We generate the data in the same fashion as in the previous section. For each method of selection, we fixed the maximum number of covariates selected by the program and run our test on the solutions they produced. The results are displayed in Figures 3.2, 3.3 and 3.4.

Generally, we may say that the experiments confirm the test is reasonably powerful. In scenario 1 in particular, all the estimators perform well at selecting important variables but the performances of the 2-steps IHT are slightly below the ones of the other estimators and this is detected by the test. Note that our experiment does *not* constitute a valid comparison between the performances of the estimators considered. We did not look for comparability in tuning the algorithms and it was useful to our experiment that some algorithms underperform in order to observe how the proposed testing procedure reacted to these differences of efficiency (that may or may not be intrinsic to the algorithms).

As shown in the previous section, the power of the test decreases when the difficulty of estimating the optimal direction in step 1.(d) of Algorithm 3 increases. As before, this may be caused by decreasing the SNR or making Σ somehow more complicated to estimate. In addition to this, the results in Scenarios 2 and 3 highlight another cause of loss of power. Recall that to estimate the optimal direction in step 1.(d) of Algorithm 3, one needs to solve a penalized/constrained least squares problem. Moreover, to obtain a selection of variables, one also solves a penalized/constrained least squares. This is the case in our experiments and the results show that, in those cases, the power of the test may decrease when the algorithms employed to solve the two problems behave too much alike. Heuristically we may say that every algorithm has its blind spots that depend on its particular sensitivity to the data. When the

selection of variables and the estimation of the optimal direction are operated using, say, the same algorithm, the same blind spot that affected the former operation, affects the latter operation. In other words, if some important variables have not been selected by the algorithm in the selection step, it is more likely not to be selected in the estimation of the optimal direction if the two problems are solved by algorithms with similar sensitivity to the data. This situation may be observed on the results of the experiment in Scenarios 2 and 3. In the experiments, we solved the problem of estimating the optimal direction via LR. This is why in Scenario 3, the testing procedure is less powerful when it comes to testing H_0 for the selections of the variable made by LR and more powerful when it comes to testing H_0 for selections made by SDAR and 2−steps IHT. In Scenario 2, the selections of SDAR and LR look alike (this may be checked via various metrics) thus SDAR and LR display a similar sensibility to the data. As a consequence, the power of testing H_0 for the selections made by these algorithms decreases, while the power of testing the Oracle's selection and the selection made by 2-steps IHT remains reasonably high. This insight is of great importance for practitioners since, fortunately, this loss of power may be avoided by using different types of algorithms for the selection of variables and for the estimation of the optimal direction.

Figures 3.5, 3.6 and 3.7 show that the testing procedure do indeed control for the level of Type I error in finite sample.

3.4.3 Application: assessing the quality of the selection prior to nowcasting macroeconomic time series

Motivations

Key economic variables such as GDP, aggregate private consumption, or sectoral added value are typically published by economic institutions at low frequency (quarterly) and with delay. Early predictions and estimates of these key variables may be achieved by using the information provided in data published at higher frequencies (*e.g.* monthly). This problem is known as *nowcasting* and is common and important in many institutions such as central banks and research centers (*e.g.* Giannone et al. (2008), Banbura et al. (2010), Bok et al. (2018)). One line of research in nowcasting and forecasting focuses on the use of large and high-dimensional datasets (see *e.g.* Stock and Watson (2002), Stock and Watson (2006), Barhoumi et al. (2008), Bok et al. (2018)). However, both theoreticians and practitioners argue that increasing the number of predictors may hurt the prediction (*e.g.* Boivin and Ng (2006), Caggiano et al. (2011)). They also demonstrate that selecting a subset of variables from a large dataset and then forecasting using the selected subset of predictors lead to better forecast (e.g. Boivin and Ng (2006), Bai and Ng (2008), Caggiano et al. (2011), Kim and Swanson (2014)). There is no con-

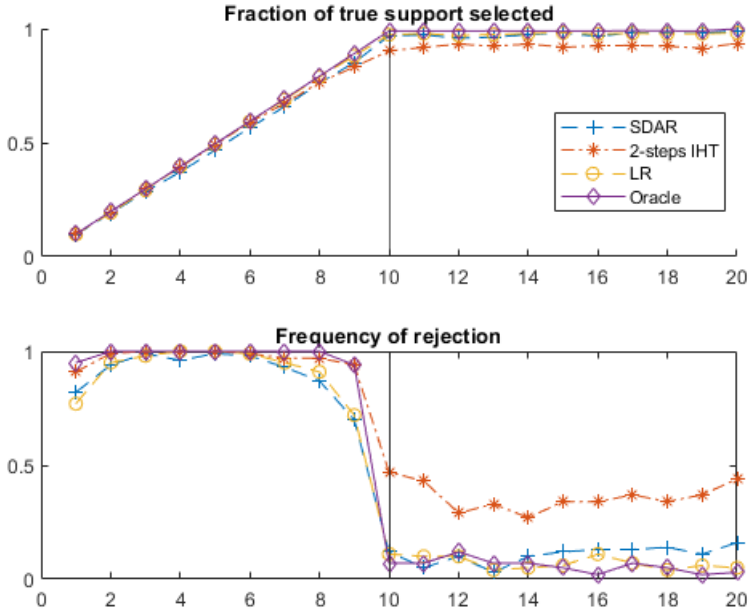

Figure 3.2: Scenario 1

(Top) Fraction of the true support of β selected by the three estimators as well as by the Oracle estimator. (Bottom) Frequency of rejection of H_0 based on the selection operated by four estimators.

sensus on how to select the predictor but penalized/constrained least squares programs are definitely one of the main methods in the toolbox of practitioners (*e.g.* Bai and Ng (2008), Bessec (2013), Piette et al. (2014), Girardi et al. (2017), Cepni et al. (2020)). Often those programs provide different selections of predictors. Moreover, most of these programs depend on tuning parameters that play a key role in the outcome of the variable selection. Hence arises the question of the statistical validity of a particular selection.

Application: predictor selection for the nowcasting of Belgian aggregate private consumption

We consider GDP nowcasting in Belgium. Four procedures for variable selection are used to select sets of predictors of the first difference of the seasonally adjusted *log* GDP measured in constant price. The potential predictors are 471 monthly time series. These time series are publicly available and belong to the set used in the quarterly nowcast of the GDP provided by the Institute of Economic and Social Research from the Université Catholique de Louvain (Atanasi et al., 2020). They contain a sample of business and consumer surveys (for a discussion on the role of survey data as a predictor of macroeconomic

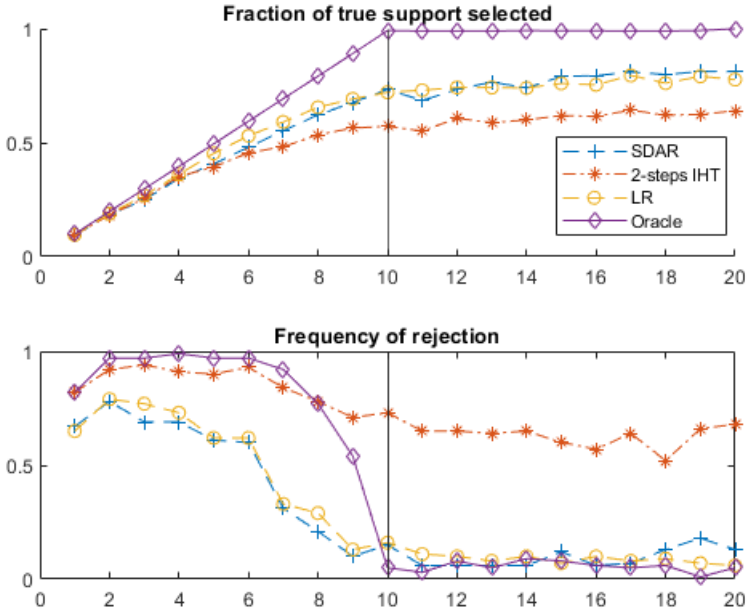

Figure 3.3: Scenario 2

(Top) Fraction of the true support of β selected by the three estimators as well as by the Oracle estimator. (Bottom) Frequency of rejection of H_0 based on the selection operated by four estimators.

times series see *e.g.* Piette et al. (2014) or Girardi et al. (2016) and reference thereafter), economic indices as well as real, financial and monetary variables. The sample period range from the first quarter of 2004 to the second quarter of 2022 both for monthly and quarterly data. All the data are downloaded from Macrobond's data services.

When monthly time series are available quarterly time series are constructed by taking the average over each quarter. We transform the obtained time series as in Piette et al. (2014), that is the logarithm is taken for every strictly positive times series and apply a first difference to every times series displaying a unit-root. Finally, we z-score all the covariates.

We run Elastic-net (Zou and Hastie, 2005), SDAR, 2-steps IHT and LR to select at least k predictors ($1 \leq k \leq 48$) and compute the $p-$values for all the selections from Algorithm 3 with $m = 1000$ and probability of type I error bounded by 0.05. We also compute the R^2 of the post-selection linear regression (that is the projection of the explained variable on the selected covariates) for each selection and plot the $p-$value against the R^2 in Figure 3.8.

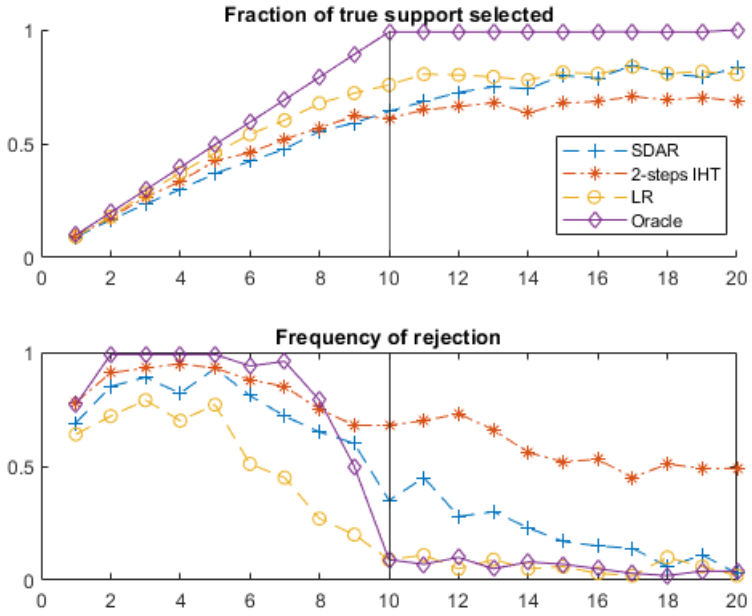

Figure 3.4: Scenario 3

(Top) Fraction of the true support of β selected by the three estimators as well as by the Oracle estimator. (Bottom) Frequency of rejection of H_0 based on the selection operated by four estimators.

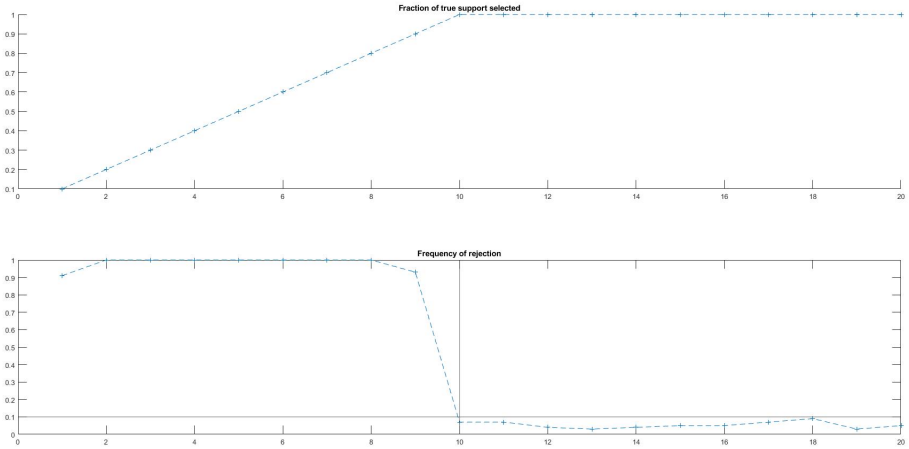

Figure 3.5: Illustration of the control on the error of Type I (scenario 5)

Illustration of the control on the error of Type I in scenario 5. The selection considered is the Oracle' selection.

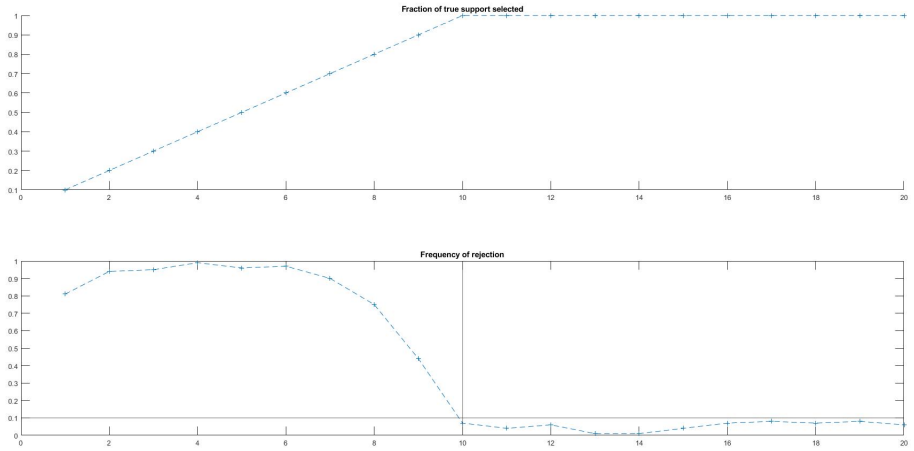

Figure 3.6: Illustration of the control on the error of Type I (scenario 6)

Illustration of the control on the error of Type I in scenario 6. The selection considered is the Oracle' selection.

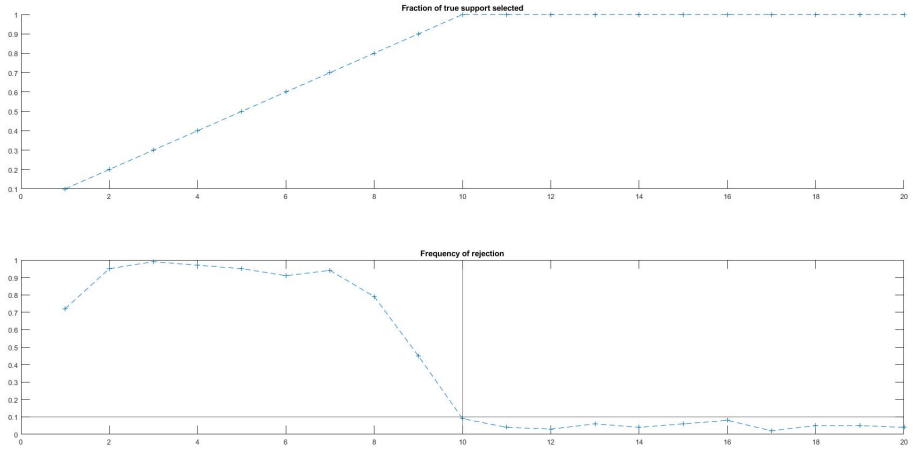

Figure 3.7: Illustration of the control on the error of Type I (scenario 7)

Illustration of the control on the error of Type I in scenario 7. The selection considered is the Oracle' selection.

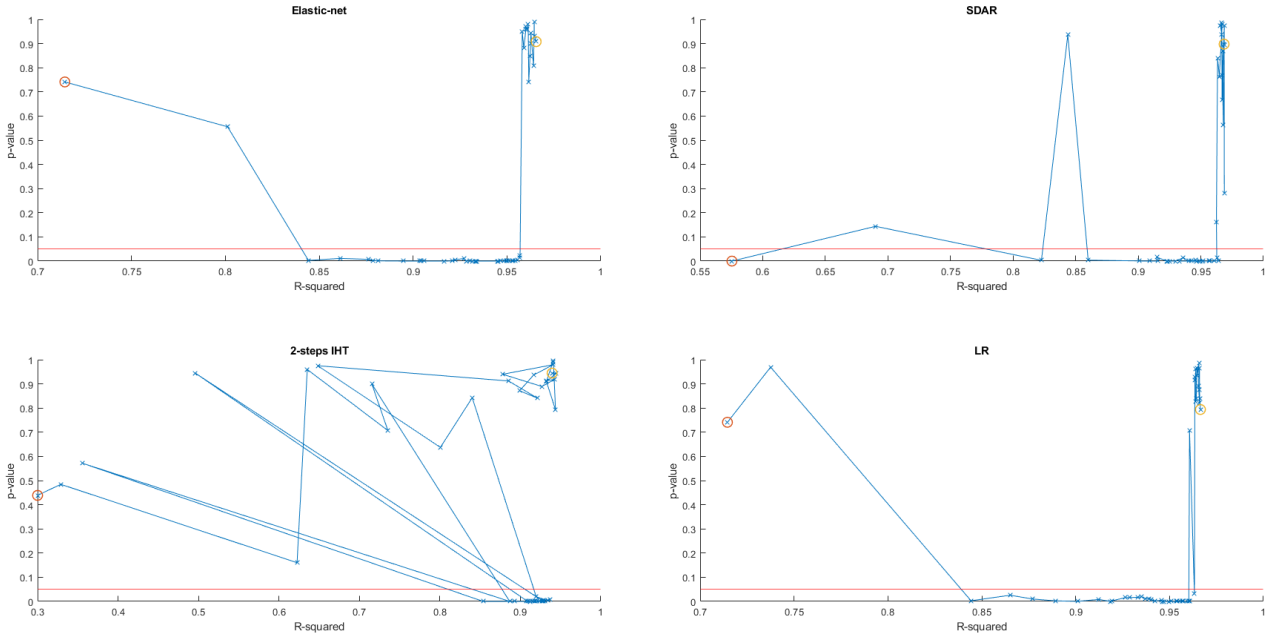

Figure 3.8: Results of the goodness-of-fit tests on real data

$p-$value plotted against R^2 for 48 selections operated by the estimator named in the title of the sub-graph. The model with one variable is surrounded by a red circle, and the model with 48 variables is surrounded by an orange circle. Each dot is connected by a line to the dots corresponding to selections one $k-1$ and $k+1$ selected variables. The red horizontal line corresponds to a $p-$value of 0.05.

Figure 3.8 displays interesting patterns. Firstly, for each estimator, there are $\bar{k} \in \mathbb{N}$ such that for all $k \geq \bar{k}$ the selections have high $p-$value and high R^2. This is a desirable feature that is expected from a selection encompassing all the important variables of a linear model. Secondly, for each estimator (except 2-step IHT), the selections for $5 < k < \bar{k}$ are consistently rejected by the test. For $k \geq \bar{k}$ they are consistently no longer rejected. This behavior is desirable since it displays the consistency of the test across the sequence of selection and indicates the test is indeed able to capture a qualitative change in the selections. It also suggests considering as fake negatives the non-rejection of H_0 for selections operated by LASSO and SDAR for $k = 1, 2$. The increased probability of fake negative for low k had already been observed in controlled experiments in the previous section. Thirdly, there is coherency between the results of the four different estimators since for each estimator $\bar{k} \in \{31, ..., 35\}$.

Those findings indicate that our statistical test is able to provide coherent and meaningful information concerning the selection of variables on real datasets. They suggest choosing the set of predictors of the Belgian GDP that correspond to selections with $k \geq \bar{k}$. This suggestion is non-trivial since operating the selection via a cross-validated Elastic-net only leads to selecting 11 covariates, a model that is rejected by our test and is far below $\bar{k}$.

3.5 Conclusion

In this chapter, we test if a given subset of the covariates of a high-dimensional linear model contains all the important regressors (H_0). Testing this hypothesis is important in the context of a high-dimensional linear *sparse* model where the existence of such subset is assumed *a priori*.

Our theoretical contributions are twofold. On the one hand, we provide theoretical results that reduce the problem of testing H_0 to the problem of testing the location of the mean of some measurable vector. One of our theorems proves a novel identification condition for the parameter of a sparse linear model that allows for collinearity in the set of non-important covariates. This last result is of independent interest. On the other hand, we provide a test based on multiple splitting (MPT) (Liu et al., 2022). Other regularization strategies for testing the location of high-dimensional mean vectors are available (see Huang et al. (2022) for a recent review) and may be as well employed to solve the problem. We validate the power of the proposed test via Monte Carlo experiments and apply our method to a real dataset by tackling the problem of selecting predictors to nowcast the Belgian GDP.

In this chapter, we have assumed the exogeneity of the parameters. However the theoretical study of endogeneity, as well as of instrumental variables (IV) regressions, in high dimensions are questions at the edge of the research frontier in econometrics (*e.g.* Belloni et al. (2012), Fan and Liao (2014), Chernozhukov et al. (2015*b*), Chernozhukov et al. (2015*a*), Gold et al. (2020)).

Further research might explore the possibility to provide an extension of our theoretical results to the case of an IV regression and study the opportunity it would open for inference.

3.6 Appendix

3.6.1 Preliminary results

Theorem 5 *Let $n \in \mathbb{N}$ and $p : \mathbb{N} \to \mathbb{N}, p(n) = p$ such that $p \to \infty$ as $n \to \infty$. For any n, let $U_1^{(n)}, ..., U_n^{(n)}$ be a sequence of $\mathbb{R}^p-$valued* independent *random vectors and let $\mu_i^{(n)} := \mathbb{E}(U_i^{(n)})$. Moreover for any n and any $v_n \in \mathbb{R}^p$ let $B_{n,v}^2 = \sum_{i=1}^n \mathbb{E}|v_n^\top (U_i^{(n)} - \mu_i^{(n)})|^2$. If*

$$\sup_{v_n \in \mathbb{R}^p \setminus \{0_p\}} \frac{1}{B_{n,v}^2} \sum_{i=1}^n \mathbb{E}\left(|v_n^\top (U_i^{(n)} - \mu_i^{(n)})|^2 \cdot \mathbf{1}_{\{|v_n^\top (U_i^{(n)} - \mu_i^{(n)})| > \varepsilon B_{n,v}\}} \right) \to 0 \tag{3.6.1}$$

as $n \to \infty$, then

$$\sum_{i=1}^n U_i^{(n)} \text{ is } \mathcal{AN}\left(\sum_{i=1}^n \mu_i^{(n)}, \sum_{i=1}^n \mathbb{C}ov(U_i) \right)$$

Proof:

Observe that to any sequence $\{v_i \in \mathbb{R}^p, i = 1, 2, ...\}$ correspond to a triangular array of the form

$$\begin{aligned} &v_1^\top U_1^{(1)} \\ &v_2^\top U_1^{(2)}, v_2^\top U_2^{(2)} \\ &v_3^\top U_1^{(3)}, v_3^\top U_2^{(3)}, v_3^\top U_3^{(3)} \\ &... \end{aligned}$$

For any of such sequence, the Lindeberg-Feller Central Limit Theorem for double array states that under the Lindeberg condition

$$\frac{1}{B_{n,v}^2} \sum_{i=1}^n \mathbb{E}\left(|v_n^\top (U_i^{(n)} - \mu_i^{(n)})|^2 \cdot \mathbf{1}_{\{|v_n^\top (U_i^{(n)} - \mu_i^{(n)})| > \varepsilon B_{n,v}\}} \right) \to 0$$

as $n \to \infty$ we have

$$\sum_{i=1}^n v_n^\top U_i^{(n)} \text{ is } \mathcal{AN}\left(\sum_{i=1}^n v^\top \mu_i^{(n)}, \sum_{i=1}^n v_n^\top \mathbb{C}\text{ov}(U_i) v_n \right).$$

(see *e.g.* page 31 in Serfling (1980)). Since assumption (3.3.1) ensures that such Lindenberg conditions hold for *all* sequences $v_1, v_2, v_3, ...$ the result follows. □

3.6.2 Proof of Theorem 4

Since $\mathbb{E}(y|x_\circ, w) = x_\circ^\top \beta_\circ + w^\top \beta_w$ we obtain the following system of equation

$$\begin{cases} \mathbb{E}(x_\circ y) = \mathbb{E}(x_\circ x_\circ^T)\beta_\circ + \mathbb{E}(x_\circ w^\top)\beta_w \\ \mathbb{E}(wy) = \mathbb{E}(wx_\circ^T)\beta_\circ + \mathbb{E}(ww^\top)\beta_w \end{cases} \tag{3.6.2}$$

To prove that statement 1 implies statement 2 it suffices to consider $w^\top \beta = 0$ in the system (3.6.2) and to substitute $\beta_\circ$.

To prove that statement 2 implies statement 1, observe that from the first equation of system (3.6.2) we have

$$\beta_\circ = \mathbb{E}(x_\circ x_\circ^\top)^{-1}\mathbb{E}(x_\circ y) - \mathbb{E}(x_\circ x_\circ^\top)^{-1}\mathbb{E}(x_\circ w^\top)\beta_w \tag{3.6.3}$$

substituting $\beta_\circ$ in the second equation of system (3.6.2), we have

$$\mathbb{E}(wy) = \mathbb{E}(wx_\circ^\top)(\mathbb{E}(x_\circ x_\circ^\top)^{-1}\mathbb{E}(x_\circ y) - \mathbb{E}(x_\circ x_\circ^\top)^{-1}\mathbb{E}(x_\circ w^\top)\beta_w) + \mathbb{E}(ww^\top)\beta_w$$

rearranging the terms we get

$$\mathbb{E}(wy) - \mathbb{E}(wx_\circ^\top)\mathbb{E}(x_\circ x_\circ^\top)^{-1}\mathbb{E}(x_\circ y) = \\ \mathbb{E}(ww^\top)\beta_w - \mathbb{E}(wx_\circ^\top)\mathbb{E}(x_\circ x_\circ^\top)^{-1}\mathbb{E}(x_\circ w^\top)\beta_w.$$

By statement 2 the left-hand side of the previous equation is zero, hence we have

$$\big(\mathbb{E}(ww^\top) - \mathbb{E}(wx_\circ^\top)\mathbb{E}(x_\circ x_\circ^\top)^{-1}\mathbb{E}(x_\circ w^\top)\big)\beta_w = 0 \tag{3.6.4}$$

We use this equation to show that $\beta_w \in \ker(\mathbb{E}(ww^\top))$

Assume that $\beta_w \notin \ker(\mathbb{E}(ww^\top))$, then by Equation (3.6.4) have

$$\mathbb{E}(ww^\top)\beta_w = \mathbb{E}(wx_\circ^\top)\mathbb{E}(x_\circ x_\circ^\top)^{-1}\mathbb{E}(x_\circ w^\top)\beta_w \neq 0$$

but this implies there exists $v \in \{\text{range}(\mathbb{E}(ww^\top)) \cap \text{range}(\mathbb{E}(wx_\circ^\top))\} \setminus \{0\}$ which is false by assumption. Hence, $\beta_w \in \ker(\mathbb{E}(ww^\top))$.

Since $\mathbb{E}(ww^\top\beta) = 0$, for all non-random $v \in \mathbb{R}^{p-k}$ we have

$$\begin{aligned} \langle \mathbb{E}(ww^\top \beta_w), v\rangle_2 &= \mathbb{E}(\langle w\langle w, \beta_w\rangle_2 v\rangle_2) \\ &= \mathbb{E}(\langle w, v\rangle_2 \langle w, \beta_w\rangle_2) \\ &= 0 \end{aligned}$$

and in particular this is true for all v such that $\langle w, \beta_w\rangle_2 = \langle w, v\rangle_2$ almost surely, hence we have $\mathbb{E}(\langle w, \beta_w\rangle_2^2) = 0$ which implies that $\langle w, \beta_b\rangle_2 = 0$.

The proof that statement 2 implies statement 3 is simply considering Equation (3.6.3) while $w^\top \beta_w = 0$. Finally, observed from the same Equation that statement 3 combined with the assumption $\mathbb{E}(wx_\circ^\top)$ implies statement 1. □

3.6.3 Proof of Proposition 8

Let $U = (y_1 - x_{\circ 1}^\top \beta_\circ, ..., y_n - x_{\circ n}^\top \beta_\circ)^\top$. We have that

$$\begin{aligned}\sqrt{n}T_n &= n^{-1/2} W^\top (Y - X_\circ (X_\circ^\top X_\circ)^{-1} X_\circ^\top Y) \\ &= n^{-1/2} (W^\top X_\circ \beta_\circ + W^\top U - W^\top X_\circ \beta_\circ - W^\top X_\circ (X_\circ^\top X_\circ)^{-1} X_\circ^\top U) \\ &= n^{-1/2} (W^\top U - W^\top X_\circ (X_\circ^\top X_\circ)^{-1} X_\circ^\top U).\end{aligned}$$

Let $\Gamma_n := W^\top X_\circ (X_\circ^\top X_\circ)^{-1}$. Since $\sum_{i=1}^n \mathbb{E}\|x_i\|^2 = o(n^2)$, $\text{rank}(\mathbb{E}(x_{\circ 1} x_{\circ 1}^\top)) = k$ and $k = o(n)$ the weak Law of Large Number for independent variables combined with Slutsky's lemma ensures that we have $\Gamma_n \xrightarrow{p} \Gamma := \Sigma_w^n (\Sigma_{x_\circ}^n)^{-1}$ (see e.g. Serfling (1980)). Let $Q_n := (-\Gamma_n, I_{(p-k)\times p-k}) \in \mathbb{R}^{p-k\times p}$ then we have that $Q_n \xrightarrow{p} Q := (-\Gamma, I_{p-k\times p-k})$. Now let $X := (Z, W)$ and observe that

$$\begin{aligned}\sqrt{n}T_n &= n^{-1/2} Q_n X^\top U \\ &= n^{-1/2} Q_n \sum_{i=1}^n X_i U_i.\end{aligned}$$

Then by Theorem 3.3.1, that is a multivariate CLT for double asymptotic, we have that

$$n^{-1/2} Q_n \sum_{i=1}^n X_i U_i \text{ is } \mathcal{AN}\left(0, \frac{1}{n} Q \sum_{i=1}^n \sigma_i^2 \mathbb{C}\text{ov}(x_i) Q^\top\right).$$

□

Four

Multivariate Locally Stationary Wavelet Processes

4.1 Introduction

The analysis of time series often requires the assumption of *(second-order* or *covariance) stationarity* (*e.g.*, Brillinger (2001)). Let $X(t)$ with $t \in \mathbb{Z}$ be an univariate, *zero-mean*, real-valued, time series. In general the *autocovariance* of $X(t)$, that is $C_X(t,\tau) = \mathbb{C}\text{ov}(X(t), X(t+\tau))$ for $\tau \in \mathbb{Z}$, is a function of both the time t and lag τ. The *covariance stationarity* of the time series is the assumption that autocovariance does not vary with t. Such stationary stochastic processes may be written as

$$X(t) \overset{L^2}{=} \int_{-\pi}^{\pi} A(\omega) \exp(i\omega t) d\zeta(\omega)$$

where $A(\omega)$ is the amplitude of the process and $d\zeta(\omega)$ is an orthogonal increment process (e.g. Priestley (1981)). Moreover, under mild conditions, the autocovariance has the following representation, which is indeed invariant in t

$$C_X(t,\tau) = C_X(\tau) = \int_{-\pi}^{\pi} f(\omega) \exp(i\omega\tau) d\omega,$$

where f_X is the *spectral density* of $X(t)$.

Despite the fact that stationarity (here and thereafter in the sense of second-order stationarity) greatly facilitate the mathematical treatment of times series, in applied research many data sets do not display such regularity. This is especially true in economics where the inherent evolution of the data-generating process happens at a slow rate (e.g. incremental technical/technological change) but also abruptly (e.g. the so-called technical/ technological/ social revolutions). Furthermore, besides the evolution of the data generating process *itself*, most economic data, when they are considered for a sufficiently long enough period of time, display change in their

second-order structure, with periods of high volatility succeeding to periods of low volatility (and vice-versa). These changes may be inherent to the data-generating process that more or less regularly undergoes periods of financial or economic crisis, wars, financial bubbles, *etc.* Consequently, it is of great practical interest to go beyond the assumption of stationarity in the statistical analysis of economic data.

There is no unique way to relax the assumption of stationarity. One important contribution is made by Dahlhaus (1997) who introduces a class of *local stationary* processes written

$$X_n(t) = \int_{-\pi}^{\pi} A(\frac{t}{n}, \omega) \exp(i\omega t) d\zeta(\omega) \tag{4.1.1}$$

for $t = 0, ..., n-1$, $n \in \mathbb{N}$ and where the transfer function, $A(z, \omega)$, is defined on $(0, 1) \times [-\pi, \pi)$. In this setting letting $n \to \infty$ does not have the usual meaning of obtaining observations further in the future, but means we observe a given segment of time at higher and higher frequency, or in other words, that we gain more information about the local structure (i.e. for a segment of time) of $A(z, \omega)$.

Later, the notion of locally stationary process has been fruitfully employed by Nason, Von Sachs and Kroisandt (2000) to introduce a class of locally stationary *wavelet* processes, which replaced the harmonic system $\{\exp(i\omega t)\}$ in (4.1.1) by a non-decimated discrete wavelet system $\{\psi_{j,k}(t)\}$ (discussed in section 2 of this chapter) and that writes

$$X_n(t) = \sum_{j=-\infty}^{-1} \sum_{k=0}^{n-1} W_j(z) \varepsilon_{j,k} \psi_{j,k}(t) \tag{4.1.2}$$

for $t = 0, ..., n-1$ observations and where $z = k/n$ is the rescaled time, and $\{\varepsilon_{j,k}\}$ are independent standard normal increments. Because wavelet systems are well localized in frequency and *in time*, they appear more natural for modeling time-varying spectra of nonstationary processes. However, to use non-decimated discrete wavelet systems to model time series one needs the pay the price of abandoning the identification of the parameter $W_j(z)$. This is due to the fact that non-decimated wavelet systems are overcomplete and hence do not permit to represent a signal uniquely. Nevertheless, Nason et al. (2000) proves that even if the time series representation in the wavelet basis is not unique, the representation of the autocovariance function of the time series is asymptotically identified. Hence when the second-order structure of the time series is the object of interest, using non-decimated discrete wavelet systems is more than appropriate. This conclusion is pushed further in Van Bellegem and von Sachs (2008) that enlarges the class of locally stationary wavelet process for it to contain processes whose spectral density may change very suddenly in time. Finally, we mention that Fryzlewicz, Van Bel-

legem and Von Sachs (2003) considers the problem of forecasting univariate locally stationary wavelet processes.

The present and the following chapters resume the work initiated by Koch (2015) to extend the model of Nason et al. (2000) in two directions. First, allowing the increments to be correlated *across scales*, that is allowing the data generating process to be such that $\mathbb{C}\text{ov}(\varepsilon_{j,k}, \varepsilon_{i+j,k}) = \Gamma_{i,j}(k/n) \neq 0$. Second, extending the class of locally stationary processes to encompass multivariate processes while allowing the increment involved in the representation of each constituent of the multivariate time series to be correlated.

In Park, Eckley and Ombao (2014) a class of multivariate locally stationary processes is introduced with the uncorrelated random increments across scales and across constituents of the multivariate time series. In Cho and Fryzlewicz (2015) a more general multivariate locally stationary process is introduced but the increment remains uncorrelated across scales. Not allowing the increment to be correlated across scales is restrictive. Heuristically it means that the short-term random shocks that affect a time series are *not* correlated with the long terms random shocks. Formally, it means that some serious constraints are imposed on the second order-structure of the time series, for instance, it imposes that

$$\mathbb{C}\text{ov}\big(X_1(t), X_2(t-\tau)\big) = \mathbb{C}\text{ov}\big(X_1(t), X_2(t+\tau)\big)$$

where $X_1(t), X_2(t)$ are two times series and $t, \tau \in \mathbb{Z}$.

