Nonparametric Detection and Estimation of Highly Oscillatory Signals

Thesis by

Hannes Helgason

In Partial Fulfillment of the Requirements
for the Degree of
Doctor of Philosophy

California Institute of Technology
Pasadena, California

2008
(Defended August 30, 2007)

Fyrir Gerði og Iðunni Önnu.

Acknowledgements

I would like to express my deepest gratitude to my advisor, Prof. Emmanuel Candes, for his support, guidance, and friendship throughout my studies at Caltech. His insights into science and contagious enthusiasm have been a great inspiration to me. I also extend my gratitude to his wife Chiara and the people in his research group, including Laurent Demanet, Paige Randall, Peter Stobbe, Justin Romberg, and Lexing Ying.

I would like to thank Gary Lorden, Houman Owhadi, Yaser Abu-Mostafa, and Alan Weinstein for serving on my thesis committee, Philip Charlton for his collaboration, and the people at LIGO. I would also like to acknowledge all the great professors whom I took classes from while at Caltech, in particular Oscar Bruno and P.P. Vaidyanathan. I owe special thanks to Dr. Askell Hardarson, who had a great influence on me and opened my eyes to the joy and beauty of mathematics during my teenage years at Flensborg.

During the summer of 2006 I had the fortunate opportunity to be a research advisor for the RIPS program at IPAM, UCLA. I would like to thank the talented students in my research group, the team at NASA Goddard for providing us with an interesting and challenging research project, and the remarkable staff at IPAM, in particular Prof. Mark Green and Stacey Beggs.

The Applied and Computational Mathematics department would not be the same without the terrific work done by Sheila Shull, Chad Schmutzer, and Sydney Garstang. I am forever thankful for Sheila's support and kindness during my stay at Caltech.

My stay at Caltech would not have been the same without all the great friends I have had there. These include Guillaume, Angel, Alan & Allison, Stephane & Marta, Lisa, Fady, Oliver, Jon & Kaithlyn, Donal, Vala, and Kristjan, who have been like my second family while abroad.

Most of all I would like to thank my family in Iceland: My parents, Helgi & Guðmunda, for giving me the opportunities they never had, my brothers, Vignir & Heimir, and most

importantly, my wife Gerður and daughter Iðunn Anna, who have been an endless source of inspiration, love and support. I dedicate this thesis to them.

Abstract

This thesis considers the problem of detecting and estimating highly oscillatory signals from noisy measurements. These signals are often referred to as chirps in the literature; they are found everywhere in nature, and frequently arise in scientific and engineering problems. Mathematically, they can be written in the general form $A(t)\exp(\imath\lambda\,\varphi(t))$, where λ is a large constant base frequency, the phase $\varphi(t)$ is time-varying, and the envelope $A(t)$ is slowly varying. Given a sequence of noisy measurements, we study two problems seperately: 1) the problem of testing whether or not there is a chirp hidden in the noisy data, and 2) the problem of estimating this chirp from the data.

This thesis introduces novel, flexible and practical strategies for addressing these important nonparametric statistical problems. The main idea is to calculate correlations of the data with a rich family of local templates in a first step, the multiscale chirplets, and in a second step, search for meaningful aggregations or chains of chirplets which provide a good global fit to the data. From a physical viewpoint, these chains correspond to realistic signals since they model arbitrary chirps. From an algorithmic viewpoint, these chains are identified as paths in a convenient graph. The key point is that this important underlying graph structure allows to unleash very effective algorithms such as network flow algorithms for finding those chains which optimize a near optimal trade-off between goodness of fit and complexity.

Our estimation procedures provide provably near optimal performance over a wide range of chirps and numerical experiments show that both our detection and estimation procedures perform exceptionally well over a broad class of chirps. This thesis also introduces general strategies for extracting signals of unknown duration in long streams of data when we have no idea where these signals may be. The approach is leveraging testing methods designed to detect the presence of signals with known time support.

Underlying our methods is a general abstraction which postulates an abstract statistical problem of detecting paths in graphs which have random variables attached to their vertices. The formulation of this problem was inspired by our chirp detection methods and is of great independent interest.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This thesis considers the problem of detecting and estimating highly oscillatory signals from noisy measurements. The statistical model assumes that the data is sampled uniformly and is of the form

$$y_k = \alpha f_k + z_k, \quad k = 0, 1, \ldots, N-1; \tag{1.1}$$

where (f_k) is an unknown vector with samples of a signal $f(t)$ of interest, (z_k) is a random vector with mean zero and a known distribution, and α is an unknown scalar. We will study two different statistical questions:

- **Detection.** *Is there a signal hiding in the noise?* More formally, given the data as in (1.1), we would like to conduct a hypothesis test to decide between

  $$H_0 : \alpha = 0 \quad \text{(data is only noise)}$$

  vs. the alternative

  $$H_1 : \alpha \neq 0 \quad \text{(data is only noise)}.$$

- **Estimation.** *How well can we recover the signal from the noise?* Given data of the form (1.1), we wish to recover f. The goal is to minimize the error of the estimation as measured by a some loss function, which quantifies how far away the estimated signal is from the truth.

We would like to stress that the object f is assumed to be unknown in the sense that it does not depend upon a small number of parameters. In fact, we assume that f belongs to an infinitely dimensional functional class. These problem settings are usually referred to as *nonparametric testing* and *nonparametric estimation*, as opposed to the classical finite-parametric case. The study of these problems depends heavily on the class of unknown signals one considers. The classical results for the nonparametric estimation problem assume the objects are spatially homogeneous and obey smoothness conditions. The estimation procedures often rely on local averages of the noisy measurements [75]. In the early 1990s, Donoho and his collaborators established important links between nonparametric statistical estimation, wavelets and modern harmonic analysis in general [35, 36]. This has initiated a whole new industry of the exploration of new statistical methods based on recent advances in applied harmonic analysis (for surveys on this topic, see [21, 53]). Theoretical studies for nonparametric detection have, however, not reached the same level of maturity, and have focused on similar smoothness classes (e.g., Sobolev spaces) as in the theory of nonparametric estimation. See, for example, [50] and references therein.

The signals we wish to study, so-called *chirps*, are very different than those that have been studied in nonparametric statistics in the past, and therefore call for new methodologies.

1.1 Chirps

The objects of our main focus, chirps, are oscillatory signals whose frequency can vary with time, and are in fact parts of people's daily life. English-speaking people should be familiar with the everyday use of the word, as the word "chirp" is often used to signify the sound made by a bird singing. Another notable example of a chirp is the sound when a fire engine or a train passes by. The sound of the siren or the train whistle becomes higher in pitch as the vehicles approach us, and lower as they move away. This is the result of the relative motion of the sound source to the observer; the well-known *Doppler effect* we all read about in science textbooks at an early age.

In mathematical terms, we can describe chirps as oscillatory signals whose frequency changes with time and take the general form

$$f(t) = A(t)\cos(\lambda\phi(t)); \tag{1.2}$$

where t represents time, $\lambda > 0$, and the amplitude A and phase ϕ are time varying. For now, we think about A being smooth and the oscillation degree λ as being large. That is, A is an "envelope" in which $\cos(\lambda\phi(t))$ oscillates.

There are plenty of examples of chirps in the nature. Besides the before mentioned example of birds, another example is echolocation in some species of mammals. Bats and dolphins have evolved sonar systems to locate food and navigate through skies and the ocean [67, 73, 74, 79]. Some species of bats are known to emit chirps to help them navigate and humpback whales are known to sing long and complex songs to communicate with other whales. Better tools for analyzing the songs of humback whales could help us to learn more about the habits of these species [78].

In science and technology chirps are used, for example, in remote sensing, radar, and sonar. We can learn about velocity and location of moving objects by emitting electromagnetic waves towards them and recording the echo. As in the case of the train and the siren, the Doppler effect implies that the received signal has a frequency evolving with time, which is, for example, the principle behind automobile radars. Another example is that doppler radar return from a small piece of ice floating in an ocean environment is chirp-like [46]. In communication systems, information is transmitted by a transmitting chirps, $f(t) = \cos(\theta(t))$, where the sender encodes information by phase and frequency modulation. This is done by tuning $\theta(t)$ and, in turn, changing the rate of oscillation of $f(t)$ [70].

The final example we wish to mention is gravitational waves in astrophysics, whose existence has been predicted by the theory of general relativity. These propagating waves are disturbances in the curvature of space-time caused by the motion of matter. An example of a source of gravitational waves is a system of two massive objects (e.g., black holes or neutron stars) orbiting each other. General relativity predicts that this system loses energy in the form of gravitational radiation, which causes the objects spiral in towards each other, eventually merging in a violent event. These type of gravitational waves can be modeled as chirps [4, 5]. There are observations that binary pulsar systems are losing energy at the rate predicted by the theory of general relativity, providing strong evidence that gravitational waves truly exist [47, 76]. Other than that, gravitational waves have not been verified directly by experiments [80]. Enormous effort is being put today in verifying the existence of gravitational waves directly by measurements based on interferometry. The biggest effort is the Laser Interferometric Gravitational-wave Observatory (LIGO) [1]. Other gravitational wave detectors are VIRGO and GEO in Europe, and TAMA in Japan. Nonparametric detection and estimation strategies could provide a helpful tool for data exploration in gravitational wave searches.

1.2 Time-Frequency Analysis

The field of time-frequency analysis dates back to Dennis Gabor in 1946 [41] and Jean Ville in 1948 [82]. It considers the problem of developing mixed representations of a signal in terms of a double sequence of elementary signals, each being localized in time and frequency. Ville was motivated by music and took the example of a musical passage where a note at certain frequency can only be heard when it is played. Although the musical signal could be represented mathematically by Fourier analysis, such a representation can only tell which notes have been played but cannot say at which time during the passage. Ville then imagined that we could define an *instantaneous frequency* of a signal as a function of time, which describes the structure or frequency – in the usual Fourier sense – of the signal at a given instant. Gabor was the first to introduce time-frequency wavelets [41], which are now called Gabor wavelets, and his idea was to divide an oscillatory signal, or a wave, into little wavelets or time-frequency atoms. These are functions defined by

$$w(t) = g((t - t_0)/s)e^{\imath\omega t}, \tag{1.3}$$

where g is a Gaussian window, $s > 0$, and $t_0, \omega \in \mathbb{R}$; we can interpret ω as the average frequency of the wavelet, and $t_0 - s$ and $t_0 + h$ as the start and finish of the "note" which it plays. This defines a collection Ω of *time-frequency atoms*, and the idea is to write a signal f as a series $\sum_j \alpha_j w_j$, where α_j is a scalar and $w_j \in \Omega$. Since Gabor's proposal, other collections Ω have been developed, such as Liénard's time-frequency atoms, Malvar-Wilson wavelets, and chirplets (see [52]). With all these collections to choose from, we can pose two questions:

(i) To study a given class of signals, which collection should we choose?

(ii) For a given Ω, how should we find the "best" representation, i.e., a linear combination of time-frequency atoms, of the signals of interest?

There is no simple answer to the first question, but the usual criterion is that the time-frequency atoms should have the same "appearance" as the signal (or pieces of it) [52]. People have proposed several answers to the second question. For example, Coifman and Wickerhauser's best-basis algorithm [28] (1992), Mallat's matching pursuit algorithm [58] (1993), and the basis pursuit [27] (1998). At this point, we would like to mention that the statistical methods introduced in this thesis, could provide – as a side-product – a new answer to the second question, at least in the case of finding good approximations of chirp signals. This methodology is quite different from what these previous methods have proposed.

1.2.1 Instantaneous frequency

When listening to music we hear tones, or frequencies, changing with time. Ville's idea was to define the notion of *instantaneous frequency* of chirps, which could describe their rate of oscillation at each instant. A cosine of the form

$$f(t) = a \cdot \cos(\omega_0 t + \theta_0) = a \cdot \cos(\phi(t)),$$

where $\omega_0, \theta_0, a \in \mathbb{R}$, has frequency ω_0, which is the derivative of its phase $\phi(t)$. To generalize this, we consider (1.2) and define its instantaneous frequency as

$$\omega(t) = \lambda \phi'(t),$$

where we adapt the sign of $\phi(t)$ such that $\omega(t) > 0$. This can be an ill-defined term, since the pair (A, ϕ) in (1.2) is not necessarily unique; a simple example is to take $A(t) = \cos(ct)$, $c \in \mathbb{R}$, and switch the roles of $A(t)$ and $\cos(\lambda\phi(t))$ in (1.2). To get a unique representation, we consider the analytic part f_a of f, which is defined from the Fourier transform: $\hat{f}_a(\omega) = 2\hat{f}(\omega)$ if $\omega \geq 0$, and 0 otherwise. This complex signal can be represented uniquely in the form

$$f_a(t) = A(t) \exp(i\phi(t)),$$

where $A(t)$ is a real-valued function with $A(t) > 0$ for all t. By the construction of f_a, we have a unique representation

$$f(t) = Re(f_a) = A(t) \exp(i\phi(t)).$$

But this does not necessarily mean that ϕ' really gives some information on how the frequency content of f changes with time. We need conditions for the variations in the envelope $A(t)$ to not get in the way of determining the instantaneous frequency. Conditions where this holds can be described roughly as:

$$\left|\frac{A'(t)}{A(t)\phi'(t)}\right| \ll 1, \qquad \left|\frac{\phi''(t)}{(\phi'(t))^2}\right| \ll 1. \tag{1.4}$$

The first condition requires the envelope $A(t)$ to change little on a scale given by the local "period," $1/\phi'(t)$, and the second puts a bound on how fast $\phi'(t)$ can change in time. These conditions have been used to define chirps in signal processing (see [25]). Now consider the windowed Fourier transform defined by

$$Sf(u,\xi) = \langle f, g_{s,u,\xi}\rangle, \qquad g_{s,u,\xi}(t) = s^{-1/2} g((t-u)/s) e^{i\xi t},$$

where g is a smooth symmetric window supported in $[-1/2, 1/2]$. If the conditions (1.4) hold, $|Sf(u,\xi)|$ would have large values along the curve $\xi = \varphi'(u)$ in the "time-frequency" plane (u,ξ), and decay away from it.

1.2.2 Chirplets

Multiscale chirplets were used in [23] in the context of chirp detection and are an ingredient in the methods presented in this thesis. Prior to that that, Candés used windowed multiscale chirplets in [20] for the estimation of chirps. However, the term "chirplet" was introduced by Haykin and Mann, who wrote the first published reference to chirplets in [60] (see also [12, 61]). They proposed a new transform, which in its simplest form for signals of dimension one, is based on windowed chirplets

$$w_\lambda(t) = g((t-t_0)/s) e^{\imath(\omega t + \delta t^2)}, \quad \lambda = (s, t_0, \omega, \delta), \tag{1.5}$$

where g is a Gaussian window, $s > 0$, and $t_0, \omega, \delta \in \mathbb{R}$. They define the chirplet transform of a signal f as the collection of inner products $\langle f, w_\lambda\rangle$. These chirplets were motivated by Gabor's wave atoms, which are a of the same form as (1.5) but with the chirping parameter δ equal to 0. More general sets of chirplets can be constructed by considering chirp atoms having polynomial phase (e.g., piecewise cubic, piecewise quadratic, etc.), and even sinusoidally varying phase. What is new in the methodology we proposed in [20], and separates our constructions from theirs, is the notion of the *chirplet graph*: the idea of modeling chirps by looking at chirplets as nodes in a graph with directed edges connecting the vertices.

1.3 The Detection Problem

1.3.1 Measures of performance

In hypothesis-testing problems there are two types of errors we can make:

1. *False Detection:* Decide that H_1 is true when H_0 is true; i.e., decide there is a signal hiding in the data when the measurements are only noise.

2. *False Dismissal:* Decide that H_0 is true when H_1 is true; i.e., decide the data is only noise when there is a signal hiding in it.

In the literature these are usually referred to as *Type I* and *Type II* errors, respectively. Because the data is subject to stochastic noise, we measure the average performance of a decision rule. That is, the probability of committing either of these errors.

The probability of Type I error is well-defined in our statistical problem (1.1) since the distribution of the data under H_0 is fully determined. On the other hand, the probability of Type II error depends on the signal that is hiding in the noise and the signal level α. The two main approaches in statistical decision theory to deal with this are the *Minimax* and *Bayesian* paradigms, described in Section 9.2.

1.3.2 Current detection strategies

Since the problem of study involves nonparametric classes of chirps, we cannot assume that we can approximate the family of signals with finite collections of the form $\{f_\theta : \theta \in \Theta\}$. But if that was the case, a popular detection strategy based on the GLRT paradigm, so-called "method of matched filters" [65], could be applied. In the case of additive Gaussian noise, the decision would be based on the statistic

$$T^* = \max_{\theta \in \Theta} \frac{|\langle y, f_\theta \rangle|^2}{\|f_\theta\|^2},$$

and T^* compared to a fixed threshold. The problem with this method is that unless the set Θ is moderate, the cost of computing T^* could be enormous. Also, if we do not have reliable parametric models for approximating the unknown waveforms, the method could lack robustness.

In the last decades, researchers in the field of time-frequency analysis have assembled a great collection of tools. Various methods utilizing these tools have been proposed for detecting chirps in noisy data. Not all of these methods are designed for detecting nonparametric classes of chirps, and some are instead aimed at overcoming some of the computational burden and robustness issues of matched filters. This can come at the cost of sacrificing statistical sensitivity.

For example [51, 63] propose to look for ridges in the time-scale plane. First the continuous time wavelet transform $W(s,t)$, where $s > 0$ is scale and t is time, is computed. Then the goal is to search for a curve $\rho(s)$ along which the sum of $\int |W(s,\rho(s))|^2 ds/s$ is maximum. Searching the whole space of curves is virtually impossible, so the method is restricted to parametric "power-law chirps" of the form $f(t) = (t_0 - t)_+^\alpha \cos(2\pi F_\beta (t_0 - t)^{\beta+1})$, where α and β are unknown and F_β is some constant. Thus, the method only considers a parametric problem. It is more robust than the method of matched filters for this parametric class, but instead sacrifices power and is less sensitive.

A similar method, based on searching for ridges in the time-frequency plane, is proposed in [25]. This method focuses also on power-law chirps. It uses time-frequency distributions to try localizing the unknown signal in the time-frequency plane. The choice of a suitable time-frequency distribution depends on the parameters α and β for the power-law chirp. For example, although the Wigner-Ville distribution is ideal for linear chirps, it works badly for hyperbolic chirps (a hyperbolic chirp is a function of the form $f(t) = \cos(a/(b-t))$, for $0 \leq t < \beta$). In this case another time-frequency distribution needs to be chosen. The interferences in the Wigner-Ville distribution are also problematic. When attempting to suppress them using averaging methods, we sacrifice time-frequency resolution. This affects the performance of the detection strategy [25].

To name a few nonparametric methods, we would like to mention [5]. It starts off by getting a time-frequency portrait of the data using the Wigner-Ville distribution. This gives rise to a time-frequency distribution $\rho(t,\omega)$ for discrete values of times $t_0 \leq t \leq t_1$ and frequencies $\omega_0 \leq \omega \leq \omega_1$. Then it considers the points (t,ω) as being "pixels" in a region $[t_0,t_1] \times [\omega_0,\omega_1]$ and $\rho(t,\omega)$ the pixel level at (t,ω). Then it searches for ridges or edges in this portrait, based on ideas that have been used for finding edges in noisy images. Those pixels with levels exceeding a threshold are tagged as "ridge points" and the statistical test is based on the length of the longest ridge. Because of the use of the Wigner-Ville distribution there will be interferences and even a clean signal would take nonzero values in regions which have nothing to do with its frequency content. Again, one has to suppress the interferences which causes loss in resolution and smearing of the true ridges. Another important issue is that although the chirp may not be locally detectable, it could still be detectable by other methods. This could mean that most of the true ridge points will not be tagged. An example which might demonstrate this can be found in the numerical simulations for the GLRT in the square lattice in Chapter 9.

Finally, while working on this project, we became aware of the recent and independent work of Chassande-Mottin and Pai [26], who propose a detection method that is similar in spirit to what we present in Chapter 3. However, we are neither aware of their method having been extended to deal with long streams of data as we have done for our method and is described in Chapter 4, nor do we know, whether any statistical theory has been developed as we do here in Chapters 7 and 8. We discuss the connections and differences between the two methods in Section 3.5.1.

1.4 The Estimation Problem

1.4.1 Measures of performance

To quantify the performance of an estimation procedure we need to choose a *loss* function. For example, in image or speech processing we would wish to have a quantitative way of measuring visual and audio degradation. Understandably, such perceptual measures can be hard to model mathematically. Perhaps the most popular choice is to use a quadratic loss function

$$l(f, \hat{f}) = \frac{1}{N}\|f - \hat{f}\|_{\ell_2}^2 = \frac{1}{N}\sum_{k=0}^{N-1} |f_k - \hat{f}_k|^2;$$

because of the $1/N$ factor, we can interpret this as the quadratic error per coordinate. Besides making sense intuitively, this loss function is mathematically attractive for its simplicity, nice properties, and geometric interpretation, and it has often proven to lead to useful practical procedures. With a loss function in hand, we define the *risk* of the estimator as $E(l(f, \hat{f}))$, where the expectation is taken with respect to the distribution of the noise process (z_k). We wish to recover f with a small mean-squared error (MSE)

$$MSE(f, \hat{f}) := E\left[\frac{1}{N}\sum_{k=0}^{N} |f_k - \hat{f}_k|^2\right]. \tag{1.6}$$

1.4.2 Current estimation strategies

Perhaps our benchmark method for the estimation of chirps is the procedure Candès introduced in [20] for estimation of nonparametric classes of chirps. We include a description of this in Section 5.6. Other possible methods that could be suitable for our setup are procedures based on time-frequency dictionaries. For example thresholding in "best-basis" using local cosines [83].

1.5 Detection vs. Estimation in the Nonparametric Case

As we discussed in the beginning of the introduction, the theory of nonparametric estimation is fairly developed, and studies under this setting have generated a large literature over the last few decades. Nonparametric testing problems have, however, not drawn as much attention. This situation is very different from the classical parametric case, where these problems where studied in parallel. In fact, for nonparametric setups as in (1.1), we often have situations were for a certain range of signal strengths, the signals can be detected with full power, while the accuracy of any estimator would be intolerable (see [50]).

1.6 The Goal

Our goal is to introduce new tools to solve applied statistical problems involving chirps. Therefore we not only wish to find statistical procedures that have provable, very good statistical performance (hopefully optimal or near-optimal) for the problems of interest, we also aim at designing methods that are practical in the sense that we can use fast algorithms to apply them to data. Satisfying both of these requirements simultaneously is challenging.

Our methodology is partly motivated by the idea of time-frequency representations. Chirps with smoothly evolving instantaneous frequency have a simple local structure, and in our case we will model them as having linear instantaneous frequency at a small enough scale (i.e., as chirplets). At a global scale, our models are chains of chirplets which form a chirp with polygonal instantaneous frequency. We can think about the chirplets as vertices in a graph with edges between chirplets whose instantaneous frequency is such that their juncture does not differ by much. The statistical methods based on this framework search for paths in the graph which give a good trade-off between goodness of fit and complexity. The graph structure provides us both with the possibility of designing rapidly computable methods, and some ground to do theoretical studies.

1.7 Influences

The main inspiration for our methods comes from [32]. It introduced a framework for detecting curves in noisy images using chains of beamlets, which are line segments occupying a range of dyadic locations, scales, and orientations. Another close relation is [11], where a graph structure was used to detect high concentrations of points along smooth curves in a background of uniform random points. As far as we know, ideas of this type can be traced back to Sha'ashua and Ullman [72], who proposed to use a graph structure for searching for salient structures in images. Our detection methods are different than those above and also consider different types of data models.

1.8 Organization

The thesis can be divided into two parts. The first part introduces the methodology behind our statistical procedures for detecting and estimating chirps. As a "proof-of-concept" that the methods are powerful and practical, we present results from numerical experiments. In the second part we develop statistical theory as an attempt to rigorously quantify, in a precise mathematical sense, the performance of the methods. This leads us to the study of an abstract statistical problem that is by itself of great interest and importance. A brief description of the organization of the thesis is as follows:

- **Part I: Methodology**
  - **Chapter 2:** Description of multiscale chirplets and the definition of the chirplet graph
  - **Chapter 3:** Introduction of methods for detecting chirps on a interval of data, supplemented with numerical simulations

  - **Chapter 4:** Extension of the methods in Chapter 3 for detecting chirps of unknown support in long data streams. Includes numerical simulations.
  - **Chapter 5:** Estimation procedures for noisy chirps based on similar ideas as the statistical tests in Chapter 3. Includes a short section with simulations.

- **Part II: Theory**
  - **Chapter 6:** Mathematical description of a rich nonparametric class of chirps for theoretical study. Investigations of approximation properties of chirplet paths for this class. These results are needed for developing statistical theory in subsequent chapters, but are also of independent interest.
  - **Chapter 7:** Study of the theoretical performance of the estimator presented in Chapter 5. Shows that it is optimal for estimating the chirps defined in Chapter 6 in the case when the regularity of the chirps is assumed to be known. In the case when the regularity is unknown, we show that the method is near-optimal, in the sense that it comes within a logarithmic factor of the ideal risk.
  - **Chapter 8:** Study of statistical properties of the method presented in Chapter 3, giving conditions for guaranteed good performance
  - **Chapter 9:** Study of a new abstract detection problem of detecting paths in graphs. The formulation of the problem was motivated by the test procedure we designed for chirp detection.

The appendices include supplementary material for the chapters, such as proofs of lemmas and theorems, and information regarding the numerical experiments. Appendix F is a short theoretical study on the problem of detecting sinusoids of unknown support and frequency in long streams of data. It identifies the threshold of detectability for this problem.

1.9 Credits

Chapter 2 is based on joint work with Emmanuel Candés and was published in [23]. Section 3.2 is included almost unchanged from [23]. It was written jointly with Emmanuel and is mostly his prose based on a draft by me.

Chapter 3 is based on joint work with Emmanuel Candés and Philip Charlton and published in [23]. Section 3.4 is included almost unchanged from [23] and written jointly with Emmanuel. It is Emmanuel's prose that is included, while the simulations were run by me. Emmanuel wrote Section 3.1.1 and Section 3.5, which are taken from [23]. Section 3.3 was written jointly by me and Emmanuel, but his writing is what is included. However, Section 3.3.6 is recent work written by me and did not appear in the paper.

Chapter 4 is joint work with Emmanuel Candés and Philip Charlton. The material in the chapter is being prepared for submission to a journal. Philip Charlton deserves the credit for the gravitational wave signal model used in the simulations and my gratitude for allowing me to use his text with the description of this model in Appendix E. He also provided the code for generating these waveforms which was based on a Maple program, supplied by Warren Anderson, for simulating binary black hole coalescence.

Chapter 9 is joint work with Emmanuel Candés, Ery Arias-Castro, and Ofer Zeitouni. Parts of this work were published in [9]. In particular Section 9.6, which is included in the thesis almost unchanged from the publication, was written jointly with Emmanuel and Ery.

Appendix B is joint work with Emmanuel Candés and written jointly with him. It was sent to the editor of The Journal of Applied and Computational Harmonic Analysis with our revision for [23] prior to its publication.

Chapter 2

The Chirplet Graph

We start off by describing the main mathematical architecture behind the statistical procedures that will be presented for chirp detection and estimation. First we define a dictionary of multiscale chirp atoms called *chirplets* which provide good local approximations of a wide range of chirps. Then we introduce the notion of a *chirplet graph*, which is the essential ingredient of our methods, allowing us to build rapidly computable statistical estimators and detectors, using efficient algorithms from the literature of network flow algorithms. The graph is also a useful vehicle for establishing theoretical results, and makes the methods quite general and extendable to other statistical problems.

2.1 Multiscale Chirplets

Although the chirps we wish to consider are too complex to be modeled by a simple parametric class of functions, their local structure might be simple enough to be captured well by a simple parametric model. Our methods use a dictionary of *multiscale chirplets*, which are oscillating functions, supported on dyadic time-segments of varying lengths and whose phase varies quadratically with time. Such functions provide a good local approximation of a wide range of chirps.

To be more specific, assume we work in the time interval $[0, 1]$ and the data is evenly sampled. Let $I = [t_0, t_1]$ be a dyadic interval such that for any integer $j \geq 0$,

$$I = [k2^{-j}, (k+1)2^{-j}), \quad k = 0, 1, \ldots, 2^j - 1.$$

The *multiscale chirplet dictionary* is the family of functions defined by

$$c_{I,\mu}(t) := e^{\imath\left(a_\mu(t-t_0)^2/2+b_\mu(t-t_0)\right)}\, 1_I(t)\, |I|^{-1/2}, \tag{2.1}$$

where $(a_\mu, b_\mu) \in \mathcal{M}_j$ is a discrete collection of parameters; a_μ is called the *slope* of the chirplet and b_μ is called the *frequency offset* of the chirplet. We will use the term *scale* interchangeably for the time support of a chirplet, $|I| = 2^{-j}$, or simply the the index j – the meaning should be clear from the context. The parameters a_μ and b_μ may depend on scale and any prior information we have about the chirps of interest. For example, the range of the slope parameter a_μ may depend on how rapidly the instantaneous frequency of the chirps under consideration can change. Note that chirplets are normalized such that

$$\|c_{I,\mu}\|_{L_2} = 1.$$

The chirplet dictionary has elements of various durations, locations, average frequencies and chirp rates. It is convenient to think of chirplets as line segments in the time-frequency plane, as a time-frequency portrait, such as the Wigner-Ville distribution, would reveal (see, for example, [39, 57]). One can think of the "instantaneous frequency" of a chirplet as being linear and equal to $a_\mu t + b_\mu$ over its time duration. Figure 2.1 shows a diagrammatic representation of two chirplets at different scales.

2.1.1 Discrete chirplets

For evenly sampled signals, discrete chirplets are the discrete-time waveforms

$$c_{I,\mu}[n] = c_{I,\mu}(n/N)/\sqrt{N}, \tag{2.2}$$

with $n = 0, \ldots, N-1$. Note that these signals are normalized such that

$$\|c_{I,\mu}\|_{\ell_2} = 1.$$

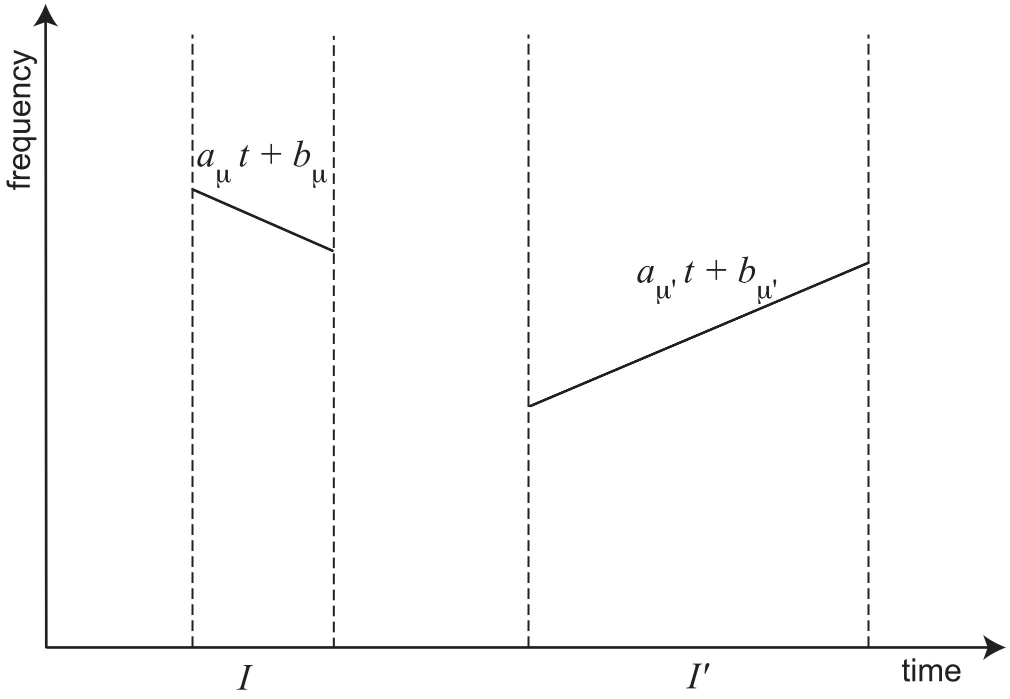


Figure 2.1: A diagrammatic representation of two chirplets in the time-frequency plane

For a fixed set of chirplet parameters (a_μ, b_μ) and a collection of chirplet scales, we define the *chirplet analysis* or *chirplet transform* of a signal f of length N as the set of inner products [1]

$$\langle f, c_{I,\mu} \rangle = \sum_{n=0}^{N-1} f[n] c^*_{I,\mu}[n],$$

with all the elements in the chirplet dictionary.

2.1.2 Recursive dyadic partitions

Partitions of the base interval $[0, 1]$ into dyadic intervals are usually called *Recursive Dyadic Partitions* (RDPs), and are quite classical in the literature of time-frequency analysis. We give here a formal definition of RDPs and the special case of *balanced* RDPs which will be used later in the thesis:

Definition 1. *A recursive dyadic partition (RDP) is any partition $\mathcal{P}$ constructed according to the following rules:*

- $\mathcal{P} = \{[0, 1]\}$ *is an RDP;*
- *Let $\mathcal{P} = \{I_1, \ldots, I_m\}$ be an RDP. Then a partition obtained by splitting any interval $I_j \subset \mathcal{P}$ into two adjacent dyadic interval and leaving the others the same is also an RDP.*

We say that $\mathcal{P}$ is a balanced *RDP (BRDP) if it is an RDP such that any two adjacent intervals $I, I' \in \mathcal{P}$ obey*

$$\frac{\max(|I|, |I'|)}{\min(|I|, |I'|)} \leq 2.$$

Villemoes uses the definition of BRDPs in [83] when introducing adapted bases of local cosines which satisfy a uniform bound on their time-frequency concentration thanks to this condition. He argues that this restriction on the allowed segmentations is not a big price to pay, considering the sizes of two adjacent intervals in an RDP can only differ greatly at special dyadic locations. Candés also uses BRDPs in [20] for his construction of libraries of tight frames of multiscale chirplets (described in the Appendix). In our case BRDPs can also be useful since they decrease the number of edges in the chirplet graph without sacrificing much (or in some cases, any,) adaptivity for a wide range of chirps (see Section 6.2.1.1).

[1] Here z^* stands for the complex conjugate of the complex number z.

2.2 The Chirplet Graph

The motivation behind the graph structure is to use a linear combination of chirplets to model a chirp, $f(t) = A(t)\cos(\lambda\varphi(t))$ or $f(t) = A(t)\exp(i\lambda\varphi(t))$, by approximating its instantaneous frequency $\lambda\varphi'(t)$ by a piecewise linear curve in the time-frequency plane. A *chirplet graph* $G = (V, E)$ is a set of vertices (or nodes), V, and edges (or arcs), E. Each vertex corresponds to a chirplet indexed by $v = (I, \mu)$; I being its time support and $\mu = (a_\mu, b_\mu)$ being the index for the slope and frequency offset parameters for a chirplet as described in (2.1). Restricting us to chirps with smooth phase and well-defined instantanous frequency leads us to imposing some natural conditions for the edges in the chirplet graph to satisfy:

1. Two chirplets can only be connected if they have adjacent supports in time.
2. Two chirplets are connected if the difference in frequency at the juncture is small.
3. Two chirplets are connected if the difference in their slopes is not too large.

The rules for connecting nodes in the chirplet graph are called *connectivity constraints*, or *connectivities*. Figure 2.2 gives diagrammatic examples of some admissible and inadmissible connectivities in the chirplet graph.

2.3 Discretization

For the sake of concreteness, we will describe a particular discretization of the chirplet parameters and connectivities for the chirplet graph. A similar configuration will later be used in some of our numerical experiments. In practice one might of course wish to choose another discretization, based on the application and any prior knowledge about the unknown signals. Some more rigorous guidelines for choosing configurations can be found in later chapters of the thesis; in particular, in Chapter 6.

Let's assume the signal is evenly sampled and of dyadic length $N = 2^J$ for some positive integer J. Consider the portion $[0, 1] \times [-\pi, \pi]$ of the "time-frequency" plane. For each dyadic interval $I = [t_I, t_{I'}] = [k2^{-j}, (k+1)2^{-j}]$, mark out two vertical lines at the endpoints of I similar to what is done on Figures 2.2 and 2.3. Place ticks at spacing $2\pi/N$ along these vertical lines. Then connect the ticks between the two vertical lines and let each such line

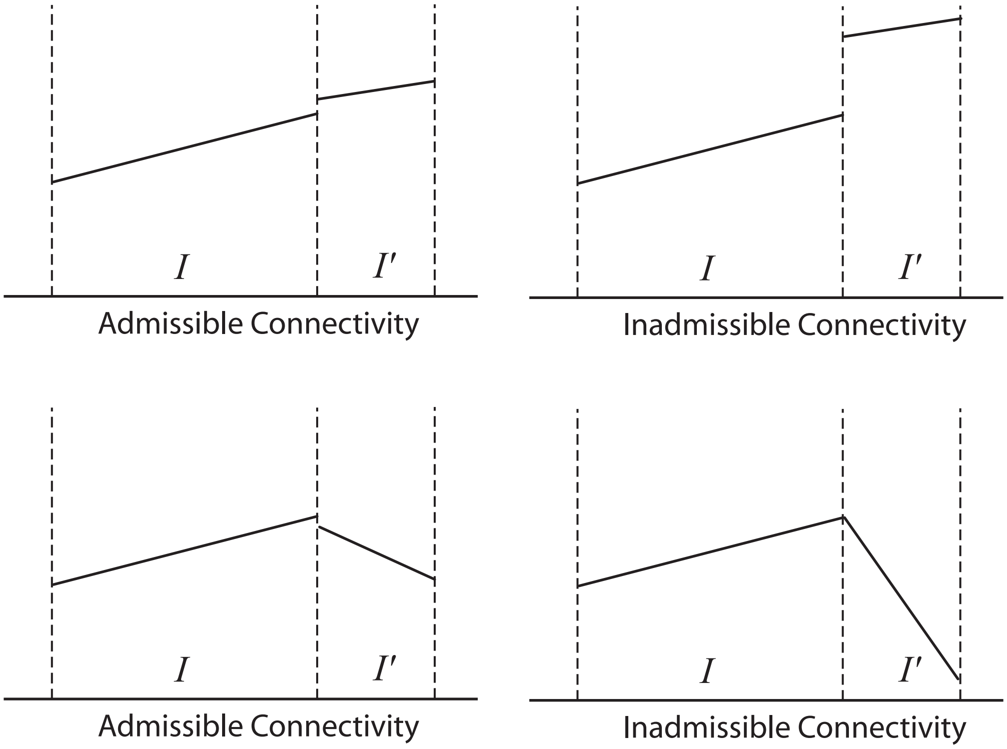


Figure 2.2: Examples of connectivities in the chirplet graph. Each line segment represents the instantaneous frequency of a chirplet. Chirplets may not be connected when the difference in offsets and/or slopes is large.

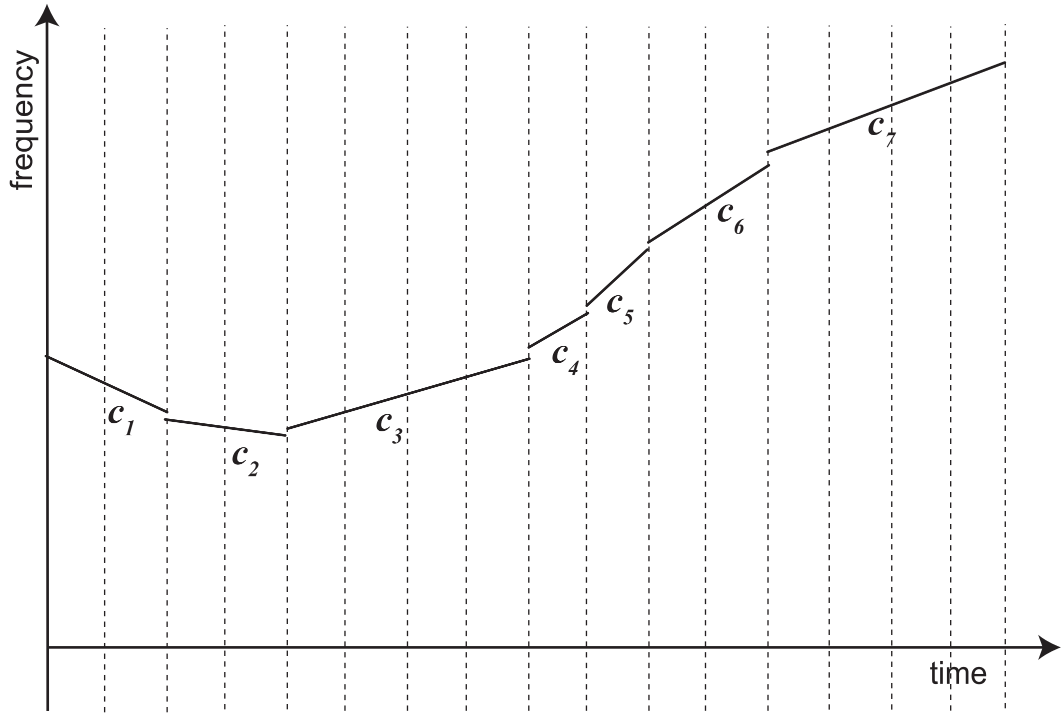


Figure 2.3: Diagrammatic representation in the time-frequency plane of the instantaneous frequency along a chirplet path

segment correspond to a chirplet; the position of the tick mark at the start of the line segment is the parameter b_μ in (2.1), and the slope is a_μ. This builds a dictionary of chirplets. For sake of simplicity, let's suppose we only have to create such lines with a slope—in absolute value—less or equal to 2π to "capture" the phase of the unknown chirp. By counting, the number of slopes is about $2N \cdot 2^{-j}$, and therefore the number N_j of chirplets per dyadic interval obeys

$$N_j = \#\,\text{offsets } \times \#\,\text{slopes} \approx N \times 2N/2^j.$$

At scale 2^{-j} we have 2^j dyadic intervals, and therefore about

$$2^j\, N_j \asymp 2N^2$$

chirplets. Thus, considering a range of scales $j \in \{0, \ldots, N-1\}$ we see that the size of this chirplet dictionary is

$$O\left(N^2 \log N\right).$$

It is clear that one can use the FFT to compute the chirplet coefficient table. For example, with the above discretization, it is possible to compute all the coefficients against chirplets "living" in the fixed interval $[k2^{-j}, (k+1)2^{-j})$ in $O(N_j \log(N/2^j))$ flops so that the computational complexity of the chirplet transform is $O(N^2 \log^2 N)$. There are many other possible discretizations and the experienced reader will also notice that, for regular discretizations, the complexity will scale as $O(M_N \log N)$, where here and below M_N is the number of chirplets in the dictionary. In summary, the computational cost is at most of the order $O(\log N)$ per chirplet coefficient.

Chapter 3

Detection of Chirps by Chirplet Path Pursuit

This chapter considers the problem of detecting chirps from noisy measurements. Suppose we have noisy sampled data

$$y_k = \alpha\, f_k + z_k, \qquad k = 1, \ldots, N, \tag{3.1}$$

where the unknown vector (f_k) consists of sampled values $f_k = f(t_k)$ of a signal of interest $f(t), t \in [0,1]$ belonging to a class of functions $\mathcal{F}$. (z_k) is a zero-mean random sequence, not necessarily i.i.d. but with a known distribution. Based on the observations (y_k), one would like to decide whether or not a signal is hiding in the noise; i.e., we would like to test the null hypothesis

$$H_0 : \alpha = 0 \text{ (noise only)}$$

against the alternative

$$H_1 : \alpha \neq 0 \text{ (signal is buried in noise).}$$

It is important to emphasize that we assume that we may not be able to model the functions from $\mathcal{F}$ by a parametric model depending upon a small number of parameters. This puts us in the situation of *nonparametric testing.* Although we will be considering $\mathcal{F}$ as being a set of chirps, the methodology we are about to present can be deployed to other detection problems; such as detection of curves in two-dimensional data.

3.1 Detection Statistics

We now describe the complete algorithm for searching for chirps through the data. To explain our methodology, it might be best to first focus on the case of additive Gaussian white noise

$$y_i = \alpha S_i + z_i, \quad i = 1, \dots, N, \quad z_i \text{ i.i.d. } N(0, \sigma^2).$$

We wish to test $H_0 : \alpha = 0$ against $H_1 : \alpha \neq 0$. A general strategy for testing composite hypotheses is the so-called *Generalized Likelihood Ratio Test* (GLRT). We suppose that the set of alternatives is of the form λf where λ is a scalar and f belongs to a subset $\mathcal{F}$ of unit vectors of $\mathbf{R}^N$, i.e., obeying $\|f\| = 1$ for all $f \in \mathcal{F}$ (unless specified otherwise, $\| \cdot \|$ is the usual Euclidean norm). In other words, the alternative consists of multiples of a possibly exponentially large set of candidate signals. In this setup, the GLRT takes the form

$$\max_{\lambda \in \mathbf{R}, f \in \mathcal{F}} \frac{L(\lambda f; y)}{L(0; y)}, \tag{3.2}$$

where $L(\lambda f; y)$ is the likelihood of the data when the true mean vector is λf. In the case of additive white noise, a simple calculation shows that the GLRT is proportional to

$$\max_{\lambda \in \mathbf{R}, f \in \mathcal{F}} \quad e^{-\|y - \lambda f\|^2 / 2\sigma^2} = \max_{f \in \mathcal{F}} \quad e^{-\|y - \langle y, f \rangle f\|^2 / 2\sigma^2},$$

since for a fixed $f \in \mathcal{F}$, the likelihood is maximized for $\lambda = \langle y, f \rangle$. It then follows from Pythagoras' identity $\|y - \langle y, f \rangle f\|^2 = \|y\|^2 - |\langle y, f \rangle|^2$ so that the GLRT is equivalent to finding the solution to

$$\max_{f \in \mathcal{F}} \quad |\langle y, f \rangle|^2,$$

and comparing this value with a threshold.

3.1.1 The Best Path statistic

Supplied with a chirplet graph, a reasonable strategy would be to consider the class of signals which can be rewritten as a superposition of piecewise linear chirps

$$f(t) = \sum_{v \in W} \lambda_v f_v(t),$$

where W is any path in the chirplet graph and (λ_v) is any family of scalars, and apply the GLRT principle. In this setup, the GLRT is given by

$$\max_{W} \max_{(\lambda_v)} e^{-\|y-\sum_{v\in W}\lambda_v f_v\|^2/2\sigma^2} = \max_{W} \max_{(\lambda_v)} \prod_{v\in W} e^{-\|y_v-\lambda_v f_v\|^2/2\sigma^2},$$

where for each $v=(I,\mu)$, y_v is the vector $(y_t)_{t\in I}$, i.e., the portion of y supported on the time interval I. Adapting the calculations detailed above shows that the GLRT is then equivalent to

$$\max_{W} \sum_{v\in W} |\langle y, f_v\rangle|^2. \tag{3.3}$$

In words, the GLRT simply finds the path in the chirplet graph which maximizes the sum of squares of the empirical correlation coefficients. As a side remark, we note that the value of (3.3) does not change if one adjusts phase offsets c_μ, with $f_v(t) = |I|^{-1/2}\, e^{i(a_\mu t^2/2+b_\mu t+c_\mu)} 1_I(t)$, so that the phase $\sum_I(\frac{1}{2}a_\mu t^2 + b_\mu t + c_\mu)1_I(t)$ is continuous. The situation for real-valued signals is a little different and is discussed in Appendix A. In this case, imposing continuity implies a substantially greater computational complexity without improving the detection.

A major problem with the approach (3.3) is that the GLRT will naively overfit the data. By choosing paths with shorter chirplets, one can find chirplets with increased correlations (one needs to match data on shorter intervals), and as a result the sum $\sum_{v\in W}|\langle y, f_v\rangle|^2$ will increase. In the limit of tiny chirplets, $|\langle y, f_v\rangle|^2 = \|y_v\|^2$ which gives

$$\max_{W} \sum_{v\in W} |\langle y, f_v\rangle|^2 = \|y\|^2,$$

and one has a perfect fit! There is an analogy with model selection in multiple regression where one improves the fit by increasing the number of predictors in the model. Just as in model selection, one needs to adjust the goodness of fit with the complexity of the fit.

Let W be a fixed path of length $|W|$. Then under the null hypothesis, $\sum_{v\in W}|\langle y, f_v\rangle|^2$ is distributed as a chi-squared random variable with $|W|$ degrees of freedom. Thus for fixed paths, we see that the value of the sum of squares along the path grows linearly with the length of the path. In some sense, the same conclusion applies to the maximal path; i.e., the value of the sum of squares along a path of a fixed size ℓ also grows approximately linearly with ℓ, with a constant of proportionality *greater than 1.* An exact quantitative statement would be rather delicate to obtain, in part because of the inherent complexity of the chirplet graph, but also because it would need to depend on the special chirplet discretization. We refer the reader to [9].

The above analysis suggests taking a test statistic of the form

$$Z^* = \max_W \frac{\sum_{v \in W} |\langle y, f_v \rangle|^2}{|W|}, \tag{3.4}$$

which may be seen as a perhaps unusual trade-off between the goodness of fit and the complexity of the fit. This is of course motivated by our heuristic argument, which suggests that under the null hypothesis, the value of the best path of length ℓ, the quantity

$$T_\ell^* := \max_{|W| \le \ell} \sum_{v \in W} |\langle y, f_v \rangle|^2, \tag{3.5}$$

grows linearly with ℓ, and is well concentrated around its mean by standard large deviation inequalities. In other words, with $Z_\ell^* := T_\ell^*/\ell$, one would expect Z_ℓ^* to be about constant under H_0, at least for ℓ sufficiently large. This would imply that if we ignored paths of small length, one would expect—owing to sharp concentration—$Z^* = \max_\ell Z_\ell^*$ to be about constant under H_0. Therefore, a possible decision rule might be to reject H_0 if Z^* is large.

Numerical simulations confirm that under the null, T_ℓ^* grows linearly with ℓ, but they also show—as expected—deviations for small values of ℓ (see Figure 3.1). For example, with the discretization discussed in Section 2.3, $\mathcal{E} Z_\ell^*$ seems to be decreasing with ℓ. With this discretization, Z^* is also almost all the time attained with paths of length 1 (one single chirplet) so that Z^* is almost always equal to Z_1^*. If we were to set a threshold based on the quantile of the null distribution of Z^* which basically coincides with that of Z_1^*, we would lose some power to detect the alternative. Suppose, indeed, that there is signal. Then the signal may be strong enough so that the observed value of Z_ℓ^* for some ℓ may very well exceed the appropriate quantile of its null distribution, hence providing evidence that there is signal, but too weak for the observed Z^* to exceed the appropriate quantile of its null distribution. Hence, we would have a situation where we could, in principle, detect the signal, but would fail to do so because we are using a low-power test statistic which is not looking in the right place.

A more powerful approach in order to gather evidence against the null consists of looking at the Z_ℓ^*s for many different values of ℓ, and finding one which is unusually large. Because

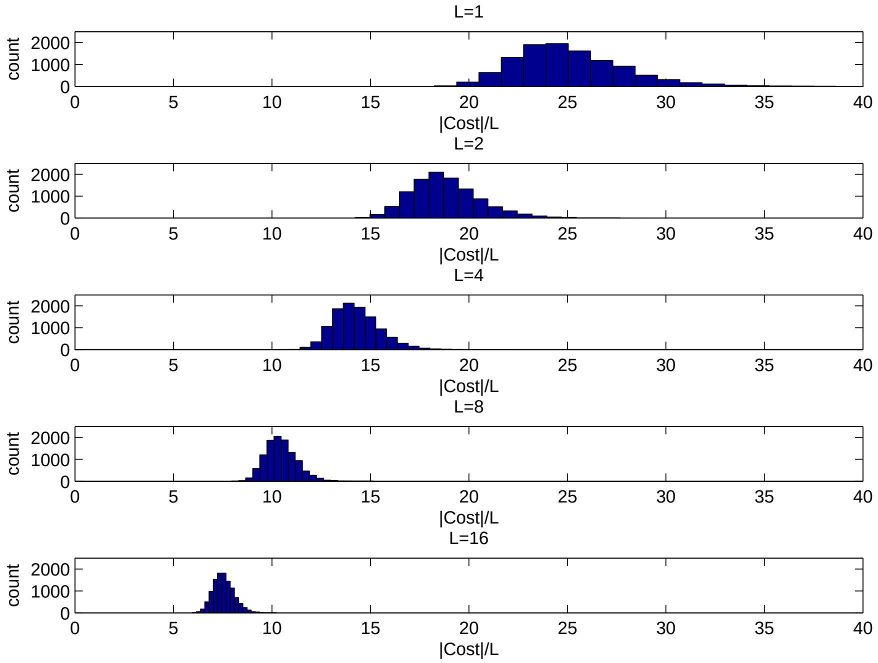


Figure 3.1: Null distribution of Z^*_ℓ for values of ℓ equal to 1, 2, 4, 8, 16. The mean and standard deviation are decreasing with ℓ.

we are now looking at many test statistics simultaneously, we need a multiple comparison procedure which would deal with issues arising in similar situations, e.g., in multiple hypothesis testing [84]. For example, suppose we are looking at k values of ℓ and let $q_\ell(\alpha)$ be the αth quantile of the distribution of Z^*_ℓ. Then to design a test with significance level α, one could use the Bonferroni procedure and reject the null if one of the Z^*_ℓs exceeds $q_\ell(1-\alpha/k)$ (informally, one would test each hypothesis at the α/k level). The Bonferroni method is known to be overly conservative in the sense that it has low power and a better approach is to conduct an α-level test is as follows:

1. Calculate the p-values for each of the Z^*_ℓ and find the minimum p-value.

2. Compare the observed minimum p-value with the distribution of the minimum p-value under the null hypothesis.

In the first step, we are choosing the coordinate of the multivariate test statistic that gives the greatest evidence against the null hypothesis. In the second step, we compare our test statistic with what one would expect under the null. We call this the *Best Path (BP) test/statistic*. In Section 3.4, we will see that this simple way of combining the information in all the coordinates of the multivariate test statistic enjoys remarkable practical performance.

At this point, one might be worried that the computational cost for calculating the Z_ℓ^*s is prohibitive. This is not the case. In fact, besides having sound statistical properties, the BP test is also designed to be rapidly computable. This is the subject of Section 3.2.

3.1.2 Why multiscale chirplets?

If one were to use monoscale chirplets, i.e., a set of chirplets living on time intervals of the form $[k2^{-j}, (k+1)2^{-j})$ for a *fixed* scale 2^{-j}, then all the paths would have the same length (equal to 2^j) and the issue of how to best trade-off between the goodness of fit and the complexity of the fit would, of course, automatically disappear. One could then apply the GLRT (3.3), which is rapidly computable via dynamic programming, as we will see in Section 3.2.1.

Multiscale chirplets, however, provide a much richer structure. Whereas a monoscale approach imposes the use of templates of the same length everywhere, the multiscale approach offers the flexibility to use shorter templates whenever the instantaneous frequency exhibits a complicated structure, and longer templates whenever it exhibits a simpler structure. In other words, the multiscale chirplet graph has the advantage of automatically adapting to the unknown local complexity of the signal we wish to detect. Moreover, with monoscale models, one would need to decide which scale to use and this may be problematic. The best scale for a given signal may not be the best for a slightly different signal, so that the whole business of deciding upon a fixed scale may become rather arbitrary. We are of course not the first to advocate the power of multiscale thinking as most researchers in the field have experienced it (the list of previous “multiscale successes” is very long by now and ever increasing). Here, we simply wish to emphasize that the benefits of going multiscale largely outweigh the cost.

3.2 Best Path algorithms

This section presents an algorithm for computing the Best Path statistic, which requires solving a sequence of optimization problems over all possible paths in the chirplet graph; for each ℓ in a discrete set of lengths, we need to solve problem (3.5). Although the number of paths in the graph is exponential in the sample size N, the BP statistic is designed in such way that it allows the use of network flow algorithms for rapid computation. We will find out that the complexity of the search is of the order of the number of arcs in the chirplet graph times the maximum length of the path we are willing to consider. Later in this section, we will discuss proxies for the Best Path statistic with even more favorable computational complexities.

Before we begin, we assume that all the vertices in the chirplet graph are labeled and observe that the chirplet graph is a directed and acyclic graph, meaning that the vertices on any path in the graph are visited only once (i.e., the graph contains no loops). Suppose that two vertices v and w are connected, then we let $C(v, w)$ be the cost associated with the arc (v, w), which throughout this section is equal to the square of the chirplet coefficient at the node w, $C(v, w) = |\langle y, f_w \rangle|^2$. (We emphasize that nothing in the arguments below depends on this assumption.) To properly define the cost of starting-vertices, we could imagine that there is a *dummy vertex* from which all paths start and which is connected to all the starting-vertices in the chirplet graph. We put $|E|$ and $|V|$ to denote the number of arcs and vertices in the graph under consideration.

3.2.1 Preliminaries

An important notion in graph optimization problems is that of *topological ordering.* A topological ordering of a directed acyclic graph is an ordering of the vertices(v_i), $i = 1, \dots, |V|$, such that for every arc (v_i, v_j) of the graph, we have $i < j$. That is, a topological ordering is a linear ordering of all its vertices such that the graph contains an edge (v_i, v_j), then v_i appears before v_j in the ordering. From now on, we will use the notations i or v to denote vertices and (i, j) or (v, w) to denote edges interchangeably.

Labeling chirplets in the chirplet graph is easy. We move along the time axis from left to right, taking the smallest possible time step (depending on the smallest allowable scale) and label all the chirplets starting from the current position on the time axis; all these chirplets are not connected to each other and, therefore, we may order them freely. Any chirplet starting at a later time will receive a larger topological label and, therefore, the chirplets are arranged in topological order.

Suppose we wish to find the so-called shortest path in the chirplet graph, i.e., solving the optimization problem (3.3). (In the literature on algorithms this is called the shortest path, because by flipping the cost signs and interpreting the costs as distances between nodes this is equivalent to finding the path along which the sum of the distances is minimum.) To find the shortest path, one can use Dijkstra's algorithm, which is known to be a good algorithm [2]. We let $i = 0$ be the source or dummy node and $d(v)$ be the value of the maximum path from the source to node v. Below, the array pred will be a list of the predecessor vertices in the shortest path. That is, if $\text{pred}(j) = i$, then the arc (i, j) is on the shortest path.

Algorithm for shortest path in a chirplet graph:

- Set $d(s) = 0$ and $d(i) = 0$ for $i = 1, \ldots, |V|$.
- Examine the vertices in topological order. For $i = 1, \ldots, |V|$:
  - Let $A(i)$ bet the set of arcs going out from vertex i.
  - Scan all the arcs in $A(i)$. For $(i, j) \in A(i)$, if $d(j) < d(i) + c(i, j)$, set $d(j) = d(i) + c(i, j)$ and $\text{pred}(j) = i$.

Since every arc is visited only once, this shows that the maximum path in the chirplet graph can be found in $O(|E|)$ where we recall that $|E|$ is the number of edges in the graph.

3.2.2 The Best Path algorithm

The idea of solving a Shortest Path problem using updated costs can be used to solve a Lagrangian relaxation of the Constrained Shortest Path problem. This approach is well known in the field of Network Flows. Solving the problem (3.5) for every possible length would give us the points defining the convex hull of the achievable paths, i.e., the convex hull of the points $(|W|, C(W))$, where $C(W)$ is the cost of the path W. A point on the convex hull is a solution to

$$\max_{W} \sum_{v \in W} |\langle y, f_v \rangle|^2 - \lambda \, |W|, \tag{3.6}$$

where λ is some positive number, which can be solved by the Dijkstra's algorithm by setting $\tilde{C}(v, w) = C(v, w) - \lambda$. Then one could try to solve a series of problems of this type for different values of λ to hunt down solutions of the Constrained Shortest Path problem for different values of length. There are many proposed rules in the literature for updating λ but nothing with guaranteed efficiency.

