Granger Causal Network Learning and the Depth Wise Grouped LASSO

by

Ryan Kinnear

A thesis
presented to the University Of Waterloo
in fulfilment of the
thesis requirement for the degree of
Master of Applied Science
in
Electrical and Computer Engineering

Waterloo, Ontario, Canada, 2017

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

In this thesis we study the notion of Granger-causality, a statistical concept originally developed to estimate causal effects in econometrics. First, we suggest a more general notion of Granger-causality in which to frame the proceeding practical developments. And second, we derive a proximal optimization algorithm to fit large and sparse vector autoregressive models, a task closely connected to the estimation Granger-causality amongst jointly wide sense stationary process. Experimental results from our so called "Depth Wise Grouped LASSO" convex program are obtained for both simulated data, as well as Canadian meteorology data. We conclude by discussing some applications and by suggesting future research questions.

Acknowledgements

I must extend the utmost gratitude to Professor Mazumdar for his unwavering support and guidence, without which this work would never have come to be. And moreover, to my friends, peers, and professors for their inspiration and mentorship. Finally, to my parents for their support and encouragement.

To my friends and family

Table of Contents

List of Figures

List of Abbreviations

$\mathbb{P}$-a.s.	Almost surely with respect to the probability measure $\mathbb{P}$..
ADMM	Alternating Direction Method of Multipliers.
AR	Auto Regressive.
DW	Depth-wise.
DWGLASSO	Depth-wise Grouped LASSO.
GLASSO	Grouped LASSO.
LASSO	Least Absolute Shrinkage and Selection Operator.
LMMSE	Linear Minimum Mean Squared Error.
MCC	Matthew's Correlation Coefficient.
OLS	Ordinary Least Squares.
OLST	Ordinary Least Squares with Tikhonov Regularization.
p.d.	Purely Deterministic.
p.n.d.	Purely Non-deterministic.
SCM	Structural Causal Model.
WSS	Wide Sense Stationary.

List of Symbols

$\mathbb{B}(c, r)$	Ball of radius r centered at the point c. The underlying normed space is to be understood from context, but is usually standard Euclidean space. sort.
$\mathbf{0}_n$	Column vector of size n containing only 0.
$\mathbf{1}$	Logical indicator function. $\mathbf{1}(P) = 1$ if proposition P is True, and $\mathbf{1}(P) = 0$ otherwise.
$\mathbf{1}_n$	Column vector of size n containing only 1.
$B^\star$	The statistically optimal parameters for a loss function.
B	Arragement of coefficients of a $\mathsf{VAR}(p)$ model. See definition 3.1.
$\widetilde{B}_{ij}$	Vector of filter coefficients from process $x_j(t)$ to $x_i(t)$ in a $\mathsf{VAR}(p)$ model. See definition.
$R(\tau)$	Covariance matrix of a WSS process $\mathbb{R}(\tau) = \mathbb{E}[X(t)X(t-\tau)^\mathsf{T}]$.
$\overline{\mathbf{conv}}$	Closed convex hull of a set.
$\mathbf{dom}\ f$	The domain of the function f.
$\delta(t)$	The Dirac delta function. $\delta(0) = 1, \delta(t) = 0\ \forall t \neq 0$.
$\mathbb{E}$	Mathematical Expectation.
$G^\star$	The adjacency matrix of the true underlying causality graph. With respect to which modeling space must be inferred by context.
$\mathbf{H}_{t,p}$	Hilbert space linearly generated by the past of p samples of an n dimensional process $x(t)$. $\mathbf{H}_{t,p} = \mathsf{cl}\{\sum_{\tau=1}^{p} A(\tau)x(t-\tau) \mid A(\tau) \in \mathbb{R}^{n\times n}\}$.
I_n	Identity matrix of size $n \times n$.

$\otimes$	Kronecker product of matrices. $A \in R^{m\times n}, B \in R^{p\times q} \implies (A \otimes B) \in \mathbb{R}^{mp\times nq}$ with the $p \times q$ block of $(A \otimes B)$ in the i, j position being given by $a_{i,j}B$..
L_2^n	n-fold Cartesian Product of L_2.
$L_2(\Omega, \mathcal{F}, \mathbb{P})$	Hilbert space of square integrable random variables over $(\Omega, \mathcal{F}, \mathbb{P})$. Usually abbreviated to L_2.
$\|\| \cdot \|\|$	An abstract norm.
n	The dimension of a process $x(t)$, or the number of nodes in a causality graph.
$(\Omega, \mathcal{F}, \mathbb{P})$	Our underlying probability space with sample space Ω, σ-algebra $\mathcal{F}$ and probability measure $\mathbb{P}$.
$\mathbb{P}$	A Probability Measure.
prox_f	Proximal operator of a function f. See definition 3.7.
p	The lag length used in an autoregressive model $\mathsf{VAR}(p)$.
$\mathsf{sgn}(\cdot)$	$\mathsf{sgn}(a) = 1$ if $a > 0$, $\mathsf{sgn}(a) = -1$ if $a < 0$, $\mathsf{sgn}(a) = 0$, otherwise.
T	The transpose of a matrix A is written A^{T}.
τ	A time lag or time difference.
Var	The variance $\text{Var}X = \mathbb{E}[X - \mathbb{E}X]^2$.
$\mathsf{VAR}(p)$	Abbreviation for "Vector Auto Regression of order p". A vector random process is modeled as $x(t) = \sum_{\tau=1}^{p} x(t-\tau) + e(t)$.
$\mathbf{X}_t$	A space generated by the past of an n-dimensional stochastic process $x(t)$, usually $\mathbf{X}_t = \mathbf{H}_t \triangleq \mathsf{cl}\{\sum_{\tau=-\infty}^{t-1} a_\tau^{\mathsf{T}} x(\tau) \| A(\tau) \in \mathbb{R}^{n\times n}\}^n$.
$\mathcal{Z}$	A more convenient arrangement of data for subgradient descent and a distributed formulation of DWGLASSO. See definition 3.2.
Z	The natural arrangement of data for estimating $\mathsf{VAR}(p)$ models via least squares. See definition 3.1.

Geologic history shows us that life is only a short episode between two eternities of death, and that, even in this episode, concious thought has lasted and will last only a moment. Thought is only a gleam in the midst of a long night. But it is this gleam which is everything. - Henri Poincaré, The Value of Science, 1905

Chapter 1

Introduction

1.1 The Philosophy of Causality

The philosophical study of causality dates back at least 2400 years to the time of Plato who stated "everything that becomes or changes must do so owing to some cause; for nothing can come to be without a cause" [1]. Humanity's understanding of causality has changed dramatically since the time of the ancient Greeks, and while this thesis is in no way a philosophical work, we take some time in this introductory chapter to cover the basic conceptions of causation in order to make more clear the nature of our work. Our main source is the short text by Mumford [2] (and some references therein) which lays out a brief and accessible introduction to the philosophy of causality.

Early Beginnings According to Mumford, it was shortly after Plato's statement that Aristotle developed his well known theory of "The Four Causes" [3]. Aristotle's notion of "causality" was a metaphysical theory focused on the determination of what it is that brings anything into "being", and as such touches on much more than what we today understand as causality. In particular, The Four Causes dealt first with the "material cause", which is that which brings something into being, second with the nature of the object's creator, called the "efficient cause", thirdly with the purpose of an object, or it's "final cause", and finally with the "formal cause", which is simply the object's distinguishing feature. Aristotle's efficient cause is the only one having some resemblance to our modern notion.

Following the early beginnings of the ancient Greeks, the modern philosophy of causality began it's development around the time of the age of enlightenment and the scientific revolution with Hume, Spinoza, Locke, and others. And, while a great number of philosophical schools have dealt with causality, particularly the stoics, these thinkers have had a profound influence on the modern (western) understanding of causation.

The principle aims of philosophical theories of causality are to answer questions about what causation *is* and whether or not such a thing even *exists*. Although it is naturally clear in the minds of most men what they mean by "cause", whether or not one has a consistent theory of causality which can be formulated in one of the world's natural languages is another question entirely. Indeed, Bertrand Russel even claimed that there was no such thing as cause, stating: "The law of causality, I believe, like much of a bygone age, surviving, like the monarchy, only because it is

erroneously supposed to do no harm" [4].

Consistent Regularity Revolutionary at the time, David Hume's theory of causation asserted simply that causation was consistent regularity [2]. For example, if it is always the case that a certain sound proceeds the collision of two billiard balls, then it must be that this collision is what causes the sound. This at first seems to be perfectly acceptable reasoning. But, if this really is what defines causality, then our modern conception is in trouble. For example, there is scientific consensus that smoking *causes* lung cancer, but there are many people who have been regular smokers that never developed cancer. Since there is no consistent regularity here, does it mean that science misuses the notion of cause? Quite the contrary, the evidence showing that smoking causes lung cancer seems to be well in tune with our intuition, so from this perspective, it seems clear that causality should be something more than merely consistent regularity.

Physicalist Causation Another view of causation, which seems particularly appealing to engineers and others with a technical background, is one inspired by physics. It has been suggested that what we observe as causation is fundamentally a result of the laws of mechanics acting at the atomic level. This theory is reasonable to follow if we consider again the example of billiard balls, and furthermore the same idea can be used via a reductionist approach to speak of much more complex situations, our previous example of smoking being a cause of lung cancer for instance.

Mumford points out however that the physicalist view of causation starts to run into some difficulties if we were to talk about the cause of the Russian revolution, or the beginning of World War I. Indeed, it is argued that the assassination of Franz Ferdinand caused the outbreak of World War I, and while this may be disputed, no one could seriously argue that it was really caused by the complex mechanistic evolution of particles. This simply does not jive with our common notion of causality. Not only that, it would require one to seriously call into question our own free will. Finally, the modern understanding of quantum mechanics seems to put a nail in the coffin of this physicalist theory of causality.

Counterfactual Dependence Consider the question of necessity. For an event to occur, is it's cause sufficient? Is the cause necessary? Or can there be uncaused events? Can a cause sometimes fail to produce it's effect? These questions may lead into the idea that causation is defined by counterfactual dependence: if B cannot occur unless A has first occurred, then A must cause B. This notion is close to how causality is determined in the sciences (section 1.2), but may still not hold water in philosophy. Indeed, everything you have ever done is counterfactually dependent upon your birth, but to seriously contest that it is your birth which caused all of these things seems to be rather far fetched.

Furthermore, similarly to the physicalist theory, some answers to the questions above may again bring our free will into question. Even if the apparent randomness of quantum mechanics provides the world with uncaused events, does that make us a slave to chance? These are important philosophical questions connected with causality, but certainly will not be of great importance throughout the following chapters of this thesis.

Pluralism To round out our brief philosophical discussion we take note of the view of pluralism. As we have seen, there are a variety of sensible theories which attempt to define causality, but none of them seem to fit our intuition in every case. Perhaps our intuitive notion of cause is not actually a single thing and our language has only misled us to believe that it is. Perhaps the physical theory of causality is well applicable to situations involving mechanics, as in the sun causing the orbit of the planets, and another theory of causality is applicable when we consider what caused the Russian revolution, and yet another when we say that smoking causes lung cancer. Unfortunately, pluralism seems a somewhat intellectually lazy way of escaping the problem of determining what causality *really is*, and hence the debate continues.

1.2 Causality in Science

Sufficient Reason If we return to our opening quote from Plato in section 1.1, is it indeed the case that nothing can come about without a cause? Closely connected to this idea is the *principle of sufficient reason*, originating with Leibniz or perhaps Spinoza [5]. This is vaguely a restatement of Plato's belief that everything must have a cause. More precisely, if something *is*, *occurs*, or *is true* then the principle asserts that there is a sufficient explanation as to *why*. A belief in this principle has been critical for the sciences, and it's importance is often cited for example in the early 20^{th} century writings of the great French mathematician Henri Poincaré [6].

If there were no causal connections between events or processes, science could make no progress. We would not even be able to form collections of facts, as in "mercury is liquid above $234K$ (at 1 atm)", let alone more general laws of nature regarding for example why, how, or under what circumstances matter transforms from solid to liquid. Fortunately, science has enjoyed a great deal of success dispensing with much of the philosophical issues of causality, taking as it's starting point the supposition that there are indeed laws of nature, and furthermore that these laws are unchanging through time. While it seems a daunting task to attempt to prove that this is the case, the myriad of results suggests we can safely proceed. Similarly to why one should not refuse to breathe before they prove it's importance for a continued existence, we should not shun practical notions of causality before coming to agreement on the concept's deep philosophical meaning.

Causation in Physics One of Bertrand Russell's reasons for denying the existence of causation is that the mathematical laws of physics are entirely symmetric. Taking the most common example (e.g. see [7]) of Newton's law $F = ma$, it is common to say that a force *causes* an acceleration. But, we can just as easily write $a = \frac{F}{m}$, yet no one would suggest that an acceleration causes the mass. Science follows to some extent the intuitions of the scientist, and it's success is judged by it's capacity for making predictions; "a force causes an acceleration" produces intuitively pleasing and accurate models of the world, so there is little reason to dispense with the notion of cause. However, it is indeed true that physical laws exhibit symmetry, and physicists do take great care to avoid making causal claims in writing. Yet at the same time, almost every physicist surely makes use of the language of causality in their ordinary discourse.

Proving Causation In contrast to physics, there are many scientists in the fields of biology, medicine, and statistics that do attempt to make serious causal claims in their professional writing. We have seen one already: "smoking causes lung cancer", and at this point, the causal link between smoking and cancer is universally accepted [1].

How do these scientists come to agreement about causation? That is, how do they prove that something causes something else? It is well understood that correlation does not suffice to prove causation, for example, the level of mercury in a tube clearly does not cause a particular ambient pressure level. In order to make claims of causation acceptable to scientists we perform controlled trials. For example, in order to determine the efficacy of a drug in treating a disease, we provide a drug to one group of people having the disease, and a placebo to another, medically similar group of individuals who also have the disease. We compare the outcomes of these two treatments and the second group provides us a means of answering the counter-factual question: "had the first group not been given the drug, what would have been the outcome?". It is in this way that we can conclude that, if the group receiving the drug had a higher recovery rate, that the drug must have been the *cause* of their recovery.

This brings us back to the philosophical discussion of causality. Scientists use counter-factual dependence to test for causal links, but we ask again, is counter-factual dependence the *essence* of cause? Philosophically, the answer seems to be in the negative. Everything you have done since birth is counter-factually dependent on your birth, on your parents having met, on your parent's births, etc... But it makes no sense to suggest that your birth was the *cause* of your actions. Science sidesteps the philosophy and relies to some extent on human intuition. We deem it acceptable in many cases to use counter-factual dependence as a proof of causation because it fits our intuition, and because it has produced reliable and testable results.

1.2.1 The Causal Calculus of Judea Pearl

The work of Judea Pearl [8] has attempted to give the science of causality a concrete calculus. Pearl has argued that the correct way for the sciences to reason rigorously about causation is through a formalism of "Structural Causal Models" (SCMs). An SCM is a triple (X, V, F) where X is a set of "exogenous" variables, determined by nature or factors otherwise outside the model, V is a set of endogenous variables determined by variables in X or V, and F is a set of functions in which f_i assigns a deterministic value to V_i given the values of a subset of $X \cup V$ called the "parents" of V_i. Probabilities are naturally incorporated into this model by assigning a probability measure over the exogenous variables X. Through the formalism of SCMs, scientists can reason about interdependence, confounding, intervention, counterfactuals, etc...

One of the main distinctions between this type of causal modeling and the modern and widely applied methods of statistical learning is that the former seeks not only to make *predictions* about new data points given a training set (the primary domain of statistical learning), but also to predict what would happen given some *intervention* (e.g. when we control the value of a particular variable in an exper-

[1] An interesting historical note is that even Ronald Fisher denied that smoking has a causal link to cancer

iment) and to answer *counter-factual* questions (e.g. had we not given treatment, would the patient have died?). It is argued by Pearl that traditional probabilistic models are sufficient for prediction, but not for reasoning about intervention or counter-factuals.

The key point about Pearl's work is that it seeks to provide a rigorous calculus for causal reasoning on a model of the world. The construction of the model is left to the domain expert.

1.3 Granger Causality

Pearl's conception of causality, based on functional models, competes with the school of thought that argues causation belongs in a probabilistic framework. One particularly well known contributor to this school of thought is Clive Granger [9]. The main idea of Granger's concept of causality is that if from one event, A, we gain knowledge of another, B, where this information is not available from anywhere else, then the former event A must have some causal impact on the latter B.

While we will see that Granger's and Pearl's concept of causality bear little in common, the two have been in agreement that the statistical literature lacks a concrete calculus for causation. Granger in 1980 states "[The statistical] textbooks, having given a cautionary warning about causality, virtually never go on with a positive statement of the form 'the procedure to test for causality is..."' [9].

One of the advantages of Granger's work on causality is that it directly incorporates *time*, a feature of causation critical to much of our intuition about cause and effect, while Pearl's causality does not. Furthermore, it isn't immediately clear how to bring time directly into the picture of Pearl's causal calculus.

1.3.1 Granger's Axioms

Granger presents three axioms of causality, which we will present with mathematical notation, but in an informal manner. A more careful treatment of the definitions is given in chapter 2.

Consider two processes $X(t)$ and $Y(t)$ (these may simply be i.i.d. samples from an experiment), and then with a slight abuse of notation on the time variable t let $\mathcal{F}_t^X$ and $\mathcal{F}_t^Y$ denote all of the information provided to us by X and respectively Y up to time t. Then denote $\mathcal{F}_t$ as all of the information available anywhere (including from $X(t)$ and $Y(t)$) up to time t. We state Granger's three axioms:

1. The past and the present may cause the future, but the future may not cause the past.

2. $\mathcal{F}_t$ contains no redundant information. That is, if $X(t)$ is related in a deterministic and invertible way to $Y(t)$ then only the information produced by one of these processes needs be present in $\mathcal{F}_t$.

3. All causal relations remain constant in direction throughout time.

The first of these axioms is widely accepted, although the reliance on it is criticized by Pearl who suggests that it "excludes a priori the analysis of cases in which

the temporal order is not well-defined" ([8] pg. 250). To this we would simply respond the opposite: Pearl's models of causality do not naturally take into account temporal order, a critical part of much of our causal intuition.

The second axiom is essentially technical in nature, and ensures that removing $\mathcal{F}_t^Y$ from $\mathcal{F}_t$ removes *all* of the information provided by Y.

The third axiom is certainly necessary if we are to make any meaningful conclusions. If we were to allow the possibility that causal laws changed according to the caprice of God, there would again be no science. Without it, all we could say is that "at the particular time of our experiment, A had a causal impact on B".

1.3.2 Defining Causality

Granger's ideas about causality can be written a little bit more mathematically, although still informally; our goal is to convey intuition. We will write $\mathcal{F}_t \setminus \mathcal{F}_t^X$ (analagously for Y) to be the information available from anywhere, *except* what is available from the process Y (recall the second axiom). We note further that it is possible to characterize everything about the processes $X(t)$ and $Y(t)$ with statements of the form $\mathbb{P}\{X(t) \in A\}$, for sets A. Granger's definition of Causality is that $Y(t)$ causes $X(t+1)$ if

$$\exists E \text{ s.t. } \mathbb{P}\{X(t+1) \in E|\mathcal{F}_t\} \neq \mathbb{P}\{X(t+1) \in E|\mathcal{F}_t \setminus \mathcal{F}_t^Y\}. \tag{1.1}$$

This says that knowledge of $Y(t), Y(t-1), \ldots$ provides us with knowledge about $X(t+1)$ that is not available anywhere else.

We modify Granger's definition slightly in order to talk about Y causing X, rather than $Y(t)$ causing $X(t+1)$. The modification is of little consequence in the end since we will later generally suppose that processes are stationary (see remark 5, and the following). We will say that Y Granger-Causes X if

$$\exists t, \exists E \text{ s.t. } \mathbb{P}\{X(t+1) \in E|\mathcal{F}_t\} \neq \mathbb{P}\{X(t+1) \in E|\mathcal{F}_t \setminus \mathcal{F}_t^Y\}. \tag{1.2}$$

The notion of 1.2 differs only slightly from that of 1.1, the intuition is the same but we are enabled to speak of the process Y causing the *process* X, rather than the process Y causing particular samples of X.

Probability and Causality It is widely (not universally, although most exceptions belong to the realm of philosophy and not to that of science) accepted that notions of causality must be connected with probability. The fact is that smoking increasing the probability of developing lung cancer, and this is enough for us to say that there is a causal link. Again, perhaps this does not reach the *essence* of causation, but the idea of using probability to describe causation is well established amongst scientists. There are other probabilistic definitions of causation aside from Granger's, but discussing them here begins to take us too far afield.

Pearl asserts that definitions of causality must rely on an extension of classical probability theory via his structural causal models in order to incorporate intervention and counterfactuals. We are in agreement that Granger's definition is not enough to get at the true idea of causation but this is not our goal, Granger's notion still enjoys significant applicability.

1.3.3 Prima Facie Causation

The easiest way to criticize a definition of causality based on 1.2 is that it provides no room for scientific application as we do not have access to $\mathcal{F}$. In order to move closer to an applicable definition, Granger proceeds by attempting to operationalize 1.2. Suppose we observe the process $x(t) = [x_1(t)\ x_2(t)\ \dots\ x_n(t)]^\mathsf{T}$. We denote all the information we obtain from observing x up to time t as $\mathcal{F}_t^x$ then Granger says x_j is a *prima facie cause* of x_i with respect to x if

$$\exists t,\ \exists E \text{ s.t. } \mathbb{P}\{x_i(t+1) \in E|\mathcal{F}_t^x\} \neq \mathbb{P}\{x_i(t+1) \in E|\mathcal{F}_t^x \setminus \mathcal{F}_t^{x_j}\}. \tag{1.3}$$

That is, x_j provides information about x_i which is not available elsewhere in our data set. This definition is similar to 1.2, but it is weaker. Consider a situation in which X fully determines Y which in turn has a causal effect on Z. Then, Y causes Z with respect to (Y, Z), but not with respect to (X, Y, Z), since knowing X tells us everything we need.

Pearl has pointed out that as a definition of causality, 1.3 is circular. How do we decide which processes to observe? We would do so based on which processes we believe will be causally connected. But how do we decide which processes are causally connected? By application of 1.3. By using the phrase *prima facie*, it is admitted that we have strayed from attempting to formulate a true definition of causality and towards something that is operationally useful. Clearly, 1.3 cannot discern true causality, but it is enough for *plausible causal discovery*. That is, it provides us with a useful tool for investigating causal questions. Processes which exhibit prima facie causation according to equation (1.3) are reasonable candidates for closer consideration.

1.3.4 Granger Causality

The notion most generally referred to as "Granger Causality", and originally alluded to by Wiener, follows. We let $\xi^2[x_i(t+1)|\mathcal{F}_t^x]$ denote the variance of the *linear* minimum mean square error estimator of $x_i(t+1)$ given the history of x, and $\xi^2[x_i(t+1)|\mathcal{F}_t^x \setminus \mathcal{F}^{x_j}]$ the same quantity without the inclusion of x_j in the inforation set. The last assumptions necessary to provide a truly operational notion is wide sense stationarity and ergodicity (see section 2.1.3), because in order to estimate the aforementioned variances in practice, we need a sufficient number of samples from a statistically consistent process. These assumptions consequently set aside the "$\exists t$" portion of the definition, as the variance is unchanged across t. Finally, we say x_j Granger-Causes x_i if

$$\xi^2[x_i(t+1)|\mathcal{F}_t^x] < \xi^2[x_i(t+1)|\mathcal{F}_t^x \setminus \mathcal{F}_t^{x_j}]. \tag{1.4}$$

This follows the same intuition as the previous notions, except now the definition is operational - the tools for estimating these quantities in practice are well understood. The classical formulation of Granger Causality was cemented by Geweke [10] [11] and are elaborated in chapter 3.

1.4 Plausible Causal Discovery and Thesis Outline

In stating equation (1.4) it is clear that we have strayed rather far from what could be considered the *essence* of causality, or even a method for testing true causation. Indeed, Granger Causality is a merely *statistical* notion. However, the fact that it is computable given only observational data and requires no experimental design, intervention, or control confers some utility, particularly in the age of "big data".

Deficiencies We must take care to note some of the immediately obvious deficiencies. The first issue is that we can't reasonably apply Granger-Causality to datasets that cannot be modeled with traditional econometric means, that is, with linear models. Extensions to Granger-Causality have been explored in the literature, and we touch on this issue in chapter 2. Secondly, the set of measured processes is extremely important. It is possible for a Granger-Causal link to disappear amongst processes once we condition on a third. Furthermore, given a high dimensional data set (when the number of processes is comparable to or exceeds the number of samples) care must be taken to minimize the amount of spurious links.

Benefits We can also note some of the immediately obvious benefits. First and foremost, as has already been mentioned, Granger-Causality can be estimated from merely observational data and is hence applicable to situations in which it is impossible or infeasible to interact with the system under study. Modern science is inundated with copious amounts of data, and Granger-Causality provides a tool to help tease out important relationships. The naive approach to this so called "plausible causal discovery" is to look merely at the correlations between data. After all, both correlation and Granger-Causality are linear measures of the relationships between time series. But, it is immediate to see that Granger-Causality is much more powerful than mere correlation. Granger-Causality may be detected between time series that are not at all correlated, and additionally Granger-Causality can deal jointly with a large number of processes, whereas correlation is a merely pairwise (and undirected) concept.