As demonstrated below, it is possible to allow the increments involved in the representation of each constituent of the multivariate time series to be correlated across constituents of the time series *and* across scales by introducing the concept of Cross-Correlation of Wavelet Functions (CCWF). Because non-decimated discrete wavelet systems are overcomplete, relaxing the assumption of independence of the increment across scales makes it necessary to be able to measure the level of redundancy among sets of non-decimated discrete wavelet functions. This is what is achieved by the CCWF that are introduced in Section 3 of the chapter.

Section 4 of this chapter provides the definition of the *Multivariate Locally Stationary Wavelet process* (MLSW). There, it is also demonstrated that the spectrum of the process has a unique asymptotic representation in the CCWF domain.

4.2 Wavelet Analysis

In this section, we review some results on wavelet systems. We begin by recalling the concept of multiresolution analysis and show it leads to the dilation of the scaling function. Then we introduce the notions of wavelet functions, non-decimated wavelet functions, and discrete wavelet functions. A complete introduction to the subject may be found in Daubechies et al. (1992).

4.2.1 Multiresolution analysis of $L^2(\mathbb{R})$

The multiresolution analysis (MRA) of $L^2(\mathbb{R})$ constitutes the starting point for the construction of an orthonormal wavelet basis. An MRA of $L^2(\mathbb{R})$ is a sequence of closed vector subspace V_j of $L^2(\mathbb{R})$ such that

1. $V_j \subset V_{j+1}$ (nestedness)
2. $\cup_{j\in\mathbb{Z}} V_j = L^2(\mathbb{R})$ (density)
3. $\cap_{j\in\mathbb{Z}} V_j = \{0\}$ (separation)
4. for all $f \in L^2(\mathbb{R})$ and all $j \in \mathbb{Z}$ we have $f(x) \in V_j \Leftrightarrow f(2x) \in V_{j+1}$ (scaling)
5. for all $f \in L^2(\mathbb{R})$ and all $k \in \mathbb{Z}$ we have $f(x) \in V_0 \Leftrightarrow f(x-k) \in V_0$ (translation invariance)
6. there exists a function $\phi \in V_0$, with $\int_{\mathbb{R}} \phi(x)dx \neq 0$, called the scaling function, such that the system $\{\phi(x-k)\}_{k\in\mathbb{Z}}$ of shifted functions is an orthogonal basis of V_0.

We assume here and thereafter that all scaling functions $\phi(x)$ are real-valued functions with $\int_{\mathbb{R}} \phi(x)dx = 1$. Moreover, referring to the scaling property and the translation invariance property of the MRA, we define the *dilated-translated functions* of the scaling function $\phi_{j,k}(x) \in V_j$ as

$$\phi_{j,k}(x) = 2^{j/2}\phi(2^j x - k).$$

where j and k are respectively the indices of dilatation and translation.

4.2.2 Construction of wavelets from a multiresolution analysis

This section explains how to derive wavelet systems from a MRA.

Since $\phi \in V_0$, which is included in V_1, we can decompose ϕ in the basis $\{\phi_{1,k}\}$ of V_1:

$$\phi = \sum_{k\in\mathbb{Z}} h_k \phi_{1,k}$$

with $h_k = \langle \phi, \phi_{1,k} \rangle_2$. By expanding this decomposition we obtain

$$\phi(x) = \sqrt{2} \sum_{k\in\mathbb{Z}} h_k \phi(2x - k).$$

with $x \in \mathbb{R}$. The coefficients $..., h_{-1}, h_0, h_1, ...$ are the *scaling filter* of the MRA. In general, one can write for every scale j and every shift k

$$\phi_{j,k}(x) = \sum_{l \in \mathbb{Z}} a_{k,l} \phi_{j+1,l}(x) \tag{4.2.1}$$

with $x \in \mathbb{R}$ and $a_{k,l} = \langle \phi_{j,k}, \phi_{j+1,l} \rangle_2$. Interestingly we have that

$$\begin{aligned} a_{k,l} &= \langle \phi_{j,k}, \phi_{j+1,l} \rangle_2 \\ &= 2^{j+1/2} \int_{\mathbb{R}} \phi(2^j x - k) \phi(2^{j+1} x - l) dx \\ &= 2^{1/2} \int_{\mathbb{R}} \phi(u) \phi(2u - (l - 2k)) du \\ &= \int_{\mathbb{R}} \phi(u) \phi_{1,l-2k}(u) du \\ &= h_{l-2k} \end{aligned}$$

where the third equality follows by a change of variable and the fourth line follows from the definition of the dilated-translated function. Hence the coefficients $a_{k,l}$ can be expressed in terms of elements of the scaling filter, and Equation (4.2.1) can be rewritten as

$$\phi_{j,k}(x) = \sum_{l \in \mathbb{Z}} h_{l-2k} \phi_{j+1,l}(x). \tag{4.2.2}$$

Hence the basis of V_j can be expressed as a linear combination of the basis in V_{j+1}. Consequently, we can recursively apply the same reasoning to write the basis of any element of the sequence of space $\{V_j\}_j \in \mathbb{Z}$ in terms of the basis of the remaining elements. This reasoning lead to the *dilatation equation* of the function ϕ that is Equation (4.2.2) rewritten as

$$\phi(2^j - k) = 2^{1/2} \sum_{l \in \mathbb{Z}} h_{l-2k} \phi(2^{j+1} x - l) \tag{4.2.3}$$

for every $j, k \in \mathbb{Z}$.

The dilatation equation is directly linked to the construction of wavelet functions. Consider any $f(x) \in V_j$ for $j \in \mathbb{Z}$, we then have $f(x) \in V_{j+i}$ for any $i \in \mathbb{N}$ and we can express $f(x)$ as a linear combination of the basis function of V_{j+i} that is $\{\phi_{j+i,k}\}_{k \in \mathbb{Z}}$. However, to express $f(x)$ in terms of the basis function of V_{j+i} when i is a negative integer, we need an additional type of function: the wavelet functions, which we now define.

Definition 4 *(Wavelet functions) Suppose that $\{V_j\}_j \in \mathbb{Z}$ forms a multiresolution analysis for $L^2(\mathbb{R})$ with scaling function $\phi(x)$ that satisfies the dilation*

equation given in (4.2.3). *The function mother wavelet function* $\psi(x) \in V_1$ *is defined as*

$$\psi(x) := 2^{1/2} \sum_{l \in \mathbb{Z}} g_l \phi(2x - l)$$

and the functions wavelet function $\psi_{j,k}(x) \in V_{j+1}$ *are given by*

$$\psi_{j,k}(x) := 2^{j/2} \psi(x - 2^{-j}k), \tag{4.2.4}$$

that is

$$\psi_{j,k}(x) = 2^{1/2} \sum_{l \in \mathbb{Z}} g_{l-2k} \phi_{j+1,l}(x) \tag{4.2.5}$$

for all $j, k \in \mathbb{Z}$, *where the* wavelet filter, $\{g_l\}_{l \in \mathbb{Z}}$ *are given by*

$$g_l = (-1)^l h_{1-l}.$$

Denote by W_j the orthogonal complement of the subspace V_j in V_{j+1}, the Meyer-Mallat theorem states that the system $\{\psi_{j,k}\}_{j,k \in \mathbb{Z}}$ is a basis for W_j (e.g. Theorem 7.2 at page 271 in Mallat (1999)). Hence, for any $j \in \mathbb{Z}$ we can write any functions $f(x) \in V_{j+1}$ as

$$f(x) = \sum_{k \in \mathbb{Z}} a_k \phi_{j,k}(x) + \sum_{k \in \mathbb{Z}} b_k \psi_{j,k}(x) \tag{4.2.6}$$

where $\sum_{k \in \mathbb{Z}} a_k \psi_{j,k}(x)$ is the approximation of $f(x)$ in V_j and $\sum_{k \in \mathbb{Z}} b_k \psi_{j,k}(x)$ is in W_j. Finally, given that $\psi_{j,k} \in V_j$, we can apply the same reasoning recursively to write $f(x)$ as an infinite linear combination of wavelet functions that lie in $W_j, W_{j-1}, W_{j-2}, ...$ respectively.

Example 1 *(Haar wavelet system.) The simplest example of a wavelet system starts with the* Haar scaling function $\phi(x) = \mathbf{1}_{[0,1[}(x)$. *This scaling function leads to the scaling filters*

$$h_k = \begin{cases} 2^{-1/2} & \text{if } k = 1, 2 \\ 0 & \text{otherwise.} \end{cases} \tag{4.2.7}$$

The closed-form expression of the resulting mother wavelet function is given by

$$\psi(x) = -\mathbf{1}_{[0,1/2[}(x) + \mathbf{1}_{[1/2,1[}(x)$$

and the Haar wavelet system is generated according to

$$\psi_{j,k}(x) = 2^{j/2} \psi(x - 2^j k)$$

where $j = -1, -2, ...$ *and* $k \in \mathbb{Z}$.

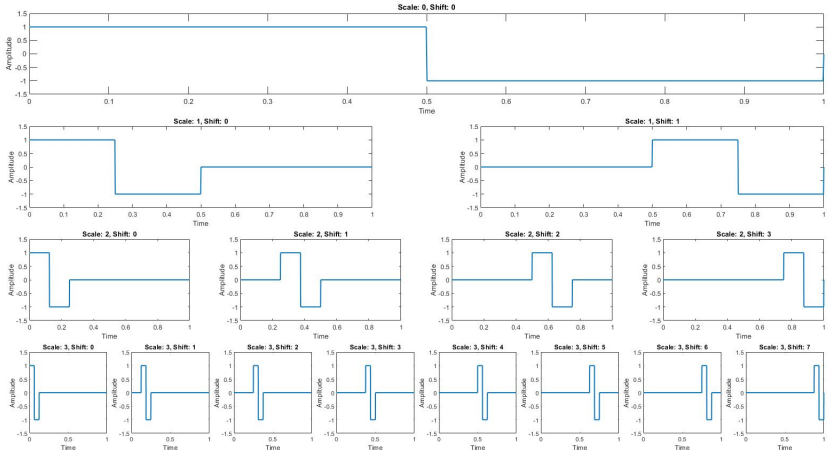

Figure 4.1: Illustration of the Haar system

4.2.3 The discrete wavelet transform

In the previous section, we have studied how wavelet systems can be used to represent any continuous function in $L^2(\mathbb{R})$. Now we study how wavelet may be employed to represent *discrete* signals.

Consider any deterministic sequence $s_0, ..., s_{n-1}$ in l_2 where $n = 2^J$ for some $J \in \mathbb{N}$, and define the continuous function

$$f(x) = \sum_{k \in \mathbb{Z}} c_{0,k} \phi(x - k) \tag{4.2.8}$$

with $x \in \mathbb{R}$, $c_{0,k} = s_k$ for $k = 0, ..., n-1$. The function $f(x)$ is in the space V_0 of the MRA of $L^2(\mathbb{R})$. Since W_{-1} is the orthogonal complement of V_{-1} in V_0, we can use Equation (4.2.6) to decompose $f(x)$ as

$$f(x) = \sum_{k \in \mathbb{Z}} c_{-1,k} \phi_{-1,k}(x) + \sum_{k \in \mathbb{Z}} d_{-1,k} \psi_{-1,k}(x) \tag{4.2.9}$$

where $c_{-1,k} = \langle f, \phi_{-1,k} \rangle_2$ is the *scaling coefficients* at scale j and shift k and $d_{-1,k} = \langle f, \psi_{-1,k} \rangle_2$ is the *wavelet coefficient* scale j and shift k. Hence, by Equation (4.2.2), the orthonormality of the scaling function implies that

$$c_{-1,k} = \sum_{l \in \mathbb{Z}} h_{l-2k} c_{0,l}$$

and by Equation (4.2.5), following the same reasoning, we obtain

$$d_{-1,k} = \sum_{l \in \mathbb{Z}} g_{l-2k} c_{0,l}.$$

for $k = 0, ..., n/2-1$. The two previous equations show how the coefficients at scale $j = -1$ may be computed from the coefficients at scale $j = 0$. Observe that there are half as many coefficients at scale -1 than there are at scale 0. This important feature is known as the *decimation*.

The *discrete wavelet function* at scale -1 and shift 0 is given by the wavelet filter $\{g_t\}_{t\in(0,...,L[h])}$, that is given by

$$\psi_{-1,0}(t) = g_t \tag{4.2.10}$$

for $t \in \mathbb{Z}$ where $L[h]$ denotes the length of the support of the scaling filter (from Example 1, we see that $L[h] = 2$). Consequently the support of the wavelet function at scale -1 is given by $(0, ..., L[h] - 1)$.

Next, we follow an iterative reasoning and consider the sequence of coefficient $c_{-1,0}, ..., c_{-1,n/2-1}$. Define the continuous function

$$f_{-1}(x) = \sum_{k\in(0,...,n/2-1)} c_{-1,k}\psi_{1,k}(x)$$

for $x \in \mathbb{R}$. And, analogously to above, since W_{-2} is the orthogonal complement of V_{-2} in V_{-1}we may write

$$f_1(x) = \sum_{k\in\mathbb{Z}} c_{-2,k}\phi_{-1,k}(x) + \sum_{k\in\mathbb{Z}} d_{-2,k}\psi_{-1,k}(x) \tag{4.2.11}$$

And by the same reasoning as above we may express the coefficient at scale $j = -2$ in terms of the coefficients at scale $j = -1$, that is

$$c_{-2,k} = \sum_l h_{l-2k}c_{-1,l}$$

and

$$d_{-2,k} = \sum_l g_{l-2k}c_{-1,l}.$$

for $k = 0, ..., n/4 - 1$. But given the above relation between the coefficient at scale -1 and the ones at scale 0, we may as well represent the coefficient at scale -2 in terms of the ones at scale 0. By doing so we obtain

$$c_{-2,k} = \sum_m c_{0,m} \sum_l h_{m-2l}h_{l-2k}c_{-1,l}$$

and

$$d_{-2,k} = \sum_m c_{0,m} \sum_l h_{m-2l}g_{l-2k}.$$

for $k = 0, ..., n/4 - 1$.

The *discrete wavelet function* at level -2 and shift 0 is given by

$$\begin{aligned}\psi_{-2,0}(t) &= \sum_{l\in\mathbb{Z}} h_{t-2l}\psi_{-1,0}(l) \\ &= \sum_{l\in\mathbb{Z}} h_{t-2l}g_l\mathbf{1}_{0,...,L[h]-1}(l) \\ &= \sum_{l\in\mathbb{Z}} h_{t-2l}g_l\mathbf{1}_{0,...,3(L[h]-1)}(l)\end{aligned}$$

where the third line follows by noting that $h_{t-2l} = 0$ if $t < 2l$ or $t > L[h]-1+2l$, and that $g_l = 0$ if $l < 0$ or $l > L[h] - 1$.

By iterating in this procedure we express the complete deterministic sequences $s_0, ...s_{n-1}$ in terms of wavelet coefficient and discrete wavelet functions. Hence we obtain the following recursive definition of the *discrete wavelet function*

$$\begin{aligned}\psi_{-1,0}(t) &= g_t && \text{for } t \in \mathbb{Z} \\ \psi_{j-1,0}(t) &= \sum_{l\in\mathbb{Z}} h_{t-2l}\psi_{j,0}(l) && \text{for } t \in \mathbb{Z} \text{ and } j = -1, -2, ...\end{aligned}$$

The discrete wavelets are defined as convolutions of the scaling and wavelet filters. The length of the support of the wavelet at scale j, denoted L_j is given by a multiple of the length of the support of the scaling filter, that is

$$L_j = (2^{-j} - 1)(L[h] - 1) + 1. \tag{4.2.12}$$

Example 2 *(Discrete Haar Wavelet). Geometrically, the non-shifted discrete Haar Wavelets may be represented by the vectors*

$$\begin{aligned}\psi_{-1,0} &= 2^{-1/2}(-1, 1) \\ \psi_{-2,0} &= 2^{-1}(-1, -1, 1, 1) \\ \psi_{-3,0} &= 2^{-3/2}(-1, -1, -1, -1, 1, 1, 1, 1) \\ &... \end{aligned}$$

It is important to note that, except for Haar Wavelets, the discrete wavelets are not just the sampled version of the associated continuous wavelet.

4.2.4 Nondecimated discrete wavelets

Observe that from Equation (4.2.4) in Definition 4, the wavelet functions are given by $\psi_{j,k}(x) := 2^{j/2}\psi(x - 2^{-j}k)$. This definition implies that wavelet can only be shifted by 2^{-j}. The key point of *non-decimated* wavelet function is to

construct wavelets $\psi_{j,k}$ from the motherwavelet ψ that can be shifted to *any location* (e.g. Berkner and Wells Jr (2002)).

To do so, let $\{\psi_{j,0}\}_{j=-1,-2,...}$ be a system of discrete wavelet functions constructed from the mother wavelet ψ by the procedure described in the previous section. Then the *non-decimated* discrete wavelet functions are given by

$$\psi_{j,k}(t) = \psi_{j,0}(t-k)$$

for $j = -1, -2, ...$ and $k \in \mathbb{Z}$.

Example 3 *(Non-decimated discrete Haar System) The non-decimated, discrete Haar System is given by*

$$\psi_{j,0}(t) := 2^{-j/2}\mathbf{1}_{\{0,1,...,2^{-j-1}-1\}}(t) - 2^{j/2}\mathbf{1}_{\{2^{-j-1},...,2^{-j}-1\}}(t)$$

for $j = -1, -2, ...$ and $t \in \mathbb{Z}$. The shifted non-decimated discrete wavelets are given by

$$\psi_{j,k}(t) := \psi_{j,0}(t-k). \tag{4.2.13}$$

One important feature of the *non-decimated* wavelet systems, is the absence of *decimation*, that is, the number of coefficients at each scale, does not depend on the wavelet scale. This implies that the number of non-decimated discrete wavelet functions at each scale is always larger than in the decimated case. Because the decimated system is linearly independent, this cannot be the case for non-decimated systems, hence shifting wavelet to any location (and not just by shifts of 2^j) leads to redundancy in the system. Consequently, the representation of a signal in a non-decimated wavelet domain is not unique. The next section introduces the notion of *Cross-correlation of wavelet functions* that allows to measure the degree of redundancy in a non-decimated discrete wavelet system.

4.3 The Cross-Correlation of Wavelet Functions

Consider two discrete, normalized, signals in times, say, $a(t)$ and $b(t)$ with $t \in \mathbb{Z}$. The cross-correlation between these two signals, which we denote by $C_{a,b}$ is defined as

$$C_{a,b}(\tau) = \sum_{t\in\mathbb{Z}} a(t)b(t+\tau) \text{ with } \tau \in \mathbb{Z}.$$

This object is a sliding inner product between the two signals and gives a measure of the similarity between the signals as a function of the lag τ. As

such, the well-known notion of *autocorelation* may be defined as the cross-correlation of a signal with itself.

The idea behind the notion of Cross-correlation of wavelet functions (CCWF), is to obtain a measure of the similarity between two wavelet functions as a function of a lag *and* as a function of the difference of scale between the two wavelets.

The following definition introduces the notion of Cross-correlation of wavelet functions.

Definition 5 (Cross-correlation of wavelet functions - CCWF) *Let the family* $\{\psi_{j,\tau}(t)\}_{\tau\in\mathbb{Z}}$ *with* $j=-1,...,-J$ *and* $J\in\mathbb{N}_+$ *be a collection of non-decimated discrete wavelet functions. Denote as*

$$\Psi_J=\{\tau\to\Psi_{j,l}(\tau),\tau\in\mathbb{Z}\}_{j,l=-J}^{-1} \tag{4.3.1}$$

the collection up to scale J*, of the Cross-correlation of wavelet functions that are defined by*

$$\Psi_{j,l}(\tau)=\sum_{t\in\mathbb{Z}}\psi_{j,0}(t)\psi_{l,\tau}(t) \quad \textit{with } \tau\in\mathbb{Z} \tag{4.3.2}$$

for $l,j=-1,...,-J$*.*

At each given scale, there are J cross-correlations of wavelet functions, hence Ψ_J contains J^2 elements.

Furthermore, we note that for all $j=-1,-2,...$ the functions $\Psi_{j,j}$ are identical to the autocovariance wavelet function (ACWF) studied in Saito and Beylkin (1993).

4.3.1 Linear independence of the cross-correlation of wavelet functions

In this section, we provide conditions that ensure the linear independence of the CCWF. In other words we establish that (under those conditions) for any $J\in\mathbb{N}_+$, a collection of CCWF Ψ_J forms a basis for a subset of $\{\tau\to\mathbb{R}|\tau\in\mathbb{Z}\}$. This means that there are no elements in Ψ_J that may be represented as a linear combination of the others elements in Ψ_J, or formally that for any $j,l=-1,...,-J$

$$\Psi_{j,l}(\tau)=\alpha_{j,l}\Psi_{j,l}(\tau)+\sum_{(r,s)\neq(j,l)}\alpha_{r,s}\Psi_{r,s}(\tau) \quad \text{for every } \tau\in\mathbb{Z}$$

only if every $\alpha_{r,s}=0$ and $\alpha_{j,l}=1$.

This result is important because in the next section, we will represent functions as linear combinations of CCWF. The linear independence of the CCWF ensures that this representation is unique.

The result is given in the following theorem.

Theorem 6 *For $J = 1, 2, ...,$ let Ψ_J be a collection of cross-correlation of wavelet functions as in Definition 5 and suppose the scaling filters of the underlying wavelet functions satisfies*

$$\sum_{k\in\mathbb{Z}} h_k h_{k+2\delta} = \begin{cases} 1 & \text{if } \delta = 0 \\ 0 & \text{if } \delta \in \mathbb{Z}\setminus\{0\}, \end{cases} \tag{4.3.3}$$

and

$$\sum_{k\in\mathbb{Z}} h_{k+2\delta} g_k = 0 \tag{4.3.4}$$

for any $\delta \in \mathbb{Z}$.

Then the elements in Ψ_J are linearly independent.

The proof of this result requires some intermediate lemmas. We state and comment on those lemmas and then prove the above result.

We start with the following lemma that states that, except for $j = l$, the CCWF are not symmetric.

Lemma 13 *Let $J \in \mathbb{N}$ and let Ψ_J be a collection of Cross-correlation of wavelet functions as in Definition 5. Then*

$$\Psi_{j,l}(\tau) = \Psi_{l,j}(-\tau) \quad \text{for } j, l = -1, ..., -J$$

or, equivalently, by defining the matrices $\tilde{\Psi}_J(\tau) = (\Psi_{-j,-l}(\tau)_{j,l=-J,...,-1} \in \mathbb{R}^{J\times J}$ for every $\tau \in \mathbb{Z}$, we have

$$\tilde{\Psi}_J(\tau) = \tilde{\Psi}_J(-\tau)^\top.$$

Moreover, we have

$$\Psi_{j,l}(\tau) = 0 \text{ if } \begin{cases} \tau \le -L_j - 1 \\ \tau > L_l \end{cases}$$

for any $j = -1, -2, ...,$ where L_j is the length of the support of the underlying non-decimated discrete wavelet functions at scale j (as stated in Equation (4.2.12)).

The proof is given in Appendix.

The CCWF does not have, in general, a closed-form expression. Nevertheless, in some particular cases, there exist sets of points for which the CCWF admits a closed-form expression. The following lemma deals with this problem for the case when $j = l$, that is when the CCWF reduced to the autocorrelation wavelet function. However, even in this case, the CCWF admits a close form expression only on a subset of its support.

Lemma 14 *Let $J \in \mathbb{N}_+$ and let $\Psi_J = \{\Psi_{j,l}\}_{j,l=-J}^{-1}$ be a collection of Cross-correlation of wavelet functions as defined in Definition 5 and suppose the scaling filters of the underlying wavelet functions satisfy assumption (4.3.3). Then we have that, for $j = -1, -..., -J$ and $m \in \mathbb{Z}$,*

$$\Psi_{j,j}(2^{-j-1}m) = \sum_{l\in\mathbb{Z}} g_l g_{l-m}$$

The proof is given in Appendix. The above result is actually a particular case of Lemma 16 (below) that finds closed-form expressions of the CCWF even for $j \neq l$. However, the proof of both Lemmas 14 and 16 are technical, hence we treat the simpler case before the general case. To understand the proof of Lemma 16 one is encouraged to start by following the proof of Lemma 14.

The next lemma states that for any $j, l = -1, -2, ...,$ there exists a non-empty set of values of $\tau \in \mathbb{Z}$ such that $-L_j - 1 < \tau \leq L_{j+i}$ for which $\Psi_{i,j}(\tau) = 0$.

Lemma 15 *Let $J \in \mathbb{N}_+$ and let $\Psi_J = \{\Psi_{j,l}\}_{j,l=-J}^{-1}$ be a collection of Cross-correlation of wavelet functions as defined in Definition 5 and suppose the scaling filters of the underlying wavelet functions satisfy assumptions (4.3.3) and (4.3.4).*

Then for any $j, l = -1, ..., -J$, every $m \in \mathbb{Z}$, and any $r, s \in \mathbb{N}$

$$\Psi_{j,l}(2^r(2^l - 1)m) = 0 \text{ if } j \geq -r. \tag{4.3.5}$$

The proof is given in Appendix.

Lemma 14 provides a closed-form expression for values in the support of $\Psi_{j,j}$. The following lemma generalizes the result to the general CCWF $\Psi_{j,l}$.

Lemma 16 *Let $J \in \mathbb{N}$ and let $\Psi_J = \{\Psi_{j,l}\}_{j,l=-J}^{-1}$ be a collection of Cross-correlation of wavelet functions as defined in Definition 5 and suppose the scaling filters of the underlying wavelet functions satisfy assumptions (4.3.3) and (4.3.4). Then for $-J \leq j \leq l \leq -1$ we have*

$$\Psi_{j,l}(-2^{-j+1}(2^{-l+j+1} - 1)(L[h] - 1)) = g_{L[h]-1} g_0 h_{L[h]-1}^{-l+j}$$

where $L[h]$ is the length of the support of the scaling filters of the underlying wavelet functions.

The proof is given in Appendix.

We are now ready to prove Theorem 6.

Proof of Theorem 6 To show the linear independence of $\Psi_{r,s}$ with $\{\Psi_{j,l}\}_{j,l=-J}^{-1}$ it is enough to show that, for some $\tau^\star \in \mathbb{Z}$ $\Psi_{r,s}(\tau^\star) \neq 0$ while $\Psi_{j,l}(\tau^\star) = 0$ for all $j, l = -1, ..., J$, because in this case there exists no linear combination of $\{\Psi_{j,l}\}_{j,l=-J}^{-1}$ that is equal to $\Psi_{r,s}$ for every $\tau \in \mathbb{Z}$.

First we show that $\Psi_{-J,-J}$ is linearly independent from $\Psi_J \setminus \{\Psi_{-J,-J}\}$. From Lemma 16, we have

$$\Psi_{-J,-J}(-2^{J+1}(L[h]-1)) = g_{L[h]-1}g_0 \neq 0.$$

However by Lemma 15, we have for all $j = -1, ..., -J+1$ and $l = -1, ..., -J$ that

$$\Psi_{j,l}(-2^{J+1}(L[h]-1)) = 0,$$

moreover Lemma 14 ensures that for $l = -1, ... - J$

$$\Psi_{-J,l}(-2^{J+1}(L[h]-1)) = \Psi_{l,-J}(2^{J+1}(L[h]-1)),$$

and Lemma 15 provides

$$\Psi_{l,-J}(2^{J+1}(L[h]-1)) = 0$$

for $l = -1, ...-J+1$. Hence we have shown that $\Psi_{-J,-J}$ is linearly independent from $\Psi_J \setminus \{\Psi_{-J,-J}\}$.

Second, we show that $\Psi_{-J,-J+1}$ is linearly independent from $\Psi_J \setminus \{\Psi_{-J,-J+1}\}$. we already showed that $\Psi_{-J,-J+1}$ is linearly independent from $\{\Psi_{-J,-J}\}$. Hence $\{\Psi_{-J,-J+1}, \Psi_{-J,-J}\}$ is a basis. Moreover, by Lemma 13 and Lemma 16

$$\begin{aligned}\Psi_{-J,-J+1} = (2^{-J}(2^2-1)(L[h]-1)) &= \Psi_{-J+1,-J}(-2^{-J}(2^2-1)(L[h]-1)) \\ &= g_0 g_{L[h]-1} h_{L[h]-1} \neq 0.\end{aligned}$$

However, by Lemma 15 for $j = 1, ..., -J+2$ and $l = -1, ..., -J$

$$\Psi_{j,l} = (2^{-J}(2^2-1)(L[h]-1)) = 0,$$

and for $j = -J+1, -J$ and $l = -1, ..., -J+2$ by Lemma 13 and Lemma 15

$$\begin{aligned}\Psi_{j,l} = (2^{-J}(2^2-1)(L[h]-1)) &= \Psi_{l,j}(-2^{-J}(2^2-1)(L[h]-1)) \\ &= 0.\end{aligned}$$

hence we have shown that $\Psi_{-J,-J+1}$ is linearly independent from $\Psi_J \setminus \{\Psi_{-J,-J+1}\}$

Third, suppose that for some $s = 1, ..., J-2$, $\Psi_{-J,-J+s+1}$ is linearly independent from $\{\Psi_{-J,-J+u}\}_{u=0}^{J-s}$. Hence $\{\Psi_{-J,-J+s+1}, \Psi_{-J,-J+u}\}_{u=0}^{J-s}$ forms a basis. We show that, in addition, $\Psi_{-J,-J+s+1}$ is linearly independent from $\Psi_J \setminus \{\Psi_{-J,-J+s+1}\}$. By Lemma 13 and Lemma 16

$$\begin{aligned}\Psi_{-J,-J+s+1} = (2^{J-s}(2^{s+2}-1)(L[h]-1)) &= \Psi_{-J+1,-J}(-2^{J-s}(2^{s+2}-1)(L[h]-1)) \\ &= g_{L[h]-1} g_0 h_{L[h]-1}^{s+1} \neq 0.\end{aligned}$$

However, by Lemma 15 for $j = 1, ..., -J+s+2$ and $l = -1, ..., -J$

$$\Psi_{j,l} = (2^{J-s}(2^{s+2}-1)(L[h]-1)) = 0,$$

and for $j = -J+s+1, -J$ and $l = -1, ..., -J+s+2$ by Lemma 13 and Lemma 15

$$\begin{aligned}\Psi_{j,l} = (2^{J-s}(2^{s+2}-1)(L[h]-1)) &= \Psi_{l,j}(-2^{J-s}(2^{s+2}-1)(L[h]-1)) \\ &= 0.\end{aligned}$$

hence we have shown that $\Psi_{-J,-J+s+1}$ is linearly independent from $\Psi_J \setminus \{\Psi_{-J,-J+s+1}\}$. Furthermore, since we have already proved that the above is true for $s = 1$, then by induction it is true for all $s = 1, ..., -J-2$.

By a symmetric argument we can show that $\Psi_{J+u,J}$ is linearly independent of $\Psi_J \setminus \{\Psi_{J+u,J}\}$ for every $u = 1, ..., -J-1$ while, in the same time showing that $\{\{\Psi_{-J+u,-J}\}_{u=0}^{J-1}, \{\Psi_{-J,-J+v}\}_{v=0}^{J-1}\}$ is a basis.

Consequently, if $\Psi_J \setminus \{\{\Psi_{-J+u,-J}\}_{u=0}^{J-1}, \{\Psi_{-J,-J+v}\}_{v=0}^{J-1}\}$ is a basis, then Ψ_J is a basis.

Observe that for $J = 2$, the set $\Psi_J \setminus \{\Psi_{-J,-J+1}, \Psi_{-J+1,-J}, \Psi_{-J,-J}\} = \{\Psi_{-1,-1}\}$ that is a basis.

Finally, since all of the above is true for any finite J, it is true for $J+1$, and by induction we obtain the result for $J = 1, 2, ...$ □

4.4 Multivariate Locally Stationary Wavelet Process

We are now ready to define the class of Multivariate Locally Stationary Processes.

Definition 6 *A sequence of doubly indexed stochastic processes in* $L^2(\Omega, \mathcal{A}, \mathbb{P})$

$$\left\{(X_n^{(1)}(t), ..., X_n^{(p)}(t))^\top\right\}_{t=0}^{n-1} \qquad n \in \mathbb{N}_+$$

is in the class of Multivariate Locally Stationary Processes of depth $J \in \mathbb{N}_+$ *and of* order $q \in \{0, 1, ..., J\}$ *if, for every* $u, v = 1, ..., p$ *and every* $t = 0, ..., n-1$*, there exists a representation*

$$X_n^{(u)}(t) = \sum_{j=-J}^{-1} \sum_{k=0}^{n-1} W_j^{(u)}(z) \varepsilon_{j,k}^{(u)} \psi_{j,k}(t)$$

where the equality holds up to every set of probability zero and where

1. $z = k/n$;
2. *the collection* $\left\{\{\psi_{j,k}\}_{j=-\infty}^{-1}\right\}_{k=0}^{n-1}$ *is a set of* discrete non-decimated wavelet functions;
3. $W_j^{(u)}(z) \in \mathbb{R}$ *is finite for every* $u = 1, ..., p$ *and every* $t = 0, ...n-1$.

4. *the collection* $\{\{\varepsilon_{j,k}^{(u)}\}_{j=-J}^{-1}\}_{k=0}^{n-1}$ *follows a multivariate Gaussian distribution and every one of its elements are standardized (zero means, unit variance) random variables;*

5. $\mathbb{C}ov(\varepsilon_{j,k}^{(u)}, \varepsilon_{l,r}^{(v)}) = \begin{cases} \Gamma_{j,l}^{(u,v)}(z) & \text{if } |j-l| \leq q \text{ and } k = r \\ 0 & \text{either.} \end{cases}$

From now on, all the equalities involving random variables are supposed to hold, up events of probability zero. Moreover, to simplify the derivations we introduce the following notations.

Notation 3 *For* $n, J \in \mathbb{N}_+$, $t, k = 0, ..., n-1$, $u, v = 1, ..., p$, *let*

1. $W_J^{(u)}(z) := \text{diag}\big(W_1^{(u)}(z)), ..., W_J^{(u)}(z)\big) \in \mathbb{R}^{J \times J}$;
2. $\psi_{k;J}(t) := (\psi_{-1,k}(t), ..., \psi_{-J,k}(t))^\top \in \mathbb{R}^J$;
3. $\varepsilon_{k;J}^{(u)} := (\varepsilon_{-1,k}^{(u)}, ..., \varepsilon_{-J,k}^{(u)})^\top \in \mathbb{R}^J$;
4. $\Gamma_J^{u,v} =: (\Gamma_{i,j}^{(u,v)})_{i,j=-1,...,-J} \in \mathbb{R}^{J \times J}$

Hence if $\big\{(X_n^{(1)}(t), ..., X_n^{(p)}(t))^\top\big\}_{t=0}^{n-1}$ for $n \in \mathbb{N}_+$ is in the class of (J,q)-MLSW processes (multivariate locally stationary wavelet process of depth J and order q), then for every $u = 1, ..., p$ we can write

$$X_n^{(u)}(t) = \sum_{k=0}^{n-1} \big(\varepsilon_{k;J}^{(u)}\big)^\top W_J^{(u)}(z)\psi_{k;J}(t).$$

The following assumption imposes a regularity condition on the covariance of a $(J,q) - MLSW$ process.