Perhaps surprisingly, although the Constrained Shortest Path problem is in general NP-complete for noninteger times; we can solve it in polynomial time by changing the Shortest

Path algorithm only slightly [54]. We let $i = 0$ be the source node and $d(i, k)$ be the value of the maximum path from the source to node i using exactly k arcs, where k ranges from 0 to $\ell_{\max}$. As before, we denote by $\text{pred}(i, \ell)$ the vertex which precedes vertex i in the tentative best path of length ℓ from the source node to vertex k. The following algorithm solves the Constrained Shortest Path problem in about $O(\ell_{\max} |E|)$, where $\ell_{\max}$ is the maximum number of vertices allowed in the path and $|E|$ is the number of edges in the graph.

Best Path Algorithm:

- Set $d(0, \cdot) = 0$ and $d(i, \cdot) = 0$ for $i = 1, \ldots, |V|$.
- Examine vertices in topological order. For $i = 1, \ldots, |V|$:
  - Let $A(i)$ be the set of arcs going out from vertex i.
  - Scan arcs in $A(i)$. For $(i, j) \in A(i)$, for $k = 1, \ldots, \ell_{max}$, if $d(j, k) < d(i, k-1) + c(i, j)$, set $d(j, k) = d(i, k-1) + c(i, j)$ and $\text{pred}(j, k) = i$.

This algorithm is slightly more expensive than the Shortest Path algorithm since it needs to keep track of more distance labels. The memory storage requirement is of size $O(|V| \times \ell_{\max})$ for storing the distance labels and the predecessor vertices. If we want to include all possible lengths in (3.5) so that $\ell_{\max}$ is of about size N in the chirplet graph, then the memory would scale as $O(N \times M_N)$, where M_N is the number of chirplets.

3.2.3 Variations

There are variations on the BP statistic which have lower computational costs and storage requirements, and this section introduces one of them. Instead of computing (3.5), we could solve the Minimum-Cost-to-Time Ratio problem (MCTTR)

$$\max_{W \in \mathcal{W}_k} \sum_{v \in W} \frac{|\langle y, f_v \rangle|^2}{|W|}, \tag{3.7}$$

where for each k, $\mathcal{W}_k$ is a subset of all paths in the chirplet graph. A possibility is to let $\mathcal{W}_0$ be the set of all paths, $\mathcal{W}_1$ be the set of paths which cannot use chirplets at the coarsest scale, $\mathcal{W}_2$ be the set of paths which cannot use chirplets at the two coarsest scales, and so on. Hence the optimal path solution to (3.7) is forced to traverse at least 2^k nodes. In this way, we get a family of near-optimal paths of various lengths. There is an algorithm which allows computing the MCTTR for a fixed k by solving a sequence of Shortest Path problems (see Section 3.2). This approach has the benefit of requiring less storage, namely, of the order of $O(|V|)$, and for each k, the computational cost of computing the Best Path is typically of size $O(|E|)$.

3.2.4 MCTTR algorithms

In this section, we briefly argue that one can compute the MCTTR introduced in Section 3.2.3 efficiently [2, 29]. Assume that $p^\star$ is the maximum value of $\sum_{i \in W} c(i,j)/|W|$ (with optimal solution $W^\star$) and that we have a lower bound p_0 on $p^\star$ (a trivial lower bound for the chirplet problem is $p_0 = 0$). Suppose that W_0 solves the Shortest Path (SP) problem with modified costs $c_0(i,j) = c(i,j) - p_0$. Then there are three possible cases, and we will rule one out:

1. $\sum_{W_0} c_0(i,j) < 0$. Then $\sum_W c_0(i,j) \le \sum_{W_0} c_0(i,j) < 0$ for all paths W and $\sum_{W^\star} c(i,j)/|W^*| < p_0 \le p^\star$. This is a contradiction and this case never comes up.

2. $\sum_{W_0} c_0(i,j) = 0$. Then $\sum_W c_0(i,j) \le \sum_{W_0} c_0(i,j) = 0$ and, hence, $\sum_W c(i,j)/|W| \le p_0$ for all paths W. We conclude that $p_0 = p^\star$.

3. $\sum_{W_0} c_0(i,j) > 0$. Then $\sum_{W_0} c(i,j)/|W_0| > p_0$ and we have a tighter lower bound on p^*. Take $p_1 = \sum_{W_0} c(i,j)/|W_0|$ and repeat with the new costs $c_1(i,j) = c(i,j) - p_1$.

The MCTTR algorithm solves a sequence of SP problems, and visits a subset of the vertices on the boundary of the convex hull of the points $(|W|, C(W))$ until it finds the optimal trade-off. The number of vertices is, of course, bounded by the maximum possible length $\ell_{\max}$ of the path. In practice, the MCTTR converges after just a few iterations—between 4 and 6 in our simulations.

Note the Shortest Path algorithm we have presented relies heavily on the fact that our graph is acyclic. Were it not, we could not hope to solve the Constrained Shortest Path problem in polynomial time. This is well known to be an NP-hard optimization problem in general (see [2]). We have only described algorithms that are needed for our statistical procedures. Because of the graph structure, a wealth of algorithms from the literature for networks is at our disposal.

3.3 Extensions

Thus far, we have considered the detection problem of chirps with slowly time-varying amplitude in Gaussian white noise; in this section, we discuss how one can extend the methodology to deal with a broader class of problems.

3.3.1 Colored noise

We consider the same detection problem (3.1) as before but we now assume that the noise z is a zero-mean Gaussian process with covariance Σ. Arguing as in Section 3.1, the GLRT for detecting an alternative of the form λf where $\lambda \in \mathbf{R}$ and f belongs to a class of normalized templates is of the form

$$\min_{\lambda \in \mathbf{R},\, f \in \mathcal{F}} \quad e^{-(y-\lambda f)^T \Sigma^{-1} (y - \lambda f)/2},$$

which simplifies to

$$\max_{f \in \mathcal{F}} \quad \frac{|y^T \Sigma^{-1} f|^2}{f^T \Sigma^{-1} f}. \tag{3.8}$$

Note that the null distribution of $|y^T \Sigma^{-1} f|^2 / f^T \Sigma^{-1} f$ follows a chi-square distribution with one degree of freedom.

Our strategy then parallels that used in the white noise model. We define new chirplet costs by

$$C(v) = \frac{|y^T \Sigma^{-1} f_v|^2}{f_v^T \Sigma^{-1} f_v}, \tag{3.9}$$

and compute a sequence of statistics by solving the Constrained Shortest Path problem

$$T_\ell^* := \max_W \sum_{v \in W} C(v), \qquad |W| \le \ell. \tag{3.10}$$

Note that we still allow ourselves to call such statistics T_ℓ^* since they are natural generalizations of those introduced earlier. We then form the family $Z_\ell^* := T_\ell^* / \ell$ and find the Best Path by applying the multiple comparison procedure of Section 3.1.1. In short, everything is identical but for the cost function, which has been adapted to the new covariance structure. In particular, once the new costs are available, the algorithm for finding the best path is the same and, therefore, so is the computational complexity of the search.

3.3.2 Computation of the new chirplet costs

In the applications we are most interested in, the noise process is stationary and we will focus on this case. It is well known that the Discrete Fourier Transform (DFT) diagonalizes the covariance matrix of stationary processes so that

$$\Sigma = F^* D F, \quad D = \mathrm{diag}(\sigma_\omega^2),$$

where F is the N by N DFT matrix, $F_{kt} = \exp(-\imath 2\pi kt/N)/\sqrt{N}$, $0 \le k, t \le N-1$, and $\sigma_1^2, \dots, \sigma_N^2$ are the eigenvalues of Σ.

To compute the chirplet costs, we need to evaluate the coefficients $y^*\Sigma^{-1} f_v$. Observe that

$$y^*\Sigma^{-1} f_v = \tilde{y}^* f_v, \qquad \tilde{y} = \Sigma^{-1} y = F^* D^{-1} F y.$$

In other words, we simply need to compute $\tilde{y}$ and apply the discrete chirplet transform. The cost of computing $\tilde{y}$ is negligible since it only involves two 1D FFT of length N and N multiplications. Hence, calculating all the coefficients $y^*\Sigma^{-1} f_v$ takes about the same number of operations as applying the chirplet transform to an arbitrary vector of length N.

To compute the costs, we also need to evaluate $f_v^*\Sigma^{-1} f_v$, which can of course be done offline. It is interesting to notice that this can also be done rapidly. We explain how in the case where the discretization is that introduced in Section 2.3. First, observe that for any pair of chirplets f_v, f_w which are time-shifted from one another, we have

$$f_v^*\Sigma^{-1} f_v = f_w^*\Sigma^{-1} f_w,$$

since Σ^{-1} is time invariant. Thus we only need to consider chirplets starting at $t = 0$. Second, letting $(\hat{f}[\omega])_{0 \le \omega \le N-1}$ be the DFT of $(f[t])_{0 \le t \le N-1}$

$$\hat{f}[\omega] = \frac{1}{\sqrt{N}} \sum_{t=0}^{N-1} f[t]\, e^{-\imath 2\pi \omega t/N},$$

we have that

$$f^*\Sigma^{-1} f = \sum_{\omega=0}^{N-1} |\hat{f}[\omega]|^2/\sigma_\omega^2.$$

All the chirplets associated with the fixed time interval $[0, 2^{-j})$ are of the form

$$f_{a,b}[t] = |I|^{-1/2}\, e^{\imath 2\pi (bt/N + a(t/N)^2/2)}\, 1_I(t),$$

where $b = 0, 1, \ldots, N-1$, and a is a discrete set of slopes of cardinality about $N/2^j$. Now the modulation property of the DFT gives $\hat{f}_{a,b}[\omega] = \hat{f}_{a,0}[\omega - b]$ and so we only need to compute the DFT of a chirplet with zero frequency offset. This shows that for a fixed slope, we can get all the coefficients corresponding to all offsets by means of the convolution

$$f^*_{a,b}\Sigma^{-1} f_{a,b} = \sum_{\omega=0}^{N-1} |\hat{f}_{a,0}[\omega - b]|^2/\sigma^2_\omega,$$

which can be obtained by means of 2 FFTs of length N. With the assumed discretization, there are about $N/2^j$ slopes at scale 2^{-j} and so computing $\hat{f}_{a,0}[\omega]$ for all slopes has a cost of at most $O(N^2/2^j \cdot \log N)$ flops. Hence the total cost of computing all the coefficients $f^*_v \Sigma^{-1} f_v$ is at most $O(N^2 \log N)$ and is comparable to the cost of the chirplet transform.

3.3.3 Varying amplitudes

We are still interested in detecting signals of the form $S(t) = A(t)\exp(\imath\lambda\varphi(t))$, but $A(t)$ is such that fitting the data with constant amplitude chirplets may not provide local correlations as large as one would wish; one would also need to adjust the amplitude of the chirplet during the interval of operation.

To adapt to this situation, we choose to correlate the data with templates of the form $p(t)\, e^{\imath\varphi_v(t)} 1_I(t)$, where $p(t)$ is a smooth parametric function (e.g., a polynomial of a degree at most 2), and $e^{\imath\varphi_v(t)} 1_I(t)$ is an unnormalized chirplet. The idea is, of course, to look for large correlations with superpositions of the form

$$\sum_{v \in W} p_v(t) \tilde{f}_v(t), \quad \tilde{f}_v(t) = e^{\imath\varphi_v(t)}\, 1_I(t).$$

Fix a path W. In the white noise setup, we we would select the individual amplitudes p_v to minimize

$$\sum_{v \in W} \sum_{t \in I} |y_v(t) - p_v(t)\tilde{f}_v(t)|^2, \tag{3.11}$$

and for each chirplet, p_v would be adjusted to minimize $\sum_{t \in I} |y_v(t) - p_v(t)\tilde{f}_v(t)|^2$. Put

$\tilde{y}_v(t) = y_v(t)\exp(\imath\varphi_v(t))$ and let P denote the projector onto a small dimensional subspace S of smooth functions over the interval I (e.g., the space of polynomials of degree 2); if $b_1(t), \ldots, b_k(t)$ is an orthobasis of S, then P^* is the matrix with the b_is as columns. The minimizer p_v is then given by $P\tilde{y}_v$ and it follows from Pythagoras' identity that $\|\tilde{y}_v - P\tilde{y}_v\|^2 = \|\tilde{y}_v\|^2 - \|P\tilde{y}_v\|^2$. We introduce some matrix notations and let $\Phi_v = \mathrm{diag}(e^{\imath\varphi_v(t)})$ so that $\tilde{y}_v = \Phi_v^* y_v$. Then one can apply the same strategy as before, but with chirplet costs equal to

$$C(v) = \|P\tilde{y}_v\|^2 = \|A_v\, y\|^2, \qquad A_v = P\Phi_v^*. \tag{3.12}$$

It follows from this equation that the complexity of computing these costs is of the same order as that of computing the chirplet transform.

Suppose now that the covariance is arbitrary, then one chooses p_v solution to

$$\min_{p \in S}\ (y - \Phi_v p)^* \Sigma^{-1} (y - \Phi_v p) = y^* y - y^* \Sigma^{-1} A^* (A \Sigma^{-1} A^*)^{-1} A \Sigma^{-1} y,$$

so that the general chirplet cost is of the form

$$C(v) = y^* \Sigma^{-1} A^* (A \Sigma^{-1} A^*)^{-1} A \Sigma^{-1} y, \qquad A_v = P\Phi_v^*. \tag{3.13}$$

3.3.4 Computing the general chirplet costs

We briefly argue that the number of flops needed to compute all the costs (3.13) is of the same order as that needed for the original chirplet transform. Rewrite the cost (3.13) as

$$C(v) = x_v^* B_v^{-1} x_v \qquad x_v = A_v \Sigma^{-1} y, \quad B_v = A_v \Sigma^{-1} A_v^*.$$

Then all the x_vs and all the B_v^{-1}s can be calculated rapidly. Once x_v and B_v are available, computing $x_v B^{-1} x_v$ is simply a matter of calculating $B_v^{-1} x$—either a small matrix multiplication or the solution to a small linear system depending on whether we store B_v or B_v^{-1}—followed by an inner product.

We begin with the x_vs. We have already shown how to apply Σ^{-1} rapidly by means of the FFT (see Section 3.3.2). With $\tilde{y} = \Sigma^{-1} y$, the jth coordinate of x_v is given by

$$\sum_t \tilde{y}(t) b_j(t) \overline{f_v(t)}.$$

We then collect all the x_vs by multiplying the data with the appropriate basis functions and taking a chirplet transform. If we have k such basis functions per interval, the number of flops needed to compute all the x_vs is about k times that of the chirplet transform.

We now study B_v. Note that for each v, B_v is a Hermitian k by k matrix and so that we only need to store $k(k+1)/2$ entries per chirplet; e.g., 3 in the case where $k = 2$, or 6 in the case where $k = 3$. Also in the special case where $k = 1$ (constant amplitude), P is the orthogonal projection onto the constant function equal to one and $B_v = n_I^{-1}(f_v^* \Sigma^{-1} f_v)$, where n_I is the number of discrete points in the interval I. Computing B_v is nearly identical to computing $f_v^* \Sigma^{-1} f_v$, which we already addressed. First, by shift invariance, we only need to consider chirplet indices starting at time $t = 0$. Second, we use the diagonal representation of Σ^{-1} to write the (i, j) entry of B_v as

$$\sum_{\omega=0}^{N-1} \widehat{f_v b_i}[\omega]\, \overline{\widehat{f_v b_j}[\omega]}\, \sigma_\omega^{-2}.$$

Two chirplets f_v and f_w at the same scale and sharing the same chirp rate differ by a frequency shift ω_0 so that $\widehat{f_w b_\ell}[\omega] = \widehat{f_v b_\ell}[k - \omega_0]$. Again, one can use circular convolutions to decrease the number of operations. That is, we really only need to evaluate B_v for chirplets starting at $t = 0$ and with vanishing initial frequency offset. In conclusion, just as in the special case and for the discretization described in Section 2.3, one can compute all the B_vs in order $O(N^2 \log N)$ flops. To be more precise, the cost is here about $k(k+1)/2$ that of computing $f_v^* \Sigma^{-1} f_v$ for all chirplets.

3.3.5 Real-valued signals

In our simulations, we considered the detection of complex-valued chirps and we now rapidly discuss ways to extend the methodology to real-valued data where the signal is of the form $S(t) = A(t) \cos(\lambda \varphi(t))$ with unknown phase and amplitude. Again, the idea is to build a family of real-valued chirplets which exhibit good local correlations with the signal. To do this, we could consider chirplets with quadratic phase $a_\mu t^2/2 + b_\mu t + c_\mu$ and build a graph in which connectivities impose regularity assumptions on the phase function. The downside with this approach is that for each chirplet, one would need to introduce the extra phase-shift parameter c_μ, which would increase the size of the dictionary and of the graph. This is not desirable.

A much better strategy is as follows: we parameterize chirplets in the same way with $v = (I, \mu)$ where I is the time support of a chirplet and $a_\mu t + b_\mu$ is the instantaneous frequency, and define the chirplet cost by

$$C(v) := \max_c \frac{\left|\sum_{t \in I} y_t \cos(a_\mu t^2/2 + b_\mu t + c)\right|^2}{\sum_{t \in I} \cos^2(a_\mu t^2/2 + b_\mu t + c)}. \tag{3.14}$$

That is, we simply select the phase shift which maximizes the correlation (note that with complex data, the corresponding ratio $|\sum y_t \exp(\imath(a_\mu t^2/2 + b_\mu t + c))|^2 / \sum |\exp(\imath(a_\mu t^2/2 + b_\mu t + c))|^2$ is, of course, independent of c). One can use simple trigonometric identities and write the numerator and denominator in (3.14) as

$$A^2 \cos^2 c - 2AB \sin c \cos c + B^2 \sin^2 c, \quad C^2 \cos^2 c - 2D \sin c \cos c + E^2 \sin^2 c,$$

where with $\varphi_\mu(t) = a_\mu t^2/2 + b_\mu t$,

$$A + \imath B = \sum y_t \, e^{\imath \varphi_\mu(t)},$$

and

$$C^2 = \sum \cos^2 \varphi_\mu(t), \quad D = \sum \cos \varphi_\mu(t) \sin \varphi_\mu(t), \quad E^2 = \sum \sin^2 \varphi_\mu(t).$$

Note that $A + \imath B$ is nothing else than the chirplet coefficient of the data and C, D, and E can be computed off-line. There is an analytic formula for finding the value of $\cos c$ (or $\sin c$) that maximizes the ratio as a function of A, B, C, D, and E; see Appendix A for details. This extends to the more sophisticated setups discussed in Section 3.3.

Finally, there are further approximations which one could use as well. Observe the expansion of the denominator in (3.14)

$$\sum_{t \in I} \cos^2(\varphi_\mu(t) + c) = |I|/2 + 1/2 \sum_{t \in I} \cos(2\varphi_\mu(t) + 2c),$$

where $|I|$ is here the number of time samples in I. Then for most chirplets (when the support contains a large number of oscillations), the second term in the right-hand side is negligible compared to the first. Assuming that the denominator is about equal to $|I|/2$ for all phase shifts c, we would then simply maximize the numerator in (3.14). A simple calculation shows that $\exp(\imath c) = (A + \imath B)/\sqrt{A^2 + B^2}$ and

$$C(v) \approx 2 \left| \sum_{t \in I} y_t e^{-\imath(a_\mu t^2 + b_\mu t)} / \sqrt{|I|} \right|^2$$

(the "≈" symbol indicates the approximation in the denominator). Hence, the real-valued cost is just about twice the usual complex-valued cost.

3.3.6 A subtlety for rapidly changing noise spectra and real-valued data

In the case of complex-valued data and white noise the chirplet costs all have the same distribution under H_0, independent of location, scale, slope, and frequency offset. This helps to make the chirplet graph "homogeneous" as far as the costs are concerned, although, of course, the graph topology itself could be a source of some inhomogeneity and cause the costs to be correlated. In cases where the chirplet costs do not all have the same distribution there could be chance that the best path through the graph would on average prefer to go through a certain set of vertices where extreme values would be more probable under H_0.

In the case of real-valued data and colored noise, the chirplet costs do not, in general, have exactly the same distribution under H_0. Therefore it is important to understand in which circumstances the distribution of the chirplet costs can vary substantially and cause the chirplet graph to have "preferred" set of vertices for the best paths to go through under H_0.

Consider a data sequence (y_t), or in vector notation $y = [y_0, \dots, y_{N-1}]^T$, where y is a real-valued multivariate Gaussian random vector such that $y \sim N(0, \Sigma)$. Let x be a complex-valued $N \times 1$ vector, which we will take as a discrete chirplet sampled from a function $f(t) = 1_I(t) \exp(i\varphi(t))$, where the interval I is the support of the chirplet and $\varphi(t) = a/2t^2 + bt$, for some fixed real scalars a and b. We want to investigate the effect of the covariance matrix Σ on the expected value and variance of the chirplet cost:

$$C_x(y) = \frac{|y^* \Sigma^{-1} x|^2}{x^* \Sigma^{-1} x}.$$

Since

$$E|y^* \Sigma^{-1} x|^2 = x^* \Sigma^{-1} \left(E y y^* \right) \Sigma^{-1} x = x^* \Sigma^{-1} x,$$

the expected value of the chirplet cost $C_x(y)$ under H_0 is

$$E[C_x(y)] = 1,$$

independent of the covariance Σ or the chirplet x. To calculate the variance of the chirplet cost we introduce the variables

$$g_r := \Sigma^{-1/2}x_r \quad \text{and} \quad g_r := \Sigma^{-1/2}x_i,$$

where $x_r := Re(x)$ and $x_i := Im(x)$ are the real and imaginary parts of the chirplet x. The multivariate Gaussian vector y can be written in the form $y = \Sigma^{1/2}z$, where $z \sim N(0, I_{N\times N})$. This allows us to write

$$y^*\Sigma^{-1}x = z^*\Sigma^{-1/2}x = z^*g_r + iz^*g_i.$$

Then,

$$|z^*\Sigma^{-1/2}x|^2 = (z^*g_r)^2 + (z^*g_i)^2,$$

and

$$E|z^*\Sigma^{-1/2}x|^4 = E(z^*g_r)^4 + E(z^*g_i)^4 + 2E\left((z^*g_r)^2(z^*g_i)^2\right).$$

Note that

$$z^*g_r \sim N(0, \|g_r\|^2), \quad \text{and} \quad z^*g_i \sim N(0, \|g_i\|^2),$$

where $\|g_r\|^2 = x_r^*\Sigma^{-1}x_r$ and $\|g_i\|^2 = x_i^*\Sigma^{-1}x_i$. To carry the calculations further we will use the following well-known facts for Gaussian random variables:

$$E[Z^4] = 3, \quad Z \sim N(0,1),$$

and

$$E[U^2V^2] = Var(U)Var(V) + 2\left(Cov(U,V)\right)^2,$$

for two jointly Gaussian random variables U and V. [1] Denote the correlation between z^*g_r

[1] To prove the latter equality we let $U \sim N(0,\sigma_1^2)$ and $V \sim N(0,\sigma_2^2)$ with $\rho = Cov(U,V)/(\sigma_1\sigma_2)$. Without loss of generality we can assume $\rho \geq 0$ so we can write $U = \sqrt{1-\rho}\cdot\sigma_1 Z_1 + \sqrt{\rho}\cdot\sigma_1 Z_3$ and $V = \sqrt{1-\rho}\cdot\sigma_2 Z_2 + \sqrt{\rho}\cdot\sigma_2 Z_3$ where Z_1, Z_2, and Z_3 are i.i.d. standard normal random variables. The equality follows immediately.

and z^*g_i by

$$\rho := \frac{Cov(z^*g_r, z^*g_i)}{\|g_r\|\|g_i\|} = \frac{g_r^*g_i}{\|g_r\|\|g_i\|}.$$

Then,

$$E|z^*\Sigma^{-1/2}x|^4 = 3\|g_r\|^4 + 3\|g_i\|^4 + 2\|g_r\|^2\|g_i\|^2(1+2\rho^2).$$

and we have established:

Lemma 1. *Let $y \sim N(0,\Sigma)$ be a random vector in $\mathbf{R}^N$ and x be a fixed vector in $\mathbf{C}^N$. Then the random variable $C_x(y) = \frac{|y^*\Sigma^{-1}x|^2}{x^*\Sigma^{-1}x}$ satisfies*

$$E[C_x(y)] = 1,$$

and

$$Var[C_x(y)] = \frac{3\|g_r\|^4 + 3\|g_i\|^4 + 2\|g_r\|^2\|g_i\|^2(1+2\rho^2)}{(\|g_r\|^2 + \|g_i\|^2)^2} - 1.$$

If we can do the approximation $\|g_r\| \approx \|g_i\|$ we would have

$$Var[C_x(y)] \approx \frac{6\|g_r\|^4 + 2\|g_r\|^4(1+2\rho^2)}{4\|g_r\|^4} - 1 = 1 + \rho^2.$$

We can expect such an approximation to hold at least for chirplets at high frequency and with large time support (due to the Riemann-Lebesgue lemma). This suggests that the correlation between z^*g_r and z^*g_i could be a source of difference in the distribution of chirplet costs. That is,

$$\rho^2 = \left(\frac{g_r^*g_i}{\|g_r\|\|g_i\|}\right)^2 = \frac{\left(x_r^*\Sigma^{-1}x_i\right)^2}{(x_r^*\Sigma^{-1}x_r)(x_i^*\Sigma^{-1}x_i)}.$$

If we assume the noise to be stationary, the covariance matrix Σ can be approximately diagonalized with the Fourier matrix F so $\Sigma \approx FDF^*$, where D is a diagonal matrix with the eigenvalues, or the spectrum, of Σ on the diagonal. In the case when the noise is circular stationary an exact equality holds, or $\Sigma = FDF^*$. This shows how the variance of the chirplet cost depends, approximately, on the smoothness of the noise spectrum and the frequency support of the chirplet. If the noise spectrum is roughly equal to a constant D_θ

over the frequency support of the chirplet x, then

$$x_r^* \Sigma^{-1} x_i \approx (F x_r)^* D^{-1} (F x_i) \approx D_\theta^{-1} x_r^* x_i.$$

Similarly, $x_r^* \Sigma^{-1} x_r \approx D_\theta^{-1} x_r^* x_r$ and $x_i^* \Sigma^{-1} x_i \approx D_\theta^{-1} x_i^* x_i$. That would give

$$\rho^2 \approx \left(\frac{|x_r^* x_i|}{\|x_r\| \cdot \|x_i\|} \right)^2 .$$

Since $|x_r^* x_i|$ would usually be small compared to $\|x_r\| \cdot \|x_i\|$, this would cause ρ^2 to be small and the variance of the chirplet cost to be close to 1. However, if the noise spectrum varies greatly over the frequency support of the chirplet x, these heuristics do not hold and the variance of the chirplet cost could be greater than 1. We have verified this numerically using the formula of the variance from Lemma 1. This difference in variance potentially degrades the performance of the method, since under H_0 the chirplet costs might tend to be bigger around frequencies where the noise spectrum changes rapidly, resulting in a bigger value of the BP statistic. This is behavior has been observed in our numerical experiments involving rapidly changing noise spectra.

In practice, if possible, one might want stay away from frequency bands where the noise spectrum changes dramatically and restrict the chirplet graph to regions where it is slowly changing. This has proven to work well in our numerical experiments. Another unexplored possibility would be to try correcting for the difference in variance by normalizing the chirplets, by using the formula from Lemma 1, such that all the chirplet costs under H_0 have unit-variance.

3.4 Numerical Simulations

We now explore the empirical performance of the detection methods proposed in this paper. To this end, we have developed *ChirpLab*, a collection of Matlab routines that we have made publicly available (see Section 10.1). For simplicity, we use a chirplet dictionary with the discretization discussed in Section 2.3. We also consider a slightly different chirplet graph which assumes less regularity about the instantaneous frequency of the unknown chirp; namely, two chirplets are connected if and only if they live on adjacent time intervals and if the instantaneous frequencies at their juncture coincide. In practical situations such as gravitational wave detection, the user would be typically given prior information about the signal she wishes to detect which would allow her to fine-tune both the discretization and the connectivities for enhanced sensitivity. We will discuss these important details in a separate publication. Our goal here is merely to demonstrate that the methodology is surprisingly effective for detecting a few unknown test signals.

3.4.1 The basic setup

We generated data of the form

$$y_i = \alpha S_i + z_i, \qquad i = 0, 1, \ldots, N-1,$$

where (S_i) is a vector of equispaced time samples of a complex-valued chirp, and where (z_i) is a complex-valued white noise sequence: $z = z^0 + \imath\, z^1$ where z^0 and z^1 are two independent vectors of i.i.d. $N(0, 1/2)$ variables. Note that $E|z_i|^2 = 1$ and $E\|z\|^2 = N$. In this setup, we define the SNR as the ratio

$$\mathrm{SNR} = \frac{\|\alpha S\|}{\sqrt{N}}. \tag{3.15}$$

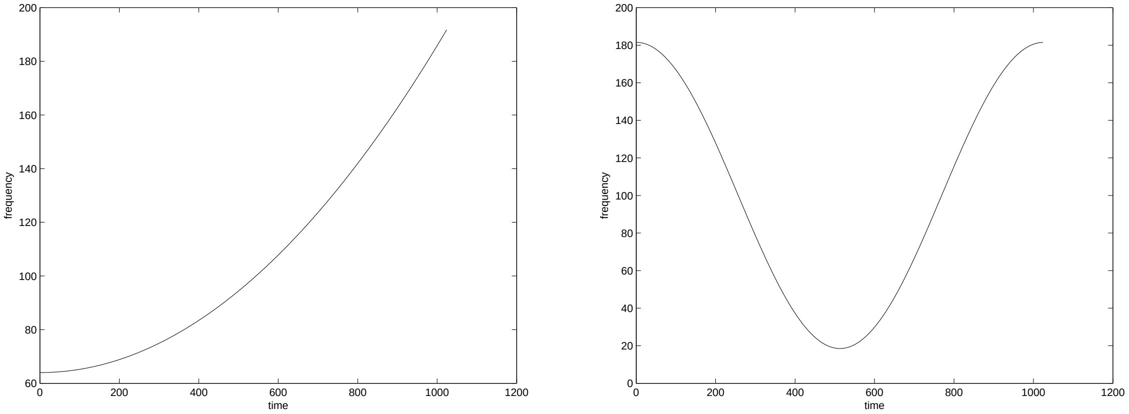
We have chosen to work with complex-valued data and want to emphasize that we could just as well perform simulations on real-valued data and detect real-valued signals (see Appendix A for details). In all our experiments, the signal S obeys the normalization $\|S\| = \sqrt{N}$ so that the parameter α actually measures the SNR. We considered signals of size $N = 512, 1024, 2048, 4096$. The chirps are of the form

$$S(t) = A(t) e^{\imath N \varphi(t)}, \tag{3.16}$$

and sampled at the equispaced points $t_i = i/N$, $i = 0, 1, \ldots, N-1$. We considered two test signals.

1. A *cubic phase chirp* with constant amplitude:

$$A(t) = 1, \qquad \varphi(t) = t^3/24 + t/16.$$



(a)

(b)

Figure 3.2: Instantaneous frequency $\varphi'(t)$ of the chirps under study: (a) Cubic phase chirp, (b) Cosine phase chirp

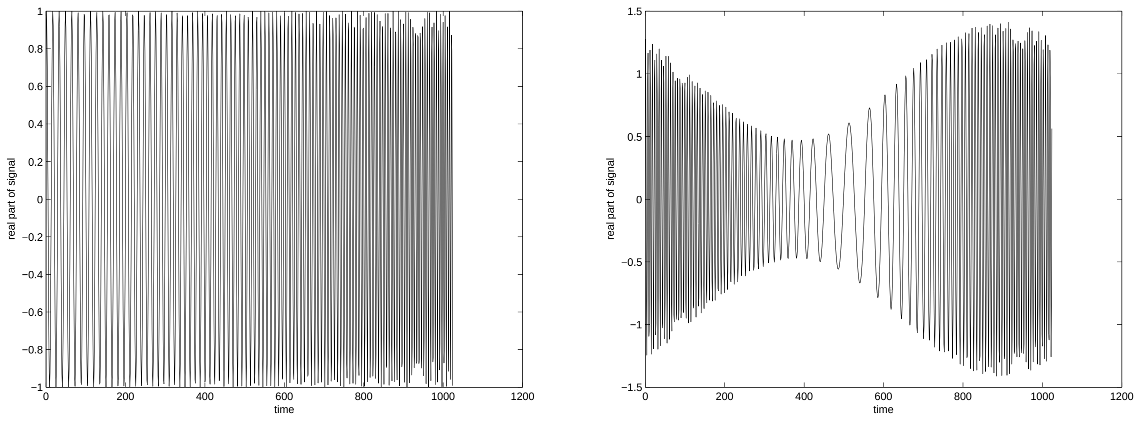
2. A *cosine phase chirp* with slowly varying amplitude:

$$A(t) = 2 + \cos(2\pi t + \pi/4), \qquad \varphi(t) = 2\pi\left(\sin(2\pi t)/4\pi + 200\pi t/1024\right).$$

Note that because of the factor N in the exponential (3.16), we are not sampling the same signal at increasingly fine rates. Instead, the instantaneous frequency of S is actually changing with N and is equal to $N\varphi'(t)$ so that the signal may oscillate at nearly the sampling rate, no matter what N is. Figures (3.2) and (3.3) show the rescaled instantaneous frequency, $\varphi'(t)$ and the real part of the signals under study for $N = 1024$.

For detection, we use the BP test statistic introduced in Section 3.1.1 with $\{1, 2, 4, 8, 16\}$ as our discrete set of path lengths. We estimated the distribution of the minimum P-value under the null hypothesis via Monte Carlo simulations. For the most part of the performance analysis, we selected a detection threshold giving a probability of false detection (Type I error) equal to 5% (.05 significance level). In the literature of gravitational wave detection, one typically considers much lower probabilities of false alarm and this is the reason why we also report on experiments with a probability of Type I error set at .05%, i.e., an average of only 5 false alarms in 10,000.

- For signal lengths $N = 512, 1024, 2048$, we randomly sampled about 100,000 realizations of white noise to compute the detection threshold (the quantile of the minimum P-value distribution). For $N = 4096$, we used 250,000 realizations of white noise.



(a) (b)

Figure 3.3: Real part $A(i/N)\cos(N\varphi(i/N))$, $i = 0, \ldots, N-1$ of the of the chirps under study: (a) Cubic phase chirp, (b) Cosine phase chirp. The cosine phase chirp has a slowly varying amplitude. Note that the instantaneous frequency depends on the sample size N.

- For each signal length, each signal, and each SNR, we sampled the data model about 1,000 times in order to compute detection rates, or (equivalently) the so-called power curves.

In these simulations, we only considered chirplets with positive frequencies and for the larger signal sizes, $N = 2048, 4096$, we restricted ourselves to discrete frequencies on the interval $\{0, \ldots, N/4 - 1\}$ to save computational time. In all cases the slope parameters a_μ of the chirplets (see equation (2.2)) ranged from $-\pi N$ to πN, with a discretization at scale 2^{-j} of the form $a_\mu = 2\pi N(-1/2 + k \cdot m 2^{j-J})$ where $J = \log_2 N$; $m = 1$, $k \in \{0, \ldots, 2^{J-j}\}$ for signal lengths $N = 512, 1024, 2048$ and $m = 4$, $k \in \{0, \ldots, 2^{J-j-2}\}$ for $N = 4096$. This ensures that any endpoint of a dyadic interval is an integer multiple of 2π.

The scales considered ranged from the coarsest 2^0 to 2^{-s} with $s = 6$ for $N = 512, 1024$, $s = 5$ for $N = 2048$ and $s = 4$ for $N = 4096$ (the motivation again is speed). In practice, these parameters would depend upon the application and would need to be selected with care. Tables 3.1 and 3.2 show the correlation between the waveforms and the best chirplet path with a fixed length. Although we use a coarser discretization and fewer scales when $N = 4096$, the correlation is still very high, at least for path lengths 8 and 16. Table 3.3 shows the correlations between the cosine phase chirp and chirplets with adapted amplitudes. As expected, the correlation increases.

signal length N	$\ell = 1$	$\ell = 2$	$\ell = 4$	$\ell = 8$	$\ell = 16$
512	0.0718	0.4318	0.7126	0.9905	0.9982
1024	0.0453	0.2408	0.5784	0.9814	0.9981
2048	0.0306	0.1643	0.5107	0.9469	0.9976
4096	0.0229	0.0953	0.4265	0.8158	0.9917

Table 3.1: Correlations between the cosine phase signal and the best chirplet path with fixed lengths $\ell \in \{1, 2, 4, 8, 16\}$

signal length N	$\ell = 1$	$\ell = 2$	$\ell = 4$	$\ell = 8$	$\ell = 16$
512	0.2382	0.8733	0.9903	0.9979	0.9999
1024	0.1498	0.6575	0.9883	0.9985	0.9997
2048	0.0932	0.3836	0.9671	0.9976	0.9995
4096	0.0590	0.2373	0.8734	0.9903	0.9971

Table 3.2: Correlations between the cubic phase signal and the best chirplet path with fixed lengths $\ell \in \{1, 2, 4, 8, 16\}$

$d :$ degree of polynomials	$\ell = 1$	$\ell = 2$	$\ell = 4$	$\ell = 8$	$\ell = 16$
$[2, 1, 1, 1, 1, 1]$	0.1481	0.2852	0.5699	0.9612	0.9995
$[2, 2, 2, 1, 1, 1]$	0.1481	0.3337	0.5999	0.9612	0.9995
$[2, 2, 2, 2, 2, 2]$	0.1481	0.3337	0.6122	0.9823	0.9999

Table 3.3: Correlations between the cosine phase signal and the best chirplet path with fixed lengths $\ell \in \{1, 2, 4, 8, 16\}$ (chirplets with varying amplitude). $N = 2048$. The first column indicates the degree of the polynomial used to fit the amplitude. The entry d_j in $d = [d_0, d_1, \ldots, d_5]$ is the degree of the polynomial at scale 2^{-j}.

3.4.2 Results from simulations

To measure the performance of the BP statistic, we first fix the probability of Type I error at 5% and estimate the detection rate, the probability of detecting a signal when there is signal. We compute such detection curves for various SNRs (3.15). To limit the number of computations we focus on a small set of signal levels around the transition between a poor and a nearly perfect detection.

Figures 3.4 and 3.5 present results of a simulation study and display the power curves for both chirps and for various sample sizes. Of course, as the sample size increases, so does the sensitivity of the detector (even though the signal is changing with the sample size). We also note that the detection of the cubic phase chirp is slightly better than that of the cosine phase chirp, which was to be expected since the cubic phase chirp is slightly less complex. (Simulations where one also adapts the amplitude give similar results.)

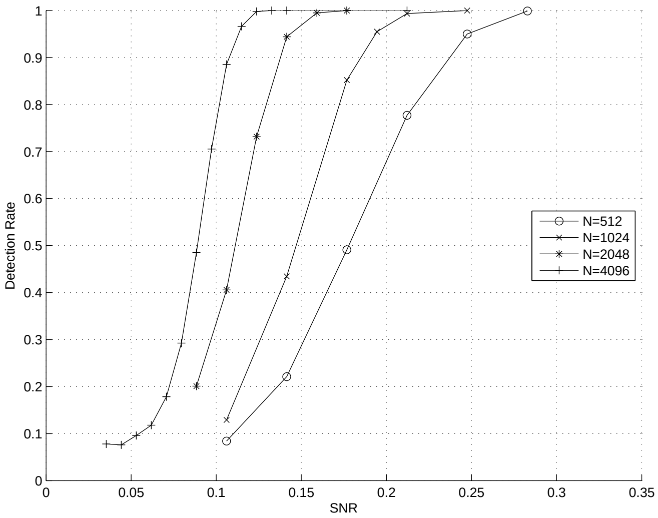


Figure 3.4: Detection rates of the cubic phase chirp with the BP method. The probability of Type I error is fixed at 5%.

Consider the cosine phase chirp with time-varying amplitude and a sample size N equal to $4,096$. Then the SNR for a detection level in the 95% range is about .12. This means that one can reliably detect an unknown chirp of about this complexity when the amplitude of the noise is about 8 times that of the unknown signal. When the probability of Type I error is orders of magnitude smaller, we expect the detection curves on Figures 3.4 and 3.5 to translate to the right since the H_1-acceptance region shrinks a little. Figure 3.6 plots the detection rate for a probability of Type I error fixed at 0.05%. The level of detectability

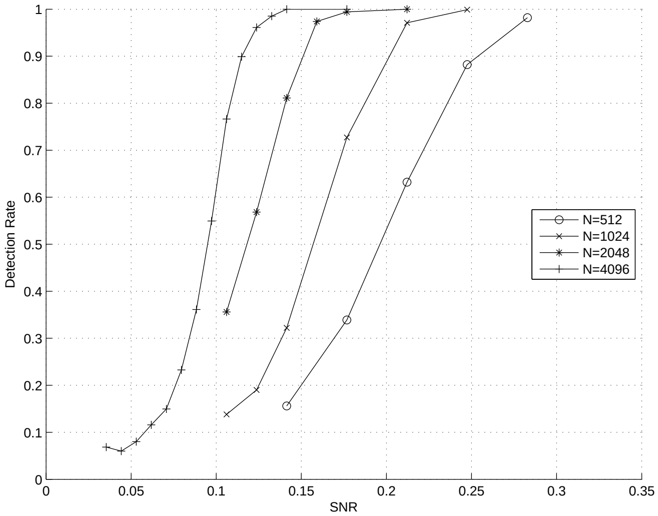


Figure 3.5: Detection rates of the cosine phase chirp with the BP method. The probability of Type I error is fixed at 5%.

does not change much.

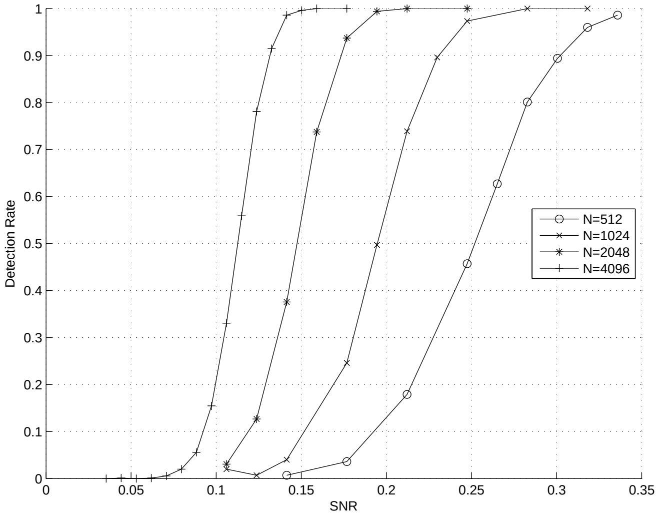


Figure 3.6: Detection rates of the cosine phase chirp with the BP method. The probability of Type I error is fixed at 0.05%.

It is interesting to study the performance gain when we increase the signal length. Fix a detection rate at 95% at the 5% significance level, and plot the SNR that achieves this rate against the sample size N. Figure 3.7 shows the base-2 logarithm of the estimated SNR (using a simple linear interpolation of the power curves) versus the logarithm of the sample size. The points roughly lie on a line with slope -0.4 (fixing a probability of Type I error at 0.05% also gives a line with slope about -0.4 and we omit the plot); as we double the signal length from N to $2N$, the SNR required to achieve a 95% detection rate is about $2^{-0.4} \approx 0.76$ times that required to achieve the same detection rate for the signal length N. In a parametric setting, we would asymptotically expect a slope of -0.5. The fact that the slope is slightly higher than this is typical of nonparametric detection problems which deal with far richer classes of unknown signals [50].

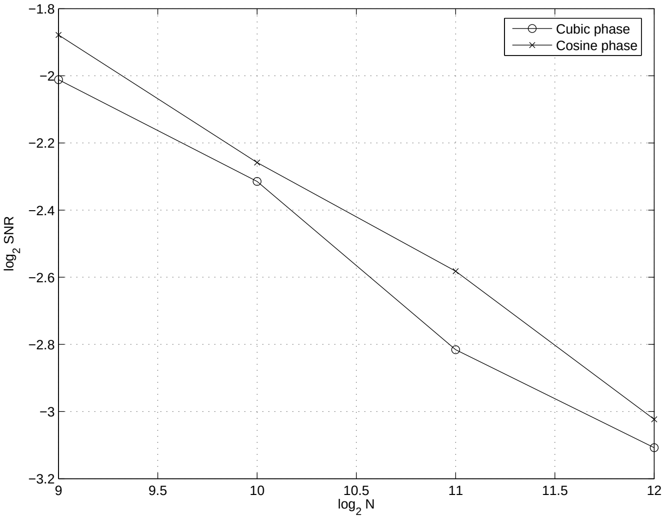


Figure 3.7: Log-log (base-2) plot of the estimated SNR (for both chirps) at the 95% detection rate versus signal length N. Again the probability of Type I error is fixed at 5%. In both cases, the slope is approximately equal to -0.4.

3.4.2.1 Comparison with the detection of a known signal

In order to see how sensitive our test statistic really is, it might be instructive to compare the detection rates with those one would achieve if one had full knowledge about the unknown signal. We then consider a simple alternative

$$H_1 : y = \alpha S_0 + z,$$

where the signal is known. That is, if there is signal, we know *exactly* what it looks like. The standard likelihood ratio test (LRT) gives the optimal test in terms of maximizing the power of detection at a given confidence level. A simple calculation shows that at the 5% level, the power function of the LRT is equal to $\Phi(1.65 - \text{SNR}\sqrt{2N})$, where Φ is the cumulative distribution of a standard normal. Figure 3.10 shows this power curve together with those obtained via the BP test for a sample size $N = 4096$. The horizontal gap between curves indicates the ratio between SNRs to achieve the same detection rate. Consider a detection level equal to about 95%. Our plot shows that one can detect a completely unknown signal via the BP statistic with the same power that one would get by knowing the signal *beforehand*, provided that the amplitude is about 3 times as large. Figure 3.8 shows a comparison of the receiver operating characteristic curves (ROC) for the BP test and the cosine phase chirp at SNR = 0.124, and the LRT at SNR = 0.042. The figure shows that the 3-to-1 ratio holds over a very wide range of significance levels. Note that this ratio is small and may be thought of as the price one has to pay for not knowing in advance what it is.

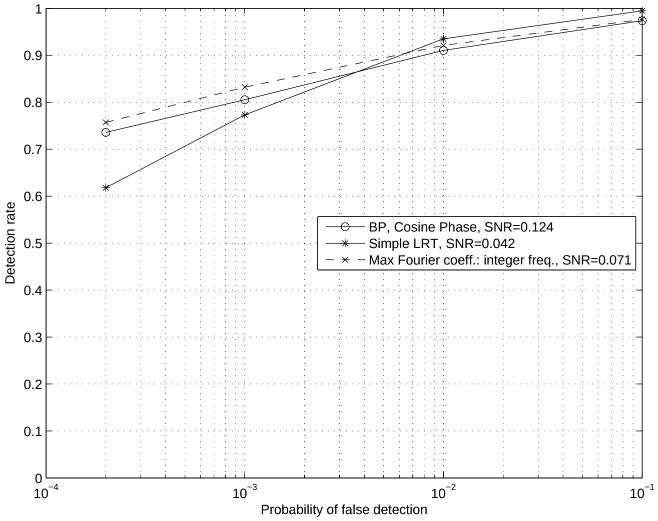


Figure 3.8: Comparison of ROCs for the LRT, GLRT (based on the maximum modulus of the Fourier coefficients) for monochromatic sinusoids when the unknown sinusoid has integer frequency, and the BP test when the unknown signal is the cosine phase chirp. The signal length is $N = 4096$. The x-axis is plotted on a log (base-10) scale.

3.4.2.2 Detection of a monochromatic sinusoid

To appreciate the performance of the BP statistic, it might be a good idea to study a more subtle problem. Suppose that the unknown signal is a monofrequency sinusoid. If there is signal, we know it is of the form $S(t) = \exp(\imath\omega t + \phi)$, where the frequency ω and the phase shifts are unknown. Consider the simpler case where for a discrete signal of length N, $\omega = 2\pi k/N$ with $k \in \{0, \dots, N-1\}$ is one of the N Nyquist frequencies. Letting y^0 and

y^1 be the real and imaginary parts of the data y, the GLRT would maximize

$$\sum_{0 \le t \le N-1} y_t^0 \cos(2\pi kt/N + \phi) + y_t^1 \sin(2\pi kt/N + \phi),$$

over $k = 0, 1, \ldots, N-1$ and $\phi \in [0, 2\pi]$. One can take the maximum over ϕ and check that the GLRT is equivalent to maximizing

$$\left| \sum_{0 \le t \le N-1} y_t e^{-\imath 2\pi kt/N} \right|$$

over k. Thus, the GLRT has a simple structure. It simply computes the DFT of the data, and compares the maximal entry of the response with a threshold. (Note the resemblance of this problem to the famous problem of testing whether the mean of a Gaussian vector is zero vs. an alternative which says that one of its component is nonzero.)

We could also make the problem a tiny bit harder by selecting the frequency arbitrarily, i.e., not necessarily a multiple of 2π but anything in the range $[0, 2\pi N]$. In this case, the method described above would be a little less efficient, since the energy of the signal would not be concentrated in a single frequency mode but spill over into neighboring frequencies. The GLRT would ask to correlate the data with the larger collection of monofrequency signals, which in practice we could approximately achieve by oversampling the DFT (e.g., we could select a finer frequency discretization so that the correlation between the unknown monochromatic signal we wish to detect and the closest test signal exceeds a fixed tolerance, e.g. .90 or .99).

We compare the detection rate curve for detecting (i) a monochromatic sinusoid with integer frequency and (ii) a monochromatic sinusoid with arbitrary frequency using the maximum absolute DFT coefficient on one hand, and the BP test on the other hand. The signals in (i) and (ii) are equispaced samples from $S_1(t) = \exp\left(\imath 2\pi \frac{N}{8} t\right)$ and $S_2(t) = \exp\left(\imath 2\pi (\frac{N}{8} + \frac{1}{2}) t\right)$. The signal length N is equal to 4096. Figure 3.9 displays the detection rates. Consider the 95% detection rate. Then for (i) the SNR for the BP test is about 20% higher than that for the GLRT. In (ii) the SNR is only 8% higher. Also, at this detection level, the ratio between the SNRs for the cosine phase chirp and the monofrequency is about 1.75. Figure 3.8 reveals that this ratio holds over a wide range of significance levels. These results show that "the price we pay" for being adaptive and having the ability to detect a rich class of chirping signals is low.

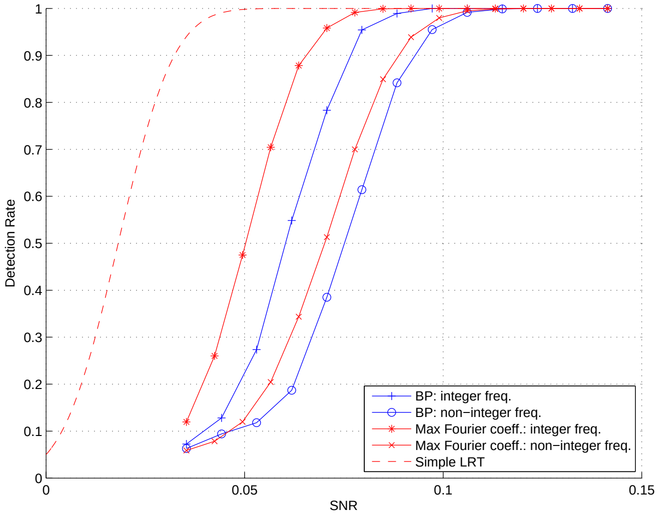


Figure 3.9: Comparison of the BP and GLRT (based on the maximum modulus of the Fourier coefficients) for monochromatic sinusoids. The probability of Type I error is set at 5%

3.4.2.3 Detection of a linear chirp

To study "the price of adaptivity", we also consider the problem of detecting linear chirps. Suppose that the unknown signal consists of sampled values of a linear chirp of the form $S(t) = \exp(\imath 2\pi N \varphi(t))$, where $\varphi(t) = at^2/2 + bt + c$. Here, $N = 4096$ and the coefficients a, b, c are adjusted so that the unknown linear chirp is a complex multiple of a chirplet at the coarsest scale (the GLRT is then the BP test restricted to paths of length 1). In the simulations, we selected a chirp with $a = 1/8$, $b = 1/16$, and $c = 0$ so that the instantaneous frequency $N\varphi'(t)$ increased linearly from 256 to 768. Figure 3.10 displays the detection rates for the GLRT and the BP test with $\{1, 2, 4, 8, 16\}$ as path lengths. The detection rates for the BP test and the GLRT are almost the same; the ratio between the SNRs required to achieve a detection rate of about 95% is about 1.05. This shows the good adaptivity properties of the BP test. For information, the plot also shows that one can detect a completely unknown signal via the BP statistic with the same power that one would get for detecting *a linear chirp* via the GLRT, provided that the amplitude is about 1.5 times as large.

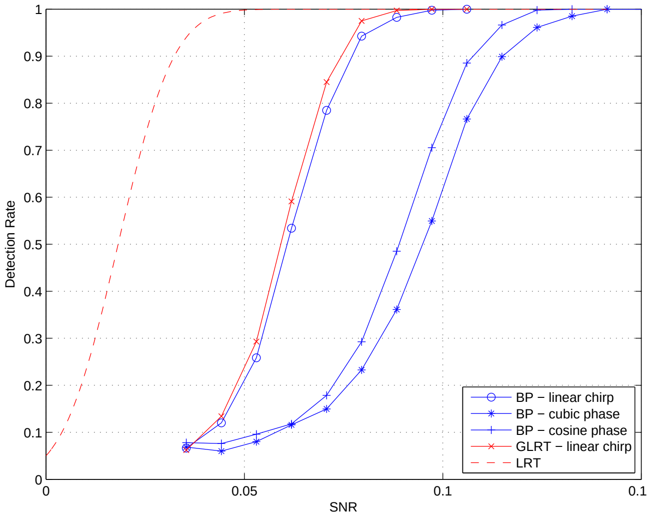


Figure 3.10: Comparison of the BP and GLRT detection rates over a set of linear chirps. The probability of Type I error is set at 5%. Detection rates are plotted along with the detection rates for the cubic and cosine phase chirps.

3.4.3 Empirical adaptivity on a simulated gravitational wave

Earlier, we argued that the GLRT or the method of matched filters would need to generate exponentially many waveforms to provide good correlations with the unknown signal of interest. The idea underlying the chirplet graph is that one can get very good correlations by considering a reasonably sized dictionary and considering correlations with templates along a path in the graph. Figure 3.11 shows the real part of a "mock" gravitational waveform whose instantaneous frequency and amplitude increase roughly as a power law. The waveform is $S(t) = A(t)e^{\imath\varphi(t)}$ where the phase is

$$\varphi(t) = a_0(t_c - t)^{5/8} + a_1(t_c - t)^{3/8} + a_2(t_c - t)^{1/4} + a_3(t_c - t)^{1/8},$$

and the amplitude is given by $A(t) = [\varphi'(t)]^{2/3}$ (see Figure 3.11). The coefficients $a_0, \ldots, a_3$ were chosen from the post-Newtonian approximation for a binary inspiral as described in [4, 5]. The coefficient t_c is the time of coalescence. The masses of the two bodies were both chosen to be equal to 14 solar masses and the sampling rate was 2048 Hz. We studied the last 1024 samples of the waveform.

As seen in Figure 3.12, the correlation with the noiseless waveform is equal to .95 with just 4 chirplets (with linear time-varying amplitudes) and .99 with just 5 chirplets. So we would not gain much (if anything at all) by computing inner products with exponentially

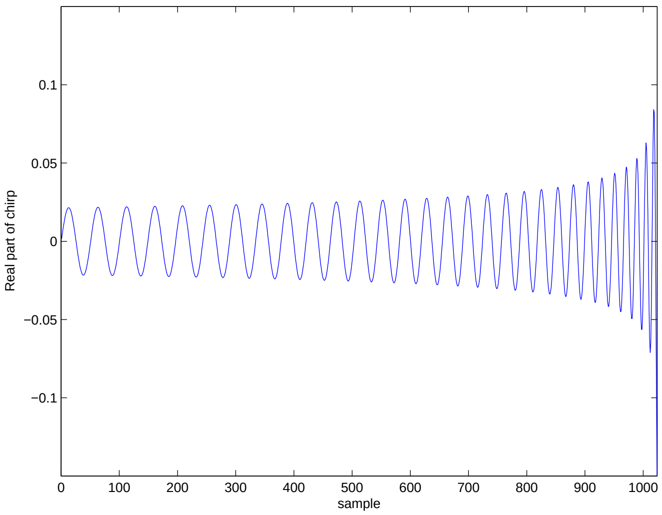


Figure 3.11: Real part of a simulated gravitational wave

many waveforms. Another interesting aspect is that the best chirplet path automatically adapts to the unknown local complexity of the signal; it uses short templates whenever required and longer templates when the signal exhibits some coherence over longer periods of time. Here, the path is refined where the instantaneous frequency starts to rise, which occurs near the end of the period under study.

3.5 Discussion

We have presented a novel and flexible methodology for detecting nonstationary oscillatory signals. The approach chains together empirical correlations to form meaningful signals which may exhibit very large correlations with the unknown signal we wish to detect. Our experiments show that our algorithms are very sensitive over very broad classes of signals.

3.5.1 Connection with other works

While working on this project ([22]) and writing [23], we became aware of the recent and independent work of Chassande-Mottin and Pai which is similar in spirit to ours [26]. In this paper, the authors also search for a chirplet chain in a graph. Despite this similitude, our approach is distinct in several aspects. First, whereas Chassande-Mottin and Pai use chirplets at a single scale, we use a multiscale dictionary which provides high flexibility and adaptivity to the unknown structure of the signal (see Section 3.1.2); the last example

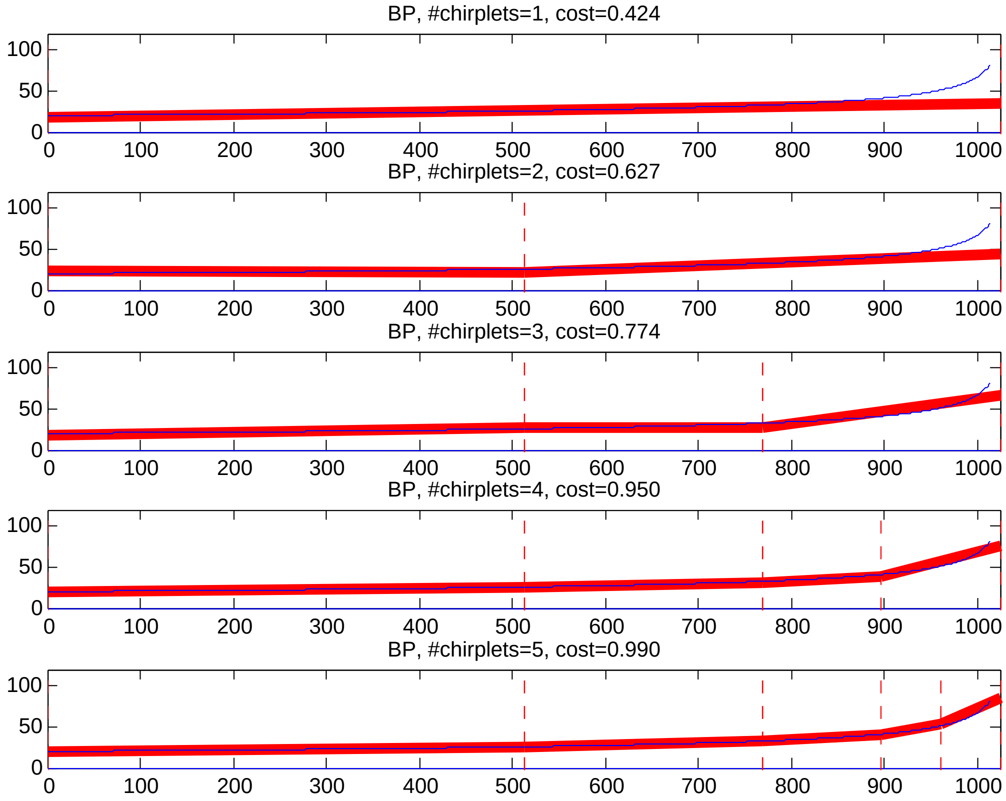


Figure 3.12: Chirplet paths returned by the BP test for path sizes equal to 1, 2, 3, 4, and 5 (the chirplets are adapted to have an amplitude varying linearly with time). The signal is a simulated gravitational wave. The cost here is simply the correlation between the waveform and the best chirplet path, so that a value of 1 indicates a perfect match. The horizontal and vertical axes indicate time and frequency. The thin line is the 'true' instantaneous frequency of the waveform. The thick line is the value of the instantaneous frequency along the path.

in Section 3.4 also clearly demonstrates the promise of the multiscale approach for the practical detection of gravitational waves. Consequently our detection strategy based on the multiple comparison between test statistics with varying complexities is of course very different. Second, while we find the best path by dynamic programming, the best chirplet chain in [26] is not the solution to a tractable optimization problem since the statistic which needs to be maximized over a set of chirplet paths is not additive. Therefore the authors need to resort to a series of approximations involving time-frequency distributions such as the WVD to obtain an approximate solution. This makes our approach also different and more general since the methodology proposed in this paper may be applied in setups which have nothing to do with chirplets and chirp detection.