Applications Applications of Granger Causality have been explored in diverse areas including medical imaging [12], neuroscience [13] [14], finance [15], genetics [16], power systems [17], process control [18], and others. These works understand that Granger Causality cannot be used to prove causal relations; it is used as a measure of system connectivity, energy or information transfer, or as a means of narrowing down the search for causation. Applications from the literature, as well as our own application results are given in chapter 4.

Contributions Our primary contribution is to provide a sensible method of regularizing time series models in the context of Granger Causality, and algorithms for carrying out this modeling in practice. Our main results are presented in chapter 3.

Chapter 2

Granger Causality - Theory

2.1 Preliminaries

Before we continue, we must recall some basic results in order to set the stage for what follows.

2.1.1 Hilbert Spaces

Hilbert spaces are a central concept in much of our theoretical development. Everything we review here is standard material.

Definition 2.1 (Hilbert Space)**.** A Hilbert space is a complete inner product space. That is, a vector space $\mathbf{H}$ equipped with inner product $\langle \cdot, \cdot \rangle$ in which Cauchy sequences in $\mathbf{H}$ converge in $\mathbf{H}$. The notion of convergence is furnished by taking the metric $d(x, y) = ||x - y||$, where the norm $|| \cdot ||$ is induced by the inner product via $||x|| = \sqrt{\langle x, x, \rangle}$.

Example 1 (Square Summable Sequences)**.** We will have occasion to refer to spaces of real, square summable sequences which we write $\ell_2(\mathbb{Z})$, and usually abbreviate to simply ℓ_2. That is, $x = (\ldots, x_{-1}, x_0, x_1, \ldots) \in \ell_2$ whenever $\sum_{t=-\infty}^{\infty} x_t^2 < \infty$. It is well known that equipping ℓ_2 with the standard inner product $\langle a, b \rangle = \sum_{t=-\infty}^{\infty} a_t b_t$ forms a Hilbert space.

We recall the projection theorem ([19], p.131), which is of fundamental importance in this thesis:

Theorem 2.1 (Projection Theorem)**.** *Let* $\mathbf{X}$ *be a closed subspace of a Hilbert space* $\mathbf{H}$*. For any* $y \in \mathbf{H}$*, there exists a unique vector* $x \in \mathbf{X}$ *denoted by* $\widehat{\mathbb{E}}[y|\mathbf{X}]$ *such that*

$$\widehat{\mathbb{E}}[y|\mathbf{X}] = \underset{x \in \mathbf{X}}{\operatorname{argmin}} \, ||x - y||.$$

Furthermore,

$$\widehat{\mathbb{E}}[y|\mathbf{X}] = x \iff \langle y - x, z \rangle = 0 \, \forall z \in \mathbf{X}.$$

Remark 1. The vector $\widehat{\mathbb{E}}[y|\mathbf{X}] \in \mathbf{X}$ is referred to as the projection of $y \in \mathbf{H}$ onto $\mathbf{X}$, and we stress the fact that this vector is unique. The second condition in the above theorem is referred to as the orthogonality condition and is often useful in computing the projection.

2.1.2 Convexity

Convex geometry, the theory of convex functions, and the application of this theory to optimization, provides some of the most commonly used mathematical tools in modern applications. It has been said by R. Rockafellar that "*the great watershed in optimization isn't between linearity and nonlinearity, but convexity and nonconvexity.*" We review in this section the key ideas that we apply throughout.

Throughout this section, the definitions rely on some underlying vector space in order for convexity to be defined. Further, the theorem 2.4 relies on an ambient normed space in order to define the norm ball $\mathbb{B}(x, r)$ centered on x and of radius r. Generally, the ambient space will be a Hilbert space, but this much structure is not strictly necessary.

Definition 2.2 (Convex sets)**.** A subset C of a vector space is convex if for every $x, y \in C$, and every $\lambda \in (0, 1)$ we have $\lambda x + (1 - \lambda)y \in C$. Illustrations for subsets of $\mathbb{R}^2$ are given in figure 2.1.

(a) A convex set (b) A non-convex set

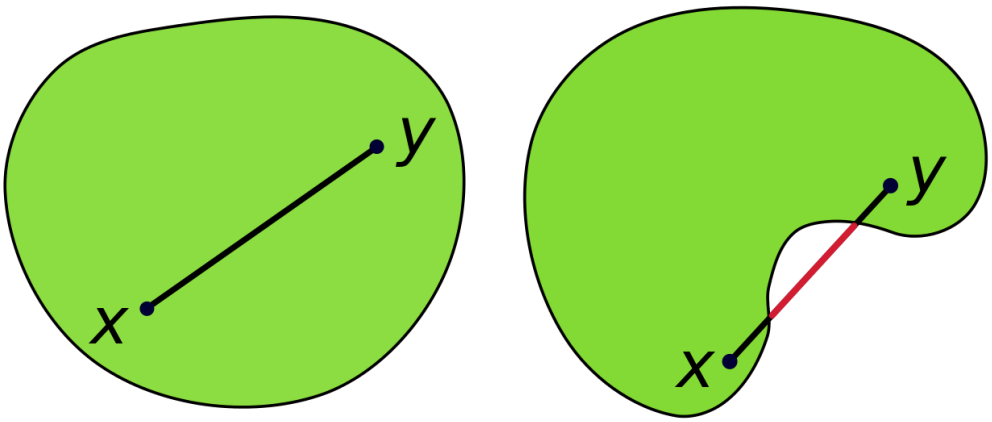


Figure 2.1: Illustrations of Convexity

An important theorem, following immediately from the definition, is that intersections of convex sets remain convex.

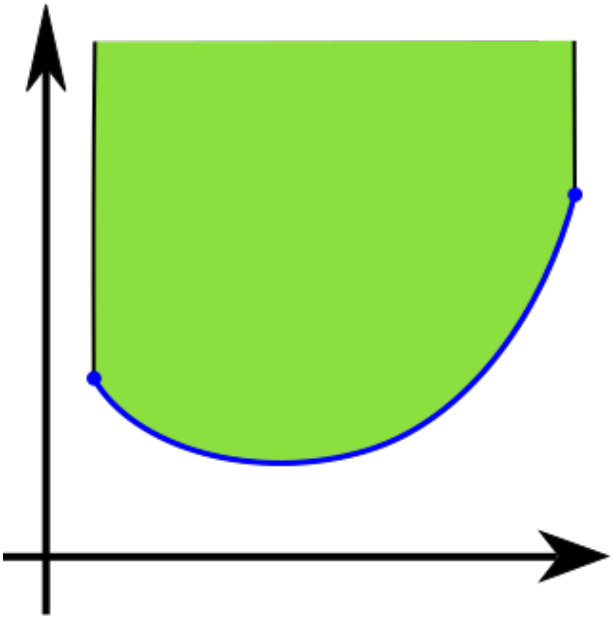
Theorem 2.2 (Intersections of Convex Sets are Convex)**.** *Let C_α be a collection of convex sets indexed by $\alpha \in \mathcal{A}$. The intersection $\bigcap_{\alpha \in \mathcal{A}} C_\alpha$ is convex.*

The idea of convexity can similarly be extended to functions.

Definition 2.3 (Convex Functions)**.** A function $f : C \to \mathbb{R}$ is convex if C is convex and for every $x, y \in C$ and every $\lambda \in (0, 1)$ we have $f(\lambda x + (1 - \lambda)y) \overset{(a)}{\leq} \lambda f(x) + (1 - \lambda)f(y)$. f is called *strictly convex* if the inequality (a) is strict. An illustration is given in figure 2.2

The connection between convex functions and convex sets is through the epigraph of f.

Figure 2.2: A Convex Function



Theorem 2.3 (Epigraphs and Convexity [20]). *A function $f : C \to \mathbb{R}$ is convex if and only if it's epigraph* ***epi*** $f = \{(x,t) \subseteq C \times \mathbb{R} \mid f(x) \leq t\}$ *is convex.*

The reason that convex functions are so important is exemplified in the following theorem, which I refer to as the fundamental theorem for convex functions:

Theorem 2.4 (Fundamental Theorem for Convex Functions [20]). *Let $f : C \to \mathbb{R}$ be a convex function. Define the set of local minimizers*

$$X^\star = \{x^\star \in C \mid \exists \epsilon > 0 \ s.t. \ f(x^\star) \leq f(x), \forall x \in \mathbb{B}(x^\star, \epsilon)\}.$$

Then for each $x^\star \in X^\star$ we have

$$f(x^\star) \leq f(x), \forall x \in C.$$

Moreover, if f is strictly convex, $X^\star$ is a singleton.

Remark 2. Theorem 2.4 tells us that the local minima of convex functions are in fact global minima. We can hence prove global statements about convex functions using only local information.

The case for strictly convex functions is important for asserting uniqueness, and in the context of minimization algorithms. It may be the case that for some iterative minimization procedure generating the sequence $x^1, x^2, \ldots$ it is known that $f(x^n) \to f^\star$, where $f^\star$ is the minimum value of $f(x)$. If f is strictly convex, this hence implies that the sequence x^n is approaching a unique minimizer.

We can generalize the projection theorem 2.1 to any closed convex subset of a Hilbert space.

Theorem 2.5 (Convex Projections [21], [22]). *Let $\mathbf{X}$ be a non-empty closed convex subset of a Hilbert space $\mathbf{H}$. For any $y \in \mathbf{H}$, there exists a unique vector $x \in \mathbf{X}$ denoted by $\widehat{\mathbb{E}}[y|\mathbf{X}]$ such that*

$$\widehat{\mathbb{E}}[y|\mathbf{X}] = \underset{x \in \mathbf{X}}{argmin} \ ||x - y||.$$

A sufficient condition for uniqueness in a Banach space is for the norm to be strictly convex. However, uniqueness fails in the simple case of the 1 norm in $\mathbb{R}^2$.

2.1.3 Probability

Throughout, we will take $(\Omega, \mathcal{F}, \mathbb{P})$ as our underlying probability space and work primarily with the standard space of real valued, and square integrable random variables, denoted $L_2(\Omega, \mathcal{F}, \mathbb{P})$, and usually abbreviated simply to L_2.

Definition 2.4 (L_2)**.** Given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, the $\mathcal{F}$-measurable function (random variable) $x : \Omega \to \mathbb{R}$, with expected value $\mathbb{E}x \triangleq \int_\Omega x(\omega)\mathsf{d}\mathbb{P}(\omega)$ is an element of $L_2(\Omega, \mathcal{F}, \mathbb{P})$ whenever $\mathbb{E}x^2 = \int_\Omega x^2 \mathsf{d}\mathbb{P} < \infty$. Convergence in this space is referred to as "mean square convergence", or "convergence in mean square".

Theorem 2.6. *Equipped with the inner product $\langle x, y\rangle \triangleq \mathbb{E}[xy]$, the space L_2 of square integrable random variables is a Hilbert space, as long as we identify random variables which are equal $\mathbb{P} - a.s.$.*

This theorem is central and we provide proof, however, we lack space to provide all of the requisite background. Additional details can be found in [23].

Proof. It is clear that L_2 is a vector space, and that $\langle \cdot, \cdot \rangle$ is a bonafide inner product. We need only to verify that Cauchy sequences converge. To this end, let x_n be a Cauchy sequence in L_2, that is

$$\mathbb{E}(x_n - x_m)^2 \to 0 \text{ as } n, m \to \infty.$$

From Markov's inequality we obtain

$$\mathbb{P}(|x_n - x_m| \geq \epsilon) \leq \frac{1}{\epsilon^2}\mathbb{E}(x_n - x_m)^2 \to 0,$$

hence x_n is Cauchy in $\mathbb{P}$ and thus there is some x such that $x_n \xrightarrow{\mathbb{P}} x$. Moreover, this implies that there is a subsequence $x_{n'}$ such that $x_{n'} \xrightarrow{a.s.} x$. From this subsequence, we obtain

$$\begin{aligned}
\mathbb{E}(x_n - x)^2 &= \mathbb{E}\big(\liminf_{n' \to \infty} (x_n - x_{n'})\big)^2 \\
&\overset{(a)}{\leq} \liminf_{n' \to \infty} \mathbb{E}(x_n - x_{n'})^2 \\
&\overset{(b)}{\to} 0 \text{ as } n \to \infty
\end{aligned}$$

where (a) follows by Fatou's lemma, and (b) is by the supposed Cauchy property, hence x_n is mean square convergent.

Finally, $\infty > \mathbb{E}(x_n - x)^2 \geq |\mathbb{E}x_n^2 - \mathbb{E}x^2|$ by the reverse triangle inequality gives us $\mathbb{E}x^2 < \infty$. Thus, $x \in L_2$ and L_2 is complete.

□

Remark 3 (Vector Valued Processes)**.** In order to deal with vector valued random variables, we may take an indexed collection of n L_2 variables, and form the vector $x = [x_1, x_2, \ldots, x_n]^\mathsf{T}$. Equipping the space of all such vectors with the inner product $\langle x, y\rangle \triangleq \mathsf{tr}\ \mathbb{E}xy^\mathsf{T} = \mathbb{E}x^\mathsf{T}y$ again yields a Hilbert space, which we will refer to as L_2^n.

Definition 2.5 (Stochastic Process)**.** A stochastic process $x(\omega, t)$ on $(\Omega, \mathcal{F}, \mathbb{P})$ is a time indexed series of random variables, with t in one of $\mathbb{R}, \mathbb{Z}_+, \mathbb{Z},$ or $\{1, 2, \ldots, T\}$ being typical. We usually suppress ω in the notation and write $x(t)$.

In this thesis, we usually work in the discrete time case ($t \in \mathbb{Z}$), and this is to be understood unless otherwise stated. Furthermore, We will usually deal with sequences of random variables in L_2^n, and will write $x(t) \in L_2^n$ to mean that $x(t) = [x_1(t) \ldots x_n(t)] \in L_2^n$ for every t. We have drawn frequently upon [24] and [25] as references for this material.

Examples of stochastic processes (not necessarily in L_2) frequently encountered include sequences of *i.i.d.* random variables, Markov chains, Poisson processes, and Gaussian processes. There are a great variety of references discussing the associated theory and applications of stochastic processes in general e.g. [26], [23].

Remark 4 (Time Series Data)**.** In the context of application, "time series data" refers to finite length sequences of observed data, in which each is also associated with a time stamp indicating when the observation was made. We model such data as a finite number of observations of a stochastic process. That is, for some process $x(t)$, the collection of T samples $(x(t_0), x(t_0+1), \ldots, x(t_0+T-1)$ constitutes a set of time series data.

Stochastic processes are used as abstract models for time series data. Time series analysis, or the closely related area of signal processing, is frequently used in economic and financial modeling, as well as classical areas of electrical engineering. Any phenomena involving time indexed observations of data may be amenable to effective modeling via stochastic processes. Much of the theory is given in e.g. [27], [28].

Remark 5 (Stationarity)**.** Consider a process $x(t)$. Each $x(t_i)$ is a random quantity, but the *statistics*, (the expectation of some measurable functions of the process) e.g. $\mathbb{E}x(t)$, $\mathbb{P}(x(t) \leq r)$, $\mathbb{E}x(t)x(s)$ etc... are deterministic functions of the time parameters t, s. If moreover, these functions are constant, e.g. $\mathbb{E}x(t) = \mu$, at least over an appreciable period of time, then the process is said to exhibit some amount of *stationarity*. Stationarity assumptions are of paramount importance in practice, as we are able only to observe finite quantities of data.

The strongest form of stationarity is as follows:

Definition 2.6 (Strict Sense Stationary)**.** A process $x(t)$ is strict sense stationary (SSS) if for every $t_1, \ldots, t_n$ we have

$$\forall \tau \; F(t_1 + \tau, \ldots, t_n + \tau) = F(t_1, \ldots, t_n),$$

where F gives the joint cumulative distribution function for the process.

The natural stationarity condition for L_2^n process is "wide sense stationarity".

Definition 2.7 (Wide Sense Stationary)**.** The process $x(t) \in L_2^n$ is called wide sense stationary (WSS) if it's covariance matrix $R(s,t) = \mathbb{E}[x(s)x(t)^\mathsf{T}]$ is such that

$$R(s,t) = R(|s-t|, 0) \triangleq R(\tau) \; ; \; \tau = |s-t|.$$

An n-dimensional process $w(t)$ with $R(\tau) = \delta(\tau) I_n$ is called a normalized white noise process, and is an important example of a WSS process.

Of course, strict sense stationarity implies wide sense stationarity, but there are a great many WSS processes that are not SSS.

Remark 6 (Ergodicity)**.** Ergodicity is another important assumption in practice. Essentially, a process is ergodic if it's sample paths are consistent with it's statistics. A WSS process $x(t)$ having meaning μ and covariance $R(\tau)$ is ergodic in mean and covariance if

$$\lim_{T\to\infty}\frac{1}{T+1}\sum_{t=-T/2}^{T/2} x(t) \overset{a.s.}{=} \mu, \tag{2.1}$$

and

$$\lim_{T\to\infty}\frac{1}{T+1}\sum_{t=-T/2}^{T/2} x(t)x(t-\tau) \overset{a.s.}{=} R(\tau), \tag{2.2}$$

that is, sample statistics converge to ensemble statistics. More generally, a process is (completely) ergodic if $P(A) \in \{0,1\}$ for every set A such that $\forall\tau\ x(t) \in A \implies x(t+\tau) \in A$ [1]. This condition ensures that $x(t)$ cannot get "stuck" in some region of space, unless that region does not effect the ensemble statistics of the process.

Ergodicity is a very difficult concept to work with, both in theory and in practice, it will always be assumed in this thesis that the processes we work with are "ergodic enough", which usually comes down to 2.1, 2.2, or the analogous formulae for $t \geq 0$.

2.2 Basic Definitions of Granger Causality

While this thesis is primarily concerned with the framework of linear models and mean squared error, we will make a more general definition of Granger Causality, suggestive of possible extensions. Prior to doing so, we need to define a bit of notation.

2.2.1 Modeling Space

We will define a generic "modeling space" $\mathbf{X}_{t,p}$, as well as the canonical modeling space $\mathbf{H}_{t,p}$ that will be used primarily throughout. Intuitively, we will model a process $x(t) \in L_2^n$ as being generated by some sort of parametric system, driven by some input process. The "size" of $\mathbf{X}_{t,p}$ quantifies the expressiveness of our model. If $\mathbf{X}_{t,p}$ is too small, then it may not be capable of capturing the variation of $x(t)$.

Definition 2.8 (Modeling Space)**.** Let $x(t) \in L_2^n$ be a process, $V \subseteq \{1,2,\ldots,n\}$ a subset of the indices of $x(t)$ so that $x_V(t) = [x_{i_1}(t),\ldots,x_{i_{|V|}}]^\mathsf{T}, i_k \in V$ restricts $x(t)$ to the V sub-indices. Next, let $\mathcal{X}_{t,p}|_V = (\sigma\{x_i(t-\tau) \mid i \in V, 0 < \tau \leq p\})^n$ be the n-times Cartesian product of the filtration generated by the V subset of $x(t)$, over the past p time steps; note that the current time is *not* included. There is no issue with taking the infinite past, e.g., $\mathcal{X}_{t,\infty}|_V = \sigma\big\{\bigcup_{p>0}\mathcal{X}_{t,p}|_V\big\}$.

[1] These are called invariant sets

A generic modeling space $\mathbf{X}_{t,p}|_V \subseteq L_2^n$ is a closed convex space of random variables with components measurable with respect to $\mathcal{X}_{t,p}|_V$.

We make the abbreviations:

$$\mathbf{X}_{t,p} \triangleq \mathbf{X}_{t,p}|_{\{1,\ldots,n\}},$$
$$\mathbf{X}_t \triangleq \mathbf{X}_{t,\infty},$$
$$\mathbf{X}_t^{-i} \triangleq \mathbf{X}_t|_{\{1,\ldots,n\}\setminus\{i\}}.$$

Remark 7 (Modeling)**.** There is some freedom to choose the particulars of this set based on the known or assumed properties of $x(t)$. Intuitively, $\mathbf{X}_{t,p}$ will give the expressiveness of our model, and expanding this space, while potentially making the necessary computations much more arduous, would enable us to capture a wider variety of interactions in a system. However, if we are not interested in modeling the precise nature of $x(t)$, the simple linear space $\mathbf{H}_{t,p}$ (defined below) may still be sufficiently expressive for the detection of causal interactions.

Definition 2.9 (Predicting $x(t)$)**.** Fix a process $x(t) \in L_2^n$ and an associated modeling space $\mathbf{X}_t$. The best prediction in mean squared error, $\widehat{x}(t)$ of $x(t)$ in $\mathbf{X}_t$ is given by

$$\widehat{\mathbb{E}}[x(t)|\mathbf{X}_t] = \operatorname*{argmin}_{z\in\mathbf{X}_t} \mathbb{E}||x(t) - z||_2, \tag{2.3}$$

which is the convex projection of $x(t)$ onto $\mathbf{X}_t$ in the metric defined earlier for L_2^n.

Example 2 (Lagged Linear Combinations)**.** The linear span of the past of $x(t)$, which we will denote by $\mathbf{H}_t$, will be our canonical modeling space. Coordinate wise we have

$$\mathbf{H}_{t,p}|_V = \mathsf{cl}\left\{\sum_{\tau=1}^{p}\sum_{i\in V} b_i^{(\tau)} x_i(t-\tau) \mid b_i^{(\tau)} \in \mathbb{R}\right\}^{|V|}, \tag{2.4}$$

where we have indicated the $|V|$-fold cartesian product of sets. $\mathbf{H}_{t,p}|_V$ is a Hilbert subspace of L_2^n. In vector notation we can write

$$\mathbf{H}_{t,p}|_V = \mathsf{cl}\left\{\sum_{\tau=1}^{p} B(\tau) x_V(t-\tau) \mid B(\tau) \in \mathbb{R}^{|V|\times|V|}\right\}, \tag{2.5}$$

where $x_V(t) \in \mathbb{R}^{|V|}$ is the subvector of $x(t)$ having indices in V.

Example 3 (Logarithmic Features)**.** If we have reason to believe that $x(t)$ is not effectively modeled as a plain autoregressive process, we can try expanding our modeling space by including some additional features. For instance,

$$\mathbf{X}_{t,p}|_V = \mathsf{cl}\left\{A\mathbf{log}(1+|x_V(t-1)|) + \sum_{\tau=1}^{p} B(\tau) x_V(t-\tau) \mid A, B(\tau) \in \mathbb{R}^{|V|\times|V|}\right\}, \tag{2.6}$$

where the log function is to be interpreted as applying element-wise. Using the inequality $\mathbf{log}(1+|x|) \leq |x|$ we see that $\mathbb{E}\mathbf{log}(1+|x(t)|)_i^2 \leq \mathbb{E}|x(t)|^2 < \infty$, and hence $\mathbf{X}_{t,p} \subseteq L_2^n$. The closed space $\mathbf{X}_{t,p}$ is thus Hilbert. Furthermore, it is convex with $\mathbf{H}_{t,p} \subset \mathbf{X}_{t,p}$. Hence, $\mathbf{X}_{t,p}$ is a more expressive modeling space than $\mathbf{H}_{t,p}$.

Remark 8. The additional modeling power afforded to us by considering convex spaces comes also allows the potential to model constraints. We may know, for instance, that $x(t)$ is bounded, or we may know a priori about some of the interactions between different components of $x(t)$. If we have some prior information about $x(t)$, we are free to use the linear spaces given above, or generalizations thereof, and restrict them to some convex subset that encodes some of our prior beliefs, without impacting the definitions below. The critical feature of projecting onto a convex set is the uniqueness of the projection c.f. 2.5.

Example 4 (Convex Restriction)**.** Consider a convex set $C \subseteq \mathbb{R}^n$. The set $\mathbf{C} = \{z \in L_2^n \mid \mathbb{P}(z \in C) = 1\}$ inherits convexity from C, and so we are able to define the convex restricted space $\mathbf{X}_{t,p}^C = \mathbf{X}_{t,p} \cap \mathbf{C}$. The restricted space maintains it's convexity by virtue of the intersection property 2.2.

Example 5 (Convex Restriction of $\mathbf{H}_{t,p}$)**.** Suppose that $e(t) \in L_2^n$ is a sequence of i.i.d. random vectors. For some stable[2] matrix $A \in \mathbb{R}^{n \times n}$, define the process $x(t) = Ax(t-1) + e(t)$. If $e(t)$ is, for example, a sub-Gaussian random variable, then we will have that for some $r \geq 0$, $x(t)$ is confined to the set

$$S_t = \{Ax(t-1) + u \mid \rho(A) < 1, ||u||_2 \leq r\}$$

with high probability, where $\rho(A) = |\lambda_{\max}(A)|$ denotes the spectral radius of A. We would like to project onto the restricted set $\mathbf{H}_{t,p} \cap S_t$ in order to produce estimates consistent with our prior knowledge.