Assumption 1 *The sequence of doubly indexed stochastic processes that belongs to* $L^2(\Omega, \mathcal{A}, \mathbb{P})$

$$\big\{(X_n^{(1)}(t), ..., X_n^{(p)}(t))^\top\big\}_{t=0}^{n-1} \quad n \in \mathbb{N}_+$$

in the class of $(J,q)-MLSW$ *processes defined in Definition 6 is such that its* Evolutionary Wavelet Spectrum *(EWS), that is for every* $u, v = 1, ..., p$

$$S_J^{(u,v)}(z) := W_J^{(u)}(z)\Gamma_J^{(u,v)}W_J^{(v)}(z),$$

has, for all $j, l = -1, ..., -J$, *its elements, noted* $S_{-j,-l}^{(u,v)}(z)$, *that are Lipschitz continuous on* $(0,1)$ *with Lipschitz constants* $C_{j,l}^{(u,v)}$ *satisfying*

$$\sum_{j,l=-1,...,-J; |j-l| \leq q} C_{j,l}^{(u,v)} L_{\max\{j,l\}}^2 < \infty$$

where L_l *denotes the support of the wavelet functions at scale* l.

The class of Multivariate Locally Stationary Process in Definition 6, uses non-decimated discrete wavelets to decompose a vector of stochastic processes with respect to a set of increment processes in the time-scale plane. It is known that the system of non-decimated (discrete) wavelet is not linearly independent (e.g. Nason et al. (2000)). Hence the coefficients $W^{(u)}$ are not identified. Nevertheless, in the following, we present a theory where, even though the multiscale representation of the MLSW is not unique, the multiscale representation of the autocovariance function is asymptotically identified. This theory is an extension of the one Nason et al. (2000) and was first studied in Koch (2015).

Before diving into the asymptotic identification of the autocovariance representation, we provide some comments concerning the class of processes defined in Definition 6 that satisfy Assumption 1.

In general, wavelets at scale j have a support of length $L_j = (2^{-j}-1)(L[h]-1)+1$ where $L[h]$ is the support of the *wavelet filter* (for the Haar wavelet this quantity is equal to 2) (e.g. (Daubechies et al., 1992)). Consequently we have that $\mathbb{C}\text{ov}(X_n^{(u)}(t), \varepsilon_{j,k}^{(u)}) = W_j^{(u)}\psi_{j,k}(t)$ and is equal to zero if $t \geq k + L_j$. Moreover, we observe that the support of the wavelet increases with $-j$ meaning that the lower the scale of the wavelet is, the larger the support of the wavelet is, and the larger is the portion of $\{X_n^{(u)}(t)\}$ that is affected by $\varepsilon_{j,k}$. In other words, the MLSW provides a representation of each element of the vector of times series in terms of short-term and long-term random shock, with the shortest term shocks being $\varepsilon_{-1,k}^{(u)}$.

Each element of the Evolutionary Wavelet Spectrum (EWS) measures the local power (contribution to the variance) in a MLSW process at a particular (rescaled) time z and scales j, l. We observe that the EWS denoted $S_J^{(u,v)}(z)$ depends on k since $z = k/n$. This model the time-varying second-order structure of the vector of time series. In Assumption 1, we require the EWS function to be Lipschitz continuous with the Lipschitz constant satisfying $\sum_{j,l=-1,\ldots,-J;|j-l|\leq q} C_{j,l}^{(u,v)} L_l^2 < \infty$. Hence Definition 6 allows the rate of change of the EWS function at higher scales to be larger than the one at lower scales.

4.4.1 Covariance representation

As mentioned above, the wavelet parameters $W^{(u)}$ are not identified due to the overcompleteness of the system of *nondecimated* wavelet functions. This section shows that the time-dependent covariance between $X_n^{(u)}(t)$ and $X_n^{(v)}(t-\tau)$ converges asymptotically to a linear combination of the CCWF, with time-varying coefficients being the EWS functions. This result is used to asymptotically identify the Evolutionary Wavelet Spectrum of the time series and hence to asymptotically identify the representation of the second-order structure.

First, for all $n \in \mathbb{N}$ and all $z \in (0,1)$, let $t = [zn]$, be integer part of zn.

Moreover, $w = k/n$ and, from Definition 6 and under Assumption 1,

$$\begin{aligned}
&C_n^{(u,v)}(-\tau, z) \\
&:= \mathbb{C}\mathrm{ov}\big(X_n^{(u)}([nz]), X_n^{(v)}([nz]-\tau)\big) \\
&= \sum_{k=0}^{n-1} \mathbb{E}\big(\psi_{k;J}(t)^\top W_J^{(u)} \varepsilon_{k;J}^{(u)} (\varepsilon_{k;J}^{(v)})^\top (w) W_J^{(v)}(w) \psi_{k;J}(t-\tau)\big) \\
&= \sum_{k=0}^{n-1} \psi_{k;J}(t)^\top S_J^{(u,v)}(w) \psi_{k;J}(t-\tau)
\end{aligned}$$

Hence we have

$$\begin{aligned}
&C_n^{(u,v)}(-\tau, z) \\
&= \sum_{j,l=-1,\ldots,-J;|j-l|\le q} \sum_{k=0}^{n-1} \psi_{j,k}(t)\psi_{l,k}(t-\tau) S_{-j,-l}^{(u,v)}(w) \\
&= \sum_{j,l=-1,\ldots,-J;|j-l|\le q} \sum_{k=0}^{n-1} \psi_{j,0}(t-k)\psi_{l,0}(t-k-\tau) S_{-j,-l}^{(u,v)}(w) \\
&= \sum_{j,l=-1,\ldots,-J;|j-l|\le q} \sum_{s=-t}^{(n-1)-t} \psi_{j,0}(-s)\psi_{l,\tau}(-s) S_{-j,-l}^{(u,v)}\big(\frac{s+[zn]}{n}\big)
\end{aligned}$$

where $s := k - t$. Therefore we obtain

$$C_n^{(u,v)}(-\tau, z) = \sum_{j,l=-1,\ldots,-J;|j-l|\le q} \Psi_{j,l}^{(n)} S_{-j,-l}^{(u,v)}(z) + R^{(n,t,u,v)}(z,\tau) \tag{4.4.1}$$

with a remainder term

$$\begin{aligned}
&R^{(n,t,u,v)}(z,\tau) \qquad (4.4.2)\\
&:= \sum_{j,l=-1,\ldots,-J;|j-l|\le q} \sum_{s=-t}^{(n-1)-t} \psi_{j,0}(-s)\psi_{l,\tau}(-s)\left(S_{-j,-l}^{(u,v)}\big(\frac{s+[zn]}{n}\big) - S_{-j,-l}^{(u,v)}(z)\right)
\end{aligned}$$

and

$$\Psi_{j,l}^{(n)}(\tau) := \sum_{s=-t}^{(n-1)-t} \psi_{j,0}(-s)\psi_{l,\tau}(-s) \tag{4.4.3}$$

and the remainder term is bounded from next following result.

Proposition 9 *For any $u, v = 1, ..., p$ and any $t = 0, ..., n-1$, let $R^{(n,t,u,v)}(z, \tau)$ be as in Equation (4.4.2). Then given the assumptions in Definition 6 and Assumption 1, we have that*

$$\sum_{\tau \in \mathbb{Z}} \int_{[0,1]} |R^{(n,t,u,v)}(z, \tau) dz| = O(\frac{1}{n})$$

uniformly in u, v, t.

The proof is given in the Appendix.

We aim to obtain to represent Equation (4.4.1) more elegantly. To this aim we provide the following definition in which we arrange in vectors objects indexed by $j, l = -1, ..., J; |j - l| \leq q$.

Definition 7 *For the remaining of this thesis, we define the following three vector in $\mathbb{R}^{(2q+1)J}$ for every $u, v = 1, ..., p$, $n, J \in \mathbb{N}_+$ and every $\tau \in \mathbb{Z}$*

$$\Phi_{n,J}(\tau) := \left(\left(\left(\Psi^{(n)}_{j,l}(\tau) \right)_{l=-1,...,-J;|j-l|\leq q} \right) \right)_{j=-1}^{-J} \qquad \Psi^{(n)}_{j,l}(\tau) \text{ as in (4.4.3)}$$

$$\Phi_J(\tau) := \left(\left(\Psi_{j,l}(\tau) \right)_{l=-1,...,-J;|j-l|\leq q} \right)_{j=-1}^{-J} \qquad \Psi_{j,l}(\tau) \text{ in Def 5}$$

$$S^{(u,v)}_J(z) := \left(\left(S^{(u,v)}_{-j,-l}(z) \right)_{l=-1,...,-J;|j-l|\leq q} \right)_{j=-1}^{-J}.$$

Hence from Proposition 9 and Definition 7 we have for all $\tau \in \mathbb{Z}$

$$C^{(u,v)}_n(-\tau, z) = \Phi^{\top}_{n,J}(\tau) S^{(u,v)}_J(z) + O(\frac{1}{n}).$$

Observe that by Definition 5, for a given J there exists $N \in \mathbb{N}$ such that for all $n > N$ $\Phi_{n,J}(\tau) = \Phi_J(\tau)$ for all $\tau \in \mathbb{Z}$. This is due to the fact that the maximum support of the CCWF contained in $\Phi_{\tau,J}$ does not lie outside $(0, ..., n-1)$. Hence we define the limit (in a sense made precise below) of the covariance between $X^{(u)}_n(t)$ and $X^{(v)}_n(t - \tau)$, that we call the *local covariance function*, as

$$C^{(u,v)}(-\tau, z) := \Phi_J(\tau)^{\top} S^{(u,v)}_J(z). \tag{4.4.4}$$

The following corollary summarizes the conclusions of the above development.

Corollary 2 *Consider a sequence of elements in the class of $(J(n), q)-$MLSW processes that satisfies Assumption 1. If $J = O(\log_2(n))$ then*

$$\sum_{\tau \in \mathbb{Z}} \int_{[0,1]} |C^{(u,v)}_n(\tau, z) - C^{(u,v)}(\tau, z)|(z) dz = O(\frac{1}{n}).$$

uniformly in u, v, where $C^{(u,v)}(-\tau, z) := \Phi_J(\tau)^{\top} S^{(u,v)}_J(z)$.

Hence asymptotically, the time-varying covariance between $X_n^{(u)}(t)$ and $X_n^{(v)}(t-\tau)$ is a linear combination of the CCWF, with time-varying coefficients being the EWS function. In the next subsection, we asymptotically identify the EWS function, hence providing an asymptotically unique representation of the second-order structure of the vector of time series.

4.4.2 Asymptotic identification of the covariance representation

We aim to identify $S_J^{(u,v)}(z)$ asymptotically, starting from Equation (5.2.2) we have that

$$\begin{aligned} C^{(u,v)}(-\tau, z) &= \Phi_J(\tau)^\top S_J^{(u,v)}(z) \\ \Phi_J(\tau) C^{(u,v)}(-\tau, z) &= \Phi_J(\tau)\Phi_J(\tau)^\top S_J^{(u,v)}(z) \\ \sum_{\tau\in\mathbb{Z}} \Phi_J(\tau) C^{(u,v)}(-\tau, z) &= \left(\sum_{\tau\in\mathbb{Z}} \Phi_J(\tau)\Phi_J(\tau)^\top\right) S_J^{(u,v)}(z). \end{aligned} \tag{4.4.5}$$

We define the following matrix.

Definition 8 *For any $q, J \in \mathbb{N}$ with $q \leq J$ and $\Phi_{\tau,J}$ as in Definition 7 we define*

$$A_{q,J} := \sum_{\tau\in\mathbb{Z}} \Phi_{\tau,J}\Phi_{\tau,J}^\top.$$

This $(2q+1)J \times (2q+1)J$ is the Gram matrix of the CCWF over $\tau \in \mathbb{Z}$. Because Theorem 6 states the linear independence of the CCWF, the matrix $A_{q,J}$ is invertible for any $J = 1, 2, \ldots$ Hence from Equation (4.4.5) we obtain

$$A_q^{-1} \sum_{\tau\in\mathbb{Z}} \Phi_J(\tau) C^{(u,v)}(-\tau, z) = S_J^{(u,v)}(z).$$

Hence the EWS functions may be written as the inverse representation of the local covariance function.

We have yet to identify the EWS function since to do so we must prove that it as the image of an injection from an identified parameter of the probability distribution of our observable (and the local autocovariance function is not). Hence we denote as S_n the inverse representation of the covariance function $C_n(\tau, z)$, that is for some $J > 0$ we define

$$S_{z,J,n}^{(u,v)} := A_{q,J}^{-1} \sum_{\tau\in\mathbb{Z}} \Phi_{\tau,J} C_n^{(u,v)}(-\tau, z).$$

Let ι represent a vector of one's in $\mathbb{R}^{J(2q+1)}$. In the following, the first line is obtained by the triangle inequality; the second line is obtained by Cauchy-Schwartz inequality (on the sum for $\tau \in \mathbb{Z}$) and the fact that the absolute value

of the CCWF is bounded by one; the third line is obtained using $\|v\|_2 \leq \|v\|_1$ and applying Tonelli's theorem[1] twice; and the fourth line is obtained by applying Corollary 2.

$$\begin{aligned}
&\int_{[0,1]} \|S^{(u,v)}_{z,J,n} - S^{(u,v)}_{J}(z)\|_2 dz \\
&\leq \|A^{-1}_{q,J}\|_2 \int_{[0,1]} \| \sum_{\tau\in\mathbb{Z}} \Phi_{\tau,J}\big(C^{(u,v)}(-\tau,z) - C^{(u,v)}_n(-\tau,z)\big)\|_2 dz \\
&\leq \|A^{-1}_{q,J}\|_2 \int_{[0,1]} \| \sum_{\tau\in\mathbb{Z}} \iota|C^{(u,v)}(-\tau,z) - C^{(u,v)}_n(-\tau,z)|\|_2 dz \\
&\leq \|A^{-1}_{q,J}\|_2 \|\iota \sum_{\tau\in\mathbb{Z}} \int_{[0,1]} |C^{(u,v)}(-\tau,z) - C^{(u,v)}_n(-\tau,z)| dz\|_1 \\
&\leq \|A^{-1}_{q,J}\|_2 O(\frac{J(2q+1)}{n}).
\end{aligned}$$

Hence for $S^{(u,v)}_{z,J,n}$ to converges with $J = J(n)$ we need $2qJ/(n\delta_{J,q}) = o(1)$ where $\delta_{J,q}$ is a lower bound on the minimum eigenvalue of $A_{q,J}$. In this thesis we have not derived $(J,q) \to \delta_{J,q}$ and only know that $\delta_{J,q} > 0$ for every $J \in \mathbb{N}$ and any $q = \{1,...,J\}$ (by the linear independence of the CCWF). In the remaining of the thesis, when this is relevant, we display the rate of convergence in function of $\delta_{J,q}$ and leave the derivation of $J \to \lambda_J$ for further research. We mention that for Haar and Shannon wavelets, for $q = 0$, Nason et al. (2000) proves that there exists an absolute lower δ bound such that for all $J \in \mathbb{N}$, $\|A^{-1}_{0,J}\|_2 > \delta$.

In conclusion of this section, we state the following proposition.

Proposition 10 *Under the assumptions in Definition 6 and Assumption 1, consider the Gram matrix $A_{q,J}$ in Equation (8). If there exists $\delta > 0$ such that the smallest eigenvalue of $A_{q,J}$ is bounded below by δ for any $J \in \mathbb{Z}$, we have*

$$\int_{[0,1]} \|S^{(u,v)}_{z,J,n} - S^{(u,v)}_{J}(z)\|_2 dz = O(\delta^{-1}_{J,q}\frac{J(2q+1)}{n})$$

uniformly in u, v, where $\delta_{J,q}$ is a lower bound on the minimum eigenvalue of $A_{q,J}$ in Definition 8.

[1] Tonelli's theorem states that if m_1 and m_2 are two *finite* measure on $(\Omega_1, \mathcal{A}_1)$ and $(\Omega_2, \mathcal{A}_2)$, then for all positive function f measurable on $\sigma(\mathcal{A}_1 \times \mathcal{A}_2)$ then $\int_{\Omega_1}(\int_{\Omega_2} f dm_2)dm_1$ is equal to $\int_{\Omega_2}(\int_{\Omega_1} f dm_1)dm_2$. Here we have $\Omega_1 = (0,1)$, Ω_2 is the set of all finite norm vectors in $\mathbb{R}^{(2q+1)J}$, and Ω_3 is the subset of $\mathbb{Z}$ where $\|\Phi_J(\tau)\|_2 \neq 0$.

4.5 Appendix

4.5.1 Proof of Lemma 13

Let $u = t - \tau$ then we can write

$$\begin{aligned}
\Psi_{j,l}(\tau) &= \sum_{t\in\mathbb{Z}} \psi_{j,0}(t)\psi_{l,\tau}(t) \\
&= \sum_{t\in\mathbb{Z}} \psi_{j,0}(t)\psi_{l,0}(t-\tau) \\
&= \sum_{t\in\mathbb{Z}} \psi_{j,0}(u+\tau)\psi_{l,0}(u) \\
&= \sum_{t\in\mathbb{Z}} \psi_{j,-\tau}(u)\psi_{l,0}(u) \\
&= \Psi_{l,j}(-\tau)
\end{aligned}$$

which is the first statement of the Lemma.

Now observe that the support of $\psi_{j,0}$ is $[0, L_j - 1] \cap \mathbb{Z}$ and that the support of $\psi_{l,\tau}$ is $[-\tau, L_l - 1 - \tau] \cap \mathbb{Z}$. Then for any τ such that

$$[0, L_j - 1] \cap [-\tau, L_l - 1 - \tau] \cap \mathbb{Z} = \{0\}$$

we have $\Psi_{j,l}(\tau) = 0$. □

4.5.2 Proof of Lemma 14

The computation in this proof may be difficult to track. Hence we begin by treating some particular cases. Let $j = -1$, then we obtain from the recursive definition of the discrete wavelet function that

$$\Psi_{-1,-1}(m) = \sum_{l\in\mathbb{Z}} g_l g_{l+m}.$$

Now let $j = -2$, in this case, we have, using the recursive definition of discrete wavelet functions, that

$$\begin{aligned}
\Psi_{-2,-2}(2m) &= \sum_{t\in\mathbb{Z}} \psi_{-2,0}(t)\psi_{-2,0}(t-2m) \\
&= \sum_{t\in\mathbb{Z}} \sum_{l\in\mathbb{Z}} h_{t-2l} g_l \sum_{l'\in\mathbb{Z}} h_{t-2l'-2m} g_{l'} \\
&= \sum_{l\in\mathbb{Z}} g_l \sum_{l'\in\mathbb{Z}} g_{l'} \sum_{t\in\mathbb{Z}} h_{t-2l} h_{t-2m-2l'} \\
&= \sum_{l\in\mathbb{Z}} g_l g_{l-m}
\end{aligned}$$

where the last line follows from Assumption 4.3.3, that ensures the term given by $\sum_{t\in\mathbb{Z}} h_{t-2l}h_{t-2m-2l'} = 1$ if $l' = l - m$ and is zero otherwise.

Once again, let $j = -3$, hence by following the same reasoning we write

$$\begin{aligned}
\Psi_{-3,-3}(4m) &= \sum_{t\in\mathbb{Z}} \psi_{-3,0}(t)\psi_{-3,0}(t-4m) \\
&= \sum_{t\in\mathbb{Z}}\sum_{l_1\in\mathbb{Z}} h_{t-2l_1}\psi_{-2,0}(l_1)\sum_{l_1'\in\mathbb{Z}} h_{t-4m-2l_2'}\psi_{-2,0}(l_1') \\
&= \sum_{l_1\in\mathbb{Z}} \psi_{-2,0}(l_1)\sum_{l_1'\in\mathbb{Z}} \psi_{-2,0}(l_1')\sum_{t\in\mathbb{Z}} h_{t-2l_1}h_{t-4m-2l_1'} \\
&= \sum_{l_2\in\mathbb{Z}} g_{l_2}\sum_{l_1\in\mathbb{Z}} h_{l_1-2l_2}\sum_{l_2'\in\mathbb{Z}} g_{l_2'}\sum_{l_1'\in\mathbb{Z}} h_{l_1'-2l_2'}\sum_{t\in\mathbb{Z}} h_{t-2l_1}h_{t-4m-2l_1'}
\end{aligned}$$

using that $\sum_{t\in\mathbb{Z}} h_{t-2l_1}h_{t-4m-2l_1'} = 1$ if $l_1' = l_1 - 2m$ and zero otherwise, we have that

$$\Psi_{-3,-3}(4m) = \sum_{l_2\in\mathbb{Z}} g_{l_2}\sum_{l_2'\in\mathbb{Z}} g_{l_2'}\sum_{l_1\in\mathbb{Z}} h_{l_1-2l_2}h_{l_1-2m-2l_2'}$$

and using the fact that $\sum_{l_1\in\mathbb{Z}} h_{l_1-2l_2}h_{l_1+2m-2l_2'} = 1$ if $l_2' = l_2 - m$ we obtain

$$\Psi_{0,j}(4m) = \sum_{l_2\in\mathbb{Z}} g_{l_2}g_{l_2-m}.$$

Now our strategy is clear enough to tackle the general case, for which we have

$$\begin{aligned}
\Psi_{j,j}(2^{-j-1}m) = &\sum_{l_{-j-1}\in\mathbb{Z}} g_{l_{-j-1}} \prod_{u=2}^{-j-1}\left(\sum_{l_{-j-u}\in\mathbb{Z}} h_{l_{-j-u}-2l_{-j-u+1}}\right) \\
&\sum_{l'_{-j-1}\in\mathbb{Z}} g_{l'_{-j-1}} \prod_{u=2}^{-j-1}\left(\sum_{l'_{-j-u}\in\mathbb{Z}} h_{l'_{-j-u}-2l'_{-j-u+1}}\right) \\
&\sum_{t\in\mathbb{Z}} h_{t-2l_1}h_{t-2^{-j-1}m-2l_1'}
\end{aligned} \tag{4.5.1}$$

and since, firstly by assumption (4.2.6) we have that, $\sum_{t\in\mathbb{Z}} h_{t-2l_1}h_{t-2l_1'-2^{j-1}m} = 1$ if $l_1' = l_1 - 2^{-j-2}m$ and since secondly, assumption (4.2.6) provides that

$$\sum_{l_\alpha\in\mathbb{Z}} h_{l_\alpha-2l_{\alpha+1}}h_{l_\alpha-2^{-j-2-\alpha}m-2l'_{\alpha+1}} = \begin{cases} 1 & \text{if } l'_{\alpha+1} = l_{\alpha+1} - 2^{-j-2-\alpha}m \\ 0 & \text{otherwise.} \end{cases}$$

for $\alpha \in \{1, ..., -j-2\}$, then Equation (4.5.1) regress and provides the result we were aiming for, that is $\Psi_{j,j}(2^{-j-1}m) = \sum_{l\in\mathbb{Z}} g_l g_{l-m}$. □

4.5.3 Proof of Lemma 15

From the proof of Lemma 14, it should be clear how, from the recursive definition of discrete wavelet functions, we can express the CCWF in terms of the scaling and wavelet filters, by the following equation

$$\begin{aligned}\Psi_{j,w}(\tau) = &\sum_{l_{-j-1}\in\mathbb{Z}} g_{l_{-j-1}} \prod_{u=2}^{-j-1} \left(\sum_{l_{-j-u}\in\mathbb{Z}} h_{l_{-j-u}-2l_{-j-u+1}} \right) \\ &\sum_{l'_{-w-1}\in\mathbb{Z}} g_{l'_{-w-1}} \prod_{u=2}^{-w-1} \left(\sum_{l'_{-w-u}\in\mathbb{Z}} h_{l'_{-w-u}-2l'_{-w-u+1}} \right) \\ &\sum_{t\in\mathbb{Z}} h_{t-2l_1} h_{t-\tau-2l'_1}\end{aligned}$$

which, if $w \leq j$, can be rewritten as

$$\begin{aligned}\Psi_{j,w}(\tau) = &\sum_{l_{-j-1}\in\mathbb{Z}} g_{l_{-j-1}} \prod_{u=2}^{-j-1} \left(\sum_{l_{-j-u}\in\mathbb{Z}} h_{l_{-j-u}-2l_{-j-u+1}} \right) \\ &\sum_{l'_{-w-1}\in\mathbb{Z}} g_{l'_{-w-1}} \prod_{u=2}^{-w+j+1} \left(\sum_{l'_{-w-u}\in\mathbb{Z}} h_{l'_{-w-u}-2l'_{-w-u+1}} \right) \\ &\prod_{u=-w+j+2}^{-w-1} \left(\sum_{l'_{-w-u}\in\mathbb{Z}} h_{l'_{-w-u}-2l'_{-w-u+1}} \right) \\ &\sum_{t\in\mathbb{Z}} h_{t-2l_1} h_{t-\tau-2l'_1}\end{aligned}$$

and by a change of variable and rearranging the terms, we have

$$\begin{aligned}\Psi_{j,w}(\tau) = &\sum_{l_{-j-1}\in\mathbb{Z}} g_{l_{-j-1}} \sum_{l'_{-w-1}\in\mathbb{Z}} g_{l'_{-w-1}} \prod_{u=2}^{-w+j+1} \left(\sum_{l'_{-w-u}\in\mathbb{Z}} h_{l'_{-w-u}-2l'_{-w-u+1}} \right) \\ &\prod_{u=2}^{-j-1} \left(\sum_{l_{-j-u}\in\mathbb{Z}} h_{l_{-j-u}-2l_{-j-u+1}} \right) \\ &\prod_{u=2}^{-j-1} \left(\sum_{l'_{-j-u}\in\mathbb{Z}} h_{l'_{-j-u}-2l'_{-j-u+1}} \right) \\ &\sum_{t\in\mathbb{Z}} h_{t-2l_1} h_{t-\tau-2l'_1}\end{aligned}$$

which, by applying the same reasoning that in the proof of Lemma 14, collapses to

$$\Psi_{j,w}(\tau) = \sum_{l_{-j-1}\in\mathbb{Z}} g_{l_{-j-1}} \sum_{l'_{-w-1}\in\mathbb{Z}} g_{l'_{-w-1}} \prod_{u=2}^{-w+j+1} \left(\sum_{l'_{-w-u}\in\mathbb{Z}} h_{l'_{-w-u}-2l'_{-w-u+1}} \right)$$

with $l'_{j-1} = l_{-j-1} - 2^{j+1}\tau$. But we have that for $u = -w + j + 1$ that $-w - u = -j - 1$. Hence we may write

$$\Psi_{j,w}(\tau) = \sum_{l_{-j-1}\in\mathbb{Z}} g_{l_{-j-1}} \sum_{l'_{-w-1}\in\mathbb{Z}} g_{l'_{-w-1}} \prod_{u=2}^{-w+j} \left(\sum_{l'_{-w-u}\in\mathbb{Z}} h_{l'_{-w-u}-2l'_{-w-u+1}} \right) \sum_{l'_{-j-1}\in\mathbb{Z}} h_{l'_{-j-1}-2^{j+1}\tau-2l'_{-j}}$$

by applying $l'_{j-1} = l_{-j-1} - 2^{j+1}\tau$, rearranging the terms, and letting $\gamma = l_{j-1}$ we get

$$\Psi_{j,w}(\tau) = \sum_{l'_{-w-1}\in\mathbb{Z}} g_{l'_{-w-1}} \prod_{u=2}^{-w+j} \left(\sum_{l'_{-w-u}\in\mathbb{Z}} h_{l'_{-w-u}-2l'_{-w-u+1}} \right) \sum_{\alpha\in\mathbb{Z}} g_\alpha h_{\alpha-2^{j+1}\tau-2l'_{-j}}.$$

Now, since any element of $\mathbb{N}$ may be written as $2^\alpha(2^\beta - 1)$ for some unique $\alpha, \beta \in \mathbb{N}$, for any non zero $m \in \mathbb{Z}$, let

$$\tau = 2^\alpha(2^\beta - 1)m$$

for any $\alpha, \beta \in \mathbb{N}$. We observe that under those conditions, τ may be any elements in $\mathbb{Z}$. Hence we have that

$$2^{j+1}\tau = 2^{j+1+\alpha}(2^\beta - 1)m$$

hence $2^{j+1}\tau$ is even only if $j \geq -\alpha$. Suppose this is the case, then by Assumption 4.3.4, we have that

$$\sum_{\alpha\in\mathbb{Z}} g_\alpha h_{\alpha-2^{j+1}\tau-2l'_{-j}} = 0$$

hence $\Psi_{j,w}(\tau) = 0$.

Now suppose that $j \leq w$. Since we have by Lemma 13, $\Psi_{j,w}(\tau) = \Psi_{w,j}(-\tau)$, we have proven our thesis. □

4.5.4 Proof of Lemma 16

Suppose that $w \leq j$, then from the proof of Lemma 15, we have that

$$\Psi_{j,w}(\tau) = \sum_{l'_{-w-1} \in \mathbb{Z}} g_{l'_{-w-1}} \prod_{u=2}^{-w+j} \left(\sum_{l'_{-w-u} \in \mathbb{Z}} h_{l'_{-w-u} - 2l'_{-w-u+1}} \right) \sum_{\alpha \in \mathbb{Z}} g_\alpha h_{\alpha - 2^{j+1}\tau - 2l'_{-j}}.$$

Let $m = L[h] - 1$, then by definition of wavelet filter, we have that $g_{l'_{-w-1}}$ is non-zero if and only if

$$0 \leq l'_{-w-1} \leq m.$$

Moreover, by definition of the scaling filter, for all $u = 2, ..., -w + j$ we have that $h_{l'_{-w-u} - 2l'_{-w-u+1}}$ is non-zero, if and only if

$$0 \leq l'_{-w-u} - 2l'_{-w-u+1} \leq m$$

which provides

$$2l'_{-w-u+1} \leq l'_{-w-u} \leq 2l'_{-w-u+1} + m. \tag{4.5.2}$$

Now suppose that

$$l'_{-w-1} > m$$

then we have that $h_{l'_{-w-2} - 2l'_{-w-1}} = 0$ if

$$l'_{-w-2} > 2m + m = 3m.$$

Suppose now that $l'_{-w-2} > 3m$, then we have that $h_{l'_{-w-3} - 2l'_{-w-2}} = 0$ if

$$l'_{-w-3} > 6m + m = 7m.$$

In general, we have that for $u = 2, ..., -w+j$ we have that $h_{l'_{-w-u} - 2l'_{-w-u+1}} = 0$ if

$$l'_{-w-u} > (2^u - 1)m. \tag{4.5.3}$$

Moreover, observe that $g_\alpha h_{\alpha - 2^{j+1}\tau - 2l'_{-j}}$ is non-zero, if and only if

$$2^{j+1}\tau \leq \alpha - 2l'_{-j} \leq 2^{j+1}\tau + m \quad \text{and} \quad 0 \leq \alpha \leq m.$$

$g_\alpha h_{\alpha - 2^{j+1}\tau - 2l'_{-j}} = 0$ if

$$2l'_{-j} < -2^{j+1}\tau - m.$$

For any $\gamma, \beta \in \mathbb{N}$, let $\tau = -2^{\gamma}(2^{\beta}-1)m$. Let $\gamma = 2^{-j-1}$ we have

$$l'_{-j} < (2^{\beta-1}-1)m.$$

and by letting $\beta = -w+j+1$

$$l'_{-j} < (2^{-w+j}-1)m$$

(observe that τ is now fixed and equal to $-2^{-j+1}(2^{-w+j+1}-1)m$.

Now observe that $l'_{-w-u} = l'_{-j}$ if $u = -w+j$, hence we have two conditions for the term $g_{\alpha}h_{\alpha-2^{j+1}\tau-2l'_{-j}} = 0$, namely, the above inequality and Inequality (4.5.3), hence

$$l'_{-j} > (2^{-w+j}-1)m \text{ or } l'_{-j} < (2^{-w+j}-1)m \Rightarrow g_{\alpha}h_{\alpha-2^{j+1}\tau-2l'_{-j}} = 0.$$

This implies that

$$\sum_{\alpha\in\mathbb{Z}} g_{\alpha}h_{\alpha-2^{j+1}\tau-2l'_{-j}} = g_0 h_m.$$

at $\tau = -2^{-j+1}(2^{-w+j+1}-1)m$. Moreover since Inequality 4.5.2 implies that $h_{l'_{-w-u}-2l'_{-w-u+1}}$ unless

$$2l'_{-w-u+1} \leq l'_{-w-u} - m \leq 2l'_{-w-u+1}$$

letting $l'_{-w-u} = (2^{u}-1)m$ implies that

$$2l'_{-w-u+1} \neq (2^{u-1}-1)m \Rightarrow h_{l'_{-w-u}-2l'_{-w-u+1}} = 0$$

hence by induction, we have

$$\Psi_{j,w}(-2^{-j+1}(2^{-w+j+1}-1)m) = g_m g_0 h_m^{-w+j}$$

□

4.5.5 Proof of Proposition 9

Recall that

$$R^{(n,t,u,v)}(z,\tau)$$
$$= \sum_{j,l=-1,\ldots,-J;|j-l|\leq q} \sum_{s=-t}^{(n-1)-t} \psi_{j,0}(-s)\psi_{l,\tau}(-s)\left(S^{(u,v)}_{-j,-l}(\left(\frac{s+[zn]}{n}\right) - S^{(u,v)}_{-j,-l}(z)\right)$$

Since we assumed the EWS functions, $S^{(u,v)}_{-j,-l}(z)$, to be Lipschitz continous on $(0,1)$ with Lipschitz constants $C^{(u,v)}_{j,l}$, we have, for all i,j,u,v,k in their respective domains, that $\int_0^1 |S^{(u,v)}_{-j,-l}(\frac{s+[zn]}{n})) - S^{(u,v)}_{-j,l-}(z)|dz \leq n^{-1}|s|C_{j,l}$. Hence, we obtain the following bound

$$\sum_{\tau\in\mathbb{Z}}\int_{[0,1]}|R^{(n,t,u,v)}(z,\tau)dz|\leq \sum_{j,l=-1,\ldots,-J;|j-l|\leq q}\sum_{\tau\in\mathbb{Z}}\sum_{s=-t}^{(n-1)-t}2|s|\psi_{j,0}(-s)\psi_{l,\tau}(-s)|\frac{C_{i,j}}{n}$$

Furthermore, since outside of their supports the wavelet functions take the value zero, we have that

$$\sum_{s=-t}^{(n-1)-t}|s|\psi_{j,0}(-s)\psi_{l,\tau}(-s)|\leq 2L_{\min(j,l)}\sum_{s=-t}^{(n-1)-t}\psi_{j,0}(-s)\psi_{l,\tau}(-s)|$$

where L_j denotes the support of the wavelet functions at scale j and we obtain the following bound

$$\sum_{\tau\in\mathbb{Z}}\int_{[0,1]}|R^{(n,t,u,v)}(z,\tau)dz|\leq \sum_{j,l=-1,\ldots,-J;|j-l|\leq q}\frac{2C_{i,j}L_{\min(j,l)}}{n}\sum_{\tau\in\mathbb{Z}}\sum_{s=-t}^{(n-1)-t}|\psi_{j,0}(-s)\psi_{l,\tau}(-s)|$$

Finally, for any $\tau\in\mathbb{Z}$, by Cauchy-Schwartz inequality, we have that

$$\lim_{n\to\infty}\sum_{s=-t}^{(n-1)-t}|\psi_{j,0}(-s)\psi_{l,\tau}(-s)|\leq 1$$

which implies that (since wavelets are zero outside of their support)

$$\sum_{\tau\in\mathbb{Z}}\sum_{s=-t}^{(n-1)-t}|\psi_{j,0}(-s)\psi_{l,\tau}(-s)|\leq 4L_{\min(j,l)}$$

and consequently, we have

$$\sum_{\tau\in\mathbb{Z}}\int_{[0,1]}|R^{(n,t,u,v)}(z,\tau)dz|\leq\frac{4}{n}\sum_{j,l=-1,\ldots,-J;|j-l|\leq q}C_{i,j}^{(u,v)}L_{\min(j,l)}^2$$

where the left-hand side convergence to zero as $n\to\infty$ since in Assumption 1 we assumed that the Lipschitz constants $C_{i,j}^{(u,v)}$ satisfy

$$\sup_{u,v=1,\ldots,p}\sum_{j,l=-1,\ldots,-J;|j-l|\leq q}C_{j,l}^{(u,v)}L_l^2<\infty.$$

□

Five

Inference in the Multivariate Locally Stationary Wavelet model and application

In the previous chapter, we define the class of *Multivariate Locally Stationary Processes* (MLSW) and demonstrate that its *Evolutionary Wavelet Spectrum* (EWS, see Definition 6) is asymptotically identified thanks to the linear independence of the Cross-correlation of wavelet functions (CCWF) (see Definition 5 and Theorem 6).