Finally, the aforementioned reference does not address the problem of detecting chirps with a time varying amplitude, and also assumes that the noise in the data is white or has been “whitened” in some fashion (the detection method in [26] requires white noise). In contrast, the statistics in this paper have a natural interpretation in terms of likelihood ratios, and can be adapted effortlessly to more sophisticated setups in which the noise may be colored and in which the amplitude may also be rapidly varying and so on. Only the chirplet costs need to be changed while other algorithms remain the same.

Chapter 4

Two-Stage Best Path Test

The BP test assumes that the chirp is either present at all times or the data is pure noise. Even though our method might be able to detect a signal that is present in a considerable portion of the data, this detection problem is a bit idealistic. It is easy to imagine practical situations where one collects a long stream of data, which is typically just noise, and somehow has to weed through it to look for evidence of signal presence. A concrete example is the search for gravitational waves we discussed in the Introduction. This chapter extends the BP test to deal with this situation. The results from numerical experiments which are presented in Section 4.6 are promising.

4.1 Setup

Before presenting a methodology to deal with these situations, we state the abstract model problem we will assume throughout this chapter. Assume we have a long stream of sampled data

$$y_k = \lambda f_k + z_k, \quad k \in \mathcal{I} := \{0, 1, \ldots, M-1\}; \tag{4.1}$$

where $z = (z_k)$ is a vector of random errors distributed according to some known multivariate Gaussian distribution, and λ is an unknown real scalar. $f = (f_k)$ is a vector of uniform samples of an unknown continuous function f_c belonging to a known class of signals, $\mathcal{F}$, where $f_k = f_c(k/N_s)$, $k \in I$, and $1/N_s$ is the frequency of sampling. The signals in $\mathcal{F}$ have a time support contained in I, bounded between two known constants, L_{min} and L_{max}:

$$1 \leq L_{min} \leq |supp(f)| \leq L_{max} \leq M.$$

Given y_k, the goal is to decide between,

$$H_0 : \lambda = 0, \quad \text{i.e., the data is only noise,}$$

and

$$H_1 : \lambda > 0, \quad \text{i.e., there is a signal } s \in \mathcal{F} \text{ somewhere in the data.}$$

Or in plain words, we would like to detect a presence of a signal in a long stream of data where the duration of the signal is shorter than the data at hand and we do not know where it starts or ends. Although we are interested in the situation where $\mathcal{F}$ is a class of chirps, the ideas and methodology we are about to present could be applied to various other problems. The case we are most interested in is when $\mathcal{F}$ is a nonparametric class of functions such that, $|supp(f)| \ll M$ for every $f \in \mathcal{F}$ (i.e., signals that cannot be described with few parameters, and whose duration is much shorter than the length of the data stream). Thus, we are faced with several challenges:

1. The unknown signal belongs to a nonparametric class of functions.
2. The support length of the signal is unknown.
3. The position of the signal in the data stream is unknown.

Is it possible to design a method that can efficiently deal with all these situations, is statistically powerful, and is computationally feasible? The goal of this chapter is to provide a method that satisfies this and provide numerical evidence in its support.

4.2 Method of Matched Filtering

Perhaps the most popular method for detection problems of the form (4.1) is the method of matched filtering we briefly mentioned in the introduction. Assume noise vector $z = (z_k)$ in (4.1) is multivariate Gaussian with mean zero and covariance matrix Σ; i.e., $z \sim N(0, \Sigma)$. Assume that $\lambda > 0$ and that $\mathcal{F}$ is a parametric class of signals which is approximated by a set $\tilde{\mathcal{F}}$. Then the GLRT test would reject H_0 for big values of

$$T = \max_{f \in \tilde{\mathcal{F}}} \frac{|y^T \Sigma^{-1} f|^2}{f^T \Sigma^{-1} f},$$

where we write y and f as column vectors. This statistic is the maximum (noise-weighted) correlation of the signals in $\tilde{F}$ with the data. For each $f \in \tilde{F}$ which has a support much shorter than the data, the discrete set would typically include every time translation of f. In the case of white noise, i.e., $\Sigma = \sigma^2 I_M$, we could calculate the inner product $y^T f$ for every time translation of f very rapidly by a fast convolution. If Σ is circulant, as in the case of circular stationary noise, Σ^{-1} is translation invariant and we can also calculate the inner products rapidly by means of the FFT algorithm. Thus, the cost due to every template f, in the template bank is about $O(M \log M)$. Then the cost of calculating T for a template bank with K templates would be

$$K \times O(M \log M).$$

4.3 Motivation

Let $\mathbf{J}$ be the set of all intervals in $\mathcal{I}$ of lengths L_{min} through L_{max}. Suppose we have a statistic $T(J)$ available to help us decide whether an interval $J \subset \mathcal{I}$ contains a signal $s \in F$ with $supp(s) \approx J$. Assume big values of $T(J)$ are evidence against the null hypothesis H_0. Then we could imagine scanning the data, by calculating $T(J)$ for every possible interval $J \in \mathbf{J}$. To make a decision based on the values of these statistics we would need to put the comparisons on a common statistical scale. If the distribution of $T(J)$ is independent of J under H_0, we could design a test that would reject H_0 if

$$T^* := \max_{J \in \mathbf{J}} T(J)$$

exceeds a threshold. Otherwise, if we knew or could estimate the distributions of $T(J)$, $\forall J \in \mathcal{I}$, under H_0, we could base the decision on comparing P-values; e.g., reject H_0 when

$$T^* := \min_{J \in \mathbf{J}} P_{H_0} \left(T(J) \geq T_{obs}(J) \right) \tag{4.2}$$

is small enough, where $T_{obs}(J)$ is the observed value of $T(J)$. In both cases, we base our decision on the statistic $T(J)$ that gives us the strongest evidence against H_0. Postponing the important question of whether such a procedure would be good at all from a statistical point of view, we will instead focus on the computational aspect.

Since we are interested in the situation $L_{min} \ll M$, $L_{max} \ll M$, we would have about $(L_{max} - L_{min})M$ intervals to consider, possibly causing the "brute force" test we just described to be computationally prohibitive unless $T(J)$ is very simple, such as in the method of matched filtering which we will discuss in Section 4.2.

4.3.1 Detection of intervals with elevated mean

Scanning every possible interval, as in (4.2), is perhaps an overkill. This could especially be the case if the statistic $T(J)$ is adaptive, in the sense, that it does not rely on the interval J matching the unknown signal's support exactly.

Instead, we might consider a smaller set of intervals that try to approximate every possible interval in $\mathbf{J}$. In [10], Arias-Castro, Donoho, and Huo show that any interval can be approximated by a set of intervals that are short chains of dyadic intervals. They use this fact to design a near-optimal detection strategy for solving the following problem of detecting intervals of elevated mean: Let the data be of the form

$$y_k = \lambda \cdot 1_{\{a \le k < b\}} + z_k, \qquad k = 0, \ldots, M-1;$$

where $z = (z_k)$ is a vector of i.i.d. standard normal random variables, and where the endpoints a, b of the interval, obeying $0 \le a < b \le M-1$, and the signal amplitude $\lambda > 0$ are assumed to be unknown. They studied the conditions of asymptotic minimax detectability (see Section 9.2 for a definition of these terms). Applying the GLRT principle for this problem, we get the statistic

$$X^* = \max_{0 \le a < b \le M} \langle X, \xi_{a,b} \rangle,$$

where $\xi_{a,b}(k) = \frac{1_{\{a \le k < b\}}}{\sqrt{b-a}}$. At the two extremes (i) $b - a = M$, and (ii) $b - a = 1$, we have two classical hypothesis-testing problems: Under H_0 all the Gaussians are i.i.d. with mean zero, and the alternative H_1 is either

(i) all the Gaussians have a common mean greater than zero; or

(ii) one of the Gaussians has a positive mean, while the others have mean zero. I.e., there is a "spike" at an unknown location in the data stream.

The thresholds of asymptotic minimax detectability for the signal level $\lambda = \lambda_M$ as $M \to \infty$, are

$$(i)\ \lambda_M \sim 1/\sqrt{M}, \quad \text{and} \quad (ii)\ \lambda_M \sim \sqrt{2\log(M)}.$$

As the authors point out, it appears to be most natural to work with the normalized amplitude $A = \lambda/\sqrt{b-a}$ for studying the threshold of detectability. We immediately see, because of (ii), that A cannot grow slower than $\sqrt{2\log(M)}$. The authors show that if A grows slightly faster, or $A = \sqrt{2(1+\eta)\log(M)}$ for any $\eta > 0$, the GLRT is asymptotically powerful. This is the optimal behavior, since if A grows like $\sqrt{2(1-\eta)\log(M)}$, every sequence of tests is asymptotically powerless. The number of intervals is $O(M^2)$ and a straightforward implementation would require $O(M^2)$ operations to calculate X^*. However, using an idea considering a smaller set of *extended dyadic intervals*, the authors show that there is an algorithm that requires $O(M)$ operations that is asymptotically powerful if $A = \sqrt{2(1+\eta)\log(M)}$, achieving the same asymptotic statistical performance as the full GLRT.

The framework we are about to describe relies partly on their set of extended dyadic intervals which we will describe next.

4.3.2 Extension of dyadic intervals

We will adopt from [10] the definition of the set of extended of dyadic intervals and the definition of the measure of *affinity* between two intervals.

Definition 2. *The measure of affinity between intervals I and J is defined by*

$$\rho(I,J) = \frac{|I \cap J|}{\sqrt{|I|}\sqrt{|J|}}.$$

Obviously $0 \leq \rho(I,J) \leq 1$, with $\rho(I,J) = 1$ if and only if I and J are identical, and 0 if their supports are disjoint. We will consider dyadic intervals contained in $\mathcal{I}$, just as we did in the construction of the multiscale chirplets:

Definition 3. *If $M = 2^J$, a* dyadic subinterval *is an interval of the form,*

$$I_{s,k} = \{k2^s, \ldots, (k+1)2^s - 1\},$$

where $0 \leq s \leq J$ and $0 \leq k \leq 2^{J-s} - 1$. The cardinality of an interval $I_{s,k}$ is $|I_{s,k}| = 2^s$.

The next definition, taken from [10], describes a method of chaining together dyadic intervals in a systematic way.

Definition 4 (l-level extension)**.** *We say that the interval J_l is an l-level extension if it is constructed in the following way:*

1. *Start from a base J_0 which is either a dyadic interval $I_{s,k}$ or the union of two adjacent dyadic intervals at the same scale, $I_{s,k}$ and $I_{s,k+1}$ where k is odd (if k is even, then the union would be equal to a dyadic interval at scale index $s+1$).*

2. *At stages $m = 1, \ldots, l$ extend J_{m-1} to J_m by attaching a dyadic interval of length $2^{-m}|I_{s,k}|$ at either or both ends of J_{m-1} or by doing nothing so that $J_m = J_{m-1}$.*

The collection of all l-level extensions of the dyadic interval I is denoted by $\mathbf{J}_l[I]$ and the collection of all l-level extensions is denoted by $\mathbf{J}_{M,l}$.

Figure 4.1 shows an example of l-level extensions for $l = 1$ and the base interval $J_0 = I_1$ being either a dyadic interval of length 2^s or a union of two dyadic intervals.

The set of extended dyadic intervals provides a very good approximation of the set of all intervals as the following lemma from [10] describes:

Lemma 2.

$$\#\mathbf{J}_{M,l} \leq M4^{l+1}$$

$$\rho^*_{M,l} = \min_{I \in \mathcal{I}} \max_{J \in \mathbf{J}_{M,l}} \rho(I, J) \geq 1/\sqrt{1 + 2 \cdot 2^{-l}}.$$

This lemma tells us that using a small set of extended intervals, we can approximate any interval in $\mathcal{I}$ well.

For the problem we have in mind, we focus on signals with support length in the range $[L_{min}, L_{max}]$. Therefore, we only need to consider a small set of scales for the dyadic intervals that we wish to extend. The number of dyadic intervals of length 2^s is $M/2^s$ and the number of l-level extensions per dyadic interval does not exceed $2{\cdot}4^l$. The number of l-level extensions for dyadic intervals of length 2^s does therefore not exceed $M \cdot 2^{2l-s+1}$.

4.4 Framework for Signal Detection When Support is Unknown

Lets consider the detection problem (4.1). Assume $T_1(J)$ and $T_2(J)$ are two statistics for determining how likely it is that a signal $s \in \mathcal{F}$ is contained in the subinterval J. The scales of the dyadic intervals to consider are determined by L_{min} and L_{max}. A possible choice is the set of dyadic intervals of length 2^s where $s = \lfloor \log_2 L_{min} \rfloor - 1, \ldots, \lceil \log_2 L_{max} \rceil$. This ensures that in any interval of length L, $L_{min} \leq L \leq L_{max}$, we could find a dyadic interval from this set whose length is roughly equal to $L/2$. The general description of the method is as follows. Fix the maximum extension level l_{max}. Then the procedure is based on two separate stages:

1. *First stage*: For each dyadic interval $I_{s,k}$ we wish to consider, calculate $T_1(I_{s,k})$ and "tag" $I_{s,k}$ as promising if the value exceeds a predescribed threshold. Construct a list of promising intervals $\mathcal{J} = \{J_1, \ldots, J_n\}$. Note that n could be random; it depends on how we tag intervals as promising.

2. *Second stage*: Extend the promising intervals:

   (a) Take $I \in \mathcal{J}$.

   (b) For $l = 1, \cdots, l_{max}$, calculate $T_2(J)$ for each $J \in \mathbf{J}_l[I]$.

3. *Decision*: Use the results from second stage to decide whether the data stream contains a signal. This would have to be done by some multiple comparison procedure since the decision has to be based on values of many statistics. A possible approach would be to take the minimum P-value in the second stage and compare it to the distribution of the minimum P-value under H_0 like we did for the BP test.

Remarks:

- T_1 and T_2 could be quite different and not necessarily lead to equally powerful tests if the signal support was known. The purpose of the first stage is to weed out as much of the data stream that "obviously" appears to contain only noise, and leave only the part that could potentially include a signal. The emphasis here is more on the speed of computation rather than detection sensitivity. If T_1 allows us to weed out with high confidence, say, 90% of the dyadic intervals, this could be a considerable gain in speed. It could allow us to spend greater time applying a more expensive and sensitive detection procedures on the 10% of the intervals that are left. Whether this is possible, and how simple T_1 can be, depends of course on the set $\mathcal{F}$ of unknown signals.

- Tagging intervals in the first stage could be done by using P-values or a fixed threshold. An interval J would be tagged if $T_1(J)$ has a P-value small enough, or its value exceeds a threshold.

- If for some reason the null distribution of $T_1(I_{s,k})$ is not available, we could tag based on the ordering of the observed value of the statistics. For each interval length 2^s, we would choose the number, q_s, of intervals that will be tagged promising. The choice of a suitable q_s would typically need to be based on numerical simulations (Monte Carlo). In order for the method to work well, the parameters (q_s) need to be chosen such that with overwhelming probability at least one dyadic interval with significant overlap with the support of the signal is labeled as promising. The computational resources at hand might put an upper limit on (q_s). As a last resort, one could choose (q_s) close to this upper limit.

4.5 Two-Stage BP Test for Chirp Detection

Consider the detection problem (4.1) where $\mathcal{F}$ is a set of chirps. Calculating a multiscale BP statistic for every possible interval in a long data stream might be prohibitively expensive in practice. Instead we will introduce a *two-stage BP test* for chirp detection, based on the framework described in Section 4.4. We design the statistics T_1 and T_2 based on the BP test we discussed in the previous chapter. This provides us with rapid algorithms, adaptivity, and flexibility.

4.5.1 Statistic for the first stage

The purpose of T_1 is to weed out as much as possible of the data that appears to only contain noise. Its design depends highly on how wildly the instantaneous frequency of the chirps can change over their time support. We mention two choices which both allow for rapid calculation of T_1, but there could be other possibilities.

Choice 1: Maximum Chirplet Coefficient. If the class of signals we wish to detect consists of chirps whose frequency evolution is very smooth, we might have a chance of identifying promising intervals by looking at the maximum chirplet cost on each interval:

$$T_1(I_{s,k}) = \text{maximum chirplet cost on interval } I_{s,k}.$$

This would be the case when the second derivative of the phase is almost constant for a considerable portion of the signal duration, causing the instantaneous frequency to change almost linearly with time.

Choice 2: Monoscale Chirplet Analysis. When the chirps exhibit a more complex structure, we could choose $T_1(I_{s,k})$ to be the value of the best path in a monoscale chirplet graph whose topology is simple. To speed up computations, it is possible to reuse calculations (see Section 4.5.4). The choice of scale depends on the class of signals we wish to be able to detect. Instead of sticking to one scale, we could consider a few of them for more adaptivity.

4.5.2 Statistic for the second stage

In the second stage the search has been focused on relatively few candidate intervals. Therefore, we can afford to apply more sophisticated methods. We will take T_2 to be based on the BP test for a multiscale chirplet graph. The discretization and allowable scales of chirplets in the chirplet graph used for T_2 could be made to depend on the length of the dyadic interval being extended.

4.5.3 A sample configuration for the two-stage BP test

Here we provide a concrete example which will serve as the base for our numerical simulations in Section 4.6.

The maximum extension level for the second stage is $l_{max} = 1$. Assume that using T_1 in the first stage we have tagged a dyadic interval $I_{s,k}$ as promising. Figure 4.1 shows all the interval extensions for this dyadic interval. If k is even, the number of interval extensions is 4, and if k is odd, it is 8. In the second stage we would therefore need to calculate 4 or 8 different statistics T_2.

Consider the intervals in part (a) of Figure 4.1. Then the chirplet graph topology for intervals I_1 and I_2 up to time $t_{1,0}$ would be identical. Also, for I_2 we would force any chirplet path to have chirplets ending and starting at time $t_{1,0}$. The same holds for intervals I_1 and I_2 in part (b) of the figure. The portion of the chirplet graph from $t_{1,0}$ to $t_{1,1}$ would be identical to the first half of the chirplet graph for interval I_1.

The topology of the chirplet graphs for intervals I_2 and I_3 were taken to be identical (these intervals are both equally long). The part of the chirplet graph for interval I_4 from $t_{0,1}$ to $t_{1,0}$ was taken to be identical to the graph for I_3. The topology of the part of the chirplet graph from $t_{1,0}$ to $t_{1,1}$ was taken to be identical to the part from $t_{0,1}$ to $t_{0,0}$.

Note that with this choice of chirplet graphs for the interval extensions we are making the method somewhat asymmetric and not completely multiscale. For example, although there is a chirplet path of length 1 for the base interval, every path for its extensions has to constitute at least 2 chirplets. For the case when the base interval $I_{s,k}$ is doubled, the paths on the intervals I_2 and I_3 have to have at least 3 chirplets, and the paths for I_4 have to have 4 chirplets or more.

The main reason for this choice is to minimize computational burden since this way we can reuse the chirplet costs for the base interval and our computation of the Best Path. If computational cost is not of much concern it would be possible to get around the restrictions these choices impose.

4.5.3.1 Speculations about computational cost

Assume q_s is the fraction of dyadic intervals of length 2^s that are tagged as promising in the first stage so the total number of tagged intervals is therefore

$$q_s M/2^s.$$

For a data sequence of length N, we would typically have an upper bound on the total computational complexity of the BP test of the order $N^2(\log N)^2$. The cost is dominated by the calculation of the chirplet coefficients, and this is the order of the complexity for a dense chirplet graph including every possible scale. Thus, the complexity of applying T_2 to all the extended dyadic intervals of length $N = 2^s$ does not exceed

$$O(q_s M/N \times N^2(\log N)^2) = O(q_s M 2^s s^2).$$

If we are in a situation where $N \ll M$, the computational cost of the second stage for the BP test does not exceed approximately M. The first stage uses a cheaper statistic and could be configured so its computational cost does at least not exceed that of the second stage.

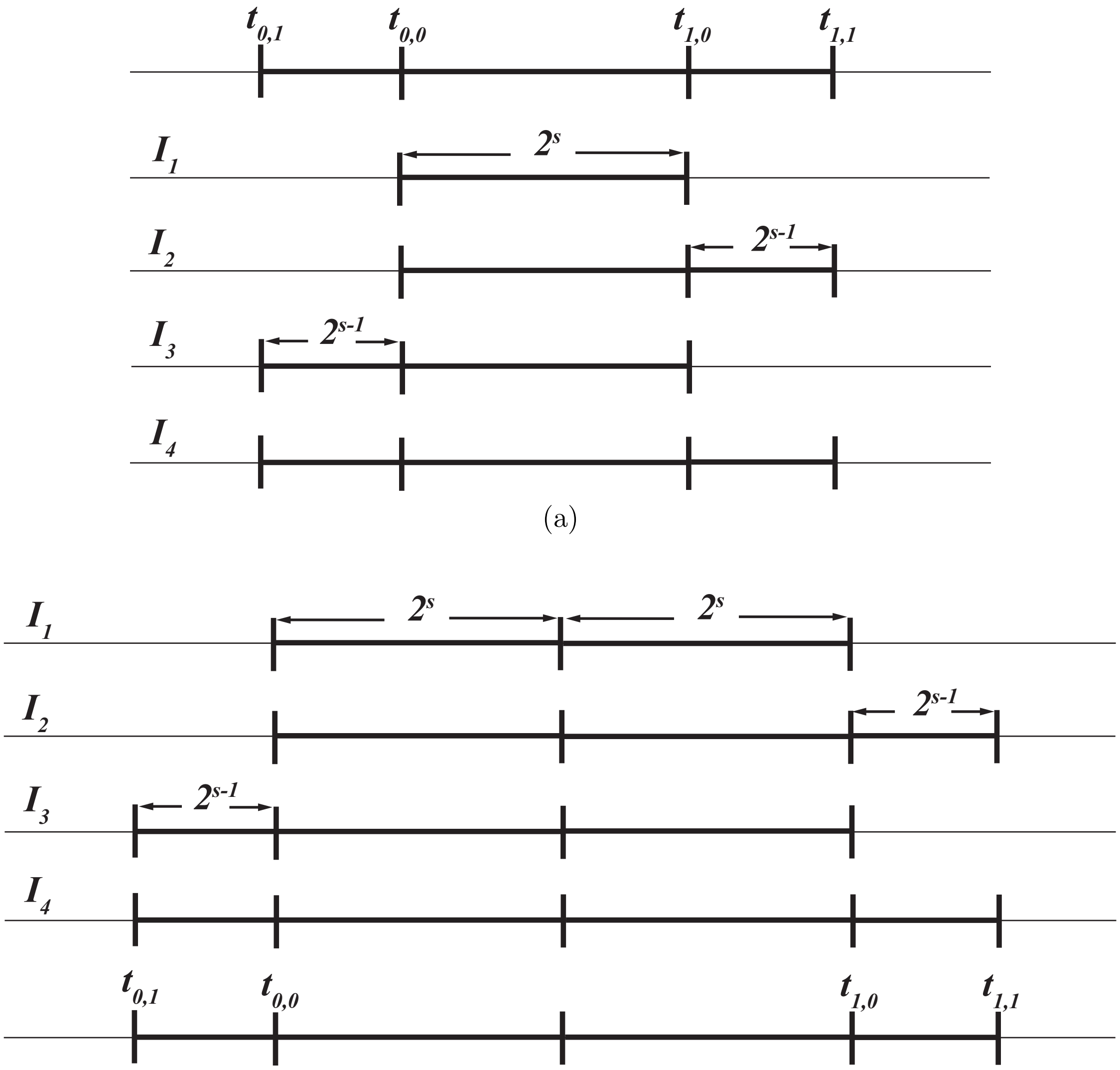
Recall that the cost of applying the method of matched filtering with a template bank of cardinality K is about $K \times M \log M$. If our considerations above hold, we expect the two-stage BP test to be much faster than even a moderately sized template bank. This is, in fact, what we have experienced empirically.

4.5.4 Speeding up computations for the two-stage BP test

It is possible to reuse calculations to speed up the computations of the two-stage BP test statistic. Consider a dyadic interval I that has been tagged as promising and we wish to extend. If the same discretization and chirplet graph is used in all of the cases, then we could reuse calculations when calculating the best path for the different intervals under consideration. First of all we could reuse the cost of calculating the chirplet cost. Recall that this is the dominant factor in the computational cost. Secondly, the nature of the BP algorithm also allows us to reuse a lot of calculations. If we fix a starting point in time, t_0, and march forward, the algorithm will provide us with the value of the Best Path starting at time t_0 to any vertex at time greater than t_0. Assume we want the values of the Best Paths starting at time t_0 and ending at times $t_1, \ldots, t_k$. Further assume that chirplets in the path can end at these times and $t_0 < t_1 < \ldots < t_k$. Then we can find the value of all the Best Paths starting from t_0 to $t_1, \ldots, t_k$, respectively, with only one sweep of the BP algorithm through the chirplet graph. The cost of the algorithm is the same as for calculating the Best Path from t_0 to t_k, the Best Paths at the intermediate times come for free.

4.6 Numerical Simulations

Since one of the targeted applications for our detection procedures is the search of gravitational waves, we choose to demonstrate the two-stage BP test in a setup which is meant to resemble the situation in the LIGO detectors. This is of course an idealistic academic model of the real situation, and the ultimate test would be to apply our methodology to real data. However, we hope that these results will show that these tests are powerful and have good potential for being useful in practical applications. The code developed for generating these



(a)

(b)

Figure 4.1: The promising interval I_1 and its extensions

results is part of *Chirplab*, which is publicly available (see Section 10.1).

4.6.1 Sampling model

To show the connection between the discrete model we use for numerical simulations and a possible data acquisition process, we present a classical sampling model for situations where analog measurements are converted into discrete sequences [64]. We imagine we have an apparatus which senses a continuous-time data stream

$$x_c(t) = s(t) + Z(t),$$

where $s(t)$ is either an unknown chirping signal of finite time-support or identically zero, and $Z(t)$ is a stationary Gaussian process with mean zero and a known power spectrum $P(\omega)$. We also assume that, in the frequency-domain, most of the energy of $s(t)$ is contained in a frequency band $[-\Omega, \Omega]$ for some positive number Ω. Figure 4.2 shows a diagram of the sampling model we are about to describe. First the signal $x_c(t)$ is fed into an analog antialiasing filter with a frequency response $H_a(\omega)$. This is to remove noise in the higher frequency range which otherwise, after sampling, would be aliased into the low-frequency range. To prevent this aliasing, we need to force the input signal to be bandlimited to frequencies below one-half of the sampling rate. We choose the sampling interval T based on the frequency support of the signal $s(t)$, that is, according to Nyquist, to satisfy $2\pi/T \geq 2\Omega$. Therefore, ideally, we might want an anti-aliasing filter with a frequency response

$$H_a(\omega) = 1_{|\omega| \leq \pi/T}(\omega).$$

But sharp analog filters are difficult and expensive to implement. In practice, the signal can be over-sampled at, say, the sampling interval $T' = T/D$, where D is a positive integer. The antialiasing filter filter is designed to have a gradual cutoff with significant attenuation at the frequency π/T'. If we assume that $H_a(\omega) = 1$ for $\omega \in [-\Omega, \Omega]$ and zero outside of the band $[-\pi/T', \pi/T']$ (although, in reality, it would only be approximately zero in that band), the output signal from the antialiasing filter is

$$x_a(t) = s(t) + Z_a(t),$$

where $Z_a(t)$ is a stationary Gaussian process with mean zero and the power spectrum

$$P_a(\omega) = |H_a(\omega)|^2 P(\omega).$$

Then the sampled sequence resulting from the analog-to-digital conversion (A/D conversion) is

$$x[n] = x_a(nT') = s(nT') + z_a[n],$$

where $z_a[n]$ is a discrete-time Gaussian process with mean zero and the power spectrum $P_a^d(\omega) = 1/T' P_a(\omega/T')$, for $\omega \in [-\pi, \pi]$ (note that this power spectrum is 2π-periodic since now we have a discrete-time signal). Finally we can reduce the sampling rate by using a sharp antialiasing digital filter with a frequency response $H^d(\omega)$ with cutoff at π/D followed by downsampling by a factor D. Ideally, $H^d(\omega) = 1_{|\omega| \leq \pi/D}(\omega)$, in order to leave the deterministic part of the signal $x[n]$ unchanged. In that case, the resulting signal is

$$y[n] = s(nDT') + z[n] = s(nT) + z[n],$$

where $z[n]$ is a discrete-time Gaussian process with mean zero and the power spectrum

$$\begin{aligned} P^d(\omega) &= \frac{1}{D}|H^d(\omega/D)|^2 P_a^d(\omega/D) = \frac{1}{DT'}|H^d(\omega/D)|^2 P_a(\omega/(DT')) \\ &= \frac{1}{T}|H^d(\omega/D)|^2 |H_a(\omega/T)|^2 P(\omega/T), \quad \omega \in [-\pi, \pi]. \end{aligned}$$

Using ideal filters we get

$$P^d(\omega) = \frac{1}{T} P(\omega/T), \quad \omega \in [-\pi, \pi].$$

If we know the power spectrum of the noise in the continuous-time data stream $x_c(t)$, the above expression gives us the power spectrum of the noise in the sampled sequence which can then be used for constructing the covariance for the discrete-time noise process. The autocorrelation for the discrete random sequence $z[n]$ is

$$R[m] := E(z[n]z[n+m]) = \frac{1}{2\pi}\int_{-\pi}^{\pi} P^d(\omega) e^{i\omega m} d\omega.$$

$$\underset{x_c(t)}{\longrightarrow}\boxed{\text{Antialiasing filter}}\underset{x_a(t)}{\rightarrow}\boxed{\text{A/D conversion}}\underset{x[n]}{\longrightarrow}\boxed{\text{Antialiasing filter}}\underset{x_{lp}[n]}{\longrightarrow}\boxed{\downarrow D}\underset{y[n]}{\longrightarrow}$$

Figure 4.2: Sampling model for simulations for the two-stage BP test

The integral can be approximated by the trapezoidal rule giving us

$$R[m] \approx \frac{1}{N}\sum_{k=-N/2}^{N/2-1} P^d\left(\frac{2\pi k}{N}\right)e^{i2\pi\frac{km}{N}}, \tag{4.3}$$

for an even integer N. This gives a relation between the autocorrelation of the discrete-time noise process $z[n]$ and the power spectrum via the discrete Fourier transform. For simulation purposes we can therefore simply generate the sequence $y[n]$ directly. Below is a description of how we simulated data based on this sampling model:

- We generated $y[n]$ directly by simulated blocks of data, $y[n]$, $0 \leq n \leq N-1$, where $N = 2^{16} = 65,536$, unless specified otherwise.
- The sampling rate in the A/D conversion is $16,384$Hz (T'=1/16,384 s).
- $x[n]$ is downsampled by a factor $D = 8$ so the deterministic signal in the data is effectively sampled at the rate $2,048$Hz.
- We assumed the frequency response of the antialiasing filters to be ideal.
- The colored noise is simulated as circular noise on each block of length N. (This creates a slight disconnection between the sampling model and the simulated data but makes the simulation process easier.)

Switching from bracket notation to subscripts, the data model for the simulated data is:

$$y_k = \lambda s_k + z_k, \qquad k = 0, 1, \ldots, N-1, \tag{4.4}$$

where $N = 2^{16} = 65,536$, λ is a non-negative scalar, (s_k) is a vector of equispaced time samples of a real-valued chirp with support smaller than N, and (z_k) is a real-valued noise sequence sampled from a multivariate normal distribution $N(0,\Sigma)$. Since we chose to simulate circular noise, the covariance matrix Σ is circulant and therefore diagonalizable in the

Fourier basis. Thus, the noise is completely characterized by the eigenvalues, or *power spectrum*, $P = (P_k)_{1 \le k \le N}$, of the covariance matrix. We use (4.3) for the relationship between the covariance of the simulated noise and the spectrum in the data model.

4.6.2 Noise model

The power spectrum in the model (4.4) is based on a fit to the LIGO-I one-sided power spectral density given in [43]

$$\begin{aligned} S(f) = S_0 \big[&(4.49 f/f_{\text{ref}})^{-56} + 0.16(f/f_{\text{ref}})^{-4.52} \\ &+ 0.52 + 0.32(f/f_{\text{ref}})^2\big], \end{aligned} \tag{4.5}$$

where $f_{\text{ref}} = 150\text{Hz}$. This fit is only valid for frequencies above $f_s = 40\text{Hz}$, so to mimic high-pass filtered data, we roll off $S(f)$ below 20Hz. When calculating the BP statistics we only search for paths in the region where frequency exceeds f_s. For our simulation purposes, the exact value of the scaling factor S_0 does not matter, since the test signals $s = (s_k)$ will be normalized with respect to the noise spectrum. If we scale the noise spectrum, the test signal would be scaled with the same factor. Figure 4.3 shows a plot of the noise curve the power spectrum was sampled from; again, the scale on the y-axis does not matter. The highest frequency index, $k = N/2$, corresponds to the frequency $f = 1,024\text{Hz}$ in the power spectral density $S(f)$.

4.6.3 Definition of signal-to-noise ratio (SNR)

We will use the LIGO convention of the definition of SNR. For a real-valued signal $s = (s_k)$, written as a column vector, the SNR for the data $y_k = \lambda s_k + z_k$ as described above is:

$$\text{SNR} = \sqrt{(\lambda s)^T \Sigma^{-1} (\lambda s)} = \lambda \sqrt{s^T \Sigma^{-1} s}.$$

If we assume the signal s to be normalized such that $s^T \Sigma^{-1} s = 1$, then the SNR is simply equal to the signal level λ. Note that if we whiten the data, by multiplying both sides of equation (4.4) by $\Sigma^{-1/2}$, we get the equivalent data model

$$\tilde{y}_k = \lambda \tilde{s}_k + \tilde{z}_k, \qquad k = 0, 1, \ldots, N-1,$$

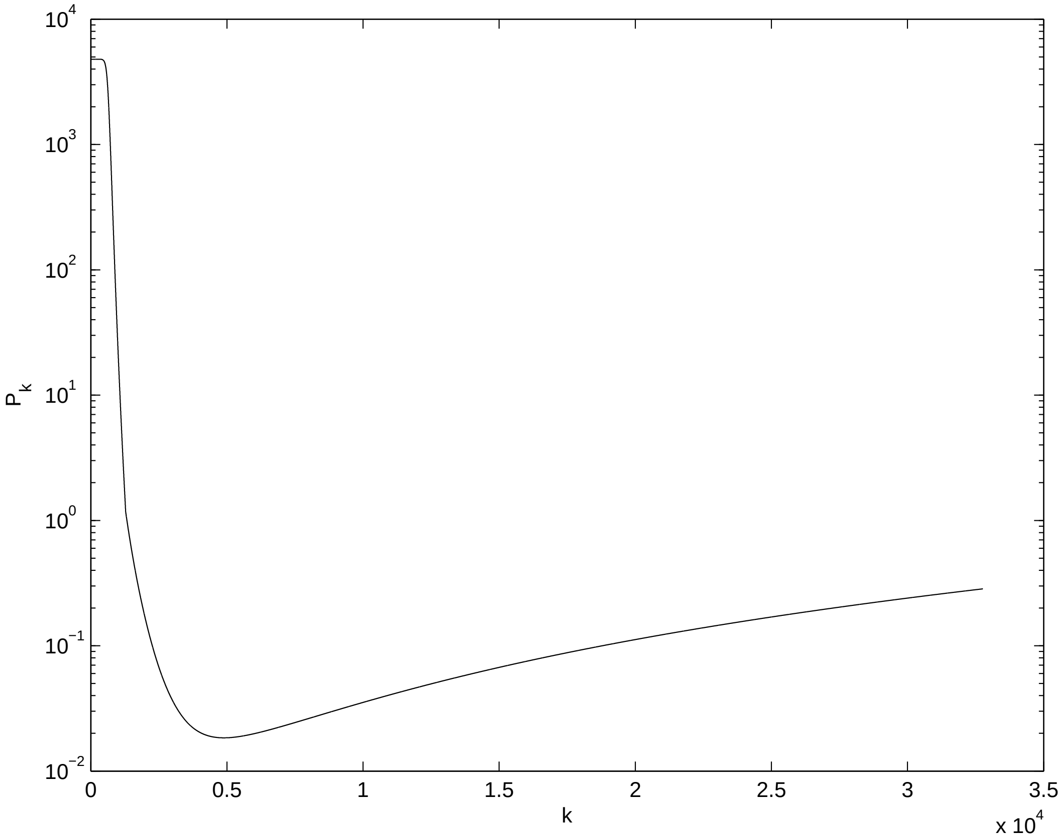


Figure 4.3: The power spectrum $P = (P_k)$ used for generating the noise. Plotted for $k = 1, \ldots, N/2$

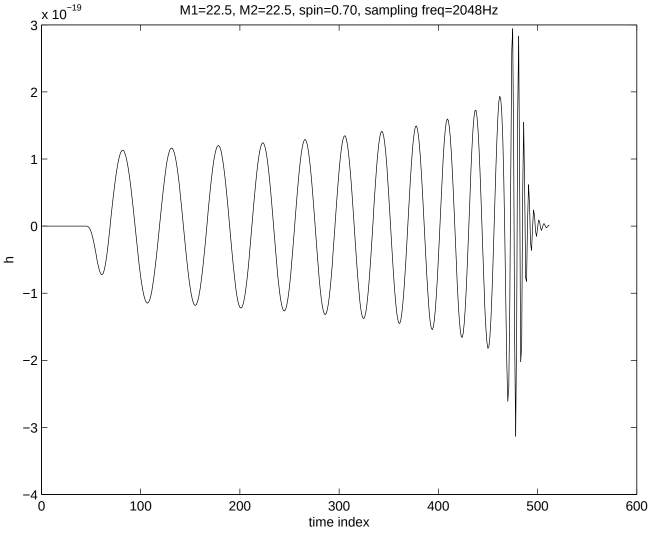


Figure 4.4: Simulated BBH coalescence with total mass $M = m_1 + m_2 = 45M_{\odot}$

where $\tilde{y} = \Sigma^{-1/2}y$, $\tilde{s} = \Sigma^{-1/2}s$, and $\tilde{z} = \Sigma^{-1/2}z \sim N(0, I)$. Then the normalization $s^T\Sigma^{-1}s = 1$ above is equivalent to taking $\|\tilde{s}\|^2 = 1$ for the transformed data. A common definition of SNR for the data model with the white noise is $\|\lambda\tilde{s}\|/\sqrt{N}$, so the LIGO convention for SNR differs by a factor $1/\sqrt{N}$.

4.6.4 Test signals: Simulated gravitational waves

Since we are trying to mimic LIGO data, we tested our methods using simulated gravitational wave signals. Detailed description of how these signals were constructed can be found in Appendix E. These test signals depend on three parameters, (m_1, m_2, a), which determine the shape and effective support of the signal [1]. m_1 and m_2 are the masses of two rotating bodies (measured in units of solar mass $M_{\odot}$), and a is a so-called spin parameter. The *total mass* of the system is $M = m_1 + m_2$.

Three different test signals were chosen with the following parameters (i) $(m_1, m_2, a) = (22.5, 22.5, 0.7)$, (ii) $(m_1, m_2, a) = (15, 15, 0.7)$, and (iii) $(m_1, m_2, a) = (10, 10, 0.7)$, yielding signals with three different support lengths. Figures 4.4, 4.5, and 4.6, show plots of these signals.

[1] By *effective support*, we mean the part of the signal which is oscillating higher than 40 Hz, since below that frequency the noise is overwhelming.

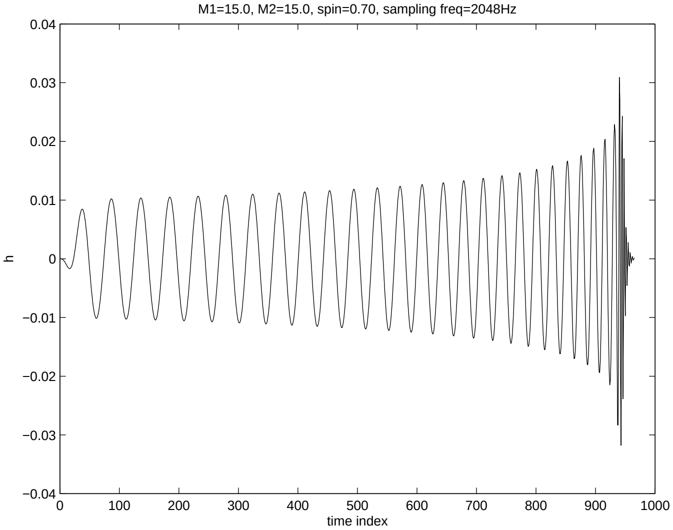


Figure 4.5: Simulated BBH coalescence with total mass $M = m_1 + m_2 = 30M_\odot$

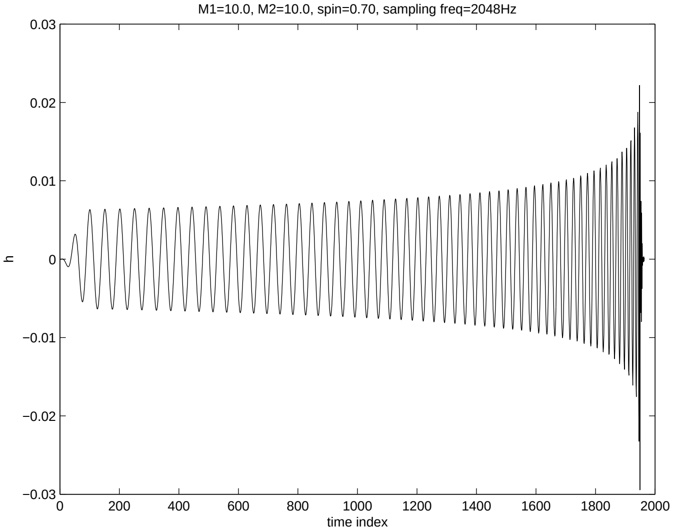


Figure 4.6: Simulated BBH coalescence with total mass $M = m_1 + m_2 = 20M_\odot$

4.6.5 Number of simulations

Number of simulations for BP test and known support: Since the support of the signals are different we used chirplet graphs of different time-supports for each case, tailored to each of the signals (i), (ii), and (iii).

- For each graph length, we randomly sampled 100,000 realizations of noise.
- For each signal and each signal level, we sampled the data model 10,000 times in order to compute detection rates.

Number of simulations for unknown support (two-stage BP): The same statistical procedure was performed on every realization.

- We randomly sampled about 150,000 realizations of circular stationary noise vectors of length 65,536 and performed the two-step BP on it in order to estimate a minimum P-value distribution (see description below).
- For each signal and each signal level, we sampled the data model 1,000 times in order to compute detection rates.

4.6.6 Configuration for the two-stage BP test

The first step in the two-stage BP method consisted of calculating the BP statistic on all dyadic intervals of lengths $L_1 = 2^7$, $L_2 = 2^8$, and $L_3 = 2^9$, using a monoscale chirplet graph G_I. This choice of lengths configures the test to detect signals of supports 128 through 1024, after the extension. The intervals corresponding to about 10% of the most extreme statistics were tagged as promising. That is, 50 for L_1, 30 for L_2, and 12 for L_3. Then every tagged interval was extended for extension level $\ell = 1$ and a multiscale BP statistic calculated for each interval extension. Let $G_{II}(L)$ be the set of chirplet graphs for the extensions for a tagged interval of length L. We used different configurations of $G_{II}(L)$ for different values of L. The configurations for G_I and $G_{II}(L)$ can be found in Appendix D in a format suitable for *Chirplab* (see Section 10.1).

Decision rule: The decision rule was based on the statistic for minimum P-value considerations for the value of the statistics in $G_{II}(L)$. We compared the values of statistics which were calculated for the same "seed" interval length L and chirplet graph topology in $G_{II}(L)$, and took the largest one. This gave us a vector V of statistics. The decision rule was based on estimating the P-value of every entry in V and picking the smallest one, P_{min}. Then we compared P_{min} to an empirical null distribution for this variable.

4.6.7 Results from simulations

We now explore the empirical performance of the detection methods proposed in this paper. To this end, we have developed *ChirpLab*, a collection of Matlab routines that we have made publicly available (see Appendix 10.1).

4.6.7.1 Two-stage BP test vs. BP test which assumes support is known

To investigate the price the two-stage BP test pays for not knowing the support of the unknown signal, we compared its performance with the BP test. Figures 4.7 and 4.8 show a comparison of the two methods for two different test signals. In each case we generated a data sequence of length $65,536$ and placed the signal so that the time index where it starts is at 10,000. The two-stage BP test was applied to the whole data stream while the BP test was applied to the portion of the data stream where the signal had its support.

The plots of the detection probabilities show that the price the two-stage BP test pays for not knowing the support in advance is low - or roughly, to achieve the same performance as the BP test which knows where to look in the data stream, the SNR needs to be about 10% higher. If we believe that the BP test is a powerful method for detecting unknown chirps for known support, this is a good indication that the two-stage BP test is performing very well.

4.6.7.2 Performance of two-stage BP test as length of data increases

It is important to have a feeling for how much the performance of the two-stage BP test degrades as the length of the data stream increases. Longer data provides more places for a signal to hide and our method would lose power. Figure 4.9 shows a comparison of the method for two different data lengths: $M = 2^{16} = 65,536$ and $M = 2^{17} = 131,072$. The only difference in the configuration of the method for $M = 2^{17}$ is that we doubled the number of intervals tagged as promising in the first stage. As a result, analyzing the data of length $M = 2^{17}$ is twice as expensive, in terms of computational cost, as analyzing the

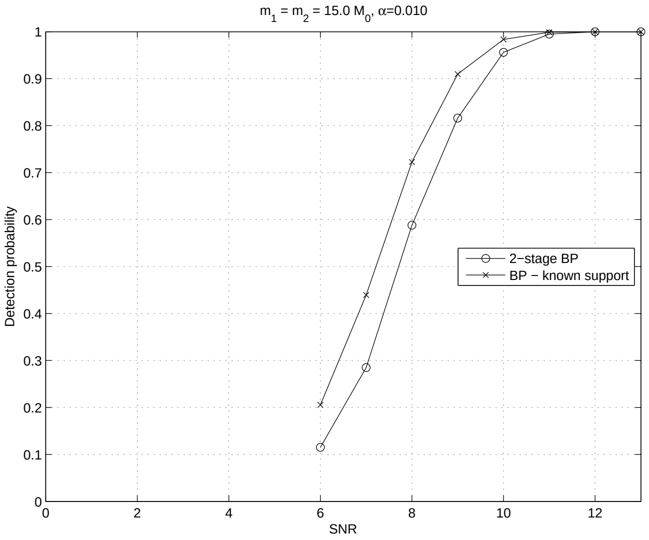


Figure 4.7: Comparison of performance for the BP test assuming the support to be known and two-stage BP test for the BBH coalescences with total mass $M = 30M_\odot$. (a) Detection probability as a function of SNR. The probability of Type I error is fixed at 1%.

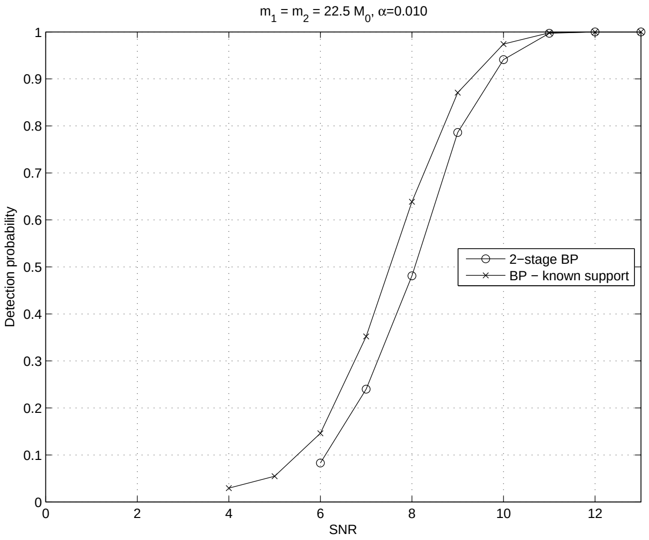
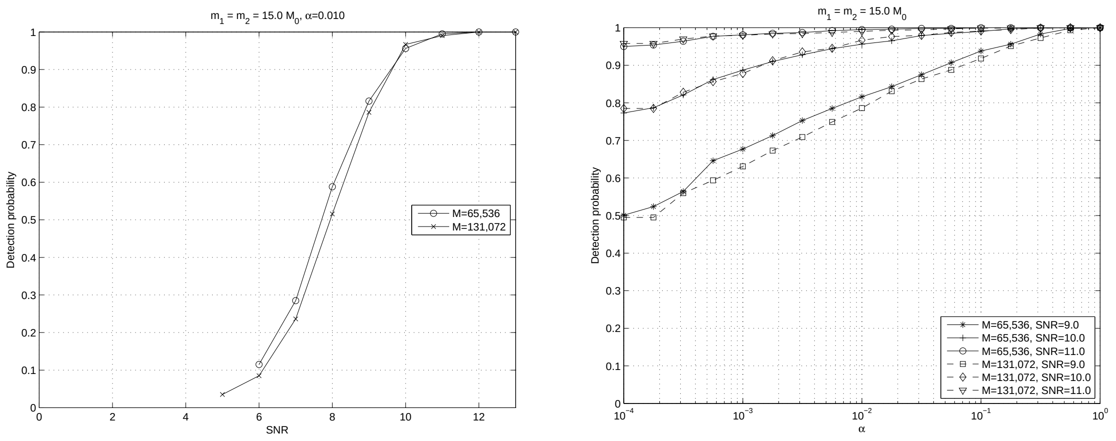


Figure 4.8: Comparison of performance for the BP test assuming the support to be known and two-stage BP test for the BBH coalescences with total mass $M = 45M_\odot$. (a) Detection probability as a function of SNR. The probability of Type I error is fixed at 1%.



(a) (b)

Figure 4.9: Comparison of performance of two-stage BP for the data lengths $M = 2^{16} = 65,536$ and $M = 2^{17} = 131,072$. The test signal is a gravitational wave from a BBH coalescence with total mass $M = m_1 + m_2 = 30M_{\odot}$. (a) Detection probability as a function of SNR. The probability of Type I error is fixed at 1%. (b) Plots of ROC for different values of SNR

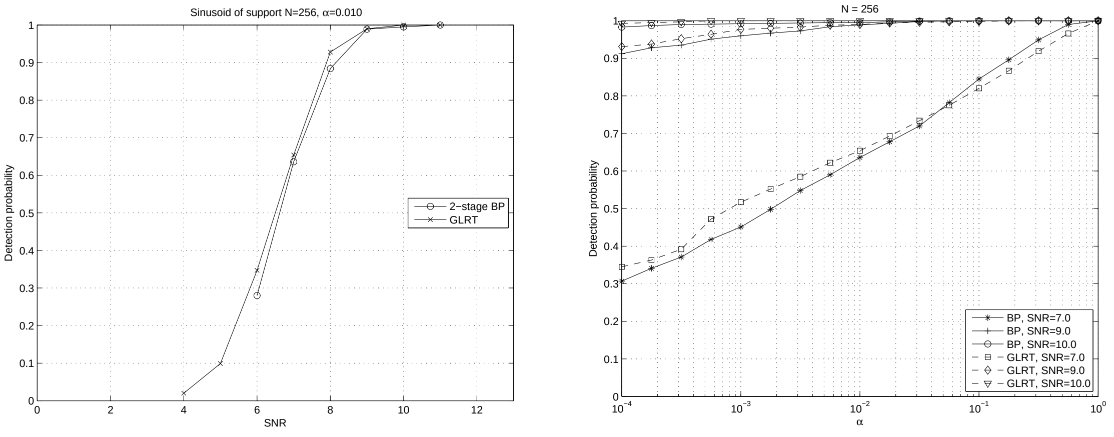
data of length $M = 2^{16}$. The plots show that there is hardly any difference in performance as the length of the data is increased.

4.6.7.3 Detection of a sinusoid with unknown support and frequency

Suppose the unknown signal is a monofrequency local sinusoid with an unknown frequency and phase offset. It is windowed by a smooth function of finite support $N \in \{256, 384, 512\}$. Let the set of frequencies be $\omega \in \Omega_N := \{2\pi m : m \in \{\lfloor 40/2048N \rfloor, \ldots, N/2-1\}\}$. Then we compare the two-stage BP test with the GLRT in Section 4.2 local sinusoids as the bank of templates:

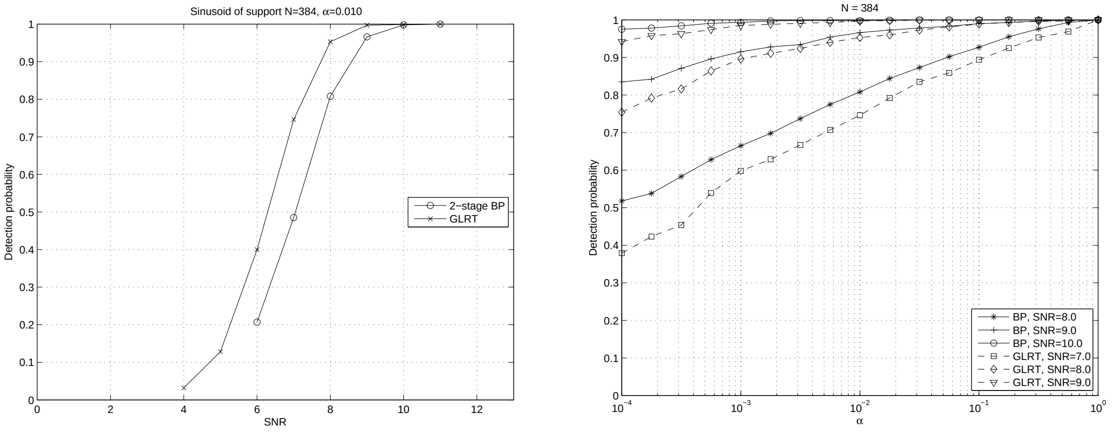
$$\{f : f_k = 1_{I_N}(k) e^{\imath\omega_N k/N}, I_N \text{ is an interval of length } N \in \{256, 384, 512\}, \text{ and } \omega_N \in \Omega_N\}.$$

Figures 4.10, 4.11, and 4.12, show comparisons of the two-stage BP test with the template bank of local sinusoids. The test signals were local sinusoids of lengths $N = 256, 384, 512$, and with frequencies $\omega = 2\pi \cdot N/4$. As we see, the two-stage BP test does surprisingly well at detecting these local sinusoids compared to a method that knows the signal form in advance.



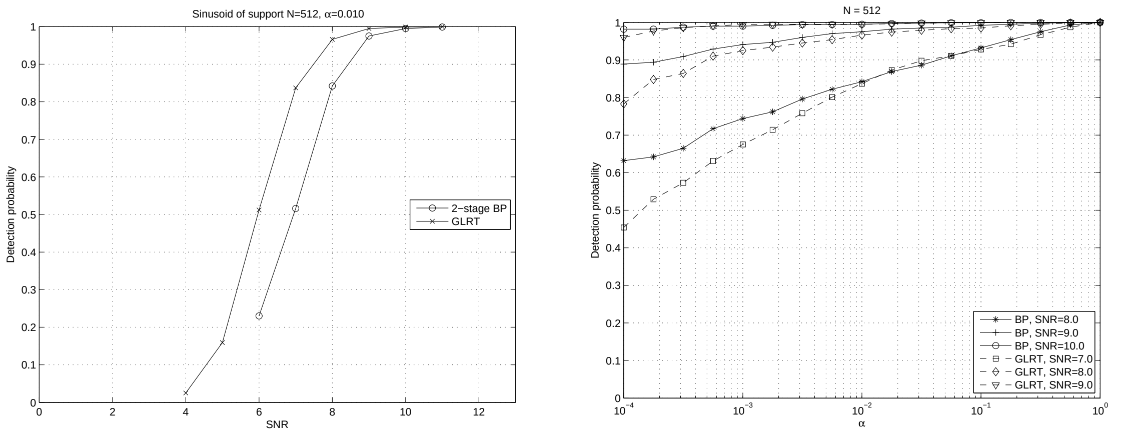
(a) (b)

Figure 4.10: Comparison of performance for GLRT and two-stage BP test for sinusoid of support $N = 256$. (a) Detection probability as a function of SNR. The probability of Type I error is fixed at 1%. (b) Plots of ROC for different values of SNR



(a) (b)

Figure 4.11: Comparison of performance for GLRT and two-stage BP test for sinusoid of support $N = 384$. (a) Detection probability as a function of SNR. The probability of Type I error is fixed at 1%. (b) Plots of ROC for different values of SNR



(a) (b)

Figure 4.12: Comparison of performance for GLRT and two-stage BP test for sinusoid of support $N = 512$. (a) Detection probability as a function of SNR. The probability of Type I error is fixed at 1%. (b) Plots of ROC for different values of SNR

4.6.7.4 Comparison of two-stage BP test with matched filtering

Here we show a comparison of the performance of the two-stage BP test with a GLRT, or matched filtering, for simulated gravitational waves of binary black hole coalescences.

Three separate template banks were designed for detecting the signals: BANK-1 for signal (i), BANK-2 for signal (ii), and BANK-3 for signal (iii).

- BANK-1: Templates with $m_1, m_2 \in [20.5, 30.4]$, $a \in [0.18, 0.98]$. Discretization spacing for the masses was $\Delta m = 0.9$, and for the spin parameter $\Delta a = 0.06$. This gave a bank of 1014 templates.

- BANK-2: Templates with $m_1, m_2 \in [13.5, 20.4]$, $a \in [0.18, 0.98]$. Discretization spacing for the masses was $\Delta m = 0.3$, and for the spin parameter $\Delta a = 0.16$. This gave a bank of 1800 templates.

- BANK-3: Templates with $m_1, m_2 \in [9.5, 13.4]$, $a \in [0.18, 0.98]$. Discretization spacing for the masses was $\Delta m = 0.12$, and for the spin parameter $\Delta a = 0.4$. This gave a bank of 1800 templates.

Every possible integer time translation of each template was correlated with the data stream as described in Section 4.2.

Comment about the "effective dimensionality" of the template banks: Before we proceed to the comparison, it is important to point out that the templates within each bank are highly correlated. Each bank searches over a very small subset in the space of chirps, and comparing the performance of matched filtering with that of two-stage BP must be considered unfair. The latter method is designed to search over a broad set of chirps while the first is targeted at a small space of parameterized functions.

To attempt to reveal something about the "effective dimensionality" of the template banks, we will compare them with a template bank designed for searching for sinusoids. Consider the following data:

$$y_k = \lambda s_k + z_k, \quad k = 0, \ldots, N-1,$$

where $z = (z_k)$ is a sequence of i.i.d. $N(0,1)$ random variables, $\lambda \in \mathbb{R}$, and the unknown signal $s = (s_k)$ belongs to a class of signals, $\mathcal{F}$. We also assume the support of s_k being known and approximately equal to N. The goal is to test, based on $y = (y_k)$,

$$H_0 : \lambda = 0 \quad \text{vs.} \quad H_1 : \lambda \neq 0.$$

We consider the case when the data length is $N = 512$ and compare two possible sets of functions $\mathcal{F}$:

(i) $\mathcal{F}_1 =$ the set of normalized templates in BANK-1.