However, S_t is not in general a convex set[3]. In order to obtain well defined projections, we propose two convex approximations of S_t, the first corresponding roughly to the proposal given in the related work of [29]:

$$\underline{C_t} = \{Ax(t-1) + u \mid ||A|| \leq \gamma, ||u||_2 \leq r\} \tag{2.7}$$

$$\overline{C_t} = \{Ax(t-1) + u \mid \mathsf{tr}|A| \leq n\gamma, ||u||_2 \leq r\}, \tag{2.8}$$

where the convexity of these sets follows directly from the definition. Furthermore, both $\underline{C_t}$ and $\overline{C_t}$ are closed. By virtue of their convexity, and of the inequalities $\rho(A) \leq ||A||$ and $\mathsf{tr}|A| \leq \rho(A)$ (where $|A|$ denotes element-wise absolute value) we have the sequence of inclusions

$$\underline{C_t} \subseteq \overline{\mathbf{conv}}\ S_t \subseteq \overline{C_t}.$$

The projections

$$\widehat{\mathbb{E}}[x(t)|\mathbf{H}_{t,p} \cap \underline{C_t}],\ \widehat{\mathbb{E}}[x(t)|\mathbf{H}_{t,p} \cap \overline{C_t}] \tag{2.9}$$

are thus well defined, and serve as an approximation to the ill-defined quantity "$\widehat{\mathbb{E}}[x(t)|\mathbf{H}_{t,p} \cap S_t]$".

[2] A stable matrix is one in which all of it's eigenvalues are contained strictly within the unit circle. A recursive process

$$x(t) = Ax(t-1) + e(t)$$

remains bounded for bounded $e(t)$ if and only if A is stable.

[3] $\rho(A)$ is convex when A is symmetric.

2.2.2 Defining Granger Causality

Suppose we have an underlying model space $\mathbf{X}_{t,p}$. With this space understood, we will write $\widehat{x}(t) \triangleq \widehat{\mathbb{E}}[x(t)|\mathbf{X}]$, and $\widehat{x}_i(t)$ to be the i^{th} component thereof. Let

$$\xi[x_i(t)|\mathbf{X}] \triangleq \mathrm{Var}\{x_i(t) - \widehat{x}_i(t)\}, \tag{2.10}$$

which is the error variance between $x_i(t)$ and the i^{th} component $\widehat{x}_i(t)$ of the projection $\widehat{x}(t)$ onto $\mathbf{X}$. If the estimate $\widehat{x}(t)$ is unbiased, that is, $\mathbb{E}\widehat{x}(t) = \mathbb{E}x(t)$ then this reduces to the mean squared error.

We now define four quantities, inspired by [10], but giving the notion a greater level of generality:

Definition 2.10 (Measures of Interaction)**.** Given the modeling space $\mathbf{X}_{t,p}$, and the related spaces as described above in section 2.2.1, we define the four fundamental quantities

$$\begin{aligned}
F_{x_j \to x_i}(t) &= \xi[x_i(t+1)|\mathbf{X}_{t,p}^{-j}] - \xi[x_i(t+1)|\mathbf{X}_{t,p}], \\
F_{x_i \to x_j}(t) &= \xi[x_j(t+1)|\mathbf{X}_{t,p}^{-i}] - \xi[x_j(t+1)|\mathbf{X}_{t,p}], \\
F_{x_i \leftrightarrow x_j}(t) &= \xi[x_i(t+1)|\mathbf{X}_{t,p}] - \xi[x_i(t+1)|\mathbf{X}_{t,p}, \mathbf{X}_{t+1,p+1}|_j] \\
&\quad + \xi[x_j(t+1)|\mathbf{X}_{t,p}] - \xi[x_j(t+1)|\mathbf{X}_{t,p}, \mathbf{X}_{t+1,p+1}|_i], \\
F_{x_i,x_j}(t) &= \xi[x_i(t+1)|\mathbf{X}_{t,p}^{-j}] - \xi[x_i(t+1)|\mathbf{X}_{t,p}, \mathbf{X}_{t+1,p+1}|_j] \\
&\quad + \xi[x_j(t+1)|\mathbf{X}_{t,p}^{-i}] - \xi[x_j(t+1)|\mathbf{X}_{t,p}, \mathbf{X}_{t+1,p+1}|_i],
\end{aligned} \tag{2.11}$$

where $\widehat{\mathbb{E}}[x(t)|\mathbf{X}_1, \mathbf{X}_2] \triangleq \widehat{\mathbb{E}}[x(t)|\overline{\mathbf{conv}}\ \mathbf{X}_1 \cup \mathbf{X}_2]$, the projection onto the closure of the convex hull[4] of $\mathbf{X}_1 \cup \mathbf{X}_2$. Given sufficient stationarity assumptions, these quantities will not vary with t.

The quantity $F_{x_j \to x_i}$ is directly analogous to the original idea of Granger-causality, and measures the "energy" transfer from x_j to x_i strictly forward in time. The quantity $F_{x_i \leftrightarrow x_j}$ gives a measure of instantaneous feedback between the two processes by taking into account the current sample of x_j. Finally F_{x_i,x_j} gives a measure of the total interaction between x_i and x_j. We have the fundamental decomposition, analogous again to that given by Geweke [10]:

Proposition 2.1. *The quantities given in equation (2.11) satisfy the decomposition*

$$F_{x_i,x_j|x}(t) = F_{x_i \leftrightarrow x_j|x}(t) + F_{x_i \to x_j|x}(t) + F_{x_j \to x_i|x}(t). \tag{2.12}$$

Proof. Immediate from equation (2.11) □

Remark 9. The proposition 2.1 verifies that the notions of Granger-causality can be sucessfully generalized to a more sophisticated modelling space while still maintaining it's fundamental character.

We are now in a position to state our main definition:

[4]The convex hull of a set is the smallest convex set which contains it.

Definition 2.11 (Granger-causality)**.** If $\exists t \in \mathbb{T}$ such that

$$F_{x_j \to x_i}(t) > 0 \tag{2.13}$$

we say that x_j Granger-causes x_i with respect to $\mathbf{X}$, and write $x_j \underset{\mathbf{X}}{\longrightarrow} x_i$.

The intuition behind this definition is detailed in section 1.3, but essentially it says that x_j provides information about x_i which is not available from any other component of x.

The preceding discussion suggests the possibility to discuss the idea of Granger-causality for a fairly large class of models. However, the definition of 2.11 is obviously not easy to work with. For most of our work we will restrict ourselves to the special case in which $x(t)$ is wide sense stationary with $\mathbf{X}_t = \mathbf{H}_t$ (see example 2 which yields a more standard definition of Granger-causality.

Definition 2.12 (Granger-causality)**.** Fix some $t \in \mathbb{T}$. If $x(t) \in L_2^n$ is wide sense stationary, then x_j Granger-causes x_i with respect to $\mathbf{H}$ if

$$\xi[x_i(t)|\mathbf{H}_t] < \xi[x_i(t)|\mathbf{H}_t^{-j}] \tag{2.14}$$

Unless otherwise specified, definition 2.12 (equation 2.14) is what we mean by "Granger-causality" in the remainder of this thesis. This notion is in fact what was first proposed by Granger in [30].

In the sequel we will have occasion to deal with a small set of named scalar L_2 processes, say $x(t), y(t), z(t)$. In this case we will write $x \underset{(w,x,y)}{\longrightarrow} y$ to indicate that x Granger-causes z with respect to the Hilbert space linearly generated by the past of the combined process (w, x, y).

2.2.3 Basic Properties

A number of important properties of Granger-causality follow directly from the Hilbert space structure. We first define some notation. Denote by $g_i(u) \in \ell_2(\mathbb{N})$ the sequences of coefficients of $i = 1, 2, \ldots, n$ linear, causal, and invertible filters. Then define a sequence of *diagonal* matrices $H(u) \in \mathbb{R}^{n \times n}$ such that $H_{ii}(u) = g_i(u)$. H acts on processes $x(t) \in L_2^n$ via convolution: $(H * x)(t) \triangleq \sum_{u=0}^{\infty} H(u)x(t-u)$. And, convolution with the diagonal matrix of inverse filters $H^{-1}(u)$ yields by definition: $(H^{-1} * H * x)(t) = x(t)$. Define by $H * \mathbf{H}_t$ the Hilbert space generated by the past of the filtered processes $(H * x)(t)_i$.

Proposition 2.2 (Invariance to Causal, Invertible, LSI Filtering)**.** *Let $x(t) \in L_2^n$ and $H(u) \in \mathbb{R}^{n \times n}, u \in \mathbb{N}$ denote the diagonal matrix of coefficients of n causal and invertible LSI filters. Then $H * \mathbf{H}_t = \mathbf{H}_t$ and if $x_j \underset{\mathbf{H}}{\longrightarrow} x_i$, then $g_j * x_j \underset{H*\mathbf{H}}{\longrightarrow} g_i * x_i$.*

Essentially, since the Hilbert space H_t contains all of the linear combinations of the past of $x(t)$, it also contains the inverse filters.

Proof. $H * H_t \subseteq H_t$ is evident. To show the converse let $z \in H_t$. Then, $z = \sum_{i=1}^n (A * x)(t)_i$ for some sequence of (dense) matrices $A(\tau) \in \ell_2^{n \times n}(\mathbb{N})$ hence $\sum_{i=1}^n (H^{-1} * A * x)(t)_i \in H_t$ and further $\sum_{i=1}^n (H * H^{-1} * A * x)(t)_i = \sum_{i=1}^n (A * x)(t)_i = z \in H * \mathbf{H}_t$. □

Remark 10. The utility of this result is in justifying Granger-causal inference for systems which are only observable after filtering, and to justify any filtering operations as preprocessing steps.

2.3 Granger Causality Graphs

The graphical structure of Granger Causality is obvious from the definition. We will define a "Granger-causality Graph" $\mathcal{G} = (E, V)$ as follows: take each component process $x_i(t)$ of $x(t)$ to be a vertex in V. The edge (j, i) is present in E if $x_j \underset{\mathbf{X}}{\longrightarrow} x_i$. It is generally the case that we have measurements of $x(t)$, and hence V is known. Determining the edge set E however, requires us to detect Granger-causal relations between the component processes of $x(t)$. Since V is fixed, it is natural to use an adjacency matrix representation of the graph $\mathcal{G}$, which will be denoted G. Estimating this adjacency matrix is our primary interest.

Remark 11. The inferred causality graph is of course dependent upon the particular modeling space that is chosen. That being said "Granger-causality Graph" almost always refers to the graph induced by definition 2.12 using the canonical modeling space **H**. It is a question for future research about the relationships between graphs induced by different modeling spaces. A particularly interesting avenue is to investigate when causality with respect to the simplest space **H** is sufficient to recover the graph induced by a more sophisticated modeling space.

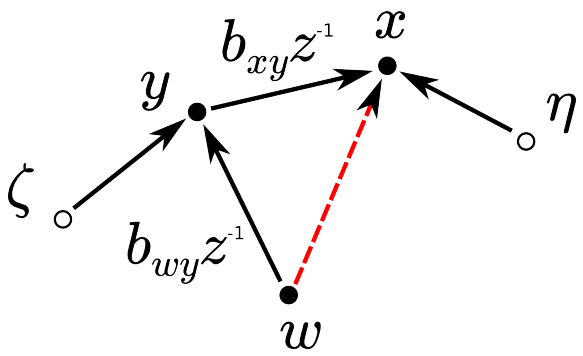
2.3.1 Pairwise Granger-Causality

Given $x \in L_2^n$ we will say that x_j pairwise Granger-causes x_i if $x_j \underset{(x_j, x_i)}{\longrightarrow} x_i$. An interesting question to ask is about the relationship between pairwise Granger-causality and Granger-causality with respect to all the observed information. Unfortunately, pairwise Granger-causality is not sufficient to conclude Granger-causality, and it is not necessary either. The importance of the caveat "with respect to **X**" in definition 2.12 was discussed by Granger in [9].

Suppose there is an underlying Granger-causality graph $\mathcal{G} = (E, V)$. If we construct an edge set E_p via pairwise tests, then neither $E_p \subseteq E$ (sufficiency) nore $E \subseteq E_p$ (necessity) need to hold.

The insufficiency of pairwise testing can be seen from figure 2.3.

Figure 2.3: Pairwise Granger-Causality is not Sufficient



The system described by this graph involves 5 wide sense stationary L_2 processes w, x, y, ζ, η in which we suppose ζ, η are white, uncorrelated, and unobserved driving processes, and w, x, y are observed. We use z^{-1} to indicate a one step lag. The

equations from which this graph is derived are

$$\begin{aligned} x(t) &= b_{xy}y(t-1) + \eta(t), \\ y(t) &= b_{yw}w(t-1) + \zeta(t). \end{aligned} \tag{2.15}$$

The red dashed edge from w to x in figure 2.3 would be incorrectly inferred via pairwise testing. That is

$$w \underset{(w,x)}{\longrightarrow} x,$$

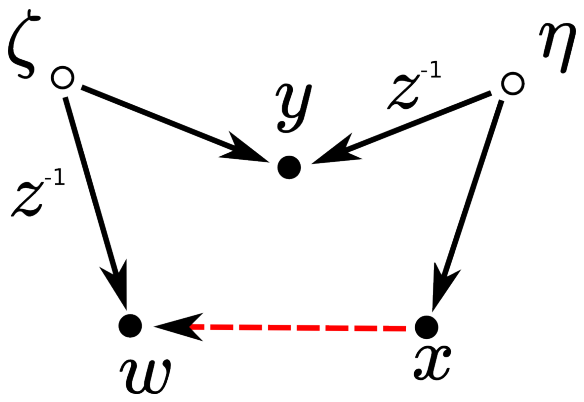
but

$$w \underset{(w,x,y)}{\not\longrightarrow} x.$$

The issue with pairwise testing in this case is that w has an indirect effect on x which is detected by pairwise testing, but is irrelevant when y is taken into account.

Similarly, the non-necessity of pairwise testing can be illustrated as in 2.4.

Figure 2.4: Pairwise Granger-Causality is not Necessary



In this case we have:

$$\begin{aligned} x(t) &= \eta(t), \\ w(t) &= \zeta(t-1), \\ y(t) &= \zeta(t) + \eta(t-1). \end{aligned} \tag{2.16}$$

The red dashed line in 2.4 would not be present in a pairwise test with respect to (x, w), but it is present with respect to (x, y, w) since w can be described as $w(t) = y(t-1) - x(t-2)$ indirectly through the noise processes η and ζ.

Remark 12. There are surely additional assumptions we may be able to impose on our system to rule out the example of figure 2.4, and these pairwise graphs may still be of use. We have not pursued this avenue in great detail, and here focus on joint inference on the entire graph.

2.4 Time Series Models

In this section we review The Wold decomposition, a critically important theorem in the analysis of L_2 stochastic processes, which serves as the motivation and justifi-

cation for the use of autoregressive modeling in Granger-Causal analysis. We recall the key aspects of the theory [24], [31].

Intuitively, the Wold decomposition theorem tells us that "almost" every stationary process in L_2 admits a representation as causally filtered white noise, and that "almost" each of these processes in turn admit an autoregressive representation. To elucidate exactly which proccesses admit this representation, we require the following notion of a purely nondeterministic (p.n.d) process:

Definition 2.13 (Purely Nondeterministic Process)**.** Let $x(t)$ be an n-dimensional w.s.s. L_2 process, and let $\mathbf{H}_t$ be the Hilbert space generated by the past of x as in equation (2.5). If there exists a k-dimensional ($k \leq n$) L_2 normalized white noise process which is jointly stationary with x[5] generating the hilbert space $\mathbf{W}_t$ such that

$$\mathbf{H}_t = \mathbf{W}_t \; ; \; \forall t \in \mathbb{Z}, \tag{2.17}$$

then $x(t)$ is a purely nondeterministic process.

Further, we can also define purely deterministic (p.d.) processes

Definition 2.14 (Purely Deterministic Process)**.** Suppose $x(t)$ is as in definition 2.13. Then $x(t)$ is purely deterministic if

$$\mathbf{H}_t = \mathbf{H}_\infty \; ; \; \forall t \in \mathbb{Z}. \tag{2.18}$$

The idea of a deterministic process is that it is perfectly predictable given only a single observation. For example, a pure sinusoid whose phase is determined by a single random variable Θ is a classic example of a purely deterministic process.

We can now state the Wold decomposition theorem:

Theorem 2.7 (Wold Decomposition)**.** *Let $y(t)$ be an n-dimensional w.s.s. L_2 process. Then there exists a p.n.d. process $x(t)$ and a p.d. process $z(t)$ such that*

$$y(t) = x(t) + z(t)$$

Further, $x(t)$ admits the representation:

$$x(t) = \sum_{\tau=0}^{\infty} M(\tau) w(t-\tau) \tag{2.19}$$

where $w(t)$ is a k-dimensional[6] ($k \leq n$) normalized white noise process and $\forall \tau \; M(\tau) \in \mathbb{R}^{n \times k}$ with $M(\tau)_{ij} \in \ell_2(\mathbb{Z})$.

It is possible to specify $x(t)$ and $z(t)$ in terms of Hilbert space projections, but this is not necessary for our purposes. The point of this theorem is that a very wide variety of stochastic processes can be written as a sum of some "trend" $z(t)$ (which we would subtract) and a linearly filtered white noise $w(t)$ called the "innovations" of the process. Moreover, we will usually assume that $k = n$, as this simply eliminates the uninteresting case where $x(t)$ is a.s. confined to a strict subspace of $\mathbb{R}^n$.

[5] x and w are jointly stationary if $(x(t)\, w(t))$ is stationary

[6] k is the dimension of the Hilbert space spanned by y.

2.4.1 Inverting 2.19

The Wold decomposition is an extremely powerful theorem, but equation 2.19 in it's current state is not very useful for our applications, since we have no possibility to observe directly the noise $w(t)$ (although it can be calculated via linear projections). In this section, we take note of the conditions for inverting 2.19 into an autoregressive version.

Consider the spectrum $S_x(\lambda)$ of the WSS process $x(t)$, where $\lambda \in [-\pi, \pi)$. If there is some constant $c > 0$ such that, for λ almost everywhere

$$c^{-1} I \preceq S_x(\lambda) \preceq cI \tag{2.20}$$

then equation (2.19) is invertible and hence we get the autoregressive representation:

$$x(t) = \sum_{\tau=1}^{\infty} B(\tau) x(t-\tau) + e(t), \tag{2.21}$$

where $e(t)$ is a temporally uncorrelated, though not necessarily white sequence, which is also uncorrelated with $x(t-\tau)$, for any $\tau \geq 1$. See [31] p.78.

2.4.2 Granger Causality in Autoregressive Models

Suppose we have a wide sense stationary process $x \in L_2^n$. Suppose further that $x(t)$ is generated by the autoregressive model 2.21 in which $e(t)$ is a zero mean serially uncorrelated sequence, which is also uncorrelated with $\{x(t-\tau) \mid \tau \geq 1\}$, having autocorrelation sequence

$$\mathbb{E}[e(t)e(t-\tau)^{\mathsf{T}}] \triangleq R_e(\tau) = \delta(\tau) R_e. \tag{2.22}$$

The Hilbert space projections in this model are trivial:

Proposition 2.3 (Hilbert Space Projections for AR Models)**.** *Let $x \in L_2^n$ be a wide sense stationary stochastic process generated by the autoregressive model 2.21. Let $\mathbf{H}_t$ be the Hilbert space generated by the past of $x(t)$ as in 2.5. Then $\widehat{x}_i(t) = \sum_{\tau=1}^{\infty} \sum_{k=1}^{n} B_{ik}^{(\tau)} x_k(t-\tau)$.*

Proof. Fix some t and suppose $\widehat{x}_i(t) = \sum_{\tau=1}^{\infty} \sum_{k=1}^{n} \tilde{B}_{ik}^{(\tau)} x_k(t-\tau)$, for some square summable sequence $\tilde{B}_{ik}$. Then

$$\mathbb{E}\Big[\big(x_i(t) - \sum_{\tau=1}^{\infty}\sum_{k=1}^{n} \tilde{B}_{ik}^{(\tau)} x_k(t-\tau)\big)^2\Big] = \mathbb{E}[e_i(t)^2] + \mathbb{E}\Big[\big(\sum_{\tau=1}^{\infty}\sum_{k=1}^{n} (B_{ik}^{(\tau)} - \tilde{B}_{ik}^{(\tau)}) x_k(\tau)\big)^2\Big],$$

since $e_i(t)$ is uncorrelated with $x(t-\tau)$. This error is minimized by the choice $\tilde{B}_{ik}^{(\tau)} = B_{ik}^{(\tau)}$ and the proposition follows by the uniqueness of Hilbert space projections. □

Granger causality admits a simple and intuitive characterization in autoregressive models. The following proposition simply tells us that $x_j(t)$ Granger-causes $x_i(t)$ if the LSI filter (see 2.24) $\widetilde{\mathsf{B}}_{ij}(z)$ is non-zero. We prove the proposition for autoregressive models of infinite order, the finite order case is obviously a specialization thereof.

Proposition 2.4 (Granger-causality for AR models)**.** *Suppose $x(t) \in L_2^n$ is a 0-mean stochastic process generated by the infinite autoregressive model 2.21, where the autocovariance function of $e(t)$ is given by 2.22. Let $\mathbf{H}_t$ be the Hilbert space generated by the past of $x(t)$ as in 2.5. Then, $x_j(t)$ Granger-causes $x_i(t)$ with respect to $\mathbf{H}$ if and only if $\exists \tau_0 \in \{1, 2, \ldots, p\}$ such that $B_{ij}^{(\tau_0)} \neq 0$.*

Proof. The condition for Granger-Causality for 0-mean processes $x_j \xrightarrow[\mathbf{x}]{} x_i$ is given as

$$\mathbb{E}|x_i(t) - \widehat{x}_i(t)|^2 < \mathbb{E}|x_i(t) - \widehat{\mathbb{E}}[x(t)|\mathbf{H}_t^{-j}]_i|^2$$

Using proposition 2.3 this is equivalent to

$$\mathbb{E}|x_i(t) - \sum_{\tau=1}^{\infty}\sum_{k=1}^{n} B_{ik}^{(\tau)} x_k(\tau)|^2 < \mathbb{E}|x_i(t) - \sum_{\tau=1}^{\infty}\sum_{k\neq j} B_{ik}^{(\tau)} x_k(\tau)|^2.$$

If there were no τ_0 such that $B_{ij}^{(\tau_0)} \neq 0$ then the above strict inequality would infact be an equality, a contradiction. Conversely, since $B_{ik}^{(\tau)}$ provides the best linear estimate of $x_i(t)$ from $x(t)$, if there is some τ_0 such that $B_{ij}^{(\tau_0)} \neq 0$ then above strict inequality must hold, otherwise $B_{ij}^{(\tau_0)} = 0$ would provide an equivalent or superior prediction, contradicting uniqueness or optimality. □

Finally we see that we can obtain the adjacency matrix G of the Granger-causality graph directly from the coefficient matrices as $G_{ij} = \mathbf{1}(\exists \tau \text{ s.t. } B_{ji}^{(\tau)} \neq 0)$. Alternatively, $G_{ij}^{\mathsf{T}} = \mathbf{1}(\exists \tau \text{ s.t. } B_{ij}^{(\tau)} \neq 0)$, the transpose of G, indicates "Granger-caused by" relations.

2.5 Finite Autoregressive Models

In this section we develop a few cursory properties of finite AR models which proved important in experimental simulations. Note that the dimensions and the layout of the following matrices are important to keep in mind.

The Wold decomposition theorem 2.7, and it's inverted autoregressive formulation 2.21 suggests that we model $x(t)$ as an autoregressive system.

In order to derive models that are more amenable to fitting with real data, we will restrict the autoregressive order to p. Since we know that for "most" L_2 processes, the series in 2.21 is ℓ_2 convergent, the $B(\tau)$ matrices when τ is large, have a small effect. So it is reasonable to suppose that many processes in practice can be modeled by finite autoregressions. This is common practice is the literature, and the simple autoregressive model admits a wide array of generalizations.

Our model will take the following recursive form, with boundary conditions $x_i(t) = 0 \, \forall t \leq 0$

$$x_i(t) = \sum_{j=1}^{n}\sum_{\tau=1}^{p} B_{ij}^{(\tau)} x_j(t-\tau) + e_i(t), \tag{2.23}$$

where $e_i(t)$ is called the i^{th} input, and may be random. And, $B_{ij}^{(\tau)} \in \mathbb{R}$ is the τ-lag coefficient from process j to process i.

It is natural to place all n^2 coefficients corresponding to a particular lag τ into a matrix:

$$B(\tau) = \begin{bmatrix} B_{11}^{(\tau)} & B_{12}^{(\tau)} & \dots & B_{1n}^{(\tau)} \\ B_{21}^{(\tau)} & B_{22}^{(\tau)} & \dots & B_{2n}^{(\tau)} \\ \vdots & \vdots & \ddots & \vdots \\ B_{n1}^{(\tau)} & B_{n2}^{(\tau)} & \dots & B_{nn}^{(\tau)} \end{bmatrix}$$

If one imagines stacking into the page "depth-wise" the matrices $B(\tau)$ and then looking through this stack in the ij position, they will see the column vector denoted by $\widetilde{B}_{ij}$ and we may interpret these coefficients as an LSI filter from process j to i with z-transform $\widetilde{\mathsf{B}}_{ij}(z)$, and place these coefficients into a vector denoted by:

$$\widetilde{\mathsf{B}}_{ij}(z) = \sum_{\tau=1}^{p} B_{ij}^{(\tau)} z^{-\tau} \; ; \; \widetilde{B}_{ij} = \begin{bmatrix} B_{ij}^{(1)} \\ \vdots \\ B_{ij}^{(p)} \end{bmatrix} \tag{2.24}$$

The $\widetilde{B}_{ij}$ notation is used to remind us that $\widetilde{B}_{ij}$ is a column vector rather than a single element, and to later distinguish between two matrices B and $\widetilde{B}$.