This chapter studies a consistent estimator of the EWS in the MLSW class of process. This estimator, called the *corrected and smoothed wavelet periodogram* extends the estimator provided in Nason et al. (2000) for the smaller class of *locally stationary wavelet process*. Moreover, since the extension of the class of Locally Stationary Wavelet process into the multivariate framework opens the opportunity to study the stability of the estimation when the multivariate time series is *high-dimensional*, that is when p, the number of time series, is itself a function of the number of observations n. The estimation in this setting is achieved at the end of Section 2.

In Section 3, we consider an economic application of our methods. The application considers measuring the (lack of) coordination between Saudi Arabia and the rest of the OPEC in oil production. This application is a contribution to a large field of applied econometrics that has been puzzled since the eighties by the contradictory reports of (lack) of statistical measurement of the said coordination (e.g. Griffin (1985), Smith (2005), Alkhathlan, Gately and Javid (2014), Al Rousan, Sbia and Tas (2018)). In the past decade, this field came to recognize that changes in market conditions as well as changes in the economic strategies of the players on the market, render difficult the interpretation of statistical measurement. In particular Alkhathlan et al. (2014) argues that one needs to employ multiple economic models to analyze the behavior of Saudi Arabia at different periods of time in order to accommodate

the shifts in its behavior. We argue that the issues encountered with statistical analysis of the behavior of Saudi Arabia and OPEC can be traced back to the assumption of stationarity upon which have relied on the econometric methods employed to tackle the question up to now. As it appears clearly in Figure 1.2, the evolution of the quarterly oil production of Saudi Arabia, as well as the one of the remaining of OPEC display change in its volatility across times, a feature that is not compatible with the assumption of stationarity. We employ the methods studied in this paper to provide our own measurements of the similarity of the response of Saudi Arabia and the rest of OPEC when submitted to random shocks. To the best of our knowledge, those measurements of the coordination of Saudi Arabia with its partners in OPEC are the first that does not rely on the assumption of stationarity. Our results, displayed in Section 5, are consistent with qualitative analysis of the historical trends in the oil market.

5.1 Estimation of the Evolutionary Wavelet Spectrum in the $(J, q)-$Multivariate Locally Stationary Wavelet Model

This section introduces the estimation procedure of the EWS function. In the remaining of this chapter, we assume that for some $n \in \mathbb{N}$ we observe $\{\{X_n^{(u)}(t)\}_{t=0}^{n-1}\}_{u=1}^{p}$ that is in the class of $(J,q) - MLSW$ processes defined in Definition 5. We begin by defining the collection of wavelet coefficients

$$d_{jk;n}^{(u)} := \sum_{t=0}^{n-1} \psi_{j,k}(t) X_n^{(u)}(t) \tag{5.1.1}$$

$$\tag{5.1.2}$$

with $j = -1, ..., -J$, $k = 0, ..., (n-1)$ and $u = 1, ..., p$. Similarly, define the vector

$$d_{k;Jn}^{(u)} = \sum_{t=0}^{n-1} \psi_{k;J}(t) X_n^{(u)}(t)$$

where the collection $\left\{\{\psi_{j,k}\}_{j=-J}^{-1}\right\}_{k=0}^{n-1}$ is a set of *discrete non-decimated wavelet functions* and $\psi_{k;J}(t) = (\psi_{-1,k}(t), ..., \psi_{-J,k}(t)))^{\top} \in \mathbb{R}^J$.

Recall that the wavelet coefficients are not identified because systems of *non-decimated* discrete wavelet functions are over-complete. Nevertheless, they are the basic elements of our estimation of the EWS (which is itself asymptotically identified as demonstrated in the previous chapter).

We begin by studying a first estimator of the EWS, that is the *raw wavelet periodogram*.

5.1.1 Raw wavelet periodogram

The raw wavelet periodogram of the MLSW(q) model consists in a set of random variables denoted as $I_{jlk;n}^{(u,v)}$ with $j,l=-1,...,-J$, $u,v=1,...,p$ and $k=0,...,n-1$ such that

$$I_{jlk;n}^{(u,v)}=d_{jk;n}^{(u)}d_{l,k,n}^{(v)}\mathbf{1}_{|l-j|\leq q} \tag{5.1.3}$$

We construct the vector

$$I_{k;n}^{(u,v)}=\left(\left(I_{jlk;n}^{(u,v)}\right)_{l=-1,...,-J;|j-l|\leq q}\right)_{j=-1}^{-J} \tag{5.1.4}$$

The following result states that the raw wavelet periodogram is an asymptotically biased estimator of the EWS of the $(J,q)-$MLSW process.

Theorem 7 *Suppose that for $n,J\in\mathbb{N}_+$ and $q\in\{0,...,J\}$, $\{\{X_n^{(u)}(t)\}_{t=0}^{n-1}\}_{u=1}^{p}$ is in the class of $(J,q)-$MLSW processes defined in Definition 6. Suppose in addition that Assumption 1 holds and that J is such that $2^J(L[h]-1)<n$ (where $L[h]$ is the length of the support of the scaling filter of the wavelet functions). Then for any $u,v=1,...,p$ and any $k=0,...,n-1$, the* raw wavelet periodogram, *$I_{k;n}^{(u,v)}$ defined by Equation (5.1.4) is such that*

$$\mathbb{E}(I_{k;n}^{(u,v)})=A_{q,J}S_J^{(u,v)}(z)+R(J,q,n,z,u,v)$$

where $z=k/n$, $A_{q,J}$ defined in Definition 8, $S_J^{(u,v)}(z)$ defined in Definition 7, and

$$\int_{[0,1]}\|R(J,q,n,z,u,v)\|_2dz=O\left(\frac{J(2q+1)}{n}\right).$$

uniformly in u,v.

The proof is given in the Appendix.

The following theorem states that the variance of the raw wavelet periodogram does not vanish when $n\to\infty$.

Theorem 8 *Under the same assumptions than in Theorem 7. We have that for any $u,v=1,...,p$, any $k=0,...,n-1$ and any $j,l=-1,...,-J$ such that $|l-j|\leq q$, the* raw wavelet periodogram, *$I_{jlk;n}^{(u,v)}$ defined by Equation (5.1.3) is such that*

$$\begin{aligned}&\mathbb{V}ar(I_{jlk;n}^{(u,v)})=\\&\left(\sum_{\tau}\Psi_{j,l}\Phi_J(\tau)S_J^{(u,v)}(z)\right)^2+\left(\sum_{\tau}\Psi_{j,j}\Phi_J(\tau)S_{z,J}^{(u,u)}\right)\left(\sum_{\tau}\Psi_{l,l}\Phi_J(\tau)S_{z,J}^{(v,v)}\right)\\&+O\Big(r(J,q,n,z,u,v)\Big)\end{aligned}$$

where $z \in (0,1)$, $A_{q,J}$ *defined in Definition 8,* $S_J^{(u,v)}(z)$ *defined in Definition 7, and*

$$\int_{[0,1]} r(J,q,n,z,u,v) = O\left(\frac{1}{n}\right).$$

uniformly in u, v, J.

The proof is given in the Appendix.

5.1.2 Smoothed wavelet periodogram

The nonvanishing variance of the raw wavelet periodogram is a known feature of periodograms (even in the stationary case, see *e.g.* Brillinger (2001)). This problem is usually treated by smoothing the raw periodogram, *e.g.* in Nason et al. (2000), the authors obtain a vanishing variance by smoothing the raw wavelet periodogram before correcting its non-asymptotic bias. We apply the same strategy here. Hence we define the following *smoothed wavelet periodogram* as the moving averages of the raw wavelet periodogram.

$$I_{ijk;nM}^{(u,v)} = \frac{1}{2M+1} \sum_{s=-M}^{M} I_{ij,k+s;n}^{(u,v)} \tag{5.1.5}$$

with $j, l = -1, ..., -J$, $u, v = 1, ..., p$ and $k = 0, ..., n-1$.

Moreover, we construct the vector

$$I_{k;nM}^{(u,v)} = \left(\left(I_{jlk;nM}^{(u,v)} \right)_{l=-1,...,-J;|j-l|\leq q} \right)_{j=-1}^{-J}. \tag{5.1.6}$$

The following result is given for the expectation of the smoothed wavelet periodogram.

Theorem 9 *Under the same assumptions than in Theorem 7, we have that, for any* $u, v = 1, ..., p$ *and any* $k = 0, ..., n-1$, *the* smoothed wavelet periodogram, $I_{k;nM}^{(u,v)}$ *defined by Equation (5.1.6) is such that*

$$\mathbb{E}(I_{k;nM}^{(u,v)}) = \mathbb{E}(I_{k;n}^{(u,v)}) + O\left(\frac{M}{n}\right)$$

uniformly in u, v, J, k.

The proof is given in the Appendix.

Moreover, we provide the following result for the variance of the smoothed wavelet periodogram.

Theorem 10 *In, in addition to the assumptions in Theorem 7, for any $j,l = -1,...,-J$ there exists $K>0$, finite, such that*

$$\sup_{u,v=1,...,p}\sup_{z\in(0,1)}|S_{j,l}^{(u,v)}(z)|\leq K/L_{\min(j,l)}$$

then for any $u,v=1,...,p$ and any $k=0,...,n-1$, the smoothed wavelet periodogram, $I_{ijk;nM}^{(u,v)}$ *defined by Equation (5.1.5) is such that*

$$\mathbb{V}ar(I_{ijk;nM}^{(u,v)})=O\left(\frac{B_{j,l,q,J}}{M}\right)$$

where

$$\begin{aligned}B_{j,l,q,J}&=O\left(2^{-\min\{j,l\}}q^2\min(-j,-l)^2\right)+\\&+O\left(2^{-\min\{j,l\}}q\min(-j,-l)r(J,q,n,z,v,u)\right)\\&+O\left(2^{J+\min\{j,l\}}qr(J,q,n,z,v,u)\right)\end{aligned}$$

with

$$\int_{[0,1]}r(q,n,z,u,v)=O\left(\frac{1}{n}\right).$$

uniformly in u,v.

The proof is given in the Appendix.

To make the previous result meaningful, we provide the following Corollary

Corollary 3 *Under the same assumptions as in Theorem 10, if in addition*

$$2^Jq=O(n)\tag{5.1.7}$$

then

$$\mathbb{V}ar(I_{jlk;nM}^{(u,v)})=O\left(\frac{2^{-\min\{j,l\}}q^2\min(-j,-l)^2}{M}\right)$$

uniformly in u,v, for all $j,l=-1,...,J$, $k=0,...,n-1$, $M=0,...,(n-1)/2$.

We observe that the previous result does not provide that for $2^{J(n)}q(n)=O(n)$ the variance of the smoothed wavelet program vanishes at all scales. If, for instance, we let q be constant in n and let $J(n)=\log_2(n)$ we obtain that $\mathbb{V}\text{ar}(I_{-J,-J,k,n,M}^{(u,v)})=O(n\log_n(n)^2/M)$ that does not vanish asymptotically, even for increasing M in n.

5.1.3 Smoothed and corrected wavelet periodogram

To obtain a consistent and asymptotically unbiased estimator of the EWS functions, we apply a correction to the smoothed wavelet periodogram. Hence we define the *smoothed and corrected wavelet periodogram* by

$$\tilde{I}^{(u,v)}_{k;nM} = A^{-1}_{q,J} I^{(u,v)}_{k;nM} \tag{5.1.8}$$

for all $k = 0, ..., n-1$ and all $u, v = 1, ..., p$.

Since $A_{q,J}$ is a non-random, invertible matrix, by the bounded inverse theorem, $A^{-1}_{q,J}$ is bounded, and by Theorem 9 and 7 the following result is immediate.

Corollary 4 *Under the same assumption that in Theorem 9, the* smoothed and corrected wavelet periodogram *defined in Equation* (5.1.8) *is such that*

$$\mathbb{E}(\tilde{I}^{(u,v)}_{k;nM}) = S^{(u,v)}_J(z) + O(\frac{qJM}{n})$$

uniformly in u, v, k, *and where* $z = k/n$.

Hence $\tilde{I}^{(u,v)}_{k;nM}$ is asymptotically non biased as long as $qJM = o(n)$.

Moreover under the same set of assumption than in Corollary 3,$\delta_{J,q} > 0$ be a lower bound on the smallest eigenvalue of $A_{q,J}$, then we obtain

$$\begin{aligned}
\mathbb{E}\|\tilde{I}^{(u,v)}_{k;nM} - \mathbb{E}(\tilde{I}^{(u,v)}_{k;nM})\|_2^2 &\leq \|A^{-1}_{q,J}\|_2^2 \mathbb{E}\|I^{(u,v)}_{k;nM} - \mathbb{E}(I^{(u,v)}_{k;nM})\|_2^2 \\
&\leq (2q+1)J\|A^{-1}_{q,J}\|_2^2 \mathbb{E}\|I^{(u,v)}_{k;nM} - \mathbb{E}(I^{(u,v)}_{k;nM})\|_\infty^2 \\
&= (2q+1)J\|A^{-1}_{q,J}\|_2^2 \max_{j,l} \mathbb{V}\mathrm{ar}(I^{(u,v)}_{jlk;nM}) \\
&= (2q+1)J\|A^{-1}_{q,J}\|_2^2 \mathbb{V}\mathrm{ar}(I^{(u,v)}_{-J,-J,k,n,M}) \\
&= O(\delta^{-1}_{J,q} J^3 q^3 2^J / M)
\end{aligned} \tag{5.1.9}$$

We provide a refined version of this result in the next theorem. To do so we need some technicalities. Let $A_{j,l,r,i}(q,J) = \sum_{\tau\in\mathbb{Z}} \Psi_{j,l}(\tau)\Psi_{r,i}(\tau)$ be an element of the matrix $A_{q,J}$. Moreover let $A^{-1}_{j,l,r,i}(q,J)$ be the corresponding element in the matrix $A^{-1}_{q,J}$. Now define

$$\tilde{I}^{(u,v)}_{jlk;nM} := \sum_{r,i=-1,-...,-J;|r-i|\leq q} A^{-1}_{j,l,r,i}(q,J) I^{(u,v)}_{ri;nM}. \tag{5.1.10}$$

Then we have

$$\tilde{I}^{(u,v)}_{k;nM} = \left((\tilde{I}^{(u,v)}_{jlk;nM})_{l=-1,...,-J;|j-l|\leq q} \right)_{j=-1}^{-J}$$

almost surely.

Theorem 11 *Under the same assumptions that in Corollary 3, if in addition there exists a finite C_J such that $\sum_{r,i=-1,-...,-J;|r-i|\leq q}|A^{-1}_{j,l,r,i}(q,J)| \leq C_J L_{\min(j,l)}$ (for $A^{-1}_{j,l,r,i}(q,J)$ defined as just above) , then*

$$\mathbb{V}ar\big(\tilde{I}_{jlk;nM}(u,v)\big) = O\bigg(\frac{2^{-\min\{j,l\}}q^2\min(-j,-l)^2}{M}\bigg).$$

uniformly in u,v,k, with $\tilde{I}_{jlk;nM}$ as in Equation (5.1.10), for all $j,l=-1,...,J$, $k=0,...,n-1$, $M=0,...,(n-1)/2$.

The proof is given in the Appendix.

As it appears in Theorem 11, the convergence rate of the smoothed and corrected wavelet periodogram depends on M, the parameter of the smoothing step. Hence we can optimize the rate of convergence for M.

In the following, we consider J to be fixed. Under this condition we observe that under the assumptions of Corollary 3, by Theorem 4 and Equation (5.1.9) we have that the mean squared error of the corrected and smoothed wavelet periodogram is

$$\mathbb{E}\|\tilde{I}^{(u,v)}_{k;nM} - S^{(u,v)}_J(z))\|_2^2 = O(1/M) + O(M/n)$$

Secondly, we let $M = o(n^\gamma)$ with $0<\gamma<1$, in which case we have that

$$\mathbb{E}\|\tilde{I}^{(u,v)}_{k;nM} - S^{(u,v)}_J(z))\|_2^2 = O\Big(n^{-\gamma} + n^{\gamma-1}\Big)$$

hence the optimal rate of convergence is obtained at $\gamma = 1/2$ which gives

$$\mathbb{E}\|\tilde{I}^{(u,v)}_{k;nM} - S^{(u,v)}_J(z))\|_2^2 = O\Big(n^{-1/2}\Big).$$

Hence we state the following corollary that provides the convergence in the probability of our estimation.

Corollary 5 *Under the same set of assumption that in Corollary 3, suppose $J\in\mathbb{N}$, then we have*

$$\mathbb{P}\Big(\|\tilde{I}^{(u,v)}_{k;nM} - S^{(u,v)}_J(z)\|_2 > \epsilon\Big) = O\Big(n^{-1/2}\Big).$$

uniformly in u,v,J,k,M, where $z=k/n$, $\tilde{I}^{(u,v)}_{k;nM}$ is defined in Equation (5.1.8) and $S^{(u,v)}_J(z)$ is defined in Definition 7.

5.1.4 Estimation in High Dimensions

Up to this point, we assumed that for any n, the dimension of the vector of time series $(X_n^{(1)}(t), ..., X_n^{(p)}(t))^\top$, namely p, was fixed and given. Here we extend the results of the previous section to situations where p is a function of n. In particular we consider $p(n) = O(n^\alpha)$ where $\alpha > 0$.

Hence in this section, we consider a *sequence* in n of triply indexed stochastic processes

$$\left\{(X_n^{(1)}(t), ..., X_n^{(p_n)}(t))^\top\right\}$$

all of which satisfies Definition 6.

We observe that this setting is typical of the double asymptotic paradigm where asymptotic reasoning is conducted by letting both n and p grow to infinity. The consistency of the smoothed and corrected wavelet periodogram in this setting is given in the following result.

Theorem 12 *For $n \in \mathbb{N}$ let*

$$\mathcal{S}^{(n)} := \left\{(X_n^{(1_n)}(t), ..., X_n^{(p_n)}(t))^\top\right\}$$

be a sequence of triply-indexed stochastic processes. Assume that for each n, $\mathcal{S}^{(n)}$ is in the class of $(J(n), q)-$MLSW processes defined in Definition 6 and that satifies Assumption 1. In addition, assume there exists $J^\star$ such that for all $n \in \mathbb{N}$, $J(n) \leq J^\star$,

Finally, construct the vectors

$$\tilde{I}_{k;nM} = \left(\tilde{I}_{k;nM}^{(u,v)}\right)_{u,v=1,...,p_n}$$

and

$$S_{z,J(n)} = \left(S_{z,J(n)}^{(u,v)}\right)_{u,v=1,...,p_n}.$$

If $M = o(n^{1/2}$ then we have for all $k = 0, ..., n-1$ that

$$\mathbb{P}\left(\|\tilde{I}_{k;nM} - S_{z,J(n)}\|_2 > \epsilon\right) = O\left(\frac{p_n^2}{\sqrt{n}}\right).$$

where $z = k/n$, $\tilde{I}_{k;nM}^{(u,v)}$ is defined in Equation (5.1.8) and $S_J^{(u,v)}(z)$ is defined in Definition 7.

Proof: The result follows from the fact that

$$\mathbb{E}\|\tilde{I}_{k;nM} - S_{z,J(n)}\|_2^2 \leq p_n^2 \sup_{u,v=1,...,p} \mathbb{E}\|\tilde{I}_{k;nM}^{(u,v)} - S_{z,J(n)}^{(u,v)}\|_2^2.$$

□

5.1.5 Simulations

In this section we present a simple example of the application of the estimator on synthetic data. The data generating process is

$$X_n(t) = \sum_{j=-3}^{-1} \sum_{k=0}^{n-1} W_j(z) \varepsilon_{j,k} \psi_{j,k}(t)$$

with

$$\begin{aligned} W_{-1} &= \cos(4\pi z)^2, \\ W_{-2} &= \cos(3\pi z)^2, \\ W_{-3} &= \sin(2\pi z + 0.5)^2 \end{aligned}$$

and $\epsilon_{j,k} \sim \mathcal{N}(0,1)$ such that $\mathbb{C}\text{ov}(\epsilon_{j,k}, \epsilon_{l,k}) = 0.5^{|j-l|}$ and $\epsilon_{j,k}$ is *iid* to $\epsilon_{j,k'}$ for $k, k' = 0, ..., n-1$ and $k' \neq k$.

The target parameter is the Evolutionary Wavelet Spectrum (EWS), of the process. The estimation is conducted with the smoothing parameter fixed at $m = 2\sqrt{n}$, for depth parameter $J = 3$ and order parameter $q = 1$. At the boundaries of the time window, the smoothing step required more data than was is available. To address this problem we computed the smoothed wavelet periodogram by mirroring the data at the boundaries.

The results for $n = 10^3$ are displayed in Figure 5.1 and the results for $n = 10^4$ are displayed in Figure 5.2. The simulations show that the estimator is indeed able to estimate the Evolutionary wavelet spectrum. In particular, the shape of the EWS is well estimated for the main part of the signal. At the boundaries, the quality of the estimation deteriorates. This is most probably due to the artificial mirroring of the data required by the moving average of the smoothing step. It also appears that the quality of the estimation is not uniform across scales. This effect is completely due to the fact that we employed the same smoothing parameter for each scale.

While highlighting the validity of our theoretical results, the simulation also indicates the direction for improvement in the way the estimator is implemented. Firstly, other strategies of smoothing may be considered. Here we mirrored the data at the boundaries, there may be more efficient strategies. More generally more advanced smoothing procedures may be considered. Here we employed a simple moving average, but in Nason et al. (2000) the authors studied the DWT shrinkage and the translation invariant denoising of Coifman and Donoho (1995). Second, the simulations suggest using different smoothing parameters *for each scale*. This suggestion is in line with our assumptions that the time variations of the EWS at lower scales are smoother than at higher scales. Finally, the implementation of the estimator highlights that the quality of the results depends on good the choice of parameters. Addressing this issue would require rendering the estimation *adaptative*. This has

already been done in Van Bellegem and von Sachs (2008) in the univariate framework. The extension of our method in this direction will be addressed in a future research.

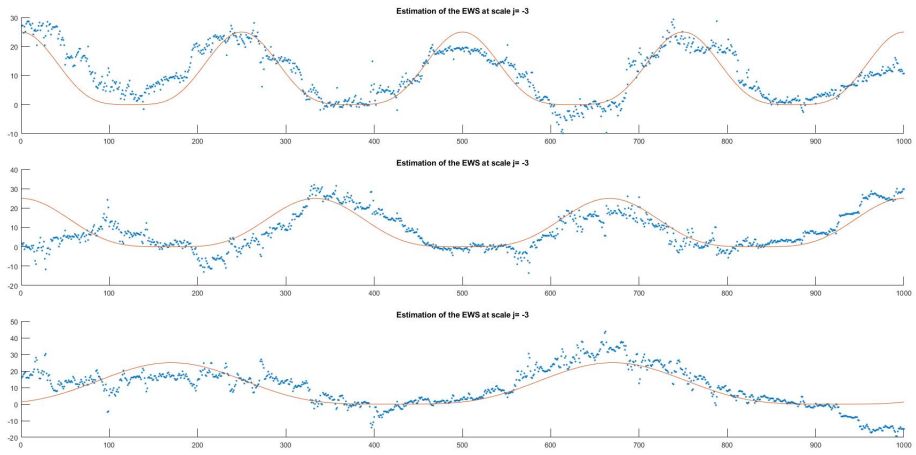

Figure 5.1: Estimation of the EWS for $n = 1000$

Estimation of Evolutionary wavelet spectrum with $n = 1000$. The red curve is the true EWS and the blue dots are the values taken by the Smoothed and Corrected Wavelet periodogram for each $k = 0, ..., n - 1$.

5.2 Application to real data: oil market

In this section, we apply the MLSW model to study the varying similarity between the oil production of Saudi Arabia the one of the remaining of the OPEC during the last half-century.

5.2.1 Motivations

The importance of oil for the world economy could not be too much emphasized. The physical properties of oil drive it to be one of the most important commodities in the world, serving both as an energy source and as a raw material for many petroleum products (gasoline, distillates such as diesel fuel and heating oil, jet fuel, petrochemical feedstocks, waxes, lubricating oils, asphalt,...). The International Energy Agency estimates that in the second decade of the 21^{th} century, oil represents around one-third of the world economy's energy mix. This global reliance on oil grew during the 20^{th} century, and, mostly because oil reserves are scarce and non-uniformly geographically distributed, the oil market has risen in parallel to become one of the most in-

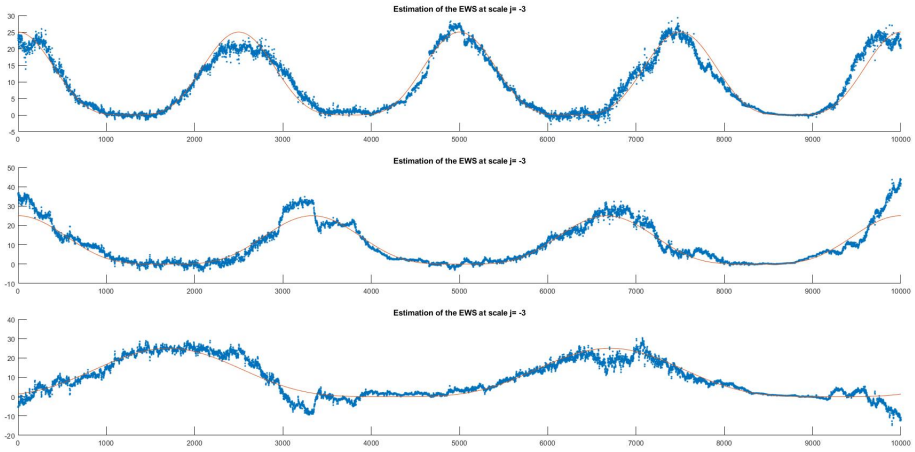

Figure 5.2: Estimation of the EWS for $n = 10000$

Estimation of Evolutionary wavelet spectrum with $n = 10000$. The red curve is the true EWS and the blue dots are the values taken by the Smoothed and Corrected Wavelet periodogram for each $k = 0, ..., n - 1$.

tegrated, influential, and financialized markets in the world economy (Yergin, 2011).

The capital intensity of the oil industry is huge. Hence since the beginning of the 20^{th} century, competition in the oil market has been structured around a small hand of major players (Yergin, 2011). The set of these major competitors change across time, as well as the distribution of power among them (Yergin, 2011). Those changes are driven by political and geopolitical shocks as well as technical discoveries (Yergin (2011), Yergin (2012)).

By 1930, seven major companies, the so-called *seven sisters*, dominated together the oil production worldwide (Yergin, 2011). The seven sisters owned jointly virtually all rights on oil production in Iraq, Iran, Saudi Arabia, and in the Persian Gulf. They coordinated investment and production to maintain prices high while minimizing royalties and taxes due to the governments of the countries in which the seven sisters operated (Yergin (2011), Colgan and Colgan (2021)). In the aftermath of the Second World War, this situation began to be fought by governments of countries wealthy in oil reserves but poor in oil revenues. This led to the nationalization of Iranian oil fields by the Iranian government in 1951 and in a series of meetings and conferences between representatives of governments of major oil-producing countries. Ultimately this chain of events led to the creation of the Organization of Petroleum Export Countries (OPEC) in 1960, and to a wave of nationalization in countries such as Lybia, Algeria, Iraq, Nigeria, Saudi Arabia, and Venezuela in the early

seventies that resulted in a shift of power from the seven sisters to the OPEC (Yergin, 2011).

All observers rapidly agreed that the creation of OPEC and the nationalization of oil fields disrupted the distribution of revenue between major oil companies and governments of oil-producing countries. In the public debate OPEC was soon viewed as a cartel. Nevertheless, in the early eighties, some economists challenged the hypothesis that OPEC possesses the ability to influence price dynamics by implementing collusive behavior (Fattouh, 2007). Different views appeared and the available empirical evidence did not help to narrow the debate (Fattouh, 2007). In his seminal paper, Griffin (1985) observes that empirical studies tend to "reach onto the shelf of economic models to select one, to validate [their] choice[s] by pointing to selected events not inconsistent with model predictions". To circumvent this defect Griffin (1985) used annual data over the years 1971–1982 in order to test alternative theories concerning OPEC. He concluded that the partial-market sharing cartel model cannot be rejected for all countries. Since then a wide and active area of research derived from the one of Griffin (1985). One of the important questions tackled by this literature is about the role of Saudi Arabia and whether it coordinates its production decision with its partners in OPEC (Alkhathlan et al., 2014). The special attention given to Saudi Arabia is backed up by historical analysis of its political role within OPEC (Yergin, 2011), of its economic policy (Alkhathlan et al., 2014), as well as by technical analysis of its production capacities that allowed Saudi Arabia to play the crucial role of *swing producer* in many occasions (Fattouh, 2007).

Many of the articles that followed the work the seminal work of Griffin (1985) have been critically reviewed by Smith (2005) who argues against drawing strong conclusions from the statistical analysis:

> "There is only weak evidence to indicate that Saudi Arabia has acted as a leader or dominant firm within the cartel, although that possibility cannot be formally rejected. If the Saudis have performed such a role, then at least one can say that it has not been executed with sufficient vigor or consistency to be clearly discernible in the data." (Quoted in Alkhathlan et al. (2014))

This conclusion has been criticized in Alkhathlan et al. (2014), who argue that the behavior of Saudi Arabia changes over time. Alkhathlan et al. (2014) reduces the conflict between statistical reports concerning the coordination of Saudi Arabia with the rest of OPEC to the matter of the time frame used in the analysis. Alkhathlan et al. (2014) reports that when studying the correlation between the export or the production of Saudi Arabia with the remaining of the OPEC, Griffin (1985) obtains a positive correlation using 1971-1983 data, that Dahl and Yucel (1991) obtains weak evidence of coordination using 1971-1987 data, that Kaufmann, Bradford, Belanger, Mclaughlin and Miki

(2008) find no correlation using 1987-2003 data and that Al-Yousef (2012) cites 1973-1994 data and obtain negative correlation.

Alkhathlan et al. (2014) argue that the correlation of Saudi Arabia's oil export with the one of the rest of the OPEC, goes through periods of moderately strong correlation followed by periods of weak or negative correlation. Hence, to statistically analyze the relationship between Saudi Arabia and OPEC, Alkhathlan et al. (2014) considered six different periods of time: 1973.Q1–1978.Q3 (positive correlation), 1978.Q3–1985.Q3 (weak or negative correlation), 1985.Q3–1990.Q2 (positive correlation), 1990.Q2–1998.Q2 (weak or negative correlation), 1998.Q2–2009.Q2 (positive correlation), 2009.Q2–2012.Q1 (weak or negative correlation). In Alkhathlan et al. (2014) the authors write

> "Saudi behavior is more complicated than can be captured by a single economic model. Such models were designed to explain firms' economic behavior in particular industries; geopolitical and security concerns are ignored. But Saudi Arabia is the leading producer of the world's most important product, subject to international pressures and security threats to its national sovereignty. Given the dramatic changes in geopolitics and in the market since 1973, it should not be surprising that Saudi behavior cannot be well explained by a single model. [...] several different models would need to be adopted, depending upon the change in circumstances."

This conclusion has been supported by others studies. Fattouh and Mahadeva (2013) argues that theoretical models do not perform well since they do not consider OPEC's varying conduct. Huppmann and Holz (2012) uses game theoretic reasoning to discuss changes in the behavior of the players in the crude oil market. More recently the work of Al Rousan et al. (2018) uses a dynamic network econometric model to measure changes in the relative influence of Saudi Arabia between 2002 and 2016.

The following application of the MLSW process to the oil market is motivated by this debate. The literature agrees that the changing environment and the changing strategies of the player in the oil market render the conclusions drawn from past statistical analysis difficult to interpret. However, the state of the art has made all analyses subject to the assumption of covariance stationarity. Alkhathlan et al. (2014) intend to circumvent the difficulty by applying different models on data from different periods of time defined *a priori* based on qualitative and historical analysis. Al Rousan et al. (2018) use a dynamic network model, but it is restricted to a short period of time.

In the following, we apply the MLSW model to estimate the time-varying similarity between the oil production of Saudi Arabia and others OPEC countries over a long period of time going from 1973 until 2022. By doing so we use a *single model* and provide a coherent response to the question of whether or not Saudi Arabia coordinates with the rest of OPEC. Because we use a

single model to produce the measurement, it becomes possible to answer the question of the relative coordination of Saudi Arabia with OPEC across time.

The measurements show that the level of coordination of Saudi Arabia with the rest of OPEC has greatly increased since the birth of the organization. However, this long-term increase was subject to shocks, and periods of massive drops in coordination are observed (especially in the aftermath of the first gulf war). The measurement allows identifying *a posteriori* 16 periods of increase, or decrease in cooperation. This segmentation of the last half century in 16 periods is coherent with a consensus analysis of the recent history of the oil market. The projection of this historical analysis into our measurement suggests that the *decisions* of Saudi Arabia (for instance in response to geopolitical events) has a great impact on the level of coordination.

5.2.2 The model

This section defines the model that is estimated below. Let $X^{(1)}(t)$ be the first log difference of the quarterly crude oil production of Saudi Arabia. Let $X^{(2)}(t)$ be the first log difference of the quarterly crude oil production of the OPEC without Saudi Arabia.