(ii) $\mathcal{F}_2 = \{f : f(t) = \cos(\omega t + \phi),\ t = k/N,\ k = 0, \ldots, N-1,\ \omega \in \Omega_{512},\ \phi \in \mathbb{R}\}$.

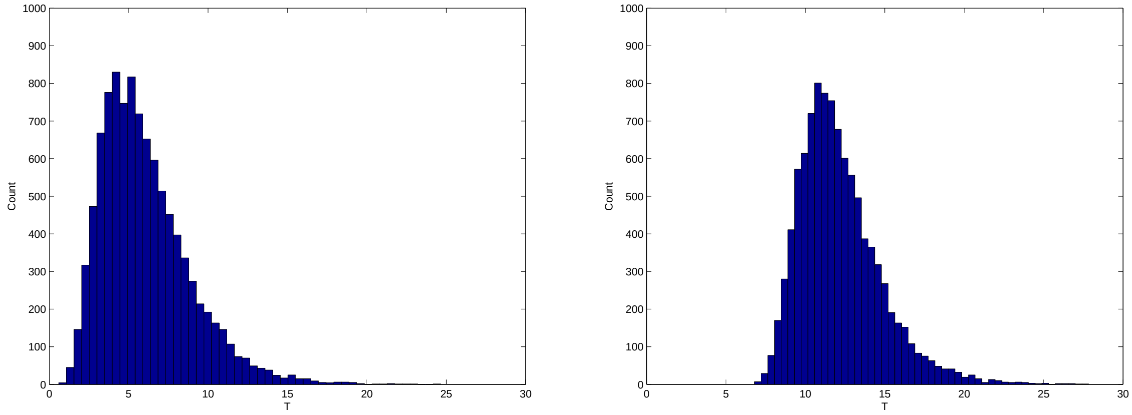
For the comparison we will use the GLRT statistic

$$T = \max_{f \in \mathcal{F}} \frac{|\langle y, f \rangle|^2}{\|f\|^2},$$

and reject H_0 for large values of T. The norms of the test functions in case (ii) are independent of the phase offset ϕ, making the GLRT statistic equivalent to

$$T = \max_{\omega \in \Omega_{512}} \frac{|\langle y, e^{\imath \omega t} \rangle|^2}{N/2}.$$

Figure 4.13 shows histograms of T under H_0 for the two choices of $\mathcal{F}$. Each histogram is based on 10,000 samples of the data vector $y = (y_k)$ under H_0. The typical values of T



(a) (b)

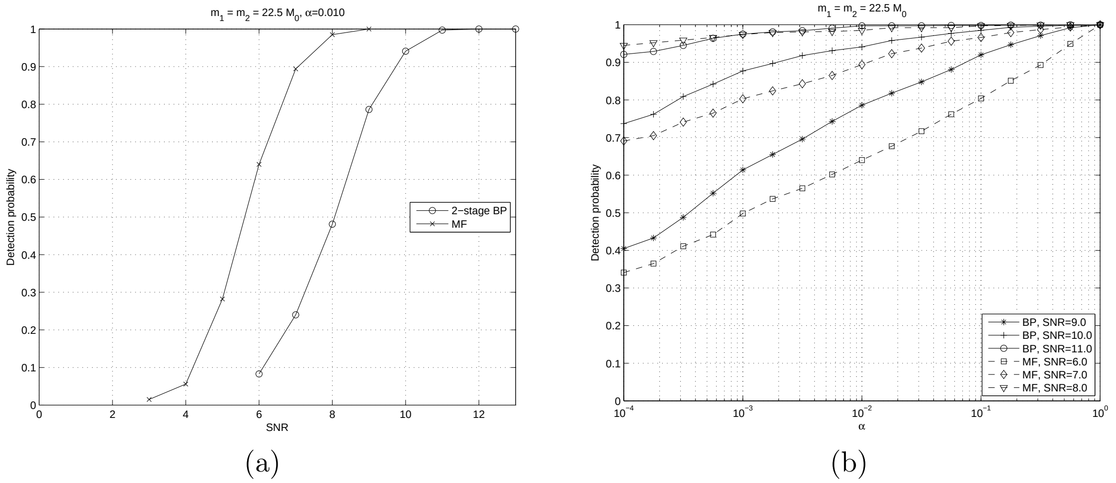
Figure 4.13: Comparison of histograms under H_0 for GLRT tests for different set of templates. Each histogram was generated using 10,000 realizations of white noise. (i) $\mathcal{F}$ is the set of templates in BANK-1, (ii) $\mathcal{F}$ is a set of sinusoids.

for case (i) are considerably smaller than for (ii); for example, the ratio of the mean of T for (ii) versus (i) is about 2. To control the tests at the same significance level, the threshold in case of (ii) needs to be higher. The difference lies in that dimensionality of $\mathcal{F}$. Even though the number of templates in BANK-1 is $|\mathcal{F}_1| \approx 1000$, and greater than the approximately 250 orthogonal sinusoids in case (ii), these test functions are highly correlated. Note that $\mathcal{F}_2$ is a small subset of the class of functions the BP test is designed to find. Therefore, we expect the performance of the GLRT with $\mathcal{F} = \mathcal{F}_1$ to be considerably better than that of the BP test when searching for signals restricted to BANK-1. The same holds for BANK-2 and BANK-3.

Figures 4.14, 4.15, and 4.16, show comparisons of performance for the two detection methods. The position of the signals was always placed around index 10,000 in the data. The SNR needs to be about 40% higher for the two-stage BP to achieve the same detection rate as matched filtering. It is important to point out that the exact same two-stage BP test was used for detecting all of the three signals, while a suitable template bank was chosen for each signal which makes the comparison ever more unfair.

4.6.7.5 Computational cost

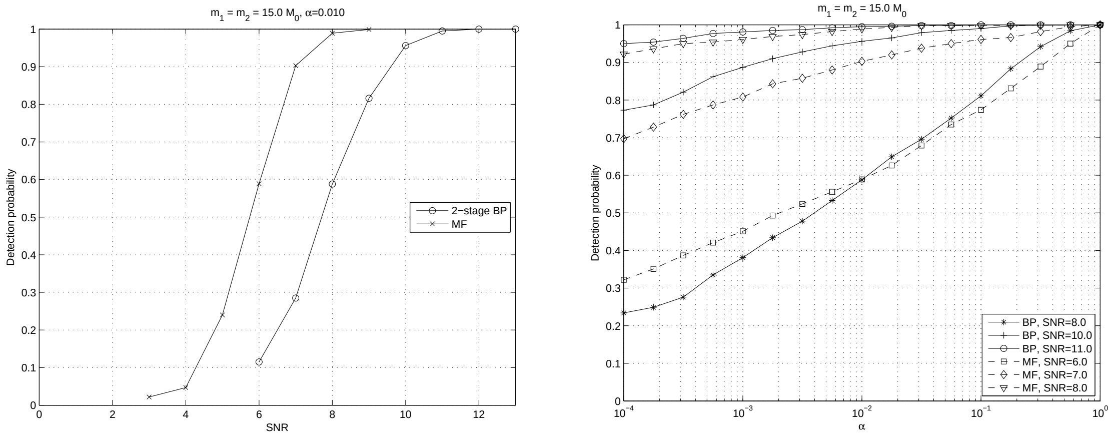
Performing the two-stage BP test for realization, i.e., block of length $M = 2^{16} = 65,536$, took about 20 seconds on a single processor on a 3.0GHz Mac Pro machine. For our sampling



(a) (b)

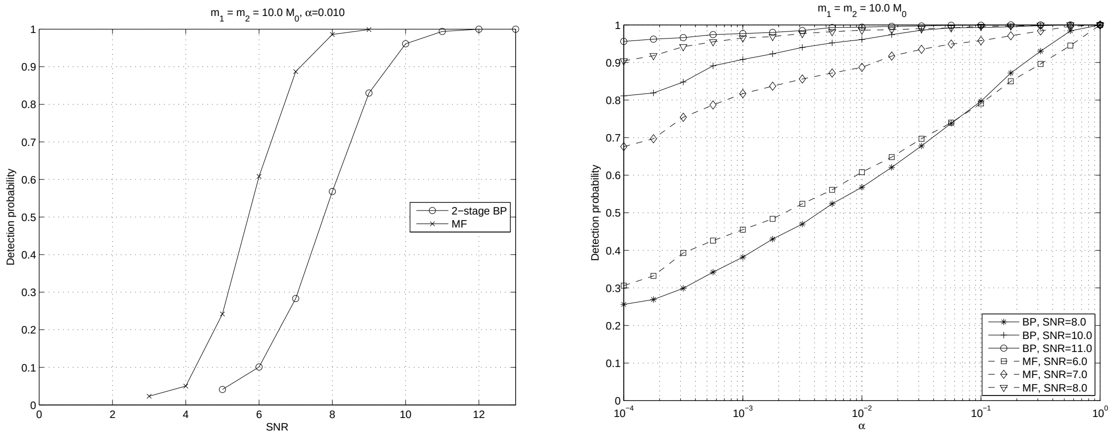
Figure 4.14: Comparison of performance for matched filtering and two-stage BP test for the BBH coalescences with total mass $M = m_1 + m_2 = 45M_{\odot}$. The set of templates is BANK-1. (a) Detection probability as a function of SNR. The probability of Type I error is fixed at 1%. (b) Plots of ROC for different values of SNR

model, the original data was sampled at 16kHz and downsampled by factor 8. Therefore, the original data segment would have been of length $2^{19} = 524,288$, which corresponds to 32 seconds. This indicates that this procedure has the potential of being applied to data at these sampling rates in real-time on a single processor.



(a) (b)

Figure 4.15: Comparison of performance for matched filtering and two-stage BP test for the BBH coalescences with total mass $M = m_1 + m_2 = 30M_\odot$. The set of templates is BANK-2. (a) Detection probability as a function of SNR. The probability of Type I error is fixed at 1%. (b) Plots of ROC for different values of SNR



(a) (b)

Figure 4.16: Comparison of performance for matched filtering and two-stage BP test for the BBH coalescences with total mass $M = m_1 + m_2 = 20M_\odot$. The set of templates is BANK-3. (a) Detection probability as a function of SNR. The probability of Type I error is fixed at 1%. (b) Plots of ROC for different values of SNR

Chapter 5

Estimation of Chirps by Chirplet Path Pursuit

This chapter considers the problem of estimating chirps from noisy data. Suppose we have noisy sampled data

$$y_k = f_k + z_k, \qquad k = 0, \dots, N-1; \tag{5.1}$$

where the unknown vector $f = (f_k)$ consists of sampled values of an object of interest $f(t), t \in [0,1]$, belonging to a class of functions $\mathcal{F}$. We assume uniform sampling such that $f_k = f(k/N)$, $k = 0, \dots, N-1$. The vector $z = (z_k)$ is a zero-mean random sequence with a known distribution, but not necessarily i.i.d. entries. In our setup, the set of signals $\mathcal{F}$ is a *nonparametric class of chirps.*

Based on the observation $y = (y_k)$, we wish to recover f the best as we can. To measure the performance of an estimator $\hat{f} = (\hat{f}_k)$ quantitatively, we could use the popular mean-squared error

$$MSE(f, \hat{f}) = E\left(\frac{1}{N}\sum_k (f_k - \hat{f}_k)^2)\right).$$

Our new estimation procedure relies on similar ideas and methodology we used for the BP test; i.e., chaining of local correlations of the data, using multiscale chirplets and the chirplet graph. This allows us build a rapidly computable and flexible estimator. Later in Chapter 7, we will show that our estimation procedures have theoretical optimality properties over a rich nonparametric class of chirps.

We finish this chapter by demonstrating the method by numerical experiments using simulated data and academic signals.

5.1 Motivation by Gaussian Model Selection

Consider data of the form as in (5.1) and where z is a vector of zero-mean i.i.d. Gaussians with variance σ^2. Assume we model the objects of interest $\mathcal{F}$ by a collection $\{\mathcal{F}_m, m \in \mathcal{M}\}$ of finite-dimensional linear spaces, where d_m is the dimension of $\mathcal{F}_m$. As the dimension d_m increases, we can say that the complexity of the models increases.

For each model m we consider the least-squares estimator $\hat{f}_m$, i.e., the solution to

$$\min_{\tilde{f}_m \in \mathcal{F}_m} \|y - \tilde{f}_m\|^2.$$

Since $\mathcal{F}_m$ is a linear space, this minimizer is equal to the projection of the data y onto $\mathcal{F}_m$. Denote the orthogonal operator for this linear projection by P_m. Then the quality or risk of the estimator based on model m, as measured by the mean-squared error, is

$$\begin{aligned} E\|f - \hat{f}_m\|^2 &= E\|f - P_m y\|^2 = E\|f - P_m f - P_m z\|^2 \\ &= \|f - P_m f\|^2 - 2E\left[Re(\langle f - P_m f, z\rangle)\right] + E\|P_m z\|^2 \\ &= \|f - P_m f\|^2 + \sigma^2 \cdot d_m, \end{aligned}$$

since $E\left[Re(\langle f - P_m f, z\rangle)\right] = 0$ and

$$E\|P_m z\|^2 = E z^T \underbrace{P_m^T P_m}_{=P_m} z = \sigma^2 \mathrm{trace}(P_m) = \sigma^2 \cdot d_m.$$

An ideal estimation procedure would be the one that minimizes this risk. We see that the error is a sum of two terms: the *squared bias*, $\|f - P_m f\|^2$, and the *variance*, $\sigma^2 \cdot d_m$. Finding the best model is therefore a search for the best trade-off between the fit to the signal and the complexity of the model. This is obviously beyond reach, since the bias term depends on the unknown signal and is therefore not available to us. Instead we seek a data-driven model selection procedure $\hat{m}$ that is close to the ideal risk

$$\inf_{m \in \mathcal{M}} E\|f - \hat{f}_m\|^2.$$

This is a well-studied problem in the literature of *Gaussian Model Selection* (see for example [18] and references therein). Most of the proposed methods for solving these problems fall under a class of *penalized approaches*; we seek to find the model $\hat{m}$ which minimizes

$$\|y - \hat{f}_m\|^2 + \Lambda(m),$$

where $\Lambda(m)$ is a nonnegative function defined on $\mathcal{M}$ and somehow measures the complexity of the model m. For our setup, model selection via penalization corresponds to penalized maximum log-likelihood and have criteria have been used for decades. The most common approach is

$$\arg\min_{m \in \mathcal{M}} \|y - \hat{f}_m\|^2 + \lambda \cdot \sigma^2 \cdot d_m, \tag{5.2}$$

where the parameter λ is either constant or depends on the sample size N. As we let λ increase, this procedure prefers "simple" models of low dimensionality; as λ gets closer to 0 we approach the method of maximum log-likelihood, inviting the risk of overfitting the data. Popular procedures such as Mallows' C_p , AIC , BIC, and RIC [3, 40, 59, 71] are all of the form (5.2) for different values of λ. Other similar approaches can be found in [13, 14, 17]. Solving (5.2) is in general NP-hard since it requires an exhaustive search over all the models. Therefore, unless the models in $\mathcal{M}$ all have a very special structure (for example, canonical) or if there are few models to consider, these procedures become virtually impossible to apply in practice.

Consider the Gaussian linear regression problems of (i) estimating the parameter $\beta \in \mathbb{R}^p$ and (ii) estimating $X\beta$, from the linear model

$$y = X\beta + z,$$

where $y \in \mathbb{R}^N$ is a vector of observations, X is an $N \times p$ predictor matrix, and $z \sim N(0, \sigma^2 I_N)$. Then the model selection procedure (5.2) can be written as

$$\arg\min_{\tilde{\beta} \in \mathbb{R}^p} \|y - X\tilde{\beta}\|^2 + \lambda \cdot \sigma^2 \cdot \|\tilde{\beta}\|_{\ell_0}, \tag{5.3}$$

where $\|\tilde{\beta}\|_{\ell_0} := \#\{k : \tilde{\beta} \neq 0\}$. As before, this problem is in general NP-hard unless X has a very special structure [1]. To overcome the computational difficulties, people have proposed to relax the ℓ_0 norm to the ℓ_1-norm $\|\tilde{\beta}\|_{\ell_1}$. This is done, for example, in the lasso [81]; see

[1] An example where this problem can be solved in practice is when X has orthogonal column vectors. This is discussed in Section 5.6.3.

also the closely related Basis Pursuit [27] and [69]. This replaces (5.3) with a linear program which can be solved in a rapid fashion thanks to progress in the field of convex optimization (see, for example, [19]).

However, although these methods seem to work well in practice, not much is known about their theoretical performance. Recently, Candés and Tao [24] introduced the Dantzig Selector for estimating β in (5.3) in the case when p, the number of explanatory variables, can be much larger than the number of observations N. This method is also based on ℓ_1-regularization and is rapidly computable using linear programming. But unlike previous methods, they also show that for design matrices X obeying a general property called a *uniform uncertainty principle* and β sufficiently sparse, the risk of their estimator comes within a factor $\log p$ of the ideal mean-squared error one would get when supplied with the information of which entries in β are nonzero, and which are above the noise level. This provides a practical and provably optimal estimation procedure.

Our estimation procedure will be based directly on (5.2). Thanks to the structure of the models we use to fit the chirps, we can solve the optimization problem exactly in a rapid fashion. Besides being practical, the estimation procedure has also very good theoretical performance (see Chapter 7). Next section describes the estimation procedure in detail.

5.2 The Best Path Estimator

We will make a distinction between real-valued and complex-valued data. Assume we have a chirplet graph $\mathcal{G}$. Let W be a chirplet path in the graph and $\{c_v\}$ the collection of chirplets on the path. Then in the case of complex-valued data we will consider estimators which are functions of the form

$$\tilde{f} = \sum_{v \in W} \alpha_v c_v, \tag{5.4}$$

and in the case of real-valued data we will consider estimators of the form

$$\tilde{f} = \sum_{v \in W} \frac{1}{2}(\alpha_v c_v + \alpha_v^* c_v^*), \tag{5.5}$$

where $\{\alpha_v\}$ is a set of complex scalars. Denote this class of functions by $\mathcal{C}$. Define the complexity functional

$$K(\tilde{f}, f) = \|\tilde{f} - f\|_2^2 + \Lambda(\tilde{f}) \tag{5.6}$$

where

$$\Lambda(\tilde{f}) = \lambda^2 N(\tilde{f}), \quad N(\tilde{f}) := |W|, \quad \lambda \in \mathbb{R}^+;$$

i.e., the complexity functional is a measure of trade-off between the fit between $\tilde{f}$ to the data f versus the number of terms, $N(\tilde{f})$, in $\tilde{f}$. As the parameter λ increases, the functional prefers functions with fewer terms. In Chapter 7, where we study theoretical properties of this estimator, we will take

$$\lambda^2 = \eta^2 \cdot (1 + \sqrt{2 \log M_N})^2,$$

for some fixed $\eta > 8$, where M_N is the number of chirplets in the graph. But for now the reader should think about λ as being a fixed positive number, chosen by the user of the estimator.

Given data $y = f + z$, our estimator $\hat{f}$ is the minimizer of the empirical complexity $K(\tilde{f}, y)$, or

$$\hat{f} = \arg\min_{\tilde{f} \in \mathcal{C}} K(\tilde{f}, y). \tag{5.7}$$

We call this estimator the *Best Path Estimator.*

5.2.1 Estimator for complex-valued chirps

Consider a fixed chirplet path W with the chirplets $\{c_v : v \in W\}$ normalized such that $\|c_v\| = 1$, i.e., with $|c_v(t)| = 1/\sqrt{|I_v|}$ (here and below the norm $\|\cdot\|$ stands for the ℓ_2-norm). For functions $\tilde{f}$ of the form (5.4), we have

$$\min_{(\alpha_v)} \|y - \tilde{f}\|^2 = \|y\|^2 - \sum_{v \in W} |\langle y, c_v \rangle|^2,$$

for any data sequence y (see Section 6.2.3). This gives

$$K(\tilde{f}, y) \geq \|y\|^2 - \sum_{v \in W} |\langle y, c_v \rangle|^2 + \Lambda(\tilde{f}),$$

with equality when the coefficients α_v in (5.4) satisfy

$$\arg \alpha_v = \arg \langle y, c_v \rangle, \tag{5.8}$$

and

$$|\alpha_v| = |\langle y, c_v \rangle|, \tag{5.9}$$

as is shown in Section 6.2.3. Thus,

$$\begin{aligned} \min_{\tilde{f} \in \mathcal{C}} K(\tilde{f}, y) &= \|y\|^2 - \left[\max_W \sum_{v \in W} |\langle y, c_v \rangle|^2 - \lambda^2 |W| \right] \\ &= \|y\|^2 - \max_W \sum_{v \in W} C_\lambda(v \mid y), \end{aligned}$$

where we define the *chirplet cost* given the data y, for the chirplet indexed with v, by

$$C_\lambda(v \mid y) = |\langle y, c_v \rangle|^2 - \lambda^2.$$

Therefore, to minimize the complexity functional we can equivalently solve the optimization problem

$$\max_W \sum_{v \in W} C_\lambda(v \mid y). \tag{5.10}$$

Once we have the optimal chirplet path W^*, our Best Path Estimator is

$$\hat{f} = \sum_{v \in W^*} \hat{\alpha}_v c_v,$$

where the $\hat{\alpha}_v$ is determined by (5.8) and (5.9). If $\{I_v : v \in W\}$ is the partition of the time axis for W^*, this would give us a piecewise constant estimate of the amplitude, or envelope, of the chirp, i.e.,

$$\hat{A}(t) = \sum_{v \in W^*} |\hat{\alpha}_v| \cdot 1_{I_v}(t) / \sqrt{|I_v|},$$

since the chirplets are normalized such that $|c_v(t)| = 1/\sqrt{|I_v|}$. Also, the phase and the instantaneous frequency of the optimal chirplet path could give us estimates for the phase and first derivate of the phase of the unknown chirp.

From our discussion in Section 3.2 we know that we can solve (5.10) rapidly using the shortest path algorithm.

5.2.2 Estimator for real-valued chirps

For real-valued data we do not enjoy the same "luck" of being able to minimize the complexity functional as simply as for the complex-valued case. However, we can still solve it exactly in a similar way if we define the chirplet costs in the graph slightly differently.

Consider a fixed chirplet path W with the chirplets $\{c_v : v \in W\}$ and let $\mathcal{P} = \{I_v : v \in W\}$ be the corresponding partition of the time axis. We can write the function in (5.5) as

$$\tilde{f}(t) = \sum_{v \in W} \rho_v \cdot \cos(\theta_v(t) + \theta_{0,v}),$$

where $\theta_v(t)$ is the phase of the chirplet c_v, and $\rho_v, \theta_{0,v} \in \mathbb{R}$. As for the complex-valued case, we have

$$\min_{(\alpha_v)} \|y - \tilde{f}\|^2 = \sum_{v \in W} \min_{\alpha_v} \|y - \tilde{f}\|_{I_v}^2 = \sum_{v \in W} \min_{\rho_v, \theta_{0,v}} \|y - \tilde{f}\|_{I_v}^2,$$

so we can handle each interval I_v separately. Think about $\theta_{0,v}$ as fixed for now. Then we have,

$$\begin{aligned} \min_{\rho_v} \|y - \tilde{f}\|_{I_v}^2 &= \min_{\rho_v} \|y\|_{I_v}^2 - 2\rho_v \langle y, \cos(\theta_v + \theta_{0,v}) \rangle_{I_v} + \rho_v \| \cos(\theta_v + \theta_{0,v})\|_{I_v}^2 \\ &= \|y\|_{I_v}^2 - \left(\frac{\langle y, \cos(\theta_v + \theta_{0,v}) \rangle_{I_v}}{\| \cos(\theta_v + \theta_{0,v})\|_{I_v}} \right)^2. \end{aligned}$$

Define the quantity

$$D(v \mid y) = \max_{\theta_{0,v}} \left(\frac{\langle y, \cos(\theta_v + \theta_{0,v}) \rangle_{I_v}}{\| \cos(\theta_v + \theta_{0,v})\|_{I_v}} \right)^2.$$

There is an analytic formula for calculating $D(v \mid y)$ and the argument $\theta_{0,v}$ which maximizes it. It uses as input the chirplet coefficient $\langle y, c_v \rangle$; see Section 3.3.5 and Appendix A for further discussion and details. The important point here is that the complexity of calculating $D(v \mid y)$ is of the same order as the computational complexity for calculating the chirplet coefficients.

Analogous to the complex-valued case, let's define the chirplet cost, given the data y, for the chirplet indexed with v, by

$$C_\lambda(v \mid y) = D(v \mid y) - \lambda^2.$$

Then, just as before, we can minimize the complexity functional by equivalently solving the optimization problem

$$\max_W \sum_{v \in W} C_\lambda(v \mid y).$$

Notice that except for the difference in the definition of the chirplet cost, the estimator has the exact same form for complex-valued and real-valued data.

5.2.3 Approximate complexity functional for real-valued chirps

Consider the complexity function for real-valued data. If we restrict ourselves to estimating chirps that are highly oscillating so their frequency support is away from frequency zero, we can choose the chirplet graph to include only chirplets at large frequencies. In that case we have an approximation for calculating $D(v \mid y)$. Note that if the $\theta_v'(t)$ is big enough for every $t \in I_v$, we have

$$\|\cos(\theta_v + \theta_{0,v})\|_{I_v}^2 \approx |I_v|/2.$$

Then

$$\begin{aligned} D(v \mid y) &\approx \max_{\theta_{0,v}} \frac{2}{|I_v|}(\langle y, \cos(\theta_v + \theta_{0,v})\rangle_{I_v})^2 = \frac{2}{|I_v|}|\langle y, \exp(i\theta_v)\rangle_{I_v}|^2 \\ &= 2|\langle y, c_v\rangle|^2, \end{aligned}$$

and the chirplet cost would be

$$C_\lambda(v \mid y) = 2|\langle y, c_v\rangle|^2 - \lambda^2.$$

5.2.4 Computing the estimator

The computation of the BP estimator consists of three independent steps:

1. Calculation of the chirplet costs $C_\lambda(v \mid y)$.

2. Minimizing the empirical complexity functional; i.e., solving (5.10).

3. Building the estimate based on the solution to step 2.

The computational cost of the last step is almost negligible compared to the other two steps since we can reuse the calculations of the chirplet coefficients to get the value of the coefficients (α_v) in the the linear combinations (5.4), (5.5), for the best chirplet path from step 2. In our discussion of the computational complexity of the BP test, we already argued that the cost of step 1 is greater than the cost of step 2. In this case, the cost of step 2 is even less than for the BP algorithm since we can use the simpler Shortest Path algorithm which is described in Section 3.2. The computational cost of the BP estimator is essentially equal to the cost of the chirplet transform.

5.2.5 A remark on imposing continuity of the phase

Note that the optimization problem does not require the phase to be continuous and on each time interval we are estimating the phase offset locally. To force the phase to be continuous, one could include a phase offset parameter in the chirplets and consider a *phaselet graph* instead, where we would look for (near) continuous paths of piecewise quadratic phase functions in a time-phase diagram, with regularity constraints to determine the topology of the graph. For this bigger graph, the computational complexity would increase. Based on our discussion in Appendix B on imposing phase continuity for the BP test, it is not immediately clear how much it would improve the estimation in situations where the noise is very strong. Besides that, the way we model the fit to the chirps introduces a discontinuity in the phase due to the piecewise constant fit to the amplitude. We could get around that by adding an extra step of estimating the amplitude of the chirp globally after we have an estimate of the phase. See Section 5.5 for further discussion on global amplitude fitting.

5.3 Choosing the Roughness Parameter λ in the Complexity Functional

The complexity functional our estimator is based on involves a parameter λ which controls the trade-off between the complexity of the estimator and the goodness-of-fit to the data. Although we know which value to use for attaining theoretical bounds, it does not necessarily mean it is the best choice for practical applications.

A popular way to choose regularization parameters of this sort is to use a graphical tool called the *L-curve* (see for example [44, 45]). To explain it, we will consider the well-known Tikhonov regularization scheme for solving ill-posed problems. Let K be a linear operator and assume we observe the data y of the form

$$y = Kf + z,$$

where z is white noise and f is the object we want to recover. Then the estimate of f using the Tikhonov regularization scheme would be the solution to the following optimization problem

$$\min_f \|Kf - y\|_2^2 + \lambda^2 \omega(f)^2, \tag{5.11}$$

where λ is a specified regularization parameter and $\omega(f)$ measures some kind of "smoothness" of the object f. In the continuous case, where f is a real-valued function on $\mathbb{R}$, one possible choice could be based on the second derivative of f, or $\omega(f) = \|f''\|_2$.

The L-curve for the Tikhonov regularization problem would be the plot of the points $(\omega(\hat{f}), \|K\hat{f} - y\|_2)$ for all valid regularization parameters λ, where $\hat{f}$ is the solution to (5.11); i.e., the plot of the regularity of the estimate versus the residual norm. It turns out that this curve very often has an L-shaped appearance, hence the name. For small values of the residual norm, $\omega(\hat{f})$ tends to be big, and as it increases, $\omega(\hat{f})$ gets smaller. Therefore, the plot displays the compromise between the minimization of these two quantities. For regularization methods with a discrete regularization parameter λ the L-curve consists of discrete set of points (an example is the Truncated Singular Value Decomposition (TSVD) in the case of linear regression [44]).

The L-curve can also be used in our estimation problem to aid in choosing a good trade-off between the goodness-of-fit of the model to the data and the complexity of the model. In the case of the BP estimation procedure, the goodness of fit is the residual norm and the complexity of the model is the number of chirplets in the estimator. Due to the Best Path algorithm in Section 3.2 we have a fast way of calculating the best fit $\hat{f}_L$ to the data, where $\hat{f}_L$ minimizes $\|y - \hat{f}_L\|^2$ over all paths with L chirplets. This allows us to plot the residual sum of squares $\|y - \hat{f}_L\|_2$ against the model complexity L. We expect to see a “kink” in the plot where the goodness of fit starts to improve slower with increasing L. The estimation could then be based on the best chirplet path corresponding to a length L close to the kink.

To demonstrate this idea, we will show an example based on the data and configurations in Section 5.7 for the noise level $\sigma = 0.03$. Figure 5.1 shows a plot of the squared error $\|f - \hat{f}_L\|^2$ against L for one realization of the noise. Notice that the error has a minimum

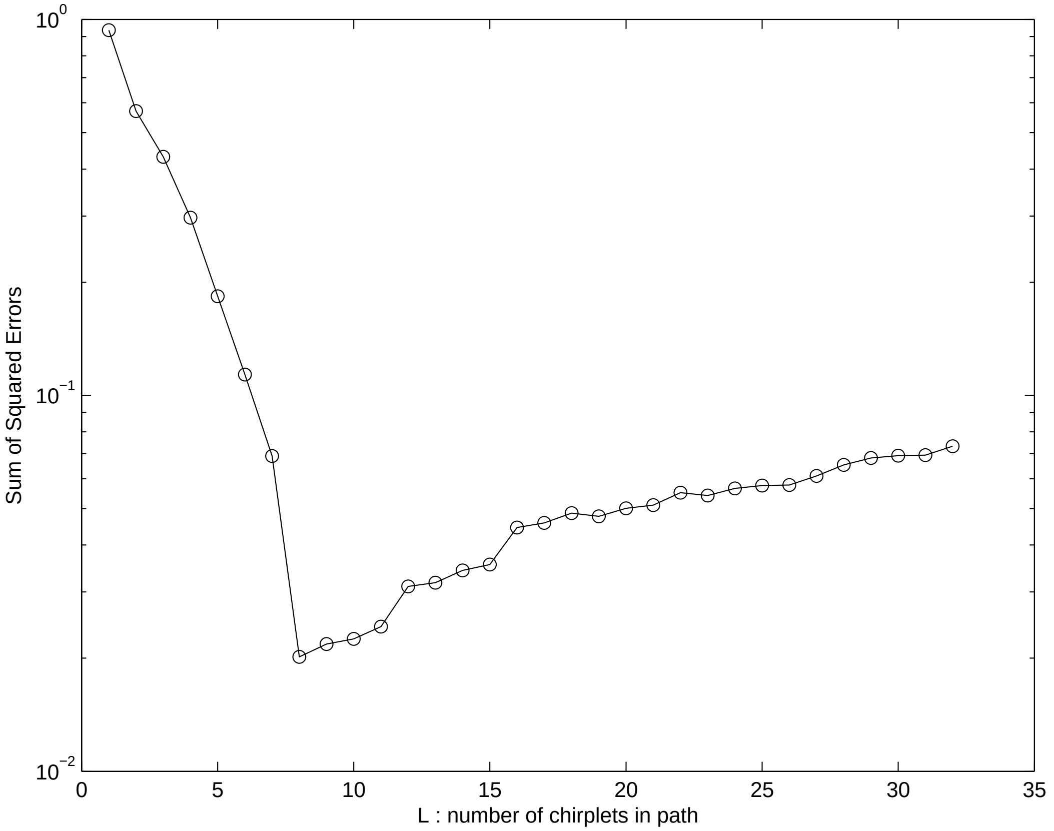


Figure 5.1: Plot of the sample squared error $\|f - \hat{f}_L\|^2$ for BP estimators of fixed number of chirplets L

for a chirplet path of length $L = 8$. Of course we could never draw such a plot in practice. Figure 5.2 shows a plot of $K_L := \|y - \hat{f}_L\|^2 + \lambda L$ versus the length L. Note that the complexity functional K from (5.6) is the minimum of this curve for the same choice of λ. In this case the complexity functional would choose the same model as the oracle that could plot the squared error. Finally, on Figure 5.3, we look at the plot of the residual sum of squares $\|y - \hat{f}_L\|^2$ against the number of chirplets in the best path. Notice the characteristic L-shape and the kink at $L = 8$, which for this realization of the noise corresponds to the model which minimizes the squared error.

5.4 Extensions

Because of the familiarity to the BP test, many of the extensions discussed in Chapter 3 could apply for the BP estimator. We could follow up on the idea of local fit of chirplets to the data and replace the chirplet costs in (5.10) by the costs proposed for colored noise and

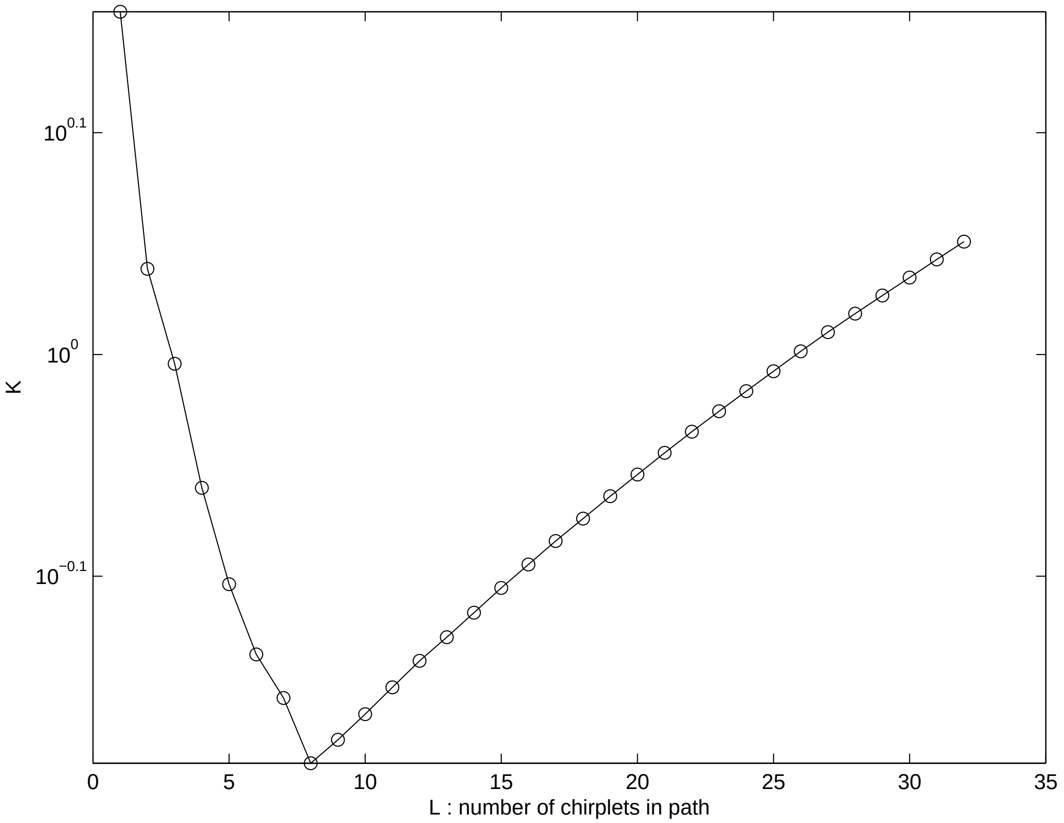


Figure 5.2: Plot of $K_L = \|y - \hat{f}_L\|^2 + \lambda L$ against the number of chirplets L in the best path

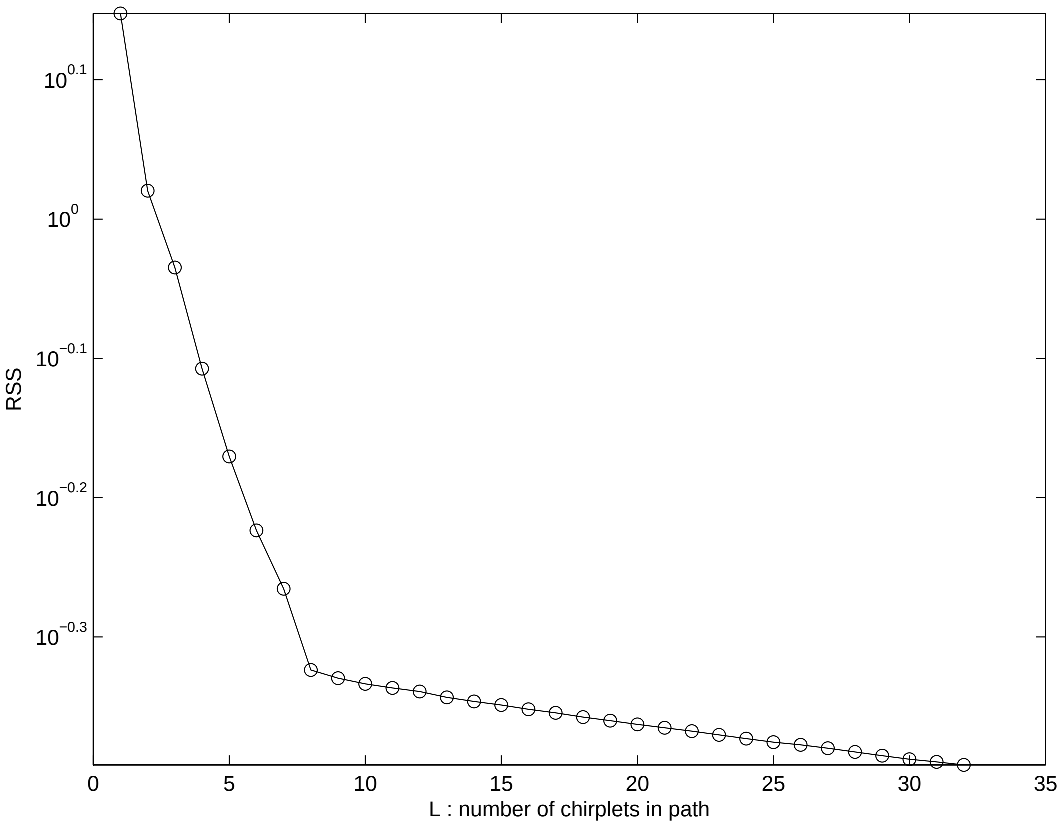


Figure 5.3: Plot of the residual sum of squares $\|y - \hat{f}_L\|^2$ for BP estimators of fixed number of chirplets L

varying amplitude as explained in Section 3.3.

5.5 Refining the Estimate of the Amplitude

Since the estimated chirp based on the Best Path is a linear combination of chirplets, the estimate could have infeasible discontinuities due to the piecewise constant estimate of the chirp's amplitude. This problem would not vanish if we instead used amplitude-modulated chirplets, giving a piecewise polynomial estimate of the amplitude. Therefore, in cases where we know the amplitude of the chirp is smooth and continuous, we propose splitting the estimation procedure into two steps to get a smoother, and hopefully better, fit to the amplitude. Suppose the unknown chirp is of the form $A(t)\exp(\imath\varphi(t))$. Then the procedure would be as follows:

Global amplitude estimation for the BP estimator:

1. Use the BP estimator to estimate the phase of the unknown chirp. This can be done based on the instantaneous frequency of the chirplet path that minimizes the complexity functional and the local maximum likelihood estimates of the phase offsets. Call the estimate of the phase $\hat{\varphi}_t$. (As discussed in 5.2.5, this phase estimate could be discontinuous).

2. Demodulate the data y using the estimate of the phase:

$$\tilde{y}_t = y_t \cdot \exp\left(-\imath\hat{\varphi}_t\right) = A(t)\exp\left(\imath(\varphi_t - \hat{\varphi}_t)\right) + \tilde{z}_t,$$

   where $\tilde{z}_t := z_t \cdot \exp\left(-\imath\hat{\varphi}_t\right) = \tilde{z}_t^0 + \imath\tilde{z}_t^1$.

3. Estimate the amplitude $a(t)$ from the demodulated data $\tilde{y}_t$ or $Re(\tilde{y}_t)$.

If $\hat{\varphi}_t$ is a good estimate of the phase, we would expect the real part of the demodulated data to satisfy the approximation

$$Re(\tilde{y}_t) \approx A(t) + \tilde{z}_t^0.$$

Assume (z_t) is a sequence of complex-valued white noise such that $z_t = z_t^0 + \imath z_t^1$ with (z_t^0) and (z_t^1) being independent sequences of i.i.d. $N(0,1)$ varables. Then $(\tilde{z}_t^0)$ is a sequence of i.i.d. $N(0,1)$ random variables. There are many methods at our disposal for estimating smooth functions from data contaminated with white noise. Which method to choose depends on the *a priori* information we have on the amplitude of the chirping signal. We will discuss two possible choices: thresholding in a Fourier basis and spline smoothing.

5.5.1 Estimating the amplitude using thresholding in the Fourier basis

One possibility would be to use Fourier approximations. $\{e_m(t) := e^{i2\pi mt}\}_{m\in\mathbb{Z}}$ is an orthonormal basis of $L_2[0,1]$, and we can decompose any function $f \in L_2[0,1]$ using its Fourier series

$$f(t) = \sum_{m=-\infty}^{\infty} \langle f, e_m\rangle e_m(t)$$

where $c_m(f) := \langle f, e_m\rangle = \int_0^1 f(t)e^{-i2\pi mt}dt$. The Fourier series defines a periodic extension of f for all $t \in \mathbb{R}$, and the decay of the Fourier coefficients $|c_m(f)|$ as m increases depends on the regularity of this extension. To prevent the extension from having singularities at $t = k$, $k \in \mathbb{Z}$, f needs to be compactly supported in $(0,1)$ or $f^{(l)}(0) = f^{(l)}(1)$, $l = 0, \ldots, l'$ for l' sufficiently large.

The linear approximation of $f \in L_2[0,1]$ by the sinusoids of the $M+1$ lowest frequencies, i.e., $e_m(t)$ for $m \in \{-M/2, \ldots, M/2\}$, is

$$f_M(t) = \sum_{m=-M/2}^{M/2} \langle f, e_m\rangle e_m(t),$$

with the approximation error

$$\|f - f_M\|^2_{L_2[0,1]} = \int_0^1 |f(t) - f_M(t)|^2 dt = \sum_{|m|>M/2} |c_m(f)|^2.$$

We can quantify this approximation error based on the regularity of f. To distinguish between the regularity of functions that are n times continuously differentiable, but not $n+1$ times, we consider the Sobolov differentiability [62]. Let $s > 0$ and define the space $\mathbf{W}^s(\mathbf{R})$ of functions $f \in L_2(\mathbf{R})$ that are s times differentiable in the sense of Sobolev, i.e., whose Fourier transform satisfies

$$\int_{-\infty}^{\infty} |\omega|^{2s}|\hat{f}(\omega)|^2 d\omega < \infty.$$

It can be verified that if $s > n + 1/2$, then f is n times continuously differentiable. Define $\mathbf{W}^s[0,1]$ as the space of functions $f \in L_2[0,1]$ that can be extended outside $[0,1]$ to functions $f \in \mathbf{W}^s(\mathbf{R})$. Then we have a classical approximation theorem (see, for example, [57]):

Theorem 1. *Let $f \in L_2[0,1]$ be compactly supported in $(0,1)$. Then $f \in \mathbf{W}^s[0,1]$ if and only if*

$$\sum_{M=1}^{\infty} M^{2s-1} \|f - f_M\|^2_{L_2[0,1]} < \infty,$$

which implies $\|f - f_M\|^2_{L_2[0,1]} = o(M^{-2s})$.

It is straightforward to construct an estimator based on Fourier approximations. Let the data be $y[k] = f[k] + z[k]$, $k = 0, \ldots, N-1$ where $(z_[k])$ is a sequence of i.i.d. standard normal random variables and $f[k] = f_c(k/N)$, where $f_c \in \mathbf{W}^s[0,1]$. Consider the discrete Fourier basis

$$\left\{ g_m[n] = \frac{1}{\sqrt{N}} \exp\left(\frac{i2\pi mn}{N}\right) \right\}_{0 \le m < N},$$

where for $m \ge N$ or $m < 0$, $g_m = g_{m'}$ where $m' = m \bmod N$. Then a Fourier estimator of $f[k]$, using the $M+1$ lowest frequencies would be

$$\hat{f}_M[k] = \sum_{m=-M/2}^{M/2} \langle y, g_m \rangle g_m[k].$$

If we know the regularity exponent s, the estimate of the approximation error from Theorem 1 can help us in choosing M. Since the estimator $\hat{f}_M$ is simply a projection of the data y onto a linear space of dimension $M+1$, we can use the decomposition of the MSE of the estimator into bias and variance term from Section 5.1:

$$E\|f - \hat{f}_M\|^2_{\ell^2} = \|f - f_M\|^2_{\ell^2} + (M+1)\sigma^2 \le C \cdot \frac{M^{-2s}}{N} + (M+1)\sigma^2,$$

where the last inequality is based on Theorem 1 and C is some constant. Assume that $\sigma^2 = 1$. Taking

$$M \sim N^{1/(2s+1)}$$

makes the bias and variance term of the same order and gives the best bound. But since we do not know the smoothness we would have to choose the parameter M adaptively; for example, by adaptive thresholding where only the low frequency coefficients that exceed a threshold are kept and the other are set to zero.

Perhaps a more serious limitation of this method is that it is not suitable for cases when the signal is non-periodic or does not have its support entirely within the data under consideration. To deal with non-periodicity we review another well-known method for estimating smooth functions; curve estimation using smoothing splines.

5.5.2 Estimating the amplitude using spline smoothing

Most of the following presentation of spline smoothing is based on the monograph by Green and Silverman [42]. Suppose that $t_1, \ldots, t_n$ are points in an interval $[a, b]$ satisfying $a < t_1 < \ldots < t_n < b$. Suppose we have observations $y_1, \ldots, y_n$. Given any twice-differentiable function f on $[a, b]$, let $S(f; \alpha)$ be the penalized sum of squares

$$S(f;\alpha) := \sum_{k=1}^{n} (y_k - f(t_k))^2 + \alpha \int_a^b \left(f''(t)\right)^2 dt,$$

where α is a positive smoothing parameter. Note that this is a functional of the form described in Section 5.3. Here, the term $\int_a^b \left(f''(t)\right)^2 dt$ measures roughness of the function f. A function g defined on $[a, b]$ is a *cubic spline* if it satisfies two conditions:

1. g is a cubic polynomial on each of the intervals $(a, t_1), (t_1, t_2), \ldots, (t_n, b)$.

2. The first and second derivative of g are continuous on $[a, b]$.

A *natural cubic spline* is a cubic spline whose second and third derivatives are zero at a and b. It can be shown that for any given set of points $(t_1, z_1), \ldots, (t_n, z_n)$, with $t_1 < \cdots < t_n$, there exists a unique natural cubic spline interpolating them.

Interestingly enough, the minimizer of $S(f; \alpha)$ is necessarily a natural cubic spline with knots at the points t_k. To see why this holds, take any curve f and let g be a natural cubic spline with knots at the points t_k and $g(t_k) = f(t_k)$ for all k. The natural cubic spline interpolant has the property of being the unique minimizer of $\int (h'')^2$ among all C^2 curves h that interpolate the data $\{(t_k, f(t_k))\}$ (see [42]) and therefore $\int (g'')^2 \leq \int (f'')^2$. Since $\sum_{k=1}^{n} (y_k - f(t_k))^2 = \sum_{k=1}^{n} (y_k - g(t_k))^2$ we have

$$S(g; \alpha) \leq S(f; \alpha)$$

with equality only if $f = g$.

Let's switch to vector notation and write $\mathbf{y} = [y(t_1), \ldots, y(t_n)]$ and $\mathbf{f} = [f(t_1), \ldots, f(t_n)]$ where f is a natural cubic spline. As shown in [42], the roughness penalty term $\int (f'')^2$ can be written as $\mathbf{f^T K f}$ where K is a matrix only depending only on the knots $t_1, \ldots, t_n$. In vector notation we obtain

$$S(f; \alpha) = (\mathbf{y} - \mathbf{f})^{\mathbf{T}}(\mathbf{y} - \mathbf{f}) + \alpha \mathbf{f^T K f},$$

which has a unique minimum obtained at

$$\hat{\mathbf{f}} = (\mathbf{I} + \alpha \mathbf{K})^{-\mathbf{1}} \mathbf{y}. \tag{5.12}$$

An algorithm due to Reinsch [66] can be used to determine the smoothing spline in $O(n)$ arithmetic operations. It uses the fact that it is possible to set up a system of linear equations involving only band matrices to determine the second derivatives of the smoothing spline at the knots t_k.

Choosing the smoothing parameter α: We need an objective method for choosing the smoothing parameter α for our estimation procedure. One of the most popular choices is cross-validation, which in our case would be as follows:

1. Fix α.

2. For every $l \in \{1, \ldots, n\}$ take the observation (t_l, y_l) from the set of the data and use the remaining data to estimate the curve. The estimate is the minimizer $\hat{f}^{(-l)}(t; \alpha)$ of the complexity functional

$$S^{(-l)}(f; \alpha) := \sum_{k \neq l} (y_k - f(t_k))^2 + \alpha \int_a^b \left(f''(t)\right)^2 dt.$$

3. Calculate the cross-validation score function

$$CV(\alpha) = n^{-1} \sum_{k=1}^{n} (y_k - \hat{f}^{(-k)}(t_k; \alpha))^2.$$

4. Repeat steps 1 through 3 for different values of α and choose the value of α that minimizes $CV(\alpha)$ for the spline smoother.

The term $(y_k - \hat{f}^{(-k)}(t_k;\alpha))^2$ measures how well the estimator $\hat{f}^{(-l)}(t;\alpha)$ predicts y_k.

It appears that for every α we would need to solve n smoothing problems to calculate $CV(\alpha)$. This would add at least an extra order of complexity to the method of smoothing splines, but luckily this is not the case. If we write the matrix for the spline smoother in (5.12) as

$$\mathbf{A}(\alpha) = (\mathbf{I} + \alpha\mathbf{K})^{-\mathbf{1}},$$

it can be shown (see [42]) that

$$CV(\alpha) = n^{-1}\sum_{k=1}^{n}\left(\frac{y_k - \hat{f}(t_k)}{1 - A_{kk}(\alpha)}\right)^2, \tag{5.13}$$

where $\hat{f}$ is the spline smoother calculated from the full data set with the smoothing parameter α and $A_{kk}(\alpha)$ is the k-th diagonal element in $\mathbf{A}(\alpha)$. There is an algorithm due to Hutchinson and de Hoog [48] for finding the diagonal elements of the matrix $\mathbf{A}(\alpha)$ that runs in $O(n)$ operations. As a result, since the Reinsch algorithm finds the values $\hat{f}(t_k)$ in $O(n)$ operations, the cross-validation score $CV(\alpha)$ can be found in $O(n)$ operations for each α.

5.6 Estimation of Chirps Using Tight Frames of Multiscale Chirplets

Here we review Candès' estimation procedure for chirps which was introduced in [20] and was shown to be theoretically near-optimal over a wide range of chirps. The method is based on thresholding in the *best empirical tight frame of chirplets*, which we will abbreviate to TBCF. Based on noisy observations as in (5.1), the goal of the method is to adaptively select a frame from a library of frames, in which is best to recover the unknown signal. Before we give a precise description of the method, we will define the chirplet tight frames.

5.6.1 Tight frames of chirplets

In $\mathbb{C}^n$, a tight frame is a collection of vectors $\{g_k\}_{k\in\Gamma}$ with the property

$$\sum_{k\in\Gamma} |\langle f, g_k\rangle|^2 = A \cdot \|f\|^2,$$

for some constant $A > 0$. Let the vectors $\{g_k\}_{k\in\Gamma}$ be the columns of a matrix Φ. For $A = 1$, this property can be expressed using matrix notation as

$$\|\Phi^* f\|^2 = \|f\|^2.$$

This isometry property provides a reconstruction formula from the frame coefficients $\{\langle f, \varphi_k\rangle\}_{k\in\Gamma}$:

$$f = \sum_{k\in\Gamma} \langle f, g_k\rangle g_k$$

(see [30, 57].) Note that the vectors g_k may be linearly dependent and their number be greater than n, the dimension of the space. Note that for a tight frame $\Phi\Phi^* = I$.

The chirplet tight frames are made from a collection of vectors which are windowed chirplets and are restricted to balanced recursive dyadic partitions (BRDPs) (see Definition 1). Let I be a dyadic inteval, $I = [k2^{-j}, (k+1)2^{-j})$, $t_I = k2^{-j}$. Let $\rho(t)$ be a smooth cutoff function satisfying $\rho(t) = 0$ for $t < -1/2$, $\rho(t) = 1$ for $t > 1/2$, and $\rho(t)^2 + \rho(-t)^2 = 1$ for $|t| \leq 1$. Define the dyadic window $w_I^{\epsilon,\epsilon'}$, localized near I, such that

$$w_I^{\epsilon,\epsilon'} = \rho\left(\frac{t-t_I}{\epsilon}\right)\rho\left(\frac{t-t_I}{\epsilon'}\right)$$

is smooth. Define the cutoff parameter $\eta_I = (\epsilon_I, \epsilon_I')$. Assume we are given a BRDP $\mathcal{P}$. Then for every ordered pair (I, I') of adjacent intervals, we choose the cutoff

$$\epsilon' = \min(|I|, |I'|).$$

Restricting us to BRPDs requires us to only consider four different windows per interval I. Note from the properties of the cutoff function ρ, the collection of windows $(w_I^{\eta_I})_{I\in\mathcal{P}}$ obeys

$$\sum_{I\in\mathcal{P}} |w_I^{\eta_I}(t)|^2 = 1. \tag{5.14}$$

Then the dictionary of multiscale chirplets is made out of functions

$$g_{a,I,n}(t) = \frac{1}{\sqrt{2|I|}} \cdot w_I^{\eta_I}(t) e^{\imath(a_I t^2/2 + \pi n t/|I|)},$$

for all dyadic intervals I, cutoffs η_I, and sequences $a = (a_I)$. The parameter a_I, or the *slope* of the chirplet, is restricted to a discrete set of values on each interval I. One possible choice for slope discretization is

$$a_I = \ell \cdot 2^j \cdot \delta_j, \quad \ell \in \mathbb{Z}, \quad \text{and } |a_I| \leq B,$$

for some fixed $B > 0$ possibly depending on the smoothness of the phase of the chirps under consideration. This choice was used in [20] for establishing theoretical results. Note the frequency offsets of these chirplets are of the form $b_I = \pi n/|I|$, for $n \in \mathbb{Z}$.

If we suppose the signal is of length N, the number of distinct elements in the library with this choice of discretization is

$$M_N = O\left(N^{4/3}\right),$$

and the number of distinct tight frames is exponential in N.

Fix a sequence of slopes $(a_I)_{I\in\mathcal{P}}$ such that on each dyadic interval, every chirplet has the same slope. Then the family of chirplets $(g_{a,I,n})_{I\in\mathcal{P},n\in Z}$ is a tight frame; for any signal f we have

$$\sum_n |\langle f, g_{a,I,n}\rangle|^2 = \int |f(t)|^2 |w_I^{\eta_I}(t)|^2 dt,$$

and from (5.14) we have

$$\sum_{I\in\mathcal{P}} \sum_n |\langle f, g_{a,I,n}\rangle|^2 = \|f\|^2.$$

Therefore we have the reconstruction formula

$$f = \sum_{I\in\mathcal{P}} \sum_n \langle f, g_{a,I,n}\rangle g_{a,I,n}.$$

5.6.2 Thresholding in a library of tight frames

Suppose we have observations $y_k = f_k + z_k$, $k = 1, \ldots, N$, where (f_k) is a signal and (z_k) is i.i.d. Gaussian white noise. Let $\mathcal{L} = \{\Phi_1, \Phi_2, \ldots, \Phi_{B_N}\}$ be a library of tight frames of chirplets. For a fixed $\Phi \in \mathcal{L}$ and some function h, let $\theta_k[\Phi] = \langle h, \varphi_k\rangle$ and define the entropy functional

$$\mathcal{E}_\Lambda(h, \Phi) = \sum_k \min(|\theta_k[\Phi]|^2, \Lambda), \tag{5.15}$$

for some parameters Λ. Let $\hat{\Phi}$ be the empirical best frame according to this entropy:

$$\hat{\Phi} = \arg\min_{\Phi \in \mathcal{L}} \mathcal{E}_\Lambda(y, \Phi),$$

and define the hard-thresholding $\eta_\tau(y) = y 1_{\{|y|>\tau\}}$. The estimator is then constructed as follows:

1. Find the empirical best frame $\hat{\Phi}$ according to the entropy $\mathcal{E}_\Lambda$.
2. Apply hard-thresholding in the empirical best frame with threshold τ:

$$\hat{y}_{I,n} = \eta_\tau(y_k[\hat{\Phi}]).$$

3. Reconstruct using the empirical best frame $\hat{\Phi} = \{g_{a,I,n}\}$:

$$\hat{s} = \sum_{I \in \mathcal{P}, n \in \mathbb{Z}} \hat{y}_{I,n} g_{a,I,n}.$$

For theoretical purposes in [20], upper bounds on the performance of the estimator were proved for the choice of parameters

$$\Lambda = t^2 \cdot \sigma^2 \cdot (1 + \sqrt{2\log(M_N)})^2, \qquad \tau = \sqrt{\Lambda}, \tag{5.16}$$

where M_N is the number of chirplets in the dictionary, σ^2 is the variance of the white noise and $t > 4$. These are not necessary the best choices in practice, since the choice of Λ could tend to be a little bit conservative. Later in our numerical experiments we will use slightly different choices for these parameters.

Although the empirical best frame needs to be found among exponentially many tight frames, the search can be done rapidly. The cost of the search for dyadic intervals of every possible scale 2^{-j}, $j = 0, \dots, \log_2 N$ is at most of the order of N operations, while the computational complexity of the chirplet analysis is $O(M_N \log N)$, where M_N is the total number of chirplets in the dictionary. Since for a typical discretization (see Chapter 2), we have $M_N = O(N^\beta)$, for some $\beta > 1$, the total cost of calculating the estimator is dominated by the chirplet analysis. The search algorithm is similar to the best-basis algorithm for cosine packets [28] and for adapted bases of local cosines [83], and relies heavily on the entropy functional being additive. See [20] for details.