Define the n-dimensional vector processes $x(t)$ and $e(t)$ at time t as

$$x(t) = \begin{bmatrix} x_1(t) \\ \vdots \\ x_n(t) \end{bmatrix}, \; e(t) = \begin{bmatrix} e_1(t) \\ \vdots \\ e_n(t) \end{bmatrix}.$$

We can now write out the AR equation 2.23 in a more compact form

$$\begin{aligned} x(t) &= \sum_{\tau=1}^{p} B(\tau) x(t-\tau) + e(t) \\ x(t) &= 0 \; \forall t \leq 0, \end{aligned} \tag{2.25}$$

which we will work with extensively.

2.5.1 Stability

Consider a first order discrete time AR model

$$x(t) = Ax(t-1) + e(t) \; ; \; A \in \mathbb{R}^{n \times n}$$

It is well known that this system is stable if and only if the spectral radius of A is strictly less than unity: $\rho(A) \triangleq |\lambda_{\max}(A)| < 1$. Due to it's importance in this thesis, we will derive a generalization to models of order p.

To this end, define the block companion matrix $C_B \in \mathbb{R}^{np \times np}$

$$C_B = \left[\begin{array}{cccc|c} B(1) & B(2) & \dots & B(p) & \\ \hline & & & & \mathbf{0}_n^{\mathsf{T}} \\ & \mathrm{I}_{n(p-1)} & & & \vdots \\ & & & & \mathbf{0}_n^{\mathsf{T}} \end{array}\right] \tag{2.26}$$

Then we can write the p order AR system in the form of a larger first order system:

$$\begin{bmatrix} x(t) \\ x(t-1) \\ \vdots \\ x(t-p) \end{bmatrix} = C_B \begin{bmatrix} x(t-1) \\ x(t-2) \\ \vdots \\ x(t-p-1) \end{bmatrix} + \begin{bmatrix} e(t) \\ \mathbf{0}_n \\ \vdots \\ \mathbf{0}_n \end{bmatrix}$$

Hence we have stability if and only if $|\lambda(C_B)| < 1$. We establish the criteria in terms of $\mathsf{B}(z) = \sum_{\tau=1}^{p} B(\tau) z^{-\tau}$ with the following proposition

Proposition 2.5. *λ is an eigenvalue of C_B if and only if $\mathsf{det}(I_n - \mathsf{B}(\lambda)) = 0$*

Proof. (necessity) Suppose $\mathsf{det}(I_n - \mathsf{B}(\lambda)) = 0$, then $\exists v \in \mathbb{C}^{np} \setminus \{\mathbf{0}_n\}$ such that $\mathsf{B}(\lambda)v = v$. Hence

$$C_B \begin{bmatrix} \lambda^{-1}v \\ \lambda^{-2}v \\ \vdots \\ \lambda^{-p}v \end{bmatrix} = \begin{bmatrix} B(\lambda)v \\ \lambda^{-1}v \\ \vdots \\ \lambda^{-(p-1)}v \end{bmatrix} = \lambda \begin{bmatrix} \lambda^{-1}v \\ \lambda^{-2}v \\ \vdots \\ \lambda^{-p}v \end{bmatrix},$$

and λ is an eigenvalue of C_B.

(sufficiency) Suppose $\lambda \in \mathbb{C}, v \in \mathbb{C}^{np}$ forms an eigen-pair of C_B. Then $C_B v = \lambda v$, and

$$\begin{bmatrix} B^{(1)}v_1 + \ldots + B^{(p)}v_p \\ v_1 \\ \vdots \\ v_{p-1} \end{bmatrix} = \lambda \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_p \end{bmatrix},$$

where each $v_i \in \mathbb{C}^n$. From the lower $n(p-1)$ rows we see that $v_{k+1} = \lambda^{-k} v_1 1$ and combining this with the equation in the top n rows we obtain $\mathsf{B}(\lambda)v_1 = v_1$ and hence $\mathsf{det}(I_n - \mathsf{B}(\lambda)) = 0$.

□

The stability criterion is an immediate corollary.

Corollary 2.1. *The autoregressive model of order p defined in equation 2.25 is stable if and only if $\mathsf{det}(I_n - \mathsf{B}(z)) \neq 0 \, \forall z \in \mathbb{C}$ such that $|z| \geq 1$.*

Chapter 3

Granger Causality - Methods

We will proceed by briefly describing some of the classical methods of statistically testing for Granger causality in section 3.1. These methods however are not well suited when the number of processes under consideration is large. That is, the condition $T >> np$ does not hold. We will develop methods based on sparsity inducing regularization and convex optimization, which we will see leads to much more workable techniques for large n.

3.1 Classical Methods

The classical methods of inference for Granger-causality were cemented by Geweke in [10], and [11]. In the case of [10], a method for testing Granger-causality between groups of processes is developed. This approach, based on a likelihood ratio test and asymptotic theory, is adequate for pairwise testing between two processes, or groups thereof. This case is not of great importance in this work, but was an important stepping stone in the development of the classical theory.

The pairwise methods of [10] are extended in [11] to the case of conditional Granger-causality, that is, Granger-causality between two processes with respect to a group of processes, in the parlance of this thesis. It is suggested to use the periodogram estimate of $\mathbb{E}x(t)x(t-\tau)^{\mathsf{T}}$ and then to solve the Yule-Walker equations, providing estimates of the coefficients in 2.25, as well as the variance matrix of the innovations. Lacking a tractable asymptotic approach, Geweke goes on to form approximate confidence intervals for test statistics used to infer Granger-causality.

3.2 The Linear Model

The Ensemble Model Suppose we have a zero mean, wide-sense stationary stochastic processes $x(t) \in L_2^n$ generated by the $\mathsf{VAR}(p)$ model

$$x(t) = \sum_{\tau=1}^{p} B^\star(\tau)x(t-\tau) + e(t).$$

The notation $B^\star$ refers to the true parameters of the model. Consider Granger-causality with respect to the modeling space $\mathbf{H}_{t,p}$ defined in section 2.2.1.

We will seek to fit this model by calculating $\widehat{\mathbb{E}}[x(t)|\mathbf{H}_{t,p} \cap \mathbf{C}_t]$, which can be done via solving

$$\underset{\substack{B(\tau)\in\mathbb{R}^{n\times n}\\ 1\leq\tau\leq p}}{\text{minimize}} \ \mathbb{E}\ J(B(1),\ldots,B(p)), \tag{3.1}$$

where

$$J(B(1),\ldots,B(p)) = ||x(t) - \sum_{\tau=1}^{p} B(\tau)x(t-\tau)||_2^2 + \lambda\Gamma(B(1),\ldots,B(p)) \tag{3.2}$$

is the Lagrangian loss function in which Γ codifies our restriction to the set $\mathbf{C}_t$.

Remark 13. In our formulation of Granger-causality (see chapter 2), this corresponds to restricting $\mathbf{H}_{t,p}$ to the subset $\mathbf{C}_t = \{\sum_{\tau=1}^{p} B(\tau)x(t-\tau) \mid \Gamma(B) \leq \widetilde{\lambda}\} \subseteq L_2^n$, where $\widetilde{\lambda}$ depends on λ. Essentially, 3.1 is a Lagrangian reformulation of the restricted problem.

The subset $\mathbf{C}_t$ is convex whenever Γ is convex, and indeed, it is generally extremely difficult to solve 3.1 unless this is the case. The analogous application of proposition 2.4 to this model will give us Granger-causality with respect to $\mathbf{H}_{t,p} \cap \mathbf{C}_t$. That is, e.g. $x_i(t) \underset{\mathbf{H}_{t,p}\cap\mathbf{C}_t}{\longrightarrow} x_j(t)$. If $\Gamma(B^\star(1),\ldots,B^\star(p)) \leq r$, then $x_i(t) \underset{\mathbf{H}_{t,p}\cap\mathbf{C}_t}{\longrightarrow} x_j(t) \iff x_i(t) \underset{\mathbf{H}_{t,p}}{\longrightarrow} x_j(t)$.

Remark 14. The question should be raised whether or not strong duality holds for the Lagrangian relaxation applied to the calculation of $\widehat{\mathbb{E}}[x(t)|\mathbf{H}_{t,p} \cap \mathbf{C}_t]$. That is, for any $\widetilde{\lambda}$ defining the restriction $\mathbf{C}_t$, does there exist a λ such that $\widehat{x}(t)$, obtained from solving 3.1, is equal to the projection $\widehat{\mathbb{E}}[x(t)|\mathbf{H}_{t,p} \cap \mathbf{C}_t]$. In all of our cases, $\mathbf{C}_t$ is formed from restricting some norm of the coefficient matrices B, in which case $B = 0$ is an easy Slater (strictly feasible) point. The question of strong duality begs a more careful consideration in general, but we do not pursue this question further.

Remark 15. The autoregressive representation (Eq. 2.21), combined with the connection to Granger-causality in proposition 2.4, tells us that we can determine every Granger-causal relation amongst $x(t)$ by solving 3.1, and that this procedure is sufficient for any Markov WSS L_2^n process. This is a fairly wide class of processes, although there are certainly processes in practice which do not satisfy these conditions, even after significant preprocessing. Expanding the applicability of these techniques is a topic of ongoing research, and our attempt to generalize the notion of Granger-Causality to general convex projections is a small step in this direction.

The best autoregressive model in our setting is given by

$$x^\star(t) = \sum_{\tau=1}^{p} B^\star(\tau)x(t-\tau), \tag{3.3}$$

where $B^\star(\tau)$ denote the minimizers of 3.1 with $\lambda = 0$. That is, the optimal parameters for the loss function over the entire ensemble. The Granger-causality graph $G^\star$ is easily inferred from $B^\star(\tau)$ via proposition 2.4.

Compacting the notation, we can rewrite equation 3.3 as

$$x^\star(t) = (B^\star)^\mathsf{T} z(t),$$

where $B^\star = [B^\star(1)\ B^\star(2)\ ...\ B^\star(p)]^\mathsf{T} \in \mathbb{R}^{np\times n}$ vertically[1] stacks all of the matrix coefficients and $z(t) = \mathbf{vec}([x(t)\ x(t-1)\ \dots\ x(t-p+1)]) \in \mathbb{R}^{np}$ vertically lays out the vectors of samples according to their lag. The problem 3.1 becomes

$$\underset{B\in\mathbb{R}^{np\times n}}{\text{minimize}}\ \mathbb{E}||x(t) - B^\mathsf{T} z(t)||_2^2 + \Gamma(B). \tag{3.4}$$

The Population Model In practice, we will draw $T + p > p$ samples[2] $x(-p+1), x(-p+2), \dots, x(T)$ from $x(t)$. The t^{th} sample of process i will be written as $x_i(t)$, which are grouped into a column vector $x(t) = [x_1(t), ..., x_n(t)]^\mathsf{T};\ t = 1, \dots, T+p$. In this setting, it is necessary to replace ensemble averages with time averages and finite approximations. That is, equation 3.4 becomes

$$\underset{B\in\mathbb{R}^{np\times n}}{\text{minimize}}\ \frac{1}{2T}\sum_{t=1}^{T} ||x(t) - B^\mathsf{T} z(t)||_2^2 + \Gamma(B). \tag{3.5}$$

The parameters $\widehat{B}$ which solve[3] the minimization problem 3.5 are themselves random, they "inherit" randomness from $x(t)$. If we are using all of $\mathbf{H}_{t,p}$ as our modeling space (in which case $\Gamma = 0$) then unless $T >> np$, the variance of $\widehat{B}$ may be extremely large, that is, small changes in the samples of $x(t)$ can lead to wildly different sets of parameters. In the context of traditional regression problems, a regularization function $\Gamma(B)$ is added to the objective, which prevents B from being "too large". Our restriction of the modeling space with $\mathbf{C}_t$, which is equivalent to adding a regularization function, accomplishes the same goal, but we have made the connection to Granger-causality more explicit.

Remark 16. The functional form of Γ can be specified depending on what type of structure we want the resulting minimizer to take. For our purposes, Γ will always be a norm, but this need not necessarily be the case in general, however, if Γ is not at least convex, 3.5 is likely to be intractable.

Furthermore, we will generally write Γ as an abstract regularizer, the particular form of which depends on the context. We will be more specific about Γ in section 3.4.

Remark 17. A significant amount of research has gone into the study of these types of problems, see e.g. [32]. Indeed, least squares problems have been studied since the time of Gauss. More generally, the formulation 3.5 is a type of "$\mathcal{M}$-estimator" see e.g. [33].

[1] We follow this layout convention so as to later be more consistent with the literature on linear regression, which usually writes $\mathbf{X}\beta$ for a data matrix $\mathbf{X}$ and coefficients β.

[2] To avoid issues with p being less than the number of samples, we are here effictively dictating that we draw at least $T \geq 1$ additional samples.

[3] It is a question whether or not there is some $\widehat{B}$ which actually attains the minimum of 3.5. We address this issue in 3.4.2

We will continue to abbreviate our notation. Define the $T \times n$ matrix of "future" data Y, and the $T \times np$ matrix of "past" data Z:

$$Y = \begin{bmatrix} x(T)^\mathsf{T} \\ x(T-1)^\mathsf{T} \\ \dots \\ x(1)^\mathsf{T} \end{bmatrix}, Z = \begin{bmatrix} x(T-1)^\mathsf{T}\ x(T-2)^\mathsf{T}\ \dots\ x(T-p)^\mathsf{T} \\ x(T-2)^\mathsf{T}\ x(T-3)^\mathsf{T}\ \dots\ x(T-p-1)^\mathsf{T} \\ \dots \\ x(0)^\mathsf{T}\ x(-1)^\mathsf{T}\ \dots\ x(-p+1)^\mathsf{T} \end{bmatrix}, \tag{3.6}$$

These matrices finally allow us to write 3.3 and 3.5 in a standard form:

Definition 3.1 (Standard Form)**.** With the arrangement of data given in equation 3.6, we can formulate the problem 3.1 in what we will refer to as standard form, as it is the most natural arrangement of the data:

$$\underset{B \in \mathbb{R}^{np \times n}}{\text{minimize}}\ \frac{1}{2T}||Y - ZB||_F^2 + \lambda\Gamma(B). \tag{3.7}$$

We can also write the model 3.3 in a slightly different manner, which emphasizes the role of each edge-wise filter $\widetilde{B}_{ij}$ as follows:

$$\begin{aligned} \widehat{x}_i(t) &= \sum_{\tau=1}^{p}\sum_{j=1}^{n} B_{ij}^{(\tau)} x_j(t-\tau) \\ &= \sum_{j=1}^{n} \widetilde{B}_{ij}^\mathsf{T} \zeta_j(t-1) \\ &= \mathcal{Z}(t-1)\widetilde{B}_i \\ \implies \widehat{x}(t)^\mathsf{T} &= \mathcal{Z}(t-1)\widetilde{B}, \end{aligned}$$

where

$$\zeta_j(t) = [x_j(t)\ x_j(t-1)\ \dots\ x_j(t-p)]^\mathsf{T}\ (p \times 1)$$

groups lags of process j, and

$$\mathcal{Z}(t) = [\zeta_1(t)^\mathsf{T}\ \dots\ \zeta_n(t)^\mathsf{T}]\ (1 \times np)$$

horizontally stacks each of these groups, and $\widetilde{B}_i,\ (np \times 1)$ and $\widetilde{B},\ (np \times n)$ are given by (recall also 2.24):

$$\widetilde{B}_i = \begin{bmatrix} \widetilde{B}_{i1} \\ \widetilde{B}_{i2} \\ \vdots \\ \widetilde{B}_{in} \end{bmatrix},\ \widetilde{B} = \begin{bmatrix} \widetilde{B}_1 & \widetilde{B}_2 & \dots & \widetilde{B}_n, \end{bmatrix} \tag{3.8}$$

which organizes all of the filter coefficients. The tilde on top of the B matrix indicates this rearrangement. Using these arrangements of data, we have the "alternate form" version of 3.1.

Definition 3.2 (Alternate Form)**.** Given the matrix of data Y from 3.6, and defining the $T \times np$ matrix $\mathcal{Z}$

$$\mathcal{Z} = \begin{bmatrix} \mathcal{Z}(T-1) \\ \mathcal{Z}(T-2) \\ \vdots \\ \mathcal{Z}(0) \end{bmatrix} \tag{3.9}$$

we have the optimization problem

$$\underset{\widetilde{B}\in\mathbb{R}^{np\times n}}{\text{minimize}} \frac{1}{2T}||Y - \mathcal{Z}\widetilde{B}||_F^2 + \lambda\Gamma(\widetilde{B}). \tag{3.10}$$

Remark 18. The differences between definition 3.1 and 3.2 are simply trivial rearrangements, but these rearrangements have a significant impact on the interpretations of the matrices B and $\widetilde{B}$, as well as on the functional form of gradients etc... in the following.

Remark 19. The Γ function appearing in equation 3.7 as well as in 3.10 need not be the same function. At this point, it stands in only as a generic "placeholder" for a regularization function, but we will use the same symbol to a particular function after definition 3.3.

3.3 Classical Approaches to the Linear Model

3.3.1 Ordinary Least Squares

The ordinary least squares (OLS) solution, corresponding to $\Gamma(B) = 0$ in equation (3.7) dates back to the beginning of the 19th century. The method is commonly attributed to Gauss, although Legendre was the first to publish the method [34]. In any case, the solution is well known, assuming that Z is full rank:

$$B_{OLS} = \Big(\frac{1}{T}Z^\mathsf{T}Z\Big)^{-1}\Big(\frac{1}{T}Z^\mathsf{T}Y\Big) \tag{3.11}$$

$$= \Big(\frac{1}{T}\sum_{t=1}^{T} z(t)z(t)^\mathsf{T}\Big)^{-1}\Big(\frac{1}{T}\sum_{t=1}^{T} z(t)x(t)^\mathsf{T}\Big) \tag{3.12}$$

Remark 20. It is important to point out that when writing down the ordinary least squares solution we are really viewing $x(t)$ as a deterministic sequence. In some sense this is perfectly correct since we have observed the sequence and seek to find a model that explains it. On the other hand, the advantage of viewing $x(t)$ as a stochastic process is that we may attempt to design a model that makes sense for the entire ensemble of possible realizations of $x(t)$. If we suppose that the noise sequence $e(t)$ in 2.21 is not merely white, but is in fact Gaussian, then the maximum likelihood estimate of B corresponds exactly to the OLS solution. Without the Gaussian supposition, the stochastic models rapidly become intractable.

Alternatively, if we restrict ourselves to the use of second order statistics, it is possible to specify conditions under which equation 3.11 will converge to the Linear Minimum Mean Square Error estimator $B_{LMMSE} = \Sigma_z^{-1}\Sigma_{zx}$ as $T \to \infty$. Hence, working in the context of OLS is a reasonable approach.

Remark 21. It is easy to check that these matrices are all correctly specified by noting the orthogonality condition $(Y - ZB) \perp Z \implies B = \Sigma_z^{-1}\Sigma_{zx}$, and then verifying that the dimensions line up.

The following results are vector generalizations of standard results for OLS in the scalar case. Asymptotic hypothesis testing corresponds essentially to the "classical" approach to Granger-causality. In order to obtain an asymptotic distribution for least squares, we need a substantial number of nontrivial conditions to be met.

OLS Asymptotics Conditions

1. Each process is WSS, 0 mean and $L_2^n(\mathbb{Z})$ (L_2 Processes)
2. $\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{\infty} z(t)z(t)^\mathsf{T} = \Sigma_z$. In $\mathbb{P}$ or a.s. (LLN 1)
3. $\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{\infty} x(t)z(t)^\mathsf{T} = \Sigma_{xz}$. In $\mathbb{P}$ or a.s. (LLN 2)
4. $\exists B, e(t)$ s.t. $x(t) = Bz(t) + e(t)$ and $\mathbb{E}[e(t)z(t)^\mathsf{T}] = 0$ (Correct Model)
5. $\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{\infty} e(t)z(t)^\mathsf{T} = 0$ (LLN 3)
6. $\exists R$ s.t. $R^{-1/2} \lim_{T\to\infty} \frac{1}{\sqrt{T}} \sum_{t=1}^{T} (\overrightarrow{e(t)z(t)^\mathsf{T}}) \sim \mathcal{N}(0, I_{n^2p})$ (Residual CLT)
7. $R(\tau) = \mathbb{E}[(\overrightarrow{e(t)z(t)^\mathsf{T}})(\overrightarrow{e(t-\tau)z(t-\tau)^\mathsf{T}})^\mathsf{T}] = \begin{cases} 0, \tau > 0 \\ R, \tau = 0 \end{cases}$ (Consistent R_T)
8. $\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{\infty} (\overrightarrow{e(t)z(t)})(\overrightarrow{e(t)z(t)^\mathsf{T}})^\mathsf{T} = R$ (LLN 4)
9. $\mathbb{E}[(\overrightarrow{z(t)z(t)^\mathsf{T}})(\overrightarrow{z(t)z(t)^\mathsf{T}})^\mathsf{T}] < \infty$ (Bounded 4^{th} Moments)

Where we use the arrow notation as a short hand for **vec** as in $\overrightarrow{x} \triangleq \mathbf{vec}(x)$.

The conditions (LLN 1), (LLN 2), (LLN 3), and (LLN 4) are standard laws of large numbers for $\Sigma_z \triangleq \mathbb{E}[z(t)z(t)^\mathsf{T}]$, $\Sigma_{xz} \triangleq \mathbb{E}[x(t)z(t)^\mathsf{T}]$, $\mathbb{E}[e(t)z(t)^\mathsf{T}]$ and R.

The condition (Correct Model) is the most difficult to verify, or the most dishonest to simply assume. The intuition is that our data must truly come from an autoregressive model. This condition is violated for example if the data generating process is nonlinear, time varying, has a moving average component, or has a model order other than p. Tests exist for attempting to check these conditions in practice [27], [35].

The condition (Residual CLT) is a central limit theorem for the process $e(t)z(t)^\mathsf{T}$ where $R \triangleq \mathbb{E}[(\overrightarrow{e(t)z(t)^\mathsf{T}})(\overrightarrow{e(t)z(t)^\mathsf{T}})^\mathsf{T}]$.

Finally, the conditions (Consistent R_T) and (Bounded 4^{th} Moments) are necessary to obtain consistent estimates of the R matrix in (Residual CLT). The supposition that $e(t)$ is a martingale difference sequence with respect to $z(t)$ and $x(t-1)$ is sufficient for (Consistent R_T).

OLS Asymptotics

Suppose throughout that (L_2 Processes) holds, and that anything that needs to be inverted can be. Now,

$$\begin{aligned} B_{OLS} &= \Big(\frac{1}{T}\sum_{t=1}^{T} x(t)z(t)^{\mathsf{T}}\Big)\Big(\frac{1}{T}\sum_{t=1}^{T} z(t)z(t)^{\mathsf{T}}\Big)^{-1} \xrightarrow{(a)} B_{LMMSE} \\ &\overset{(b)}{=} \Big(\frac{1}{T}\sum_{t=1}^{T} (Bz(t)+\epsilon(t))z(t)^{\mathsf{T}}\Big)\Big(\frac{1}{T}\sum_{t=1}^{T} z(t)z(t)^{\mathsf{T}}\Big)^{-1} \\ &\overset{(c)}{=} B + \Big(\frac{1}{T}\sum_{t=1}^{T} e(t)z(t)^{\mathsf{T}}\Big)\Big(\frac{1}{T}\sum_{t=1}^{T} z(t)z(t)^{\mathsf{T}}\Big)^{-1} \\ &\overset{(d)}{\Longrightarrow} \mathbb{E}[B_{OLS}] = \mathbb{E}[B]. \end{aligned}$$

Where (a) follows given (LLN 1), (LLN 2), and an application of Slutsky's theorem. In equality (b) $\epsilon(t)$ is the error between $x(t)$ and $Bz(t)$ and given (Correct Model) we have $\epsilon(t) = e(t)$ which is assumed for (c), and allows us to conclude the consistency (d) since $\mathbb{E}[e(t)z(t)^{\mathsf{T}}] = 0$. The consistency of B_{OLS} to some "true" underlying B is critical for hypothesis testing.