We assume that the vector of time series $(X^{(1)}(t), X^{(2)}(t))^\top$ belongs to some class of MLSW(q) processes. Moreover, we assume that each element of the increment process may be written as

$$\varepsilon_{j,k}^{(u)} = \sum_{m=1}^{J(n)p} \Delta_{m,j}^{(u)}(z) Z_{m,k}$$

where $Z_{m,k}$ are independent Gaussian random variables with mean zero and unit variance. The sequence $\{Z_{m,k}\}$ is interesting and implies that at *each point in time* and at *each scale*, there are p exogenous and independent random shocks, that generate the vector of time series. The time-varying second-order structure of the increment of the MLSW(q) process follows from the linear combination of those random shocks. Indeed we have

$$\begin{aligned}
\mathbb{C}\mathrm{ov}(\varepsilon_{j,k}^{(u)} \varepsilon_{l,q}^{(v)}) &= \mathbb{E}\Big(\sum_{m=1}^{J(n)p} \Delta_{m,j}^{(u)}(z) Z_{m,k} \sum_{m'=1}^{J(n)p} \Delta_{m',l}^{(u)}(z) Z_{m',q}\Big) \\
&= \sum_{m=1}^{J(n)p} \Delta_{m,j}^{(u)}(z) \Delta_{m,l}^{(v)}(z) \\
&= \Gamma_{j,l}^{(u,v)}
\end{aligned}$$

where the second line follows from the independence of the random shocks $Z_{m,j}$ and the third line follows from Definition 6. Consequently, by Definition

6 we have that the EWS functions takes the form

$$S_{j,l}^{(u,v)}(z) := W_j^{(u)}(z) W_l^{(v)}(z) \sum_{m=1}^{J(n)p} \Delta_{m,j}^{(u)}(z) \Delta_{m,l}^{(v)}(z). \tag{5.2.1}$$

Moreover we have that

$$X_n^{(u)}(t) = \sum_{k=0}^{n-1} \sum_{m=1}^{J(n)p} Z_{m,k} \sum_{j=-\infty}^{-1} W_j^{(u)}(z) \Delta_{m,j}^{(u)}(z) \psi_{j,k}(t).$$

In this model, we are interested in the vector

$$Y_{n,k}^{(u)} := \{\{X_n^{(u)}(t) Z_{m,k}\}_{m=1}^{J(n)p}\}_{t=0}^{n-1}$$

whose expectation is such that each of its elements provides the covariance at each point in time, between the random variable $X_n^{(u)}(t)$ and the Jp exogenous shocks at time k. As such, this expectation may be seen as a vector of impulse response functions that provides a measure of the response of a given time series at time t when it is submitted to random shocks time k.

We are interested in $Y_{n,k}^{(u)}$, because we want to measure the similarity between the response of two given time series to the Jp exogenous shocks happening at each point in time. We formalize this idea by defining the measure of similarity in the response of two random variables in terms of *cosine* of the angle between the $Y_{n,k}^{(u)}$ and $Y_{n,k}^{(v)}$ for $u \neq v$, that is

$$\cos\left(\theta\big(\mathbb{E}(Y_{n,k}^{(u)}), \mathbb{E}(Y_{n,k}^{(v)})\big)\right) = \frac{\langle \mathbb{E}(Y_{n,k}^{(u)}), \mathbb{E}(Y_{n,k}^{(v)}) \rangle_2}{\|\mathbb{E}(Y_{n,k}^{(u)})\|_2 \|\mathbb{E}(Y_{n,k}^{(v)})\|_2}$$

where

$$\begin{aligned}
&\langle \mathbb{E}(Y_{n,k}^{(u)}), \mathbb{E}(Y_{n,k}^{(v)}) \rangle_2 \\
&= \sum_{m=1}^{J(n)p} \sum_{j=-J(n)}^{-1} \sum_{j'=-J(n)}^{-1} W_j^{(u)}(z) \Delta_{m,j}^{(u)}(z) W_j'^{(v)}(z) \Delta_{m,j'}^{(v)}(z) \sum_{t=0}^{n-1} \psi_{j,k}(t) \psi_{j',k}(t) \\
&= \sum_{j=-J(n)}^{-1} \sum_{m=1}^{J(n)p} W_j^{(u)}(z) W_j^{(v)}(z) \Delta_{m,j}^{(u)}(z) \Delta_{m,j}^{(v)}(z) \\
&= \sum_{j=-J(n)}^{-1} S_{j,j}^{(u,v)}(z) \\
&= C^{(u,v)}(0, z)
\end{aligned}$$

where the first equality follows from direct computation, the second equality follows from orthonormality of the wavelet functions, and the third equality

follows from Equation (5.2.1) and the last line follows from Equation by noting that $\Phi_J(0)$ is the unit vector. Hence

$$\cos\left(\theta\big(\mathbb{E}(Y_{n,k}^{(u)}),\mathbb{E}(Y_{n,k}^{(v)})\big)\right) = \rho^{(u,v)}(z)$$

where $\rho^{(u,v)}(z)$ is the instantaneous correlation between the two variables.

Hence, by Theorem 11 and Slutsky's lemma, a consistent estimator of $\cos\left(\theta\big(\mathbb{E}(Y_{n,k}^{(u)}),\mathbb{E}(Y_{n,k}^{(v)})\big)\right)$ is given by

$$\cos(\hat{\theta}) = \frac{\sum_{j=-J}^{-1}\tilde{I}_{j,j,k,n}^{u,v}}{\left(\sum_{j=-J}^{-1}\tilde{I}_{j,j,k,n}^{u,u}\right)^{1/2}\left(\sum_{j=-J}^{-1}\tilde{I}_{j,j,k,n}^{v,v}\right)^{1/2}} \tag{5.2.2}$$

We are also interested by the spectral decomposition (scale by scale) of the second-order structure of the signals. Hence, since the cosine similarity may be interpreted as the correlation between the two signals, we extend the analysis to measure the *local coherency* of the two-time series. The concept of local coherency may be interpreted, as the *scale*-analog of the correlation between two signals (see *e.g.* Ombao and Van Bellegem (2008)) and is defined as

$$\kappa^{(u,v)}(j,z) = \frac{C^{(u,v)}(0,z)}{\sqrt{C^{(u,u)}(0,z)}\sqrt{C^{(v,v)}(0,z)}}.$$

and is consistently estimated by

$$\hat{\kappa}^{(u,v)}(j,z) = \frac{\tilde{I}_{j,j,k,n}^{u,v}}{\left(\tilde{I}_{j,j,k,n}^{u,u}\right)^{1/2}\left(\tilde{I}_{j,j,k,n}^{v,v}\right)^{1/2}}$$

for all $j = -1, ..., -J$.

5.2.3 Results

We use the data for the quarterly crude oil production (in barrels per day) of Saudi Arabia and of the rest of OPEC that is publicly available on the U.S. Energy Information Administration database. Our data range from the first quarter of 1973 to the third quarter of 2022.

The cosine similarity is estimated for every quarter between 1973Q2 to 2002Q3 using Equation (5.2.2). The results are displayed in Figure 5.3. Figure 5.5 display the corresponding Local coherency.

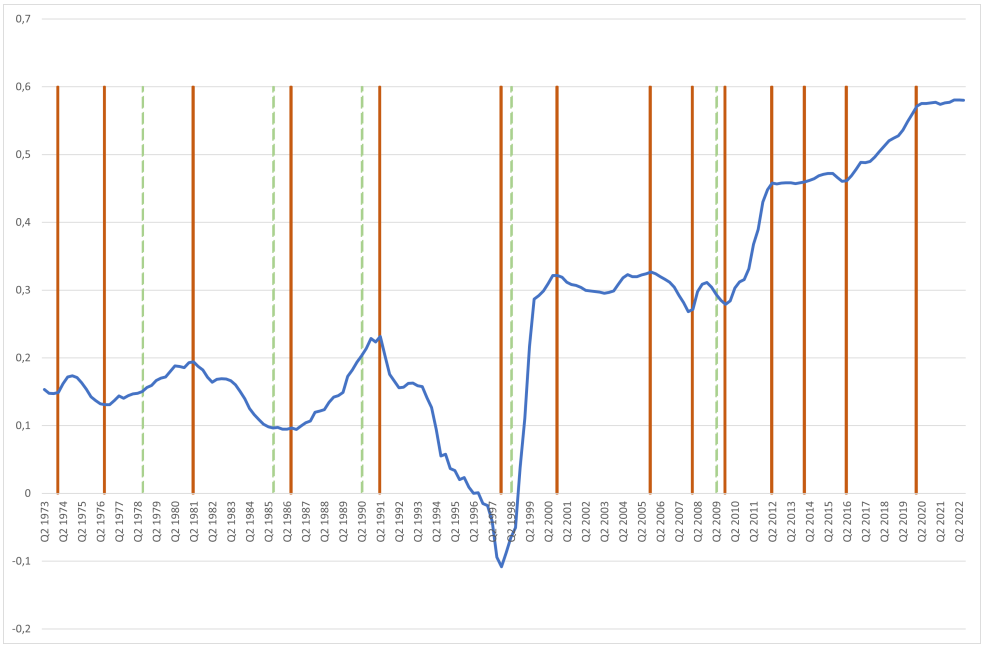

Figure 5.3: Cosine similarity

Estimation of the cosine similarity as defined in Equation (5.2.2) for every quarter between 1973Q2 and 2022Q3. The dotted green vertical line represents the separations between the periods identified by Alkhathlan et al. (2014). The red plain vertical line represents the separations between the period we identify based on our measurements.

Figure 5.3 demonstrates that the level of coordination (here and thereafter assimilated with a positive value of the cosine similarity) of Saudi Arabia with its partners in OPEC greatly varies in time. Interestingly the level at the beginning of the period is much lower (around 0.15) than at the end (nearly 0.6). Moreover, it appears that this long-term increase in cooperation is highly non-linear and goes through periods of decrease in cooperation followed by periods of increase in cooperation.

In Alkhathlan et al. (2014), the authors divided the period going from the second quarter of 1973 to the first quarter of 2012 into six different periods of time during which Saudi Arabia cooperated more, or less with its partners in OPEC. Our measurement allows us a much more fine-grained identification of the period during which the cooperation of Saudi Arabie with the rest of OPEC increased of decreased. For the period going from the first quarter of 1973 to the third quarter of 2022, we identify 16 different periods. Below we project the historical analysis of the history of OPEC provided in Alkhathlan et al. (2014) into our segmentation of time (for segments posterior to the first quarter of 2012, that are not analyzed in Alkhathlan et al. (2014), we

provide our own historical interpretations). We found that our measurement and the periods we identified are consistent with the historical facts pointed out in Alkhathlan et al. (2014) (one may even argue that the periods our measurement allowed us to identify are somewhat more consistent with the historical analysis provided by Alkhathlan et al. (2014) than ones the authors identify in their paper).

Historical interpretation of the curve in Figure 5.3 .

From 1973.Q2 to 1974.Q1 OPEC quatrupled prices and OPEC countries nationalized their oil reserves. This was called the first oil shock (Alkhathlan et al., 2014).

From 1974.Q1 to 1976.Q3: The first oil shocks induced a recession in the world economy. Demand for OPEC oil stopped growing and the organization responded by implementing a policy of proportional restriction on export. (Alkhathlan et al., 2014). This policy seems to have increased coordination between Saudi Arabia and the remaining of OPEC at the beginning of its implementation but we observe the estimated cosine similarity dropping as soon as 1975, suggesting that competition among OPEC members may have had reduced their coordination.

From 1976.Q3 to 1981.Q2 Demand for OPEC oil recovered (Alkhathlan et al., 2014). In late 1978, the Iranian Revolution shut down oil production in Iran and Saudi Arabia increased its own export to partially offset this loss. Interestingly our model does not detect any effect of the Iranian Revolution on the coordination of Saudi Arabia with the remaining of OPEC. This period is marked by an increase in coordination.

From 1981.Q2 to 1986.Q3 The second oil shock happened during the years 1979-1980. In itself, the oil shock did not undermine cooperation in OPEC. Actually OPEC implemented a policy to defend the 1979-1980 price doubling. This policy exacerbated the world recession. Many economies responded by shifting away from OPEC oil (by shifting away from oil to generate electricity and heating space, as well as by buying oil from non OPEC producers) (Alkhathlan et al., 2014). This translated into large cuts in the demand faced by OPEC. It resulted in a significant reduction in coordination between Saudi Arabia and the rest of OPEC.

From 1985.Q3 to 1991.Q2 At the end of 1985, Saudi Arabia abandoned its output-restricting policy to let crude oil price be determined by the *netback pricing* of oil products (Alkhathlan et al., 2014). This was followed by a reinforcement in OPEC coordination as it implemented production quotas to

manage the gradual increases in the export of its member and facilitate the recovery of their market share. Our model captures the effect of this policy through a significant increase in cosine similarity.

From 1991.Q2 to 1997.Q4 In August 1990 Irak invaded Kuwait and the export from both countries stopped. In response, Saudi Arabia responded by doubling its export and maintained it at this level for nearly a decade while the rest of OPEC gradually increased their export (Alkhathlan et al., 2014). Our measurements suggest this period is marked by a drastic dropped in cooperation between Saudi Arabia and the rest of OPEC. From the fourth quarter of 1996 to the third quarter of 1998, the cosine similarity is even negative, suggesting that Saudi Arabia and the rest of OPEC responded in opposite directions to the shocks.

From 1997.Q4 to 2000.Q4 In the aftermath of the Asian Financial Crisis, Saudi Arabia resumed a more active role in OPEC coordination (Alkhathlan et al., 2014). This resulted in a massive increase in the estimated cosine similarity that stabilized around a level of 0.3.

From 2000.Q4 to 2005.Q4 During this period OPEC exports kept on increasing until 2005 (Alkhathlan et al., 2014). Our measurement suggests a stabilization in the coordination of Saudi Arabia with its partners in OPEC.

From 2000.Q4 to 2003.Q1 and from 2003.Q1 to 2005.Q4 During the period going from 2000.Q4 to 2005.Q4 OPEC exports kept on increasing (Alkhathlan et al., 2014). However, our measurement detects a shift in cooperation before and after the invasion of Iraq by the United States of America. Before, our measurement suggest a slow decline in the coordination of Saudi Arabia with its partners in OPEC. After, our measurement shows an increase in cooperation.

From 2005.Q4 to 2008.Q1 In the aftermaths of China entrance into the World Trade Organization and the acceleration of its integration into the international chain of production, the demand for crude oil grew faster than OPEC and non-OPEC oil suppliers could respond (Alkhathlan et al., 2014). In response the level of price surges. That effect negatively affected the coordination of Saudi Arabia with the rest of OPEC.

From 2008.Q1 to 2009.Q4 Consequently to the Great Recession, OPEC implemented a novel policy of proportional restriction on export in order to mitigate the drop in price(Alkhathlan et al., 2014). Interestingly, the effect we measured is similar to the one measured in 1974-1975: an increase in cooperation that is followed by a drop in cooperation leaving the level of

cooperation at the end of the period at a close level than the one at the beginning of the period.

From 2009.Q4 to 2014.Q1 In the aftermath of the Great Recession, prices recovered as a consequence of growing demand from developing countries (Alkhathlan et al., 2014). This effect on price has then been greatly reinforced by the Arab Spring revolts starting in 2011 (Alkhathlan et al., 2014). During this period OPEC responded by increasing its cooperation.

From 2014.Q1 to 2016.Q2 In response to (1) the rise of non-conventional oil production in non-OPEC countries (in particular the shale oil production in the USA), (2) the slowing down of growth in China, (3) a shift of market power from Saudi Arabia to USA (Al Rousan et al., 2018) and (4) the entrance of the oil market in an exploitation phase (characterized by declining price) (Fu, 2016), Saudi Arabia implemented a *market share* strategy during which it increased exports to maintain prices low in the hope to gain market share (Behar and Ritz, 2017). This period is marked by the stability of cooperation between Saudi Arabia and the rest of OPEC.

From 2016.Q1 to 2020.Q1 The market share strategy failed because it was not able to drive shale oil producers out of the market. In response, Saudi Arabia shifts its strategy again in order to secure revenues. In a move to regain market power, OPEC implemented an agreement of cooperation with Russia and nine other non-OPEC producers (*Declaration of Cooperation OPEC and non-OPEC*, 2017). This new group is now known as OPEC+ and still exists to this day. The implementation of OPEC+ led to an important increase in the coordination of Saudi Arabia with the rest of its partners in OPEC.

From 2020.Q1 to 2022.Q3 In early 2020 the COVID-19 pandemic started. From this point, the cosine similarity stabilize at an unprecedented level (near 0.6) despite the huge recession that followed the anti-covid policies. In early 2022, Russia invaded Ukraine, the data at our disposal does not yet allow us to conclude anything about the effect of this shock on the cooperation of Saudi Arabia with the rest of OPEC.

5.2.4 Robustness analysis

The first step of our robustness analysis is to compare the measure in Figure 5.3 with the one of a more common estimator. Since the cosine similarity may be interpreted in terms of instantaneous correlation, we measure this quantity using a rolling window of the same size as the smoothing parameter, which is a more common estimator of the instantaneous correlation.

The results of this estimation are displayed in Figure 5.4. They show that measuring the instantaneous correlation using a rolling window yields a less

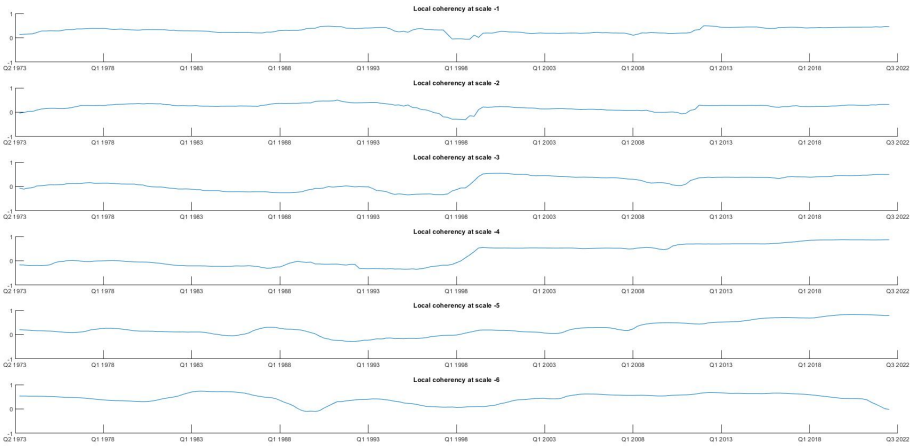

Figure 5.4: Local Coherency

Estimation of local coherency at different scales, for every quarter between 1973Q2 and 2022Q3.

smooth result. This difference is due to the fact that in the estimation conducted via our own estimator, the time series a projected onto the wavelet domain, and hence produces some loss of information. The second main difference between the two estimators is that, even if they display somewhat similar shapes, the values estimated are not of the same magnitude. This also prevails when we compare the on each graph the relative magnitude of the positive and negative spikes.

The second step of our robustness analysis is to change the size of the smoothing parameter. The estimation displayed in Figure 5.3 was obtained using a rather large smoothing parameter (*i.e.* we set the smoothing parameter to $m = 34 \sim n^{2/3}$. This choice, while providing interesting results in regard to the above historical analysis, may induce a somewhat important bias in the estimation by putting too much weight on long terms effects. Hence, in Figure 5.6 we display our results using a much shorter smoothing parameter $m = 4$, meaning we are smoothing over a year.

The estimation of the cosine similarity is displayed in Figure 5.7, the estimation of the local coherency at various scales in Figure 5.8, and the estimation of the instantaneous correlation measure via a rolling window of the same size as the smoothing parameter is displayed in Figure. The most important conclusion we draw from the result is that the time localization, as well as the relative amplitude of the shocks, are affected by this change. This difference appears in the estimation provided by the rolling window estimator and hence may be completely attributed to the change of smoothing parameter that is

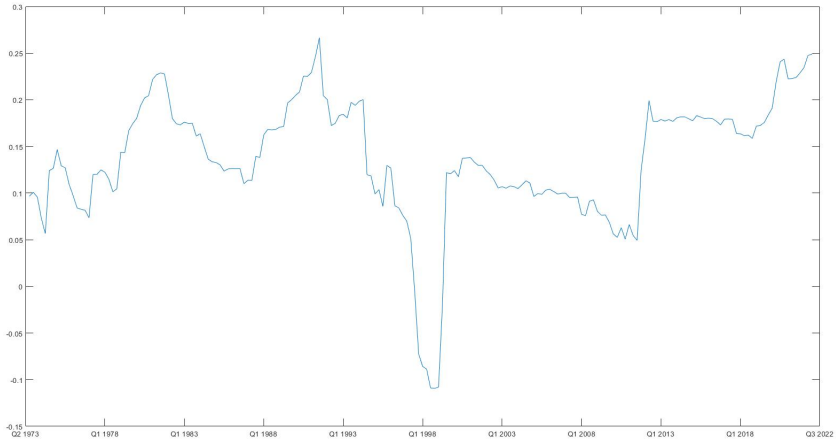

Figure 5.5: Instant correlation estimated using rolling window

Estimation of the instantaneous correlation using rolling a rolling window, for every quarter between 1973Q2 and 2022Q3 with a smoothing parameter of 4.

not an artifact of the wavelet-based estimator. Those differences highlight the importance to render our estimator *adaptative* in future research. Finally, the estimation of the local coherency appears to be more informative once we reduce the smoothing parameter.

The last step of our robustness analysis is to use higher-frequency data. Hence, we resume the estimation using monthly data and fix the smoothing parameter to have the same order of magnitude relative to the sample size, as in the estimation displayed in Figure 5.3, namely approximatively $n^{2/3}$.

The results are displayed in Figure 5.9, 5.10 and 5.11. They mostly indicate the same patterns as the estimation conducted with the quarterly data. The main and important difference comes from the *time-localization* of the shocks.

The main insight of our robustness analysis is that the smoothing step of the estimation is determinant and may provide qualitatively different results. In particular, we stress the importance of the smoothing parameter. As already suggested in the simulations, a more fine-tuned estimation should be possible by choosing a different smoothing parameter for each different scale. Moreover, a deeper analysis of the result, especially with regard to the history of the oil market, may help to select a ´´better" smoothing parameter. However, this strategy may be dubious from a scientific point of view, since it might lead to selecting the smoothing parameter based on a confirmation bias, rather than letting the data speak for itself. This conclusion stresses once again the importance of rendering the estimation adaptative, in the same fashion as in Van Bellegem and von Sachs (2008).

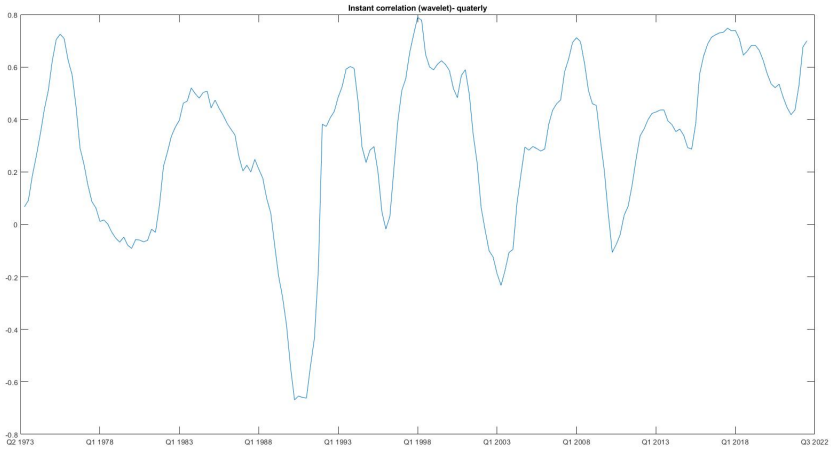

Figure 5.6: Cosine similarity estimated with a shorter smoothing parameter

Estimation of the cosine similarity as defined in Equation (5.2.2) for every quarter between 1973Q2 and 2022Q3. The estimation uses a shorter smoothing parameter than in Figure 5.3 whose value is here fixed at $m = 4$.

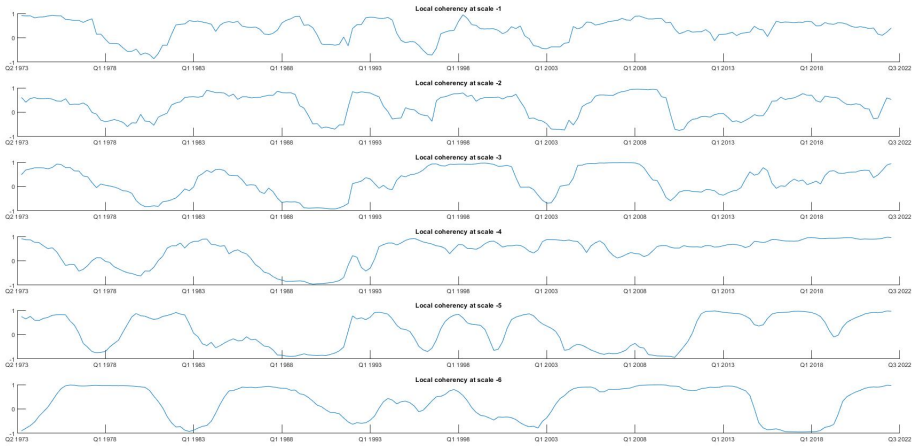

Figure 5.7: Local coherency at different scales estimated with a shorted smoothing parameter

Estimation of local coherency at different scales, for every quarter between 1973Q2 and 2022Q3. The estimation uses a shorter smoothing parameter than in Figure 5.3 whose value is here fixed at $m = 4$.

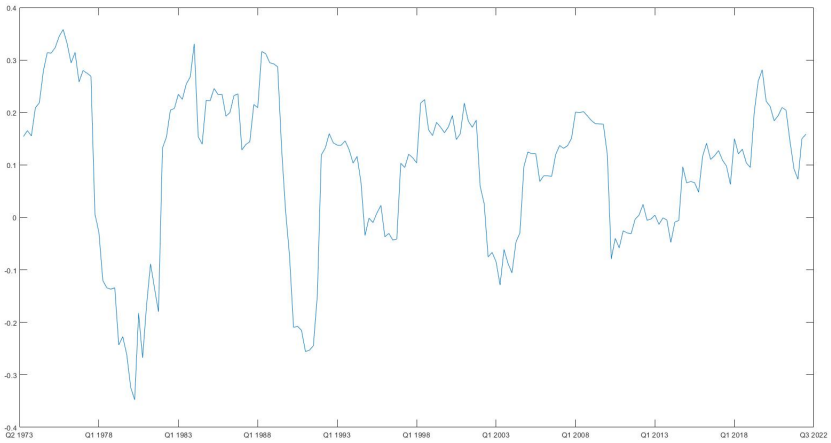

Figure 5.8: Instantaneous correlation estimated using rolling window and a shorter smoothing parameter

Estimation of the instantaneous correlation using rolling a rolling window, for every quarter between 1973Q2 and 2022Q3 with a smoothing parameter of 4.

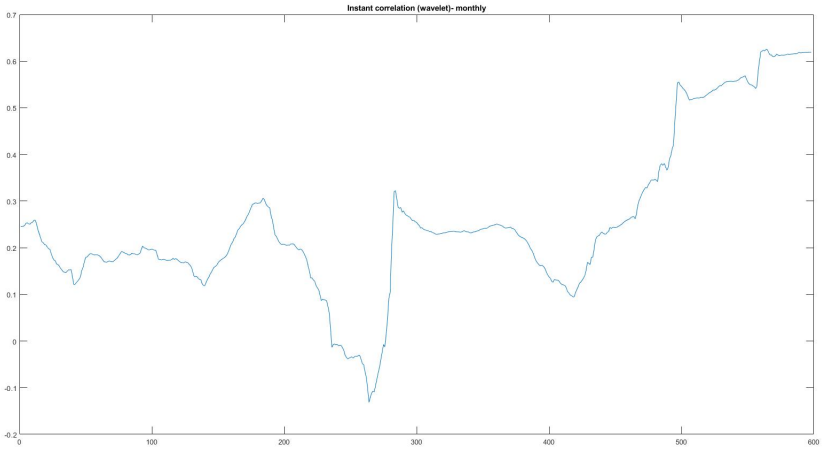

Figure 5.9: Cosine similarity estimated with monthly data

Estimation of the cosine similarity as defined in Equation (5.2.2) for every month between February 1973 and November 2022.

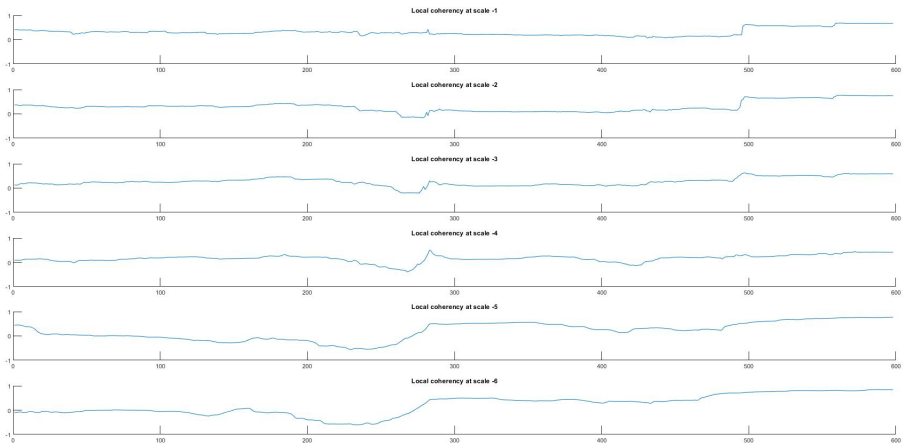

Figure 5.10: Local coherency at different scales estimated with monthly data

Estimation of local coherency at different scales, for every month between February 1973 and November 2022.

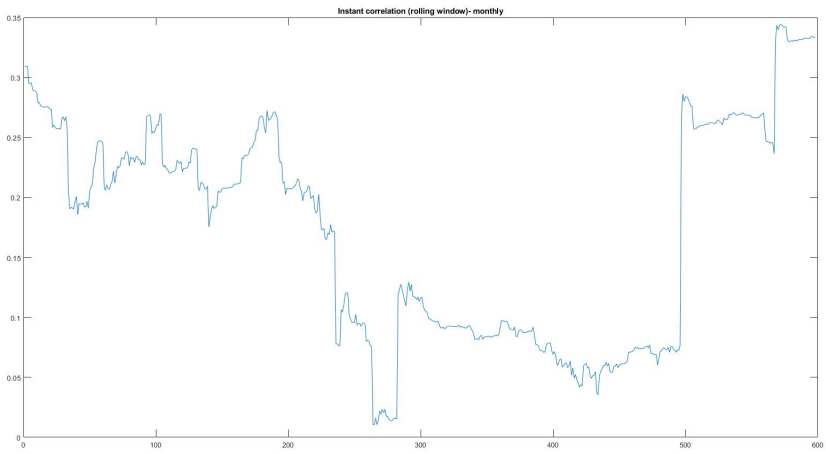

Figure 5.11: Instantaneous correlation estimated using rolling window and monthly data

Estimation of the instantaneous correlation using rolling a rolling window, for every month between February 1973 and November 2022.

5.3 Appendix

5.3.1 Proof of Theorem 7

By Definition 6 we have for $|j-l| \leq q$

$$\begin{aligned}
\mathbb{E}(I^{(u,v)}_{jlk;n}) =& \mathbb{E}\big(\sum_t \psi_{j,k}(t) X_n(t)^{(u)} \sum_{t'} X^{(u)}_n(t')\psi_{l,k}(t')\big) \\
&= \sum_t \sum_\tau \psi_k(t)\psi_k(t-\tau)\mathbb{E}(X_n(t)^{(u)} X^{(v)}_n(t-\tau)) \\
&= \sum_t \sum_\tau \psi_{j,k}(t)\psi_k(t-\tau)\mathbb{C}\mathrm{ov}(X_n(t)^{(u)} X^{(v)}_n(t-\tau)) \\
&= \sum_t \sum_\tau \psi_{j,k}(t)\psi_k(t-\tau) C^{(u,v)}_n(-\tau, z) \\
&= \sum_\tau \Psi_{j,l}(\tau) C^{(u,v)}_n(-\tau, z) \qquad (5.3.1)
\end{aligned}$$

where the last line is obtained for n high enough (namely $n > 2^J(L[h]-1)$). Hence, since we constructed the vector $I^{(u,v)}_{k;n}$ by using the same injection that we used to construct the vector $\Phi_J(\tau)$ in 7, we obtain that

$$\begin{aligned}
\mathbb{E}(I^{(u,v)}_{k;n}) &= \sum_\tau \Phi_J(\tau) C^{(u,v)}_n(-\tau, z) \\
&= \sum_\tau \Phi_J(\tau)\Phi^\top_J(\tau) S^{(u,v)}_J(z) + \sum_\tau \Phi_J(\tau)(C^{(u,v)}_n(-\tau,z) - C^{(u,v)})(-\tau,z) \\
&= A_{q,J} S^{(u,v)}_J(z) + \sum_\tau \Phi_J(\tau)(C^{(u,v)}_n(-\tau,z) - C^{(u,v)}(-\tau,z)).
\end{aligned}$$

Finally, let ι represent a vector of one's in $\mathbb{R}^{J(2q+1)}$, we have that

$$\begin{aligned}
&\int_{[0,1]} \| \sum_\tau \Phi_J(\tau)(C^{(u,v)}_n(-\tau,z) - C^{(u,v)}(-\tau,z))\|_2 dz \\
&\leq \int_{[0,1]} \| \sum_{\tau\in\mathbb{Z}} \iota |C^{(u,v)}(-\tau,z) - C^{(u,v)}_n(-\tau,z)|\|_2 dz \\
&\leq \|\iota \sum_{\tau\in\mathbb{Z}} \int_{[0,1]} |C^{(u,v)}(-\tau,z) - C^{(u,v)}_n(-\tau,z)| dz\|_1 \\
&\leq O(\frac{J(2q+1)}{n}).
\end{aligned}$$

where the first line is obtained by Cauchy-Schwartz inequality (on the sum for $\tau \in \mathbb{Z}$) and the fact that absolute value of the CCWF is bounded by one; the second line is obtained using $\|v\|_2 \leq \|v\|_1$ and applying Tonelli's theorem twice; and the fourth line is obtained by applying Corollary 2. □

5.3.2 Proof of Theorem 8

Recall that

$$\mathbb{V}\mathrm{ar}(I_{jlk;n}^{(u,v)}) = \mathbb{E}[I_{jlk;n}^{(u,v)}]^2 - \mathbb{E}(I_{jlk;n}^{(u,v)})^2$$

Observe that by Definition 6, the wavelet coefficients are jointly normally distributed. Since by Isserli's theorem we have that

$$\begin{aligned}\mathbb{E}[I_{jlk;n}^{(u,v)}]^2 &= 2\mathbb{E}(d_{jk;n}^{(u)} d_{l,k,n}^{(v)})^2 + \mathbb{E}(d_{jk;n}^{(u)} d_{jk;n}^{(u)})\mathbb{E}(d_{l,k,n}^{(v)} d_{l,k,n}^{(v)}) \\ &= 2\mathbb{E}(I_{jlk;n}^{(u,v)})^2 + \mathbb{E}(I_{j,j,k,n}^{(u,v)})\mathbb{E}(I_{l,l,k,n}^{(u,v)})\end{aligned}$$

we write

$$\mathbb{V}\mathrm{ar}(I_{jlk;n}^{(u,v)}) = \mathbb{E}(I_{jlk;n}^{(u,v)})^2 + \mathbb{E}(I_{j,j,k,n}^{(u,u)})\mathbb{E}(I_{l,l,k,n}^{(v,v)})$$

Hence by Equation (5.3.1) we obtain

$$\mathbb{V}\mathrm{ar}(I_{jlk;n}^{(u,v)}) = (\sum_{\tau} \Psi_{j,l} C_n^{(u,v)}(\tau, z))^2 + \sum_{\tau} \Psi_{j,j} C_n^{(u,u)}(\tau, z) \sum_{\tau} \Psi_{l,l} C_n^{(v,v)}(\tau, z).$$

Since

$$\sum_{\tau} \Psi_{j,l} \Phi_J(\tau) S_J^{(u,v)}(z) + \sum_{\tau} \Psi_{j,l} \big(C_n^{(u,v)}(\tau, z) - C(\tau, z)^{(u,v)} \big)$$

let $R(J, q, n, z, u, v) = \big(C_n^{(u,v)}(\tau, z) - C(\tau, z)^{(u,v)} \big)$, we then have

$$\begin{aligned}&\mathbb{V}\mathrm{ar}(I_{jlk;n}^{(u,v)}) = \\ &\Big(\sum_{\tau} \Psi_{j,l} \Phi_J(\tau) S_J^{(u,v)}(z) \Big)^2 + \big(\sum_{\tau} \Psi_{j,j} \Phi_J(\tau) S_{z,J}^{(u,u)} \big) \big(\sum_{\tau} \Psi_{l,l} \Phi_J(\tau) S_{z,J}^{(v,v)} \big) \\ &+ O\Big(R(J, q, n, z, u, v) \Big)\end{aligned}$$