5.6.3 An interpretation of the entropy functional

Consider the data of the form (5.1) where the noise $z = (z_k)$ is Gaussian white noise such that $E|z_k|^2 = \sigma^2$. For the purpose of demonstration, we assume without a loss of generality that the noise level is $\sigma = 1$. For a tight frame Φ with N rows and p columns, we have the equivalent detection problem [2]

$$\Phi^* y = \Phi^* f + \Phi^* z,$$

which can be written as

$$y_k[\Phi] = f_k[\Phi] + z_k[\Phi], \quad k = 1, \ldots, p.$$

Define the complexity functional

$$\begin{aligned} K(y, \Phi, \theta) &:= \|\Phi^* y - \theta\|^2 + \Lambda \cdot \|\theta\|_{\ell_0} \\ &= \sum_k \left[|y_k[\Phi] - \theta_k|^2 + \Lambda \cdot 1_{\{\theta_k \neq 0\}} \right], \end{aligned}$$

If $\theta_k = 0$, the k-th term in the sum equals $|y_k[\Phi]|$; otherwise it equals $|y_k[\Phi] - \theta_k|^2$, which is minimized by $\theta_k = y_k[\Phi]$. Let $(\hat{\Phi}, \hat{\theta})$ be the minimizer of $K(y, \Phi, \theta)$. Then

$$K(y, \hat{\Phi}, \hat{\theta}) = \mathcal{E}_\Lambda(y, \hat{\Phi}).$$

Hence, we can interpret the entropy as a complexity functional which trades off the fit to the data and the complexity of the fitted model as measured by the number of non-zero frame coefficients. Note that if the frame vectors in Φ were orthonormal so that $\Phi^*\Phi = I_N$, we would have

$$\|\Phi^* y - \theta\|^2 + \Lambda \cdot \|\theta\|_{\ell_0} = \|y - \Phi\theta\|^2 + \Lambda \cdot \|\theta\|_{\ell_0},$$

which is a functional of the same form as in (5.3). Therefore, as a side-note, we have shown that for Gaussian linear regression model selection where the predictor matrix is orthonormal, we can solve (5.3) rapidly.

[2] Recall that $\Phi\Phi^* = I_N$ and therefore no information is lost since we can always go back to the original model, making the two problems equivalent.

5.7 Numerical Simulation

In this section we briefly explore the empirical performance of the BP estimator. We have chosen to work with complex-valued data and white noise but could have performed simulations on real-valued data. We postpone the important investigations of the empirical performance of the estimator in the case of real-valued data and colored noise for future research. Based on our promising numerical results for the BP test in the case of colored noise and real-valued data, we expect the estimation procedure to perform for the estimation of highly oscillatory real-valued chirps.

5.7.1 The basic setup

We generated data of the form

$$y_k = f_k + z_k, \qquad k = 0, 1, \ldots, N-1;$$

where (f_k) is a vector of equispaced time samples of a complex-valued chirp, and where (z_k) is a complex-valued white noise sequence $z = z^0 + \imath\, z^1$ where z^0 and z^1 are two independent vectors of i.i.d. $N(0, \sigma^2/2)$ variables. Note that $E|z_k|^2 = \sigma^2$ and $E\|z\|^2 = N\sigma^2$. In our experiments the signal f obeys the normalization $\|f\| = \sqrt{N}$ and we vary the noise level σ. The chirp is of the form

$$f(t) = C \cdot A(t) e^{\imath N \varphi(t)}, \tag{5.17}$$

and sampled at the equispaced points $t_k = k/N$, $k = 0, 1, \ldots, N-1$; the parameter C is a normalization constant. As a test signal we considered the *cosine phase chirp* with slowly varying amplitude:

$$A(t) = 2 + \cos(2\pi t + \pi/4), \qquad \varphi(t) = 2\pi\Big(\sin(2\pi t)/4\pi + 200\pi t/1024\Big).$$

In all the simulations, we chose $N = 512$ and the normalization constant C such that $\|f\|^2_{\ell_2} = 1$. Figure (5.4) shows the real part of the signal under study.

As a measure of performance we used the mean-squared error

$$MSE(f, \hat{f}) = E\left[\frac{1}{N}\sum_k |f_k - \hat{f}_k|^2\right], \tag{5.18}$$

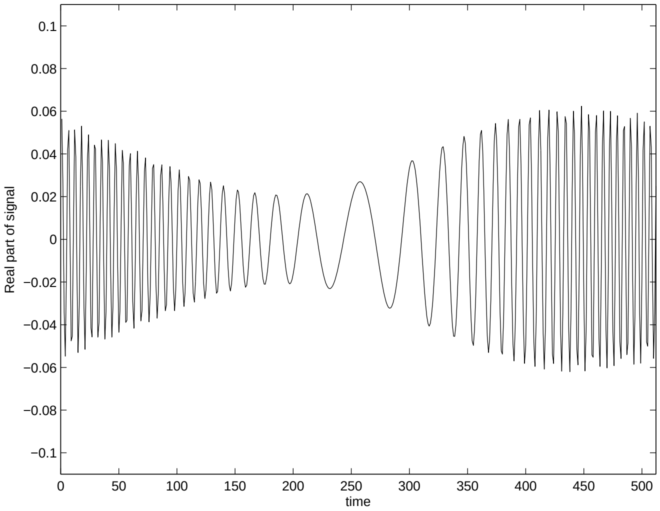


Figure 5.4: Real part $A(k/N)/\|A\| \cos(N\varphi(k/N))$, $k = 0, \ldots, N-1$ of the of the cosine phase chirp under study

where $\hat{f} = (\hat{f}_k)$ is the estimate of the unknown signal. Note the mean-squared error of estimating the signal using the data itself [3]; we have $MSE(f, y) = E\left[\frac{1}{N}\|z\|^2\right] = \sigma^2$.

In this setup, we define the Signal-to-Noise Ratio of an estimator $\hat{f}$ as

$$SNR_{db}(f, \hat{f}) = 10 \log_{10} \left(\frac{\|f\|^2}{E\|f - \hat{f}\|^2} \right). \tag{5.19}$$

SNR_{db} is measured in decibels and due to the normalization of f we have $SNR_{db}(f, y) = -10 \log_{10}(N \cdot \sigma^2)$. A high value of SNR_{db} is an indication that the estimator is good. For the sample SNR_{db} for a particular realization of the data, we drop the expectation in the definition.

We considered three estimation procedures:

1. *BP* : Best Path estimator using constant amplitude fits of chirplets; see Section 5.2.1.

2. *BPGA* : Best Path estimator based on the estimation of the chirp phase provided by standard BP estimator in Section 5.2.1, followed by a global amplitude fitting using splines as explained in Section 5.5.

3. *TBCF* : thresholding in the best chirplet tight frame; see Section 5.6.

All the methods assumed the noise variance to be known.

[3] This amounts to maximum likelihood estimation without penalization.

For the BP estimators we considered the chirplet graph such that two chirplets are connected if and only if they live on adjacent time intervals and if the instantaneous frequencies at their juncture coincide. For the discrete chirplets based on (2.1), we considered frequency offsets $b_\mu = 2\pi m$, for $m \in \{-N/2, \ldots, 0, \ldots, N/2 - 1\}$. For the BP estimator, we did not consider chirplets which had the instantaneous frequency falling out of this range. The slope parameters a_μ of the chirplets ranged from $-\pi N$ to πN with a discretization at scale 2^{-j} of the form

$$a_\mu = 2\pi N(-1/2 + k \cdot 2^{j-J}), \quad k \in \{0, \ldots, 2^{J-j}\}$$

where $J = \log_2 N$. This ensures that any endpoint of a dyadic interval is an integer multiple of 2π. The scales considered where 2^{-j}, where $j = 0, 1, \ldots, J-1$, resulting in dyadic intervals I of lengths $|I| = 2^{J-j}$. For practical purposes one might want to use a different discretization; in particular it should not be required to consider dyadic intervals of such small scales. We choose this configuration to keep things simple for our demonstration.

The parameter λ in the complexity functional (5.10) for the BP estimators was chosen to be

$$\lambda^2 = 4\sigma^2 \log(N).$$

We tried using different values for this parameter for the set of noise levels we considered. The estimator seemed to choose the same set of chirplets for values even factor 10 times different than the value given by this formula; i.e., it chose the same point on the convex hull determined by the value of the best chirplet paths of different lengths.

The smoothing parameter α for the spline smoothing for the global amplitude estimate in BPGA was chosen using cross-validation. For each realization of the data y, the cross-validation score $CV(\alpha)$ (see (5.13)) was calculated for the range of parameters

$$\alpha = 0.001 \cdot k, \quad k = 1, 2, \ldots, 50.$$

The parameter that minimizing $CV(\alpha)$ was used for the amplitude smoothing.

For the TBCF estimator we chose the same discretization of chirplet slopes as for the BP estimator. The parameters (5.16) were chosen to be

$$\Lambda = 4\sigma^2 \log(N) \quad \text{and} \quad \tau = 3\sqrt{2}.$$

For this fixed value of Λ, we tuned τ by trying out a series of different values for estimating our test signal and choosing a value that gave good results. The cutoff function for the chirplet windowing was chosen to be $\rho(t) = \sin(\pi/4(1 + \sin(\pi t)))$ for $|t| \leq 1/2$, 1 if $t > 1/2$, and 0 if $t < -1/2$.

5.7.2 Results

We compared the three estimation procedures for the test signal using different values of the noise level. Figures 5.5, 5.6, 5.7, and 5.8 show examples of estimations of the cosine phase chirp in the case of $\sigma = 0.01, 0.03, 0.05$, and 0.07, respectively. We plot the real parts of the noisy signal and the estimations for that particular realization. Figure 5.9 shows the estimation of the instantaneous frequency $\varphi'(t)$ for the same realizations. The estimation is based on the instantaneous frequency of the chirplets in the best path.

Both qualitatively and in terms of SNR_{db}, the BP estimator using the global amplitude fit performs the best. At the higher noise levels $\sigma = 0.05, 0.07$, the TBCF estimator starts to estimate the chirp as being absen,t while the other two methods give at least a qualitatively reasonable estimates - even for the extreme case $\sigma = 0.07$, where the plot of the noise does not even fit inside the portion of the graph that is displayed[4]. Notice the discontinuity of the estimate for the BP estimators in the slowly oscillatory portion of the chirp. Therefore there might be room for improving the methods for estimation at low frequencies; perhaps something along the lines of the “phaselet graph” we previously discussed.

Figure 5.9 shows that we have a very good agreement between the instantaneous frequency of the unknown chirp and the instantaneous frequency of the chirplet paths. Notice that as the noise level increases, the BP estimator uses fewer chirplets to fit the data.

Table 5.1 shows estimated $MSE(f, \hat{f})$ for a range of noise levels. The MSE was estimated by the sample mean of squared errors from 1,000 realizations of the noise sequence z. Notice that BPGA always performed best. The BP estimator performed better than TBCF for all the cases except for the lowest noise level. This is perhaps due to the crude estimation piecewise constant approximation of the amplitude.

[4] We choose to use the same scaling of the axis for all the plots.

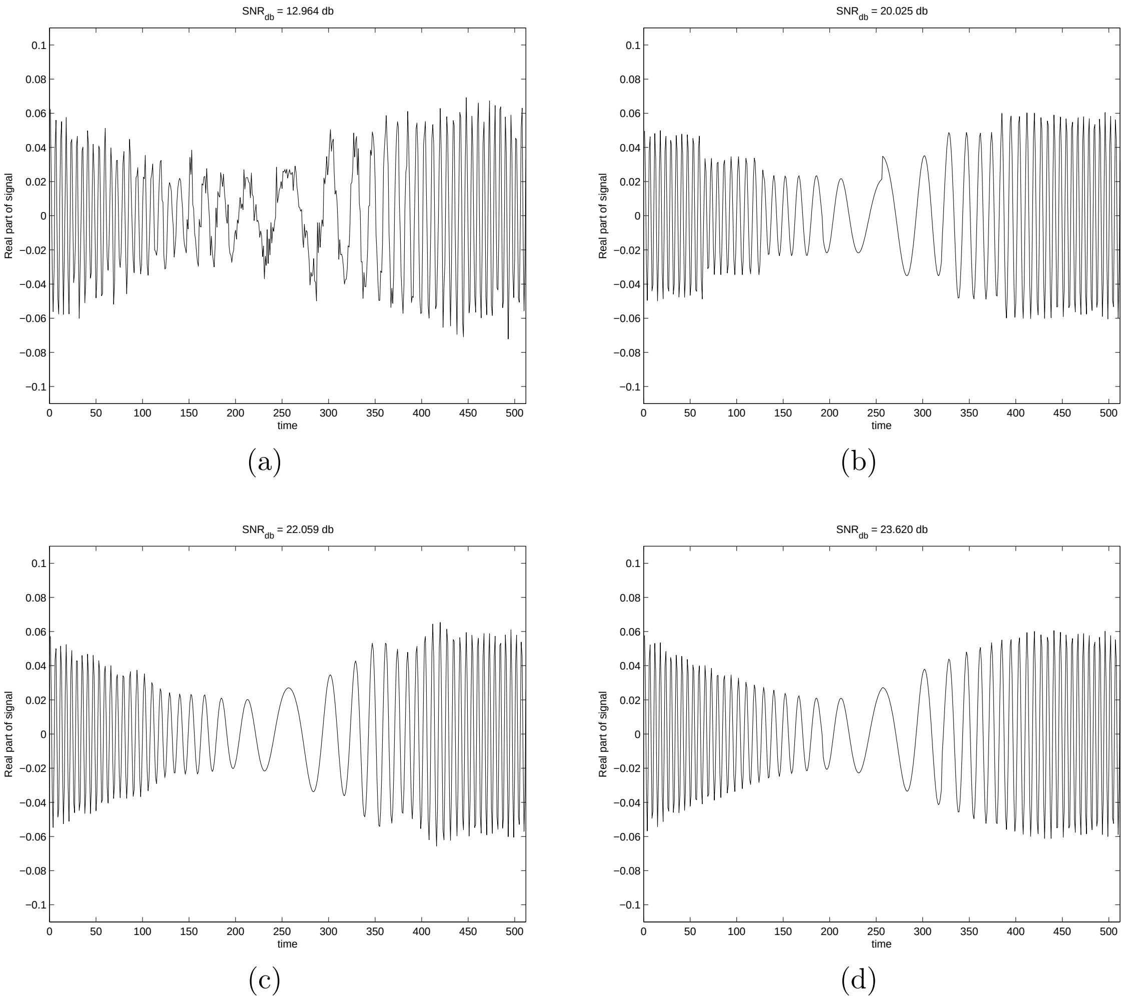
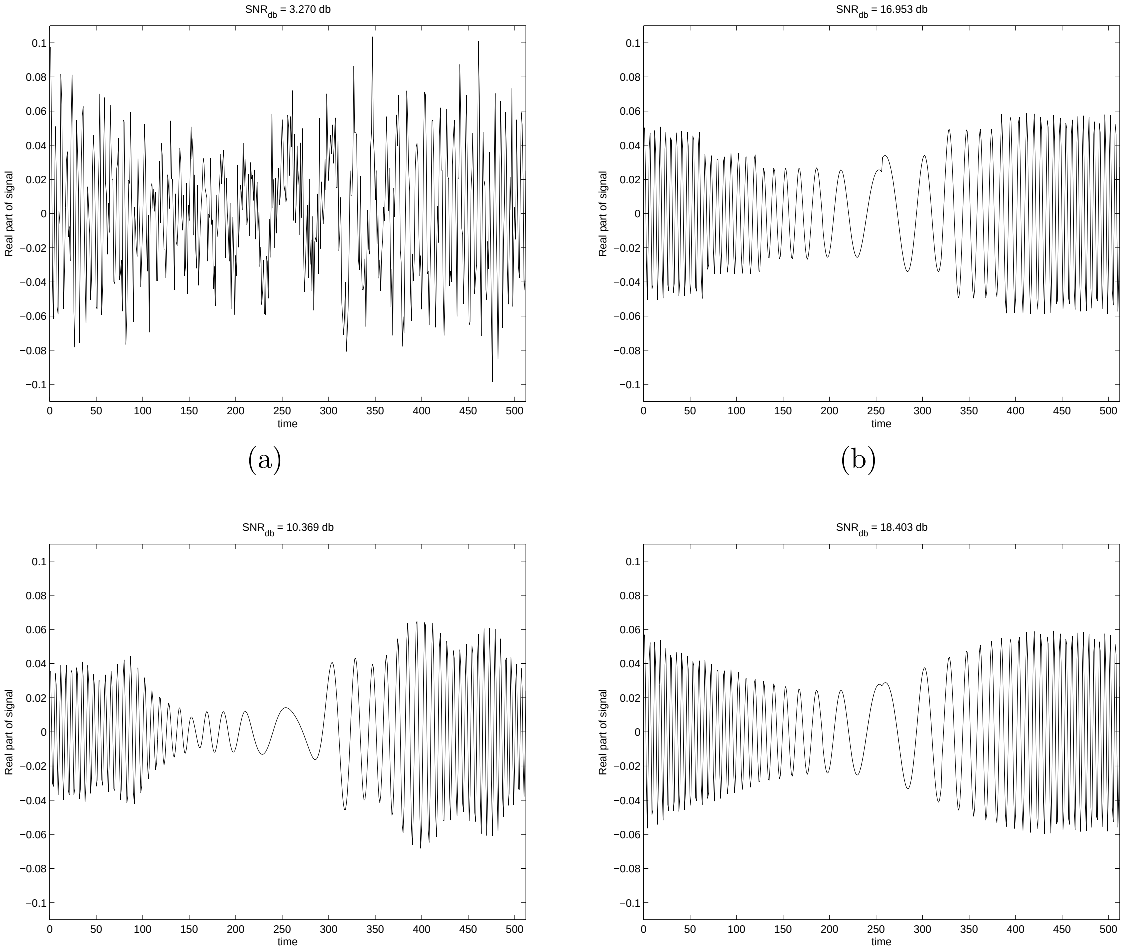


Figure 5.5: Comparison of the estimation procedures for noise level $\sigma = 0.01$: (a) Noisy signal, (b) BP estimation, (c) TBCF estimation, (d) BPGA estimation

σ	0.0100	0.0300	0.0500	0.0700	0.0900	0.1000
BP	0.0106	0.0208	0.1373	0.3452	0.6144	0.7353
TBCF	0.0071	0.0890	0.2857	0.5745	0.8906	0.9764
BPGA	0.0051	0.0146	0.1062	0.2596	0.4430	0.5607

Table 5.1: Comparison on estimated MSE for the three estimation procedures and the varying amplitude cosine chirp of length $N = 512$. Each estimate is based on 1,000 realizations of the noise.



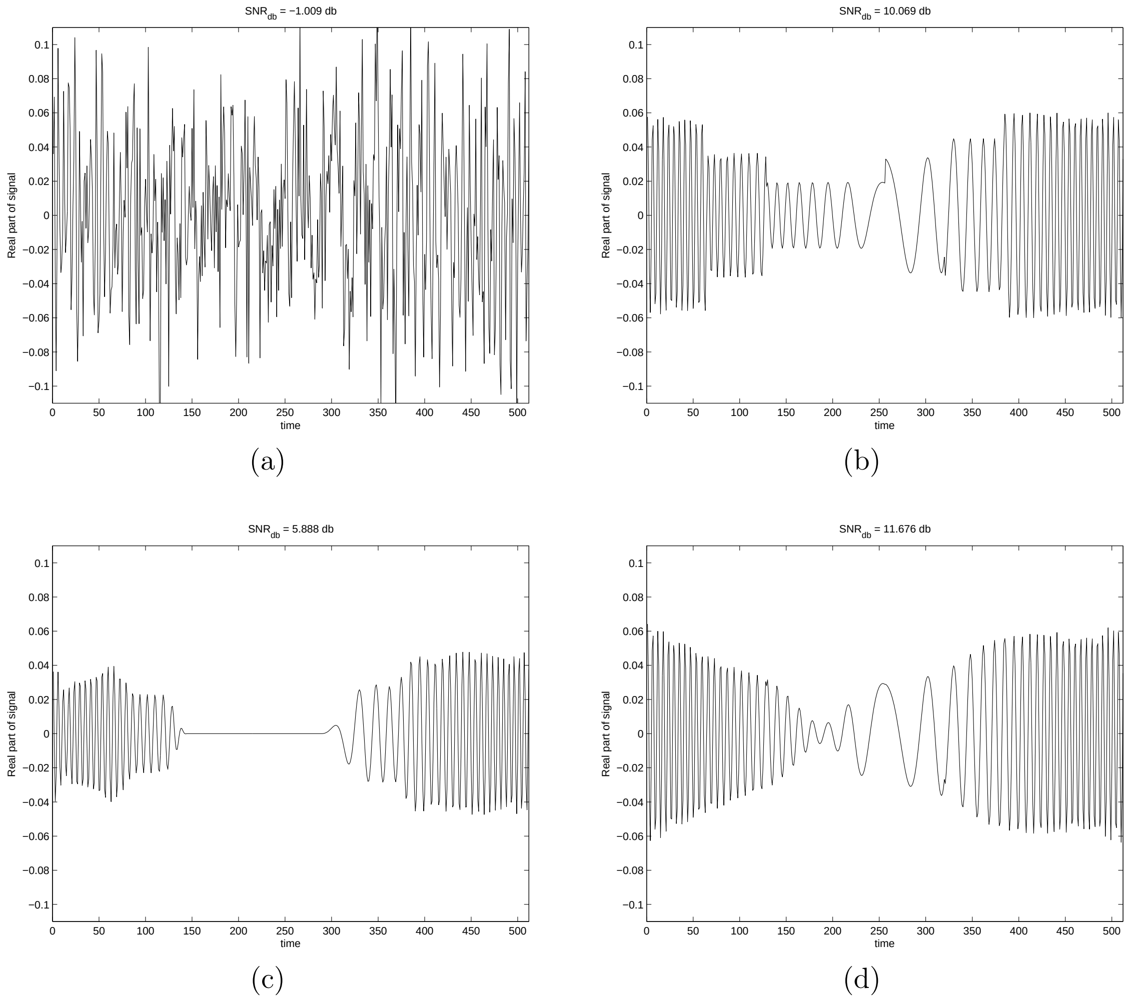
(a)

(b)

(c)

(d)

Figure 5.6: Comparison of the estimation procedures for noise level $\sigma = 0.03$: (a) Noisy signal, (b) BP estimation, (c) TBCF estimation, (d) BPGA estimation



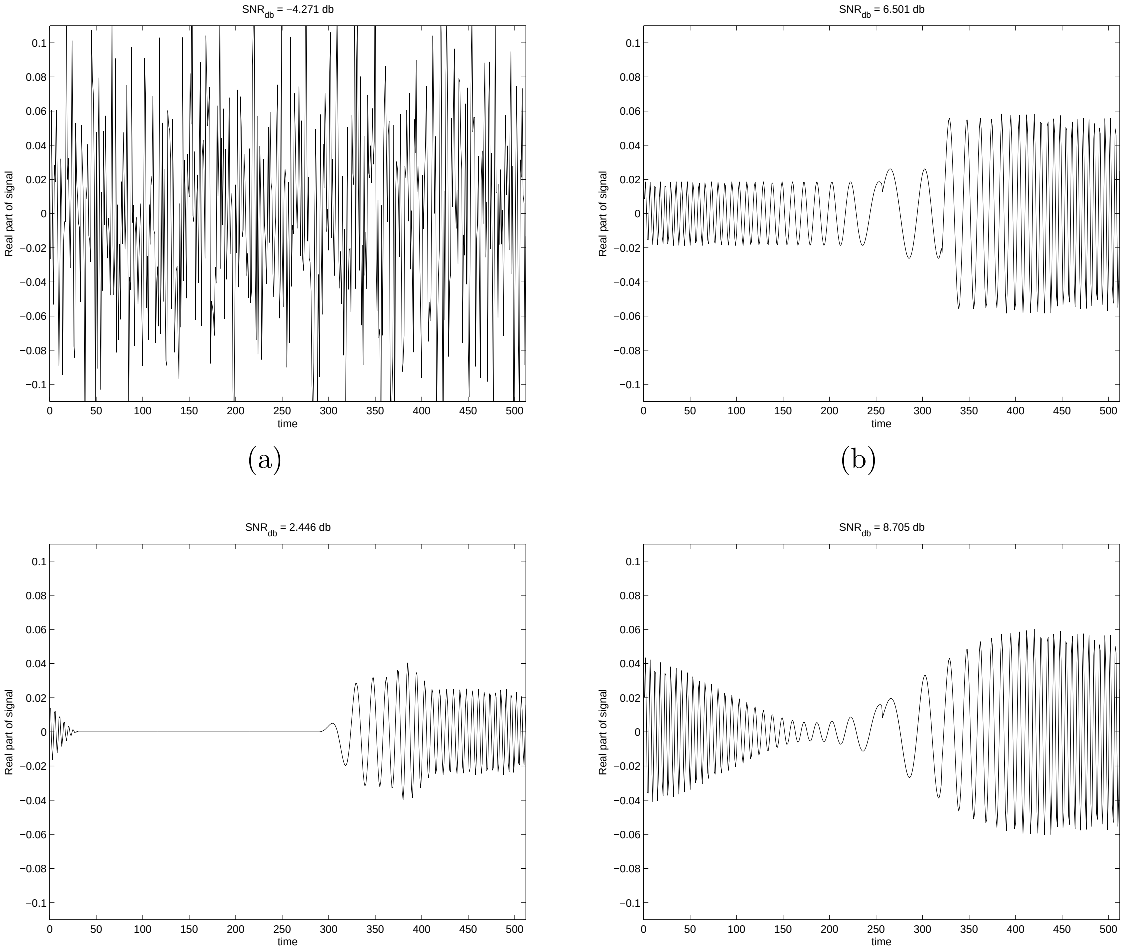
(a)

(b)

(c)

(d)

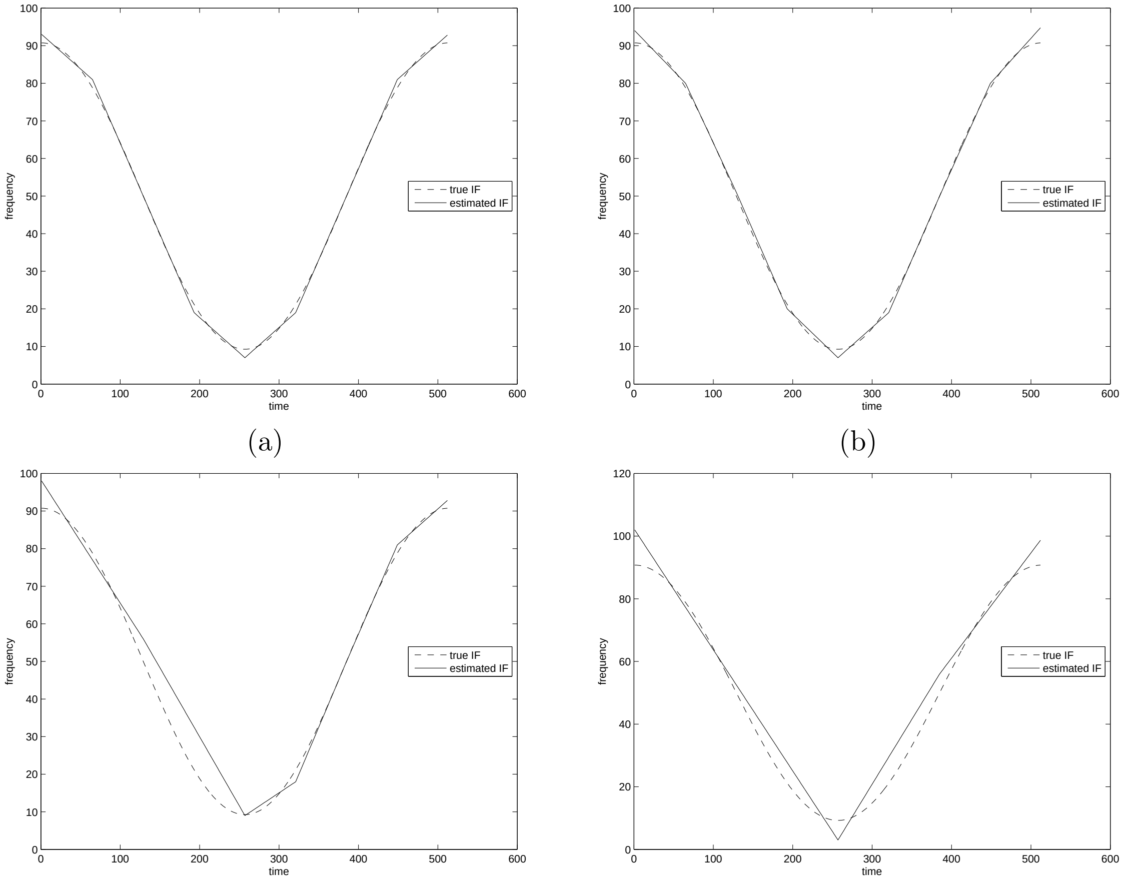
Figure 5.7: Comparison of the estimation procedures for noise level $\sigma = 0.05$: (a) Noisy signal, (b) BP estimation, (c) TBCF estimation, (d) BPGA estimation



(a) (b)

(c) (d)

Figure 5.8: Comparison of the estimation procedures for noise level $\sigma = 0.07$: (a) Noisy signal, (b) BP estimation, (c) TBCF estimation, (d) BPGA estimation



(a)

(b)

(c)

(d)

Figure 5.9: Estimation of the instantaneous frequency $\varphi'(t)$ based on the best chirplet path from the BP estimation procedure: (a) Noise level $\sigma = 0.01$, (b) Noise level $\sigma = 0.03$, (c) Noise level $\sigma = 0.05$, (d) Noise level $\sigma = 0.07$

5.7.3 Discussion

Note that TBCF uses balanced recursive dyadic partitions (BRDPs), while the BP estimator considered every possible recursive dyadic partition involving intervals of length $|I| = 2^{J-j}$, $j = 0, 1, \ldots, J-1$. Since the regularity of the chirp does not change dramatically over the interval, we would believe that adding this restriction to the BP estimator could only improve the estimation, since there would be fewer models to consider in the complexity functional - although the improvement might not be great.

We notice that the performance of the BP estimator seems to be better than that of the TBCF estimator at the noise levels we looked at. But the comparison is perhaps not fair since the BP estimator relies heavily on the assumption that the unknown signal is a single chirp, and would not be suitable for estimation problems where there was more than one chirp present in the data at the same time. However, the TBCF estimator does not make this assumption and could be used in situations where the BP estimator would be ill suited. Another point: TBCF does not rely much on the fact that the phase φ of the chirp is smooth and allows for the possibility of being extended to handle discontinuities in the instantaneous frequency, φ', due to the dyadic structure and adaptivity of the libraries of multiscale chirplets. Since the BP estimator relies on, -and exploits-, the fact that the phase is smooth, it would be difficult for it to compete with the TBCF estimator on these grounds (unless, perhaps, by making some prior assumptions about how big the jump at the discontinuity can be). Finally, if the data was only noise or the chirp "turn on and off", the TBCF would tend to correctly estimate the chirp to be zero at those places. Meanwhile, the BP estimator would always return some estimate. We could extend the BP estimator so that first we would do a BP test to see how likely there is to be a chirp in the data segment, and then follow with an estimation, or we could incorporate some kind of thresholding, setting any chirplet costs below a certain value to zero.

Chapter 6

Approximation Properties of Chirplet Paths

After introducing multiscale chirplets and the notion of a chirplet graph, we need to decide upon a discretization of the chirplets and define connectivities to fix the topology of the chirplet graph. Obviously such configurations depend on the class of functions we want to consider, such as how wildly we want to allow the chirps to behave, and so forth. Let's say we have decided upon the class of chirps; e.g., by using a mathematical description (bounds on derivatives, etc.). A natural requirement would be to configure the chirplet graph so that any chirp in the class is in some sense "close to" some chirplet path. That is, we want to know how to construct a chirplet graph so that the distance, measured using a suitable norm, from any chirp to the space of chirplet paths satisfies some prescribed bound. In the following sections we will give a precise mathematical description of such a class of chirps and give error bounds for approximation of chirps based on chirplet paths.

6.1 Mathematical Description of a Class of Chirps

We follow the definition in [20] and consider a class of chirps with some restrictions on the regularity of the amplitude A and phase φ of the chirps. The measure of roughness will be based on the Hölder regularity of these functions. For $0 < s \leq 1$, the function $g(t)$ is said to be in the Hölder class $\text{HÖLDER}^s(R)$ if $\|g\|_{L_\infty} \leq R$ and

$$|g(t) - g(t')| \leq R \cdot |t - t'|^s, \quad 0 \leq t, t' \leq 1.$$

For $s > 1$ and $m < s \leq m+1$, g is in the Hölder class $\text{HÖLDER}^s(R)$ if $\|g\|_{L_\infty} \leq R$, and

$$|g^{(m)}(t) - g^{(m)}(t')| \leq R \cdot |t - t'|^{s-m}, \quad 0 \leq t, t' \leq 1.$$

Note that if $s = 1$ the inequality above is simply the Lipschitz condition. For $s = m+1$, m positive integer, the Hölder class is the set of functions g bounded by R such that $g^{(m)}$ is Lipschitz. Denote the homogeneous Hölder norm by $\| \cdot \|_s$ where

$$\|g\|_s = \sup_{t,t'} \frac{|g^{(m)}(t) - g^{(m)}(t')|}{|t - t'|^{s-m}}.$$

For $\lambda \in [1, N]$ and $R > 0$, define the class of chirps

$$\text{CHIRP}(s, \lambda, R) := \{f : A, \varphi \in \text{HÖLDER}^s(R), |\varphi'(t)| \leq \pi, |A'(t)| \leq R\} \tag{6.1}$$

where $f(t) = A(t)\exp(i\lambda\varphi(t))$ or $f(t) = A(t)\cos(\lambda\varphi(t))$. The condition $A, \varphi \in \text{HÖLDER}^s(R)$ controls the roughness of the amplitude and phase of chirp, while $|\varphi'(t)| \leq \pi$ and $1 \leq \lambda \leq N$ ensures that the oscillation rate is bounded and does not exceed the Nyquist rate.

It is important to point out that this model is allowed to depend on the sampling rate. If A, φ, and λ were kept fixed and we let N grow, the chirps in the class under consideration would essentially become non-oscillatory and the statistical problems would correspond to estimation or detection of smooth functions. Instead we will keep the smoothness of A and φ fixed while letting λ grow as the number of samples N grows. We will mostly focus on the extreme case $\lambda = N$, which allows for oscillations almost at the sampling rate. We want to emphasize that the theoretical results on chirp estimation we will present in Chapter 7 will be nonasymptotic and in that case we can take N as being a large fixed constant.

6.2 Approximation of Chirps Using Chirplet Paths

Approximation properties are very important in statistical theory, particularly for estimation. The results below will be used for our discussion about near-optimal estimation of chirps using chirplet paths. We start with a lemma that provides an upper bound on how well in terms of the L_∞-norm the instantaneous frequency for any chirp in $\text{CHIRP}(s, N, R)$ can be approximated by an instantaneous frequency of a chirplet path. The estimates provided by this lemma are then used to establish bounds for errors in terms of ℓ_2 and L_2 norms of approximating chirps using constant amplitude chirplets and amplitude-modulated chirplets.

6.2.1 Approximation of the instantaneous frequency

Consider $f \in \mathrm{CHIRP}(s, N, R)$, where either $f(t) = A(t)\cos(N\varphi(t))$, or $f(t) = A(t)\exp(iN\varphi(t))$. The derivation of our bounds for approximating f using chirplet paths will be based on how well the instantaneous frequency, φ', of the chirp can be approximated by an instantaneous frequency of a chirplet path. Let $\mathcal{P}$ be a partition of the interval $[0, 1]$, and for $I \in \mathcal{P}$, write $I = [t_{0,I}, t_{1,I}]$. Then we have the following bounds:

Lemma 3. *Assume $\varphi \in \text{HÖLDER}^s(R)$ and that the parameters a_μ, b_μ in (2.1) are of the form*

$$b_\mu = 2\pi\Delta b \cdot m, \qquad a_\mu = 2\pi\frac{\Delta b}{|I|} \cdot l, \tag{6.2}$$

where $l, m \in \mathbb{Z}$, for some $\Delta b > 0$. Then there is a continuous broken line $\sum_{I\in\mathcal{P}}(b_\mu + a_\mu(t - t_{0,I}))1_I(t)$ such that for every $I \in \mathcal{P}$,

$$\sup_{t\in I} |\varphi'(t) - (b_\mu + a_\mu(t - t_{0,I}))| \leq R \cdot |I|^{s-1} + 2\pi\Delta b. \tag{6.3}$$

As a consequence, the piecewise quadratic phase function, $\theta(t)$, whose first derivative is equal to this broken line, i.e., $\theta'(t) = \sum_{I\in\mathcal{P}}(b_\mu + a_\mu(t - t_{0,I}))1_I(t)$, and $\theta(t_{0,I}) = \varphi(t_{0,I})$, obeys:

$$\sup_{t\in I} |\varphi(t) - \theta(t)| \leq R \cdot |I|^s + 2\pi\Delta b|I|, \tag{6.4}$$

for every $I \in \mathcal{P}$.

For these bounds to hold, it suffices to consider slope parameters (a_μ) such that

$$|a_\mu| \leq R \cdot |I|^{s-2} + 3\pi\Delta b|I|^{-1}. \tag{6.5}$$

For $s = 2$, the range

$$|a_\mu| \leq R \cdot |I|^{s-2} + 3\pi\Delta b|I|^{-1}$$

suffices.

The proof the lemma may be found in Appendix C.2 and relies on error bounds for

Taylor approximations of functions in $\text{HÖLDER}^s(R)$. Note that a similar lemma was stated and proved in [20].

Lemma 3 suggests a discretization for the chirplet graph. The bound on the maximum distance from the piecewise polynomial curve from φ' and φ in (6.3) and (6.4) is split into two terms:

1. Approximation error due to the smoothness constraint of the phase function φ.
2. Approximation error due to discretization of the frequency offset $b_{\mu,I}$.

The first error dominates the other unless we choose a fine enough discretization for the frequency offsets. How small we need to take the discretization step Δb depends on the size of the shortest interval in the partition $\mathcal{P}$ and the smoothness parameter s. Bounding the discretization step gives us:

Corollary 1. *For the frequency spacing, $\Delta b \leq \inf_{I \in \mathcal{P}} |I|^{s-1}$, the bounds are*

$$\sup_{t \in I} |\varphi'(t) - (b_\mu + a_\mu(t - t_{0,I}))| \leq C(R) \cdot |I|^{s-1}, \tag{6.6}$$

and

$$\sup_{t \in I} |\varphi(t) - \theta(t)| \leq C(R) \cdot |I|^s, \tag{6.7}$$

for every $I \in \mathcal{P}$ where we can take $C(R) = R + 2\pi$. It suffices to consider slope parameters (a_μ) such that

$$|a_\mu| \leq (C(R) + \pi) \cdot |I|^{s-2}. \tag{6.8}$$

For $s = 2$, the range

$$|a_\mu| \leq C(R),$$

suffices.

The requirement

$$|\varphi'(t)| \leq \pi,$$

for chirps in $\text{CHIRP}(s, \lambda, R)$ with phase φ, gives us a sufficient range for the frequency offsets $b_{\mu,I} = 2\pi\Delta b \cdot m$, $m \in \mathbb{Z}$; it is enough to consider integers m satisfying

$$|m| \leq \frac{1}{2\Delta b}.$$

We chose the discretization of the frequency b_μ to be independent of the length of the interval $|I|$. This simplifies the topology of the chirplet graph and therefore makes implementations of network flow algorithms a little bit easier. With this choice, we can define the following connectivity constraint and the lemma would still hold: Two chirplets, possibly at different scales, are connected if and only if they are supported on adjacent time intervals and their "instantaneous frequencies" at the juncture coincide. Another possibility would be to choose the discretization of the frequency offset to be scale dependent. Then we could still have a piecewise linear function satisfying the bounds in the lemma, but not necessarily continuous. At the breakpoints of the piecewise linear function, we would, however, have some inequalities for the distance between the endpoints. This would give us some connectivity requirements which would determine a topology for the chirplet graph.

Lemma 3 tells us how wide the range of slope parameters needs to be. The next lemma gives us guidelines for deciding the connectivities in the chirplet graph for approximating chirps from CHIRP(s, N, R) using the discretization in Lemma 3. Assume we have two chirplets on adjacent time intervals such that one ends at the same frequency as the other one starts. Then we have the following sufficient bound on the maximum difference in their slopes for the estimates (6.3) and (6.4) to still hold.

Lemma 4. *Assume $\varphi \in \textit{HÖLDER}^s(R)$. Let $\mathcal{P}$ be a partition of $[0,1)$ and the collection of parameters $(b_I)_{I\in\mathcal{P}}$ and $(a_I)_{I\in\mathcal{P}}$ be of the form as in Lemma 3 such that the continuous broken line $\sum_{I\in\mathcal{P}}(b_I + a_I(t - t_{0,I}))1_I(t)$ satisfies (6.3) and is constructed as in the proof of Lemma 3. Take $\Delta b \le \inf_I |I|^{s-1}$.Then for every two adjacent intervals, I and I', in $\mathcal{P}$,*

$$|a_I - a_{I'}| \le (2R + 6\pi) \cdot (\max\{|I|, |I'|\})^{s-2}. \tag{6.9}$$

For $s = 2$, the inequality holds with the constant $2R + 4\pi$ instead.

Proof. This lemma follows directly from the bound on the slope parameter in (6.8) and the triangle inequality $|a_I - a_{I'}| \le |a_I| + |a_{I'}|$. □

6.2.1.1 Chirplet graph topology for balanced recursive dyadic partitions

Restricting ourselves to Balanced Recursive Dyadic Partitions $\mathcal{P}$ as in Definition 1, we have that for any adjacent intervals $I, I' \in \mathcal{P}$,

$$\max\{|I|, |I'|\} \leq 2 \min\{|I|, |I'|\}.$$

Consider the case when these intervals differ in length. Without loss of generality, assume that $|I| > |I'|$ so that

$$|I'| = \frac{1}{2}|I|.$$

Let $a_I = 2\pi|I|^{s-2} \cdot l_1$ and $a_{I'} = 2\pi|I'|^{s-2} \cdot l_2$ be the slopes of the continuous broken line constructed in the proof of Lemma 3 and satisfying (6.3). Then from Lemma 4 we have

$$|a_I - a_{I'}| = 2\pi|l_1 - l_2/2^{s-2}| \cdot |I|^{s-2} \leq (2R + 6\pi) \cdot |I|^{s-2},$$

and therefore

$$|l_1 - l_2 \cdot 2^{2-s}| \leq (R/\pi + 3).$$

Hence, each chirplet only needs to be connected to a constant number of chirplets for the approximation bounds in Lemma 3 to hold.

6.2.1.2 Chirplet graph topology for monoscale analysis

If we knew the regularity s of the unknown chirp *a priori* we could use a monoscale chirplet graph where the time axis is partitioned uniformly. Assume that we use parameters of the same form as in Lemma 3 with the frequency discretization equal to $\Delta b = |I|^{s-1}$, where $|I|$ is the length of the intervals in the uniform partition. Then

$$a_\mu = 2\pi|I|^{s-2} \cdot l, \quad l \in \mathbb{Z}.$$

Take two adjacent time intervals I and I' from the partition. On these intervals, let $a_I = 2\pi|I|^{s-2} \cdot l_1$ and $a_{I'} = 2\pi|I|^{s-2} \cdot l_2$ be the slopes of the continuous broken line constructed in the proof of Lemma 3 and satisfying (6.3). By Lemma 4,

$$|a_I - a_{I'}| = 2\pi|I|^{s-2} \cdot \Delta l \leq C \cdot |I|^{s-2},$$

or,

$$\Delta l \leq C',$$

where $\Delta l = |l_1 - l_2| \in \mathbb{Z}$ and the constant C' can taken to be $C' = R/\pi + 3$, and if $s = 2$, $C' = R/\pi + 2$. Thus for monoscale analysis, every chirplet in the chirplet graph only needs to be connected to a constant number of chirplets for the bounds in Lemma 3 to hold.

6.2.2 The case $s = 2$

For chirps in CHIRP(s, λ, R) with $s = 2$, we can get the same bounds as in Lemma 3 by considering only monochromatic chirplets; i.e., chirplets whose frequency does not change over their support, or, simply said, local cosines. If we multiply each chirplet with a smooth window, we would have basis functions similar to classical Gabor Analysis (see [52] and references therein). Lemma 5 states this fact more precisely. The proof is given in Appendix C.3.

Lemma 5. *Assume $\varphi \in$ HÖLDER$^s(R)$ with $s = 2$. Consider the uniform partition $\mathcal{P}$ of $[0, 1]$ into L intervals, I_k, $k = 1, \dots, L$, each of length $|I|$. Let the parameters a_μ, b_μ in (2.1) be such that $a_\mu = 0$ and*

$$b_\mu = 2\pi\Delta b \cdot m, \tag{6.10}$$

where $m \in \mathbb{Z}$, for some $\Delta b > 0$. Then there is a piecewise constant function $\sum_{I\in\mathcal{P}} b_I 1_I(t)$ such that all of the below are satisfied:

1. *For every $I \in \mathcal{P}$,*

$$\sup_{t\in I} |\varphi'(t) - b_I| \leq R/2 \cdot |I| + \pi\Delta b. \tag{6.11}$$

2. *For two adjacent intervals, I and I',*

$$|b_I - b_{I'}| \leq R \cdot |I| + 2\pi\Delta b. \tag{6.12}$$

3. *The piecewise linear phase function, $\theta(t)$, satisfying $\theta'(t) = \sum_{I\in\mathcal{P}} b_I 1_I(t)$, and $\theta(t_{0,I}) = \varphi(t_{0,I})$, obeys:*

$$\sup_{t\in I} |\varphi(t) - \theta(t)| \leq R/2 \cdot |I|^2 + \pi\Delta b|I|, \tag{6.13}$$

for every $I \in \mathcal{P}$.

We also give a lemma for the specific case $R = 1$, which will be used in a later chapter. Its trivial proof is given in Appendix C.4.

Lemma 6. *Assume same conditions as in Lemma 5, but with*

$$b_{\mu,I_k} = 2\pi|I| \cdot m, \text{ if } k \text{ is odd}, \quad b_{\mu,I_k} = 2\pi|I| \cdot (m + 1/2), \text{ if } k \text{ is even}, \tag{6.14}$$

where $m \in \mathbb{Z}$. Then there is a piecewise constant function $\sum_{I\in\mathcal{P}} b_I 1_I(t)$ such that all of the below are satisfied:

1. *For every $I \in \mathcal{P}$,*
$$\sup_{t\in I} |\varphi'(t) - b_I| \leq (1/2 + \pi) \cdot |I|. \tag{6.15}$$

2. *Let I and I' be two adjacent intervals with $b_I = 2\pi|I| \cdot m_1$ and $b_{I'} = 2\pi|I| \cdot (m_2 + 1/2)$, $m_1, m_2 \in \mathbb{Z}$. Then,*
$$m_1 = m_2, \quad \text{or} \quad m_1 = m_2 - 1. \tag{6.16}$$

3. *The piecewise linear phase function, $\theta(t)$, satisfying $\theta'(t) = \sum_{I\in\mathcal{P}} b_\mu 1_I(t)$, and $\theta(t_{0,I}) = \varphi(t_{0,I})$, obeys:*
$$\sup_{t\in I} |\varphi(t) - \theta(t)| \leq (1/2 + \pi) \cdot |I|^2, \tag{6.17}$$
for every $I \in \mathcal{P}$.

Thus for phase functions $\varphi \in \text{HÖLDER}^2(1)$, the dictionary of chirplets and the topology of the chirplet graph can be very simple for the bounds (6.15) and (6.17) to hold. Every chirplet is monochromatic and connects to two other chirplets to the right (unless, of course, it is a chirplet at the far right end in the chirplet graph).

6.2.3 Approximation using constant-amplitude chirplets

In the analysis below the norm $\|\cdot\|$ can either stand for the L_2-norm $\|f\|_{L_2}^2 = \int_0^1 |f(t)|^2 dt$ or the ℓ_2-norm $\|f\|_{\ell_2}^2 = \sum_{k=0}^{N-1} |f(k/N)|^2$. Let $f \in \text{CHIRP}(s, N, R)$ be a complex-valued chirp so that

$$f(t) = A(t) \cdot e^{iN\varphi(t)},$$

where A, φ satisfy the conditions in (6.1). Consider a chirplet graph with a uniform partition $\mathcal{P}$ of the time interval $[0, 1]$ into intervals of length $|I|$. Let the slope and frequency offset parameters of the chirplets in the graph be chosen as in (6.2) in Lemma 3 with

$$\Delta b = |I|^{s-1},$$

so that Corollary 1 holds. Let W be a chirplet path in the graph and write $\mathcal{P} = \{I_v : v \in W\}$. Then the number of chirplets in the path is

$$|W| = |I|^{-1}.$$

Let θ be a phase function corresponding to this path such that

$$\theta(t) = \sum_{v \in W} \theta_v(t) 1_{I_v}(t)$$

where $\theta_v(t)$ is the phase of the chirplet $c_v(t) = \exp(iN\theta_v(t)) \cdot 1_{I_v}(t)/\|1_{I_v}\|$, normalized such that $\|c_v\| = 1$. We wish to approximate f with a function $\tilde{f}$ of the form

$$\tilde{f} = \sum_{v \in W} \lambda_v c_v, \tag{6.18}$$

where (λ_v) is a family of complex scalars. We assume the chirplets have phase offsets equal to 0, that is, if $I_v = [t_0, t_1)$, $\theta_v(t_0) = 0$. Then the phase offset of the local fitting function $\lambda_v c_v$ is encoded in λ_v. Equivalently we could have considered λ_v to be a non-negative real number and included a phase offset parameter in the chirplet c_v.

For a fixed phase function θ, the minimum squared error for approximating f using a linear combination of chirplets as in (6.18), is

$$\begin{aligned}
\min_{(\lambda_v)} \|f - \tilde{f}\|^2 &= \min_{(\lambda_v)} \left[\|f\|^2 - \sum_{v \in W} 2Re(\langle f, \lambda_v c_v \rangle) + |\lambda_v|^2 \|c_v\|^2 \right] \\
&= \|f\|^2 + \sum_{v \in W} \min_{(\lambda_v)} \left[|\lambda_v|^2 - 2Re(\langle f, \lambda_v c_v \rangle) \right] \\
&= \|f\|^2 - \sum_{v \in W} |\langle f, c_v \rangle|^2.
\end{aligned}$$

The last equality follows from:

1. $Re(\langle f, \lambda_v c_v \rangle) \leq |\langle f, \lambda_v c_v \rangle| = |\lambda_v||\langle f, c_v \rangle|$, with equality when $\arg(\langle f, \lambda_v c_v \rangle) = 0$, or

$$\arg \lambda_v = \arg\langle f, c_v \rangle =: \gamma_v.$$

2. The minimization of

$$|\lambda_v|^2 - 2|\langle f, c_v \rangle| \cdot |\lambda_v|,$$

achieved when

$$|\lambda_v| = |\langle f, c_v \rangle|.$$

Therefore, the minimizer is of the form

$$\tilde{f} = \sum_{v \in W} \lambda_v c_v = \sum_{v \in W} e^{i\gamma_v} |\langle f, c_v \rangle| \cdot c_v.$$

Later, when we consider chirp estimation using chirplet paths, these two conditions for equality can be interpreted as the local maximum likelihood estimates (MLE) of the phase offset and the amplitude of the unknown chirp.

The next step is to get a good upper bound on the approximation of the chirp; i.e., establish an upper bound for

$$\min_{W} \min_{(\lambda_v)} \|f - \tilde{f}\|^2 = \min_{W} \|f\|^2 - \sum_{v \in W} |\langle f, c_v \rangle|^2. \tag{6.19}$$

In our chirplet graph, there is a chirplet path W' with a phase function θ such that θ' and φ' satisfy inequality (6.3) from Lemma 3. An important observation is that the approximation error does not depend on the phase offset of the chirplets. Therefore, to establish an upper bound on the right-hand side of (6.19), we can assume $\theta_v(t_0) = \varphi(t_0)$ for every interval $I_v = [t_0, t_1]$ in the partition $\mathcal{P}$. Thus we can assume that θ and φ satisfy (6.4) from Lemma 3. Write $\delta = \varphi - \theta$ and let $\bar{\delta}$ be the upper bound in inequality (6.7); i.e.,

$$\bar{\delta} = C \cdot |I_v|^s, \tag{6.20}$$

with $C = C(R) = R + 2\pi$. Furthermore, assume

$$\bar{\delta} \leq N^{-1}.$$

Define $\langle\cdot\rangle_I$ and $\|\cdot\|_I$ by $\langle f, g\rangle_I = \langle f, g1_I\rangle$ and $\|f\|_I^2 = \langle f, f\rangle_I$ for any two functions f and g. As an upper bound for (6.19) we will use

$$\|f\|^2 - \sum_{v\in W'} |\langle f, c_v\rangle|^2 = \sum_{v\in W'} \|f\|_{I_v}^2 - |\langle f, c_v\rangle|^2.$$

It is clear that it suffices to bound each term in the sum separately. We start by establishing a lower bound on $|\langle f, c_v\rangle|^2$. The elementary inequalities $|Re(z)| \le |z|$ for all $z \in \mathbb{C}$, and $\cos(x) \ge 1 - x^2/2$ for all $x \in \mathbb{R}$, give

$$\begin{aligned} |\langle f, c_v\rangle|^2 &= |\langle Ae^{N\delta}, 1_{I_v}/\|1_{I_v}\|\rangle|^2 \ge (\langle A\cos(N\delta), 1_v/\|1_{I_v}\|\rangle)^2 \\ &\ge (1 - N^2\bar{\delta}^2/2)^2 \cdot (\langle |A|, 1_{I_v}/\|1_{I_v}\|\rangle)^2 \\ &\ge (1 - N^2\bar{\delta}^2) \cdot (\langle |A|, 1_{I_v}/\|1_{I_v}\|\rangle)^2. \end{aligned}$$

We can bound $(\langle |A|, 1_{I_v}/\|1_{I_v}\|\rangle)^2$ by the Cauchy-Schwartz inequality:

$$(\langle |A|, 1_{I_v}/\|1_{I_v}\|\rangle)^2 \le \|A\|^2 \|1_{I_v}/\|1_{I_v}\|\| = \|A\|^2.$$

Until now the bounds have been independent of the type of norm. Consider the two cases:

1. *Assume we are using the L_2-norm*: Then $\|1_{I_v}\| = \sqrt{|I_v|}$. Bounding $(\langle |A|, 1_{I_v}/\|1_{I_v}\|\rangle)^2$ by the Cauchy-Schwartz inequality as above, gives,

$$\begin{aligned} \|f\|_{I_v}^2 - |\langle f, c_v\rangle|^2 &\le \|A\|_{I_v}^2 - (\langle A, 1_{I_v}/\sqrt{|I_v|}\rangle)^2 + N^2\bar{\delta}^2 \cdot (\langle |A|, 1_{I_v}/\sqrt{|I_v|}\rangle)^2 \\ &\le \|A\|_{I_v}^2 - (\langle A, 1_{I_v}/\sqrt{|I_v|}\rangle)^2 + N^2\bar{\delta}^2 \cdot \|A\|_{I_v}^2 \\ &\le \|A\|_{I_v}^2 - (\langle A, 1_v/\sqrt{|I_v|}\rangle)^2 + C^2 \cdot N^2|I_v|^{2s} \cdot \|A\|_{I_v}^2. \end{aligned}$$

Observe that the first two terms are simply the minimum squared error of approximating $A(t)$ with a constant function on the interval I_v. Therefore, Taylor's formula and the requirement $|A'| \le R$ give us the bound

$$\|A\|_{I_v}^2 - (\langle A, 1_{I_v}/\sqrt{|I_v|}\rangle)^2 \le R^2|I_v|^2\|1_{I_v}\|_{I_v}^2 = R^2|I_v|^3.$$

We also have the trivial bound

$$\|A\|_{I_v}^2 - (\langle A, 1_{I_v}/\sqrt{|I_v|}\rangle)^2 \leq \|A\|_{I_v}^2.$$

Finally we have

$$\begin{aligned}\min_W \min_{(\lambda_v)} \|f-\tilde{f}\|_{L_2}^2 &\leq |I_v|^{-1}\cdot(R^2|I_v|^3) + \|A\|_{L_2}^2 \cdot C^2 \cdot N^2 |I_v|^{2s} \\ &= R^2|I_v|^2 + \|A\|_{L_2}^2 \cdot C^2 \cdot N^2 \cdot |I_v|^{2s}.\end{aligned}$$

2. *Assume we are using the ℓ_2-norm*: Let N_I be the number of points in each interval $I \in \mathcal{P}$. Then the normalization constant for the chirplets is

$$\|1_{I_v}\| = \sqrt{N_I}.$$

The same arguments as above give,

$$\|f\|_{I_v}^2 - |\langle f, c_v\rangle|^2 \quad \leq \quad \|A\|_{I_v}^2 - (\langle A, 1_{I_v}/\sqrt{N_I}\rangle)^2 + C^2 \cdot N^2 |I_v|^{2s} \|A\|_{I_v}^2,$$

and

$$\|A\|_{I_v}^2 - (\langle A, 1_{I_v}/\sqrt{N_I}\rangle)^2 \leq R^2|I_v|^2 N_I.$$

Since the total number of points is $N = N_I|I|^{-1}$, we get

$$\begin{aligned}\min_W \min_{(\lambda_v)} \|f-\tilde{f}\|_{\ell_2}^2 &\leq |I_v|^{-1}\cdot(R^2|I_v|^2 N_I) + \|A\|_{\ell_2}^2 \cdot C^2 \cdot N^2 \cdot |I_v|^{2s} \\ &= R^2 N|I_v|^2 + \|A\|_{\ell_2}^2 \cdot C^2 \cdot N^2 \cdot |I_v|^{2s}.\end{aligned}$$

We summarize these results in a theorem:

Theorem 2. *Suppose $f \in$ CHIRP(s, N, R) and consider the chirplet graph $\mathcal{G}_N$ with a uniform partition of the time interval $[0,1]$ such that each interval is of length $|I|$, the slope and offset parameters are discretized as in Lemma 3, and*

$$C \cdot N \cdot |I|^s \leq 1. \tag{6.21}$$

Then there exists a chirplet path W in the chirplet graph $\mathcal{G}_N$ and a family of complex scalars

$(\lambda_v)_{v \in W}$ *such that the linear combination of chirplets* $\tilde{f} = \sum_{v \in W} \lambda_v c_v$*, satisfies:*

$$\|f - \tilde{f}\|^2 \leq \min(\|A\|^2, R^2|I|^2) + \|A\|^2 \cdot C^2 \cdot N^2 \cdot |I|^{2s}, \tag{6.22}$$

where the constant $C = C(R)$ *is as in Corollary 1. In particular, since* $\|A\|_{L_\infty} \leq R$*, we have*

$$\|f - \tilde{f}\|^2 \leq R^2|I|^2 + R^2 \cdot C^2 \cdot N^2 \cdot |I|^{2s}.$$

The norm is either the L_2*-norm or the Euclidian norm defined by* $\|f\|^2 = 1/N \cdot \sum_{k=0}^{N-1} |f(k/N)|^2$.

We also have established an important bound for the sum of the chirplet costs along a path:

Corollary 2. *Assume the conditions in Theorem 2 hold.*

- ℓ_2-norm*: If*

  $$\|A\|_{\ell_2} \geq \sqrt{N} R|I|,$$

  then there exists a chirplet path W *in the graph that*

  $$\sum_{v \in W} |\langle f, c_v \rangle|^2 \geq \|A\|_{\ell_2}^2 \cdot \left(1 - C^2 \cdot N^2 \cdot |I|^{2s}\right) - NR^2|I|^2.$$

- L_2-norm*: If*

  $$\|A\|_{L_2} \geq R|I|,$$

  then there exists a chirplet path W *in the graph that*

  $$\sum_{v \in W} |\langle f, c_v \rangle|^2 \geq \|A\|_{L_2}^2 \cdot \left(1 - C^2 \cdot N^2 \cdot |I|^{2s}\right) - R^2|I|^2.$$

The constant $C = C(R)$ *can be chosen to be the same as in Corollary 1.*

The first term of the bound (6.22) in Theorem 2 can be interpreted as the error due to approximating the amplitude of the chirp by a piecewise constant function. The second term is due to how well we can approximate the instantaneous frequency of the chirp. Note that the bound is trivial unless the norm (or energy) of the chirp, $\|A\|$, satisfies

$$\|A\| \geq R|I|. \tag{6.23}$$

Then the second term controls the error of approximation if the following condition is satisfied:

$$\|A\|^2 \cdot C^2 \cdot N^2 \cdot |I|^{2s} \geq R^2 \cdot |I|^2,$$

or

$$|I| \geq C_1 \cdot N^{-1/(s-1)},$$

for the constant $C_1 = (R^2/(\|A\|^2 \cdot C^2))^{1/(2(s-1))}$. Note that condition (6.21) imposes

$$|I| \leq C_2 \cdot N^{-1/s},$$

for $C_2 = 1/C$. Since $s \geq 2$, $N^{-1/s} \geq N^{-1/(s-1)}$, and therefore these two conditions do not need to contradict each other. Notice that the lower bound requirement has a constant where the norm of the chirp, $\|A\|$, appears in the denominator.