Continuing the calculations

$$\begin{aligned} B_{OLS} - B &= \Big(\frac{1}{T}\sum_{t=1}^{T} e(t)z(t)^{\mathsf{T}}\Big)\Big(\frac{1}{T}\sum_{t=1}^{T} z(t)z(t)^{\mathsf{T}}\Big)^{-1} \\ \overset{(a)}{\Longrightarrow} \sqrt{T}(\overrightarrow{B}_{OLS} - \overrightarrow{B}) &= \Big[\Big(\frac{1}{T}\sum_{t=1}^{T} z(t)z(t)^{\mathsf{T}}\Big)^{-1} \otimes I_n\Big] R^{1/2}\Big[\frac{1}{\sqrt{T}} R^{-1/2} \sum_{t=1}^{T} \overrightarrow{e(t)z(t)^{\mathsf{T}}}\Big] \\ &\overset{(b)}{\Longrightarrow} \sqrt{T}\Big[\Sigma_z \otimes I_n\Big] R^{-1/2}(\overrightarrow{B}_{OLS} - \overrightarrow{B}) \to \mathcal{N}(0, I_{n^2p}). \end{aligned} \tag{3.13}$$

Where (b) is conditional upon (Residual CLT) and we have applied the identity $\mathbf{vec}(ABC) = (C^{\mathsf{T}} \otimes A)\mathbf{vec}(B)$ in (a). Since we do not have direct access to Σ_z, it must be estimated using (LLN 1):

$$\widehat{\Sigma}_z = \frac{1}{T}\sum_{t=1}^{T} z(t)z(t)^{\mathsf{T}}.$$

The covariance matrix R is also unavailable to us, and it is rather more difficult to estimate. We will employ estimates of $e(t)$, denoted $\widehat{e}(t)$ to do so. Let

$$\widehat{e}(t) = x(t) - \widehat{x}(t) = e(t) - (B_{OLS} - B)z(t).$$

Then,

$$\begin{aligned}
\widehat{R}_T &= \frac{1}{T}\sum_{t=1}^{T}(\overrightarrow{\widehat{e}(t)z(t)^\intercal})(\overrightarrow{\widehat{e}(t)z(t)^\intercal})^\intercal \\
&= \frac{1}{T}\sum_{t=1}^{T}\Big[(\overrightarrow{e(t)z(t)^\intercal})(\overrightarrow{e(t)z(t)^\intercal})^\intercal - (\overrightarrow{e(t)z(t)^\intercal})(\overrightarrow{(B_{OLS}-B)z(t)z(t)^\intercal})^\intercal \\
&- (\overrightarrow{(B_{OLS}-B)z(t)z(t)^\intercal})(\overrightarrow{e(t)z(t)^\intercal})^\intercal + (\overrightarrow{(B_{OLS}-B)z(t)z(t)^\intercal})(\overrightarrow{(B_{OLS}-B)z(t)z(t)^\intercal})^\intercal\Big] \\
&\overset{(a)}{\to} R.
\end{aligned}$$

Where we get the limit (a) by applying (LLN 4) to the first term (which converges to R), (LLN 3), (LLN 4), and Slutsky's theorem to the middle two terms (converging to 0), and finally the (Bounded 4^{th} Moments) to the last term, which hence will go to 0.

Finally, substituting this estimate in for R in equation 3.13 we obtain:

$$\begin{aligned}
\sqrt{T}\Big[\Big(\frac{1}{T}\sum_{t=1}^{T}z(t)z(t)^\intercal\Big)^{-1}\otimes I_n\Big]\Big[\frac{1}{T}\sum_{t=1}^{T}\big(\overrightarrow{\widehat{e}(t)z(t)^\intercal}\big)\big(\overrightarrow{\widehat{e}(t)z(t)^\intercal}\big)^\intercal\Big]^{-1/2}(\overrightarrow{B}_{OLS}-\overrightarrow{B}) \\
\longrightarrow \mathcal{N}(0, I_{n^2p}) \text{ as } T\to\infty.
\end{aligned} \tag{3.14}$$

From this asymptotic distribution, it is possible to form hypothesis tests for the Granger-causality adjacency matrix G. However, two huge issues with this approach are immediately apparent:

1. The estimate B_{OLS} has very high variance unless $T >> np$ (it is not stable).
2. The necessary conditions are extensive

This straightforward approach is not at all appropriate for estimating causality graphs with a large number of nodes.

Tikhonov Regularization

The standard approach to deal with the issue of stability for B_{OLS} is referred to as either Tikhonov regularization, or ridge regression (we abbreviate as "LST" for "least squares Tikhonov"). In this approach we use the squared 2-norm as a regularizer in 3.5, that is, use $\Gamma(B) = ||B||_F^2$. This discourages the B matrix from being "too large", and again the solution is well known:

$$B_{LST}^{\lambda} = \big(Z^\intercal Z + \lambda I\big)^{-1}\big(Z^\intercal Y\big)$$

While this modification enables us to make stable estimates of model parameters from observed data, the estimate is biased: $\mathbb{E}[B_{LST}^{\lambda}] \neq B$ (unless $\lambda = 0$), and hence a lot of standard hypothesis testing machinery is invalid. Tikhonov regularization is an highly applicable technique for regression problems in particular when the end goal is forecasting. In the context of Granger-causality however, since B_{LST}^{λ} is entirely dense, an additional step is required to choose how to infer the presence or absence of edges in the Granger-causality graph. We next explore some alternatives, in which the inferred model is naturally sparse.

3.3.2 The LASSO

Take an example of economic forecasting. We may be interested in, for example, making predictions about the price of corn at some time in the future given data about interest rates, employment, the price of oil, sunspot activity, and any number of other features. If we are so industrious as to collect a great number of features, say, n of them, then it is almost certain that some of these features will have no value whatsoever for the task of predicting corn prices. In this case, it would be naturally desirable to narrow down the number of features we use to the ones that enable us to make the most accurate predictions, or at least to provide a parsimonious model. This is the problem of "best subset selection".

A rather straightforward approach to the best subset selection problem is to fit all of the models that are possible given the features we observe, and then use some criteria to choose which one is best. If we apply some basic combinatorial reasoning, it becomes clear that there are in fact 2^n such models, and even for a modestly sized n, to attempt to examine all of them would rapidly exhaust the patience of even the most diligent of economists.

Example 6 (Best Subset Selection)**.** In our context, we could attempt to formulate the best subset selection problem by restricting our modeling space to $\mathbf{H}_{t,p} \cap \mathbf{C}_{t,r}$, where

$$\mathbf{C}_{t,r} = \{\sum_{\tau=1}^{p} B(\tau)x(t-\tau) \mid \sum_{\tau=1}^{p} \gamma_0(B(\tau)) \leq r\},$$

and $\gamma_0(B)$ is the number of nonzero coefficients of the matrix B. The corresponding (Lagrangian) formulation uses $\Gamma(B) = \gamma_0(B)$ which effectively codifies our desire for more or less sparse coefficient matrices. Unfortunately, $\mathbf{C}_{t,r}$ is not convex, and best subset selection is almost entirely intractable for even a modest number of parameters.

Sparsity inducing regularization is an approximation to the best subset selection problem for fitting regression models in which we make the a priori supposition that many of the entries in the coefficient matrices are 0. It is known that $\Gamma_{LASSO}(B) = ||B||_1 \triangleq \sum_{i,j,\tau}^{n,n,p} |B_{ij}^{(\tau)}|$ results in the minima of the program having many entries exactly equal to 0, as seen in the following example. This technique is referred to as LASSO (Least Aboslute Shrinkage and Selection Operator) regression, and has been used at least since [36]. Related theoretical properties of sparsity were studied in the landmark paper of Candés [37]. See also the recent book [38].

Example 7 (LASSO)**.** Consider an ordinary regression problem where we are given a set of input/output pairs (not necessarily time series) $\{(y^{(t)}, x^{(t)}) \in \mathbb{R} \times \mathbb{R}^n, t = 1, 2, \ldots, T\}$, and we want to fit the model $\widehat{y} = x^{\mathsf{T}}\beta$. Let $\mathbf{y}$ and $\mathbf{X}$ be the natural arrangements of the data so that the i^{th} row of $\mathbf{X}$ corresponds to samples $x^{(i)}$. If $\mathbf{X}$ is full rank, and $T \geq n$, the OLS approach is to use $\beta_{OLS} = (\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}y$. The LASSO problem is to instead solve the convex program

$$\underset{\beta \in \mathbb{R}^n}{\text{minimize}} \ \frac{1}{2T}||\mathbf{y} - \mathbf{X}\beta||_2 + \lambda||\beta||_1, \tag{3.15}$$

using the 1-norm regularizer $||\beta||_1 = \sum_{i=1}^{n} |\beta_i|$. If we consider the function $\gamma_q(\beta) = \sum_{i=1}^{n} |\beta_i|^q$, the smallest value of q such that γ_q is convex, is $q = 1$, in which

case $\gamma_1(\beta) = ||\beta||_1$. From this perspective, the LASSO problem 3.15 is a convex relaxation of the best subset selection problem.

Like best subset selection, it can be shown that we obtain a sparse solution from solving 3.15; many of the entries of $\widehat{\beta}$ are 0. Furthermore, this convex program still admits a unique solution (under mild conditions) even when $n >> T$, a case in which the OLS solution does not even exist. A particular and illustrative special case is easy to analyze, as seen in the next example.

Example 8 (Orthgonal Design LASSO ([38]))**.** Consider the setting of the previous example. If the design matrix $\mathbf{X}$ is orthogonal and normalized such that $\frac{1}{T}\mathbf{X}^\mathsf{T}\mathbf{X} = I$, then we can obtain an illustrative closed form solution.

We wish to minimize the objective function $J(\beta) = \frac{1}{2T}||\mathbf{y} - \mathbf{X}\beta||_2^2 + \lambda||\beta||_1$. To this end, consider the subdifferential (see definition 3.4) of L:

$$\partial J(\beta) = -\frac{1}{T}\mathbf{X}^\mathsf{T}(\mathbf{y} - \mathbf{X}\beta) + \lambda\partial||\beta||_1. \tag{3.16}$$

The vector $\widehat{\beta}$ is a minimizer of J if and only if $0 \in \partial J(\widehat{\beta}) \iff \frac{1}{T}\mathbf{X}^\mathsf{T}\mathbf{y} - \widehat{\beta} \in \lambda\partial||\widehat{\beta}||_1$. If we then apply the normalization $\frac{1}{T}\mathbf{X}^\mathsf{T}\mathbf{X} = I$, and consider the j^{th} component individually this is equivalent to:

$$\frac{1}{T}\mathbf{x}_j^\mathsf{T}\mathbf{y} - \widehat{\beta} \in \lambda\partial|\widehat{\beta}_j|, \forall 1 \leq j \leq n,$$

where $\partial|\widehat{\beta}_j|$ is given in example 9. Hence, we have

$$\widehat{\beta}_j = 0 \iff \frac{1}{T}\mathbf{x}_j^\mathsf{T}\mathbf{y} \in [-\lambda, \lambda] \iff |\frac{1}{T}\mathbf{x}_j^\mathsf{T}\mathbf{y}| \leq \lambda.$$

And, in the case where $\widehat{\beta}_j \neq 0$, $\widehat{\beta}_j = \frac{1}{T}\mathbf{x}_j^\mathsf{T}\mathbf{y} - \lambda\mathsf{sgn}(\beta_j)$. Reasoning by cases for $\frac{1}{T}\mathbf{x}_j^\mathsf{T}\mathbf{y} > \lambda$ and $\frac{1}{T}\mathbf{x}_j^\mathsf{T}\mathbf{y} < -\lambda$ we obtain:

$$\widehat{\beta}_j = \begin{cases} \frac{1}{T}\mathbf{x}_j^\mathsf{T}\mathbf{y} - \lambda, & \frac{1}{T}\mathbf{x}_j^\mathsf{T}\mathbf{y} > \lambda \\ 0, & \frac{1}{T}\mathbf{x}_j^\mathsf{T}\mathbf{y} \in [-\lambda, \lambda] \\ \frac{1}{T}\mathbf{x}_j^\mathsf{T}\mathbf{y} + \lambda, & \frac{1}{T}\mathbf{x}_j^\mathsf{T}\mathbf{y} < -\lambda \end{cases},$$

or, $\widehat{\beta}_j = \mathsf{sgn}(\mathbf{x}_j^\mathsf{T}\mathbf{y})(|\frac{1}{T}\mathbf{x}_j^\mathsf{T}\mathbf{y}| - \lambda)_+$, which we finally write in vector form as

$$\widehat{\beta} = \mathcal{S}_\lambda(\frac{1}{T}\mathbf{X}^\mathsf{T}\mathbf{y}), \tag{3.17}$$

an operation called "soft thresholding".

Remark 22. The above example is illustrative, as in our case the data (design) matrix need not be orthogonal, and we are working with a matrix of variables, rather than a vector. This latter difference is essentially trivial since $XB = (I \otimes X)\mathbf{vec}B$, the LASSO formulation applies equally as well to a matrix variable as a vector. The issue with non orthogonal data matrix is more significant as there is no longer a closed form solution, but fast iterative algorithms can be derived to solve 3.15, and it can be shown that the resulting minimizer is sparse.

Remark 23. The sparsity inducing LASSO is attractive for detecting potential Granger-causal relations amongst a large number of processes when we do not have enough data to reasonably apply asymptotic estimation methods, since the LASSO automatically selects a subset of variables.

3.4 Depth Wise Grouped LASSO (DWGLASSO)

The group LASSO (GLASSO) is a method for encouraging simultaneous sparsity of groups of variables. This is in contrast to the overall "random" sparsity pattern of the classical LASSO. We use a norm

$$\Gamma_{GLASSO}(B) = \sum_{g \in \mathbf{G}} ||B_g||_2$$

where $\mathbf{G}$ specifies a set of groups, and $B_g \in \mathbb{R}^{|g|}$ is a vector constructed from B_{ij} according to $(i,j) \in g$. This approach was first proposed in [39], and the un-squared 2 norm on the groups induces group-wide sparsity where every entry B_{ij} having $(i,j) \in g$ is simultaneously set to 0. Depending on the grouping structure, this can be simple (if we group only on the rows or columns of B) or rather complicated if we want to pick the groups from arbitrary (or at least less-structured) locations in B.

For application to causality graphs, it makes sense to form the groups from edges in the graph. When we interpret an edge from process j to process i as a filter

$$\mathsf{B}_{ij}(z) = \sum_{\tau=1}^{p} B_{ij}^{(\tau)} z^{-\tau},$$

and we want to encourage either $\mathsf{B}_{ij}(z) = 0$ or $B_{ij}^{(\tau)} \neq 0$, $\tau = 1, 2, ..., p$ so that in the former case we conclude that there is no causal relation.

Recall the vector $\widetilde{B}_{ij} = [B_{ij}^{(1)}, ..., B_{ij}^{(p)}]^{\mathsf{T}}$ specifying the coefficients of the edgewise filters, (2.24, 3.8). With this notation we define a particular regularization function. The same regularization function was also employed by [40].

Definition 3.3 (Depth Wide Group LASSO Regularizer)**.** We define the DW-GLASSO regularizer as follows:

$$\Gamma_{DW}(B) = \sum_{i=1}^{n} \sum_{j=1}^{n} ||\widetilde{B}_{ij}||_2. \tag{3.18}$$

which is a group LASSO penalty function. That is, an L_1 norm of the L_2 norms. This is sometimes written $||\widetilde{B}||_{1,2}$.

Remark 24. Since we employ different rearrangements of the autoregressive coefficients, e.g. as B in the standard form 3.7 and as $\widetilde{B}$ in the alternate form 3.10, we remark that the definition of Γ_{DW} is to always sum the coefficients of the edge wise filters as in equation (3.18), regardless of whether we write $\Gamma_{DW}(B)$ or $\Gamma_{DW}(\widetilde{B})$. The subdifferential (see section 3.4.1) hence needs to be interpreted correctly in context, the matrix $\Phi \in \partial\Gamma_{DW}(B)$ is different than $\Phi \in \partial\Gamma_{DW}(\widetilde{B})$ in it's layout.

Finally, we will tend to simply write Γ, rather than Γ_{DW} hereafter.

Remark 25. It is natural to think of the matrix B as being 3-dimensional with the $B^{(\tau)}$ matrices being stacked depth-wise. The vector $\widetilde{B}_{ij}$ can then be viewed as extending into the page at location (i,j) of B. For this reason, we refer to this particular grouping structure as "depth-wise" and hence refer to 3.5 with $\Gamma \triangleq \Gamma_{DW}$ as the Depth-Wise Group LASSO (DWGLASSO) problem.

3.4.1 Properties of Γ_{DW}

In this section we derive some fundamental properties of Γ_{DW} including the dual norm, the subdifferential, and induced matrix norms. We will work with Γ as a function $\Gamma : \mathbb{R}^{np} \to \mathbb{R}$ where we have n groups of size p, as this elucidates the fundamental properties. Modification to any particular arrangement of matrices in the input space is straightforward, and again, must be viewed in context.

Define

$$\Gamma(x) = \sum_{j=1}^{n} ||\bar{x}_j||_2, \tag{3.19}$$

where $\bar{x}_j = [x_{(j-1)p+1}\ x_{(j-1)p+2}\ \dots x_{jp}]^\mathsf{T} \in \mathbb{R}^p$ denotes for $1 \leq j \leq n$ the n groups of size p from the vector x.

It is immediate to verify that Γ is a norm on $\mathbb{R}^{np}$. We have $\Gamma \geq 0$, the triangle inequality $\Gamma(x+y) \leq \Gamma(x)+\Gamma(y)$, and $\Gamma(\alpha x) = |\alpha|\Gamma(x)$ from the analogous properties of $||\cdot||_2$. The point separation property $\Gamma(x) = 0 \iff x = 0$ follows as well, but we stress that this is the case only if all of the n groups are included in Γ.

From here it also follows that Γ is convex, as this is a property true of any norm.

Subdifferential

In this section, the subdifferential of Γ, denoted by $\partial\Gamma$, will be viewed as a set valued mapping from $\mathbb{R}^{np}$ into the power set thereof. This is a generic arrangement of the matrix entries which serves to elucidate the structure of the subdifferential, recalling the discussion in definition 3.3, the arrangement must be interpreted in context.

The subgradient is a generalization of the gradient to non-differentiable functions and has important properties and applications in convex analysis [41] [20].

Definition 3.4 (Subdifferential)**.** For a function $f : \mathbb{R}^n \to \mathbb{R}$ (which need not be differentiable), the subdifferential $\partial f : \mathbb{R}^n \to \mathcal{P}(\mathbb{R}^n)$ is defined as

$$\partial f(x) = \{\phi \in \mathbb{R}^n \mid f(y) \geq f(x) + \phi^\mathsf{T}(y - x),\ \forall y \in \textbf{dom}\ f\}, \tag{3.20}$$

which gives the normal vectors for hyperplanes supporting the epigraph of f. Each element of this set is called a subgradient.

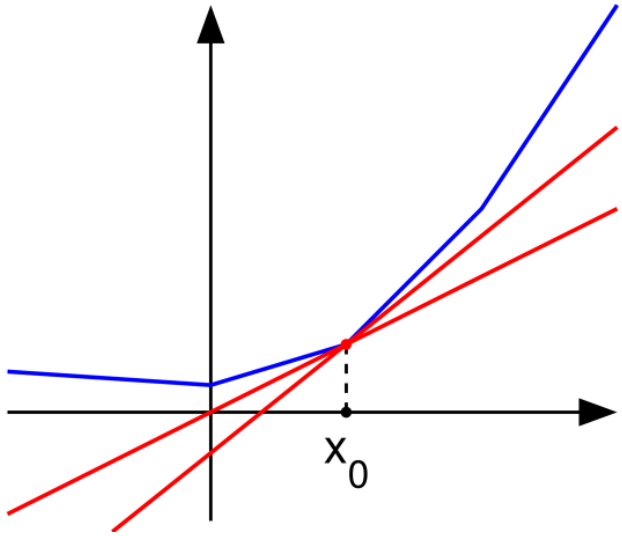
Essentially, the subdifferential of a function at a point x is defined through the set of all affine minorants at that point. See figure 3.1

Theorem 3.1 (Subdifferentials of Convex Functions [20])**.** *For any convex function $f : \mathbb{R}^n \to \mathbb{R}$, the subdifferential $\partial f(x)$ is a non-empty, compact, and convex subset of $\mathbb{R}^n$, for every $x \in$ **dom** f. Furthermore, at points where f is differentiable we have $\partial f(x) = \{\nabla f(x)\}$, that is, the only subgradient is the gradient.*

Example 9 (Absolute value)**.** The most natural example of a subgradient is furnished by the absolute value function $f(x) = |x|$, whose subdifferential is easily seen to be

$$\partial|x| = \begin{cases} \{-1\}, & x < 0 \\ [-1, 1], & x = 0 \\ \{1\}, & x > 0 \end{cases}$$

Figure 3.1: A convex function with non-differentiable "kinks". Examples of subgradients at x_0 are shown in red.



In general, characterizing the subdifferential of a convex function is very difficult and one must often be satisfied with obtaining just a single subgradient. In our case however, the non-overlapping group structure makes it relatively easy to obtain a complete characterization of $\partial\Gamma$. To this end, define the finite family of n functions $\eta_j : \mathbb{R}^{np} \to \mathbb{R}$ by $\eta_j(x) = ||\bar{x}_j||_2$, for $1 \leq j \leq n$. Note that each η_j is a semi-norm, as the point separation property $\eta_j(x) = 0 \iff x = 0$ is not satisfied. Using the linearity property for subdifferentials $\partial\Gamma(x) = \sum_{i=1}^{n} \partial\eta_j(x)$, we can obtain the subdifferential of Γ.

Proposition 3.1 (Subdifferential $\partial\Gamma$)**.** *We have $\phi \in \partial\Gamma(x)$ if and only if*

$$\bar{\phi}_j \in \begin{cases} \overline{\mathbb{B}}_p(0;1), & \bar{x}_j = 0 \\ \bar{x}_j/||\bar{x}_j||_2, & \textit{otherwise} \end{cases}$$

$\forall\ 1 \leq j \leq n$, and $\phi = [\bar{\phi}_1\ \ldots\ \bar{\phi}_n]^\mathsf{T}$. The notation $\overline{\mathbb{B}}_p(c;r)$ denotes the closed euclidean ball in $\mathbb{R}^p$ centered at c and with radius r.

Proof. From the definition, we can see that the subdifferential of η_j can be obtained group-component-wise as

$$\partial\eta_j(x) = [\overline{\eta_j(x)}_1\ \ldots\ \overline{\eta_j(x)}_n]^\mathsf{T},$$

where $\overline{\eta_j(x)}_i = 0$ whenever $i \neq j$, since any nonzero y that is zero inside group j will have $\eta_j(y) = 0$ and the freedom to choose y in the definition can easily lead to contradicting the inequality in the definition, unless ϕ is zero outside of group j. We then have

$$\begin{aligned} \overline{(\partial\eta_j(x))}_j &= \{\phi \in \mathbb{R}^p \mid \eta_j(y) \geq \eta_j(x) + \phi^\mathsf{T}(y - \bar{x}_j),\ \forall y \in \mathbb{R}^p\} \\ &= \{\phi \in \mathbb{R}^p \mid ||y||_2 \geq ||\bar{x}_j||_2 + \phi^\mathsf{T}(y - \bar{x}_j),\ \forall y \in \mathbb{R}^p\Big\} \\ &\overset{(a)}{=} \begin{cases} \overline{\mathbb{B}}_p(0;1),\ \bar{x}_j = 0 \\ \frac{\bar{x}_j}{||\bar{x}_j||_2},\ \text{otherwise} \end{cases}, \end{aligned}$$

where the case $\bar{x}_j = 0$ in (a) follows because for any ϕ with $||\phi|| > 1$, we could choose $y = \phi$ and obtain the contradiction $||\phi||_2 \geq ||\phi||_2^2$, while the case for $\bar{x}_j \neq 0$ is obtained by differentiation since $||\cdot||_2$, is differentiable as long as the argument is not zero.

Taking the minkowski sum of $\partial\eta_j(x)$ gives the result. □

A second useful property, which in fact holds for norms in general by Fenchel's inequality, is as follows.

Proposition 3.2. *For any $\phi \in \partial\Gamma(x)$ we have $\phi^\mathsf{T} x = \Gamma(x)$.*

Proof. We have $\phi^\mathsf{T} x = \sum_{j=1}^n \bar{\phi}_j^\mathsf{T} \bar{x}_j$. Now, if $\bar{x}_j = 0$ then clearly $\bar{\phi}_j^\mathsf{T} \bar{x}_j = 0$. On the other hand, if $\bar{x}_j \neq 0$ then $\bar{\phi}_j = \bar{x}_j / ||\bar{x}_j||_2$ and hence $\bar{\phi}_j^\mathsf{T} \bar{x}_j = ||\bar{x}_j||_2$. Taking the sum over j completes the proof. □

Dual norm $\Gamma_\star$ of Γ

The dual norm is an important concept in analysis which we will draw on later.

Definition 3.5 (Dual norm)**.** The dual norm $||\cdot||_\star$ of a norm $||\cdot||$ is given by

$$||y||_\star = \sup_{||z|| \leq 1} z^\mathsf{T} y. \tag{3.21}$$

Again, it may in general be rather difficult to evaluate this quantity for any arbitrary norm $||\cdot||$, but the case of Γ is tractable.

Proposition 3.3 (Dual norm $\Gamma_\star$)**.** *The dual norm of Γ is given by*

$$\Gamma_\star(y) = \max_{1 \leq j \leq n} ||\bar{y}_j||_2 \tag{3.22}$$

Proof. First, since we are working in a finite dimensional space, the set of all z such that $\Gamma(z) \leq 1$ is compact and secondly, the function Γ is continuous. We can hence apply the extreme value theorem and write $\Gamma_\star(y) = \max_{\Gamma(z) \leq 1} z^\mathsf{T} y$.