□

5.3.3 Proof of Theorem 9

We have

$$\mathbb{E}(I_{k;nM}^{(u,v)}) = \frac{1}{2M+1} \sum_{s=-M}^{M} \mathbb{E}(I_{k+s,n}^{(u,v)})$$

then by Theorem 7 we have

$$\mathbb{E}(I_{k+s,n}^{(u,v)}) = A_{q,J} S_{z+s/n}^{(u,v)} + R(J, q, n, z, u, v)$$

where $z = k/n$. By adding zeros we obtain

$$\mathbb{E}(I^{(u,v)}_{k+s,n}) = A_{q,J}S^{(u,v)}_z + A_{q,J}\big(S^{(u,v)}_{z+s/n} - S^{(u,v)}_z\big) + R(J,q,n,z,u,v)$$

Since we assumed the EWS functions $S^{(u,v)}_{j,l}(z)$ to be Lipschitz continuous with Lipschitz constant $C^{(u,v)}_{j,l}$, we construct the vector

$$C^{(u,v)} = (C^{(u,v)}_{j,l}) =_{i,l=-1,\dots,-J;|l-j|\leq q}$$

by using the same homeomorphisme than in Definition 5 and obtain the following bound

$$A_{q,J}\big(S^{(u,v)}_{z+s/n} - S^{(u,v)}_z\big) = \frac{s}{n}A_{q,J}C^{(u,v)}.$$

Moreover since $A_{q,J}$ is an invertible matrix, it is bounded. Hence $sn^{-1}A_{q,J}C^{(u,v)} = O(s/n)$ and

$$\mathbb{E}(I^{(u,v)}_{k+s,n}) = A_{q,J}S^{(u,v)}_z + O(\frac{s}{n}) + R(J,q,n,z,u,v).$$

Consequently

$$\mathbb{E}(I^{(u,v)}_{k;nM}) = A_{q,J}S^{(u,v)}_z + O(\frac{M}{n}) + O\bigg(R(J,q,n,z,u,v)\bigg).\square$$

5.3.4 Proof of Theorem 10

Using that for any constant c, we have

$$\mathbb{V}\text{ar}(c\sum_{i=1}^{n} v_i) = c^2\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{C}\text{ov}(v_i,v_j)$$

We bound the variance of the smoothed wavelet periodogram by

$$\mathbb{V}\text{ar}(I^{(u,v)}_{jlk;nM}) \leq \frac{1}{(2M+1)^2}\sum_{s=-M}^{M}\sum_{\kappa}|\mathbb{C}\text{ov}(I^{(u,v)}_{jl,k+s;nM}, I^{(u,v)}_{jl,k+s+\kappa;nM})|.$$

By Isserli's theorem for higher moment of multivariate normal we have

$$\begin{aligned}
&\mathbb{C}\text{ov}(I^{(u,v)}_{jl,k+s;nM}, I^{(u,v)}_{jl,k+s+\kappa;nM}) \qquad (5.3.2)\\
&= \mathbb{E}(I^{(u,v)}_{jl,k+s;nM}I^{(u,v)}_{jl,k+s+\kappa;nM}) - \mathbb{E}(I^{(u,v)}_{jl,k+s;nM})\mathbb{E}(I^{(u,v)}_{jl,k+s+\kappa;nM})\\
&= \mathbb{E}(d^{(u)}_{j,k+s;n}d^{(u)}_{j,k+s+\kappa;n})\mathbb{E}(d^{(v)}_{l,k+s;n}d^{(v)}_{l,k+s+\kappa;n})\\
&\quad + \mathbb{E}(d^{(u)}_{j,k+s;n}d^{(v)}_{l,k+s+\kappa;n})\mathbb{E}(d^{(v)}_{l,k+s;n}d^{(u)}_{j,k+s+\kappa;n}).
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}(d^{(u)}_{j,k+s;n} d^{(v)}_{l,k+s+\kappa;n}) =& \mathbb{E}\big(\sum_t \psi_{j,k+s}(t) X_n(t)^{(u)} \sum_{t'} X_n^{(u)}(t') \psi_{l,k+s+\kappa}(t') \big) \\
&= \sum_t \sum_{t'} \psi_{k+s}(t) \psi_k(t-\tau) \mathbb{E}(X_n(t)^{(u)} X_n(t-\tau)^{(u)}) \\
&= \sum_t \sum_{\tau} \psi_{j,k+s+\kappa}(t) \psi_k(t-\tau) \mathbb{C}\mathrm{ov}(X_n(t)^{(u)} X_n(t-\tau)^{(u)}) \\
&= \sum_t \sum_{\tau} \psi_{j,k+s}(t) \psi_{k+s}(t-\tau-\kappa) C_n^{(u,v)}(\tau, z) \\
&= \sum_{\tau} \Psi_{j,l}(\tau+\kappa) C_n^{(u,v)}(\tau, z) \\
&= \sum_{\tau} \Psi_{j,l}(\tau+\kappa) \Phi_J(\tau) S_J^{(u,v)}(z) + r(J,q,n,z,u,v).
\end{aligned}$$

We observe that

$$\sum_{\tau} \Psi_{j,l}(\tau+\kappa) \Phi_J(\tau) S_J^{(u,v)}(z) = \sum_{r,i} \sum_{\tau} \Psi_{j,l}(\tau+\kappa) \Psi_{r,i}(\tau) S_{r,i}^{(u,v)}(z)$$

then since

$$\sum_{\tau} \Psi_{j,l}(\tau+\kappa) \Psi_{r,i}(\tau) \leq 4 \min\{L_j, L_l, L_r, L_i\} \mathbf{1}_{|\kappa| \leq \max\{L_j, L_l, L_r, L_i\}}$$

this bound follows from the facts that (1) the absolute value of the product of wavelets is bounded by one, (2) that the maximum support of a product of n wavelet is n times the length of the of the support of the wavelet with the greater support, and (3) that for a $|\kappa|$ too high relatively to the supports of the wavelets involved in the product, the all product goes to zero.

Hence we have the bound

$$\begin{aligned}
&\mathbb{E}(d^{(u)}_{j,k+s;n} d^{(v)}_{l,k+s+\kappa;n}) \leq 4 \sum_{r,i} S_{r,i}^{(u,v)}(z) \min\{L_j, L_l, L_r, L_i\} \mathbf{1}_{|\kappa| \leq \max\{L_j, L_l, L_r, L_i\}} \\
&+ r(J,q,n,z,u,v)
\end{aligned}$$

Moreover, observe that by changing variable, we can write

$$\sum_{r,i=-1,-\ldots,-J; |r-i| \leq q} S_{r,i}^{(u,v)}(z) = \sum_{r=-J}^{-1} \sum_{i=-q}^{q} S_{r,r+i}^{(u,v)}.$$

which, combine with the above bound provide

$$\begin{aligned}
&\mathbb{E}(d^{(u)}_{j,k+s;n} d^{(v)}_{l,k+s+\kappa;n}) \leq 4 \mathbf{1}_{|\kappa| \leq \max\{L_j, L_l\}} \sum_{r=-\max(j,l)}^{-1} \sum_{i=-q}^{q} S_{r,i}^{(u,v)}(z) \min\{L_r, L_{i+r}\} \\
&+ 4 \min\{L_j, L_l\} \sum_{r'=-J}^{-\max(j,l)-1} \sum_{i'=-q}^{q} S_{r',i'}^{(u,v)}(z) \mathbf{1}_{|\kappa| \leq \max\{L_r, L_{i+r}\}} \\
&+ r(J,q,n,z,u,v).
\end{aligned}$$

This means that we can bound $|\mathbb{E}(d^{(u)}_{j,k+s;n} d^{(v)}_{l,k+s+\kappa;n})\mathbb{E}(d^{(v)}_{l,k+s;n} d^{(u)}_{j,k+s+\kappa;n})|$ in Equation (5.3.2) by

$$
\begin{aligned}
&16\mathbf{1}_{|\kappa|\leq\max\{L_j,L_l\}} \sum_{r=-\max(j,l)}^{-1} S^{(u,v)}_{r,i}(z) \sum_{i=-q}^{q} S^{(u,v)}_{r,i}(z) \min\{L_r, L_{i+r}\} \\
&\sum_{r'=-j}^{-1} \sum_{i'=-q}^{q} S^{(u,v)}_{r',i'}(z) \min\{L'_r, L_{i'+r'}\} \\
&+ 16\mathbf{1}_{|\kappa|\leq\max\{L_j,L_l\}} \min\{L_j, L_l\} \sum_{r=-\max(j,l)}^{-1} \sum_{i=-q}^{q} S^{(u,v)}_{r,i}(z) \min\{L_r, L_{i+r}\} \\
&\sum_{r'=-J}^{-\max(j,l)-1} \sum_{i'=-q}^{q} S^{(u,v)}_{r',i'}(z)\mathbf{1}_{|\kappa|\leq\max\{L_r,L_{i+r}\}} \\
&+ 4\mathbf{1}_{|\kappa|\leq\max\{L_j,L_l\}} \sum_{r=-\max(j,l)}^{-1} \sum_{i=-q}^{q} \min\{L_r, L_{i+r}\} r(J,q,n,z,v,u) \\
&+ 16\min\{L_j, L_l\}\mathbf{1}_{|\kappa|\leq\max\{L_j,L_l\}} \sum_{r'=-J}^{-\max(j,l)-1} \sum_{i'=-q}^{q} S^{(u,v)}_{r',i'}(z)\mathbf{1}_{|\kappa|\leq\max\{L_r,L_{i+r}\}} \\
&\sum_{r=-\max(j,l)}^{-1} \sum_{i=-q}^{q} S^{(u,v)}_{r,i}(z) \min\{L_r, L_{i+r}\} \\
&+ 16\min\{L_j, L_l\}^2 \sum_{r=-J}^{-\max(j,l)-1} \sum_{i=-q}^{q} S^{(u,v)}_{r',i'}(z) \\
&\sum_{r'=-J}^{-\max(j,l)-1} \sum_{i'=-q}^{q} S^{(u,v)}_{r',i'}(z)\mathbf{1}_{|\kappa|\leq\max\{L_r,L_{i+r},L'_r,L_{i'+r'}\}} \\
&+ 4\min\{L_j, L_l\} \sum_{r'=-J}^{-\max(j,l)-1} \sum_{i'=-q}^{q} S^{(u,v)}_{r',i'}(z)\mathbf{1}_{|\kappa|\leq\max\{L_r,L_{i+r}\}} r(J,q,n,z,v,u) \\
&+ 4r(J,q,n,z,u,v)\mathbf{1}_{|\kappa|\leq\max\{L_j,L_l\}} \sum_{r=-\max(j,l)}^{-1} \sum_{i=-q}^{q} \min\{L_r, L_{i+r}\} S^{(u,v)}_{r,i}(z) \\
&+ 4r(J,q,n,z,u,v)\min\{L_j, L_l\} \sum_{r'=-J}^{-\max(j,l)-1} \sum_{i'=-q}^{q} S^{(u,v)}_{r',i'}(z)\mathbf{1}_{|\kappa|\leq\max\{L_r,L_{i+r}\}} \\
&+ r(J,q,n,z,u,v) r(J,q,n,z,v,u)
\end{aligned}
$$

which simplifies in

$$
\begin{aligned}
&16\mathbf{1}_{|\kappa|\le\max\{L_j,L_l\}}\sum_{r=-\max(j,l)}^{-1}S_{r,i}^{(u,v)}(z)\sum_{i=-q}^{q}S_{r,i}^{(u,v)}(z)\min\{L_r,L_{i+r}\}\\
&\sum_{r'=-j}^{-1}\sum_{i'=-q}^{q}S_{r',i'}^{(u,v)}(z)\min\{L'_r,L_{i'+r'}\}\\
&+16\mathbf{1}_{|\kappa|\le\max\{L_j,L_l\}}\min\{L_j,L_l\}\sum_{r=-\max(j,l)}^{-1}\sum_{i=-q}^{q}S_{r,i}^{(u,v)}(z)\min\{L_r,L_{i+r}\}\\
&\sum_{r'=-J}^{-\max(j,l)-1}\sum_{i'=-q}^{q}S_{r',i'}^{(u,v)}(z)\\
&+4\mathbf{1}_{|\kappa|\le\max\{L_j,L_l\}}\sum_{r=-\max(j,l)}^{-1}\sum_{i=-q}^{q}\min\{L_r,L_{i+r}\}r(J,q,n,z,v,u)\\
&+16\min\{L_j,L_l\}\mathbf{1}_{|\kappa|\le\max\{L_j,L_l\}}\sum_{r'=-J}^{-\max(j,l)-1}\sum_{i'=-q}^{q}S_{r',i'}^{(u,v)}(z)\\
&\sum_{r=-\max(j,l)}^{-1}\sum_{i=-q}^{q}S_{r,i}^{(u,v)}(z)\min\{L_r,L_{i+r}\}\\
&+16\min\{L_j,L_l\}^2\sum_{r=-J}^{-\max(j,l)-1}\sum_{i=-q}^{q}S_{r',i'}^{(u,v)}(z)\\
&\sum_{r'=-J}^{-\max(j,l)-1}\sum_{i'=-q}^{q}S_{r',i'}^{(u,v)}(z)\mathbf{1}_{|\kappa|\le\max\{L_r,L_{i+r},L'_r,L_{i'+r'}\}}\\
&+4\min\{L_j,L_l\}\sum_{r'=-J}^{-\max(j,l)-1}\sum_{i'=-q}^{q}S_{r',i'}^{(u,v)}(z)\mathbf{1}_{|\kappa|\le\max\{L_r,L_{i+r}\}}r(J,q,n,z,v,u)\\
&+4r(J,q,n,z,u,v)\mathbf{1}_{|\kappa|\le\max\{L_j,L_l\}}\sum_{r=-\max(j,l)}^{-1}\sum_{i=-q}^{q}\min\{L_r,L_{i+r}\}S_{r,i}^{(u,v)}(z)\\
&+4r(J,q,n,z,u,v)\min\{L_j,L_l\}\sum_{r'=-J}^{-\max(j,l)-1}\sum_{i'=-q}^{q}S_{r',i'}^{(u,v)}(z)\mathbf{1}_{|\kappa|\le\max\{L_r,L_{i+r}\}}\\
&+r(J,q,n,z,u,v)r(J,q,n,z,v,u)
\end{aligned}
$$

and since $\sum_\kappa \mathbf{1}_{|\kappa|\le L_\alpha}=2L_\alpha$, summing the above quantity over κ leads to

$$
\begin{aligned}
B_{j,l,q,J} &= 32\max\{L_j, L_l\} \sum_{r=\max(j,l)}^{-1} \sum_{i=-q}^{q} S_{r,i}^{(u,v)}(z) \min\{L_r, L_{i+r}\} \\
&\sum_{r'=-\max(j,l)}^{-1} \sum_{i'=-q}^{q} S_{r',i'}^{(u,v)}(z) \min\{L'_r, L_{i'+r'}\} \\
&+ 32\max\{L_j, L_l\} \min\{L_j, L_l\} \sum_{r=-\max(j,l)}^{-1} \sum_{i=-q}^{q} S_{r,i}^{(u,v)}(z) \min\{L_r, L_{i+r}\} \\
&\sum_{r'=-J}^{-\max(j,l)-1} \sum_{i'=-q}^{q} S_{r',i'}^{(u,v)}(z) \\
&+ 8\max\{L_j, L_l\} \sum_{r=-\max(j,l)}^{-1} \sum_{i=-q}^{q} S_{r,i}^{(u,v)}(z) \min\{L_r, L_{i+r}\} r(J,q,n,z,v,u) \\
&+ 32\min\{L_j, L_l\} \max\{L_j, L_l\} \sum_{r'=-J}^{-\max(j,l)-1} \sum_{i'=-q}^{q} S_{r',i'}^{(u,v)}(z) \\
&\sum_{r=-\max(j,l)}^{-1} \sum_{i=-q}^{q} S_{r,i}^{(u,v)}(z) \min\{L_r, L_{i+r}\} \\
&+ 32\min\{L_j, L_l\}^2 \sum_{r=-J}^{-\max(j,l)-1} \sum_{i=-q}^{q} \sum_{r'=-J}^{-\max(j,l)-1} \sum_{i'=-q}^{q} \\
&S_{r,i}^{(u,v)}(z) S_{r',i'}^{(u,v)}(z) \min\left((\max\{L_r, L_{i+r}), \max(L'_r, L_{i'+r'}\}\right)\} \\
&+ 8\min\{L_j, L_l\} \\
&\sum_{r'=-J}^{-\max(j,l)-1} \sum_{i'=-q}^{q} S_{r',i'}^{(u,v)}(z) \max\{L_r, L_{i+r}\} r(J,q,n,z,v,u) \\
&+ 8r(J,q,n,z,u,v) \max\{L_j, L_l\} \sum_{r=-\max(j,l)}^{-1} \sum_{i=-q}^{q} S_{r,i}^{(u,v)}(z) \min\{L_r, L_{i+r}\} \\
&+ 8r(J,q,n,z,u,v) \min\{L_j, L_l\} \sum_{r'=-J}^{-\max(j,l)-1} \sum_{i'=-q}^{q} S_{r',i'}^{(u,v)}(z) \max\{L_r, L_{i+r}\} \\
&+ \sum_{\kappa} r(J,q,n,z,u,v) r(J,q,n,z,v,u)
\end{aligned}
$$

Remark 5 *By assumption we have that the EWS functions are such that for*

any $\alpha, \beta = 1, ..., J$, $|S^{(u,v)}_{\alpha,\beta}(z)| \leq \frac{K}{L_{-\min(\alpha,\beta)}}$ *where* $L_{\alpha-1} = 2L_\alpha$, *then we have*

$$\sum_{l=\max(j,l)}^{-1} \sum_{r=-q}^{q} |S_{r,l}(z)^{(u,v)}| \leq \frac{K}{2}(2q+1)\min(-j,-l)$$

Following the same line of though we have that the well-known formula on geometric series provides

$$\sum_{l=-J}^{\max(j,l)-1} \sum_{r=-q}^{q} |S_{r,l}(z)^{(u,v)}| \leq \frac{2K(2q+1)}{\min(L_j, L_l)}.$$

Hence by Remark 5 we have

$$\sum_{r=-\max(j,l)}^{-1} \sum_{i=-q}^{q} S^{(u,v)}_{r,i}(z)\min\{L_r, L_{i+r}\} \leq \frac{K}{2}(2q+1)\min(-j,-l)\max\{L_j, L_l\}$$

And consequently

$$\begin{aligned}
&B_{j,l,q,J} \leq 8\max\{L_j, L_l\}^3 K^2(2q+1)^2\min(-j,-l)^2 \\
&+ 64\max\{L_j, L_l\}^2 K^2(2q+1)^2\min(-j,-l) \\
&+ 16\max\{L_j, L_l\}^2 \frac{K}{2}(2q+1)\min(-j,-l) r(J,q,n,z,v,u) \\
&+ 32\min\{L_j, L_l\}^2 \sum_{r=-J}^{-\max(j,l)-1} \sum_{i=-q}^{q} \sum_{r'=-J}^{-\max(j,l)-1} \sum_{i'=-q}^{q} \\
&S^{(u,v)}_{r,i}(z) S^{(u,v)}_{r',i'}(z) \min\left((\max\{L_r, L_{i+r}), \max(L'_r, L_{i'+r'}\}\right)\} \\
&+ 16\min\{L_j, L_l\} \sum_{r'=-J}^{-\max(j,l)-1} \sum_{i'=-q}^{q} S^{(u,v)}_{r',i'}(z)\max\{L_r, L_{i+r}\} r(J,q,n,z,v,u) \\
&+ (2M+1) r(J,q,n,z,u,v) r(J,q,n,z,v,u)
\end{aligned}$$

We observe moreover we can write

$$\sum_{r=-J}^{-\max(j,l)-1} \sum_{i=-q}^{q} \sum_{r'=-J}^{-\max(j,l)-1} \sum_{i'=-q}^{q} S^{(u,v)}_{r,i}(z) S^{(u,v)}_{r',i'}(z) \min\left((\max\{L_r, L_{i+r}), \max(L'_r, L_{i'+r'}\}\right)$$

as

$$\sum_{r=-J}^{-\max(j,l)-1}\sum_{i=-q}^{q} S_{r,i}^{(u,v)}(z)\max\{L_r, L_{i+r}\}\sum_{l'\geq l}\sum_{i=-q}^{q} S_{r',i'}^{(u,v)}(z)$$
$$+\sum_{r'=-J}^{-\max(j,l)-1}\sum_{i'=-q}^{q} S_{r,i}^{(u,v)}(z)\max\{L'_r, L_{i'+r'}\}\sum_{l\geq l'}\sum_{i=-q}^{q} S_{r,i}^{(u,v)}(z)$$

which, by Remark 2 can be bounded above by

$$\sum_{r=-J}^{-\max(j,l)-1}\sum_{i=-q}^{q} S_{r,i}^{(u,v)}(z)\max\{L_r, L_{i+r}\}\sum_{l'\geq l}\sum_{i=-q}^{q}\frac{K}{L_{\min(r',l')}}$$
$$+\sum_{r'=-J}^{-\max(j,l)-1}\sum_{i'=-q}^{q} S_{r,i}^{(u,v)}(z)\max\{L'_r, L_{i'+r'}\}\sum_{l\geq l'}\sum_{i=-q}^{q}\frac{K}{L_{\min(r,l)}}$$

which can be bounded above by

$$\sum_{r=-J}^{-\max(j,l)-1}\sum_{i=-q}^{q} S_{r,i}^{(u,v)}(z)\sum_{l'\geq l}\sum_{i=-q}^{q} K$$
$$+\sum_{r'=-J}^{-\max(j,l)-1}\sum_{i'=-q}^{q} S_{r,i}^{(u,v)}(z)\sum_{l\geq l'}\sum_{i=-q}^{q} K$$

which, by Remark 2 can be bounded above by

$$K(2q+1)\left(\sum_{r=-J}^{-\max(j,l)-1}\sum_{i=-q}^{q} S_{r,i}^{(u,v)}(z)+\sum_{r'=-J}^{-\max(j,l)-1}\sum_{i'=-q}^{q} S_{r,i}^{(u,v)}(z)\right)\leq K^2(2q+1)^2.$$

Consequently we have

$$\begin{aligned}
B_{j,l,q,J} &\leq 8\max\{L_j, L_l\}^3K^2(2q+1)^2\min(-j,-l)^2\\
&+64\max\{L_j, L_l\}^2K^2(2q+1)^2\min(-j,-l)\\
&+16\max\{L_j, L_l\}^2\frac{K}{2}(2q+1)\min(-j,-l)r(J,q,n,z,v,u)\\
&+32\min\{L_j, L_l\}^2K^2(2q+1)^2\\
&+16\min\{L_j, L_l\}\sum_{r'=-J}^{-\max(j,l)-1}\sum_{i'=-q}^{q} S_{r',i'}^{(u,v)}(z)\max\{L_r, L_{i+r}\}r(J,q,n,z,v,u)\\
&+(2M+1)r(J,q,n,z,u,v)^2
\end{aligned}$$

Hence we have shown that

$$
\begin{aligned}
B_{j,l,q,J} &= O\Big(\max\{L_j, L_l\}^3 q^2 \min(-j,-l)^2 \Big) + \\
&+ O\Big(\max\{L_j, L_l\}^2 q \min(-j,-l) r(J,q,n,z,v,u) \Big) \\
&+ 16 \min\{L_j, L_l\} \sum_{r'=-J}^{-\max(j,l)-1} \sum_{i'=-q}^{q} S^{(u,v)}_{r',i'}(z) \max\{L_r, L_{i+r}\} r(J,q,n,z,v,u) \\
&+ O(M r(J,q,n,z,u,v)^2)
\end{aligned}
$$

Moreover, since we have that

$$
\sum_{r'=-J}^{-\max(j,l)-1} \sum_{i'=-q}^{q} S^{(u,v)}_{r',i'}(z) \max\{L_r, L_{i+r}\} \leq \frac{L_J}{\min(L_j, L_l)} 2K(q+1)
$$

we have

$$
\begin{aligned}
B_{j,l,q,J} &= O\Big(\max\{L_j, L_l\}^3 (2q+1)^2 \min(-j,-l)^2 \Big) + \\
&+ O\Big(\max\{L_j, L_l\}^2 (2q+1) \min(-j,-l) r(J,q,n,z,v,u) \Big) \\
&+ O\Big(\frac{L_J}{\min(L_j, L_l)} 2K(q+1) r(J,q,n,z,v,u) \Big) \\
&+ O(M r(J,q,n,z,u,v)^2)
\end{aligned}
$$

and since for any $\alpha = -1, ..., -J$ we have $L_\alpha = O(2^{-\alpha})$ we have shown that

$$
\begin{aligned}
B_{j,l,q,J} &= O\Big(2^{-\min\{j,l\}} q^2 \min(-j,-l)^2 \Big) + \\
&+ O\Big(2^{-\min\{j,l\}} q \min(-j,-l) r(J,q,n,z,v,u) \Big) \\
&+ O\Big(2^{J-\min\{j,l\}} q r(J,q,n,z,v,u) \Big) \\
&+ O(M r(J,q,n,z,u,v)^2)
\end{aligned}
$$

and since $M \leq (n+1)/2$ and then by Proposition 9 in the previous Chapter, we have

$$B_{j,l,q,J} = O\left(2^{-\min\{j,l\}} q^2 \min(-j,-l)^2\right) +$$
$$+ O\left(2^{-\min\{j,l\}} q \min(-j,-l) r(J,q,n,z,v,u)\right)$$
$$+ O\left(2^{J+\min\{j,l\}} q r(J,q,n,z,v,u)\right)$$

We note that the same bound can be derive for

$$|\mathbb{E}(d^{(u)}_{j,k+s;n} d^{(u)}_{l,k+s+\kappa;n}) \mathbb{E}(d^{(v)}_{l,k+s;n} d^{(u)}_{j,k+s+\kappa;n})|$$

in Equation (5.3.2).

In summary we have

$$\mathbb{V}\mathrm{ar}(I^{(u,v)}_{ijk;nM}) \leq \frac{1}{(2M+1)^2}(2M+1) B_{i,j,q,J} \tag{5.3.3}$$
$$= \frac{B_{i,j,q,J}}{(2M+1)}. \tag{5.3.4}$$

□

5.3.5 Proof of Theorem 11

We have

$$\mathbb{V}\mathrm{ar}(\tilde{I}_{jlk;nM}) \leq \sum_{r,i=-1,-\ldots,-J;|r-i|\leq q} |A^{-1}_{j,l,r,i}(q,J)| \mathbb{V}\mathrm{ar}(I_{r,i,n,M})^{1/2}$$
$$\sum_{r',i'=-1,-\ldots,-J;|r-i|\leq q} |A^{-1}_{j,l,r',i'}(q,J)| \mathbb{V}\mathrm{ar}(I_{r',i',n,M})^{1/2}.$$

From the proof of Theorem 10, we may write

$$\mathbb{V}\mathrm{ar}(\tilde{I}_{jlk;nM}) \leq \sum_{r,i=-1,-\ldots,-J;|r-i|\leq q} |A^{-1}_{j,l,r,i}(q,J)| B_{r,i,q,J}$$
$$\sum_{r',i'=-1,-\ldots,-J;|r-i|\leq q} |A^{-1}_{j,l,r',i'}(q,J)| B^{1/2}_{r',i',q,J}.$$

and since we assumed that $\sum_{r,i=-1,-\ldots,-J;|r-i|\leq q} |A^{-1}_{j,l,r,i}(q,J)| \leq C L_{\min(l,j)}$, then the result follows since $L_{\min(l,j)} = O(2^{\min(l,j)})$. □

5.3.6 Proof of Theorem 12

Let $V(z) = I_{n,M}(z) - S(z)$ and consider $\|V(z)^\top V(z)\|_{\max}$ where for any matrix $M \in \mathbb{R}^{l\times l}$ $\|M\|_{\max} = \max_{i,j\in\{1,...,l\}} M_{i,j}$. Since $\big(V(z)^\top V(z)\big) \in \mathbb{R}^{(p\log_2(n))\times(p\log_2(n))}$ we have

$$\|V(z)\|_2^2 \leq p\log_2(n)\|V(z)^\top V(z)\|_{\max}.$$

Observe that each element of $(V(z)^\top V(z))$ correspond a unique $(u, j, u', j') \in \big(\{1, ..., p\}\times\{-1, ..., -J\}\times\{1, ..., p\}\{\times -1, ..., -J\}\big)$. Hence by Cauchy-Schwarz inequality we can bound each component of $\mathbb{E}(V(z)^\top V(z))$ as

$$\begin{aligned}
&\sum_i \sum_v \mathbb{E}\bigg((\tilde{I}^{(u,v)}_{n,M,i,j} - S^{(u,v)}_{i,j})(\tilde{I}^{(u',v)}_{n,M,i,j'} - S^{(u',v)}_{i,j'})\bigg) \\
&\leq \sum_i \sum_v \bigg(\mathbb{E}[\tilde{I}^{(u,v)}_{n,M,i,j} - S^{(u,v)}_{i,j}]^2 \mathbb{E}[(\tilde{I}^{(u',v)}_{n,M,i,j} - S^{(u,v)}_{i,j})]^2\bigg)^{1/2}
\end{aligned}$$

where

$$\mathbb{E}[\tilde{I}^{(u',v)}_{n,M,i,j} - S^{(u,v)}_{i,j}]^2 = \mathbb{E}(\tilde{I}^{(u',v)}_{n,M,i,j} - S^{(u,v)}_{i,j})^2 + \mathbb{V}\mathrm{ar}(\tilde{I}^{(u',v)}_{n,M,i,j}).$$

Hence by Theorem 11, we have that

$$\mathbb{E}[\tilde{I}^{(u',v)}_{n,M,i,j} - S^{(u,v)}_{i,j}]^2 = O(\frac{M^2\log_2^2(n)}{n^2}) + O(\frac{\log_2(n)}{M})$$

which implies

$$\|V(z)\|_{\max} = O\bigg(p\log_2^2(n)(\frac{M^2}{n^2} + \frac{1}{M\log_2(n)}\bigg)$$

and hence

$$\|V(z)\|_2^2 = O\bigg(p^2\log_2^4(n)(\frac{M^2}{n^2} + \frac{1}{M\log_2(n)}\bigg)$$

which is the result we intended to prove. □

Six

From double asymptotics to functional regression

Abstract

First, during the production of this thesis, questions arise concerning the kind of models one use to model high-dimensionality. As we discuss in the introduction of this chapter, there are two main approaches. Hence we questioned the links between the two approaches.

Second, in the approach of high-dimensionality considered in this thesis, one conducts asymptotic reasoning with both the number of observations and the number of covariates diverging. As we touch briefly in the second section of Chapter 2, this double asymptotic regime implies that objects of interest (such as the parameter of interest), change when one let the number of observation increase. Hence we asked: what does the *consistency* of an estimator mean in this context?

Third, at the very end of the production of this thesis, we came to understand (better late than never) that statistical analysis does not require the data (the observables) to be *random variables*. Actually, the data can be considered as mere *number tables* without affecting our ability to statistically analyze them. It is possible to do so using the concept of *statistical experiments*. In return, the mathematical analysis of estimators, test statistics,... and their qualities, may be conducted in the abstract structure of statistical experiments. In this paradigm, one cannot rely on a *true* probability to measure events. Hence we asked: under which probability measure is it possible to state that an abstract estimator converges to a parameter?

The present chapter connects these three questions. Providing early results to answer them, it opens the path for further research.

6.1 Introduction

There are two main types of statistical experiments that encompass high-dimensionality.

The first one, that is the one we considered so far in this thesis, supposes the number of covariates, p, to be a function of the number of observations, n (*e.g.* Stock and Watson (2002)). More formally, in this setting, one considers a sequence of data $\{X_n \in \mathbb{R}^{n\times p(n)}\}_{n\in\mathbb{N}}$ and a sequence of statistical experiments $\{\mathcal{E}_n\}_{n\in\mathbb{N}}$ that if the observations are supposed *iid*, may be written as

$$\{\mathcal{E}_n\}_{n\in\mathbb{N}} = \{\mathbb{R}^{n\times p(n)}, \overbrace{\sigma(\mathbb{R}^{p(n)}) \times ... \times \sigma(\mathbb{R}^{p(n)})}^{n \text{ times}}, \left(\mathbb{P}_{\theta_n}\right)^n | \theta_n \in \Theta_n\}_{n\in\mathbb{N}} \quad (6.1.1)$$

where $\sigma(\mathbb{R}^{p(n)})$ is the sigma-algebra generated by $\mathbb{R}^{p(n)}$, $\mathbb{P}_{\theta_n}$ is a familly of probability law on $\mathbb{R}^{p(n)}$ with parameter θ_n and Θ_n is the parameter space.

In a second setting high dimensionality is encompassed by letting the dimensionality of each observation to be infinite in all the statistical experiments. In other words, in this second setting, it is considered that the observables are vectors of infinite dimension or functions. This is for instance the case when one assumes the observables are *iid* $L^2(0,1)-$valued random functions as in functional regressions (e.g. Florens and Van Bellegem (2015)).

A simple functional regression may be written as

$$\begin{aligned} \mathbb{E}_\theta(y|x) &= \langle x, \beta\rangle_{L^2} \\ &= \int_{(0,1)} x(\xi)\beta(\xi)d\xi \end{aligned} \quad (6.1.2)$$

where β is a non-random function in $L^2(0,1)$ and x is a random function valued in $L^2(0,1)$. In this case, the sequence of statistical experiments may be written as

$$\{\mathcal{E}_n\}_{n\in\mathbb{N}} = \{\mathbb{R}^{n\times L^2(0,1)}, \overbrace{\sigma\big(L^2(0,1)\big) \times ... \times \sigma\big(L^2(0,1)\big)}^{n \text{ times}}, \left(\mathbb{P}_{\theta}\right)^n | \theta \in \Theta\}_{n\in\mathbb{N}}$$

In this above approach, which we call the *infinite dimensional* setting, one works with functional spaces instead of working with finite dimensional vector spaces. However, besides the technical difficulties coming from this fact, the infinite-dimensional setting remains an instance of the standard *single asymptotic* statistical experiment in which the dimensionality of the covariates is fixed, and only the number of observations grows along the sequence of statistical experiments. Consequently, as in any sequence of statistical experiments in single asymptotic that suppose that the observations are *iid*, the parameter space Θ as well as the parameter θ *do not need to depend on* n[1]. Observe that this quality is lost in the double-asymptotic case described in Equation (6.1.1). There if one focuses on, for instance, the mean, say $\mu_{\theta_n} \in \mathbb{R}^{p(n)}$, one observes

[1] In the context of heteroscedasticity θ may be a function of n, however, the parameter space Θ does *not need* to depend on n.

μ_{θ_n} *must* changes for every n, *even when the observations are supposed to be* iid.

One consequence of this fact is that in the double asymptotic approach (when both p and n diverge), asymptotic reasoning is different than in the standard single asymptotic regime (when only n diverges). For instance the usual notion of *consistency* of an estimator (in the sense of convergence in probability to the parameter of interest) needs to be modified.