As a second corollary of Theorem 2, we have:

Corollary 3 (m-term approximation)**.** *Suppose $f \in$ CHIRP(s, N, R). Assume the chirplet graph implied in Theorem 2 where the uniform partition of $[0, 1)$ consists of $m = |I|^{-1}$ intervals such that*

$$C_1 \cdot N^{-1/(s-1)} \leq |I| \leq C_2 \cdot N^{-1/s}. \tag{6.24}$$

Then there is a linear combination $\tilde{f}$ of chirplets along a path in the graph obeying

$$\|f - \tilde{f}\|^2 \leq C \cdot N^2 m^{-2s},$$

for a constant C depending on R.

As for Theorem 2, the norm is either the L_2-norm or the Euclidian norm defined by $\|f\|^2 = 1/N \cdot \sum_{k=0}^{N-1} |f(k/N)|^2$. Note that the lower bound in condition (6.24) puts a restriction on how well we can approximate chirps from CHIRP(s, N, R) using chirplet paths with constant amplitude chirplets. This problem vanishes if we use amplitude-modulated chirplets, as we can see in the next section. However, for the purpose of estimation in Chapter 5, we will see that constant-amplitude chirplets suffice for estimating the chirps in the class CHIRP(s, N, R).

Approximating chirps of a lower oscillation degree: We also have similar approximation bounds for chirps in CHIRP(s, λ, R):

Theorem 3. *Suppose $f \in$ CHIRP(s, λ, R) and consider the chirplet graph $\mathcal{G}_N$ with a uniform partition of the time interval $[0, 1]$ such that each interval is of length $|I|$, the slope and offset parameters are discretized as in Lemma 3, and*

$$C \cdot \lambda \cdot |I|^s \leq 1. \tag{6.25}$$

Then there exists a chirplet path W in the chirplet graph $\mathcal{G}_N$ and a family of complex scalars $(\alpha_v)_{v \in W}$ such that the linear combination of chirplets $\tilde{f} = \sum_{v \in W} \alpha_v c_v$, satisfies:

$$\|f - \tilde{f}\|^2 \leq \min(\|A\|^2, R^2 |I_v|^2) + \|A\|^2 \cdot C^2 \cdot \lambda^2 \cdot |I_v|^{2s},$$

where the constant $C = C(R)$ is as in Lemma 3. The norm is either the L_2-norm or the Euclidian norm defined by $\|f\|^2 = 1/N \cdot \sum_{k=0}^{N-1} |f(k/N)|^2$.

The proof is only a minor modification of the proof for Theorem 2 and will be omitted.

6.2.4 Approximation using amplitude-modulated chirplets

Below we will give bound on the accuracy of approximating chirps from CHIRP(s, N, R) using amplitude-modulated chirplets. Let $f \in$ CHIRP(s, N, R) be a complex-valued chirp such that

$$f(t) = A(t) \cdot e^{iN\varphi(t)},$$

where $A, \varphi \in \text{HÖLDER}^s(R)$ and $0 < A < R$.

Consider the same chirplet graph as was used in Section 6.2.3 for the approximation of chirps using constant-amplitude chirplets. Define $\mathcal{C}$ to be the set of functions of the form

$$\tilde{f}(t) = \sum_{v \in W} p_v(t) c_v(t)$$

where $c_v(t) = e^{i\theta_v(t)} 1_{I_v}(t)$ is an unnormalized chirplet, $p_v(t)$ is a smooth parametric function belonging to some class, S_v, of smooth functions supported on I_v, and W is a chirplet path in the chirplet graph. For our purposes, $p_v(t)$ is a polynomial of degree at most 2, so $\tilde{f}(t) = \sum_{v \in W} (\alpha_{v,0} + \alpha_{v,1} t + \alpha_{v,2} t^2) c_v(t)$ where $\{\alpha_{v,l}, l = 0, 1, 2\}$ is a set of complex scalars.

It will become apparent shortly the reason why these coefficients are chosen to be complex and not real-valued, but one of the reasons is that it allows for local fitting of the phase offset of the chirp. Let $\{b_{v,1},\ldots,b_{v,k}\}$ be an orthonormal basis for S_v so the members of the class $\mathcal{C}$ are functions

$$\tilde{f}(t)=\sum_{v\in W}\sum_{l=1}^{k}\alpha_{v,l}b_{v,l}(t)c_v(t). \tag{6.26}$$

Denote the set of all possible amplitude-modulated chirplets by

$$\mathcal{M}=\{b_l(t)c(t): l=1,\ldots,k, c(t) \text{ is a member of our chirplet dictionary}\}.$$

Denote the number of elements in $\mathcal{M}$ by M_n (it is equal to k times the number of elements in the chirplet dictionary).

For approximating f we will use a linear combination of amplitude-modulated chirplets as in (6.26). That is, we will consider approximating functions of the form

$$\tilde{f}(t)=\sum_{v\in W}\sum_{l=1}^{k}\alpha_{v,l}b_{v,l}(t)c_v(t).$$

Take the phase function $\theta:=\sum_{v\in W}\theta_v(t)$ in $\tilde{f}$ as fixed. Since the minimum squared error can be decomposed as

$$\min_{(p_v)}\|f-\tilde{f}\|^2=\sum_{v\in W}\min_{p_v}\|Ae^{iN\varphi}-p_ve^{iN\theta_v}\|_{I_v}^2,$$

we will focus our attention on the approximation error on one of the intervals, I_v. Write

$$\delta(t):=\varphi(t)-\theta(t).$$

Restricted to the inverval I_v, the approximation is of the form $\tilde{f}(t)=\sum_{l=1}^{k}\alpha_l b_l(t)c_v(t)$. Since $\{b_1,\ldots,b_k\}$ are orthonormal, we get

$$\begin{aligned}\min_{p_v}\|Ae^{iN\varphi}-p_ve^{iN\theta_v}\|_{I_v}^2 &= \min_{p_v}\|Ae^{iN\delta}-p_v\|_{I_v}^2=\min_{(\alpha)}\|Ae^{iN\delta}-\sum_{l=1}^{k}\alpha_l b_l\|_{I_v}^2\\ &= \|A\|_{I_v}^2-\sum_{l=1}^{k}|\langle Ae^{iN\delta},b_l\rangle_{I_v})|^2.\end{aligned}$$

If we had chosen the α_ls to be real-valued, we would have gotten $Re(\langle Ae^{iN\delta}, b_l\rangle_{I_v}))^2$ instead of $|\langle Ae^{iN\delta}, b_l\rangle_{I_v})|^2$. But because of the absolute value, the terms in the sum do not change if we exchange $e^{iN\delta}$ for $e^{i(N\delta+\theta_0)}$ for any $\theta_0 \in \mathbb{R}$. Therefore, we can assume that $\delta = \varphi - \theta$ satisfies the inequality (6.4) in Lemma 3. Let $\bar{\delta}$ be the upper bound as defined in (6.20) and assume that

$$\bar{\delta} \leq N^{-1}.$$

Define the constant $C = C(R)$ to be as in Lemma 3. Using again the elementary inequalities $Re(z) \leq |z|$, for all $z \in \mathbb{C}$, and $\cos(x) \geq 1 - x^2/2$, for all $x \in \mathbb{R}$, we get

$$\begin{aligned}
\min_{p_v} \|Ae^{iN\varphi} - p_v e^{iN\theta_v}\|^2_{I_v} &\leq \|A\|^2_{I_v} - \sum_{l=1}^{k} [\langle A\cos(N\delta), b_l\rangle_{I_v}]^2 \\
&\leq \|A\|^2_{I_v} - \sum_{l=1}^{k} [\langle A(1 - N^2\bar{\delta}^2/2), b_l\rangle_{I_v}]^2 \\
&= \|A\|^2_{I_v} - (1 - N^2\bar{\delta}^2/2)^2 \sum_{l=1}^{k} [\langle A, b_l\rangle_{I_v}]^2 \\
&\leq \|A\|^2_{I_v} - (1 - N^2\bar{\delta}^2) \sum_{l=1}^{k} [\langle A, b_l\rangle_{I_v}]^2 \\
&= \|A\|^2_{I_v} - \sum_{l=1}^{k} [\langle A, b_l\rangle_{I_v}]^2 + N^2\bar{\delta}^2 \sum_{l=1}^{k} [\langle A, b_l\rangle_{I_v}]^2.
\end{aligned}$$

The first two terms are simply the squared error of approximating $A(t)$ with its projection onto S_v; i.e.,

$$\|A\|^2_{I_v} - \sum_{l=1}^{k} [\langle A, b_l\rangle_{I_v}]^2 = \min_{p_v} \|A - p_v\|^2_{I_v}. \tag{6.27}$$

Since $\min_{p_v} \|A - p_v\|^2_{I_v} \geq 0$, this also allows us to bound the last term as follows:

$$N^2\bar{\delta}^2 \sum_{l=1}^{k} [\langle A, b_l\rangle_{I_v}]^2 \leq N^2\bar{\delta}^2 \|A\|^2_{I_v} \leq N^2 \cdot C^2 \cdot |I_v|^{2s} \cdot \|A\|^2_{I_v}.$$

Let $I_v = [t_0, t_1]$ and let $m \in \mathbb{Z}$ such that $m < s \leq m + 1$. Then, according to Lemma 16 in Appendix C.1, we can write

$$A(t) = \sum_{k=0}^{m-1} A^{(k)}(t_0)(t - t_0)^k + \epsilon(t), \quad \forall t \in I_v$$

where $|\epsilon(t)| \leq K \cdot |I_v|^s$ for a an integer $K \leq \|A\|_s \leq R$. Since A is twice differentiable we can bound the error in (6.27) by replacing p_v with $\sum_{k=0}^{m-1} A^{(k)}(t_0)(t-t_0)^k$, and get

$$\min_{p_v} \|A - p_v\|_{I_v}^2 \leq \|\epsilon\|_{I_v}^2 \leq R^2 \cdot |I_v|^{2s+1}. \tag{6.28}$$

We also have the trivial bound

$$\min_{p_v} \|A - p_v\|_{I_v}^2 \leq \|A\|_{I_v}^2.$$

Putting everything together gives

$$\min_{p_v} \|Ae^{iN\varphi} - p_v e^{iN\theta_v}\|_{I_v}^2 \leq \min(\|A\|_{I_v}^2, R^2|I_v|^{2s+1}) + N^2 \cdot C^2 \cdot |I_v|^{2s} \|A\|_{I_v}^2,$$

and therefore

$$\min_{(p_v)} \|f - \tilde{f}\|^2 \leq \min(\|A\|^2, R^2|I_v|^{2s}) + \|A\|^2 \cdot C^2 \cdot N^2 \cdot |I_v|^{2s}.$$

Thus we have proved the theorem:

Theorem 4. *Suppose $f \in CHIRP(s, N, R)$ such that $f(t) = A(t)\exp(iN\varphi(t))$ with $A, \varphi \in HÖLDER^s(R)$ and $0 < A < R$. Consider the chirplet graph $\mathcal{G}_N$ with a uniform partition of the time interval $[0,1]$ such that each interval is of length $|I|$, the slope and offset parameters are discretized as in Lemma 3, and*

$$C \cdot N \cdot |I|^s \leq 1. \tag{6.29}$$

Then there exists a chirplet path W in the chirplet graph $\mathcal{G}_N$ and a set of second-order polynomials $\{p_v, v \in W\}$ such that the combination of amplitude modulated chirplets $\sum_{v \in W} p_v(t) c_v(t)$ satisfies:

$$\|f - \tilde{f}\|^2 \leq \min(\|A\|^2, R^2|I_v|^{2s}) + \|A\|^2 \cdot C^2 \cdot N^2 \cdot |I_v|^{2s},$$

where the constant $C = C(R)$ is as in Lemma 3. The norm is either the L_2-norm or the Euclidian norm defined by $\|f\|^2 = 1/N \cdot \sum_{k=0}^{N-1} |f(k/N)|^2$.

We can look at the right-hand side of the inequality in Theorem 4 as a separation of the approximation error into two terms:

- Approximation error due to the local fitting of amplitude, $A(t)$.
- Approximation error due to the oscillating part of the chirp, $\exp(i\varphi(t))$.

It is clear that the latter error dominates the first one; i.e., by how well we can approximate the instantaneous frequency of the chirp. If we instead consider chirps from the class CHIRP(s, λ, R), the approximation error obeys,

$$\min_{(p_v)} \|f - \tilde{f}\|^2 \leq \min(\|A\|^2, R^2|I_v|^{2s}) + \|A\|^2 \cdot C(R) \cdot \lambda^2 \cdot |I_v|^{2s},$$

where we would require $\bar{\delta} \leq \lambda^{-1}$. If we let $\lambda \to \infty$ as the sample size N increases, we also have that error due to local fitting of amplitude being dominated by the latter term. Otherwise if $\lambda = O(1)$, the two errors are of the same order.

6.2.5 Approximation of real-valued chirps

Consider the case of real-valued chirps

$$f(t) = A(t)\cos(N\varphi(t)),$$

where $A, \varphi \in \text{HÖLDER}^s(R)$. Let $\tilde{f}$ be a linear combination of chirplets satisfying the conditions in Theorem 2 for the complex-valued chirp $A(t)\exp(iN\varphi(t))$. By the triangle inequality, we get

$$\begin{aligned} \|f - 1/2(\tilde{f} + \tilde{f}^*)\| &\leq \frac{1}{2} \cdot \|A\exp(iN\varphi) - \tilde{f}\| + \frac{1}{2} \cdot \|A\exp(-iN\varphi) - \tilde{f}^*\| \\ &= \|A\exp(iN\varphi) - \tilde{f}\|, \end{aligned}$$

and therefore the same type of upper bound holds for the approximation of real-valued chirps.

Chapter 7

Theoretical Performance of the Best Path Estimator

In Chapter 5 we introduced a flexible procedure for estimating chirps from noisy measurements. Thanks to the underlying graph structure, we have fast network flow algorithms at our disposal which make the estimator rapidly computable. Since results from preliminary numerical experiments are promising, we believe this methodology has a potential of being useful for practical purposes. The purpose of this chapter is to demonstrate that this estimation procedure has very good theoretical performance and possesses optimality properties, at least in the case of white noise.

We will focus on the class $\mathrm{CHIRP}^s(R)$ of chirps defined in Chapter 6 and introduced in [20]. Assuming the smoothness parameter s is unknown but restricted to the interval $[2, 3]$, our estimator is near-optimal over this class of chirps in the presence of additive Gaussian white noise; it comes within a factor $\log(N)$ within the worst-case mean-squared error, where N is the length of the signal. If we assume the smoothness parameter to be known, the estimator is essentially optimal in the sense that it comes within a constant factor times the worst-case mean-squared error. The theoretical results rely on the approximation results from Chapter 6 and on the concentration of measure phenomena for Lipschitz functionals over Gaussian fields.

7.1 Preliminaries from Statistical Decision Theory

We will take the decision theoretical approach to quantify the performance of estimators for the problem of recovering functions $f \in \mathcal{F}_N$ from sampled data

$$y_k = f_k + z_k, \quad k = 0, 1, \ldots, N-1; \tag{7.1}$$

where $f_k = f(k/N)$ and z_k is a stochastic sequence of zero mean and with known distribution. Our goal is to minimize the error of the estimation as measured by the average squared error loss

$$MSE(\hat{f}, f) = E\left[\frac{1}{N}\sum_k (f_k - \hat{f}_k)^2\right] \tag{7.2}$$

where N is the number of samples and $\hat{f}$ is the estimator of f. This *risk* depends on the true signal which is unknown. To control the risk for any $f \in \mathcal{F}_N$ we wish to minimize the maximum risk:

$$R(\mathcal{F}_N, \hat{f}) = \sup_{f \in \mathcal{F}_N} MSE(\hat{f}, f).$$

This gives us a mathematical way to compare estimators quantitatively. We say the estimator $\hat{f}$ is better than the estimator $\tilde{f}$ in the *minimax* sense if

$$R(\mathcal{F}_N, \hat{f}) \leq R(\mathcal{F}_N, \tilde{f}).$$

The *minimax risk* $R^*(\mathcal{F}_N)$ is the lower bound for the risk of all estimators:

$$R^*(\mathcal{F}_N) = \inf_{\hat{f}} \sup_{\mathcal{F}_N} R(\mathcal{F}_N, \tilde{f}).$$

Usually it is hard to find estimators that attain this minimax risk exactly, and this is certainly not trivial in our estimation problem. Besides that, finding out what $R^*(\mathcal{F}_N)$ actually is can be quite challenging. Instead, one usually tries to establish a lower bound on this risk. A common technique for doing this – which we will describe in the next subsection– is based on an important result from decision theory. It compares the risks from the minimax and Bayesian viewpoints.

7.1.1 Bounding the minimax risk using bayes priors

A Bayesian approach to the estimation problem assumes that the unknown function f is a random variable taking values in the set $\mathcal{F}_N$. Let π be the distribution of the signals f drawn from $\mathcal{F}_N$. Then the *Bayes risk* of an estimator $\hat{f}$ for the mean squared error risk is defined to be the expected risk

$$R(\hat{f}, \pi) = E_\pi \left[MSE(\hat{f}, f) \right],$$

and the minimum Bayes risk is defined by:

$$B(\pi) := \inf_{\hat{f}} R(\hat{f}, \pi).$$

The *Bayes estimator* $\tilde{f}$, which yields the minimum Bayes risk in the case of a squared error loss, is the conditional expectation of the randomly drawn function f given the data y; also called the *posterior mean*:

$$\tilde{f} = E_\pi(f \mid y).$$

The Bayes risk for squared error loss is the posterior variance of f given y.

The next theorem relates the minimax risk and the minimum Bayes risk. Its proof can be found in [57]:

Theorem 5. *For any choice of prior π obeying $\pi(\mathcal{F}_N) = 1$,*

$$R^*(\mathcal{F}_N) \geq B(\pi).$$

A distribution π which satisfies $B(\pi) = R^*(\mathcal{F}_N)$ is called a *least favorable prior* distribution. A technique for bounding the minimax risk by the help of this theorem could be to seek a prior π that is close to being a least favorable prior, and hopefully simple enough so we can evaluate the Bayes risk (or at least establish a good lower bound on it). Once we have a good lower bound on the minimax risk, we try to establish a good upper bound on the maximum risk of our estimation procedure, ideally one of the same order as the lower bound on the best risk we could achieve. An estimator that is of the same order as the minimax risk is said to be *optimal*, and an estimator which achieves it within a logarithmic factor is said to be *near-optimal*. For further details about this decision theoretical framework, see

[49, 68] and perhaps [57].

7.2 The Data Model for Chirp Estimation

Our theory will assume the data is of the form (7.1), where the class of unknown signals is $\mathcal{F} = \mathrm{CHIRP}^s(R)$ where $s \in [2,3]$. The noise vector $z = (z_k)$ is assumed to be of the form $z = z^0 + \imath z^1$ where z^0 and z^1 are two independent vectors of i.i.d. $N(0, 1/2)$ variables.

The strategy described in Section 7.1.1 is precisely the one we will use to show that our method for estimating chirps is optimal when the regularity parameter $s \in [2,3]$ is known, and near-optimal when it is unknown. A good lower bound on the minimax risk has in fact been already established by Candès in [20], and therefore our main work is to get a good upper bound for our estimator.

7.3 A Lower Bound on the Minimax Risk for Chirp Estimation

Here is the lower bound for estimating chirps that Candès established in [20]:

Theorem 6. *If $s \in [2,3]$, then for a constant C,*

$$M^*(N, \mathcal{F}_N) := \inf_{\hat{f}} \sup_{\mathcal{F}_N} MSE(\hat{f}, f) \geq C \cdot N^{\frac{-2(s-1)}{2s+1}}. \tag{7.3}$$

For completeness, we include a detailed proof of this theorem in Appendix C.5 which simply reproduces Candès' arguments.

7.4 An Upper Bound on the Risk of the BP estimator

Recall the definition of the complexity functional $K(\tilde{f}, f)$ from (5.6) and the definition of the class of functions $\mathcal{C}$ from Section 5.2:

$$K(\tilde{f}, f) = \|\tilde{f} - f\|_2^2 + \Lambda(\tilde{f}),$$

and $\mathcal{C}$ is the set of linear combinations of chirplets such that

$$\tilde{f} = \sum_{v \in W} \alpha_v c_v, \qquad \tilde{f} = \sum_{v \in W} \frac{1}{2}(\alpha_v c_v + \alpha_v^* c_v^*),$$

in the case of complex-valued or real-valued data, respectively. $\{\alpha_v\}$ is a set of complex scalars and W is any chirplet path in the chirplet graph. For our theoretical considerations, we will take

$$\Lambda(\tilde{f}) = \lambda^2 N(\tilde{f}), \quad N(\tilde{f}) := |W|,$$

with

$$\lambda^2 = \eta^2 \cdot (1 + \sqrt{2 \log M_N})^2,$$

for some fixed $\eta > 8$, where M_N is the number of chirplets in the graph.

7.4.1 Discretization for the BP estimator

For our theoretical treatment we need to choose a suitable discretization for the chirplet graph and chirplet parameters. Assume the number of samples is dyadic, $N = 2^J$, for some positive integer J. For the partitions of the time interval $[0, 1)$ we will consider dyadic time intervals of lengths 2^{-j} for $j = \lfloor \log_2(N)/3 \rfloor, \ldots, \lceil \log_2(N)/2 \rceil$, i.e.,

$$j = \lfloor J/3 \rfloor, \ldots, \lceil J/2 \rceil.$$

On a dyadic interval I, the slope and frequency offset parameters $a_{\mu,I}, b_{\mu,I}$ for a chirplet as in (2.1) are chosen to be of the form

$$b_{\mu,I} = 2\pi \Delta b \cdot m, \qquad a_{\mu,I} = 2\pi \frac{\Delta b}{|I|} \cdot l, \tag{7.4}$$

where $\Delta b = 2^{-\lceil J/2 \rceil}$. This choice ensures that the condition $\Delta b \leq \inf_{I \in \mathcal{P}}$ holds for any possible partition $\mathcal{P}$ using the set of dyadic intervals we have described. Therefore, the conditions for Lemma 3 hold. As a consequence our dyadic intervals $|I|$ satisfy

$$C \cdot N^{-1/2} \leq |I| \leq C' \cdot N^{-1/3}.$$

Thus, we have the approximation bound from Corollary 3 at our disposal.

7.4.2 Oracle inequality for the BP estimator

For a fixed signal f, denote the *theoretical complexity* by $K_0 := K(f_0, f)$ where f_0 is the minimizer

$$f_0 = \arg\min_{\tilde{f} \in \mathcal{C}} K(\tilde{f}, f).$$

Denote the *empirical complexity* by $\hat{K} := K(\hat{f}, f)$, where $\hat{f}$ is the BP estimation given the data $y = f + z$. Note that $\hat{K}$ is a random variable. Theorem 7 below, gives us an oracle inequality relating the expected value of the empirical to the theoretical complexity. Theorems of similar forms for complexity functionals can be found in the literature (see, for example, [20, 21, 34]).

Theorem 7. *Suppose $y = f + z$, where $f \in$ CHIRP and $z = (z_k)$ is a vector of i.i.d. standard Gaussian, either complex-valued so that $z_k = (z_k^1 + iz_k^2)/\sqrt{2}$ with z_k^1 and z_k^2 i.i.d. $N(0,1)$, or real-valued so that $z_k \sim N(0,1)$. Select $\lambda^2 = \eta^2 \cdot (1 + \sqrt{2 \log M_N})^2$ with $\eta > 8$. Let $\hat{f}$ be the minimizer of the empirical complexity $K(\tilde{f}, y)$, and K_0 be the minimum theoretical complexity. Then,*

$$E\hat{K} \leq C(\eta) \cdot \left(\lambda^2 + K_0\right),$$

where $C(\eta) = 2e \cdot (1 - 8/\eta)^{-1}$.

See Appendix C.6 for the proof of this theorem.

7.4.3 Bounding the ideal risk

Let f be the unknown chirp and let W be any chirplet path in the graph. Denote the projection of a vector y on the linear span of $\{c_v : v \in W\}$ by

$$P_W y := \sum_{v \in W} \langle y, c_v \rangle c_v.$$

From our discussion in Section 5.1, we have a bias-variance decomposition of the risk $R(W)$ of estimating f using the projection P_W:

$$R_W(f) = E\|f - P_W y\|^2 = \|f - P_W f\|^2 + |W|.$$

Denote the *ideal risk* of estimating f by linear combinations of chirplets along paths in the graph by

$$R_{IDEAL}(f) := \min_{W} R_W(f).$$

Corollary 3 provides us with an upper bound on R_{IDEAL}. In our graph there is a chirplet path W_m^* of length m, for the uniform partition with interval lengths $|I| \sim m^{-1}$, such that

$$R_{IDEAL}(f) \leq C(R) \cdot N^3 \cdot m^{-2s} + m,$$

where $C(R)$ is a constant of the same size as in the corollary. This inequality gives us the best bound when choosing the partition such that

$$m \sim C(R) \cdot N^3 \cdot m^{-2s}, \quad \text{or} \quad m \sim (C(R))^{1/(2s+1)} \cdot N^{3/(2s+1)}.$$

Therefore

$$R_{IDEAL}(f) \leq ((C(R))^{1/(2s+1)} + 1) \cdot N^{3/(2s+1)}, \tag{7.5}$$

where $C(R)$ includes a small correction factor of order 1.

7.4.4 The upper bound

Using the oracle inequality, we can show that our best path estimator (5.7) nearly achieves the lower bound in Theorem 6:

Theorem 8. *Assume the degree of regularity* $s \in [2, 3]$, *but it is otherwise unknown. Then the best path estimator* (5.7) *obeys the inequality*

$$\sup_{f \in \mathcal{F}_N} MSE(f, \hat{f}) \leq C(R) \cdot \log N \cdot N^{-\frac{2(s-1)}{2s+1}},$$

for a constant $C(R)$.

Proof of Theorem 8. The proof is based on the complexity bound from Theorem 7 and the upper bound on the ideal risk as in (7.5). Pick a signal $f \in \mathcal{F}_N$. Since

$$\|f - \hat{f}\|^2 \leq \|f - \hat{f}\|^2 + \lambda^2 \cdot N(\hat{f}) = \hat{K},$$

we have

$$\begin{aligned} E\|f-\hat{f}\|^2 &\leq C(\eta)\cdot\left(\lambda^2+\min_{\tilde{f}\in\mathcal{C}}\left(\|\tilde{f}-f\|_2^2+\lambda^2\cdot N(\tilde{f})\right)\right) \\ &\leq C(\eta)\cdot 2\lambda^2\cdot\min_{\tilde{f}\in\mathcal{C}}\left(\|\tilde{f}-f\|_2^2+N(\tilde{f})\right) \\ &= C(\eta)\cdot 2\lambda^2\cdot R_{IDEAL}(f) \\ &\leq C(R)\cdot\log N\cdot N^{3/(2s+1)}. \end{aligned}$$

Here we have fixed $\eta > 8$ and absorbed all the constants and the term $(C(R))^{1/(2s+1)}+1)$ from the bound on the ideal risk into the constant $C(R)$. The logarithmic term comes from the fact that the number of elements in $\mathcal{M}$ is $M_N = O(N^\alpha)$ for a positive constant α, and therefore $\lambda^2 = \eta^2\cdot(1+\sqrt{2\log M_N})^2 = O(\log N)$. Thus,

$$MSE(f,\hat{f}) = E\left[1/N\cdot\|f-\hat{f}\|^2\right] \leq C(R)\cdot\log N\cdot N^{-2(s-1)/(2s+1)}.$$

□

We can interpret the logarithmic term in the bound as a "price for adaptivity" we need to pay for not knowing the regularity s of the unknown chirp. Presently it is not known whether it is possible to get away with paying such a factor and it remains to either (i) get a better lower bound, possibly involving a logarithmic factor, and/or (ii) show there is an adaptive estimation procedure achieving the existing – or a better – lower bound.

We would like to point out that the chirp estimation procedure Candés proposed in [20] (which is described in Section 5.6) has an upper bound of the same rate when the regularity of the chirp is assumed to be unknown. Therefore we have not shown that the BP estimator is better than thresholding in the best chirplet tight frame (TBCF) for this class of chirps. Also, as pointed out in Chapter 5, we can easily think of situations in practice where the TBCF estimator would be better suited than the BP estimator; in particular when there is more than one chirp present at the same time in the data. However, these theoretical results give a good supplement to the methodology and numerical results presented in Chapter 5, and give a quantitative benchmark for other chirp estimation methods to compare to.

7.5 The BP Estimator is Optimal When Regularity is Known

We finish this chapter by showing that the BP estimator is optimal if we assume further information about the unknown chirp. Consider the same data model as described in Section 7.2 and assume this time that the regularity $s \in [2, 3]$ is known to the scientist. Then we have a sharp bound upper bound on the risk of the BP estimator which achieves the lower bound on the minimax risk within a constant:

Theorem 9. *Assume the degree of regularity $s \in [2, 3]$ and is known. Then there is a monoscale chirplet graph with uniform partition such that every chirplet path is of length L. Then for every chirp $f \in CHIRP^s(R)$, the best path estimator (5.7) obeys*

$$E\|f - \hat{f}\|^2 \leq \left(1 + \frac{\exp(-\gamma L)}{(1 - \exp(-\gamma L))^3}\right)(1 - 8/\eta)^{-1}\lambda^2 L,$$

and therefore, by taking $L \sim N^{3/(2s+1)}$,

$$\sup_{f \in \mathcal{F}_N} MSE(f, \hat{f}) \leq C \cdot N^{-\frac{2(s-1)}{2s+1}},$$

for a suitable constant C.

For the proof of this theorem see Appendix C.7.

Chapter 8

Bounds on Statistical Performance of the Best Path Test

Here we will identify rates for the signal strength where the BP test is guaranteed to be asymptotically powerful. The signals we will consider are chirps belonging to the class CHIRP(s, N, R). We will not provide rates for the lower bound –i.e., situations for which any sequence of tests is asymptotically powerless– and therefore we cannot claim that guaranteed rates are the limits of performance in this problem. To move towards deeper understanding of our methodology, we will try to connect the BP test and the chirplet graph with an underlying abstract statistical problem. This abstraction is the problem of detecting paths of "unusual behavior" in a graph $\mathcal{G}$, where a random variable X_v is associated with each vertex $v \in \mathcal{G}$. In such problems, the typical situation could be such that all the X_vs have the same distribution F_0, and the goal is to decide whether there is a set of connected vertices in $\mathcal{G}$ whose distribution is different from F_0. The abstract problem will be studied further in Chapter 9.

8.1 An Upper Bound for Detection of Chirps of Known Regularity

We consider the detection of chirps from CHIRP(s, N, R) and assume the regularity parameter $s \in [2, 3]$ to be known to the scientist. Let the data be

$$y_k = \alpha_N f_k + z_k, \qquad k = 0, 1, \ldots, N-1; \tag{8.1}$$

where $f_k = f_c(k/N)$ is a vector of equispaced time samples of a function $f_c \in \text{CHIRP}(s, N, R)$, and where $z = (z_k)$ is a sequence random variables such that $z = z^0 + iz^1$ where z^0 and z^1 are two independent vectors of i.i.d. $N(0, 1/2)$ variables. Assume the vector $f = (f_k)$ has unit-norm, $\|f\| = 1$. Consider the sequence of hypothesis tests $(H_{0,N})$ versus $(H_{1,N})$, where we wish to test the null hypothesis

$$H_{0,N} : \alpha_N = 0$$

against

$$H_{1,N} : \alpha_N \neq 0.$$

Given a sequence of BP tests (T_N) that we are about to describe, we want to find a sufficient rate of growth for α_N such that the tests are asymptotically powerful for this problem; i.e., such that

$$P_{H_{0,N}}\{T_N \text{ rejects } H_0\} + P_{H_{1,N}}\{T_N \text{ accepts } H_0\} \to 0.$$

The lower bound consideration, which we will not study, would be to find rates of growth for α_N such that

$$P_{H_{0,N}}\{T_N \text{ rejects } H_0\} + P_{H_{1,N}}\{T_N \text{ accepts } H_0\} \to 1,$$

as $N \to \infty$.

8.1.1 Discretization and connectivities

For a fixed signal length N, we pick a uniform partition $\mathcal{P}_N$ of $[0, 1)$ such that every interval has length $|I|$. For the slope and phase offset parameters in the chirplet dictionary we will use the same discretization as described in Lemma 3. For the connectivities we will consider the monoscale chirplet graph described in Section 6.2.1.2, where two chirplets are connected if and only if their time supports are adjacent, and if their instantaneous frequencies coincide at their juncture. According to Section 6.2.1.2, we can let every chirplet in this graph connect to only a constant number of chirplets and the approximation bound in Lemma 3 would still hold. Denote this chirplet graph by $\mathcal{G}_N = \mathcal{G}_N(L_N)$, where $L_N = |I|^{-1}$ is the number of chirplets in each path. Then the number of paths in $\mathcal{G}_N$ does not exceed

$$p_N \cdot e^{\gamma L_N},$$

for a suitable constant $\gamma > 0$, independent of the sample size N, and $p_N = o(N^\beta)$, for some $\beta > 0$ independent of N; indeed, once we have chosen a starting chirplet, we have $L_N - 1$ steps where we choose from a constant number of chirplets to connect to. Finally, the test we wish to consider is the *monoscale BP test*

$$T_N^* = T_N^*(|I|) = \max_{W \in \mathcal{G}_N} \sum_{v \in W} |\langle y, c_v \rangle|^2, \tag{8.2}$$

where the maximum is taken over all the chirplet paths in $\mathcal{G}_N$.

8.1.2 Behaviour of the monoscale BP test under H_0

The following theorem gives a bound on T_N^* when the data is pure noise:

Theorem 10. *Pick $\eta > 0$. Then,*

$$P_{H_0}\left(T_N^* > \left(1 + \sqrt{2(\gamma + \eta)}\right)^2 \cdot L_N\right) \le 2p_N \cdot \exp(-\eta \cdot L_N),$$

for any monoscale chirplet graph $\mathcal{G}_N$ which has total number of chirplet paths less than or equal to $K_N = p_N e^{\gamma L_N}$.

Proof. Pick a chirplet graph $\mathcal{G}_N$ which has total number of chirplet paths less than or equal to $K_N = p_N e^{\gamma L_N}$. Let W be a path in the graph and let P_W be the projection of the data onto the span of $\{c_v : v \in W\}$. Since the c_vs are orthonormal, the projection of the data y is,

$$P_W y = \sum_{v \in W} \langle y, c_v \rangle c_v.$$

Note that

$$\|P_W y\|_2^2 = \sum_{v \in W} |\langle y, c_v \rangle|^2, \tag{8.3}$$

which is exactly the sum of the chirplet costs along the chirplet path W.

Observe that $\|P_W z\|_2$ is a Lipschitz functional on a Gaussian field with Lipschitz constant 1 (see the proof of Lemma 22). Note that $E\|P_W z\|_2^2 = |W| = L_N$. Then we can use

Lemma 28 to bound the sum of the costs along a chirplet path under H_0: for any $t > 0$,

$$P\left(\|P_W z\|_2 \geq t + \sqrt{|W|}\right) \leq 2\exp(-t^2/2).$$

Pick $\eta > 0$ and let $t = \xi\sqrt{L_N}$ where

$$\xi = \sqrt{2 \cdot (\gamma + \eta)}.$$

Then, using identity (8.3), we have

$$P\left(\sum_{v \in W} |\langle z, c_v \rangle|^2 \geq (1+\xi)^2 \cdot L_N\right) \leq 2\exp(-\xi^2 \cdot L_N/2),$$

and the result follows from a simple union bound:

$$\begin{aligned} P\left(\max_W \left[\sum_{v \in W} |\langle z, c_v \rangle|^2\right] \geq (1+\xi)^2 \cdot L_N\right) &\leq K_N \cdot 2\exp(-\xi^2 \cdot L_N/2) \\ &\leq 2p_N \cdot \exp((\gamma - \xi^2/2) \cdot L_N) \\ &= 2p_N \cdot \exp(-\eta \cdot L_N). \quad \Box \end{aligned}$$

If $L_N \sim N^\xi$ and $p_N = o(N^\beta)$ for some $\xi, \beta > 0$, this theorem provides us with a useful bound for studying the asymptotics of the test as $N \to \infty$. It suggests we should compare T_N^* to a threshold $\tau_N \sim L_N$. Below are a couple of observations regarding Theorem 10 and its proof we would like to mention:

1. Theorem 10 holds also in the case of real-valued data $y = f + z$, where z is a sequence of i.i.d. standard Gaussians.

2. The arguments in the proof of Theorem 10 do not depend on the partition of the time interval $[0, 1)$ being uniform – $\|P_W z\|$ would still be a Lipschitz functional on a Gaussian field with Lipschitz constant 1 if the partition is non-uniform.

3. The theorem also holds in the case of colored noise and we define the chirplet costs as

$$|\langle \Sigma^{-1} y, c_v \rangle|^2 = |y^* \Sigma^{-1} c_v|^2,$$

where Σ is the covariance matrix of the noise, and y and c_v are written as column

> vectors. In this case the chirplets would be normalized such that
>
> $$c_v^* \Sigma^{-1} c_v = 1.$$
>
> With this choice of inner product, $E|\langle z, c_v \rangle|^2 = 1$, and therefore $E\|P_W z\|_2^2 = |W|$ for any chirplet path W, just as before. The norm of the projector, $\|P_W \cdot\|_2$, is still a Lipschitz(1) functional. We could even extend the result to amplitude-modulated chirplets. It would give us the same rate but different constants.

We could have used bounds for moderate deviations for χ^2 to get a similar result as in Theorem 10, but that would only work in the case when the noise sequence $z = (z_k)$ is complex-valued i.i.d. standard Gaussian (i.e., $z = z^0 + iz^1$ where z^0 and z^1 are two independent vectors of i.i.d. $N(0, 1/2)$ variables). In the case of real-valued Gaussian noise, the sum of the chirplet costs along a path would not necessarily be distributed as a χ^2 random variable.

8.1.3 Bounding the monoscale BP test under H_0 using moderate deviations for χ^2

Here we provide an alternative way of bounding the monoscale BP statistic in the case of complex-valued noise. We also give a lower bound on T_N^* which tells us that L_N is the correct rate the threshold T_N^* should be compared to. The bound is based on the following lemma which can be found in [55]:

Lemma 7. *Let $W_d \sim \chi_d^2$ be distributed as a chi-squared random variable with d degrees of freedom. Then for each $t > 0$,*

$$P(W_d - d \geq t\sqrt{2d} + t^2) \leq e^{-t^2/2} \quad \textit{and} \quad P(W_d - d \leq -t\sqrt{2d}) \leq e^{-t^2/2}.$$

Let (c_v) be a sequence of chirplets along a chirplet path W. Let $z = (z_k)$ be a complex-valued random sequence such that $z = z^0 + iz^1$ where z^0 and z^1 are two independent vectors of i.i.d. $N(0, 1/2)$ variables. Then, according to Lemma 29,

$$\langle z, c_v \rangle = U_v^1 + iU_v^2$$

where (U_v^1) and (U_v^2) are independent sequences of i.i.d. $N(0, 1/2)$ random variables. Then it is clear that

$$2|\langle z, c_v \rangle|^2 \sim \chi_2^2$$

and

$$2 \sum_{v \in W} |\langle z, c_v \rangle|^2 \sim \chi_{2L}^2,$$

where we have written $L = |W|$. Assume there are less than $K_N \cdot e^{\gamma L}$ chirplet paths of length L. Take some $\eta > 0$. Then the first inequality of Lemma 7 with $t = \sqrt{2(\gamma + \eta)L}$ gives

$$P\left(2 \sum_{v \in W} |\langle z, c_v \rangle|^2 \geq 2L + 2\sqrt{2(\gamma + \eta)}L + 2(\gamma + \eta)L \right) \leq \exp(-(\gamma + \eta)L),$$

or

$$P\left(\sum_{v \in W} |\langle z, c_v \rangle|^2 \geq CL \right) \leq \exp(-(\gamma + \eta)L)$$

where $C = 1 + (\gamma + \eta) + \sqrt{2(\gamma + \eta)}$. Finally, by using a union bound, we have:

Lemma 8. *Pick $\eta > 0$. Then there is a constant $C > 0$, such that*

$$P_{H_0}\left(T_N^* \geq CL_N\right) \leq p_N e^{-\eta L_N},$$

for any monoscale chirplet graph $\mathcal{G}_N$ which has total number of chirplet paths less than or equal to $K_N = p_N e^{\gamma L_N}$, where $P_N = o(N^\beta)$ for some $\beta > 0$. C can taken to be $C = 1 + (\gamma + \eta) + \sqrt{2(\gamma + \eta)}$.

Notice that the constant here is slightly better than in the bound of Theorem 10. The lemma on moderate deviations of χ^2 variables also gives us, with an overwhelming probability, that T_N^* cannot grow slower than L_N under H_0:

Lemma 9. *For any $t > 0$,*

$$P_{H_0}(T_N^* \leq L_N - t\sqrt{L_N}) \leq e^{-t^2/2}.$$

Proof. Pick any chirplet path W of length L in the graph and let P_W be as defined in the proof of Theorem 10. Then, as we showed before the last lemma, $2\|P_W z\|^2 \sim \chi_{2L}^2$ and the second inequality in Lemma 7 gives us

$$P_{H_0}(\|P_W z\|^2 \leq L - t\sqrt{L}) \leq e^{-t^2/2}$$

for any $t > 0$. On the event $A := \{T^* \leq \tau\}$, the event $B := \{\|P_W z\|^2 \leq \tau\}$ must hold, so $P(A) \leq P(B)$ which gives

$$P_{H_0}(T^* \leq L - t\sqrt{L}) \leq e^{-t^2/2},$$

for any $t > 0$. □

8.1.4 Bounds for the monoscale BP test under H_1

Theorem 11. *Fix $s \in [2, 3]$ and let $\rho_N = N^{\frac{1}{2s}}$. Then there exists a sequence of monoscale BP tests (T_N^*) and a constant $B > 0$, such that for any sequence of signal amplitudes (α_N) satisfying*

$$\lim_{N\to\infty} \alpha_N/\rho_N \geq B,$$

the sequence of tests (T_N^*) is asymptotically powerful for detection problem (8.1)*.*

Proof. Let $f = (f_k)$ be the unknown chirp. We will choose a chirplet graph $\mathcal{G}_N$ as described in Section 8.1.1, but postpone the choice of the interval length $|I|$ for now. By the definition of the statistic T_N^*, we have

$$T_N^* \geq \sum_{v\in W} |\langle y, c_v\rangle|^2,$$

for any path W in the graph. Therefore,

$$P(T_N^* \leq \tau) \leq P\left(\sum_{v\in W} |\langle y, c_v\rangle|^2 \leq \tau\right), \tag{8.4}$$

for any $\tau \in \mathbb{R}$. The next step is to choose a path W that gives a good bound for the right-hand side of this inequality.

Assume that

$$\alpha_N \geq \sqrt{N} R|I|. \tag{8.5}$$

Then, as a result of our choice of chirplet graph, there is a chirplet path W^* in $\mathcal{G}_N$ such that Corollary 2 from Section 6.2.3 holds, i.e., *for every $f \in CHIRP(s, N, R)$* there exists a chirplet path W^* in $\mathcal{G}_N$ such that

$$\sum_{v \in W^*} |\langle f, c_v \rangle|^2 \geq \alpha_N^2 \cdot \left(1 - C^2 \cdot N^2 \cdot |I|^{2s}\right) - NR^2|I|^2,$$

where the constant $C = C(R)$ is as in Corollary 1. Choose the interval length $|I|$ small enough such that

$$C^2 \cdot N^2 \cdot |I|^{2s} = \epsilon,$$

with $0 < \epsilon < 1$, i.e,

$$|I| = \epsilon / C^2 \cdot N^{-1/s}. \tag{8.6}$$

Then,

$$\sum_{v \in W^*} |\langle f, c_v \rangle|^2 \geq \alpha_N^2 \cdot (1 - \epsilon) - NR^2|I|^2.$$

For this choice of interval length, the number of chirplet paths in the graph $\mathcal{G}_N$ is

$$K_N = B \cdot \exp\left(\gamma \cdot N^{1/s}\right),$$

for some positive constants B and γ (depending upon C and ϵ). Also, the number of chirplets in each path is

$$L_N = |I|^{-1} = A \cdot N^{1/s}, \tag{8.7}$$

where $A = C^2/\epsilon$. Pick $\eta > 0$ and choose the sequence of thresholds (τ_N) for the tests T_N^* to be

$$\tau_N = (1 + \xi)^2 L_N, \quad \text{where } \xi = \sqrt{2(\gamma + \eta)}.$$

Then by Theorem 10,

$$P_{H_0}(T_N^* > \tau_N) \to 0, \quad \text{as } N \to \infty.$$

The triangle inequality for vectors in $\mathbb{C}^n$, (i.e., the special case of $p = 2$ and the counting measure in Minkowski's inequality) gives

$$\begin{aligned} \left(\sum_{v \in W^*} |\langle f, c_v \rangle|^2\right)^{1/2} &= \left(\sum_{v \in W^*} |\langle y, c_v \rangle + (-\langle z, c_v \rangle)|^2\right)^{1/2} \\ &\leq \left(\sum_{v \in W^*} |\langle y, c_v \rangle|^2\right)^{1/2} + \left(\sum_{v \in W^*} |\langle z, c_v \rangle|^2\right)^{1/2}. \end{aligned}$$

This, and the bound on the cost along the path W^*, gives

$$\left(\sum_{v\in W^*} |\langle y, c_v\rangle|^2\right)^{1/2} \geq \sqrt{\alpha_N^2 \cdot (1-\epsilon) - NR^2|I|^2} - \left(\sum_{v\in W^*} |\langle z, c_v\rangle|^2\right)^{1/2},$$

provided α_N is big enough such that $\alpha_N^2 \cdot (1-\epsilon) - NR^2|I|^2 \geq 0$. Using (8.4), we get for any $\tau \in \mathbb{R}^+$,

$$\begin{aligned} P(T_N^* \leq \tau) &\leq P\left(\sum_{v\in W^*} |\langle y, c_v\rangle|^2 \leq \tau\right) \\ &\leq P\left(\sqrt{\alpha_N^2 \cdot (1-\epsilon) - NR^2|I|^2} - \left(\sum_{v\in W^*} |\langle z, c_v\rangle|^2\right)^{1/2} \leq \sqrt{\tau}\right) \\ &= P\left(\left(\sum_{v\in W^*} |\langle z, c_v\rangle|^2\right)^{1/2} \geq \sqrt{\alpha_N^2 \cdot (1-\epsilon) - NR^2|I|^2} - \sqrt{\tau}\right). \end{aligned}$$

Pick the constant D big enough so that for all

$$\alpha_N \geq D \cdot \sqrt{\tau_N},$$

the inequality

$$\alpha_N^2 \cdot (1-\epsilon) - NR^2|I|^2 \geq 4\tau_N$$

holds. Such a constant exists, since for $|I| \sim N^{-1/s}$, we have $\tau \sim N^{1/s}$ and

$$NR^2|I|^2 \sim N^{(s-2)/s} = O(N^{1/s})$$

for all $s \in [2,3]$. This also ensures that condition (8.5) is satisfied. This gives,

$$\begin{aligned} P(T_N^* \leq \tau) &\leq P\left(\left(\sum_{v\in W^*} |\langle z, c_v\rangle|^2\right)^{1/2} \geq \sqrt{4\tau_N} - \sqrt{\tau_N}\right) \\ &= P\left(\sum_{v\in W^*} |\langle z, c_v\rangle|^2 \geq \tau_N\right) \to 0, \quad \text{as } N \to \infty. \end{aligned}$$

Thus, since $\tau_N \sim N^{1/s}$, we have that the sequence of tests is asymptotically powerful if $\alpha_N \geq D' N^{1/(2s)}$, where D' is a constant. □

8.2 An Upper Bound on Detection When the Chirp Regularity is Unknown

Similar bound for guaranteed detection holds in the case when the regularity of the chirp is unknown; i.e., the chirp belongs to CHIRP(s, N, R), but we only know that $s \in [2, 3]$. The reasoning we will give can be made more rigorous but instead we will attempt to keep things simple to convey the idea behind the reasoning.

One strategy that works is based on considering a collection of monoscale chirplet graphs which are defined in the same way as earlier, when s was assumed to be known. Let $T^*(L)$ be the value of the best path in the chirplet graph $\mathcal{G}_L$ with a uniform partition of L segments, each of length $|I| = L^{-1}$. The number of chirplet paths in $\mathcal{G}_L$ does not exceed $e^{\gamma_L L}$ for some constant $\gamma_L > 0$ independent of the signal length N. Let $\mathcal{L} = \{L_0, L_0 + 1, \ldots, L_1\}$, where $L_0 = \lfloor N^{1/3} \rfloor$ and $L_1 = \lceil N^{1/2} \rceil$. Define the test statistic

$$T_N^{**} = \max_{L \in \mathcal{L}} \frac{T^*(L)}{L}. \tag{8.8}$$

This is a statistic of the same form as the *Minimum-Cost-To-Time-Ratio* (MCTTR) statistic discussed in the chapter where the BP test was introduced. By Theorem 10, we have

$$P_{H_0}\left(T^*(L)/L > C_\eta(L)\right) = P_{H_0}\left(T^*(L) > C_\eta(L) \cdot L\right) \le 2 \cdot \exp(-\eta \cdot L),$$

for any $\eta > 0$ and a suitable constant $C_\eta(L) > 0$ independent of N. Let $C = \max_L C_\eta(L)$. Then, using the last inequality and a union bound, we get

$$\begin{aligned} P_{H_0}\left(T^{**} > C\right) &\le \sum_{L \in \mathcal{L}} P_{H_0}\left(T^*(L)/L > C\right) \le \sum_{L \in \mathcal{L}} 2 \exp(-\eta \cdot L) \\ &\le 2L_1 \cdot \exp(-\eta \cdot L_0). \end{aligned}$$

Since $L_1 \sim N^{1/2}$ and $L_0 \sim N^{1/3}$, this probability tends to 0 as $N \to \infty$. Therefore, comparing T^{**} to a constant of size at least as big as C, we can ensure the probability of type I error tends to 0 as $N \to \infty$. The next step is to use the same argument as in Theorem 11. If the regularity of the unknown signal is s, we know there is a graph in our collection with partions of length $|I| \approx N^{-1/s}$. For any path W in this graph and a threshold

τ (could be constant),

$$P(T^{**} \leq \tau) \leq P\left(\|P_W y\|^2/|W| \leq \tau\right) = P\left(\|P_W y\|^2 \leq \tau|W|\right). \tag{8.9}$$

Taking τ to be a constant bigger than C, the far right-hand side of (8.9) is of the same form as we had in the proof of Theorem 11. Therefore, analogous arguments can be used here to show that this probability goes to zero for a signal rate $\alpha \sim N^{1/(2s)}$ for the constant threshold τ big enough. Therefore, (8.8) is asymptotically powerful for detecting chirps of unknown smoothness $s \in [2, 3]$, provided that the signal amplitudes (α_N) satisfy $\lim_{N\to\infty} \alpha_N / N^{\frac{1}{2s}} \geq B$ for a suitable constant $B > 0$.

The same statistic would also be asymptotically powerful under these same conditions if we had considered dyadic lengths instead; i.e.,

$$\mathcal{L} = \{2^j : \lfloor N^{1/3} \rfloor \leq 2^j \leq \lceil N^{1/2} \rceil\}.$$

The reason being that similar approximation bounds hold for approximating the chirps with chirplet paths in graphs restricted to this set. Also, the same threshold works since we are considering a subset of the statistics we had before. This would provide us with a test procedure that would be computationally faster, since the number of different graphs to consider would be $O(\log N)$ instead of $O(N^{1/2})$.

Another type of statistic which has the same guaranteed performance, is based on considering chirplet graphs $\mathcal{G}$ with balanced recursive dyadic partitions (BRDPs). It is enough to consider interval lengths satisfying $|I| = 2^{-j}$ such that $\lfloor N^{-1/2} \rfloor \leq 2^{-j} \leq \lceil N^{-1/3} \rceil$. We can discretize the slopes and the offsets of the chirplet dictionary such that the required approximation bounds for the arguments of Theorem 11 are satisfied, and such that the number of chirplet paths of length L in the graph is $e^{\gamma L}$ –thanks to the BRDP condition. Then we define the MCTTR statistic

$$T^{***} = \max_{W \in \mathcal{G}} \frac{\sum_W |\langle y, c_v \rangle|^2}{|W|}. \tag{8.10}$$

Due to the bound on the number of chirplet paths in $\mathcal{G}$ of fixed length L, we can show that T^{***} is greater than some fixed constant with probability going to zero as $N \to \infty$. The argument is similar to the one for (8.8).

8.3 Is the Upper Bound Tight?

At this point it is natural to ask whether $\alpha_N \sim N^{1/(2s)}$ is in fact the optimal rate for asymptotically minimax detection. That is, whether every sequence of tests is asymptotically powerless if

$$\lim_{N\to\infty} \alpha_N/N^{1/(2s)} < C, \quad \text{or} \quad \alpha_N = o\left(N^{1/(2s)}\right)$$

for some suitable constant C. The intuition behind the derivation for the upper bound is that the best chirplet path is achieved for chirplet paths that are close to the true instantaneous frequency of the unknown signal. If this was correct, one might expect the upper bound to be tight enough. However, we will see evidence in the next chapter that this is might not be the right intuition: at critical signal levels the best chirplet path typically deviates considerably away from the "true path," while the BP test can still detect the unknown signal with almost full power. This could mean, that we have a regime where the signal is *easily detectable* while it is *not estimable*; we could tell there is a chirp in the data but we cannot be sure what it looks like. Such a remarkable phenomenon would be impossible in classical parametric statistics because of the duality between statistical estimation and detection, but has been found in various nonparametric statistical problems (see, for example, [33, 50]).

Based on numerical experiments for an abstraction of our detection problem, we have reasons to believe there exists a rate $\rho_N = o(N^{1/(2s)})$ such that if $\alpha_N \sim \rho_N$, any sequence of BP tests is asymptotically powerless, while there is a different sequence of tests that is asymptotically powerful for this sequence signal levels α_N. In other words, that the BP test might not be asymptotically optimal (although, it could certainly be a very powerful test for practical applications). We will discuss this further in the next chapter and in the concluding remarks of the thesis.

8.4 The Abstraction Underlying the BP Test

There is an underlying abstraction behind our detection problem and the chirplet graph. Consider the case when the unknown signal is of the form $f(t) = \exp(iN\varphi(t))$ where φ is a Lipschitz function with a Lipschitz constant 1. Then Lemma 6 from Section 6.2.2 gives us directions on how to configure the chirplet graph so that there is a chirplet path through the graph that correlates well with the unknown chirp. The following two lemmas give us estimates for the size of the chirplet coefficients close to and away from the instantaneous frequency of the chirp (see Appendix C.8 and Appendix C.9 for proofs).

Lemma 10. *Let $f(t) = \exp(iN\varphi(t))$ with $\varphi \in$ HÖLDER$^2(R)$ and $c(t) = 1_I(t)\exp(iNbt)$. Assume that for all $t \in I$, $|\varphi'(t) - b| \leq \Delta\omega$ where $0 < \Delta\omega \leq \sqrt{2}N^{-1}|I|^{-1}$. Then*

$$|\langle f, c\rangle| \geq \left(1 - (N\Delta\omega|I|)^2/2\right) \cdot \|I\|.$$

Lemma 11. *Let $f(t) = \exp(iN\varphi(t))$ with $\varphi \in$ HÖLDER$^2(R)$ and $c(t) = 1_I(t)\exp(iNbt)$. Assume there exists some $\Delta\omega > 0$ such that $|\varphi'(t) - b| \geq \Delta\omega$ for all $t \in I$. Then*

$$|\langle f, c\rangle| \leq \frac{1}{N\Delta\omega}\left(2 + R \cdot \frac{|I|}{\Delta\omega}\right).$$

Assume the same chirplet coefficient discretization and chirplet graph as in Lemma 6 from Section 6.2.2. This is a very simple configuration: every chirplet is monochromatic and connects to two other chirplets to the right, one with a slightly higher frequency and the other with a slightly lower frequency (Figure 9.1 in the next chapter can give a diagrammatic idea of the topology of this graph). Choose the interval length in the uniform partition such that

$$|I| \sim N^{-1/2}$$

and define

$$\Delta\omega := (1/2 + \pi)|I| < \sqrt{2}N^{-1}|I|^{-1}.$$

Consider the data

$$y_t = \mu f_t + z_t,$$

where $\mu > 0$, $f_t = \exp(iN\varphi(t))/\|I\|$ and z_t is a sequence of noise.

According to Lemma 6 there exists a chirplet path W in the graph, such that $|\varphi'(t) - b_v| \leq \Delta\omega$ for all $t \in I_v$, where I_v is the support of the chirplet $c_v(t) = \exp(iNb_vt) \cdot 1_{I_v}(t)/\|I\|$ for $v \in W$. Then according to Lemma 10, we have

$$|\langle f, c_v\rangle| \geq (1 - \epsilon), \quad \text{for } v \in W,$$

where $0 < \epsilon < 1$. Pick a chirplet $v \in W$ of frequency $b_v = 2\pi|I| \cdot m_1$, (or $b_v = 2\pi|I| \cdot (m_1 +$

$1/2)$), for some integer m_1. For a chirplet c_u with the same support but different frequency $b_u = 2\pi|I| \cdot m_2$, (or $b_v = 2\pi|I| \cdot (m_2 + 1/2)$), we have

$$|b_v - b_u| \geq C_1 \cdot \Delta m|I|,$$

where $\Delta m = |m_1 - m_2| \in \mathbb{Z}^+$ and C_1 is some positive constant. Then

$$|\varphi'(t) - b_v| > C_2 \Delta m|I|,$$

for some constant C_2. Therefore, by Lemma 11,

$$|\langle f, c_u\rangle| \leq C_3 \cdot \frac{1}{\Delta m},$$

for some constant C_3. Note that

$$|\langle y, c_v\rangle|^2 = (\mu\beta_1 + W_1)^2 + (\mu\beta_2 + W_2)^2,$$

where W_1, W_2 are i.i.d. $N(0, 1/2)$ and $\langle f, c_v\rangle = \beta_1 + i\beta_2$. Then we have:

- Under H_0, the chirplet costs at the nodes in the graph are distributed as $\frac{1}{2}\chi^2_2$.
- Under H_1, there is a path W through the graph nodes such that $E|\langle y, c_v\rangle|^2 \approx \mu + 1$ for $v \in W$ and $E|\langle y, c_v\rangle|^2$ decreases as we move away from W.

For our choice of the discretization, the chirplets are all close to being orthogonal, so the chirplet costs are close to being independent. Also, under H_1 we have for $\mu = o(1)$ that away from W, the chirplet costs are roughly identically distributed as the chirplet costs under H_0.

This leads us to an abstract formulation of our detection problem. Consider a graph $\mathcal{G}$ and associate a random variable $Y_v = \beta_v + X_v$ to each node $v \in \mathcal{G}$, where $X_v \sim F$ with $E(X_v) = 0$. Assume that the random variables are independent. Then we consider two situations:

- H_0 : all the nodes have mean zero; i.e., $\beta_v = 0$ for all $v \in \mathcal{G}$.
- H_1 : there is a sequence of connected nodes W in $\mathcal{G}$ such that

$$\beta_v = \mu, \quad \text{for } v \in W, \quad \text{and} \quad \beta_v = 0, \quad \text{for } v \notin W.$$

What would be a good test for deciding between H_0 and H_1 and what are the limits of performance for this problem? This is the topic of the next chapter.