Now consider the following inequality,

$$\begin{aligned}
\Gamma_\star(y) &= \max_{\Gamma(z) \leq 1} z^\mathsf{T} y \\
&= \max_{\Gamma(z) \leq 1} \sum_{i=1}^n \bar{z}_i^\mathsf{T} \bar{y}_i \\
&\overset{(a)}{\leq} \max_{\Gamma(z) \leq 1} \sum_{i=1}^n ||\bar{z}_i||_2 ||\bar{y}_i||_2 \\
&\leq \max_{\Gamma(z) \leq 1} \max_{1 \leq j \leq n} ||\bar{y}_j||_2 \sum_{i=1}^n ||\bar{z}_i||_2 \\
&= \max_{\Gamma(z) \leq 1} \max_{1 \leq j \leq n} ||\bar{y}_j||_2 \Gamma(z) \\
&\overset{(b)}{\leq} \max_{1 \leq j \leq n} ||\bar{y}_j||_2,
\end{aligned}$$

where (a) follows by the Cauchy-Schwarz inequality and (b) is by $\Gamma(z) \leq 1$. This inequality can be achieved by setting $\bar{z}_{j^*} = \bar{y}_{j^*} / ||\bar{y}_{j^*}||_2$ where j^* is the index that achieves $\max_{1 \leq j \leq n} ||\bar{y}_j||_2$, and $\bar{z}_i = 0$ for every other index. □

Proposition 3.4. *For any $\phi \in \partial\Gamma(x)$ we have $\Gamma_\star(\phi) \leq 1$.*

Proof. This is immediate from equation 3.22 and proposition 3.1. □

Induced Matrix Norms

The concept of a matrix (or more generally operator) norm induced by another norm is important in order to bound the possible value of expressions taking the form $||Ax||$.

Definition 3.6 (Induced Norm)**.** For a matrix $A \in \mathbb{R}^{n\times m}$, the norm $||A||_{\alpha,\beta}$ induced by the norms $||\cdot||_\alpha$ and $||\cdot||_\beta$ is defined as

$$||A||_{\alpha,\beta} = \sup_{||x||_\alpha \leq 1} ||Ax||_\beta. \tag{3.23}$$

We abbreviate $||A||_{\alpha,\alpha}$ by $||A||_\alpha$.

For our purposes, we are interested in $||\cdot||_\Gamma$ and $||\cdot||_{\Gamma_\star}$. In particular, since (from proposition 3.4) we have $\Gamma_\star(\phi) \leq 1$ for any $\phi \in \partial\Gamma(x)$, the term $||A||_{\Gamma_\star}$ can serve as a uniform bound on $\Gamma_\star(A\phi)$.

It is necessary first to establish some additional notation. For a matrix $A \in \mathbb{R}^{np\times mp}$ we will denote by $\bar{A}_i$ $(1 \leq i \leq n)$ the $p \times mp$ submatrix of A whose upper left most element is given by $A_{(i-1)p+1,1}$, by $\bar{A}^j$ $(1 \leq j \leq m)$ the $np \times p$ submatrix whose upper left most element is give by $A_{1,(j-1)p+1}$ and by $\bar{A}^j_i$ the obvious $p \times p$ submatrix. In this way, the whole matrix A may be formed as a block matrix containing each of the mn square submatrices $\bar{A}^j_i$, or of the n "fat" rectangular $\bar{A}_i$ matrices, or of the m "skinny" rectangular matrices $\bar{A}^j$. Essentially, these submatrices make it easy to consider grouped structure.

Proposition 3.5 (Matrix norm $||\cdot||_{\Gamma,\Gamma_\star}$)**.** *For a matrix $A \in \mathbb{R}^{np\times mp}$ we have*

$$||A||_{\Gamma,\Gamma_\star} = \max_{\substack{1\leq i\leq n\\ 1\leq j\leq m}} \sigma_{\max}(\bar{A}^j_i) \tag{3.24}$$

Proof.

$$
\begin{aligned}
||A||_{\Gamma,\Gamma_\star} &= \sup_{\Gamma(x)\leq 1} \Gamma_\star(Ax) \\
&= \max_{\Gamma(x)\leq 1} \max_{1\leq i\leq n} ||\bar{A}_i x||_2 \\
&= \max_{1\leq i\leq n} \max_{\Gamma(x)\leq 1} ||\sum_{j=1}^{m} \bar{A}_i^j \bar{x}_j||_2 \\
&\overset{(a)}{=} \max_{1\leq i\leq n} \max_{\substack{||\rho||_1=1 \\ \rho\geq 0}} \max_{\substack{||u_j||_2=1 \\ 1\leq j\leq m}} ||\sum_{j=1}^{m} \rho_j \bar{A}_i^j u_j||_2 \\
&\leq \max_{1\leq i\leq n} \max_{\substack{||\rho||_1=1 \\ \rho\geq 0}} \sum_{j=1}^{m} \rho_j \max_{||u_j||_2=1} ||\bar{A}_i^j u_j||_2 \\
&= \max_{1\leq i\leq n} \max_{\substack{||\rho||_1=1 \\ \rho\geq 0}} \sum_{j=1}^{m} \rho_j \sigma_{\max}(\bar{A}_i^j) \\
&\overset{(b)}{=} \max_{\substack{1\leq i\leq n \\ 1\leq j\leq m}} \sigma_{\max}(\bar{A}_i^j)
\end{aligned}
$$

where in (a) we are replacing the maximization over $\Gamma(x) \leq 1$ by maximization first over the direction of each of the m groups of x followed by maximization over the allocation to each of these directions, the vector $\rho \in \mathbb{R}^m$ provides a convex combination. The equality in (b) is obtained by allocating the entire weight to the largest singular value in the preceding sum.

This final quantity can be attained with $\Gamma(x) \leq 1$ by choosing x so that $\bar{x}_j$ is the singular vector corresponding to the maximizing submatrix and 0 elsewhere.

□

Proposition 3.6 (Matrix norm induced by $\Gamma_\star$)**.** *For a matrix $A \in \mathbb{R}^{np\times mp}$*

$$
||A||_{\Gamma_\star} = \max_{1\leq i\leq n} \sum_{j=1}^{m} \sigma_{\max}(\bar{A}_i^j), \tag{3.25}
$$

where $\sigma_{\max}$ gives the largest singular value of a matrix.

Proof.

$$\begin{aligned}
||A||_{\Gamma_\star} &= \sup_{\Gamma_\star(x)\leq 1} \Gamma_\star(Ax) \\
&\overset{(a)}{=} \sup_{\Gamma_\star(x)\leq 1} \sup_{\Gamma(z)\leq 1} z^\mathsf{T} Ax \\
&= \max_{\Gamma_\star(x)\leq 1} \max_{\Gamma(z)\leq 1} \sum_{i=1}^{n}\sum_{j=1}^{m} \bar{z}_i^\mathsf{T} \bar{A}_i^j \bar{x}_j \\
&\overset{(b)}{\leq} \max_{\Gamma_\star(x)\leq 1} \max_{\Gamma(z)\leq 1} \sum_{i=1}^{n}\sum_{j=1}^{m} \sigma_{\max}(\bar{A}_i^j)||\bar{z}_i||_2||\bar{x}_j||_2 \\
&\overset{(c)}{=} \max_{1\leq i\leq n} \max_{\Gamma(z)\leq 1} \sum_{j=1}^{m} \sigma_{\max}(\bar{A}_i^j)||\bar{z}_i||_2 \\
&= \max_{1\leq i\leq n} \sum_{j=1}^{m} \sigma_{\max}(\bar{A}_i^j),
\end{aligned}$$

where in (a) we have applied the definition of the dual norm, (b) follows by the application of Cauchy-Schwarz followed by the operator norm bound, (c) follows by taking each $||\bar{x}_j||_2 = 1$ (recall equation 3.22), and the final equality by assigning all of the available weight to the maximizing group of z.

We can achieve this bound by choosing the groups of x and z to correspond to the maximizing left and right singular vectors of the $p \times p$ sub matrices. □

Proposition 3.7 (Matrix norm induced by Γ)**.** *For a matrix $A \in \mathbb{R}^{np\times mp}$*

$$||A||_\Gamma = \max_{1\leq j\leq m} \sum_{i=1}^{n} \sigma_{\max}(\bar{A}_i^j), \tag{3.26}$$

where $\sigma_{\max}$ gives the largest singular value of a matrix.

Proof. First, $\Gamma(Ax) = \sup_{\Gamma_\star(z)\leq 1} z^\mathsf{T} Ax$, that is, the dual of the dual is the original. After this, the proof is entirely analogous to proposition 3.25. □

Proposition 3.8 (Matrix norm induced $||\cdot||_{\Gamma_\star,\Gamma}$)**.** *For a matrix $A \in \mathbb{R}^{np\times mp}$*

$$||A||_{\Gamma_\star,\Gamma} = \sum_{i=1}^{n}\sum_{j=1}^{m} \sigma_{\max}(\bar{A}_i^j), \tag{3.27}$$

Proof. Again, apply the fact that $(\Gamma_\star)_\star = \Gamma$). The rest is similar to the proof of proposition 3.6. □

The preceding propositions display pleasing symmetry amongst the grouped norm Γ, it's dual, and the four naturally induced matrix norms. We will finish this section concerning Γ_{DW} with the observation that, for any matrix $A \in \mathbb{R}^{np\times mp}$:

$$||A||_{\Gamma,\Gamma_\star} \leq \left\{ \begin{array}{c} ||A||_\Gamma \\ ||A||_{\Gamma_\star} \end{array} \right\} \leq ||A||_{\Gamma_\star,\Gamma}. \tag{3.28}$$

Where we have indicated that $||A||_\Gamma$ and $||A||_{\Gamma_\star}$ need not be comparable.

3.4.2 Existence, Uniqueness, and Consistency

The question of whether or not the DWGLASSO problem has a unique minimizer is important for us, since the minimizer is our primary interest, and not simply the minimum value, or a sufficiently useful set of parameters. We want to be able to speak of *the* solution to the DWGLASSO problem. We will consider the objective function

$$J(B) = \frac{1}{2T}||Y - BZ||_F^2 + \Gamma_{DW}(B), \tag{3.29}$$

consisting of the loss function $L(B) = \frac{1}{2T}||Y-BZ||_F^2$ and the regularizer $\Gamma_{DW}(B)$.

This is the standard form formulation, but since $\mathcal{Z}$ is simply a permutation of the columns of Z, the results of this section hold also for the alternate formulation.

Proposition 3.9 (Existence)**.** *The objective function $J(B)$ of equation 3.29 has a minimum value $\widehat{J} = \inf_B J(B)$, which is attained for some $\widehat{B}$.*

Proof. Since $J \geq 0$, we must have $\widehat{J} = \inf J \geq 0$. Secondly, J is the sum of two norms and is hence a coercive (and continuous) function, that is, $J(B) \to \infty$ as $||B||_F \to \infty$. This implies that, since $0 \in \mathbf{dom}\ J$ and thus $\widehat{J} \leq ||Y||_F^2$, that the set $\{B \mid J(B) \leq ||Y||_F^2\}$ is compact. Application of the extreme value theorem over this set implies the existence of a $\widehat{B}$ such that $J(\widehat{B}) = \widehat{J}$. □

The case of uniqueness is more nuanced, particularly if we consider the "high dimensional" regime where $np > T$.

Proposition 3.10 (Uniqueness)**.** *If $\mathsf{rk} Z = np$, then J has a unique minimizer $\widehat{B}$.*

Proof. Consider only the loss function L. This function is differentiable, and it's Hessian is given by $\frac{1}{T}Z^\mathsf{T}Z$. Recall that Z is a real valued $(T \times np)$ matrix. If $\mathsf{rk}\ Z = np$ (that is, Z is full rank), then $\frac{1}{T}Z^\mathsf{T}Z \succ 0$. The positive definite-ness of a function's Hessian is a standard condition to verify strict convexity, in which case J has a unique minimizer (see 2.4) since the sum of a convex and a strictly convex function is strictly convex. □

Remark 26. In the case for which we have more samples than there are parameters $(T \geq np)$, then barring trivial pathologies causing a rank deficiency, proposition 3.10 will always apply. The issues occur when $T < np$, in which case $\mathsf{rk} Z \leq T < np$.

The Case $T < np$

Uniqueness When $T < np$, proposition 3.10 will never apply. However, it is still possible that there exists a unique minimizer. In the case of the LASSO in classical regression, these issues are well understood (see [38], [33] and [42]). The crux of the matter is that the design matrix must satisfy a "mutual incoherence" condition. Recalling the LASSO example 7, mutual incoherence posits the existence of some $c > 0$ such that

$$\max_{j \in S^\mathsf{c}} ||(\mathbf{X}_S^\mathsf{T}\mathbf{X}_S)^{-1}\mathbf{X}_S^\mathsf{T}\mathbf{x}_j||_1 \leq 1 - c, \tag{3.30}$$

where S denotes the support set of the optimal (for the ensemble) parameter vector $\beta^\star$. The intuition is that there is minimal "mixing", or "approximate orthogonality" between the data from indices in S and those in S^c. The ideas surrounding mutual incoherence are pervasive in the literature on sparsity.

Consistency In regression it is typical to show that mutual incoherence holds with high probability for the model under consideration and then to apply it to establish consistency in the sense that, (with high probability) $\text{supp}(\widehat{B}) = \text{supp}(B^\star)$, where $\text{supp}(\cdot)$ refers to the support of the parameter vector.

In the case of our DWGLASSO program, it is reasonable to believe that similar results should apply. Indeed, the analogous proof methods work when applied to deterministic Z matrices, and an analogous incoherence condition can be stated. The question is whether or not a high probability result holds in our data generating model, given that $e(t)$ is Gaussian, or even sub-Gaussian or sub-Exponential. Unfortunately, we are not in a position to include such a result in this thesis, but this is the topic of our current research.

3.5 Algorithms for DWGLASSO

Since the DWGLASSO problem does not have a closed form solution in general, it is necessary to apply an iterative numerical algorithm. There are a wide array of algorithm design methods available for solving convex optimization problems, even when the functions involved are not convex. We discuss two of these algorithms here, beginning with a straightforward subgradient descent algorithm in 3.5.1, and then developing a more sophisticated proximal method in 3.5.2.

3.5.1 Subgradient Descent

Subgradient descent is a method for non-differentiable optimization analogous to regular gradient descent. We have drawn primarily upon [43] as a reference.

Suppose we have a convex, though not necessarily differentiable, objective function $J : \mathbb{R}^n \to \mathbb{R}$. To minimize an objective $J(x)$, subgradient descent dictates that we perform the following iteration, after arbitrary initialization of $x^{(0)} \leftarrow x_0$:

$$x^{(k+1)} \leftarrow x^{(k)} - \alpha_k \phi_J(x^{(k)}), \tag{3.31}$$

where α_k is a predetermined sequence of step size parameters, and $\phi_J(x) \in \partial J(x)$ is an arbitrary subgradient. This method is exactly analogous to gradient descent.

Remark 27. Interestingly, the proof of convergence (see [43]) for subgradient descent shows only that the distance between the current iterate x_k and the optimal x^* decreases monotonically $||x^{(k+1)} - x^*|| \leq ||x^{(k)} - x^*||$, rather than a monotonic decrease in $|J(x^{(k+1)}) - J^*|$ as with standard gradient descent. This is because subgradients are not necessarily descent directions for J.

Consider the alternate form optimization problem of 3.2. We have the objective function

$$J(\widetilde{B}) = \frac{1}{2T}||Y - \mathcal{Z}\widetilde{B}||_F^2 + \lambda \Gamma_{DW}(\widetilde{B}),$$

which we wish to minimize over $\widetilde{B} \in \mathbb{R}^{np \times n}$. There is no closed form solution to this problem, and the non-differentiability of Γ_{DW} at $\widetilde{B} = 0$ adds an additional annoyance. This task is however amenable to subgradient descent. The subdifferential of $J(\widetilde{B})$ is given by:

$$\partial J(\widetilde{B}) = -\frac{1}{T}\mathcal{Z}^{\mathsf{T}}(Y - \mathcal{Z}\widetilde{B}) + \lambda \partial \Gamma_{DW}(\widetilde{B}) \tag{3.32}$$

$$= \frac{1}{T}\big(\mathcal{Z}^{\mathsf{T}}\mathcal{Z}\widetilde{B} - \mathcal{Z}^{\mathsf{T}}Y\big) + \lambda \partial \Gamma_{DW}(\widetilde{B}). \tag{3.33}$$

Hence, we have the proposition:

Proposition 3.11 (Subgradient Descent for DWGLASSO)**.** *After arbitrary initialization of $\widetilde{B}^{(0)} \in \mathbb{R}^{np \times n}$, define the iterations, jointly and simultaneously carried out for each $i \in \{1, 2, ..., n\}$:*

$$\widetilde{B}_i^{(k+1)} \leftarrow \widetilde{B}_i^{(k)} - \frac{1}{k}\Big(\frac{1}{T}\big(\mathcal{Z}^{\mathsf{T}}\mathcal{Z}\widetilde{B}_i^{(k)} - \mathcal{Z}^{\mathsf{T}}y_i\big) + \lambda\phi(\widetilde{B}_i^{(k)})\Big), \tag{3.34}$$

Where ϕ specifies a subgradient as given in proposition 3.1, and y_i is the i^{th} column of Y. If the minimum value of 3.10 with regularizer Γ_{DW} then

$$J(\widetilde{B}^{(k)}) \to \widehat{J} \text{ as } k \to \infty.$$

Furthermore, if there is a unique minimizer $\widehat{\widetilde{B}}$ then

$$||\widetilde{B}^{(k)} - \widehat{\widetilde{B}}||_F^2 \to 0 \text{ as } k \to \infty.$$

Before we prove this proposition, we start with a lemma:

Lemma 3.1. *The function $J(\beta) = ||y - Z\beta||_2^2 + \Gamma_{DW}(\beta)$ for $\beta \in \mathbb{R}^{np}$ is Lipschitz continuous over the region $||\beta||_2 < C$ with parameter $\max(C\sigma_1(Z)^2 + ||Z^{\mathsf{T}}y||_2, 1)$.*

Proof. Firstly, we can see that Γ_{DW} is Lipschitz with parameter 1 by the reverse triangle inequality $||x||_2 - ||x'||_2 \leq ||x - x'||_2$.

Second, we will use the differentiability of $L(\beta) = ||y - Z\beta||_2^2$, combined with the fact that (see [19], Lemma 2.18) $|L(\beta) - L(\beta')| \leq M||\beta - \beta'||_2$ where

$$M = \sup_{||\beta||_2 < C} ||\frac{\partial L}{\partial \beta}(\beta)||_2.$$

We have $\frac{\partial L}{\partial \beta}(\beta) = -Z^{\mathsf{T}}(y - Z\beta)$ and hence

$$\begin{aligned}
\sup_{||\beta||_2 < C} ||\frac{\partial L}{\partial \beta}(\beta)||_2 &= \sup_{||\beta||_2 < C} ||Z^{\mathsf{T}}Z\beta - Z^{\mathsf{T}}y||_2 \\
&\leq \sup_{||\beta||_2 < C} ||Z^{\mathsf{T}}Z\beta||_2 + ||Z^{\mathsf{T}}y||_2 \\
&\overset{(a)}{\leq} C\sigma_1(Z)^2 + ||Z^{\mathsf{T}}y||_2
\end{aligned}$$

where (a) follows from the operator norm for $Z^{\mathsf{T}}Z$.

Finally, the sum of two Lipschitz continuous functions is similarly a Lipschitz continuous function having parameter equal to the maximum of the two former Lipschitz parameters. □

We now proceed with the proof of proposition 3.11.

Proof. The objective function given in the alternate formulation 3.10 is separable over the columns of Y and $\widetilde{B}$ (indeed, this is the primary motivation for this formulation), so the iterations of 3.34 work to solve the n separated problems via subgradient descent.

Since the objective is coercive, there must be some $C_i \geq 0$ such that the solution $\widehat{\widetilde{B}_i}$ to the i^{th} separated problem has $||\widehat{\widetilde{B}_i}||_2 \leq C$. If we apply the lemma 3.1 over this set, we see that the relevant objective function is Lipschitz continuous.

The convergence of the objective value and the parameters now follows from the convergence theorem for subgradient descent [43]. □

3.5.2 Alternating Direction Method of Multipliers (ADMM)

Alternating Direction Method of Multipliers (ADMM) is a method for non-differentiable convex optimization. We have drawn upon [44] and [45] as references for these techniques. Suppose our goal is to solve the following optimization problem:

$$\begin{aligned} \underset{x,z}{\text{minimize}} \quad & f(x) + g(z) \\ \text{subject to} \quad & A_x x + A_z z = c, \end{aligned} \tag{3.35}$$

where f and g must be convex, but need not be differentiable.

To solve the problem 3.35, ADMM stipulates that we perform the following iterations:

$$\begin{aligned} x^{k+1} &\leftarrow \underset{x}{\text{argmin}} \Big[f(x) + \frac{1}{2}\mu ||A_x x + A_z z^k - c + u^k||_2^2 \Big] \\ z^{k+1} &\leftarrow \underset{z}{\text{argmin}} \Big[g(z) + \frac{1}{2}\mu ||A_x x^{k+1} + A_z z - c + u^k||_2^2 \Big] \\ u^{k+1} &\leftarrow u^k + A_x x^{k+1} + A_z z^{k+1} - c. \end{aligned} \tag{3.36}$$

We will see in the sequel that introducing the proximity operator of a convex function is very useful for working with the so called "consensus form" of 3.35.

In the following, we use ψ to denote a function whose range is $\mathbb{R} \cup \{\infty\}$ and which is convex, closed (meaning that **epi** ψ is a closed set), and proper (it's domain is non-empty). We will abbreviate these three conditions and simply refer to the functions as convex.

Definition 3.7 (Proximity Operator)**.** The proximity operator of a convex function $\psi : \mathbb{R}^d \to \mathbb{R}$ at a point v (if ψ has matrix domain then $||\cdot||_2$ becomes $||\cdot||_F$) is defined as the function $\text{prox}_{\mu\psi} : \mathbb{R}^n \to \mathbb{R}^n$,

$$\text{prox}_{\mu\psi}(v) = \underset{x \in \mathbb{R}^n}{\text{argmin}} \big(\psi(x) + \frac{1}{2\mu} ||x - v||_2^2 \big),$$

where $\mu > 0$ is a parameter.

The most straightforward interpretation of the proximity operator is that we optimize ψ in a trust region around v.

A very important property of the proximity operator is the Moreau decomposition:

Theorem 3.2 (Moreau Decomposition)**.**

$$v = \text{prox}_{\psi}(v) + \text{prox}_{\psi^*}(v), \tag{3.37}$$

where

$$\psi^*(x) = \sup_{z \in \mathbb{R}^n} (x^\mathsf{T} z - \psi(z)) \tag{3.38}$$

is the Fenchel conjugate of ψ*.*

Remark 28. The Moreau decomposition tells us that if we can calculate the proximity operator of ψ, then we can also calculate that of ψ^*, and vice versa. The usefulness of this property cannot be understated.

Remark 29. There is a subtlety in the Moreau decomposition in which we actually need the conjugate of $\mu\psi$ in order to work with proximity operators when the parameter $\mu \neq 1$. Fortunately, $(\mu\psi)^*$ is closely related to ψ^*:

$$\begin{aligned}(\mu\psi)^*(x) &= \sup_z (x^\mathsf{T} z - \mu\psi(z)) \\ &= \mu \sup_z (\frac{1}{\mu} x^\mathsf{T} z - \psi(z)) \\ &= \mu\psi^*(x/\mu)\end{aligned} \tag{3.39}$$

If we now specialize $A_x = I$, $A_z = -I$ and $c = 0$ in 3.35, we transform the problem into the consensus form, in which ADMM now provides an algorithm for the unconstrained problem of minimizing $h(x) = f(x) + g(x)$. One of the primary advantages of ADMM is that if $h(x)$ is difficult to minimize when viewed as a whole, but we can easily evaluate the proximity operators of f and g separately, then h can be minimized by application of the consensus form ADMM, which comes down to the evaluation of proximity operators:

$$\begin{aligned}x^{k+1} &\leftarrow \text{prox}_{\mu f}(z^k - u^k) \\ z^{k+1} &\leftarrow \text{prox}_{\mu g}(x^{k+1} + u^k) \\ u^{k+1} &\leftarrow u^k + x^{k+1} - z^{k+1}\end{aligned} \tag{3.40}$$

Turning to DWGLASSO (in standard form), we have $J(B) = \frac{1}{2T}||Y - ZB||_F^2 + \lambda\Gamma_{DW}(B) \triangleq f(B) + g(B)$. In order to apply ADMM to this cost function we need simply to obtain the proximity operators of f and g (recall that prox is easily generalized to matrix arguments).

Proposition 3.12 (Proximity Operator of $f(B) = \frac{1}{2T}||Y - ZB||_F^2$)**.**

$$\text{prox}_{\mu f}(V) = (\frac{1}{T} Z^\mathsf{T} Z + \frac{1}{\mu} I)^{-1} (\frac{1}{T} Z^\mathsf{T} Y + \frac{1}{\mu} V).$$

Note that the matrix inverse should not actually be calculated in practice, an LU factorization of the positive definite matrix $(Z^\mathsf{T} Z + \frac{1}{\mu} I)$ should be cached and the proximity operator evaluated via back substitution.