Commonly, that is in the single asymptotic setting, one states that an estimator is *consistent* if

> "The sequence of estimators converges in probability to the true value of the parameter." (page 3 in van der Vaart (2000))

This concept of consistency, which is essentially the same as that in Amemiya (1985), suppose that the data $\{X_n \in \mathbb{R}^{n\times p}\}_{n\in\mathbb{N}}$ are, in actuality, *random functions* with domain in a *true* probability space, whose probability measure determines (induce) the parameter of the probability law of the observables. However, in the double asymptotic regime, the (*assumed* and unknown) data generating process is such that, without strong prior, the probability law of the (*iid*) covariates *must* change at every n (every new (set of) observation). Hence it becomes unclear what one means by the *true* (value) of *the* parameter.

In the context of (single asymptotic) statistical experiments, the above concept of consistency may be refined by stating that, $h_n : \mathbb{R}^{n\times p} \to \Theta$ is a *consistent* estimator of an identified parameter, say $g(\theta)$ for g some function Θ, if, *for every* $\theta \in \Theta$, and every $\delta, \epsilon > 0$ there exists $N \in \mathbb{N}$ such that for every $n > N$

$$\mathbb{P}_\theta(\|h_n - g(\theta)\| > \delta) < \epsilon$$

see *e.g.* Definition 8.1, page 54 in Lehmann and Casella (1998). In the double asymptotic regime however, the parameter space is a function of n and the definition needs to be interpreted (and modified accordingly) *a minima* with care for this fact.

For instance, one may state that a sequence of estimators say, $h_n : \mathbb{R}^{n\times p(n)} \to \Theta_n$ is consistent for an identified *sequence* of parameters of interest, say $g_n(\theta_n)$ (for a sequence of functions g_n on $\{\Theta_n\}_{n\in\mathbb{N}}$), if for every $\delta, \epsilon > 0$, there exists (an absolute) $N \in \mathbb{N}$ such that for every $n > N$

$$\sup_{\theta_n \in \Theta_n} \mathbb{P}_{\theta_n}(\|h_n - g_n(\theta_n)\| > \delta) < \epsilon.$$

This notion is stronger than the one given by Lehmann and Casella (1998) in the context of single-asymptotics, since, in the above definition, the convergences in probability need to happen *uniformly* over the whole sequence $\Omega_{N+1}, \Omega_{N+2}, ...$ where in Lehmann and Casella (1998) the definition of consistency only requires convergence in probability *point-wise* in θ, for every

$\theta \in \Theta$. In the third section of this chapter, we study notions of very weak consistency, weak consistency, and strong consistency in sequences of statistical experiments both in the single and double asymptotic regimes.

A second feature of the type of asymptotic reasoning conducted in the double-asymptotic setting comes from the fact that, as we mentioned above, the parameter of interest is actually a *sequence* of parameters of interest. However, in general, the *limit* of this sequence is unspecified and actually, does not need to exist (nor does the limit of its estimators). For instance, consider the parameter of interest is the inverse of the covariance of the covariates. At the limit, this parameter is generally not well defined (unbounded). Putting aside (for now) that this raises the question of the usual definition of the *almost-sure* convergence in the double asymptotic regime, our emphasis on the undetermination of the limit of the parameter of interest come from the fact that we may not be ready (yet) to abandon completely the idea of *true* parameter of (*true*) probability law of the observables (as in aforementioned Amemiya (1985), van der Vaart (2000)). If we inverse the heuristic in the above quote of van der Vaart (2000), the *true parameter* is the one to which converges a consistent estimator. Hence it is the limit of the sequence of parameters of interest. Consequently, if it exists, this limit, the *true parameter* of interest, can only live in a space of infinite dimension. Hence defining the true parameter in the double-asymptotic regime calls to define a limit space of infinite dimension. Or, by a continuity argument, it calls to connect conceptually the double-asymptotic regime with the infinite-dimensional regime.

We believe there are at least two routes toward this unification of the two regimes. Firstly going from the double asymptotic and studying its limit by finding constraints on the sequences of objects (parameters, estimators, $\sigma-$algebras, probabilities,...), ensuring some desired properties (such as existence) at their limit. Or secondly, starting with the infinite-dimensional spaces, defining objects in this space and then studying their projections in a sequence of finite dimensional spaces of increasing dimensions.

Here we provide early results to tackle the second case. Those results are derived to answer a simple question concerning the estimation of the functional parameters in a functional regression defined in L^2. Interestingly this question is directly linked to the problem of the connection between the double asymptotic approach of high dimensions and the infinite-dimensional approach.

To formulate the said question, we consider the methodology employed by Florens and Van Bellegem (2015) to study functional regressions. There, everything happens in the infinite-dimensional space. That is where the parameter of interest, say β as in Equation (6.1.2), is defined and identified, so it is where the definition of its estimator, say $\hat{\beta}$, as well as the proof of its consistency, takes place.

However, in practice, that is when one implements the functional estimators $\hat{\beta}$, one typically does not work with functional objects (here and thereafter

in the sense of *continuous* functions on $(0,1)$). This is *a minima* because *digital computers* cannot store in memory such objects (even if it is possible to memorize the parameters on which depends a function, the image of the functions itself cannot be stored), and *a fortiori* applying mathematical operations to them. Furthermore, even if *analog computers* require functional input/output, their current development and deployment are far from allowing many (if not any) researchers to compute functional estimators such as $\hat{\beta}$.

What we actually *do* observe and work with are *discretizations* of such an x, say x_d. Moreover, covariates of the type of x_d are not only *discrete*, they are also *finite dimensional* (because storing an infinite dimensional vector, and not a way to construct this vector, in a computer would require an infinite amount of memory, and hence of energy). From those *finite dimensional* observations of x_d one can then compute a finite-dimensional estimator $\hat{\beta}_d$. But then the following question arises: how good is this estimator? The answer to this question starts by recognizing that the dimension of $\hat{\beta}_d$, say p, is determined by the discretization of x. Hence, the convergence in probability of $\hat{\beta}_d$ (sequence of vectors) to β (functional) might be obtained by discretizing x on a finer and finer grid, that is by letting both p and n (the number of observation of x_d) diverge (given that $\mathbb{R}^\infty$ is dense in $\mathbb{R}^{(0,1)}$).

We observe that *this type of reasoning is typical of the double asymptotic regime*. Consequently, we ask the question: *could we not, instead of studying the estimation of β via estimators defined in L^2, study the estimation of β with estimators defined in a double an asymptotic approach*? Are there gains, in terms of convergence rate, to do so? Would it not make the functional regression more accessible to a wider audience?

The second section of this chapter provides early results to answer these questions. The third section analysis sequence of iid statistical experiments in the single and double asymptotic regimes to provide notions of very weak consistency, weak consistency, and strong consistency in sequences of statistical experiments.

6.2 From double asymptotic to functional regression

6.2.1 A simple theory for approximation of square-integrable functions

In this section, we provide a theory of approximation in L^2 based on simple functions. Our goal by doing so, and that is attained at the end of this section, is to be able to express the L^2 distance between any function in L^2 and any *continued vector* from $\mathbb{R}^p$ to L^2 (a notion that we made clear below) as the sum between the Euclidean distance between the discretized version of this

function and the vector, plus, the error of approximation of the function made by the discretization step.

We start by studying an operator of *continuation*, that is an operator taking as input a vector in $\mathbb{R}^p$, and that transforms it into a continuous function of $[0,1)$. From there we demonstrate that the adjoint of this operator operates a discretization).

Definition 9 *Let μ be the* Lebesgue measure *on $[0,1)$, we define the operator $D_p : \mathbb{R}^p \to ([0,1),\mu)$ by*

$$\big(D_p v\big)(u) := \sum_{m=1}^{p} v_m p \mathbf{1}(u)_{[(m-1)/p,m/p)}.$$

This operator is a *continuation* operator. We study some of its properties below. Before doing so we derive its adjoint that operates a discretization (that is, it takes a function in $L^2[0,1)$ and transforms it into a vector in $\mathbb{R}^p$. Before deriving the adjoint, we introduce the following inner product on $\mathbb{R}^p$.

Definition 10 *For any $p \in \mathbb{N}$, let $u, v \in \mathbb{R}^p$, then defined the following inner product on $\mathbb{R}^p$*

$$\langle u, v\rangle_{l_p} = \langle u\sqrt{p}, v\sqrt{p}\rangle_2.$$

Lemma 17 *Let $D_p^\star$ be the the adjoint operator of D_p, then $D_p^\star : ([0,1),\mu) \to (\mathbb{R}^p, \langle\cdot,\cdot\rangle_{l_p})$ is such that for all $w \in L^2([0,1),\mu)$*

$$(D_p^\star w)_m = \langle \mathbf{1}_{[(m-1)/p,m/p)}, w\rangle_{L^2}$$

where $(D_p^\star w)_m$ is the element on the m^{th} line of $C^\star w \in (\mathbb{R}^p, \langle\cdot,\cdot\rangle_{l_p})$.

Proof.

$$\begin{aligned}
\langle D_p v, w\rangle_{L^2} :&= \int_{[0,1)} \big(D_p v\big)(u) w(u) d\mu(u) \\
&= \int_{[0,1)} \sum_{m=1}^{p} v_m p \mathbf{1}(u)_{[(m-1)/p,m/p)} w(u) d\mu(u) \\
&= \sum_{m=1}^{p} v_m \int_{[0,1)} p \mathbf{1}(u)_{[(m-1)/p,m/p)} w(u) d\mu(u) \\
&= \sum_{m=1}^{p} v_m \langle p \mathbf{1}_{[(m-1)/p,m/p)}, w\rangle_{L^2} \\
&= \langle \sqrt{p} v, \sqrt{p} D_p^\star w\rangle_2 \\
&= \langle v, D_p^\star w\rangle_{l_p}
\end{aligned}$$

□

Hence, we do have that $D_p^\star$ is a discretization operator and it has a closed-form expression. In the following, we provide some easily obtainable properties of D_p and $D_p^\star$.

Lemma 18 *The following relationships hold true*

1. $D_p^\star = D_p^{-1}$ *on* $(\mathbb{R}^p, \langle\cdot,\cdot\rangle_{l_p})$
2. $\forall p < \infty, (D_p^\star)^{-1} \neq D_p$
3. $\forall v \in \mathbb{R}^p, \|D_p v\|_{L^2} = \|v\|_{l_p}$
4. $\forall v \in L^2[0,1), \|D_p^\star v\|_{l_p} = \|v\|_{L^2}$
5. $\lim_{p\to\infty} D_p = \lim_{p\to\infty}(D_p^\star)^{-1}$ *up to set of measure zero.*

Proof: 1. and 2. are obvious.

3. is obtained by

$$\|D_p v\|_{L^2}^2 := \langle D_p v, D_p v\rangle = \|v\|_{l_p}^2.$$

4. is obtained by

$$\begin{aligned}&\|D_p^\star v\|_{l_p}^2\\ &= \frac{1}{p}\sum_{m=1}^{p}\langle\sqrt{p}\mathbf{1}_{[(m-1)/p,m/p)}, v\rangle_{L^2}^2\\ &= \int v^2 d\mu = \|v\|_{L^2}\end{aligned}$$

To obtain 5. observe that because $\mathbb{R}^\infty$ is dense in $[0,1)\times\mathbb{R}$ (by this we mean that there are $\aleph_0$ elements in a vector of $\mathbb{R}^\infty$, hence that each element of any of these vectors may be associated with a rational number, that are dense in $[0,1)$) we have

$$\lim_{p\to\infty} D_p^\star f = f \quad \text{up to a set of measure zero.}$$

for any f in $L^2[0,1)$. Moreover, since simple functions are dense in L^2 we have that

$$\lim_{p\to\infty} D_p f = f \quad \text{up to a set of measure zero,}$$

Hence we have that

$$\lim_{p\to\infty} D_p = \lim_{p\to\infty} D_p^\star f = \lim_{p\to\infty} D_p f = f \quad \text{up to set of measure zero.}$$

for any $f \in L^2$. □

Now we give the following proposition that states that the distance between a function in L^2 and a continued vector from $\mathbb{R}^p$ may be written as the sum of the distance between the vector and the discretized function and the error of approximation.

Proposition 11 *Let $\beta \in L^2([0,1),\mu)$ and $\hat{\beta} \in (\mathbb{R}^p, \langle\cdot,\cdot\rangle_{l_p})$.*

$$\|\beta - D_p\hat{\beta}\|_{L^2}^2 = \|D_p^\star\beta - \hat{\beta}\|_{l_p}^2 + \|\beta - D_pD_p^\star\beta\|_{L^2}^2$$

Proof: By direct computation and rearrangement we obtain

$$\begin{aligned}
&\|\beta - D_p\hat{\beta}\|_{L^2}^2 \\
&= \|D_p\hat{\beta}\|_{L^2}^2 - 2\langle\hat{\beta}, D_p^\star\beta\rangle_{l_p} + \|\beta\|_{L^2}^2 \\
&= \|\hat{\beta}\|_{l_p}^2 - 2\langle\hat{\beta}, D_p^\star\beta\rangle_2 + \|\beta\|_{L^2}^2 \\
&= \|\hat{\beta}\|_{l_p}^2 - 2\langle\hat{\beta}, D_p^\star\beta\rangle_{l_p} + \|\beta - D_pD_p^\star\beta + D_pD_p^\star\beta\|_{L^2}^2 \\
&= \|\hat{\beta}\|_{l_p}^2 - 2\langle\hat{\beta}, D_p^\star\beta\rangle_{l_p} + \|\beta - D_pD_p^\star\beta\|_{L^2}^2 + 2\langle\beta - D_pD_p^\star\beta, D_pD_p^\star\beta\rangle_{L^2}^2 + \|D_pD_p^\star\beta\|_{L^2}^2 \\
&= \|\hat{\beta}\|_{l_p}^2 - 2\langle\hat{\beta}, D_p^\star\beta\rangle_{l_p} + \|\beta - D_pD_p^\star\beta\|_{L^2}^2 + 2\langle\beta, D_pD_p^\star\beta\rangle_{L^2} - \|D_pD_p^\star\beta\|_{L^2}^2 \\
&= \|\hat{\beta}\|_{l_p}^2 - 2\langle\hat{\beta}, D_p^\star\beta\rangle_{l_p} + \|\beta - D_pD_p^\star\beta\|_{L^2}^2 + 2\langle D_p^\star\beta, D_p^\star\beta\rangle_{l_p} - \|D_p^\star\beta\|_{L^2}^2 \\
&= \|\hat{\beta}\|_{l_p}^2 - 2\langle\hat{\beta}, D_p^\star\beta\rangle_{l_p} + \|\beta - D_pD_p^\star\beta\|_{L^2}^2 + \|D_p^\star\beta\|_{l_p}^2 \\
&= \|D_p^\star\beta - \hat{\beta}\|_{l_p}^2 + \|\beta - D_pD_p^\star\beta\|_{L^2}^2
\end{aligned}$$

□

6.2.2 Estimating a functional regression via a sequence of high-dimensional linear models

In this section, we consider a random function X that is in $L^2[0,1)$. As we argue in the introduction, even if X might be *observable*, it might not be *observed* because of the type of technology used to conduct or record the observations. Hence we consider also that for some $p \in \mathbb{N}$, we observe some discretized version of X that is $D_p^\star X$ (where $D_p^\star$ is as defined above).

In this context, we are interested in the linear projection of some random variable Y onto the space of linear functions of X that we assume to be the conditional expectation of Y given X. Nevertheless, because we do not have access to observations of X, we cannot directly estimate this projection. Instead, we estimate the orthogonal projection of Y onto the space of linear functions of $D_p^\star X$.

In the next result, which is the main result of this section, we quantify the L^2 error of this procedure of estimation of $\mathbb{E}(Y|X)$. Here we state the result and add some comments below.

Theorem 13 *Let $(\Sigma, \mathcal{A}, \mathbb{P})$ be the probability space of observable events. Let Y and X be two measurable functions defined by*

$$Y : \Sigma \to L^2(\mathbb{R})$$
$$X : \Sigma \to L^2[0,1).$$

Moreover, assume that there exists a unique[2] *$\beta \in L^2[0,1)$ such that*

$$\mathbb{E}(Y|X) = \langle X, \beta \rangle_{L^2}$$

and let $\Omega := \mathbb{E}(X\langle X, \cdot\rangle_{L^2})$ be the covariance operator of X. Finally for some $\lambda > 0$, let

$$\theta := \min_{b \in \mathbb{R}^p} \|D_p^\star \mathbb{E}(XY) - D_p^\star \Omega D_p\|_2^2 + \lambda \|b\|_{l_p}^2$$

then, for any $\hat{\theta} \in \mathbb{R}^p$

$$\mathbb{E}\|\beta - D_p\hat{\theta}\|_{L^2}^2 \leq \mathbb{E}\|\theta - \hat{\theta}\|_{l_p}^2 + O\left(\|\beta - D_p D_p^\star \beta\|_{L^2}^2\right) + O\big(g(\lambda)\big)$$

where $g(\lambda) = \lambda^2 \|(D_p^\star \Omega D_p + \lambda I)^{-1}\|_{l_p}^2$

In the above theorem, β is the parameter of interest, and $D_p\hat{\theta}$ is its estimator in L^2. This estimator is the *continued* version of $\hat{\theta}$ that is an estimator of θ, which itself acts as a proxy target for β. The mean squared error of the estimator, $\|\beta - D_p\hat{\theta}\|_{L^2}^2$, is bounded by the sum of three terms

$$\underbrace{\mathbb{E}\|\theta - \hat{\theta}\|_{l_p}^2}_{\text{error of estimation}} + \underbrace{O\left(\|\beta - D_p D_p^\star \beta\|_{L^2}^2\right)}_{\text{error of approximation}} + \underbrace{O\big(g(\lambda)\big)}_{\text{biais of regularization}}.$$

The dimension of θ, p, is determined by the discretization of X by $D_p^\star$. Typically, for a given number of observations, by increasing p we reduce the *error of approximation* but increase the *error of estimation*. Hence there is a trade-off between these two terms that might be optimized for p by assuming more structure on the model and deriving the minimax rate of convergence of the two types of error as a function of p and n. We stress here that the estimation of θ by $\hat{\theta}$ is done in the double asymptotic regime. By assuming for instance that θ is $s-$sparse (in addition to some mild technical requirements) we have that $\|\theta - \hat{\theta}\|_2^2 = O_p(s \log(p/s)/n)$ (Raskutti et al., 2011). Moreover, by assuming that β belongs to some subset of L^2, say that β is Lipschitz continuous, it should be possible to derive its minimax error of approximation

[2] An identification condition is assumed here. For instance we may assume that $\ker\big(\mathbb{E}(X\langle X, \cdot\rangle)\big) = \{0\}$.

in terms of p. Once this work is done, one may simply find the optimal level of discretization and of regularization as functions of the number of observations. Nevertheless, we need to stress that θ is not independent of β. They are, on the contrary, deeply related in the sense that θ is equal to the sum of the discretized version of β, that is $D_p^\star\beta$, and a bias (as it is clear from the proof below). Hence one needs to check that the assumptions one makes on both these parameters are compatible with each other. *A fortiori* it would be deeply interesting to find that, if β belong to some class of function in L^2, that implies then θ possess some nice properties (such as sparsity). Doing so would make us advance a bit further toward building a bridge between the double asymptotic regime and the infinite-dimensional regime.

To see why pursuing this research agenda further allows connecting the double asymptotic setting with the infinite dimensional setting, we also observe that, as it is clear from the proof below, $D_p^\star\Omega D_p$ is the covariance matrix of $D_p^\star X$. Hence if this covariance matrix is full rank and if we assume $\mathbb{E}(Y|D_p^\star X)$ is linear in $D_p^\star X$, we may set $\lambda = 0$ to actually have that $\mathbb{E}(Y|D_p^\star X) = \langle D_p^\star X, \theta\rangle$ where θ is identified. Those assumptions may appear strong when we start our train of thought from the functional regression model but they are typically the ones we formulate on high-dimensional linear models defined the in the double asymptotic regime. Hence, by identifying what kind of X are compatible with such assumptions, we could be able to think about high-dimensional covariates as discretized versions of some function in a subclass of L^2. Putting aside the philosophical implication of such a point of view would provide us a framework to determine the limits of objects defined in the double asymptotic regime.

In the remainder of this section, we prove the above result.

Proof of Theorem 13

We assumed that there exists a unique non random $\beta \in L^2$ such that

$$\mathbb{E}(Y|X) = \langle X, \beta\rangle_{L^2}.$$

Moreover, since X is in L^2 then Ω is a bounded operator with an unbounded generalized inverse. Observe that we can write for $\lambda \geq 0$ and I_p being the identity vector in $(\mathbb{R}^p, \langle\cdot,\cdot\rangle_{l_p})$ and I_{L^2} being the identity function in L^2.

$$
\begin{aligned}
\mathbb{E}(Y|X) &= \langle X, \beta\rangle_{L^2} \\
X\mathbb{E}(Y|X) &= X\langle X, \beta\rangle_{L^2} \\
\mathbb{E}(\mathbb{E}(XY|X)) &= \mathbb{E}(X\langle X, \cdot\rangle_{L^2})\beta \\
\mathbb{E}(XY) &= \Omega\beta \\
D_p^\star\mathbb{E}(XY) &= D_p^\star\Omega(I_{L^2} + D_pD_p^\star - D_pD_p^\star)\beta \\
D_p^\star\mathbb{E}(XY) &= D_p^\star\Omega D_pD_p^\star\beta + D_p^\star\Omega(I_{L^2} - D_pD_p^\star)\beta \\
D_p^\star\mathbb{E}(XY) &= (D_p^\star\Omega D_p + \lambda I_{l_p} - \lambda I_{l_p})D_p^\star\beta + D_p^\star\Omega(I_{L^2} - D_pD_p^\star)\beta \\
D_p^\star\mathbb{E}(XY) &= (D_p^\star\Omega D_p + \lambda I_{l_p})D_p^\star\beta - \lambda D_p^\star\beta + D_p^\star\Omega(I_{L^2} - D_pD_p^\star)\beta \quad (6.2.1)
\end{aligned}
$$

Now suppose for a moment that for all the $p \in \mathbb{N}$ there exists a non necessarily unique $\theta \in \mathbb{R}^p$ such that

$$\mathbb{E}(Y|D_p^\star X) = \langle D_p^\star X, \theta\rangle_2.$$

In this case, we could write

$$
\begin{aligned}
\mathbb{E}(Y|D_p^\star X) &= \langle D_p^\star X, \theta\rangle_2 \\
D_p^\star\mathbb{E}(X\mathbb{E}(Y|D_p^\star X)) &= D_p^\star\mathbb{E}(X\langle D_p^\star X, \theta\rangle_2) \\
D_p^\star\mathbb{E}(XY) &= D_p^\star\mathbb{E}(X\langle X, D_p\theta\rangle_2) \\
&= D_p^\star\mathbb{E}(X\langle X, \cdot\rangle_2)D_p\theta \\
&= D_p^\star\Omega D_p\theta
\end{aligned}
$$

If $D_p^\star\Omega D_p$ was invertible (observe it has a matrix representation), we could have had written

$$\theta = (D_p^\star\Omega D_p)^{-1}D_p^\star\mathbb{E}(XY).$$

However nothing guarantees the invertibility of the covariance matrix of $D_p^\star X$, nor is there guarantee that the conditional expectation of Y given $D_p^\star X$ is linear in $D_p^\star X$. Hence, instead, we identify θ as

$$
\begin{aligned}
\theta :&= \min_{b\in\mathbb{R}^p} \|D_p^\star\mathbb{E}(XY) - D_p^\star\Omega D_p\|_{l_p}^2 + \lambda\|b\|_{l_p}^2 \\
&= (D_p^\star\Omega D_p + \lambda I_{l_p})^{-1}D_p^\star\mathbb{E}(XY).
\end{aligned}
$$

By combining this equation with Equation (6.2.1) we obtain that

$$
\begin{aligned}
\theta &= (\Omega D_p + \lambda I_{l_p})^{-1}\Big((D_p^\star\Omega D_p + \lambda I_{l_p})D_p^\star\beta - \lambda D_p^\star\beta + D_p^\star\Omega(I_{L^2} - D_pD_p^\star)\beta\Big) \\
&= D_p^\star\beta + (D_p^\star\Omega D_p + \lambda I_{l_p})^{-1}\Big(D_p^\star\Omega(I_{L^2} - D_pD_p^\star)\beta + \lambda D_p^\star\beta\Big).
\end{aligned}
$$

Hence we have written θ as the sum between the discretized version of β and a bias introduced by the discretization. Moreover, by Lemma 18 we have that

$$\lim_{\lambda \to 0} \lim_{p \to \infty} \theta = \beta \text{ up to set of measure zero.}$$

For finite p and $\lambda > 0$, let $\hat{\theta}$ be an estimator of θ, we can then use Proposition 11 to obtain

$$\begin{aligned}
&\|\beta - D_p\hat{\theta}\|_{L^2}^2 \\
&= \|D_p^\star\beta - \hat{\theta}\|_{l_p}^2 + \|\beta - D_pD_p^\star\beta\|_{L^2}^2 \\
&= \|\theta - \hat{\theta} - (D_p^\star\Omega D_p + \lambda I)^{-1}\Big(D_p^\star\Omega(I_{L^2} - D_pD_p^\star)\beta + \lambda D_p^\star\beta\Big)\|_{l_p}^2 + \|\beta - D_pD_p^\star\beta\|_{L^2}^2 \\
&\leq \|\theta - \hat{\theta}\|_{l_p} \\
&+ \|\beta - D_pD_p^\star\beta\|_{L^2}^2\Big(1 + \|(D_p^\star\Omega D_p + \lambda I_p)^{-1}\|_{l_p}^2\|\Omega\|_{L^2}^2\|D_p^\star\|_{l_p}^2\Big) \\
&+ \lambda^2\|D_p^\star\beta\|_{l_p}^2\|(D_p^\star\Omega D_p + \lambda I_{l_p})^{-1}\|_{l_p}^2.
\end{aligned}$$

Since Ω is a bounded operator, $\|D_p^\star\|_{l_p} = 1$ and that $\|(D_p^\star\Omega D_p + \lambda I)^{-1}\|_{l_p}^2$ is bounded, we have that

$$\|\beta - D_p\hat{\theta}\|_{L^2}^2 \leq \|\theta - \hat{\theta}\|_{l_p} + O\Big(\|\beta - D_pD_p^\star\beta\|_{L^2}^2\Big) + O\big(\lambda\|(D_p^\star\Omega D_p + \lambda I_{l_p})^{-1}\|_{l_p}^2\big)$$

□

6.3 Consistency in two types of sequences of statistical experiments

In this section, we analyze two kinds of sequences of statistical experiments with *iid* variables: one where only the number of observations diverges (single asymptotic), one where both the number of variables and the number of observations diverges (the double asymptotic). Our final goal is to formulate concepts of consistency in the latter type of statistical experiment.

6.3.1 Simple asymptotic, independent and identically distributed data

Let $p \in \mathbb{N}$ and suppose a sequence of data, say $\{X_n \in \mathbb{R}^{n \times p}\}_{n \in \mathbb{N}}$. The statistical analysis of this sequence requires a sequence of statistical experiments, that is

$$\{\mathcal{E}_n\}_{n \in \mathbb{N}} = \{S_n, \sigma(S_n), \mathbb{P}_{\theta_n} | \theta_n \in \Theta_n\}_{n \in \mathbb{N}}$$

where $S_n = \mathbb{R}^{n\times p}$, $\sigma(S_n)$ is the $\sigma-$algebra generated by S_n and $\{\mathbb{P}_{\theta_n}|\theta_n \in \Theta_n\}$ is a family of probability measure depending on the parameter $\theta_n \in \Theta_n$.

Example 4 *Let $p \in \mathbb{N}$, a Gaussian statistical experiment is such that*

$$\mathcal{E}_n = \{\mathbb{R}^{n\times p}, \sigma(\mathbb{R}^{n\times p}), \mathcal{N}_{n\times p}(\mu_n, \Sigma_n)|\mu_n \in \mathbb{R}^{n\times p}, \Sigma_n \in (\mathbb{R}^{n\times p} \times \mathbb{R}^{n\times p})\}.$$

for all $n \in \mathbb{N}$.

For cross-sectional data, one often assumes that *the observations are independent and identically distributed*. Formally, this means that the sequence of statistical experiments becomes

$$\{\mathcal{E}_n\}_{n\in\mathbb{N}} = \{\times_n S, \times_n \mathcal{S}, \prod_n \mathbb{P}_\theta|\theta \in \Theta\}_{n\in\mathbb{N}} \tag{6.3.1}$$

where $S = \mathbb{R}^p$, $\times_n S_n = \mathbb{R}^{n\times p}$, $\mathcal{S}$ is the $\sigma-$algebra generated by $\mathbb{R}^p$ and $\times_n \mathcal{S}$ is n times the cross product of $\mathcal{S}$.

Observe that in this case the *parameter space*, does not depend on n.

Example 4 (continued) *For any $p \in \mathbb{N}$, a* iid *Gaussian statistical experiment is such that*

$$\mathcal{E}_n = \{\mathbb{R}^{n\times p}, \times_n \sigma(\mathbb{R}^p), \prod_n \mathcal{N}_p(\mu, \Sigma)|\mu \in \mathbb{R}^p, \Sigma \in (\mathbb{R}^p \times \mathbb{R}^p)\}.$$

for all $n \in \mathbb{N}$.

Suppose that for some function g well defined on the parameter space, we have $g(\theta)$ that is the parameter of interest and suppose it is identified. We observe that for a given such a g, the parameter of interest is well defined on the sequence of statistical experiments. In the following, we assume that $g : \Theta \to \Lambda \subseteq \Theta$ to simplify the discussion.

Example 4 (continued) *For any $p \in \mathbb{N}$, and all $n \in \mathbb{N}$, a parameter of interest is $\mu \in \mathbb{R}^p$.*

Suppose now an estimator for this parameter. An estimator is a function from the sample space to the space of parameter of interest, hence suppose we have $h_n : \times_n S \to \Lambda \subseteq \Theta$.

Example 4 (continued) *For any $p \in \mathbb{N}$, and all $n \in \mathbb{N}$, the estimator is*

$$\mathbb{R}^{n\times p} :\to \mathbb{R}^p, s_n \to h_n(s_n) = \frac{1}{n} s_{i;n}^\top \iota$$

where $s_{i;n}^\top$ is the i^{th} line of the vector s_n and where ι is a vector of ones.

We observe we can define a class of such estimators that takes values in $\Lambda \subseteq \Theta$

$$\mathcal{H} := \{\mathcal{H}_n := \{\times_n S \to \Lambda \subseteq \Theta, \times_{i=1}^n s_i = h_n(\times_{i=1}^n s_i) = \hat{\lambda}_n\} | n \in \mathbb{N}\}.$$

We may want to characterize estimators as random variables, hence suppose that for any $\theta \in \Theta$ and all $h_n \in \mathcal{H}$

$$\forall \lambda \in \sigma(\Lambda), h_n^{-1}(\lambda) \in \times_n \mathcal{S},$$

in this case $\forall \lambda \in \sigma(\Lambda), h_n^{-1}(\lambda) = (\alpha_1, ..., \alpha_n) \in \times_n \mathcal{S}$. Hence for all $h_n \in \mathcal{H}$ and all $\theta \in \Theta$ we can define the following probability measure

$$\forall \lambda \in \sigma(\Lambda), \mathbb{P}_{h_n}^{(\theta)}(\lambda) := \prod_{i=1}^n \mathbb{P}_\theta\big(\{\alpha_i | h_n^{-1}(\lambda) = (\alpha_1, ..., \alpha_n)\}\big)$$

that is the likelihood function.

Consequently, we can define the following class of probability spaces

$$\{\Lambda, \sigma(\Lambda), \mathbb{P}_{h_n}^{(\theta)}\}_{h_n \in \mathcal{H}, \theta \in \Theta}.$$

We are interested in defining the notions of consistence for a sequence of estimators in $\mathcal{H}$. To do so, for any parameter of interest g, we define a class of *loss functions*[3]

$$\mathcal{L}_g := \{L_{g(\theta)} : \mathcal{H} \to I = \mathbb{R}^+; h_n \to L_{g(\theta)} = \|h_n - g(\theta)\|_2 | \theta \in \Theta\}$$

We observe that this class of loss function *does not depends on* n.

Example 4 (continued) *For any $p \in \mathbb{N}$, we have*

$$\mathcal{H} := \{\mathbb{R}^{n \times p} \to \mathbb{R}^p | n \in \mathbb{N}\}$$

and one loss function is

$$L_\mu : \mathcal{H} \to \mathbb{R}_+, h_n \to L_\mu(h_n) = \|h_n - \mu\|_2$$

where $\mathbb{R}_+$ is the positive part of the real line.

Finally, we can consider loss functions as random variables by supposing that for all $L_{g(\theta)} \in \mathcal{L}_g$

$$\forall i \in \mathcal{I}, L_{g(\theta)}^{-1}(i) \in \sigma(\Lambda)$$

[3]Here we consider a class of loss function defined in terms of l_2-norms but this is probably a bad choice and we should consider others class of loss-functions.

where $\mathcal{I}$ is the $\sigma-$algebra generated by I. We can now define, for all $\theta \in \Theta$ and all $h_n \in \mathcal{H}$, the following probability measure

$$\begin{aligned}\forall i \in \mathcal{I}, \mathbb{P}_{L_{g(\theta)},h_n}(i) &= \mathbb{P}_{h_n}^{(\theta)}(\{\hat{\lambda}_n | L_{g(\theta)}(\hat{\lambda}_n) \in i\}) \\ &= \prod_{i=1}^{n} \mathbb{P}_\theta\big(\{\alpha_i | L_{g(\theta)}(\hat{\lambda}_n) \in i; h_n^{-1}(\hat{\lambda}_n) = (\alpha_1, ..., \alpha_n)\}\big).\end{aligned}$$

We make two observations. First $\mathbb{P}_\theta\big(\{\alpha_i | L_{g(\theta)}(\hat{\lambda}_n) \in i; h_n^{-1}(\hat{\lambda}_n) = (\alpha_1, ..., \alpha_n)\}\big)$ does depends on $h_n \in \mathcal{H}$, it is only through its argument, the probability $\mathbb{P}_\theta$ depends only on $\theta \in \Theta$ and does not depends on n. Third, at this point, it becomes possible to speak about the quality of an estimator, for instance, class the sequence of estimators

$$\{\underset{h_n \in \mathcal{H}_n}{\text{argmax}} \min_{L_{g(\theta)} \in \mathcal{L}_g} \mathbb{P}_{L_{g(\theta)},h_n}\}_{n \in \mathbb{N}} \subset \mathcal{H}$$

seems pretty desirable.

In the following, we defined three notions of consistency.