Chapter 9

Path Detection in Graphs

Consider a graph with a set of vertices and oriented edges connecting pairs of nodes according to some prescribed rule. Associate a random variable to each vertex. This chapter studies the problem of deciding between two hypothesis:

- Under H_0, the random variables have a common distribution F_0.
- Under H_1, there is an unknown path of connected nodes in the graph along which the random variables have a common distribution F_1, different from F_0. Away from the path the distribution is F_0.

We will consider the case when F_0 is the standard normal distribution and F_1 is a distribution of a normal variable with mean $\mu > 0$ and variance 1, and all the random variables are independent. We pose the questions: (i) for which values of μ can we detect a presence of a path in the graph, (ii) for which values is it impossible for any method, and (iii) what are the methods that achieve the limits of detectability?

9.1 The Setup

The answers to statistical questions regarding path detection in graphs depend, of course, on the type of graph one considers. We will consider two graphs that are very closely related. Below we let m be a positive integer.

- **Square lattice.** The first model is a directed graph with $m \times m$ nodes. We index the nodes by (j, k), $j, k \in \{1, \dots, m\}$ and there is an arc emanating from the node (j_1, k_1) to (j_2, k_2) if and only if $j_2 - j_1 = 1$ and $|k_2 - k_1| = 1$ modulo m. See Figure 9.1 for a sketch of the nodes and edges for this graph.

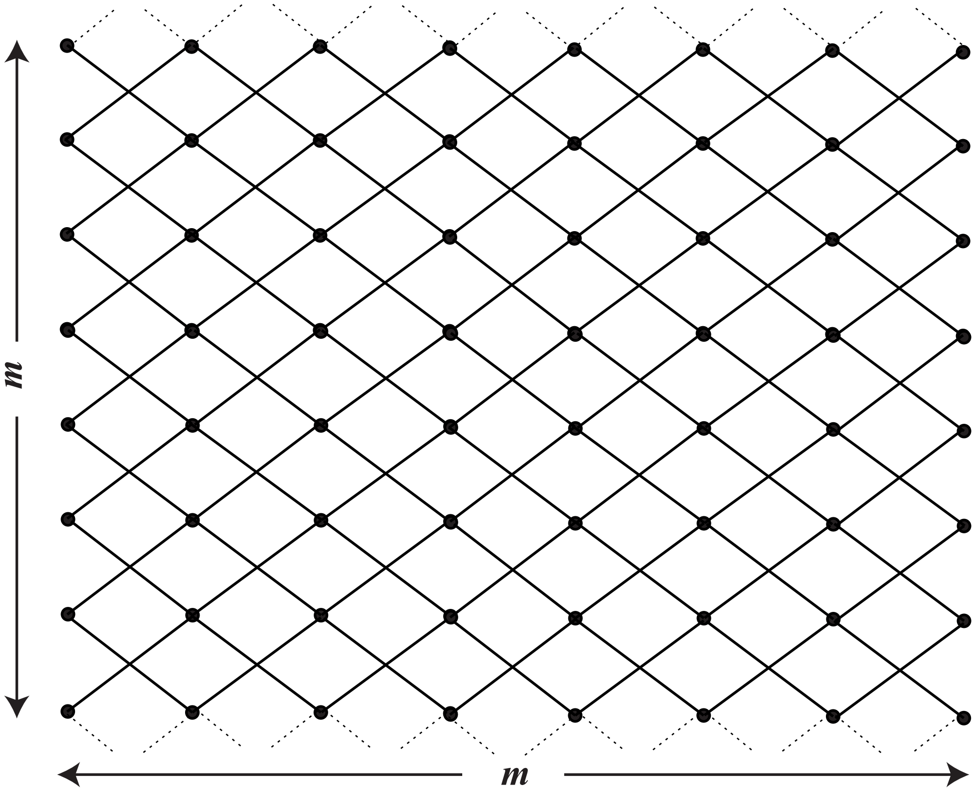


Figure 9.1: Representation of the square lattice. The dotted lines represent the periodic connectivity requirement.

- **Triangular lattice.** The second graph is a triangular lattice with vertices

$$V = \{(i, j) : 0 \leq i,\ -i \leq j \leq i \text{ and } j \text{ has the parity of } i\},$$

  and with oriented edges $(i, j) \rightarrow (i + 1, j + s)$ where $s = \pm 1$. We call $(0, 0)$ the *origin* of the graph.

Define the *length* of a path as the number of nodes the path visits. For the square lattice, the nodes indexed with $(1, k)$, $k = 1, \ldots, m$, are called *start-nodes* and the nodes indexed with (m, k), $k = 1, \ldots, m$, are called *end-nodes*. A path in the square lattice is a set of connected vertices starting from a start-node and ending at an end-node. Let $\mathcal{P}_m^s$ denote the set of all such paths. In the case of the triangular lattice, we let $\mathcal{P}_m^t$ be the set of paths in the graph starting at the origin and are of length m. (We drop the superscript when there is no ambiguity.) Notice the periodicity in the definition of the square lattice and that $\mathcal{P}_m^t$ can be considered to be a subgraph of $\mathcal{P}_{2m}^s$.

For a graph with the set of vertices V and paths $\mathcal{P}_m$, associate a random variable X_v to each vertex $v \in V$. Based on the observation $\{X_v : v \in V\}$ we wish to test the following two hypotheses:

- Under H_0, all the X_vs are i.i.d. $N(0,1)$.
- Under $H_{1,m}$, all the X_vs are independent; there is an unknown path $p \in \mathcal{P}_m$ along which the X_vs are i.i.d. $N(\mu_m, 1)$, $\mu_m > 0$, while they are i.i.d. $N(0,1)$ away from the path.

To study the quality of the tests for this problem we need to choose a criterion. Statistical decision theory provides us with the minimax and Bayesian paradigms which we will briefly review in the following section.

9.2 Statistical Preliminaries

Consider the statistical hypothesis testing problem where we want to decide between the simple hypothesis H_0 and the composite alternative hypothesis H_1, where we denote $\Theta_{1,m}$ as the set of simple hypotheses that $H_{1,m}$ is composed of. Let the decision rule T_m be a measurable function of the observed random variables, where T_m takes values in the set $\{0,1\}$. If $T_m = 0$, we decide that H_0 is true, otherwise, if $T_m = 1$, we decide that H_1 is true. There are two types of paradigms we wish to study:

- *Minimax Testing,* where we define the risk of the test T_m to be

  $$\gamma(T_m) = P(\text{Type I}) + \sup_{\theta \in \Theta_{1,m}} P_\theta(\text{Type II})$$

- *Bayesian Testing,* where we define the risk of a particular test T_m to be

  $$\gamma_\pi(T_m) = P(\text{Type I}) + E_\pi P(\text{Type II})$$

  where π is the prior on the alternative.

In our situation, $\Theta_{1,m} = \mathcal{P}_m$, T_m is a measurable function of $\{X_v\}_{v \in V}$, and π is a prior distribution on the set of paths. Recall that $P(\text{Type I}) = P_{H_0}(T_m = 1)$ and $P(\text{Type I}) =$

$P_\theta(T_m = 0)$. We say $\{T_m\}$ is *asymptotically powerful* if

$$\gamma(T_m) \to 0, \quad \text{as } m \to \infty$$

and *asymptotically powerless* if

$$\gamma(T_m) \to 1, \quad \text{as } m \to \infty,$$

where we omitted the subscript π in the case of Bayesian testing. We say the two hypotheses are *indistinguishable* if $\gamma(T_m) \to 1$ as $m \to \infty$ for every test T_m.

9.2.1 Bayesian testing

By writing $P_\pi(A) = E_\pi P(A)$, $A \in \mathcal{A}$, the Bayes problem reduces to the simple vs. simple hypothesis test, $H_0 : P_0$ vs. $H_1 : P_\pi$. In terms of the likelihood ratio $L_\pi := dP_\pi/dP_0$ the Bayes risk can be written as

$$B(\pi) = 1 - \frac{1}{2} E_{P_0}|L_\pi - 1|$$

The indistinguishability condition $B(\pi) \to 1$ holds if and only if $E_{P_0}|L_\pi - 1| \to 0$. This quantity is often difficult to investigate analytically but sometimes it is possible to get results from looking at the sufficient condition

$$E_{P_0}(L_\pi - 1)^2 \to 0.$$

Indeed, since $Var(X) \geq 0$ for any random variable X (this is essentially a consequence of Jensen's inequality) and therefore $E_{P_0}|L_\pi - 1| \leq (E_{P_0}(L_\pi - 1)^2)^{1/2}$. Thus,

$$B(\pi) \geq 1 - (E_{P_0}(L_\pi - 1)^2)^{1/2}.$$

Remark that $E_{P_0}L = 1$ so $E_{P_0}(L_\pi - 1)^2 = E_{P_0}L_\pi^2 - 1$ and the sufficient condition becomes

$$E_{P_0}L_\pi^2 \to 1.$$

9.2.2 The Bayes test for the square lattice

The best test (the one that minimizes the Bayes risk) is given by the Neyman-Pearson test:

$$L = \int \frac{dP_\theta}{dP_0} \pi(d\theta).$$

Fix a path θ. Then

$$\frac{dP_\theta}{dP_0} = \frac{\Pi_{i,j} \exp\left(-1/2(x_{ij} - \mu\theta_{ij})^2\right)}{\Pi_{i,j} \exp\left(-1/2x_{ij}^2\right)} = e^{-\frac{1}{2}m\mu^2} e^{\mu\langle x, 1_\theta\rangle}$$

and

$$L = e^{-\frac{1}{2}m\mu^2} \int e^{\mu\langle x, 1_\theta\rangle} \pi(d\theta).$$

The Bayes test rejects when L exceeds a threshold. It is hard to analyze L directly. Note that the Bayesian assumes the value of the mean μ along the unknown path to be known.

9.2.3 The Bayes test for a uniform prior on paths

Consider the square lattice and assume a uniform prior on the paths in the graph. That is, a path is a symmetric random walk with uniformly distributed start-node. To study the indistinguishability condition in the Bayes problem we study the second moment of the likelihood ratio. First we have

$$L^2 = \iint e^{-\frac{1}{2}m\mu^2} e^{\mu\langle x, 1_\theta\rangle} e^{-\frac{1}{2}m\mu^2} e^{\mu\langle x, 1_{\theta'}\rangle} \pi(d\theta)\pi(d\theta') = e^{-m\mu^2} \iint e^{\mu\langle x, 1_\theta + 1_{\theta'}\rangle} \pi(d\theta)\pi(d\theta').$$

For a random vector Z with iid $N(0,1)$ entries and a fixed vector v of the same size we have

$$E\left(e^{\mu\langle Z, v\rangle}\right) = e^{\mu^2\|v\|^2/2}.$$

Also, $\|1_\theta + 1_{\theta'}\|^2 = 2m + 2\langle 1_\theta, 1_{\theta'}\rangle$ which gives

$$\begin{aligned} E_{P_0} L^2 &= e^{-m\mu^2} \iint e^{\mu^2\|1_\theta + 1_{\theta'}\|^2/2} \pi(d\theta)\pi(d\theta') \\ &= \iint e^{\mu^2\langle 1_\theta, 1_{\theta'}\rangle} \pi(d\theta)\pi(d\theta'). \end{aligned}$$

This is the moment-generating function of the number of crossings of two independent random walks, S_i and S_i' on the graph. Then we have

$$E_{P_0}(L-1)^2 = Ee^{\mu^2 \sum_i 1_{\{S_i = S_i'\}}} - 1.$$

9.3 A Bayes Problem for the Square Lattice

Clearly our ability to detect depends on prior knowledge about the paths in the alternative hypothesis. Here we will study the case where the prior distribution of the paths in the alternative hypothesis is such that all paths are equally likely. First we choose the starting node uniformly at random and then we let the path be a random walk, where at each node it is equally likely to go up or down in the next step. That is, we want to decide between the following two hypotheses

$$H_0 : Y_{j,k} \text{ i.i.d } N(0,1) \quad \text{vs.} \quad H_\theta : Y_{j,k} \text{ i.i.d } N(\mu 1_\theta(j,k), 1)$$

where θ is a random walk path as described above, and $1_\theta(j,k)$ is 1 if the node (j,k) is on the path and 0 otherwise.

Next we want to study the threshold phenomenon in this Bayesian testing problem. That is, find the rate ϵ_m such that

- If $\mu_m/\epsilon_m \to 0$ then $\gamma \to 1$; the two hypotheses are indistinguishable.
- If $\mu_m/\epsilon_m \to \infty$ then there is a test with $\gamma \to 0$.

Theorem 12. *Consider the square lattice and assume the uniform prior on paths. If $\mu_m/(m^{-1/4}\sqrt{\log m}) \to \infty$ as $m \to \infty$, then the Bayes risk tends to 0. If $\mu_m/m^{-1/4} \to 0$ as $m \to \infty$, the Bayes risk tends to 1.*

Notice the gap between the bounds which we comment briefly on at the end of the proof of the lower bound in Section 9.3.2.

9.3.1 Proof of Theorem 12: upper bound

Let h_m be an arbitrary sequence of real numbers tending to infinity and define $B(i, h_m)$ to be a set of nodes in a strip of length m and width about $\sqrt{m}$, and centered around $j = i$;

i.e.,

$$B(i,h_m) := \{(j,k) \in V : |k-i| \leq h_m\sqrt{m}, \text{ where } k-i \text{ is calculated modulo } m\}.$$

Let $N_{m,i}$ be the number of nodes in $B(i,h_m)$ and define $B_{m,i}$ to be the normalized sum of the variables in this strip,

$$B_{m,i} = 1/\sqrt{N_{m,i}} \sum_{v \in \mathcal{S}(i,h_m)} X_v.$$

Our test statistic is the maximum of those sums,

$$T_m = \max\{B_{m,i} : i = 1,\ldots,m\},$$

and our test rejects H_0 for large values of T_m. For m random variables $Y_1,\ldots,Y_m$, such that $Y_k \sim N(0,\sigma_k^2)$, $\sigma_k \leq \sigma$ for all $k = 1,\ldots,m$, we have

$$P\left(\max\{Y_1,\ldots,Y_m\} > \sqrt{2\log(m)} \cdot \sigma\right) \leq \frac{1}{\sqrt{4\pi\log(m)}}.$$

This is a well-known inequality and follows from the Gaussian tail bound,

$$P(Y_k > u) \leq \frac{\sigma_k}{\sqrt{2\pi}u} e^{-\frac{1}{2}u^2/\sigma_k^2}, \quad \text{for all } u > 0,$$

and a union bound. Under H_0, $B_{m,i} \sim N(0,1)$, so this inequality gives us

$$P_0\left(T_m > \sqrt{2\log(m)}\right) \leq \frac{1}{\sqrt{4\pi\log(m)}}.$$

Choose the threshold $t_m = \sqrt{2\log(m)}$. The last inequality shows that $P_0(T_m > t_m) \to 0$ as $m \to \infty$. Now assume H_1 holds and let p be the hidden path. Then with high probability, the path is contained within a strip that is centered at the starting node of p.

Let the starting node of p be i and consider the strip $B(i,h_m)$. Let L_p be the number of nodes in the path p that are in this strip. Then,

$$B_{m,i} = \mu\frac{\sqrt{L_p}}{\sqrt{N_{m,i}}} + Z, \quad Z \sim N(0,1).$$

Writing $N_m = \max\{N_{m,1}, \ldots, N_{m,m}\}$, we have

$$\begin{aligned} P_p\left(T_m^* < t_m\right) &\leq P_p\left(B_{m,i} < t_m\right) \leq P_p\left(Z < t_m - \mu_m L_p/\sqrt{N_m}\right) \\ &= P\left(Z < t_m - \mu_m L_p/\sqrt{N_m} \mid L_p = m\right) P(L_p = m) \\ &\quad +P\left(Z < t_m - \mu_m L_p/\sqrt{N_m} \mid L_p < m\right) P(L_p < m) \\ &\leq P\left(Z < t_m - \mu_m \cdot m/\sqrt{N_m}\right) + P(L_p < m). \end{aligned}$$

The next steps are to show that both of the terms on the far right-hand side tend to zero for sequences (μ_m) such that $\mu_m/(m^{-1/4}\sqrt{\log m}) \to \infty$ as $m \to \infty$.

First notice that $P(L_p < m)$ is equal to the probability that the oriented symmetric random walk $(k, S_k)_{1\leq k\leq m}$, starting at i, steps outside the strip $B(i, h_m)$. We can assume the walk starts at 0 and find the probability of the event $E_m = \{\max_{1\leq k\leq m}|S_k| > h_m\sqrt{m}\}$. Recall Kolmogorov's inequality [37]:

Lemma 12 (Kolmogorov's Inequality)**.** *Let $X_1, \ldots, X_m$ be mutually independent random variables with expectations $\mu_k = E(X_k)$ and variances σ_k^2. Put $S_k = X_1 + \cdots + X_k$ and $m_k = E(S_k) = \mu_1 + \cdots + \mu_k$, $s_k^2 = Var(S_k) = \sigma_1^2 + \cdots + \sigma_k^2$. For every $t > 0$ the following inequality holds*

$$P\left(\max_{k\in\{1,\ldots,m\}} |S_k - m_k| < ts_m\right) \geq 1 - t^{-2}.$$

In case of our random walk, $m_k = 0$ and $s_m^2 = m$ and therefore $P(E_m) \leq h_m^{-2}$. Thus, for any sequnce (h_m) tending to infinity,

$$\lim_{m\to\infty} P(E_m) = 0.$$

(We could also use the reflection principle and Hoeffding's inequality to prove this.)

For the other term, note that $N_m = h_m m^{3/2}(1 + o(1))$. Then as long as $\mu_m m^{1/4} h_m^{-1/2}$ grows faster than $t_m = \sqrt{2\log(m)}$, $P\left(Z < t_m - \mu_m \cdot m/\sqrt{N_m}\right) \to 0$ as $m \to \infty$. Under the assumption $\mu_m/(\sqrt{\log(m)} \cdot m^{-1/4}) \to \infty$, we can find a sequnce $h_m \to \infty$ so that this condition holds. Therefore, the test T_m is asymptotically powerful, which completes the proof of the upper bound.

9.3.2 Proof of Theorem 12: lower bound

Recall that $\mathbf{E}_{P_0}(L-1)^2 = \mathbf{E}e^{\mu^2 \sum_i 1_{\{S_i = S_i'\}}} - 1$. In what follows, we write $t := \mu^2$. Define the event when the two random walks meet at time i as $I_i := \{S_i = S_i'\}$ and let $X_k = S_k - S_k'$, modulo m. Then for integers r, s,

$$P(X_k = r \mid X_{k-1} = s) = \begin{cases} 1/4 & \text{if } r = s \pm 1 \bmod m \\ 1/2 & \text{if } r = s \bmod m \\ 0 & \text{otherwise.} \end{cases}$$

Define

$$E_{k,j} := E\left[e^{t\sum_{i=k}^m 1_{I_i}} \mid X_k = j\right]$$

where we take the index j modulo m. Consider the following two cases:

(i) $\mathbf{j \neq 0}$: Then if $X_{k-1} = j$ we have $1_{I_{k-1}} = 0$ and

$$\begin{aligned} E_{k-1,j} &= E\left[e^{t\sum_{i=k-1}^m 1_{I_i}} \mid X_{k-1} = j\right] = E\left[e^{t\sum_{i=k}^m 1_{I_i}} \mid X_{k-1} = j\right] \\ &= \sum_{r=0}^{m-1} E\left[e^{t\sum_{i=k}^m 1_{I_i}} \mid X_{k-1} = j, X_k = r\right] P(X_k = r \mid X_{k-1} = j) \\ &= \frac{1}{4}E_{k,j+1} + \frac{1}{2}E_{k,j} + \frac{1}{4}E_{k,j-1}. \end{aligned}$$

(ii) $\mathbf{j = 0}$: Then if $X_{k-1} = j$ we have $1_{I_{k-1}} = 1$ and

$$\begin{aligned} E_{k-1,j} &= E\left[e^{t\sum_{i=k-1}^m 1_{I_i}} \mid X_{k-1} = j\right] = e^t E\left[e^{t\sum_{i=k}^m 1_{I_i}} \mid X_{k-1} = j\right] \\ &= e^t\left(\frac{1}{4}E_{k,j+1} + \frac{1}{2}E_{k,j} + \frac{1}{4}E_{k,j-1}\right). \end{aligned}$$

Next we define the column vector $E_k = [E_{k,0}, E_{k,1}, \cdots, E_{k,m-1}]^T$ and let $P = [p_{ij}]$ be an $m \times m$ matrix with 1/2 on the diagonal, 1/4 on the super- and subdiagonals, and $p_{1m} = p_{m1} = 1/4$. Let the vector $\mathbf{q} = [1, 0, \ldots, 0]^T$ be a column vector of length m and define the $m \times m$ matrix $Q = qq^T$. That is:

$$P = \begin{pmatrix} 1/2 & 1/4 & 0 & \cdots & 0 & 1/4 \\ 1/4 & 1/2 & 1/4 & \ddots & 0 & 0 \\ 0 & 1/4 & 1/2 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & 0 & \ddots & \ddots & 1/2 & 1/4 \\ 1/4 & 0 & \cdots & 0 & 1/4 & 1/2 \end{pmatrix} \qquad Q = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix}.$$

This matrix notation allows us to combine the formulas from cases (i) and (ii) into a nice recursion relation:

$$\begin{aligned} E_{k-1} &= PE_k + (e^t - 1)QPE_k \\ &= (P + \lambda QP)E_k, \end{aligned}$$

where $\lambda := e^t - 1$. Since

$$E_{m,j} = E\left[e^{t1_{I_m}} \mid X_m = j\right] = \begin{cases} e^t & \text{if } j = 0 \\ 0 & \text{otherwise} \end{cases}$$

we have $E_m = \mathbf{1} + \lambda\mathbf{q}$, and therefore

$$\begin{aligned} E_1 &= (P + \lambda QP)^{m-1}E_m \\ &= (P + \lambda QP)^{m-1}(\mathbf{1} + \lambda\mathbf{q}). \end{aligned}$$

For a vector $\mathbf{v} = [v_1, \ldots, v_n]^T$, define $Ave(\mathbf{v})$ to be the average of its entries; i.e.,

$$Ave(\mathbf{v}) := \frac{1}{n}\sum_{k=1}^{n} v_k.$$

Then by the law of total expectation

$$\begin{aligned} E\left[e^{t\sum_{i=1}^{m} 1_{I_i}}\right] &= E\left[E\left[e^{t\sum_{i=1}^{m} 1_{I_i}} \mid X_1\right]\right] = Ave(E_1) \\ &= Ave\left((P + \lambda QP)^{m-1}(\mathbf{1} + \lambda\mathbf{q})\right). \end{aligned}$$

Next we have to study the conditions on the growth of λ as $m \to \infty$ for this last quantity to tend to 1. Instead of studying $Ave\left((P+\lambda QP)^{m-1}(\mathbf{1}+\lambda\mathbf{q})\right)$, we can equivalently study $Ave\left((P+\lambda QP)^{m}(\mathbf{1}+\lambda\mathbf{q})\right)$. Indeed,

$$Ave\left((P+\lambda QP)^{m-1}(\mathbf{1}+\lambda\mathbf{q})\right) \geq Ave\left(P^{m-1}\mathbf{1}\right) = 1$$

since all the terms dropped where positive and the matrix P is stochastic. Now since

$$(P+\lambda QP)(\mathbf{1}+\lambda\mathbf{q}) = \mathbf{1} + 3/2\lambda\mathbf{q},$$

we have

$$(P+\lambda QP)^{m}(\mathbf{1}+\lambda\mathbf{q}) = (P+\lambda QP)^{m-1}(\mathbf{1}+\lambda\mathbf{q}) + (P+\lambda QP)^{m-1}1/2\lambda\mathbf{q}.$$

By dropping positive terms, we get

$$Ave\left((P+\lambda QP)^{m-1}(\mathbf{1}+\lambda\mathbf{q})\right) \leq Ave\left((P+\lambda QP)^{m}(\mathbf{1}+\lambda\mathbf{q})\right).$$

In order for this average to go to 1, we have the following sufficient condition on the rate of $\lambda = \lambda_m$. It is established in Appendix C.10.1:

Lemma 13 (A sufficient condition for the rate of λ)**.** *If* $\lambda_m = o\left(\frac{1}{\sqrt{m}}\right)$,

$$\lim_{m\to\infty} Ave\left((P+\lambda_m QP)^{m}(\mathbf{1}+\lambda_m\mathbf{q})\right) = 1.$$

Since $t = \log(1+\lambda) = \lambda + o(\lambda)$, as $\lambda \to 0$, and $\mu^2 = t$, we get that if $\mu_m = o(n^{-1/4})$,

$$E\left[e^{\mu_m^2 \sum_{i=1}^{m} 1_{I_i}}\right] \to 1, \quad m \to \infty,$$

which concludes the proof of the lower bound in Theorem 12.

We also have the following lemma which is not needed for the result and is simpler to establish than Lemma 13 (see Appendix C.10.2 for a proof). It gives us a restriction on the rate of λ for the average to go to zero:

Lemma 14 (Necessary condition for the rate of λ)**.** *Let* $\rho_m = \log(m)/\sqrt{m}$*. If* $\lim_{m\to\infty} \lambda_m/\rho_m >$

1/2, then

$$\lim_{m\to\infty} Ave\left((P+\lambda_m QP)^m(\mathbf{1}+\lambda_m \mathbf{q})\right) > 1.$$

If the upper bound we have established gives the correct rate, Lemma 14 might tell us that it could be hard to improve the lower bound based on analyzing $E_0(L_m-1)^2$ directly. Also, remember that in bounding the Bayes risk, we used the inequality $E_0|L_m - 1| \leq (E_0(L_m-1)^2)^{1/2}$ and hoped there was enough justice in the world, so that studying the variance of the Bayes statistic would still give us a tight bound (which it does in many nonparametric statistical problems [50]).

9.4 The GLRT for Path Search

Perhaps the first test that comes to mind for path detection in graphs is the likelihood ratio test (GLRT), which would reject H_0 for large values of $M_m = \max\{S_p : p \in \mathcal{P}_m\}$, where S_p is the sum of the node variables along the path p:

$$S_p = \sum_{v\in p} X_v. \tag{9.1}$$

This test makes sense intuitively, and is also attractive from a practical point of view, as it can be calculated rapidly by dynamic programming. Note that the monoscale BP test is of this form, and therefore by studying M_m we could hope to understand the BP test better.

Empirically, we have seen that for the square lattice under H_0, M_m/m gets highly concentrated around a value close to 0.87 as we let m grow. This suggests that the statistic should be compared to a constant. Using the following lemma we can identify a *sufficient condition* for the GLRT to be asymptotically powerful.

Lemma 15. *Under H_0, we have for any $\epsilon > 0$,*

$$M_m/m \leq \sqrt{2(\log 2 + \epsilon)},$$

with probability exceeding $1 - Ce^{-\lambda m}$, for some positive constants C and λ depending on ϵ.

Lemma 15 tells us that if we compare the GLRT to a constant threshold $\tau_m := m\sqrt{2\log(2+\epsilon)}$, $\epsilon > 0$, the probability of Type I error goes to zero as $m \to \infty$. Now assume that H_1 with p^*

as the true path and μ is the mean. Then

$$S_{p^*} = \mu m + \sqrt{m} Z,$$

for some random variable $Z \sim N(0,1)$. Since $M_m \geq S_{p^*}$, we have

$$P_{p^*}(M_m < \tau_m) \quad \leq \quad P_{p^*}(S_{p^*} < \tau_m) = P\left(Z < (\sqrt{2\log(2+\epsilon)} - \mu)m^{1/2}\right),$$

which tends to zero if $\mu > \sqrt{2\log(2+\epsilon)}$. This holds, no matter which path in the graph we take. Now we ask, *can the GLRT still detect when $\mu_m \to 0$ at a slow enough rate*? Although a confirmed answer remains to be given, our numerical simulations in Section 9.6 provide evidence that the answer is positive for our Bayesian setting with uniform prior on paths. They also indicate that the GLRT does not achieve the optimal threshold. However, in case of the minimax problem, the simulations hint that the answer is negative and the threshold of detectability for the GLRT is for a constant μ_m. That would also mean that, at least in the case of the triangular lattice, the GLRT is not optimal for the minimax problem, since, according to Section 9.5, there exists a test which can detect when $\mu_m \to 0$ for the minimax problem.

Proof of Lemma 15: For a random variable $X \sim N(0, \sigma^2)$ we have the classical bound

$$P(X > u) \leq \frac{\sigma}{u\sqrt{2\pi}} e^{-u^2/(2\sigma^2)}, \quad u > 0. \tag{9.2}$$

Consider first the square lattice and note that the number of paths in $\mathcal{P}_m^s$ is $m \cdot 2^{m-1}$; indeed, since there are m choices for the starting node and 2 choices at each of the $m-1$ steps after that. The sum S_p from (9.1) is distributed as $N(0, m)$. Then for any $t > 0$, inequality (9.2) with $u = mt$, and a union bound give us

$$\begin{aligned} P(S_m^* > mt) &= P(\max\{S_p : p \in \mathcal{P}_m\} > mt) \\ &\leq m \cdot 2^{m-1} \cdot \frac{\sqrt{m}}{mt\sqrt{2\pi}} \exp(-mt^2/2) \\ &= \frac{\sqrt{m}}{t2\sqrt{2\pi}} \cdot \exp(-mt^2 + 2\log(2)m/2). \end{aligned}$$

Choose $t = \sqrt{2(\log 2 + \epsilon)}$ for some $\epsilon > 0$. Then there exist positive constants C and λ such that the inequality

$$P(S_m^*/m > t) \leq C \cdot e^{-\lambda m},$$

holds [1].

The proof is basically identical for the triangular lattice where the number of paths in $\mathcal{P}_m^t$ is equal to 2^{m-1}. □

9.4.1 Numerical simulations for the GLRT in the square lattice

Here we briefly explore the behavior of the path that maximizes the sum of paths over all graphs in the graph. That is the path $p^* = \arg\max\{S_p : p \in \mathcal{P}_m\}$, so that the GLRT is $M_m = S_{p^*}$, where S_p is, as before, the sum of the node variables along the path p. We call p^* the Best Path in the graph.

We consider the Bayes problem for the square lattice of size $m \times m$ where $m = 1024$. For each realization under H_1, the starting position of the path is sampled uniformly at random and a random walk is generated for the subsequent steps of the path. We fix the probability of Type I error at 5%. Based on 10,000 realizations of the node variables $\{X_v : v \in V\}$ under H_0, we determined that this significance was achieved for threshold $\tau \approx 0.897m$, so our test compared M_m/m to the value 0.897. Figure 9.2 shows the histogram of M_m/m based on these realizations.

We are interested in knowing what the Best Path looks like under H_1 in a situation where the mean μ along the true path is critical for the GLRT; i.e., it is close to the limit of the test being able to reliably detect. Figure 9.3 shows a realization of a path drawn according to our Bayes prior. Figures 9.4, 9.5, 9.6, and 9.7 show comparisons of the real and best path for $\mu = 2, 1, 0.6$, and 0.5 respectively; the same realization of the node variables was used in every case and the true path was the path on Figure 9.3.

In all of these cases, the GLRT would have decided that there was a signal present in the graph, but only barely in the case where $\mu = 0.5$. The estimated P-values [2] for the two critical values of means, $\mu = 0.6$ and 0.5, were $4 \cdot 10^{-4}$ and 0.031 respectively. Thus, at least for $\mu = 0.6$ there is a strong evidence against H_0. Based on the histogram, we are able to detect a path for a mean that is far below the bulk of the distribution of M_m/m under H_0.

[1] This gives us some more information. Define the event $E_m := \{S_m^*/m > t\}$. Then using our bound we get that the sum $\sum_{m=1}^{\infty} P(E_m)$ is convergent, and therefore by the Borell-Cantelli lemma we have for any $\epsilon > 0$, $\limsup_{m\to\infty} S^*/m \leq \sqrt{2(\log 2 + \epsilon)}$ almost surely under H_0.

[2] Here the P-value is $P_{H_0}(M_m > M_m^{obs})$, where M_m^{obs} is the observed value of the GLRT statistic.

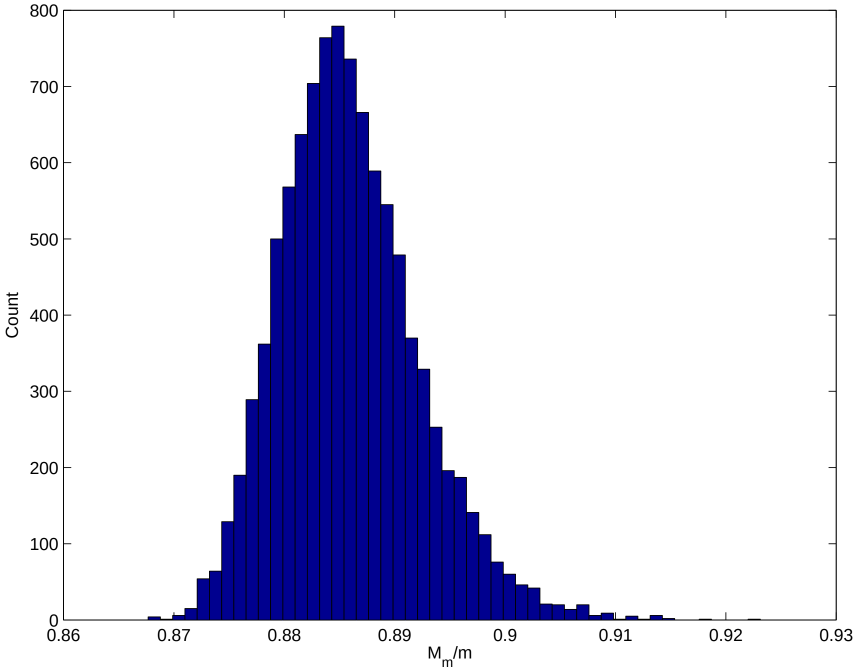


Figure 9.2: Histogram of M_m/m based on 10,000 realizations under H_0

The interesting observation is, however, that the path which provides this evidence has only about 20% nodes in common with the true path! Although we can strongly reject H_0, we would not be able to provide a reasonable estimate of the underlying path.

9.5 The Triangular Lattice

Here we will briefly review the results that were established in [9] for the triangular lattice. There the path detection problem was studied in both the minimax and Bayesian decision theoretic frameworks. In the Bayes case, where one assumes that all the paths are equally likely, the Bayes statistic is given by

$$L_m = 2^{-(m-1)} \sum_{p \in \mathcal{P}_m} e^{\mu S_p - m\mu^2/2}, \tag{9.3}$$

where S_p is defined in (9.1), and we have:

Theorem 13. *Consider the triangular lattice and assume the uniform prior on paths. If $\mu_m \, m^{1/4} \to \infty$ as $m \to \infty$, then the Bayes risk tends to 0. Conversely, if $\mu_m \, m^{1/4} \to 0$ as $m \to \infty$, the Bayes risk tends to 1.*

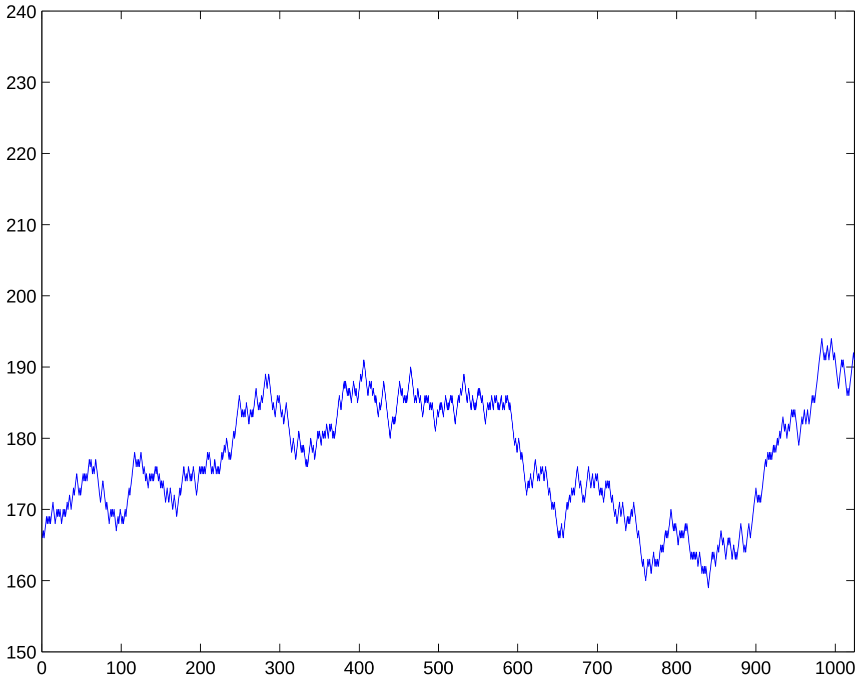


Figure 9.3: A plot of the sample path drawn from the uniform prior. Note that we only plot a portion of the vertical axis.

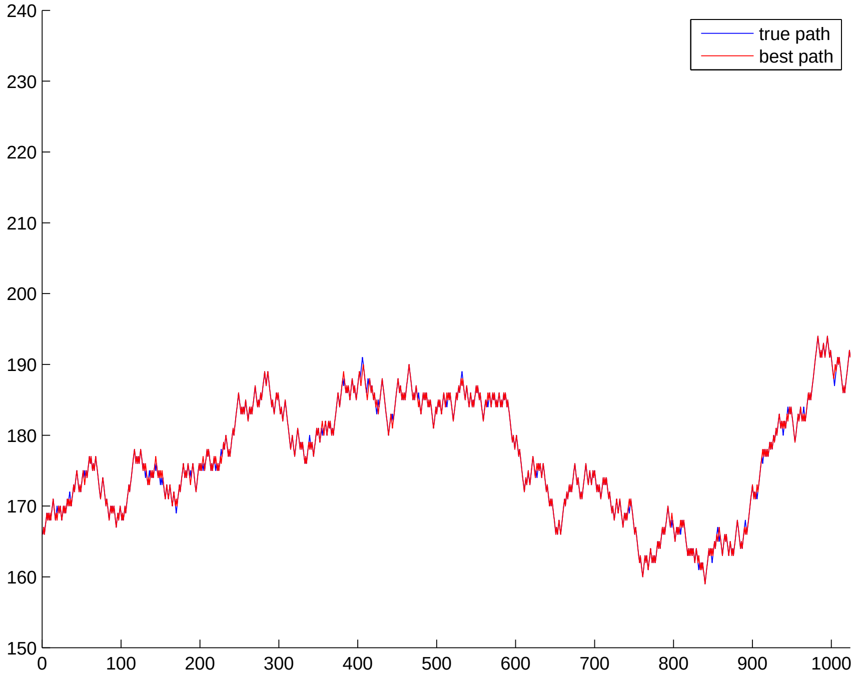


Figure 9.4: A plot of the best path and true path for $\mu = 2$. For this realization $M_m/m \approx 2.00$ and the overlap is about 96%.

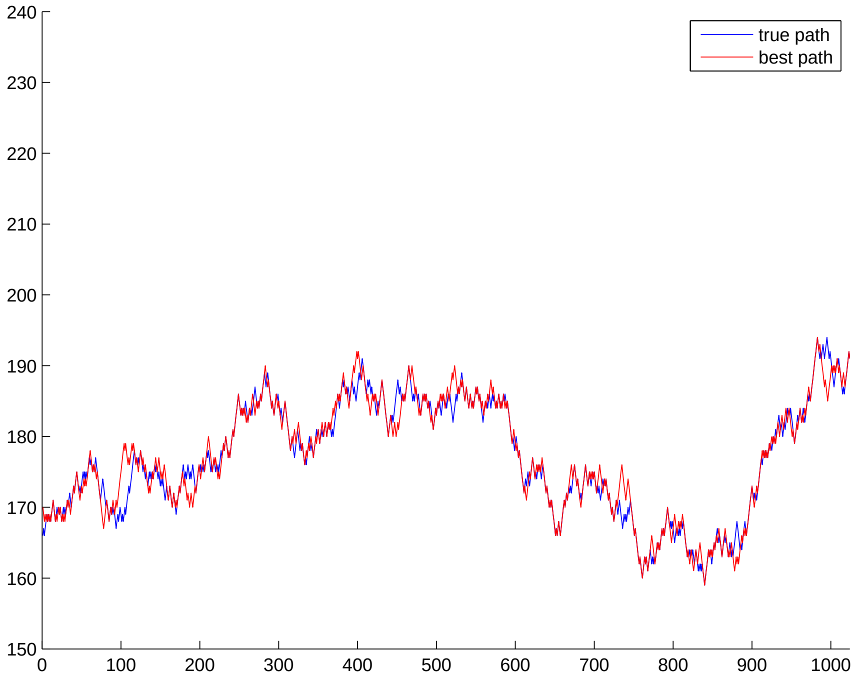


Figure 9.5: A plot of the best path and true path for $\mu = 1$. For this realization $M_m/m \approx 1.13$ and the overlap is about 70%.

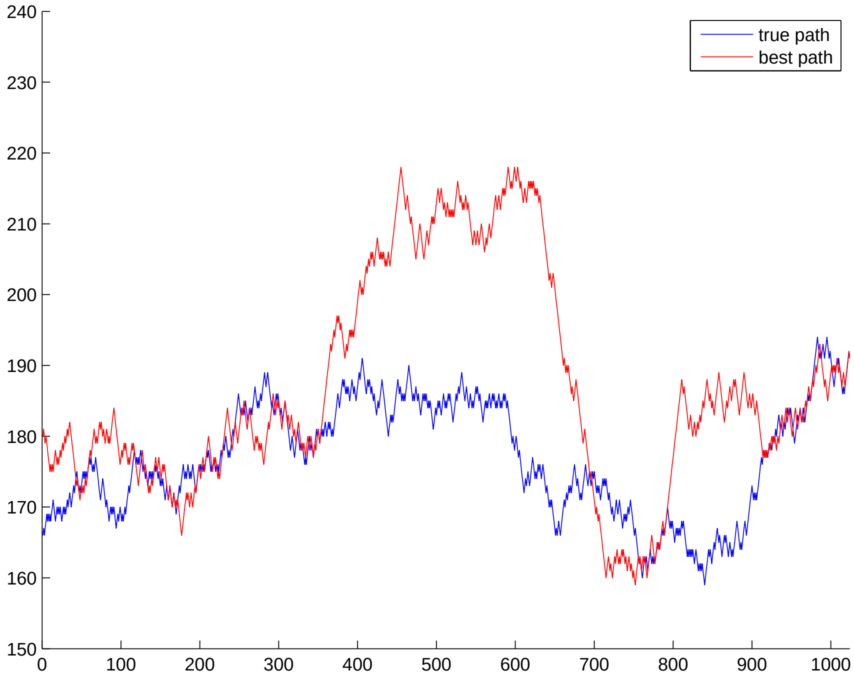


Figure 9.6: A plot of the best path and true path for $\mu = 0.6$. For this realization $M_m/m \approx 0.92$ and the overlap is about 20%.

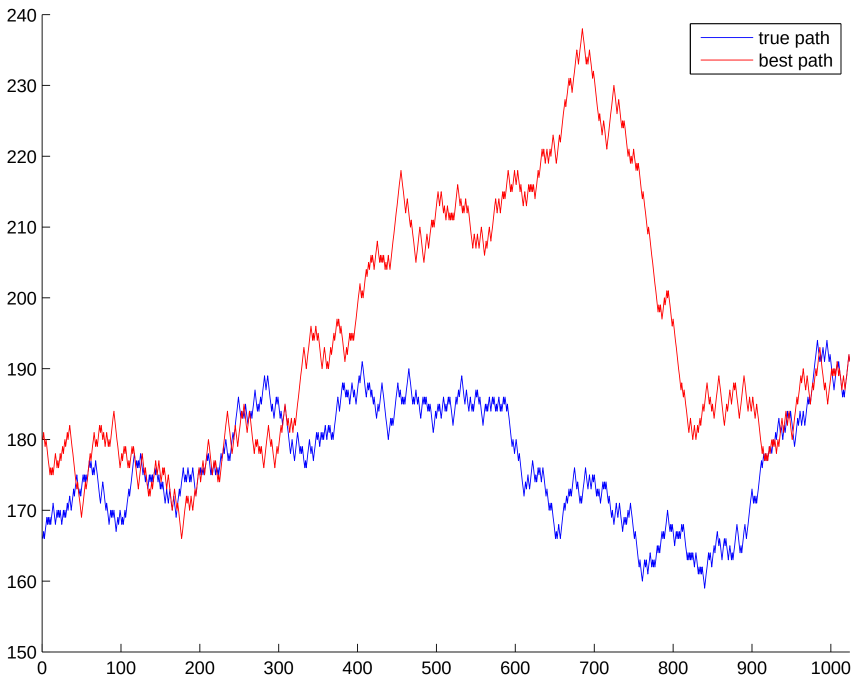


Figure 9.7: A plot of the best path and true path for $\mu = 0.5$. For this realization $M_m/m \approx 0.90$ and the overlap is only about 12.5%.

Unlike our results for the square lattice, this time we know the optimal detectability threshold; if the unknown path is chosen uniformly at random, it is possible to detect it as long as the mean of the nodes along the path exceeds $m^{-1/4}$, while no method can detect if it is below this level. Note that the Bayes test assumes the value of the mean μ along the unknown path to be known while the strip statistic used to construct the upper bound does not need that information.

The upper bound in Theorem 13 was established using a similar kind of a “strip” or “box,” as in the proof of the upper bound in Theorem 12. As expected, it is only required to sum the values of random variables in single strip centered at the origin of the lattice. Contrary to the square lattice where the starting location of the path is unknown, the test does not have to “pay the price” of a logarithmic factor for multiple comparisons of strips.

For the case for minimax detection, the following theorem was established[3]:

Theorem 14. *Consider the triangular lattice. Suppose that $\mu_m (\log m)^{1/2} \to \infty$ as $m \to \infty$. Then there is a sequence of tests which is asymptotically powerful. On the other hand,*

[3] I would like to make clear that the theoretical result for the minimax problem is due to my coauthors and I cannot claim credit for it.

suppose that $\mu_m \log m \, (\log\log m)^{1/2} \to 0$ *as* $m \to \infty$*. Then every sequence of tests* (T_m) *is asymptotically powerless.*

Theorem 14 says that it is possible to detect a path as long as $\mu_m \gg (\log m)^{-1/2}$, and impossible if $\mu_m < (\log m)^{-(1+\epsilon)}$ for each $\epsilon > 0$ and m sufficiently large. Notice the slight "gap" between the upper and lower bound that would need to be closed to decide the actual threshold of detectability. We also see that these rates are quite different than for our Bayes problem. This also tells us that the uniform prior on the paths is quite far from being a "difficult" prior, in the sense that it does not provide us with a good lower bound for the minimax problem.

To prove the upper bound in Theorem 14 the following test statistic was used:

$$T_m = \sum_{(i,j)\in V} w_{i,j} X_{i,j}, \qquad w_{i,j} := w_i = \frac{\lambda_m}{i+1}, \tag{9.4}$$

with λ_m chosen such that $\sum_i w_i = 1$. With this choice $\lambda_m \sim (\log m)^{-1}$. Thus, T_m is a simple weighted average of the values at the vertices of the graph. Since this statistic is a sum of i.i.d. Gaussian random variables, it is easy to analyze: Under H_0,

$$T_m \sim N(0, \lambda_m),$$

and for any alternative,

$$T_m \sim N(\mu_m, \lambda_m).$$

Hence, this statistic could not care less what the path looks like, only that it is contained in the triangular lattice. Therefore it does not use the information that the path is continuous! Therefore it can detect any sequence of the form $\{(i, p_i) : 0 \le i \le m-1\}$ provided that (i, p_i) is a vertex in the graph and $\mu_m (\log m)^{1/2} \to \infty$ as $m \to \infty$. This turns out to be the minimax detection threshold for detecting the presence of such sequences in the triangular lattice. It would be quite surprising if the upper bound in Theorem 14 was tight, since it would essentially mean that the continuity of the path does not really help to enhance the detection.

The lower bound in Theorem 14 uses a delicate construction of a Bayes prior based on the concept of a *predictability profile* which was first introduced in [15]. It tells how hard it is to predict a future location of a stochastic process from its current state and history. The formal definition is:

Definition 5. *The predictability profile of a stochastic process* $(S_n)_{n \geq 1}$ *is defined by*

$$\mathrm{PRE}_S(k) = \sup P(S_{n+k} = x \,|\, S_0, \dots, S_n), \tag{9.5}$$

where the supremum is taken over all positions and histories.

The strategy was then to construct a prior on the family of paths with a low predictability profile.

9.6 Numerical Simulations for Detection in the Triangular Lattice

We now explore the empirical performance of some of the detection methods that have been proposed for the case of the triangular lattice. The cost at the vertices are independent Gaussians. To measure the performance, we fix the probability of Type I error at 5% and estimate the detection rate, the probability of detecting a path when there is a path in the graph whose vertices have a non-zero mean μ. The path used in each realization will be randomly sampled according to some specific prior. The detection rates were estimated for a set of means, μ, close to the transition between a poor and a nearly perfect detection.

9.6.1 Bayesian detection under the uniform prior

We first consider detection under the uniform prior on paths. We will compare the performance of the corresponding Bayes test, the GLRT and the test based on the Strip statistic, used in the proof of the upper bound in Theorem 12. Remember that the Bayes test is optimal in this setting and that the Strip statistic was shown to achieve the optimal detection rate. However, we were unable to theoretically analyze the performance of the GLRT in this situation and would like to do so here through simulations.

9.6.1.1 Simulating the Bayes statistic

Let $Y(v) = \sum_{p \in \mathcal{P}(v)} e^{\mu S_p}$, where $\mathcal{P}(v)$ is the set of all paths in the lattice starting at $(0,0)$ and ending at the vertex v and S_p is defined in (9.1). Take a vertex $v_0 = (i,j)$ whose predecessor vertices are $v_1 = (i-1, j-1)$ and $v_2 = (i-1, j+1)$. Then,

$$Y(v_0) = e^{\mu X_{v_0}}(Y(v_1) + Y(v_2))$$

where X_{v_0} is the cost at the node v_0. Utilizing this recursion formula we can efficiently calculate the Bayes statistic (9.3). To simulate the Bayes statistic for fixed μ and m under H_0 and H_1, we used 2,000 realizations in each case.

9.6.1.2 Simulating the Strip statistic

For a graph G_m and a positive integer B the strip statistic, $T_{m,B}$, is the sum of the random variables that fall in the centered strip of length m and width $2B+1$, or

$$T_{m,B} = \sum_{0 \le i \le m-1} \sum_{j:|j| \le \min(i,B)} X_{i,j}.$$

We know the distribution of this statistic under H_0 since the variance $\sigma^2_{m,B}$ is $n_{m,B}$, the number of vertices in the strip, and therefore $T_{m,B} \sim N(0, n_{m,B})$. Under H_1 we let $R_{m,B}$ be the number of vertices inside the strip that the random walk $\{S_k\}_{0 \le k \le m-1}$ visits. If the mean along the random path is equal to μ, we observe that

$$T_{m,B} = \mu R_{m,B} + W,$$

where $W \sim N(0, n_{m,B})$ and W and $R_{m,B}$ are independent. Therefore, we simulate $T_{m,B}$ by taking one realization of $R_{m,B}$, multiplying it by μ and adding a realization of W.

To choose the width of the strip, we ran simulations for $B = \nu\sqrt{m}$ for $\nu = 0.75, 1, 2, 3$. Among these values and the graph sizes we simulated for, $B = 2\sqrt{m}$ gave the best performance. To estimate the detection rate for fixed μ and m we used 5,000 realizations of the test statistic.

9.6.1.3 Simulating the GLRT

The GLRT statistic rejects the null hypothesis for large values of $M_m = \max\{X_p : p \in \mathcal{P}_m\}$. This statistic can be calculated rapidly using dynamic programming. For each graph size, the threshold corresponding to approximately 5% Type I error probability and the detection rate for fixed μ were based on 10,000 and 1,000 realizations, respectively.

m	1,025	2,049	4,097
$\mu_{.95}$	0.37	0.31	0.26

Table 9.1: The value of μ giving detection rate about 95% using the Bayes test when the unknown path is sampled uniformly over all paths

m	1,025	2,049	4,097	8,193	16,385
$\mu_{.95}$	0.84	0.69	0.59	0.51	0.42

Table 9.2: The value of μ giving detection rate about 95% using the Strip statistic test with width $2B+1$, $B = 2\sqrt{m}$, when the unknown path is sampled uniformly over all paths

9.6.1.4 Comparing the tests

From plots of the detection rates versus the mean μ (similar to figure 9.11), we can estimate $\mu_{.95}$, the mean μ which gives the detection rate of about 95%. Tables 9.1, 9.2, and 9.3 show $\mu_{.95}$ for different graph sizes for the Bayes test, the test based on the Strip statistic, and the GLRT, respectively. As expected, the Bayes test performs better than the others, but the reader should recall that the latter two tests do not require information about the parameter μ, while the Bayes test does. Figure 9.8 shows $\log_2(\mu_{.95})$ plotted versus $\log_2(m)$ with least-squares line fit to the data. The slope of the line is about -0.255 in the case of the Bayes test and about -0.246 in the case of the test based on the Strip statistic. Both of these values are quite close to the $-1/4$ exponent in Theorem 13. In the case of the GLRT the slope is about -0.16, perhaps suggesting, that for a big enough graph, the Strip statistic test might eventually outperform the GLRT. The fitted lines through the points corresponding to the Strip statistic and the GLRT meet at approximately $m = 2^{20} \approx 10^6$, but it would be very computationally intensive to do simulations for graphs of that size. More importantly, these simulations suggest that the GLRT is only able to detect at $\mu \asymp m^{-1/6}$, and therefore does not achieve the optimal detection rate under the uniform prior on paths.

m	1,025	2,049	4,097	8,193
$\mu_{.95}$	0.46	0.40	0.36	0.33

Table 9.3: The value of μ giving detection rate about 95% using the GLRT when the unknown path is sampled uniformly over all paths

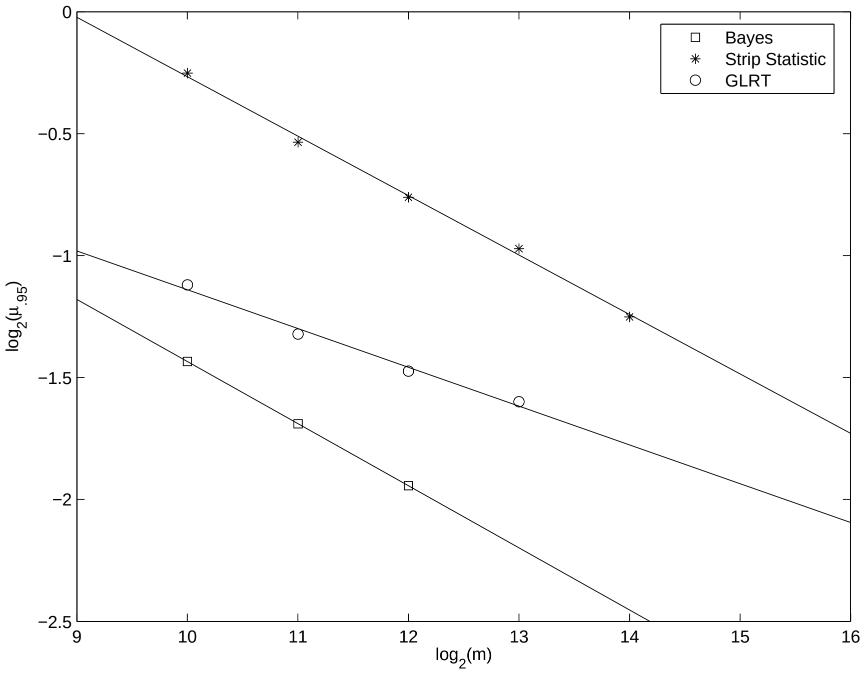


Figure 9.8: Comparison of the Bayes test, the Strip statistic test, and the GLRT under the uniform prior

9.6.2 Minimax detection

We focus on the diagonal path p, where $p_i = i, 0 \leq i \leq m-1$, as we believe this path to be most challenging for the GLRT to detect, and we compare the performance of the GLRT with the Weighted Average Statistic test (WAS) defined in (9.4). Note that it is equivalent to consider the prior that puts all its mass on this diagonal path.

Recall that, under H_0, the WAS is $\mathcal{N}(0, \lambda_m)$, and $\mathcal{N}(\mu, \lambda_m)$ under H_1, independently of the unknown path (with $\lambda_m \sim (\log m)^{-1}$). Therefore, for a power of 95% at level 5%, we need $\mu \geq 2z_{.95}\sqrt{\lambda_m}$, where $z_{.95}$ is the 95% standard normal quantile. Some power curves for the WAS are graphed in Figure 9.10. We use simulations to do the same for the GLRT in Figure 9.11, where each point on the curves is based on 1,000 realizations of the statistic.

While the power curves for the WAS clearly tend to translate to the left, this does not seem to be the case for the GLRT. This might be an indication that the detection threshold for the GLRT does not tend to zero as m increases, just as in the case of the binary tree.

m	1,025	2,049	4,097	8,193	16,385	32,769
$\mu_{.95}$	1.20	1.15	1.10	1.06	1.03	0.99

Table 9.4: The value of μ giving detection rate about 95% using the WAS test

m	1,025	2,049	4,097	8,193
$\mu_{.95}$	0.90	0.89	0.885	0.88

Table 9.5: The value of μ giving detection rate about 95% using the GLRT for detecting the diagonal path

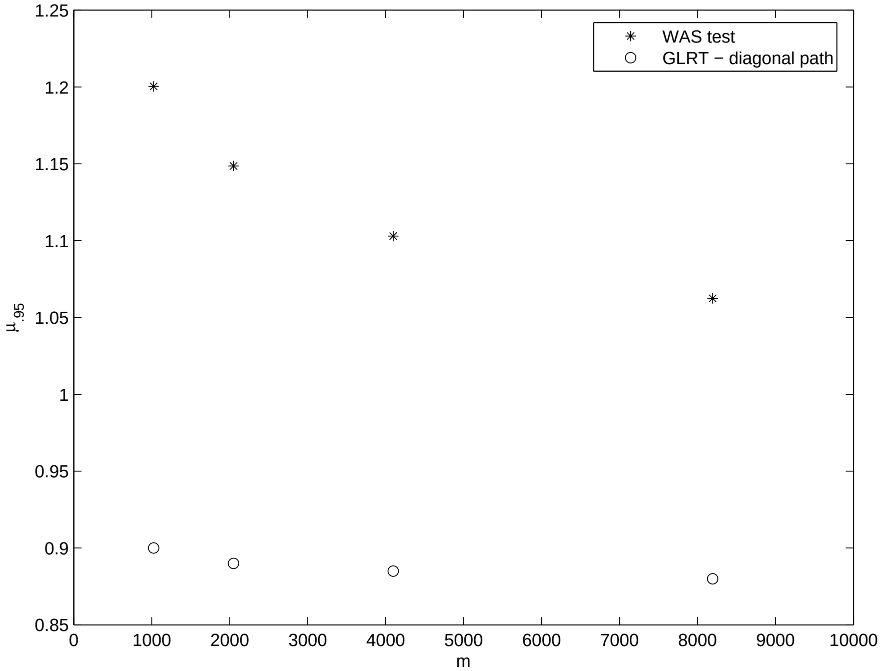


Figure 9.9: Comparison of the GLRT and the WAS when the abnormal path is the diagonal path

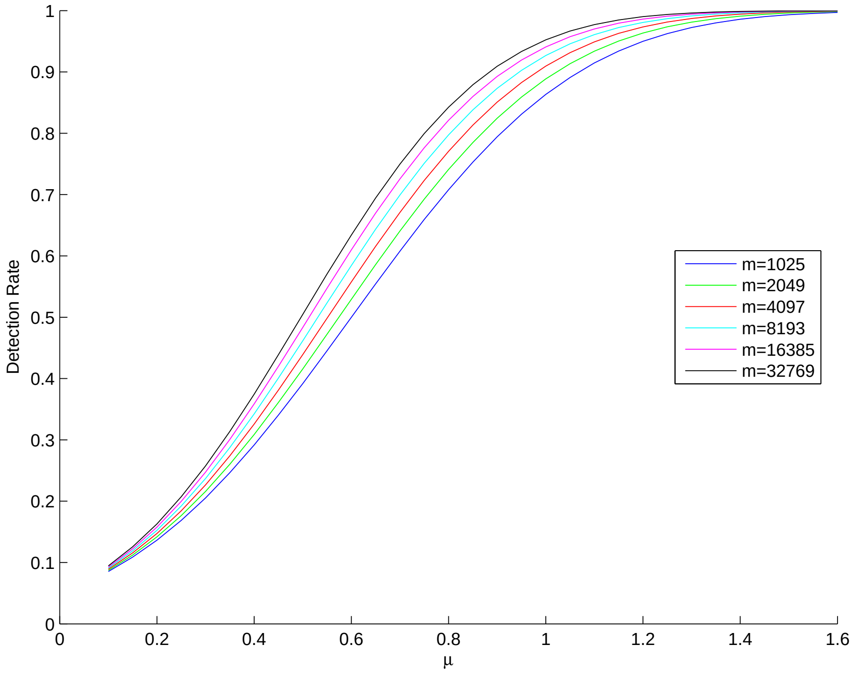


Figure 9.10: Detection rate curves for the WAS statistic for $m = 1025, 2049, 4097, 8193, 16385, 32769$. As m increases the curve moves to the left. Type I error is set to 5%.