Proof.

$$\text{prox}_{\mu f}(V) = \underset{B}{\text{argmin}}\ P_f(B),$$

where

$$P_f(B) = \Big[\frac{1}{2T}||Y - ZB||_F^2 + \frac{1}{2\mu}||B - V||_F^2\Big].$$

Since this objective is differentiable and unconstrained, it is easy to solve. We apply the method of [46] in which we first calculate a differential and then infer from it the Jacobian matrix function $\frac{\partial P_f}{\partial B} : \mathbb{R}^{np\times n} \to \mathbb{R}^{np\times n}$.

$$\begin{aligned} \mathrm{d}P_f(B) &= -\frac{1}{T}\mathrm{tr}(Y - ZB)^\mathsf{T} Z(\mathrm{d}B) + \frac{1}{\mu}\mathrm{tr}(B - V)^\mathsf{T}\mathrm{d}B \\ &= \mathrm{tr}\big[\frac{1}{\mu}B^\mathsf{T} - \frac{1}{\mu}V^\mathsf{T} - \frac{1}{T}Y^\mathsf{T}Z + \frac{1}{T}B^\mathsf{T}Z^\mathsf{T}Z\big]\mathrm{d}B \\ \implies \frac{\partial P_f}{\partial B}(B) &= \frac{1}{\mu}(B - V) + \frac{1}{T}(Z^\mathsf{T}ZB - Z^\mathsf{T}Y). \end{aligned}$$

Applying the first order optimality condition

$$\begin{aligned} &\frac{\partial P_f}{\partial B}(B^\star) = 0 \\ \implies &(\frac{1}{T}Z^\mathsf{T}Z + \frac{1}{\mu}I)B^\star = (\frac{1}{T}Z^\mathsf{T}Y + \frac{1}{\mu}V) \\ &\implies B^\star = (\frac{1}{T}Z^\mathsf{T}Z + \frac{1}{\mu}I)^{-1}(\frac{1}{T}Z^\mathsf{T}Y + \frac{1}{\mu}V), \end{aligned}$$

and since the function P_f is strongly convex, we have obtained the unique global minimizer. □

Proposition 3.13 (Proximity Operator of $g(B) = \lambda\sum_{i,j}||\widetilde{B}_{ij}||_2$ ([44]))**.**

$$\text{prox}_{\mu g}(V) = \Big[P(1)\ P(2)\ \ldots\ P(p)\Big]^\mathsf{T} \in \mathbb{R}^{np\times n},$$

where

$$P_{ij}(\tau) = \Big(1 - \frac{\lambda\mu}{||\widetilde{V}_{ij}||_2}\Big)_+ \widetilde{V}_{ij}(\tau)$$

otherwise. The notation $\widetilde{V}_{ij}$ is defined in the same way as $\widetilde{B}_{ij}$ in 2.24 and, $(a)_+ = \max(0, a)$. This is referred to as a block wise soft thresholding operation.

Proof. Let $\phi(\widetilde{B}_{ij}) = ||\widetilde{B}_{ij}||_2$, then $g(B) = \lambda\sum_{ij}\phi(\widetilde{B}_{ij})$. Since $||\cdot||_2^2$ and g both separate along columns of $\widetilde{B}$, so too does the optimization problem which defines the proximity operator 3.14 and hence we need only show that $P_{ij} = \Big(1 - \frac{\lambda\mu}{||\widetilde{V}_{ij}||_2}\Big)_+ \widetilde{V}_{ij}$.

From 3.38 we have

$$\begin{aligned}\phi^*(x) &= \sup_z(x^\mathsf{T} z - ||z||_2) \\ &= \sup_{t\geq 0}\big(t(||x||_2^2 - ||x||_2)\big),\ \{z = tx\} \\ &= I_{\mathcal{B}_2}(x),\end{aligned}$$

the indicator function for the Euclidean unit ball[4],

$$I_{\mathcal{B}_2}(x) = \begin{cases} 0, & ||x||_2 \leq 1 \\ \infty, & ||x||_2 > 1 \end{cases}$$

The proximity operator of $I_{\mathcal{B}_2}$ at v is the projection of v onto $\mathcal{B}_2$. And, projection of v onto the euclidean unit ball merely involves a rescaling of v hence,

$$\text{prox}_{(\mu\phi)^*}(\widetilde{V}_{ij}) = \begin{cases} \widetilde{V}_{ij}, & ||\widetilde{V}_{ij}||_2 \leq \mu \\ \frac{\mu\widetilde{V}_{ij}}{||\widetilde{V}_{ij}||_2}, & \text{otherwise} \end{cases}$$

where the scaling by μ comes from 3.39.

We then use the Moreau decomposition 3.37 to obtain

$$\begin{aligned}\text{prox}_{\mu\phi}(\widetilde{V}_{ij}) &= \widetilde{V}_{ij} - \text{prox}_{(\mu\phi)^*}(\widetilde{V}_{ij}) \\ &= \Big(1 - \frac{\mu}{||\widetilde{V}_{ij}||_2}\Big)_+ \widetilde{V}_{ij}.\end{aligned}$$

The additional scaling by the regularization parameter λ follows easily by rescaling μ.

□

Hence, we have algorithm 1 for fitting the model 3.5 with the DWGLASSO group structured penalty.

Computational Considerations

The computational complexity of the algorithm is actually more modest than it may initially appear. In the common case that T is much larger, or at least comparable, to np the complexity is dominated by the calculation of Σ_Z, which takes $O(n^2p^2T)$ time. In a high dimensional case in which $np > T$, the calculation of the LU factorization dominates with an $O(n^3p^3)$ time requirement. Each iteration of the repeat loop has an $O(n^2p)$ complexity. The time complexity is linear in the number of data samples, although cubic in the time lag p and number n of processes.

The choice of the parameter λ will be discussed in the sequel. The choice of μ is less important, we have found via ad-hoc tuning that $\mu = 0.1$ has worked well. It is important to note however that changing μ has an effect on the convergence criteria, and hence μ should be kept fixed if results are to be comparable.

[4]In general, the Fenchel conjugate of a norm is the indicator of the dual norm unit ball

Data: $\mu > 0, \lambda > 0, \epsilon > 0, Z \in \mathbb{R}^{T \times np}, Y \in \mathbb{R}^{T \times n}$
Result: $\widehat{B} \in \mathbb{R}^{np \times n}$ such that $\widehat{x}(t) = z(t)\widehat{B}$
Initialization: $k = 0$, $B_c^k = 0$, $B_r^k = 0$, $B_u^k = 0$
$\Sigma_Z = \frac{1}{T} Z^\intercal Z$ # variance estimate
$\Sigma_{ZY} = \frac{1}{T} Z^\intercal Y$ # covariance estimate
$L, U \leftarrow \texttt{lu\_factor}(\Sigma_Z + \frac{1}{\mu} I_{np})$
repeat
  $B_c^{k+1} \leftarrow \underset{B}{\texttt{lu\_solve}}\big\{LUB = \Sigma_{ZY} + \frac{1}{\mu}(B_z^k - B_u^k)\big\}$ # $\text{prox}_{\mu f}$
  $V \leftarrow B_r^{k+1} + B_u^k$
  for $i, j, \tau \in \{1, \ldots, n\}^2 \times \{1, \ldots, p\}$ **do**
    $V_{ij}(\tau) \leftarrow \Big(1 - \frac{\lambda\mu}{||\widetilde{V}_{ij}||_2}\Big)_+ V_{ij}(\tau)$ # $\text{prox}_{\mu g}$
  end
  $B_r^{k+1} \leftarrow V$
  $B_u^{k+1} \leftarrow B_u^k + B_c^{k+1} - B_r^{k+1}$
  $k \leftarrow k + 1$
until $\frac{1}{n^2 p}||B_c^k - B_r^k||_F^2 \leq \epsilon$ # convergence;
return B_r^k # sparse solution

Algorithm 1: DWGLASSO

3.5.3 Elastic-Net DWGLASSO

An issue with LASSO type methods is that while the objective functions are convex, they need not be strictly convex when $T < np$. This means that there may be a (convex) set containing many different global minimizers. This implies that when fitting the AR model 2.25, if $x(t)$ has a high degree of co-linearity then the estimates of $B(\tau)$ will not be stable, in the sense that one process from a group of correlated processes can be chosen, the rest being ignored. The particular minimizer chosen by the LASSO is arbitrary, and can easily vary across realizations of the data, as well as when λ varies.

One solution to this problem is given by [47] and is referred to as the elastic-net. The idea here is to blend together the Tikhonov and LASSO penalties with another parameter $\alpha \in [0, 1]$. The addition of the Tikhonov regularization has the effect of blending together correlated variables. Many desirable properties of the elastic net come from the fact that the objective function is strongly[5] convex.

It is quite easy to extend the DWGLASSO to an elastic-net variation by using the blended regularizer, with the parameter $\alpha \in [0, 1]$[6]:

$$\Gamma_{DW_e}(B) = \alpha||B||_F^2 + \Gamma_{DW}(B).$$

This gives us the cost function

$$J(B) = \frac{1}{2T}||Y - ZB||_F^2 + \lambda\big[\alpha||B||_F^2 + (1 - \alpha)\sum_{i=1}^{n}\sum_{j=1}^{n}||\widetilde{B}_{ij}||_2\big].$$

And, the modified proximity operators for ADMM follow easily.

[5] strong convexity implies strict convexity.

[6] One can alternately use $\frac{1}{2}\alpha||B||_F^2 + \Gamma_{DW}(B)$ to avoid a later factor of 2, but the formulations are equivalent.

Proposition 3.14 (Proximity Operator of $f(B) = \frac{1}{2T}||Y - BZ||_F^2 + \lambda\alpha||B||_F^2$)**.**

$$\text{prox}_{\mu f}(V) = (\frac{1}{T}Z^\intercal Z + \frac{1 + 2\mu\lambda\alpha}{\mu}I)^{-1}(\frac{1}{T}Z^\intercal Y + \frac{1}{\mu}V).$$

Proposition 3.15 (Proximity Operator of $g(B) = \lambda(1-\alpha)\sum_{i,j}||\widetilde{B}_{ij}||_2$)**.**

$$\text{prox}_{\mu g}(V) = \Big[P(1)\ P(2)\ \ldots\ P(p)\Big] \in \mathbb{R}^{np\times n},$$

where

$$P_{ij}(\tau) = \Big(1 - \frac{\mu\lambda(1-\alpha)}{||\widetilde{V}_{ij}||_2}\Big)_+ V_{ij}(\tau)$$

The above proximity operators can be substituted into the algorithm 1 to obtain an elastic net analog of our proposed method.

3.6 Simulation Results

In this section we present some simulation studies in relation to the DWGLASSO with data simulated from the model 2.23. While simulating from the true model, when application data is naturally more complex, is sometimes a pointless exercise, in our case it is an important first step because theoretical properties are not yet fully understood. Further, it is necessary to know what is the ground truth causal structure in order to draw conclusions. For application to non-synthetic data, chapter 4 briefly reviews some applications of Granger-causality from the literature, and we also present our own study.

In our simulations, the adjacency matrix G of the true underlying causal graph $\mathcal{G}$ is given by sampling from a directed random graph having n nodes and edge probability q. That is, the presence or absence the n^2 possible edges (we include self loops) are given by *i.i.d.* $\mathsf{Ber}(q)$ random variables.

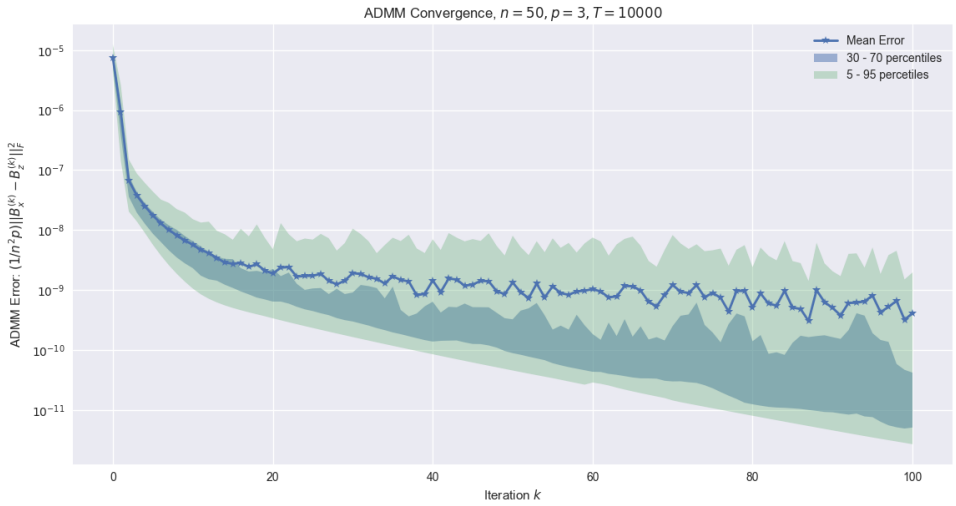
The weights of each edge filter are sampled as *i.i.d.* Gaussian random variables, where the variance of each of these is tuned so that "most" of the resulting autoregressive systems are stable (see section 2.5.1). We reject and resample any unstable models.

Remark 30. We have applied some simple methods based on the Gershgorin circle theorem in order to try to more clearly understand the relationships between the underlying graph, the distribution of the filter weights, and the resulting stability of the AR system. However, the bounds are not sharp enough to be of any real use in sampling stable systems. As far as we are aware, there are a lot of open problems standing in the way of a complete understanding of random distributions over the space of stable autoregressive systems.

3.6.1 ADMM Convergence

Recall the ADMM algorithm given in equation 3.36. Convergence can be measured by the euclidean distance between $x^{(k)}$ and $z^{(k)}$. It is typical for the iterates to rapidly converge to a modest accuracy, without making very much progress with

Figure 3.2: ADMM Convergence



further iterations. In some applications this may be a serious drawback, but this is not the case for statistical applications, since the data is already noisy and models are approximate.

In figure 3.2 we have run 50 simulations with parameters $n = 50, T = 10000, p = 3, q = 0.2, \Sigma_e = 0.1I$, which corresponds to the regime of abundant data. We have plot the average over these 50 trials as well as some bounds between percentiles. As expected, convergence to modest accuracy is rapid, with little progress thereafter. We normalize the error by the number n^2p of parameters so we are measuring the error per parameter and in this way it is possible to specify a stopping criteria which is consistent over different model sizes. e.g. one is to specify some $\epsilon > 0$ such that the algorithm halts when $\frac{1}{n^2p}||x^{(k)} - z^{(k)}||_F^2 \leq \epsilon$. We typically employ $\epsilon = 10^{-6}$ or $\epsilon = 10^{-9}$.

3.6.2 Model Consistency in Squared Error

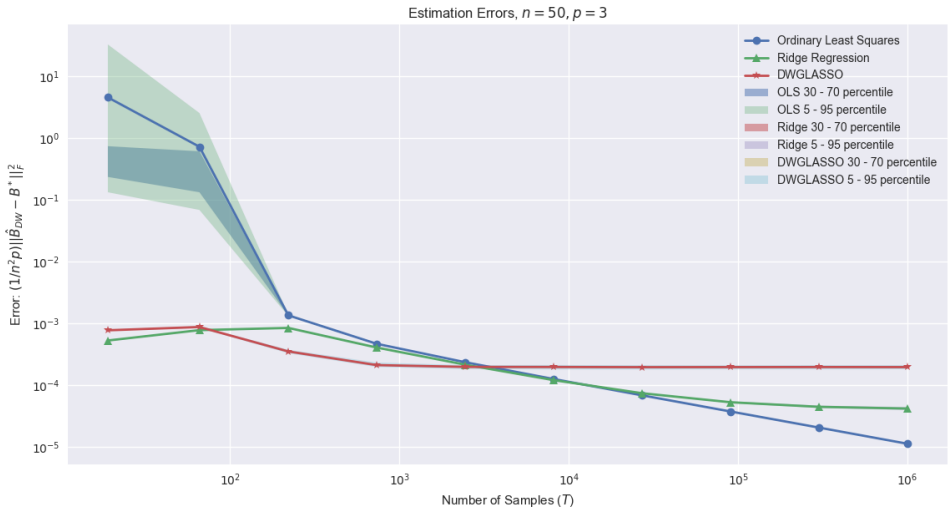
Consider the difference between the set of parameters $B^\star$ optimal for the error function over the entire ensemble and the minima $\widehat{B}$ of the finite sample DWGLASSO loss. That is, $B^\star$ minimizes $\mathbb{E}||x(t) - Bz(t)||_F^2$ and $\widehat{B}$ minimizes $\frac{1}{2T}\sum_{t=1}^{T}||x(t) - Bz(t)||_F^2 + \Gamma_{DW}(B)$.

We consider here whether or not $||\widehat{B} - B^\star||_F^2 \to 0$ as $T \to \infty$. We draw from our random model some very large number of samples, and then minimize an empirical objective using a progressively larger number of samples. We compare ordinary least squares (OLS), Tikhonov regularized least squares (OLST) and the DWGLASSO (DW).

Remark 31. Model consistency as measured by the euclidean norm is quite well understood, even for general $\mathcal{M}$-estimators (see [33]). Furthermore, convergence in norm tells us little about the recovery of the actual underlying causality graph, which is the object of our interest. We hence keep this section short.

In the case of ordinary least squares, the error between $\widehat{B}_{OLS}$ and $B^\star$ goes to zero as $T \to \infty$ under mild assumptions. However, as previous discussed, the estimate

Figure 3.3: L_2 Convergence



is highly unstable when T is small. In the case of the regularized solutions (OLST and DWGLASSO), the error will not converge to 0 when λ is fixed, it is necessary to also have $\lambda \to 0$ as $T \to \infty$.

This behaviour is observed in figure 3.3, where we have used $n = 50, p = 3, q = 0.1$ and held λ fixed. We further note that the regularized solutions are much more consistent, the percentile regions being hardly visible.

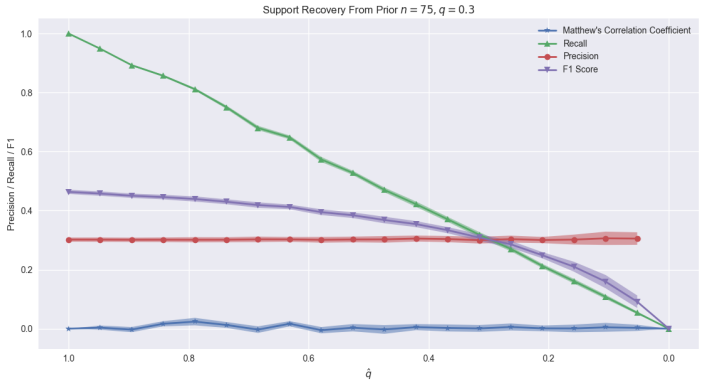
3.6.3 Model Support Recovery

As with section 3.6.2 we consider here the distance between the finite sample minimizer $\widehat{B}$ and the ensemble minimizer $B^\star$. In contrast to the previous section, we here consider the supports (the indices of the non-zero entries) of the underlying causality graph. We will refer to the causality graphs inferred from the sets of parameters as $\widehat{G}_\lambda$ and $G^\star$ in the obvious way.

For the simulations in this section, we have used $n = 75, p = 3, q = 0.3$ as well as the elastic net version of DWGLASSO with $\alpha = 0.1$ fixed. Note further that when $\lambda = 0$, the DWGLASSO solution is the same as the OLS solution, and $\widehat{G}_0$ will be completely dense. Hence, there is some λ_0 (possibly $\lambda_0 = 0$) such that $\widehat{G}_\lambda$ is dense for every $0 \leq \lambda \leq \lambda_0$. On the other hand, as $\lambda \to \infty$ there is some value λ_∞ such that $\widehat{G}_{\lambda_\infty}$ is entirely 0 for any $\lambda \geq \lambda_\infty$. In the figures of this section, we generally vary λ on a logarithmic scale between, roughly, these two quantities, $\lambda \in [\lambda_0, \lambda_\infty]$.

Our goal is to determine whether or not $\widehat{G}_\lambda$ and $G^\star$ share the same support, and when there are differences, what type of error is it (e.g. false positive or negatives). It is common to measure the quality of the estimates via the precision, recall, and $F1$ score. These are given by $\frac{TP}{TP+FP}$, $\frac{TP}{TP+FN}$, and $\frac{2TP}{2TP+FP+FN}$ respectively, where TP, TN, FP, FN are the counts of true positives, true negatives, false positives, and false negatives. The $F1$ score is the harmonic mean of the precision and recall. Intuitively, the precision measures how often our inferred edges are truly present in $G^\star$, and the recall measures how many of the true edges we discover. The $F1$ score is a reasonable combination of these two quantities into a single performance metric.

Figure 3.4: Random Guess, Base Probability $\hat{q} \in [1, 0]$



The above mentioned metrics are sensitive to the base probability q of an edge, which we have fixed at 0.3. We hence employ a final measurement of quality, Matthew's correlation coefficient (MCC), which is applicable regardless of the base probability. For the MCC measurement, a value of 0 indicates the quality of a random guess (using any probability) and a value of 1 indicates that the graph has been perfectly recovered. The definition of the MCC is given by

$$\frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}.$$

This measure is easiest to understand at a glance, as we can immediately note that "bigger is better", but the $F1$ score is also a very common metric for binary classification.

We first establish a baseline by sweeping through a probability $\hat{q}$ from 0 to 1, and form $\widehat{G}$ by uniformly at random (with probability $\hat{q}$) picking edges. The result of this scheme is given in figure 3.4 with $\hat{q}$ on the bottom axis, and with each of the above error metrics given on the vertical axis. The MCC stays close to 0, which is to be expected. On the other hand, our other error metrics can take on fairly large values simply by guessing the dense graph. This random baseline should be kept in mind while interpreting the next set of figures.

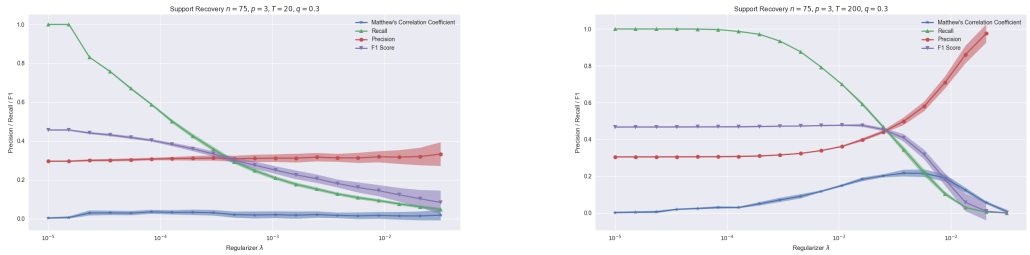
We form the figure 3.5 by sweeping λ through $[\lambda_0, \lambda_\infty]$ on the horizontal axis. It is clear that there is some value of λ which is optimal [7] for a particular instance of the problem. From the figures we can also see the tendency for the best possible performance to improve as we obtain more samples, with $T = 20000$ being enough to almost perfectly reconstruct the underlying graph.

Remark 32. For the simulations of figure 3.5 there are $n = 75$ nodes, which corresponds to as many as 5625 possible directed edges (including self edges), and with $p = 3$ there are 16875 parameters in the model. With only $T = 20$ samples, the DWGLASSO performs at best marginally better than randomly guessing, and given such a limited quantity of data, there isn't much that can be done. Given only $T = 200$ samples, a regime in which (since $\mathsf{rk} Z = np = 225$) the OLS or Tikhonov

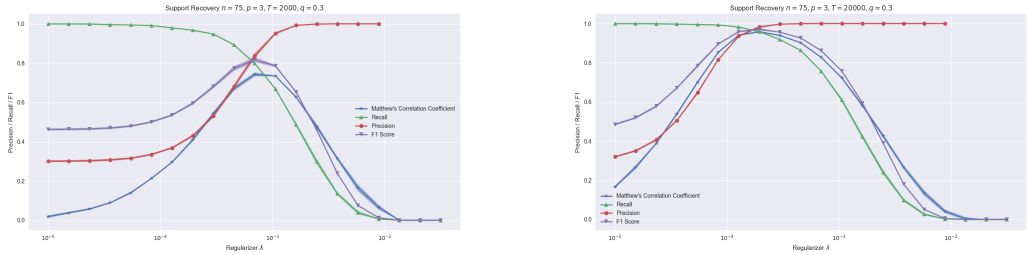
[7] The notion of "optimal" here depends on how much one cares about the difference between false negatives and false positives. And, it may also be the case that there is no one single optimal λ, but some Pareto optimal interval.

Figure 3.5: Support Recovery, $\lambda \in [10^{-5}, 10^{-1.5}]$

(a) $T = 20$ (b) $T = 200$



(c) $T = 2000$ (d) $T = 20000$



solutions still fail to even exist, we see a discernible improvement in the $F1$ score and the MCC over randomness. Observing $T = 2000$ samples (still fewer than the number of parameters) is more than enough obtain a reasonable level of accuracy, and $T = 20000$ samples is enough for almost perfect recovery.

Edge Intensity One of the most significant difficulties for support recovery is the choice of the parameter λ which leads to the best recovery. Through our experiments, it is absolutely clear that the standard method of choosing λ by cross validation on the 1-step ahead prediction task leads to a choice of λ which is much too small for the task of support recovery. And, since we don't have access to the true underlying graph a priori, it isn't clear if we can do any sort of cross validation to choose λ.