Definition 11 (Very weak consistency) (Adapter from the definition in Lehmann and Casella (1998), Def 8.1, page 54) *Given a sequence of statistical experiments $\mathcal{E}_n$ as in Equation (6.3.1), given an identified parameter of interest g, given a sequence of estimators $\{h_n \in \mathcal{H}_n\}_{n \in \mathbb{N}}$ and given a loss-function $L_{g(\theta)} \in \mathcal{L}_g$, we say that the sequence of estimator is* very weakly consistent *if for all $\theta \in \Theta$ and all $\delta, \epsilon > 0$, there exists $N \in \mathbb{N}$ such that*

$$\sup_{n > N} \mathbb{P}_{L_{g(\theta)},h_n}(L_{g(\theta)}(\hat{\lambda}_n) > \delta) < \epsilon.$$

or equivalently, if for any $\theta \in \Theta$, $\mathcal{G}_\theta$ is any neighbourhood of $g(\theta)$, there exists a $N \in \mathbb{N}$ such that

$$\sup_{n > N} \mathbb{P}_{h_n}(\mathcal{G}_\theta) = 1 - \epsilon$$

Definition 12 (Weak consistency) *Given a sequence of statistical experiments $\mathcal{E}_n$ as in Equation (6.3.1), given an identified parameter of interest g, given a sequence of estimators $\{h_n \in \mathcal{H}_n\}_{n \in \mathbb{N}}$ and given a loss-function $L_{g(\theta)} \in \mathcal{L}_g$, we say that the sequence of estimators is weakly consistent if for all $\delta, \epsilon > 0$, there exists $N \in \mathbb{N}$ such that*

$$\sup_{n > N} \sup_{\theta \in \Theta} \mathbb{P}_{L_{g(\theta)},h_n}(L_{g(\theta)}(\hat{\lambda}_n) > \delta) < \epsilon.$$

or equivalently, if for any $\theta \in \Theta$, $\mathcal{G}_\theta$ is any neighbourhood of $g(\theta)$, there exists a $N \in \mathbb{N}$ such that

$$\sup_{n > N} \inf_{\theta \in \Theta} \mathbb{P}_{h_n}(\mathcal{G}_\theta) = 1 - \epsilon$$

Definition 13 (Strong consistency) *Given a sequence of statistical experiments $\mathcal{E}_n$ as in Equation (6.3.1), given an identified parameter of interest g, given a sequence of estimators $\{h_n \in \mathcal{H}_n\}_{n\in\mathbb{N}}$ and given a loss-function $L_{g(\theta)} \in \mathcal{L}_g$, we say that the sequence of estimators is* strongly consistent *if for all $\delta > 0$, there exists $N \in \mathbb{N}$ such that*

$$\sup_{n>N} \sup_{\theta\in\Theta} \mathbb{P}_{L_{g(\theta)},h_n}(L_{g(\theta)}(h_n) > \delta) = 0 \tag{6.3.2}$$

or equivalently, if for any $\theta \in \Theta$, $\mathcal{G}_\theta$ is any neighbourhood of $g(\theta)$, there exists $N \in \mathbb{N}$ such that

$$\sup_{n>N} \inf_{\theta\in\Theta} \mathbb{P}_{h_n}(\mathcal{G}_\theta) = 1$$

6.3.2 Double-asymptotic, independent and identically distributed data

Suppose we observe a sequence of data sets, say $\{X_n \in \mathbb{R}^{n\times p(n)} | p(n) \to \mathbb{N}\}_{n\in\mathbb{N}}$. In this setup, p is a function of n. In particular, we have in mind situations where $p(n)$ is an increasing function of n. In this setup we are going to redo the same operation as in the previous section, but by taking into account that p is a function of n. The main difference with the previous setup is that, in this case, the parameter space of each statistical experiment is a function of n, even in the *iid* case. At the end of the section, it appears that, by doing so the notion of *very weak consistency* cannot be imported, but that we can still use the notion of *weak* and *strong* consistency.

The statistical analysis of the sequence of data sets $\{X_n \in \mathbb{R}^{n\times p(n)} | p(n) \to \mathbb{N}\}_{n\in\mathbb{N}}$ requires a sequence of statistical experiments, that is

$$\{\mathcal{E}_n\}_{n\in\mathbb{N}} = \{\tilde{S}_n, \sigma(\tilde{S}_n), \mathbb{P}_{\theta_n} | \theta_n \in \Theta_n\}_{n\in\mathbb{N}}.$$

As in the previous section, we assume that the observations are independent and identically distributed. By this, we mean that the sequence of statistical experiments becomes

$$\{\mathcal{E}_n\}_{n\in\mathbb{N}} = \{\times_n S_n, \times_n \mathcal{S}_n), \prod_n \mathbb{P}_{\theta_n} | \theta_n \in \Theta_n\}_{n\in\mathbb{N}} \tag{6.3.3}$$

Here we stress that even by assuming the data are *iid*, the parameter space remains a function of n. This implies firstly that the parameter of interest must be defined in every statistical experiment of the sequence. That is we need to define a class of function $\{g_n\}$ well defined on each parameter spaces Θ_n, and suppose each of its elements is identified on "their" respective statistical experiment. As before we restrict ourselves to the case

$$g_n : \Theta_n \to \Lambda_n \subseteq \Theta_n$$

for all $n \in \mathbb{N}$.

Similarly to the previous section, we define the class of estimators that take values in $\Lambda_n \subseteq \Theta_n$

$$\mathcal{H} := \{\mathcal{H}_n := \{\times_n S \to \Lambda_n \subseteq \Theta_n, h_n \to \hat{\lambda}_n\} | n \in \mathbb{N}\}.$$

We observe that $\mathcal{H}$ still *does not depend on* n but that now, both the domain and the image space of the estimators depend on n.

We still want to characterize estimators as random variables, hence suppose that for all $n \in \mathbb{N}$, for all $\theta_n \in \Theta_n$ and all $h_n \in \mathcal{H}_n$

$$\forall \lambda_n \in \sigma(\Lambda_n), h_n^{-1}(\lambda_n) \in \times_n \mathcal{S}_n,$$

hence we have that for all $n \in \mathbb{N}$, for all $h_n \in \mathcal{H}_n$ and all $\theta_n \in \Theta_n$ we can define the following probability measure

$$\forall \lambda_n \in \sigma(\Lambda_n), \mathbb{P}_{h_n}^{(\theta)}(\lambda_n) := \prod_{i=1}^{n} \mathbb{P}_{\theta_n}\big(\{\alpha_i | h_n^{-1}(\lambda_n) = (\alpha_1, ..., \alpha_n)\}\big).$$

Consequently, we can define the class of probability space

$$\Big\{\{\Lambda_n, \sigma(\Lambda_n), \mathbb{P}_{h_n}^{(\theta)}\}_{h_n \in \mathcal{H}, \theta_n \in \Theta_n}\Big\}_{n \in \mathbb{N}}$$

Now, for any sequence of parameter of interest $\{g_n\}_{n \in \mathbb{N}}$, we define a class of *loss functions*

$$\{\mathcal{L}_{g_n} := \{L_{g(\theta_n)} : \mathcal{H}_n \to I = \mathbb{R}^+, h_n \to L_{g_n(\theta_n)} = \|h_n - g_n(\theta_n)\|_2 | \theta_n \in \Theta_n\}\}_{n \in \mathbb{N}}$$

We observe that each sub-class of loss function, $\mathcal{L}_{g_n}$ *does* depends on n.

Finally, as in the previous section, we can consider loss functions as random variables by supposing that for all $n \in \mathbb{N}$, for all $L_{g(\theta)} \in \mathcal{L}_g$

$$\forall i \in \mathcal{I}, L_{g_n(\theta_n)}^{-1}(i) \in \sigma(\Lambda_n),$$

and we can define, for all $n \in \mathbb{N}$, for all $\theta_n \in \Theta_n$ and all $h_n \in \mathcal{H}_n$, the following probability measure

$$\begin{aligned}\forall i \in \mathcal{I}, \mathbb{P}_{L_{g_n(\theta_n)}, h_n}(i) &= \mathbb{P}_{h_n}^{(\theta)}(\{h_n | L_{g_n(\theta_n)}(h_n) \in i\}) \\ &= \prod_{i=1}^{n} \mathbb{P}_{\theta_n}\big(\{s_i | L_{g_n(\theta_n)}(h_n) \in i; h_n^{-1}(\lambda_n) = (\alpha_1, ..., \alpha_n)\}\big).\end{aligned}$$

Now, we are ready to import the notion of convergence in probability of an estimator to the double asymptotic setting.

Definition 14 (Weak consistency) *Given a sequence of statistical experiments $\mathcal{E}_n$ as in Equation (6.3.3), given a sequence of identified parameters of interest g_n, given a sequence of estimators $\{h_n \in \mathcal{H}_n\}_{n\in\mathbb{N}}$ and given a sequence of loss-function $L_{g_n(\theta_n)} \in \mathcal{L}_{g_n}$, we say that the sequence of estimators is weakly consistent if for all $\delta, \epsilon > 0$, there exists $N \in \mathbb{N}$ such that*

$$\sup_{n>N} \sup_{\theta_n \in \Theta_n} \mathbb{P}_{L_{g_n(\theta_n)}, h_n}(L_{g_n(\theta_n)}(\hat{\lambda}_n) > \delta) < \epsilon.$$

or equivalently, if for any $\theta_n \in \Theta_n$, $\mathcal{G}_{\theta_n}$ is any neighbourhood of $g_n(\theta_n)$, there exists a $N \in \mathbb{N}$ such that

$$\sup_{n>N} \inf_{\theta_n \in \Theta_n} \mathbb{P}_{h_n}(\mathcal{G}_{\theta n}) = 1 - \epsilon$$

Definition 15 (Strong consistency) *Given a sequence of statistical experiments $\mathcal{E}_n$ as in Equation (6.3.1), given a sequence of identified parameters of interest g_n, given a sequence of estimators $\{h_n \in \mathcal{H}_n\}_{n\in\mathbb{N}}$ and given a sequence of loss-function $L_{g_n(\theta_n)} \in \mathcal{L}_{g_n}$, we say that the sequence of estimators is strongly consistent if for all $\delta, \epsilon > 0$, there exists $N \in \mathbb{N}$ such that*

$$\sup_{n>N} \sup_{\theta \in \Theta} \mathbb{P}_{L_{g_n(\theta_n)}, h_n}(L_{g_n(\theta_n)}(h_n) > \delta) = 0 \tag{6.3.4}$$

or equivalently, if for any $\theta_n \in \Theta_n$, $\mathcal{G}_{\theta_n}$ is any neighbourhood of $g_n(\theta_n)$, there exists $N \in \mathbb{N}$ such that

$$\sup_{n>N} \inf_{\theta_n \in \Theta_n} \mathbb{P}_{h_n}(\mathcal{G}_{\theta_n}) = 1$$

Moreover, we observe that in this setting, it is not possible to ensure that $\lim_{n\to\infty} h_n$ is well defined.

Example 5 *For all $n \in \mathbb{N}$, let*

$$S_n \to \Theta_n, X_n \to h_n(X_n) = \left(\frac{1}{n} X_n^\top X_n\right)^{-1}.$$

Then in this case h_n is not well defined at the limit.

Seven

Conclusion

7.1 Overview of the contribution

The contributions of this thesis on high-dimensional data are of three types.

Modelling Chapter 2 suggests interpreting the parameter of a linear regression as the product of a direction and a magnitude. Such parametrization of the linear model has been observed in economic theory (it is for instance induced by the relativity of prices, by the Cobb-Douglass production function or it is explicitly stated in the *Cambridge Equation*), and now find a novel interpretation in econometric theory. Chapter 4 extends the class of locally stationary wavelet processes in a significant way. It is now possible to model multivariate locally stationary wavelet processes without imposing the independence on the random increments across scales.

Identification Chapter 2 identifies the direction of the parameter in a linear model as the generalized eigenspace associated with a pair of measurable matrices. Chapter 3 shows it is possible to identify the non-null parameters of a sparse linear model while allowing colinearity in the set of non-significant random variables. Chapter 4 asymptotically identifies the Evolutionary Wavelet Spectrum of a multivariate locally stationary process as a linear combination of the local covariance functions. Chapter 6 suggests that to identify the true parameter of interest of a sequence of high-dimensional statistical experiments in double asymptotic, one needs to model the limit of the sequence of statistical experiments.

Inference Chapter 2 estimates the direction of the parameter of interest in a high-dimensional sparse linear model via a new estimator of generalized sparse eigenvectors that is based on a well-established algorithm. A Central Limit Theorem is established for the output of the estimation. Chapter 3 considers testing the validity of a selection of variables in the high-dimensional sparse linear model. An application to the problem of nowcasting the Belgian GDP is

considered. Chapter 5 provides an estimator of the evolutionary wavelet spectrum of a multivariate locally stationary wavelet process allowing dependence among the increments at different scales. An application to the problem of measuring the coordination of actors in the oil market is considered. Chapter 6 suggests estimating the parameter of a functional regression via a sequence of estimators of a sequence of parameters of finite but increasing dimensions. The chapter also provides concepts of consistency that are meaningful in the context of double-asymptotics.

7.2 Possible directions for future research

This thesis suggests several open problems for further research, some of which have already been discussed in the previous chapters.

First, the Central Limit Theorem provided in Chapter 2 does not allow constructing confidence intervals because it is tied up with an unobservable bias introduced by the regularization of the problem. Further research may treat this issue by either studying the rate of convergence of the unobservable bias or by studying debiased version of the estimator.

Second, the results in Chapters 2 and 3 are derived under the assumption of the exogeneity of the parameter of interest. It would be of both theoretical and practical interest to study their extension in the context of high-dimensional instrumental sparse regression.

Third, the identification and estimation of the evolutionary wavelet spectrum of the multivariate locally stationary wavelet process of depth J and order q introduced in Chapter 4 depends on a bound on the lowest eigenvalue of the matrix $A_{q,J}$ in Definition 8. The literature has already shown that an absolute bound (valid for any J) exists when $q = 0$. It is would be of great theoretical and practical interest to obtain the same result for some $q > 0$. This may be done by studying the problem for some given q, however, an optimal result would obtain such a bound for an entire subspace of the grid of possibles (q, J).

Fourth, the identification and the estimation of the identification and estimation of the evolutionary wavelet spectrum of the multivariate locally stationary wavelet process introduce in Chapter 4 depends on an assumption formulated in terms of Lipshitz continuity of the element of the evolutionary wavelet spectrum. This assumption restricts the class of process considered to those whose amplitude varies smoothly in time. In the univariate setting, this assumption has already been relaxed by replacing it by an assumption expressed in terms of total variation of the element of the evolutionary wavelet spectrum. This relaxation allowed the class of univariate locally stationary wavelet spectrum to encompass processes whose second-order structure may change abruptly in time. A possible extension of Chapters 4 and 5 would consider doing the same for the class of multivariate locally stationary wavelet

processes.

Fifth, the simulation as well as the application in Chapter 5 indicates the influence of the smoothing parameter on the output of the estimation. Futur research should aim to render the present theory *adaptative*, *i.e.* letting the smoothing parameter be data-driven. This has already been achieved in the univariate framework in Van Bellegem and von Sachs (2008).

Sixth, the univariate class of locally stationary wavelet processes has been fruitfully employed to forecast time series. Further research may consider forecasting in the context of the class of multivariate locally stationary wavelet processes introduced in Chapter 4.

Seventh, Chapters 4 and 5 allow the representation and the estimation of the time-varying second-order structure of a vector of time series. Further research might capitalize on this ability to construct interesting models for applied macroeconometrics such as a locally stationary factor model or a locally stationary network model.

Eighth, the last chapter of this thesis opens two main paths for further research. Firstly, it would be interesting both theoretically and practically to derive an explicit rate of convergences for the bound given in Theorem 13. Moreover, it would be fruitful to obtain that kind of bound for discretization operator defined in terms of other classes of function than the class of simple functions. For instance, approximation by wavelets may be considered. Secondly, it would be of great theoretical interest to study the links existing between the double asymptotic and the infinite-dimensional approach of high dimensionality in a more abstract context than the one of linear regression.

Bibliography

Al Rousan, S., Sbia, R. and Tas, B. K. O. (2018), 'A dynamic network analysis of the world oil market: Analysis of opec and non-opec members', *Energy Economics* 75, 28–41.

Al-Yousef, N. A. (2012), 'The dominant role of saudi arabia in the oil market from 1997–2010', *The Journal of Energy and Development* 36(1), 63–84.

Alkhathlan, K., Gately, D. and Javid, M. (2014), 'Analysis of saudi arabia's behavior within opec and the world oil market', *Energy Policy* 64, 209–225.

Amemiya, T. (1985), *Advanced econometrics*, Harvard university press.

Anderson, T. W. et al. (1958), *Introduction to multivariate statistical analysis*, Wiley.

Attanasi, G., Bodart, V., Courtoy, F., Fontenay, S., Lachapelle, N., Ounnas, A., Sauvenier, M. et al. (2020), Perspectives économiques 2020-2021, Technical report, Université catholique de Louvain, Institut de Recherches Economiques et ....

Bai, J. and Ng, S. (2008), 'Forecasting economic time series using targeted predictors', *Journal of Econometrics* 146(2), 304–317. Honoring the research contributions of Charles R. Nelson.

Banbura, M., Giannone, D. and Reichlin, L. (2010), 'Nowcasting', *ECB working paper* .

Barhoumi, K., Benk, S., Cristadoro, R., Den Reijer, A., Jakaitiene, A., Jelonek, P., Rua, A., Rünstler, G., Ruth, K. and Nieuwenhuyze, C. V. (2008), 'Short-term forecasting of GDP using large monthly datasets–a pseudo real-time forecast evaluation exercise', *National Bank of Belgium Working Paper* .

Behar, A. and Ritz, R. A. (2017), 'Opec vs US shale: Analyzing the shift to a market-share strategy', *Energy Economics* 63, 185–198.

Belloni, A., Chen, D., Chernozhukov, V. and Hansen, C. (2012), 'Sparse models and methods for optimal instruments with an application to eminent domain', *Econometrica* 80(6), 2369–2429.

Belloni, A., Chernozhukov, V. and Hansen, C. B. (2013), *Inference for High-Dimensional Sparse Econometric Models*, Vol. 3 of *Econometric Society Monographs*, Cambridge University Press, p. 245–295.

Berkner, K. and Wells Jr, R. O. (2002), 'Smoothness estimates for soft-threshold denoising via translation-invariant wavelet transforms', *Applied and computational harmonic analysis* 12(1), 1–24.

Bertsimas, D., King, A. and Mazumder, R. (2016), 'Best subset selection via a modern optimization lens', *The Annals of Statistics* 44(2), 813–852.

Bessec, M. (2013), 'Short-term forecasts of french GDP: A dynamic factor model with targeted predictors', *Journal of Forecasting* 32(6), 500–511.

Blumensath, T. and Davies, M. E. (2008), 'Iterative thresholding for sparse approximations', *Journal of Fourier analysis and Applications* 14(5), 629–654.

Blumensath, T. and Davies, M. E. (2009*a*), How to use the iterative hard thresholding algorithm, *in* 'SPARS'09-Signal Processing with Adaptive Sparse Structured Representations'.

Blumensath, T. and Davies, M. E. (2009*b*), 'Iterative hard thresholding for compressed sensing', *Applied and computational harmonic analysis* 27(3), 265–274.

Boivin, J. and Ng, S. (2006), 'Are more data always better for factor analysis?', *Journal of Econometrics* 132(1), 169–194.

Bok, B., Caratelli, D., Giannone, D., Sbordone, A. M. and Tambalotti, A. (2018), 'Macroeconomic nowcasting and forecasting with big data', *Annual Review of Economics* 10, 615–643.

Box, G. E. P. and Meyer, R. D. (1986), 'An analysis for unreplicated fractional factorials', *Technometrics. A Journal of Statistics for the Physical, Chemical and Engineering Sciences* 28(1), 11–18.

Brillinger, D. R. (2001), *Time series: data analysis and theory*, SIAM.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A. et al. (2020), 'Language models are few-shot learners', *Advances in neural information processing systems* 33, 1877–1901.

Bühlmann, P. (2013), 'Statistical significance in high-dimensional linear models', *Bernoulli* 19(4), 1212–1242.

Caggiano, G., Kapetanios, G. and Labhard, V. (2011), 'Are more data always better for factor analysis? results for the euro area, the six largest euro area countries and the uk', *Journal of Forecasting* 30(8), 736–752.

Carvalho, C. M., Chang, J., Lucas, J. E., Nevins, J. R., Wang, Q. and West, M. (2008), 'High-dimensional sparse factor modeling: applications in gene expression genomics', *Journal of the American Statistical Association* 103(484), 1438–1456.

Cepni, O., Guney, I. E. and Swanson, N. R. (2020), 'Forecasting and nowcasting emerging market GDP growth rates: The role of latent global economic policy uncertainty and macroeconomic data surprise factors', *Journal of Forecasting* 39(1), 18–36.

Chen, S. S., Donoho, D. L. and Saunders, M. A. (1998), 'Atomic decomposition by basis pursuit', *SIAM Journal on Scientific Computing* 20(1), 33–61.

Chen, X., Zou, C. and Cook, R. D. (2010), 'Coordinate-independent sparse sufficient dimension reduction and variable selection', *The Annals of Statistics* 38(6), 3696–3723.

Chernozhukov, V., Hansen, C. and Spindler, M. (2015*a*), 'Post-selection and post-regularization inference in linear models with many controls and instruments', *American Economic Review* 105(5), 486–90.

Chernozhukov, V., Hansen, C. and Spindler, M. (2015*b*), 'Valid post-selection and post-regularization inference: An elementary, general approach', *Annu. Rev. Econ.* 7(1), 649–688.

Chetverikov, D., Liao, Z. and Chernozhukov, V. (2021), 'On cross-validated Lasso in high dimensions', *The Annals of Statistics* 49(3), 1300–1317.

Cho, H. and Fryzlewicz, P. (2015), 'Multiple-change-point detection for high dimensional time series via sparsified binary segmentation', *Journal of the Royal Statistical Society: Series B: Statistical Methodology* pp. 475–507.

Coifman, R. R. and Donoho, D. L. (1995), *Translation-invariant de-noising*, Springer.

Colgan, J. D. and Colgan, J. (2021), *Partial hegemony: Oil politics and international order*, Oxford University Press.

Cook, R. D. (1994), 'On the interpretation of regression plots', *Journal of the American Statistical Association* 89(425), 177–189.

Dahl, C. and Yucel, M. (1991), 'Testing alternative hypotheses of oil producer behavior', *The Energy Journal* 12(4).

Dahlhaus, R. (1997), 'Fitting time series models to nonstationary processes', *The annals of Statistics* 25(1), 1–37.

Daubechies, I. et al. (1992), 'Ten lectures on wavelets (siam, philadelphia, 1992)', *MR 93e* 42045.

Declaration of Cooperation OPEC and non-OPEC (2017).

Engle, R. F., Hendry, D. F. and Richard, J.-F. (1983), 'Exogeneity', *Econometrica. Journal of the Econometric Society* 51(2), 277–304.

Fan, J. and Li, R. (2006), Statistical challenges with high dimensionality: feature selection in knowledge discovery, *in* 'International Congress of Mathematicians. Vol. III', Eur. Math. Soc., Zürich, pp. 595–622.

Fan, J. and Liao, Y. (2014), 'Endogeneity in high dimensions', *Annals of statistics* 42(3), 872.

Fan, J., Liao, Y. and Liu, H. (2016), 'An overview of the estimation of large covariance and precision matrices', *The Econometrics Journal* 19(1), C1–C32.

Fan, J. and Lv, J. (2008), 'Sure independence screening for ultrahigh dimensional feature space', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 70(5), 849–911.

Fan, J., Lv, J. and Qi, L. (2011), 'Sparse high-dimensional models in economics', *Annu. Rev. Econ.* 3(1), 291–317.

Fattouh, B. (2007), *OPEC pricing power: The need for a new perspective*, Oxford Institute for Energy Studies.

Fattouh, B. and Mahadeva, L. (2013), 'Opec: what difference has it made?', *Annu. Rev. Resour. Econ.* 5(1), 427–443.

Fisher, I. and Brown, H. (1911), *The purchasing power of money*, Macmillan.

Florens, J., Marimoutou, V., Peguin-Feissolle, A., Perktold, J. and Carrasco, M. (2007), *Econometric Modeling and Inference*, Themes in Modern Econometrics, Cambridge University Press.

Florens, J., Mouchart, M. and Rolin, J. (1990), *Elements of Bayesian statistics*, number vol. 1 *in* 'Pure and applied mathematics, 134', New York : M. Dekker.

Florens, J.-P. and Van Bellegem, S. (2015), 'Instrumental variable estimation in functional linear models', *Journal of Econometrics* 186(2), 465–476.

Follmann, D. and Proschan, M. (2012), 'A test of location for exchangeable multivariate normal data with unknown correlation', *Journal of Multivariate Analysis* 104(1), 115–125.

Fryzlewicz, P., Van Bellegem, S. and Von Sachs, R. (2003), 'Forecasting non-stationary time series by wavelet process modelling', *Annals of the Institute of Statistical Mathematics* 55(4), 737–764.

Fu, Y. (2016), 'Steel and iron ore outlook', Goldman Sachs' Global Investment Research division. Presentation.

Ghojogh, B., Karray, F. and Crowley, M. (2019), 'Eigenvalue and generalized eigenvalue problems: Tutorial'.

Giannone, D., Lenza, M. and Primiceri, G. E. (2021), 'Economic predictions with big data: The illusion of sparsity', *Econometrica* 89(5), 2409–2437.

Giannone, D., Reichlin, L. and Small, D. (2008), 'Nowcasting: The real-time informational content of macroeconomic data', *Journal of monetary economics* 55(4), 665–676.

Girardi, A., Gayer, C. and Reuter, A. (2016), 'The role of survey data in nowcasting euro area GDP growth', *Journal of Forecasting* 35(5), 400–418.

Girardi, A., Golinelli, R. and Pappalardo, C. (2017), 'The role of indicator selection in nowcasting euro-area GDP in pseudo-real time', *Empirical Economics* 53(1), 79–99.

Gold, D., Lederer, J. and Tao, J. (2020), 'Inference for high-dimensional instrumental variables regression', *Journal of Econometrics* 217(1), 79–111.

Golub, G. and Van Loan, C. (2013), *Matrix Computations*, Johns Hopkins Studies in the Mathematical Sciences, Johns Hopkins University Press.

Greenshtein, E. (2006), 'Best subset selection, persistence in high-dimensional statistical learning and optimization under $l_1$constraint', *The Annals of Statistics* 34(5), 2367–2386.

Griffin, J. M. (1985), 'Opec behavior: a test of alternative hypotheses', *The American Economic Review* 75(5), 954–963.

Guo, Y., Zhu, Z. and Fan, J. (2020), 'Best subset selection is robust against design dependence', *arXiv preprint arXiv:2007.01478* .

Hastie, T., Tibshirani, R. and Tibshirani, R. (2020), 'Best subset, forward stepwise or lasso? analysis and recommendations based on extensive comparisons', *Statistical Science* 35(4), 579–592.

Hotelling, H. (1931), 'The Generalization of Student's Ratio', *The Annals of Mathematical Statistics* 2(3), 360 – 378.

Huang, J., Jiao, Y., Liu, Y. and Lu, X. (2018), 'A constructive approach to L_0 penalized regression', *Journal of Machine Learning Research (JMLR)* 19, Paper No. 10, 37.

Huang, Y. (2015), Projection test for high-dimensional mean vectors with optimal direction, PhD thesis. Thesis (Ph.D.)– Department of Statistics, The Pennsylvania State University at University Park.

Huang, Y., Li, C., Li, R. and Yang, S. (2022), 'An overview of tests on high-dimensional means', *Journal of Multivariate Analysis* 188, 104813. 50th Anniversary Jubilee Edition.

Huppmann, D. and Holz, F. (2012), 'Crude oil market power—a shift in recent years?', *The Energy Journal* 33(4).

Jain, P., Tewari, A. and Kar, P. (2014), On iterative hard thresholding methods for high-dimensional m-estimation, *in* Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence and K. Q. Weinberger, eds, 'Advances in Neural Information Processing Systems', Vol. 27, Curran Associates, Inc.

Jing, L., Ng, M. K. and Huang, J. Z. (2007), 'An entropy weighting k-means algorithm for subspace clustering of high-dimensional sparse data', *IEEE Transactions on knowledge and data engineering* 19(8), 1026–1041.

Jung, S., Ahn, J. and Jeon, Y. (2019), 'Penalized orthogonal iteration for sparse estimation of generalized eigenvalue problem', *Journal of Computational and Graphical Statistics* 28(3), 710–721.

Kaufmann, R. K., Bradford, A., Belanger, L. H., Mclaughlin, J. P. and Miki, Y. (2008), 'Determinants of opec production: Implications for opec behavior', *Energy Economics* 30(2), 333–351.

Kim, H. H. and Swanson, N. R. (2014), 'Forecasting financial and macroeconomic variables using data reduction methods: New empirical evidence', *Journal of Econometrics* 178, 352–367.

Klass, M. and Teicher, H. (1987), 'The central limit theorem for exchangeable random variables without moments', *The Annals of Probability* pp. 138–153.

Koch, D. (2015), Multiscale methods for the analysis of high-dimensional locally stationary time series, PhD thesis, UCL-Université Catholique de Louvain.

Lehmann, E. L. and Casella, G. (1998), *Theory of point estimation*, Springer.

Li, L. (2007), 'Sparse sufficient dimension reduction', *Biometrika* 94(3), 603–613.

Liu, W., Yu, X. and Li, R. (2022), 'Multiple-splitting projection test for high-dimensional mean vectors', *Journal of Machine Learning Research* 23(71), 1–27.

Mallat, S. (1999), *A wavelet tour of signal processing*, Elsevier.

Meeks, R. and Monti, F. (2022), 'Heterogeneous beliefs and the phillips curve', *Available at SSRN* .

Meinshausen, N. (2015), 'Group bound: confidence intervals for groups of variables in sparse high dimensional regression without assumptions on the design', *Journal of the Royal Statistical Society: Series B: Statistical Methodology* pp. 923–945.

Miller, A. (2002), *Subset selection in regression*, Vol. 95 of *Monographs on Statistics and Applied Probability*, second edn, Chapman & Hall/CRC, Boca Raton, FL.

Moghaddam, B., Weiss, Y. and Avidan, S. (2006), Generalized spectral bounds for sparse lda, *in* 'Proceedings of the 23rd International Conference on Machine Learning', ICML '06, Association for Computing Machinery, New York, NY, USA, p. 641–648.

Mouchar, M. and Oulhaj, A. (2006), 'The role of the exogenous randomness in the identification of conditional models', *Metron* 64, 253–271.

Nason, G. P., Von Sachs, R. and Kroisandt, G. (2000), 'Wavelet processes and adaptive estimation of the evolutionary wavelet spectrum', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 62(2), 271–292.

Natarajan, B. K. (1995), 'Sparse approximate solutions to linear systems', *SIAM Journal on Computing* 24(2), 227–234.

Ning, Y. and Liu, H. (2017), 'A general theory of hypothesis tests and confidence regions for sparse high dimensional models', *The Annals of Statistics* 45(1), 158 – 195.

Ombao, H. and Van Bellegem, S. (2008), 'Local coherence analysis of non-stationary signals', *IEEE Transactions on Signal Processing* 56(6), 1053.

Park, T., Eckley, I. A. and Ombao, H. C. (2014), 'Estimating time-evolving partial coherence between signals via multivariate locally stationary wavelet processes', *IEEE Transactions on Signal Processing* 62(20), 5240–5250.

Piette, C., Langenus, G. et al. (2014), 'Using brel to nowcast the belgian business cycle: the role of survey data', *Economic Review* (i), 75–98.

Pigou, A. C. (1917), 'The Value of Money', *The Quarterly Journal of Economics* 32(1), 38–65.

Priestley, M. B. (1981), *Spectral analysis and time series: probability and mathematical statistics*, number 04; QA280, P7.

Raskutti, G., Wainwright, M. J. and Yu, B. (2011), 'Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls', *Institute of Electrical and Electronics Engineers. Transactions on Information Theory* 57(10), 6976–6994.

Safo, S. E., Ahn, J., Jeon, Y. and Jung, S. (2018), 'Sparse generalized eigenvalue problem with application to canonical correlation analysis for integrative analysis of methylation and gene expression data', *Biometrics. Journal of the International Biometric Society* 74(4), 1362–1371.

Saito, N. and Beylkin, G. (1993), 'Multiresolution representations using the autocorrelation functions of compactly supported wavelets', *IEEE Transactions on Signal Processing* 41(12), 3584–3590.

Serfling, R. (1980), *Approximation Theorems of Mathematical Statistics*, Wiley Series in Probability and Statistics - Applied Probability and Statistics Section Series, Wiley.

Shah, R. D. and Bühlmann, P. (2018), 'Goodness-of-fit tests for high dimensional linear models', *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 80(1), 113–135.

Shen, X., Pan, W., Zhu, Y. and Zhou, H. (2013), 'On constrained and regularized high-dimensional regression', *Annals of the Institute of Statistical Mathematics* 65(5), 807–832.

Smith, J. L. (2005), 'Inscrutable opec? behavioral tests of the cartel hypothesis', *The Energy Journal* 26(1).

Stewart, G. (1979), 'Pertubation bounds for the definite generalized eigenvalue problem', *Linear Algebra and its Applications* 23, 69 – 85.

Stock, J. H. and Watson, M. W. (2002), 'Forecasting using principal components from a large number of predictors', *Journal of the American statistical association* 97(460), 1167–1179.

Stock, J. H. and Watson, M. W. (2006), 'Forecasting with many predictors', *Handbook of economic forecasting* 1, 515–554.

Stodden, V. (2006), Model selection when the number of variables exceeds the number of observations, PhD thesis. Thesis (Ph.D.)–Stanford University.

Sun, L., Ji, S. and Ye, J. (2009), A least squares formulation for a class of generalized eigenvalue problems in machine learning, *in* 'Proceedings of the 26th Annual International Conference on Machine Learning', ICML '09, Association for Computing Machinery, New York, NY, USA, p. 977–984.

Tan, K. M., Wang, Z., Liu, H. and Zhang, T. (2018), 'Sparse generalized eigenvalue problem: optimal statistical rates via truncated Rayleigh flow', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 80(5), 1057–1086.

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

Van Bellegem, S. and von Sachs, R. (2008), 'Locally adaptive estimation of evolutionary wavelet spectra', *The Annals of Statistics* 36(4), 1879 – 1924.

van de Geer, S. (2016), *Estimation and testing under sparsity*, Vol. 2159 of *Lecture Notes in Mathematics*, Springer, [Cham]. Lecture notes from the 45th Probability Summer School held in Saint-Four, 2015, École d'Été de Probabilités de Saint-Flour. [Saint-Flour Probability Summer School].

van de Geer, S., Bühlmann, P., Ritov, Y. and Dezeure, R. (2014), 'On asymptotically optimal confidence regions and tests for high-dimensional models', *The Annals of Statistics* 42(3), 1166 – 1202.

van der Vaart, A. W. (2000), *Asymptotic statistics*, Vol. 3, Cambridge university press.

Vershynin, R. (2012), Introduction to the non-asymptotic analysis of random matrices, *in* 'Compressed sensing', Cambridge Univ. Press, Cambridge, pp. 210–268.

Yergin, D. (2011), *The prize: The epic quest for oil, money & power*, Simon and Schuster.

Yergin, D. (2012), *The quest: Energy, security, and the remaking of the modern world*, Penguin.

Yuan, G., Shen, L. and Zheng, W.-S. (2019), A decomposition algorithm for the sparse generalized eigenvalue problem, *in* 'Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition', pp. 6113–6122.

Zhang, C.-H. and Zhang, T. (2012), 'A general theory of concave regularization for high-dimensional sparse estimation problems', *Statistical Science. A Review Journal of the Institute of Mathematical Statistics* 27(4), 576–593.

Zhang, Y., Wainwright, M. J. and Jordan, M. I. (2014), Lower bounds on the performance of polynomial-time algorithms for sparse linear regression, *in* 'Conference on Learning Theory', PMLR, pp. 921–948.

Zou, H. and Hastie, T. (2005), 'Regularization and variable selection via the elastic net', *Journal of the royal statistical society: series B (statistical methodology)* 67(2), 301–320.