As an alternative to picking a fixed λ, we propose to make a final estimate based on an "edge intensity" matrix, which we define as

$$\widehat{G}_\Lambda = \frac{1}{|\Lambda|} \sum_{\lambda \in \Lambda} \hat{G}_\lambda, \tag{3.41}$$

where $\Lambda \subseteq [\lambda_0, \lambda_\infty]$ discretizes the interval on a logarithmic scale.

The final estimate is then given by the choice of a threshold $\mathcal{T} \in [0, 1]$ as

$$\widehat{G} \leftarrow (\widehat{G}_\Lambda \geq \mathcal{T}), \tag{3.42}$$

where we interpret the inequality 3.42 as an element-wise logical operation. In figure 3.6 we plot ROC curves for this particular scheme. Finally, since an ROC curve is not a perfect measure of performance when the base rate is skewed away from 50%, we provide a similar plot of the Matthew's correlation coefficient in figure 3.7.

Figure 3.6: ROC Curves for Equation 3.42

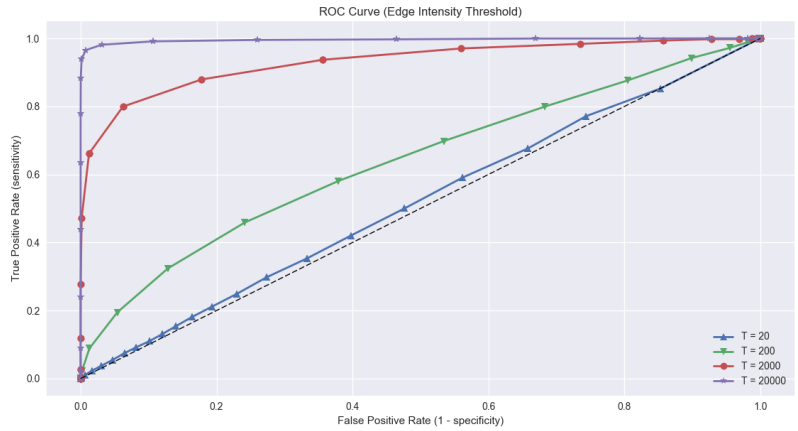
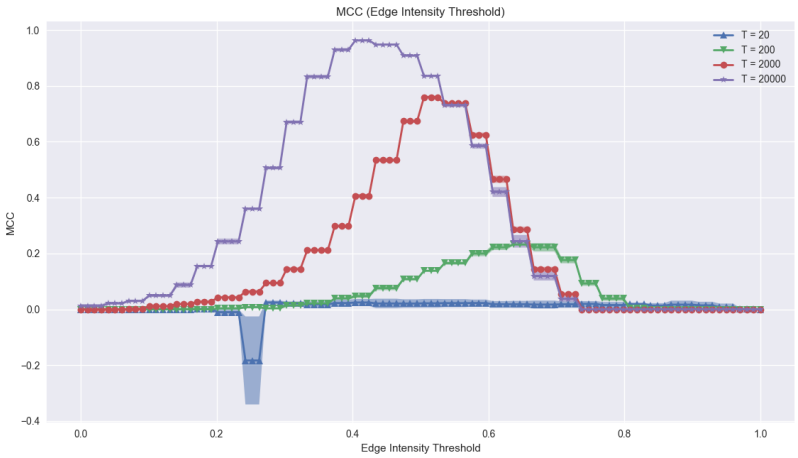


Figure 3.7: Matthew's Correlation Coefficient for Equation 3.42



Chapter 4

Applications

4.1 Applications from the Literature

In this section we touch on some of the ways in which Granger-causality is applied in practice.

4.1.1 Finance and Economics

The ideas of Granger-causality began from motivations in economics, and the potential applications are boundless, see for example the discussion at the beginning of section 3.3.2.

A particular application we would like to point out here is given in [15], which seeks to measure the connectedness of the financial sector, and to link this to systemic risk. The authors gathered data relevant to a wide variety of financial firms (e.g. banks, insurance companies, hedge funds) and applied measurements of Granger-causality to estimate the interactions between firms.

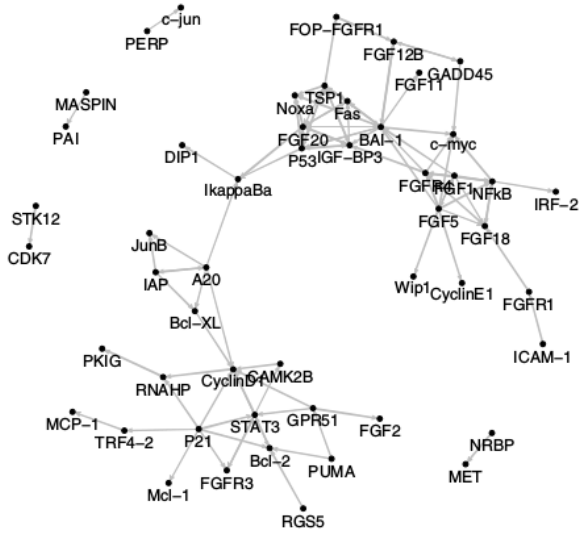
Figure 4.1 provides the results of the Granger-causal analysis in [15]. The upshot here is that an affirmative answer to the qualitative question "are financial systems more interdependent than they were in the past?" is given using Granger-causality analysis. Furthermore, the authors make useful insights using *linear* measures, even though the underlying systems and interactions are by no means linear in reality.

Figure 4.1: Qualitative Measures of Financial Sector Connectedness [15]

(a) January 1994 - December 1996

(b) January 2006 - December 2008

Figure 4.2: Gene Regulatory Network Inferred by [16] through Granger-causality and the LASSO



4.1.2 Neuroscience

In Neuroscience, researchers are ultimately interested in determining how and why the brain works. A common experimental approach to this problem is to use various methods of imaging the brain, examining blood flow for example, and to compare the observed patterns with the observable behaviour of the test subject. In this way, researchers test and hypothesize about the function of particular regions (e.g. visual cortex or auditory cortex) of the brain.

The review article [48] states the goals that "a key challenge in neuroscience and, in particular, neuroimaging, is to move beyond identification of regional activations toward the characterization of functional circuits underpinning perception, cognition, behavior, and consciousness." Granger-causality is being applied to reach these ends by identifying, via Granger-causal modeling, "functional" connections in the brain.

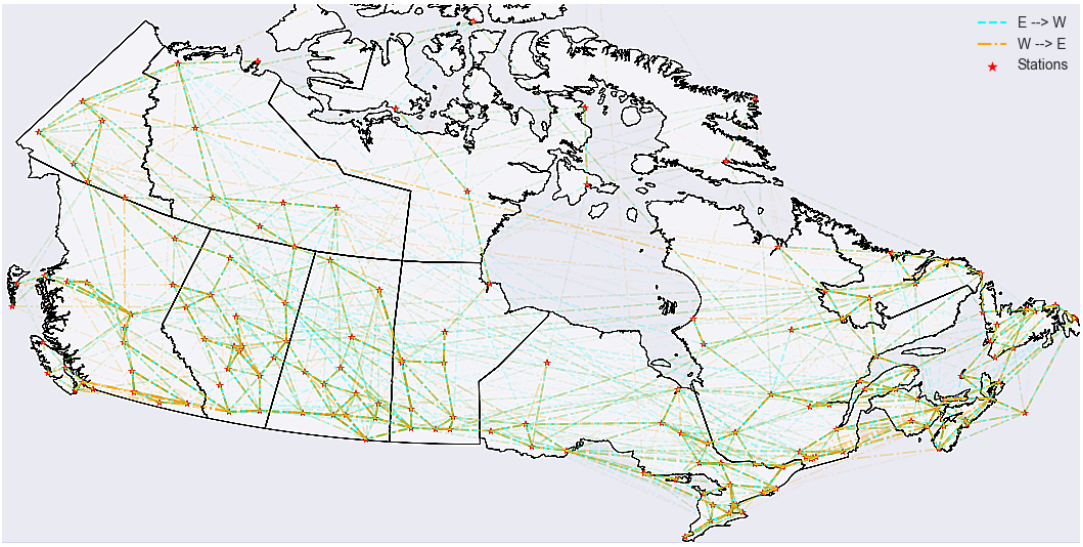
4.1.3 Biology

We consider the results of the paper by Fujita et. al. [16], who explain that "In order to understand cell functioning as a whole, it is necessary to describe, at the molecular level, how gene products interact with each other".

We first must point out that the common notion that one particular gene does one particular thing is extremely inaccurate. In reality, the vast number of genes that make up our DNA participate in highly complex interactions and feedback loops, not only producing various proteins, but also suppressing and activating other genes.

In examining gene regulatory networks, the number of genes (n) one may wish to examine may be very large, and since carrying out experiments is a rather arduous process, the number of samples (T) is likely to be very small. [16] has applied sparsity inducing regularization (LASSO) to $\mathsf{VAR}(1)$ models of gene regulatory networks, along with some additional statistical methods to try to control the false positive rate. The results are reproduced here in figure 4.2.

Figure 4.3: Inferred causality graph. Direction of each edge from west (left) to east (right) or from east to west is indicated by color and line style. The transparency of each edge is weighted by the edge intensity.



Certainly the results here must be taken with a grain of salt. As we noted, the underlying interactions are fantastically complicated, and are certainly not going to be completely captured by a VAR model. However, the results of Granger-causal analyses, like those of [16], may serve as reasonable guides for further investigation. There are a truly enormous number of possible experiments, and any techniques which help discern which ones may bear fruit deserves some consideration.

4.2 DWGLASSO Applied to CWEEDS Temperature Data

The Canadian weather, energy, and engineering data set (CWEEDS) provides (among other things) hourly temperature data between a large number of Canadian locations. We have selected $n = 165$ locations, chosen based on availability of data over a consistent span of time starting on January 1^{s} 1980, and proceeding for $T = 1600$ hours until March 6^{th}. We apply the elastic net DWGLASSO to this dataset with fixed $\alpha = 0.1, p = 2$. Note that there are a 27225 possible edges in the causality graph and 54450 model parameters, although with $np = 330$, the OLS solution still exists.

We have chosen to use temperature data because geographic considerations can give some intuition about what the "true" underlying Granger-causality graph should look like, but the data is still more realistic than the synthetic examples of section 3.6. The spirit of this application example is similar to that of the simulations in section 3.6. That is, the point is to test the potential usefulness of DWGLASSO in practice, not to discover any new or interesting patterns in weather data.

The tools we have used for this application are part of Python's scientific computing stack [49].

We have applied some limited preprocessing steps to our dataset, first interpo-

lating a number of missing data points [1], properly aligning the time stamps of each series (CWEEDS provides local times), and filtering out perfectly predictable yearly and daily temperature trends (and harmonics thereof). The last preprocessing step is important for the application of autoregressive models, since we assume there are no purely deterministic, or "trend", components in the data (see theorem 2.7 and the brief discussion that follows).

The inferred graph is given in figure 4.3. We have weighted the transparency of the edges by the edge intensity G_Λ where $\Lambda \subseteq [0.01, 10]$. There are clearly a great number of spurious edges having a low intensity (edges inferred with a small λ), but the edges with a high intensity correspond fairly consistently to weather stations in close proximity.

[1] It is important to ensure any preprocessing steps are themselves causal

Chapter 5

Conclusion

5.1 Further Research

5.1.1 Theoretical Results for Support Recovery

Consider again the true underlying causality graph $G^\star$ inferred from the set of parameters $B^\star$, optimal for the loss function over the ensemble, and the estimated graph $\widehat{G}$ obtained from the DWGLASSO method. It is an important question whether or not, and under what conditions it can be shown that $\mathbb{P}\{\widehat{G}_{\lambda_T} = G^\star\} \to 1$ as $T \to \infty$, for some sequence of λ_T. A result of this nature is given by Nardi [50] for $\mathsf{VAR}(1)$ models, which requires application only of the LASSO, rather than the grouped variation.

More generally, we seek to estimate the probability of errors $\mathbb{P}\{\gamma_0(\widehat{G}_\lambda - G^\star) \leq k\}$, and determine how this relates to the structure of the underlying graph $G^\star$, the coefficient matrices, and the statistics of the error terms $e(t)$. The highly influential paper [42] established a technique for approaching these problems referred to as the "primal dual witness" (PDW) construction which has been applied in a wide variety of papers (including [50]). Application of this proof method to the DWGLASSO problem leads to a condition analogous to the mutual incoherence condition (equation 3.30), and is also tightly connected with the dual norm and the induced operator norm of the depth wise regularizer Γ_{DW} which we have worked out in section 3.4.1. It is however, a difficult task to tie together the conditions arrived at naturally through application of the PDW to the structure of $G^\star$ and $B^\star$.

5.1.2 Choice of Hyper-parameters

The choice of the parameters λ, α, and μ in the DWGLASSO have significant effects on the inferred causality graph. It is generally reasonable to hold α and μ fixed, varying only λ, so we focus the discussion on this parameter. As we noted previously, it is known that applying cross validation on the 1-step ahead prediction task tends to yield a λ_{CV} which is much smaller than what is optimal for support recovery. A consistent, and provably statistically efficient scheme for choosing a fixed λ (or an edge intensity threshold $\mathcal{T}$) is highly desirable.

In the case of the LASSO, the error bounds can be useful for guiding the choice of λ. However, to our knowledge, the theoretical properties of cross validation schemes are not extremely well understood, especially for the support recovery task. The

choice of λ is an important question to answer.

5.1.3 Model Perturbation

One of the most commonly acknowledged limitations with standard Granger-causality is that realistic systems are not linear and are not stationary. The most immediate way to tackle this difficulty is to formulate Granger-causality for non-linear and or non-stationary processes. Alternatively, since we are interested only in the support of the underlying causality graph, we may define $G^\star$ with respect to a much more expressive model, but continue to estimate $\widehat{G}$ as if the process were indeed generated by a linear model. We can then still study conditions for which $\widehat{G} \approx G^\star$.

For instance, consider a non-stationary L_2^n process $x(t)$, with

$$\mathbb{E}x(t)x(t)^\mathsf{T} = R(t),\ \mathbb{E}x(t)x(t-1)^\mathsf{T} = r(t).$$

The statistically optimal linear estimate of $x(t)$ given $x(t-1)$ is $\hat{x}(t) = B^\star(t)x(t-1)$ with $B^\star(t) = r(t)R(t)^{-1}$. The underlying graph $G^\star(t)$ is in general time varying, but if $\mathsf{supp}(B^\star(t))$ is constant, then it is clear how to define a consistent causality graph $G^\star$.

Now, if we had assumed that $x(t)$ were actually stationary then (proceeding informally), we may have estimated by OLS

$$\widehat{B} = \Big(\frac{1}{T}\sum_{t=1}^{T} x(t)x(t-1)^\mathsf{T}\Big)\Big(\frac{1}{T}\sum_{t=1}^{T} x(t)x(t)^\mathsf{T}\Big)^{-1} \overset{(a)}{\approx} \Big(\sum_{t=1}^{T} r(t)\Big)\Big(\sum_{t=1}^{T} R(t)\Big)^{-1},$$

where in (a) we are referring to single sample estimates of the statistics. Consider then the error sequence $\Delta(t)$ such that

$$B^\star(t) = \Big(\sum_{t=1}^{T} r(t)\Big)\Big(\sum_{t=1}^{T} R(t)\Big)^{-1} + \Delta(t).$$

If $\Delta(t)$ is small for every t, then we would suspect that the correct graph structure should still be recovered, even if we applied a model which assumed $\Delta = 0$.

5.2 Summary

In this thesis we have discussed two main topics of interest to the study of causality networks.

Firstly, we have described in section 2.2.1 a framework in which a generalized Granger-causality maintains the identical spirit of much of the early work on the topic, but in which practitioners may attempt to apply more expressive models. This framework furthermore makes more explicit the connections between Granger-causality and regularized estimation of VAR models.

Secondly, we have derived in chapter 3 algorithms to fit vector auto-regressive models with the depth-wise grouped sparsity pattern (particularly well suited to working with causality graphs), and provided some experimental validation via simulation (section 3.6) and an application to weather data (section 4.2).

Bibliography

[1] Thomas Kjeller Johansen, Desmond Lee, et al. *Timaeus and Critias*. Penguin UK, 2008.

[2] Stephen Mumford and Rani Lill Anjum. *Causation: a very short introduction*. Oxford University Press, 2013.

[3] Andrea Falcon. *Aristotle on Causality*. 2006. URL: https://plato.stanford.edu/entries/aristotle-causality/.

[4] Bertrand Russell. "On the Notion of Cause". In: *Proceedings of the Aristotelian Society* 13 (1912), pp. 1–26. ISSN: 00667374, 14679264. URL: http://www.jstor.org/stable/4543833.

[5] Yitzhak Y. Melamed and Martin Lin. "Principle of Sufficient Reason". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Spring 2017. Metaphysics Research Lab, Stanford University, 2017.

[6] Henri Poincaré. *Science and hypothesis*. Science Press, 1905.

[7] Judea Pearl. "The art and science of cause and effect". In: *Causality: models, reasoning and inference* (2000), pp. 331–358.

[8] Judea Pearl. *Causality*. Cambridge university press, 2009.

[9] C.W.J. Granger. "Testing for causality: A personal viewpoint". In: *Journal of Economic Dynamics and Control* 2 (1980), pp. 329 –352. ISSN: 0165-1889. DOI: http://dx.doi.org/10.1016/0165-1889(80)90069-X. URL: http://www.sciencedirect.com/science/article/pii/016518898090069X.

[10] John Geweke. "Measurement of linear dependence and feedback between multiple time series". In: *Journal of the American statistical association* 77.378 (1982), pp. 304–313.

[11] John F. Geweke. "Measures of Conditional Linear Dependence and Feedback between Time Series". In: *Journal of the American Statistical Association* 79.388 (1984), pp. 907–915. DOI: 10.1080/01621459.1984.10477110. eprint: http://www.tandfonline.com/doi/pdf/10.1080/01621459.1984.10477110. URL: http://www.tandfonline.com/doi/abs/10.1080/01621459.1984.10477110.

[12] Olivier David et al. "Identifying neural drivers with functional MRI: an electrophysiological validation". In: *PLoS Biol* 6.12 (2008), e315. URL: http://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.0060315.

[13] Lionel Barnett and Anil K. Seth. "Detectability of Granger causality for subsampled continuous-time neurophysiological processes". In: *Journal of Neuroscience Methods* 275 (2017), pp. 93 –121. ISSN: 0165-0270. DOI: http://dx.doi.org/10.1016/j.jneumeth.2016.10.016. URL: //www.sciencedirect.com/science/article/pii/S0165027016302564.

[14] Alberto Porta and Luca Faes. "Wiener–Granger causality in network physiology with applications to cardiovascular control and neuroscience". In: *Proceedings of the IEEE* 104.2 (2016), pp. 282–309.

[15] Monica Billio et al. *Econometric Measures of Systemic Risk in the Finance and Insurance Sectors*. Working Paper 16223. National Bureau of Economic Research, 2010. DOI: 10.3386/w16223. URL: http://www.nber.org/papers/w16223.

[16] André Fujita et al. "Modeling gene expression regulatory networks with the sparse vector autoregressive model". In: *BMC Systems Biology* 1.1 (2007), p. 39. ISSN: 1752-0509. DOI: 10.1186/1752-0509-1-39. URL: http://dx.doi.org/10.1186/1752-0509-1-39.

[17] Michail Misyrlis et al. "Sparse Causal Temporal Modeling to Inform Power System Defense". In: *Procedia Computer Science* 95 (2016). Complex Adaptive Systems Los Angeles, {CA} November 2-4, 2016, pp. 450 –456. ISSN: 1877-0509. DOI: http://dx.doi.org/10.1016/j.procs.2016.09.316. URL: //www.sciencedirect.com/science/article/pii/S1877050916324899.

[18] Tao Yuan and S Joe Qin. "Root cause diagnosis of plant-wide oscillations using Granger causality". In: *Journal of Process Control* 24.2 (2014), pp. 450–459.

[19] John K Hunter and Bruno Nachtergaele. *Applied analysis*. World Scientific Publishing Co Inc, 2001.

[20] Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal. *Fundamentals of convex analysis*. Springer Science & Business Media, 2012.

[21] Ivar Ekeland and Roger Temam. *Convex analysis and variational problems*. SIAM, 1999.

[22] Jean-Pierre Aubin. *Optima and equilibria: an introduction to nonlinear analysis*. Vol. 140. Springer Science & Business Media, 2013.

[23] Geoffrey Grimmett and David Stirzaker. *Probability and random processes*. Oxford university press, 2001.

[24] Anders Lindquist and Giorgio Picci. *Linear Stochastic Systems*. Springer.

[25] T. Kailath, A.H. Sayed, and B. Hassibi. *Linear Estimation*. Prentice Hall Information and. Prentice Hall, 2000. ISBN: 9780130224644. URL: https://books.google.ca/books?id=zNJFAQAAIAAJ.

[26] Sheldon M Ross. *Introduction to probability models*. Academic press, 2014.

[27] Peter J Brockwell and Richard A Davis. *Time series: theory and methods*. Springer Science & Business Media, 2013.

[28] Vivek K Goyal Martin Vetterli Jelena Kovacevic. *Foundations of Signal Processing*. Cambridge University Press, 2014.

[29] Yuejia He, Yiyuan She, and Dapeng Wu. "Stationary-sparse causality network learning." In: *Journal of Machine Learning Research* 14.1 (2013), pp. 3073–3104.

[30] C. W. J. Granger. "Investigating Causal Relations by Econometric Models and Cross-spectral Methods". In: *Econometrica* 37.3 (1969), pp. 424–438. ISSN: 00129682, 14680262. URL: http://www.jstor.org/stable/1912791.

[31] Y. Rozanov. *Stationary Random Processes*. Holden-Day, 1967.

[32] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*. Vol. 1. Springer series in statistics Springer, Berlin, 2001.

[33] Sahand Negahban et al. "A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers". In: *Advances in Neural Information Processing Systems*. 2009, pp. 1348–1356.

[34] Stephen M. Stigler. "Gauss and the Invention of Least Squares". In: *Ann. Statist.* 9.3 (May 1981), pp. 465–474. DOI: 10.1214/aos/1176345451. URL: http://dx.doi.org/10.1214/aos/1176345451.

[35] William H Greene. "Econometric analysis 4th edition". In: *International edition, New Jersey: Prentice Hall* (2000).

[36] Robert Tibshirani. "Regression shrinkage and selection via the lasso". In: *Journal of the Royal Statistical Society. Series B (Methodological)* (1996), pp. 267–288.

[37] Emmanuel J Candès, Justin Romberg, and Terence Tao. "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information". In: *IEEE Transactions on information theory* 52.2 (2006), pp. 489–509.

[38] T. Hastie, R. Tibshirani, and M. Wainwright. *Statistical Learning with Sparsity: The Lasso and Generalizations*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis, 2015. ISBN: 9781498712163. URL: https://books.google.ca/books?id=LnUIrgEACAAJ.

[39] Ming Yuan and Yi Lin. "Model selection and estimation in regression with grouped variables". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 68.1 (2006), pp. 49–67.

[40] Stefan Haufe et al. "Sparse causal discovery in multivariate time series". In: *Proceedings of the 2008th International Conference on Causality: Objectives and Assessment-Volume 6*. JMLR. org. 2008, pp. 97–106.

[41] R Tyrrell Rockafellar and RJB Wets. *VARIATIONAL ANALYSIS*. Springer: Grundlehren der Math. Wissenschaften., 1998.

[42] Martin J Wainwright. "Sharp thresholds for High-Dimensional and noisy sparsity recovery using ℓ_1-Constrained Quadratic Programming (Lasso)". In: *IEEE transactions on information theory* 55.5 (2009), pp. 2183–2202.

[43] Stephen Boyd and Almir Mutapcic. "Subgradient methods". In: *Lecture notes of EE364b, Stanford University, Winter Quarter* 2007 (2006).

[44] Neal Parikh, Stephen Boyd, et al. "Proximal algorithms". In: *Foundations and Trends® in Optimization* 1.3 (2014), pp. 127–239.

[45] Stephen Boyd et al. "Distributed optimization and statistical learning via the alternating direction method of multipliers". In: *Foundations and Trends® in Machine Learning* 3.1 (2011), pp. 1–122.

[46] Jan R Magnus, Heinz Neudecker, et al. "Matrix differential calculus with applications in statistics and econometrics". In: (1995).

[47] Hui Zou and Trevor Hastie. "Regularization and variable selection via the elastic net". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67.2 (2005), pp. 301–320.

[48] Anil K Seth, Adam B Barrett, and Lionel Barnett. "Granger causality analysis in neuroscience and neuroimaging". In: *Journal of Neuroscience* 35.8 (2015), pp. 3293–3297.

[49] Eric Jones, Travis Oliphant, Pearu Peterson, et al. *SciPy: Open source scientific tools for Python*. 2001–. URL: http://www.scipy.org/.

[50] Y. Nardi and A. Rinaldo. "Autoregressive process modeling via the Lasso procedure". In: *Journal of Multivariate Analysis* 102.3 (2011), pp. 528 –549. ISSN: 0047-259X. DOI: http://dx.doi.org/10.1016/j.jmva.2010.10.012. URL: http://www.sciencedirect.com/science/article/pii/S0047259X10002186.