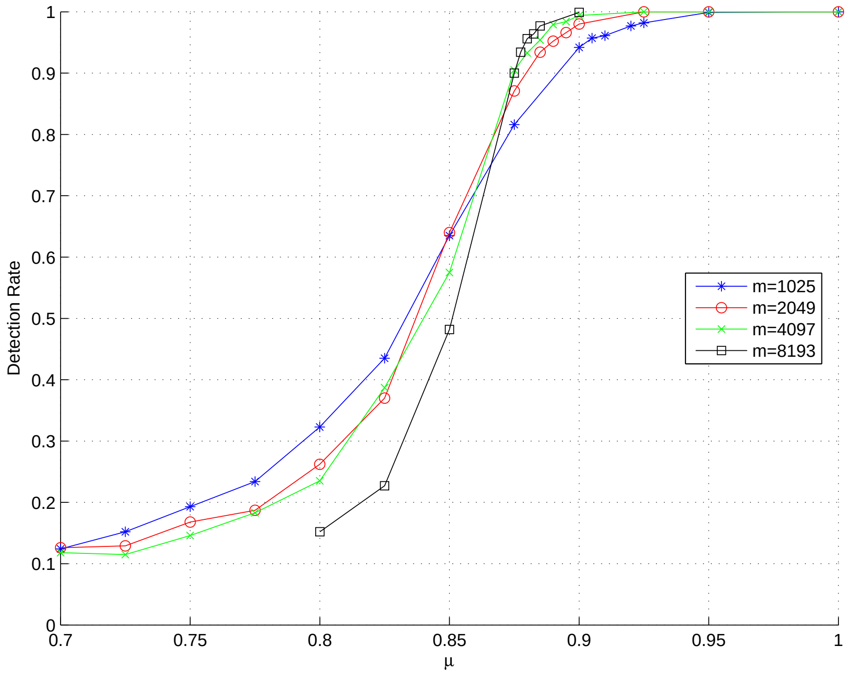


Figure 9.11: Detection rate curves for the GLRT for the diagonal path. The probability of Type I error is set to 5%.

Chapter 10

Concluding Remarks

In this last chapter of the thesis we will review our achievements and speculate about future directions and refinements. We will also partly address a question regarding the BP test which has arisen after our research on path detection in graphs; namely, whether we could design simpler and better tests than the BP test for detecting chirps.

10.1 ChirpLab

A part of this thesis was the development of a software package called *ChirpLab*, which is available at `http://www.chirplab.org`. It implements most of the algorithms proposed herein and can be used to reproduce the results. The software package also accompanies the ETD copy of this thesis (see `http://www.ndltd.org/`). This will serve as a more reliable archive and, since we expect there to be future changes to *ChirpLab*, this ensures that the original code which was used in this thesis will always be available and easily retrievable in the future.

10.2 BP Test and the Graph Problem

Our study of the abstract graph problem shows that we can achieve (near) optimal performance for path detection in graphs with very simple statistics. In the Bayesian problem for the triangular lattice, the optimal statistic is a crude sum over nodes that can be contained in a "strip." For the minimax problem, a statistic that gives a good upper bound is a simple weighted average over all the nodes in the graph. Besides that, the simulations for the graphs indicate that the GLRT does not achieve the optimal rate for detectability. For chirp detection, one might want to start searching for other alternatives than searching for the best chirplet path in the chirplet graph and perhaps look for simpler methods instead. However, there are some important remarks to make:

1. The theoretical results give optimality conditions in an asymptotic sense. For the graph sizes we considered in the numerical simulations, the GLRT was shown to perform better than the simple optimal test for the Bayes problem.

2. The graph problem is an ideal abstraction of the situation in the chirplet graph. The graph in our methodology has a much more complex topology and the costs at the nodes are not independent.

3. The typical size of a chirplet graph is *at most* moderate. In our numerical experiments we considered graphs where the length of paths never exceeded 16 chirplets.

4. We already have numerical evidence that there is not much room for improvement if the goal is to have an adaptive method method for nonparametric chirp detection. For example, in the case of detection of monochromatic and linear chirps, the price we pay for being adaptive is not great when compared to methods which rely on the prior knowledge of the chirp belonging to these small parametric classes. The methods we have proposed are also fast and flexible and a software implementation is publicly available.

10.3 Achievements

We have presented a flexible and practical methodology for estimating and detecting non-homogeneous oscillatory signals. Although the procedures were developed with chirp signals in mind, the methods could be generalized for other types of statistical problems where the idea of chaining local correlations makes sense. The estimation procedure is provably near-optimal over a wide range of chirps, and numerical experiments show that the method is promising. The numerical simulations also indicate that the BP test is very sensitive over a broad class of chirps and does not leave much room for improvement. We have extended the BP test so that with long streams of data, it can be applied rapidly without sacrificing much sensitivity. Therefore, our methods have the potential of being practical tools for solving real-world problems involving chirps.

In the search of developing theory for the detection of chirps we were lead to the study of the abstract problem of searching for paths in graphs. In collaboration with others we have pursued that problem and established theoretical results.

10.3.1 New methods for approximations with time-frequency atoms

Assume we are given a discrete collection of time-frequency atoms Ω as described in Section 1.2 and we have a class of signals we want to represent, perhaps approximately, as linear combinations of these atoms. A good representation could be one which gives the best approximation of a signal f where the number of atoms to use is fixed. If the notion of "neighboringness" between atoms, as for chirps and chirplets, makes sense for the class of signals we are considering, we could possibly utilize the ideas and algorithms presented here. We could consider an analogy to the chirplet graph, where each vertex v in the graph represent an atom w_v from Ω and the vertices are connected with edges according to some prescribed rule. Assume the atoms are normalized so that $\|w_v\| = 1$. Then, given a signal f which we wish to analyze, we could propose solving

$$\max_{W \in \mathcal{W}} \sum_{v \in W} |\langle f, w_v \rangle|^2; \tag{10.1}$$

where W is a set of connected nodes and $\mathcal{W}$ is the collection of all allowable paths. Whether we can solve the optimization problem rapidly depends on the topology of the graph, and we might need to resort to other types of network flow algorithms than we used in our methods. If we assume the topology is nice, the maximizer W^* of (10.1) would then give us a set of atoms which we could project the original function onto, giving

$$\tilde{f} = \sum_{v \in W} \langle f_v, w_v \rangle w_v,$$

as a possible approximation to f_v. Note that, since the atoms could have overlapping support, solving (10.1) is not necessarily equivalent to solving

$$\min_{W} \min_{(\alpha_v)} \|f - \sum_{v \in W} \alpha_v w_v\|^2.$$

Nevertheless, the approximation we get by solving (10.1) could still be good. Thus, our methodology provides an alternative to the best-basis algorithm [28], matching pursuit [58], and basis pursuit [27] for finding approximations of signals using time-frequency atoms.

10.4 Future directions

The methods we have developed are aimed at detecting and estimating single chirps. That is, signals of the form $f(t) = A(t)\cos(N\varphi(t))$. One could imagine practical applications where the unknown signal is a superposition of chirps with distinct amplitudes and phases, i.e., $f(t) = \sum_k A_k(t)\cos(N\varphi_k(t))$. Although the notion of a chirplet graph could still make sense, it is not obvious how one should extend the current methods to deal with this situation. Perhaps other methods based on new ideas have to be developed.

It would be interesting to explore the BP estimator further and eventually apply it to practical problems. The same holds for the BP test where perhaps the most interesting application would be gravitational wave detection. Especially since now we have a strategy to process long streams of data where the support of the chirp is unknown.

We have seen that multiscale chirplets have good approximation properties for chirps [20]. Since the dictionaries are of reasonable size and we have fast algorithms, it might be worthwhile investigating whether these mathematical tools could be practical for processing audio signals. For the sake of curiosity, we have tried applying the method of thresholding in the best chirplet frame on segments of old audio recordings and managed good qualitative results compared to other related methods that use cosine packets [16].

We would like to further study the theoretical performance of the BP test. So far we have only provided an upper bound for detection and a good lower bound is missing. Based on our studies for the abstract graph problem, we expect that the upper bound we provided for the BP test is not sharp in a Bayesian setting when every chirp is equally likely to appear in the data.

At last, we mention several open questions and extensions we could study for the problem of path detection in graphs:

- Theoretical study of the GLRT in the triangular and square lattice. From simulations we suspect that it does not achieve the optimal rate in the Bayesian and minimax settings we consider.

- Sharpening the results we already have for the Bayesian problem on the square lattice and the minimax problem for the triangular lattice. Also it would be interesting to study the minimax problem for the square lattice.

- Study variations of the problem. Such as when the unknown path could be of a length shorter than the graph and at unknown location. Another variation would be to study the problem of detecting more general sets, or regions, of connected nodes where the mean is elevated. We could imagine studying graphs with other topologies.

- Investigation of problems where the variables associated with the nodes are correlated or when the means along the unknown path are not all equal. Also, to get closer to the situation in the monoscale BP test, study the problem when means decay away from the path instead of being set to zero.

- Since our current results are asymptotic, we would also want to know what types of tests exhibit good performance for moderate sample sizes. Note that, as seen from the simulations, we can have asymptotically optimal tests (e.g., the test based on the strip statitistic) that do not perform better than the GLRT for the range of graph sizes we considered.

Appendix A

Formula for Chirplet Cost in Case of Real-Valued Signals

In Section 3.3.5 we claimed there is an analytic formula for calculating the chirplet cost

$$C(v) = \max_{\varphi_0} \frac{\left|\sum_{t\in I} y_t \cos(\varphi_\mu(t) + \varphi_0)\right|^2}{\sum_{t\in I} \cos^2(\varphi_\mu(t) + \varphi_0)}, \tag{A.1}$$

where the chirplet is indexed by $v = (I, \mu)$ and $\varphi_\mu(t) = a_\mu t^2/2 + b_\mu t$. Consider φ_μ and I fixed and write

$$f(\varphi_0) = \frac{\left|\sum_{t\in I} y_t \cos(\varphi_\mu(t) + \varphi_0)\right|^2}{\sum_{t\in I} \cos^2(\varphi_\mu(t) + \varphi_0)}.$$

Using the trigonometric identity $\cos(u + v) = \cos(u)\cos(v) - \sin(u)\sin(v)$, we have

$$g(\varphi_0) = \frac{A^2 \cos^2 \varphi_0 - 2AB \sin\varphi_0 \cos\varphi_0 + B^2 \sin^2 \varphi_0}{C^2 \cos^2 \varphi_0 - 2D \sin\varphi_0 \cos\varphi_0 + E^2 \sin^2 \varphi_0},$$

where

$$A + \imath B = \sum_{t\in I} y_t \, e^{\imath\varphi_\mu(t)},$$

and

$$C^2 = \sum_{t\in I} \cos^2 \varphi_\mu(t), \quad D = \sum_{t\in I} \cos\varphi_\mu(t) \sin\varphi_\mu(t), \quad E^2 = \sum_{t\in I} \sin^2 \varphi_\mu(t).$$

Further manipulations of $g(\varphi_0)$, using the identities

$$\cos(2u) = 2\cos^2(u) - 1 = 1 - 2\sin^2(u), \quad \sin(2u) = 2\sin(u)\cos(u),$$

give us

$$g(\varphi_0) = \frac{a\cos(2\varphi_0) + b\sin(2\varphi_0) + c}{d\cos(2\varphi_0) + e\sin(2\varphi_0) + f},$$

where

$$a = A^2 - B^2, \quad b = -2AB, \quad c = A^2 + B^2,$$

and

$$d = C^2 - E^2, \quad e = -2D, \quad f = C^2 + E^2 = |I|.$$

The critical points of g need to satisfy $g'(\varphi_0) = 0$. Differentiating with respect to φ_0, we get

$$g'(\varphi_0) = 2 \cdot \frac{(af - cd)\sin(2\varphi_0) + (ce - bf)\cos(2\varphi_0) + (ae - bd)}{(d\cos(2\varphi_0) + e\sin(2\varphi_0) + f)^2}.$$

Let consider two cases: (i) $(af - cd)^2 + (ce - bf)^2 = 0$, and (ii) $(af - cd)^2 + (ce - bf)^2 \neq 0$:

(i) Note that $af = cd$ and $ce = bf$ and since $f = |I| > 0$,

$$ae - bd = \frac{1}{f}(afe - bdf) = \frac{1}{f}(cde - cde) = 0.$$

Hence, in case (i), $g'(\varphi_0) = 0$ for all φ_0, and $g(\varphi_0)$ is constant. Since $c = A^2 + B^2 > 0$ (unless $y_t = 0$ for all $t \in I$), we have

$$\begin{aligned} g(\varphi_0) &= g(0) = \frac{a + c}{d + f} \cdot \frac{f}{c} \cdot \frac{c}{f} = \frac{af + cf}{cd + df} \cdot \frac{c}{f} = \frac{cd + cf}{cd + cf} \cdot \frac{c}{f} = \frac{c}{f} \\ &= \frac{A^2 + B^2}{C^2 + E^2} = \frac{|\sum_{t \in I} y_t \, e^{\imath\varphi_\mu(t)}|^2}{|I|}. \end{aligned}$$

(ii) We can write the numerator of $g'(\varphi_0)$ as

$$h(\varphi_0) = \frac{\alpha}{\rho}\sin(2\varphi_0) + \frac{\beta}{\rho}\cos(2\varphi_0) + \gamma, \tag{A.2}$$

where we have factored out

$$\rho = \sqrt{(af - cd)^2 + (ce - bf)^2},$$

and

$$\alpha = (af - cd)/\rho, \quad \beta = (ce - bf)/\rho, \quad \gamma = (ae - bd)/\rho.$$

> The extrema of g occur where $h(\varphi_0) = 0$. Writing $x = \cos(2\varphi_0)$ we have $\sin^2(2\varphi_0) = 1 - x^2$, and solving $h(\varphi_0) = 0$ leads to finding the solution of the quadratic equality
>
> $$\alpha^2(1 - x^2) = (\rho\gamma + \beta x)^2.$$
>
> Thus, we have simple analytic formulas for finding the values of $\cos(2\varphi_0)$ and $\sin(2\varphi_0)$ at the extrema, which can then be used to find the maximum value of g.

If we do not have to check the value of g for all the solutions of the equality above, we could go further in the analysis and, for example, look at signs of the derivative of g (i.e., the signs of h in case (ii)). But we have already made our point: It is indeed possible to calculate (A.1) using simple formulas involving A, B, C, D, and E. As we have already discussed, C, D, and E do not depend upon the data and can therefore be calculated offline. Hence, since A and B are given by the chirplet coefficient, the computational complexity of the exact calculation of the chirplet costs for real-valued data is only a small constant factor times the computational complexity of the chirplet transform.

Appendix B

Imposing Continuity of the Phase for Chirp Detection

Consider the data

$$y_t = \alpha s_t + z_t, \quad t = 0, 1, \ldots, N-1;$$

here, α is a real unknown scalar, s_t is an unknown signal of the form $s_t = \cos(\phi(t))$ where the phase ϕ is smooth, and z_t is noise which we take to be a vector of i.i.d. standard Gaussian variables. Given y_t, the goal is to decide between,

$$H_0 : \alpha = 0, \quad \text{i.e., the data is only noise,}$$

and

$$H_1 : \alpha \neq 0, \quad \text{i.e., there is a chirping signal } s_t \text{ in the data.}$$

B.1 A Test Based on Imposing Continuity of the Phase

We start by dividing the time interval $I := \{0, 1, \ldots, N-1\}$ into sub-intervals $\{I_v\}$. Let $\{\phi_v\}$ be a collection of "phaselets," where $\text{supp}(\phi_v) = I_v$ and ϕ_v is a quadratic polynomial in t on its support. For simplicity, we restrict ourselves to collections $W = \{\phi_v(t)\}$ such that $\phi_W(t) := \sum_v \phi_v(t)$ is a continuous function of t with a continuous first derivative and $\phi_W(0) = 0$. Consider now

$$f(t) = \cos(\phi_W(t) + \phi_0) = \sum_v \cos(\phi_v(t) + \phi_0) 1_{I_v}(t),$$

where ϕ_0 is a real scalar. Suppose that the set of alternatives is of the form λf where $\lambda \in \mathbb{R}$ – i.e., we have a set of chirping signals with constant amplitudes, and with phase functions which are globally C^1 and piecewise quadratic. The GLRT takes the form

$$\min_{W,\lambda,\phi_0} \|y - \lambda f\|^2 = \min_{W,\phi_0} \left[\|y\|^2 - \left(\frac{\langle y, f \rangle}{\|f\|} \right)^2 \right]$$

or

$$\max_{W,\phi_0} \left(\frac{\langle y, f \rangle}{\|f\|} \right)^2 = \max_{W,\phi_0} \frac{(\sum_v \langle y, f_v \rangle)^2}{\|f\|^2}.$$

Note that the denominator $\|f\|^2$ depends on W and ϕ_0. Therefore, in general, this test statistic is not additive and cannot be computed rapidly using ideas from dynamic programming. If we restrict ourselves to highly oscillatory signals, the norm of f becomes essentially constant and approximately equal to $N/2$; that is, independent of the value of the phase offset ϕ_0 and of the path W (due to the Riemann-Lebesgue lemma). Hence, in the high-frequency regime, the GLRT is approximately equivalent to

$$\max_{W,\phi_0} \langle y, f \rangle = \max_{W,\phi_0} \sum_v \langle y, f_v \rangle, \tag{B.1}$$

where we put $f_v(t) := \cos(\phi_v(t) + \phi_0) 1_{I_v}(t)$. Technically, we need to maximize the absolute value of the sum, but if the discretization includes f_v and $-f_v$ for each v, one can just as well work with the signed sum. Maximizing with respect to ϕ_0 gives [1]

$$\max_W |\langle y, \exp(i\phi_W) \rangle| = \max_W |\sum_v \langle y, \exp(i\phi_v) 1_{I_v} \rangle|.$$

This optimization problem cannot be solved rapidly using dynamic programming since the functional is not additive. Let us step back then and reconsider (B.1). For each fixed ϕ_0, the functional we need to maximize is additive. Therefore, by considering a discrete set of initial phase offsets and solving a sequence of optimization problems for different values of ϕ_0, one could calculate this test statistic using dynamic programming. This would of course not give the exact maximum but if the discretization of ϕ_0 is sufficiently fine, this ought to do the job. The key point here is that this test statistic is, in fact, very similar to the BP

[1] Here we have used $\langle y, f \rangle = Re\left(\langle y, e^{i(\phi_W + \phi_0)} \rangle \right) = Re\left(e^{-i\phi_0} \langle y, e^{i\phi_W} \rangle \right) \leq |\langle y, e^{i\phi_W} \rangle|$. By writing $\langle y, e^{i\phi_W} \rangle = \rho e^{i\nu}$ where $\rho = |\langle y, e^{i\phi_W} \rangle|$ we see that the equality holds when $\phi_0 = \nu$.

statistic presented in the paper. The notion of a "chirplet graph" is still valid, and one can use the same network flow algorithms, although the costs at the vertices have to be handled differently.

B.2 Calculating the Statistic

When computing the BP statistic, evaluating the costs at each vertex of the graph and running the optimization alogorithm to find the best paths are two perfectly decoupled procedures. In (B.1), although there is a coupling owing to the continuity of the phase, one can still use the same network flow algorithms. Here each node v in the chirplet graph corresponds to a phase function $\phi_v(t) = \frac{1}{2}a_v(t-t_0)^2 + b_v(t-t_0) + c_v$ supported on a time interval I_v. The parameters a_v and b_v determine the slope and frequency offset of a chirplet (the chirp rate is a_v) but now we have an extra parameter c_v to worry about, which needs to be adapted to guarantee the phase continuity. In addition to keeping track of the distance labels in the network flow algorithms, we also keep track of the phase offset c_v at each vertex. As we will see, the parameter c_v is automatically determined by the continuity requirement.

Suppose ϕ_0 is fixed, and consider two chirplets corresponding to vertices v and v', where there is an arc going from v to v'. Assume that the distance label at v is optimal, meaning that the best path up to v has been determined, and c_v is known. Note that if v is a starting node in a chirplet graph, its distance label is optimal and $c_v = \phi_0$. The cost at the vertex v is given by

$$\begin{aligned}\langle y, \cos(\phi_v)\rangle &= \mathrm{Re}\left(\langle y, e^{i(a_v/2(t-t_0)^2+b_v(t-t_0)+c_v)}\rangle\right)\\ &= \mathrm{Re}\left(e^{-ic_v}\langle y, e^{i(a_v/2(t-t_0)^2+b_v(t-t_0))}\rangle\right).\end{aligned}$$

Now let $I_v = [t_0, t_1]$ and $I_{v'} = [t_1, t_2]$ be the time supports for f_v and $f_{v'}$ with phase functions

$$\phi_v(t) = \frac{1}{2}a_v(t-t_0)^2 + b_v(t-t_0) + c_v$$

and

$$\phi_{v'} = \frac{1}{2}a_{v'}(t-t_1)^2 + b_{v'}(t-t_1) + c_{v'}.$$

The continuity of the phase, $\phi_v(t_1) = \phi_{v'}(t_1)$ imposes

$$c_{v'} = \frac{1}{2}a_v(t_1 - t_0)^2 + b_v(t_1 - t_0) + c_v.$$

Updating the distance label of v' using the cost for v tells us how to update the phase offset at v'.

Remark. We need to calculate the chirplet costs $C(v) = \langle y, e^{i(a_v(t-t_0)^2+b_v(t-t_0))}\rangle$ only once. Updating the costs based on the phase offsets is just a linear combination of the real and imaginary part of $C(v)$.

B.3 Some Criticism

- *Computational cost:* Clearly, the computation of the test statistic is more expensive than that of the BP statistic since one would need to discretize the initial phase offset ϕ_0. If the cardinality of the discrete set of values of ϕ_0 is M, the computational complexity of the statistic is M times greater. Unless the discretization is extremely coarse, this ratio is substantial.

- *Problems with colored noise:* We have assumed a white noise and chirp signals with a constant amplitude. Suppose now that the noise is colored, namely, $z := (z_0, z_1, \ldots, z_{N-1}) \sim N(0, \Sigma)$. Consider the same set of alternatives f as earlier. The GLRT gives

  $$\min_{W,\phi_0} (y - \lambda f)^T \Sigma^{-1}(y - \lambda f) = \min_{W,\phi_0} \|\tilde{y} - \lambda\tilde{f}\|^2$$

  where $\tilde{y} = \Sigma^{-1/2}y$ and $\tilde{f} = \Sigma^{-1/2}f$. Equivalently we could base our decision upon

  $$\max_{W,\phi_0} \left(\frac{\langle\tilde{y}, \tilde{f}\rangle}{\|\tilde{f}\|}\right)^2 = \max_{W,\phi_0} \frac{(y^T\Sigma^{-1}f)^2}{f^T\Sigma^{-1}f} = \max_{W,\phi_0} \frac{\left(\sum_v \langle\Sigma^{-1}y, f_v\rangle\right)^2}{f^T\Sigma^{-1}f}$$

  This test statistic looks almost like what we had earlier, with y replaced by $\Sigma^{-1}y$. The problem is that the denominator, $f^T\Sigma^{-1}f$ cannot, in general, be approximated by a constant at high frequencies. The functional will not be additive, which prevents the use of dynamic programming.

- *Restriction to high frequencies:* In order for the test statistic to be rapidly computable, we would need to restrict ourselves to highly oscillatory chirps so that $\|f\|$ is essentially constant and independent of the collection of phaselets. It is not clear what to do at low frequencies.

- *Imposing continuity of the phase does not seem to improve detection:* Our numerical simulations indicate that we do not gain much by imposing continuity of the phase as explained in this section. The BP test gives almost the same statistical performance while being computationally far less expensive, and being amenable to important extensions concerning colored noise and varying amplitude.

B.4 BP Test: Local Fit of Amplitude and Phase Offset

Now consider the case where we do not impose continuity on the phase and, instead, fit the amplitude and phase offset locally. Two key observations follow:

(i) Locally, a smooth chirp has a simple structure which asserts that at sufficiently small scale, a smooth chirp has an almost linear instantaneous frequency;

(ii) The time-frequency portrait of a smooth chirp peaks around a ridge determined by the instantaneous frequency (a smooth curve in the time-frequency plane).

The chirp detection paper proposes a two-step strategy for detection:

1. Calculate local fit of “chirplets” with data.
2. Look for good global fit by chaining together chirplets in a meaningful way.

As a measure of local fit, we could use the GLRT principle while a meaningful global fit could be based on chirplet paths in the chirplet graph. In the white noise model, we have a nice interpretation based on the GLRT. Let $W = \{\phi_v\}$ be a collection of phase functions such that $\phi_v(0) = 0$. Consider

$$f(t) = \sum_v \lambda_v f_v(t),$$

where $f_v(t) := \cos(\phi_v(t) + \phi_{v,0})1_{I_v}(t)$, and $\{\lambda_v\}$ and $\{\phi_{v,0}\}$ are collections of real scalars. Now apply the GLRT principle:

$$\min_{W,\lambda_v,\phi_{v,0}} \|y - f\|^2 = \min_{W,\lambda_v,\phi_{v,0}} \sum_v \|y1_{I_v} - \lambda_v f_v\|^2 = \min_{W,\phi_{v,0}} \|y\|^2 - \sum_v (\langle y, f_v\rangle)^2/\|f_v\|^2.$$

At high frequency, the approximation $\|f_v\|^2 \approx |I_v|/2$ is tight, and the test is nearly equivalent to

$$\max_{W,\phi_{v,0}} \sum_v |\langle y, f_v\rangle|^2/|I_v|.$$

Maximizing with respect to $\phi_{v,0}$ gives us

$$\max_W \sum_v |\langle y, |I_v|^{-1/2}\exp(i\phi_v)1_{I_v}\rangle|^2, \tag{B.2}$$

which is identical to the BP statistic, assuming the white noise model.

Remark. For colored noise, this interpretation does not hold. The GLRT asks to solve

$$\min_{W,\lambda_v,\phi_{v,0}} (y-f)^T\Sigma^{-1}(y-f) = \min_{W,\lambda_v,\phi_{v,0}} \|\Sigma^{-1/2}\tilde{y} - \sum_v \lambda_v \tilde{f}_v\|^2$$

where $\tilde{y} = \Sigma^{-1/2}y$ and $\tilde{f}_v = \Sigma^{-1/2}f_v$. Although the time supports of the f_v's are disjoint, this does not necessarily hold for the $\tilde{f}_v$'s, and one cannot take the sum outside the squared norm. We are stuck with this untractable expression. Although this type of GLRT interpretation does not motivate a candidate test statistic, one could still measure the local fit by considering $(y_v - \lambda_v f_v)^T\Sigma^{-1}(y_v - \lambda_v f_v)$, where $y_v = y1_{I_v}$. Doing so gives us the chirplet costs for colored noise as introduced in the Chirp Detection paper.

B.5 Numerical Simulation

We generated data of the form

$$y_k = \alpha s_k + z_k, \quad k = 0, 1, \ldots, N-1;$$

(s_k) is a vector of equispaced time samples of a real-valued chirp and (z_k) is a sequence of i.i.d. $N(0,1)$. We define the *Signal-to-Noise-Ratio* (SNR) as

$$SNR = \frac{\|\alpha s\|}{\sqrt{N}}.$$

The signal length is $N = 1024$ and the test signal is a *cubic phase chirp* with constant amplitude, $s(t) = \cos(\phi(t))$:

$$\phi(t) = 2\pi N(t^3/12 + t/8).$$

The signal was sampled at $t_k = k/N$, $k = 0, 1, \ldots, N-1$. We computed the test statistic (B.1) where the maximization is over a grid of phase offsets ϕ_0 with 100 equally spaced values from 0 to 2π. We call this *Test 1.* The second test we considered was based on the BP statistic as in (B.2); we call it *Test 2.*

We use the same chirplet graph for both test statistics (same chirp rates and base frequencies). For simplicity, we used a single scale of the form 2^{-s} and set $s = 2$. Thus, the time axis is divided into 4 equally long segments. We restricted ourselves to discrete frequencies b_v in the interval $\{100, \ldots, 400\}$. The slope parameters are equal to $a_v = 2\pi(-1/2 + k \cdot 2^s/N)$, where $k \in \{0, \ldots, N/2^s\}$ and the phase of the chirplet is $\phi_v(t) = a_v t^2/2 + b_v$. This discretization gives a very good correlation with the signal in the noiseless case (close to 0.95 in the case of Test 1). For each test statistic we repeated the following steps:

- Randomly sample 1,000 realizations of white noise which are used to select a detection threshold giving a probability of false detection equal to 5%.
- For each SNR, sample the data model 1,000 times in order to compute detection rates.

Figure B.1 compares the detection rates for both methods. The performance is nearly identical.

B.6 Conclusions

Our numerical simulations indicate that there is not much to gain in terms of statistical sensitivity by forcing the continuity of the phase.

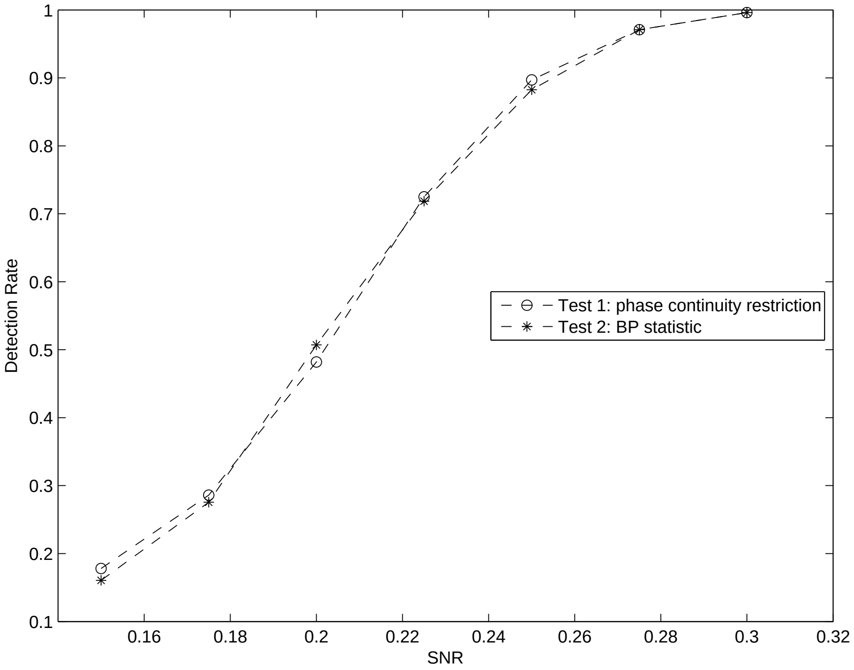


Figure B.1: Comparison of the performance of Test 1 which is based on (B.1) and Test 2, which is based on the BP statistic. The probability of Type I error is fixed at 5%.

Appendix C

Proofs and Lemmas

C.1 Taylor Approximation for Functions in HÖLDER$^s(R)$

The following lemma for Taylor approximations of functions in HÖLDER$^s(R)$ is useful.

Lemma 16. *Assume $f \in$ HÖLDER$^s(R)$ with $s \in [2,3]$ such that $m < s \leq m+1$ where m is an integer. Fix an interval $I \subset [0,1]$ and pick a $t_0 \in I$. Then*

$$f(t) = \sum_{k=0}^{m-1} \frac{f^{(k)}(t_0)}{k!}(t-t_0)^k + \epsilon(t), \quad \forall t \in I$$

where

$$|\epsilon(t)| \leq K \cdot \sup_{t \in I} |t-t_0|^s \leq K \cdot |I|^s,$$

and $K = \frac{\|f\|_s}{s}$ for $s = 2$ and $K = \frac{\|f\|_s}{s(s-1)}$ for $s \in (2,3]$. Also, for $s \in (2,3]$,

$$f'(t) = f'(t_0) + f''(t_0)(t-t_0) + \gamma(t), \quad \forall t \in I$$

where

$$|\gamma(t)| \leq K \cdot \sup_{t \in I} |t-t_0|^{s-1} \leq K \cdot |I|^{s-1},$$

and $K = \frac{\|f\|_s}{s-1}$.

Proof. The proof of this lemma relies on standard arguments from calculus.

First consider $s = 2$. Then

$$|f'(x) - f'(y)| \leq \|f\|_s \cdot |x-y|^{s-1}.$$

Let $h_y(x) = f(x) - f(y) - f'(y)(x-y)$ for some fixed y. Note that $h_y(y) = 0$ and $h'_y(x) = f'(x) - f'(y)$. Then the fundamental theorem of calculus gives

$$\begin{aligned} |h_y(x)| &= |h_y(x) - h_y(y)| = \left| \int_y^x h'_y(u) du \right| \\ &\leq \|f\|_s \cdot \left| \int_y^x |u-y|^{s-1} du \right| = \frac{\|f\|_s}{s} \cdot |x-y|^s, \end{aligned}$$

and the result follows.

Next assume $s \in (2, 3]$. Then

$$|f''(x) - f''(y)| \leq \|f\|_s \cdot |x-y|^{s-2}$$

and analogously to the result above (think of f'' in the role of f'), we get

$$|f'(x) - f'(y) - f''(y)(x-y)| \leq \frac{\|f\|_s}{s-1} \cdot |x-y|^{s-1}. \tag{C.1}$$

Let $h_y(x) = f(x) - f(y) - f'(y)(x-y) - f''(y)/2 \cdot (x-y)^2$ for some fixed y. Then since $h_y(y) = 0$ and $h'_y(x) = f'(x) - f'(y) - f''(y)(x-y)$, we get by the fundamental theorem of calculus

$$\begin{aligned} |h_y(x)| &= |h_y(x) - h_y(y)| = \left| \int_y^x h'_y(u) du \right| \\ &\leq \frac{\|f\|_s}{s-1} \cdot \left| \int_y^x |u-y|^{s-1} du \right| = \frac{\|f\|_s}{s(s-1)} \cdot |x-y|^s, \end{aligned}$$

which finishes the proof. □

C.2 Proof of Lemma 3

Consider the continuous broken line connecting the points $(t_{0,I}, b_{\mu,I})$, where $b_{\mu,I}$ is of form as in (6.2). Choose the sequence of frequency offsets $(b_{\mu,I}) = (b_I)$ such that b_I is as close to $\varphi'(t_{0,I})$ as possible. For the right endpoint of the last interval I, b_I is as close to $\varphi'(t_{1,I})$ as possible. The discretization spacing for the frequency offsets gives

$$|b_I - \varphi'(t_I)| \leq \pi \Delta b, \tag{C.2}$$

where $t_I = t_{0,I}$, unless I is the last interval, in which case $t_I = t_{1,I}$. Take two adjacent intervals I and I' in $\mathcal{P}$ and let a_I be the slope of the line segment on the interval I. It is of the requested form in (6.2), since

$$a_I = \frac{b_{I'} - b_I}{|I|} = 2\pi \frac{\Delta b}{|I|} \cdot l, \text{ for some } l \in \mathbb{Z}.$$

Now we only have to prove that the continuous broken line, $\sum_{I \in \mathcal{P}} (b_I + a_I(t - t_{0,I}))1_I(t)$, satisfies the inequality (6.3).

Let

$$p_I(t) = \varphi'(t_{0,I}) + \frac{\varphi'(t_{1,I}) - \varphi'(t_{0,I})}{|I|}(t - t_{0,I}),$$

and write

$$h(t) = \varphi'(t) - p_I(t).$$

Since $h(t_{0,I}) = h(t_{1,I}) = 0$, we have by Rolle's theorem $h'(\tau) = 0$ for some $\tau \in (t_{0,I}, t_{1,I})$. Then for $s \in (2, 3]$,

$$|h'(t)| = |h'(t) - h'(\tau)| = |\varphi''(t) - \varphi''(\tau)| \le \|\varphi\|_s \cdot |I|^{s-2}.$$

By the fundamental theorem of calculus and the triangle inequality for integrals,

$$|h(t)| = |h(t) - h(t_{0,I})| = \left| \int_{t_{0,I}}^{t} h'(t)dt \right| \le \int_{t_{0,I}}^{t} |h'(t)|dt \le \|\varphi\|_s \cdot |I|^{s-1}.$$

For $s = 2$, we have a slightly stronger inequality by Lemma 17, or

$$|h(t)| \le \frac{\|\varphi\|_s}{2} \cdot |I|^{s-1}.$$

These bounds and a repeated use of the triangle inequality give (6.3):

$$\begin{aligned}
|\varphi'(t) - (b_I + a_I(t - t_{0,I}))| &\le |\varphi'(t) - p_I(t)| + |p_I(t) - (b_I + a_I(t - t_{0,I}))| \\
&\le \|\varphi\|_s \cdot |I|^{s-1} + |\varphi'(t_{1,I}) - b_{I'}| \cdot |t - t_{0,I}|/|I| \\
&\quad + |\varphi'(t_{0,I}) - b_I| \cdot |1 - |t - t_{0,I}|/|I|| \\
&\le \|\varphi\|_s \cdot |I|^{s-1} + 2\pi\Delta b.
\end{aligned}$$

Using the bounds we have gives,

$$|a_I| \quad = \quad \left|\frac{a_I|I| + b_I - \varphi'(t_{1,I})}{|I|} + \frac{\varphi'(t_{1,I}) + b_I}{|I|}\right| \leq R \cdot |I|^{s-2} + 3\pi\Delta b \cdot |I|^{-1}.$$

For $s = 2$ we can improve the constant in the bound slightly, since,

$$\begin{aligned} |a_I| &= \left|\frac{b_{I'} - b_I}{|I|}\right| = \left|\frac{b_{I'} - \varphi'(t_{0,I})}{|I|} + \frac{\varphi'(t_{1,I}) - b_{I'}}{|I|} + \frac{\varphi'(t_{0,I}) - \varphi'(t_{1,I})}{|I|}\right| \\ &\leq R + 2\pi\Delta b \cdot |I|^{-1}. \end{aligned}$$

If $\Delta b \leq |I|^{s-1}$,

$$|\varphi'(t) - (b_I + a_I(t - t_{0,I}))| \leq (R + 2\pi)|I|^{s-1},$$

and

$$|a_I| \leq (R + 3\pi) \cdot |I|^{s-2}.$$

Consider the interval $I = [t_0, t_1)$. Assume $\theta(t) = a/2 \cdot (t - t_0)^2 + b \cdot (t - t_0) + \varphi(t_0)$ such that $|\varphi'(t) - \theta'(t)| \leq C$ for a constant $C = C(R, |I|)$, depending on R and $|I|$. Note that $\varphi(t_0) = \theta(t_0)$. Let $g(t) := \varphi(t) - \theta(t)$. Then, from the fundamental theorem of calculus,

$$\int_{t_0}^{t} g'(u)du = g(t) - g(t_0) = g(t), \quad \forall t \in I,$$

and

$$\begin{aligned} |\varphi(t) - \theta(t)| &\leq \left|\int_{t_0}^{t} g'(u)du\right| \leq \int_{t_0}^{t} |g'(u)|du \leq \sup_{t \in I} |\varphi'(t) - \theta'(t)||I| \\ &\leq C \cdot |I|, \end{aligned}$$

which finishes the proof of the lemma. □

Lemma 17. *Assume $\varphi \in \text{HÖLDER}^s(R)$ with $s = 2$. Consider the interval $I = [t_0, t_1]$ with $|I| \leq 1$ and let*

$$p_I(t) = \varphi'(t_0) + \frac{\varphi'(t_1) - \varphi'(t_0)}{|I|}(t - t_0).$$

Then for any $t \in I$,

$$|\varphi'(t) - p_I(t)| \leq \frac{R}{2}|I|.$$

Proof. Write $f(t) = \varphi'(t)$. Define the parallelogram $\delta\Omega_f$ as the boundary of the region

$$\Omega_f = \{(y,t) : |y - f(t_0)| \leq R|t - t_0| \quad \text{and} \quad |y - f(t_1)| \leq R|t - t_1| \quad t \in I\}.$$

Write $df = f(t_1) - f(t_0)$. Since $\varphi \in \text{HÖLDER}^2(R)$, $\varphi'(t) = f(t) \in \Omega_f$ for all $t \in I$. The line segment

$$p_I(t) = f(t_0) + \frac{df}{|I|}(t - t_0), t \in I,$$

is the diagonal of the parallelogram joining the points $p_0 = (t_0, f(t_0))$ and $p_1 = (t_1, f(t_1))$. The maximum distance from this line segment to $\delta\Omega_f$ is at one of the vertices of the parallelogram different from p_0 or p_1. After some algebra, one can show that this maximum distance is equal to

$$\frac{|I|}{2R}\left(R^2 - \frac{df^2}{|I|}\right),$$

Note that this is indeed a positive quantity since $|df| \leq R|I|$ by the HÖLDER requirement. This quantity is maximized with respect to df when $df = 0$. Thus, the maximum possible distance is

$$\frac{|I|}{2R}R^2 = \frac{R}{2}|I|.$$

□

C.3 Proof of Lemma 5

Proof. The proof uses the same notation as the proof of Lemma 3. Let $\bar{t} = (t_{0,I} + t_{1,I})/2$. Consider the piecewise constant function $\sum_{I\in\mathcal{P}} b_I 1_I(t)$, such that b_I is of form as in (6.10) and b_I is as close to $\varphi'(\bar{t}_I)$ as possible. From the discretization step, and since $\varphi \in \text{HÖLDER}^2(R)$, we get for every $t \in I$,

$$|\varphi'(t) - b_I| \leq |\varphi'(t) - \varphi'(\bar{t})| + \pi\Delta b \leq R/2 \cdot |I| + \pi\Delta b.$$

Note that for $t \in I$, $|t - \bar{t}| \leq |I|/2$.

Let the interval I' be adjacent to I with $t = t^*$ at their juncture. Then (6.12) follows trivially from our previous bound and the triangle inequality:

$$|b_I - b_{I'}| \leq |b_I - \varphi'(t^*)| + |\varphi'(t) - b_{I'}| \leq R \cdot |I| + 2\pi\Delta b.$$

The proof of (6.13) follows the same arguments as given in the proof of Lemma 3. □

C.4 Proof of Lemma 6

Proof. Consider the piecewise constant function $\sum_{I\in\mathcal{P}} b_I 1_I(t)$, such that b_I is of form as in (6.14). The only thing that differs from the proof of Lemma 5 is the "connectivity" bound. Let the interval I' be adjacent to I with $t = t^*$ at their juncture, $b_I = 2\pi|I| \cdot m_1$, and $b_{I'} = 2\pi|I| \cdot (m_2 + 1/2)$ with $m_1, m_2 \in \mathbb{Z}$. Then

$$|b_I - b_{I'}| \leq |b_I - \varphi'(t^*)| + |\varphi'(t) - b_{I'}| \leq (1 + 2\pi) \cdot |I|,$$

and therefore,

$$|1/2 + (m_1 - m_2)| \leq 1 + 1/(2\pi) < 1.5.$$

If $m_1 - m_2 \geq 1$ or $m_1 - m_2 \leq -2$, then $|1/2 + (m_1 - m_2)| \geq 1.5$. Hence, the only possibilities are $m_1 - m_2 = 0$ or $m_1 - m_2 = -1$. □

C.5 Proof of Theorem 6

The proof given here is essentially the same as the proof in [20] but we add a couple of elementary proofs of some technical lemmas. We chose to include the proof in the thesis for completeness. First we consider the case of real-valued data and then give an argument showing that the estimation problem for complex-valued data is almost equally as hard.

C.5.1 Real-valued data

Let φ be a smooth function obeying $\|\varphi\|_s \leq R$ and *supp*$(\varphi) \subset [0, 1]$. One concrete choice would be to use *iterative sinusoids*. The family of iterative sinusoids, $\{\beta_n, n = 0, 1, 2, ...\}$, is defined by the induction:

$$\beta_0(t) = \sin(\pi/4(1 + t)) \text{ for } t \in [-1, 1], \quad \beta_{n+1}(t) = \beta_n(\sin(\pi/2t)) \text{ for } t \in [-1, 1], n > 0.$$

These functions have the nice property that for any $n \geq 0$, β_n has $2^n - 1$ vanishing derivatives at $t = -1, 1$. For $s \in [2, 3]$ we could choose $\varphi(t) = K \cdot \beta_2(2t - 1) \cdot 1_{[0,1]}(t)$, with an appropriate constant K to satisfy the Hölder condition.

Define for $m \in \mathbb{Z}^+$,

$$\varphi_{k,m}(t) = m^{-s}\varphi(m(t-t_k)), \quad t_k = k/m, k = 0, 1, \ldots, m-1.$$

For $t \in [0,1]$, construct the phase function

$$\Phi_\xi(t) = \sum_{k=0}^{m-1} \xi_k \varphi_{k,m}(t), \quad \text{with } \xi_k \in \{0,1\}.$$

Let $p = \lfloor s \rfloor$. Since $\varphi^{(p)}_{k,m}(t) = m^{p-s}\varphi^{(p)}(m(t-t_k))$ and $p - s < 0$,

$$|\varphi^{(p)}_{k,m}(t) - \varphi^{(p)}_{k,m}(t')| \le m^{p-s} \cdot R \cdot |t-t'|^{s-p} \le R \cdot |t-t'|^{s-p}, \quad 0 \le t, t' \le 1.$$

Therefore $\|\varphi_{k,m}\|_s \le R$ and $\Phi_\xi \in \text{HÖLDER}^s(R)$. This allows us to define a collection of chirps:

$$\mathcal{H}_m = \{f(t) : f(t) = \sin(N\Phi_\xi(t)), t \in [0,1]\},$$

where $\mathcal{H}_m \subset \text{CHIRP}(s, N, R)$. For any $t \in [0,1]$, at most one term in $\Phi_\xi(t)$ is non-zero since the supports of the $\varphi_{k,m}$'s are disjoint. Therefore, if we define

$$a_{k,m}(t) = \sin(N\varphi_{k,m}(t)),$$

we can write $f \in \mathcal{H}_m$ as

$$f(t) = \sum_{k=0}^{m-1} \xi_k a_{k,m}(t).$$

Since the functions $a_{k,m}(t)$ have disjoint support, they are orthogonal.

Let $\|a_{k,m}\|^2_{\ell_2} = \sum_{l=0}^{N-1} |a_{k,m}(l/N)|^2$. Lemma 18 and Lemma 19 give size estimates for the L_2 and l_2 norms of $a_{k,m}$:

Lemma 18. *Assume $Nm^{-s} < 1$. Then,*

$$\|a_{k,m}\|^2_{L_2} \le C_1 \cdot N^2 m^{-2s-1},$$

for some fixed positive constant C_1. If in addition $Nm^{-s} \le \sqrt{3/4} \cdot \|\varphi(t)\|_{L_2} / \|\varphi(t)^2\|_{L_2}$, then

$$\|a_{k,m}\|^2_{L_2} \ge C_2 \cdot N^2 m^{-2s-1}$$

for some fixed positive constant C_2.

Proof. Recall that $\cos(x) \geq 1 - x^2/2$ for all x. Then

$$\begin{aligned}
\|a_{k,m}\|_{L_2}^2 &= \int_0^1 \sin^2(N\varphi_{k,m}(t))dt = \int_{t_k}^{t_{k+1}} \sin^2\left(Nm^{-s}\varphi(m(t-t_k))\right)dt \\
&= m^{-1}\int_0^1 \left[\frac{1}{2} - \frac{1}{2}\cos(2Nm^{-s}\varphi(t))\right]dt \\
&\leq m^{-1}\int_0^1 \frac{1}{2}\cdot\frac{(2Nm^{-s}\varphi(t))^2}{2}dt = C\cdot N^2m^{-2s-1},
\end{aligned}$$

where $C = \|\varphi(t)\|_{L_2}^2$.

For the other direction we use the inequality $\cos(x) \leq 1 - x^2/2 + x^4/24$. Then

$$\begin{aligned}
\|a_{k,m}\|_{L_2}^2 &\geq \frac{m^{-1}}{2}\int_0^1 \left[1 - \left(1 - \frac{(2Nm^{-s}\varphi(t))^2}{2} + \frac{(2Nm^{-s}\varphi(t))^4}{24}\right)\right]dt \\
&= N^2m^{-2s-1}\int_0^1 \left[(\varphi(t))^2 - \frac{4(Nm^{-s})^2(\varphi(t))^4}{3}\right]dt \\
&\geq CN^2m^{-2s-1}
\end{aligned}$$

where C is a positive constant provided that $\int_0^1 \left[(\varphi(t))^2 - \frac{4(Nm^{-s})^2(\varphi(t))^4}{3}\right]dt > 0$. This condition holds if $Nm^{-s} \leq \sqrt{3/4}\cdot\|\varphi(t)\|_{L_2}/\|\varphi(t)^2\|_{L_2}$. A sufficient condition is $\|\varphi(t)^2\|_{L_2} \leq \sqrt{3/4}\cdot\|\varphi(t)\|_{L_2}$. □

Lemma 19. *Assume $Nm^{-s} < 1$ and $m < N^{4/(2s+1)}$. Then*

$$\|a_{k,m}\|_{\ell_2}^2 \leq D_1 \cdot N^3m^{-2s-1}.$$

If in addition, $Nm^{-s} \leq \sqrt{3/4}\cdot\|\varphi(t)\|_{L_2}/\|\varphi(t)^2\|_{L_2}$, then

$$\|a_{k,m}\|_{\ell_2}^2 \geq D_2 \cdot N^3m^{-2s-1}.$$

Proof. Note that $a_{k,m}(0) = a_{k,m}(1) = 0$ so $\frac{1}{N}\|a_{k,m}\|_{\ell_2}^2$ equals the discrete approximation of $\|a_{k,m}\|_{L_2}^2$ by the trapezoidal rule. Therefore

$$|\|a_{k,m}\|_{L_2}^2 - \frac{1}{N}\|a_{k,m}\|_{\ell_2}^2| \leq C\cdot N^{-2},$$

for some constant C. Since $m < N^{4/(2s+1)}$ implies $N^{-2} < N^2m^{-2s-1}$, the discretization error is negligible as compared to the upper and lower bounds of the L_2 norm from Lemma 18. Therefore the results follow. □

Restricting ourselves to the set $\mathcal{H}_m$ of chirplets, the estimation problem becomes the problem of determining which vertex of the hypercube generated the observed data. Let P be the orthogonal projection onto the span of the $a_{k,m}$'s. Then for any estimator $\hat{f}$:

$$\|P\hat{f} - f\|^2 = \|P\hat{f} - Pf\|^2 \leq \|\hat{f} - f\|^2.$$

Therefore, we can focus on estimators which are linear combinations of $a_{k,m}$'s,

$$\hat{f}_L = \sum_{k=0}^{m-1} \hat{\xi}_k a_{k,m}.$$

Since the $a_{k,m}$'s are orthogonal and all have the same L_2 norm, we have

$$\|\hat{f}_L - f\|^2_{L_2} = \| \sum_k (\hat{\xi}_k - \xi_k) a_{k,m} \|^2_{L_2} = \|\hat{\xi} - \xi\|^2_{\ell_2} \cdot \|a_{k,m}\|^2_{L_2}, \tag{C.3}$$

so the problem reduces to that of estimating ξ. Let $Y = f + z$ be the vector of data and define $T = (T_1, \ldots, T_m)$ by

$$T_k = \langle Y, a_{k,m} \rangle / \|a_{k,m}\|^2_{\ell_2} = \langle f, a_{k,m} \rangle / \|a_{k,m}\|^2_{\ell_2} + \langle z, a_{k,m} \rangle / \|a_{k,m}\|^2_{\ell_2} = \xi_k + z'_k,$$

where $z'_k \sim N(0, 1/\|a_{k,m}\|^2_{\ell_2})$. Then $T \sim N(\xi, \sigma_m^2 \cdot I)$, where

$$\sigma_m^2 = 1/\|a_{k,m}\|^2_{\ell_2}.$$

To get a good lower bound, we choose m such that the noise level σ_m is roughly of the same size as the coordinates. If we select

$$m(N) = A \cdot N^{3/(2s+1)}$$

then $Nm^{-s} < 1$ and $m < N^{4/(2s+1)}$ so all the conditions for Lemma 19 are satisfied (if needed, we could always choose the constant A appropriately for the technical condition $Nm^{-s} \leq \sqrt{3/4} \cdot \|\varphi(t)\|_{L_2} / \|\varphi(t)^2\|_{L_2}$ to hold – another option would be to choose φ such that the condition is always true whenever $Nm^{-s} < 1$). Then

$$\sigma_m^2 = 1/\|a_{k,m}\|_{\ell_2}^2 \geq C \cdot N^{-3} m^{2s+1} = C_1 \cdot N^{-3} N^{(6s+3)/2s+1} = C_2,$$

and

$$\sigma_m^2 = 1/\|a_{k,m}\|_{\ell_2}^2 \leq C_2,$$

for some fixed constants C_1 and C_2. Or in other words, with this choice of m, $\sigma_m^2 \sim 1$.

To construct the lower bound, assume $\sigma_m^2 \sim 1$ and consider the Bayesian estimation problem of estimating $\xi = [\xi_1, \ldots, \xi_m]$, $\xi_k \in \{0, 1\}$, from the observation $T \sim N(\xi, \sigma_m^2)$ for the prior $\pi(\xi)$ where the ξ_k's are i.i.d. with $P(\xi_k = 1) = P(\xi_k = 0) = 1/2$. Let the constant B be the Bayes risk for the problem of estimating the coordinate ξ_k from $T_k \sim N(\xi_k, \sigma_m^2)$. Since we are not interested in the values of constants we do not need to know the exact value of B but only that $B = O(1)$. The Bayes estimate is $\hat{\xi} = [\hat{\xi}_1, \ldots, \hat{\xi}_m]$, where $\hat{\xi}_k = E(\xi_k|T) = E(\xi_k|T_k)$, and the Bayes risk for estimating ξ is

$$E\|\hat{\xi} - \xi\|^2 = \sum_{k=0}^{m-1} E(\hat{\xi}_k - \xi_k)^2 = m \cdot B.$$

The lower bound for this Bayes risk is a lower bound for the minimax mean-squared error. For our choice of m and using equation (C.3) this gives

$$\inf_{\hat{f}} \sup_{f \in \mathcal{F}_N} E\|f - \hat{f}\|_{L_2}^2 \geq (m \cdot B) \cdot \|a_{k,m}\|_{L_2}^2 \geq C \cdot N^{-2(s-1)/(2s+1)}, \tag{C.4}$$

for some constant C. Since $g := f - \hat{f}_L$ is zero at $t = 0, 1$, the approximation error of the trapezoidal rule gives us

$$MSE(f, \hat{f}_L) = 1/N \|g\|_{\ell_2} \geq \|g\|_{L_2} - A \cdot N^{-2},$$

for some positive constant A. The approximation error is negligible compared to the lower bound in (C.4), since for $s \in [2, 3]$, $N^{-2(s-1)/(2s+1)} > N^{-2}$. This gives us the desired lower bound in (7.3) in the case of real-valued chirps:

$$R^*(\mathcal{F}_N) \geq C \cdot N^{-2(s-1)/(2s+1)},$$

for some positive constant C.

C.5.2 Complex-valued data

To construct a lower bound in the case of complex-valued data we will use the same phase functions as before and consider the set of functions that are embedded in CHIRP(s, N, R):

$$\mathcal{H}_m = \{f(t) : f(t) = \exp(iN\Phi_\xi(t)), t \in [0,1]\}.$$

Take some $f \in \mathcal{H}_m$. Then for $t \in [k/m, (k+1)/m)$, $\Phi_\xi(t) = \xi_k \varphi_{k,m}(t)$ so $f(t) = \exp(iN\xi_k\varphi_{k,m}(t))$, and therefore $f(t) = 1$ if $\xi_k = 0$ and $f(t) = \exp(iN\varphi_{k,m}(t))$ if $\xi_k = 1$. Thus we can write each element $f \in \mathcal{H}_m$ as

$$f(t) = 1 + \sum_{k=0}^{m-1} \xi_k \left[\exp(iN\varphi_{k,m}(t)) - 1\right] 1_{[k/m,(k+1)/m)}(t), = 1 + \sum_{k=0}^{m-1} \xi_k g_{k,m}(t),$$

where

$$g_{k,m}(t) = \left[\exp(iN\varphi_{k,m}(t)) - 1\right] 1_{[k/m,(k+1)/m)}(t) = b_{k,m}(t) + ia_{k,m}(t),$$

and $b_{k,m}(t) = \cos(N\varphi_{k,m}(t)) - 1$, and as before, $a_{k,m}(t) = \sin(N\varphi_{k,m}(t))$. This shows that $\mathcal{H}_m$ is a hypercube. Note that since the $g_{k,m}$'s have disjoint supports, they are orthogonal.

Assume the data is $Y = f + z = f + z_1 + iz_2$, where z_1 and z_2 are i.i.d. random vectors of i.i.d. $N(0, 1/2)$ variables. The problem is to estimate $f = \sum_k \xi_k b_{k,m} + i \sum_k \xi_k a_{k,m}$ from the data Y. As for the real-valued data, we can focus on estimators which are linear combinations of $g_{k,m}$'s: $\hat{f}_L = \sum_{k=0}^{m-1} \hat{\xi}_k g_{k,m}$. A similar isometry as before holds in this case as well:

$$\|\hat{f}_L - f\|_{L_2}^2 = \|\hat{\xi} - \xi\|_{\ell_2}^2 \cdot \|g_{k,m}\|_{L_2}^2, \tag{C.5}$$

so the problem reduces to estimate ξ. The minimax mean-squared error is bounded by the Bayes risk for the simpler problem of estimating ξ from the observation $T = (T_1, \ldots, T_m)$ where

$$T_k = \langle Y, a_{k,m}\rangle / \|a_{k,m}\|_{\ell_2}^2 = \left[\rho_m \cdot \xi_k + z'_{1,k}\right] + i\left[\xi_k + z'_{2,k}\right];$$

$z'_{1,k}$ and $z'_{2,k}$ are i.i.d. and distributed as $N(0, \sigma_m^2)$ with

$$\sigma_m^2 = 1/(2 \cdot \|a_{k,m}\|_{\ell_2}^2),$$

and $\rho_m = \langle b_{k,m}, a_{k,m}\rangle / \|a_{k,m}\|_{\ell_2}^2$. Take $m = A \cdot N^{3/(2s+1)}$ for a suitable positive constant A, and consider the same prior as before, or $\pi(\xi)$; the ξ_k's are i.i.d. with $P(\xi_k = 1) = P(\xi_k = 0) = 1/2$. The Bayes estimate is $\hat{\xi}_k = E(\xi_k | T_k)$ with $E(\hat{\xi}_k - \xi_k)^2 = B_2$. Similar to before, the Bayes risk for estimating ξ is

$$E\|\hat{\xi} - \xi\|^2 = m \cdot B_2.$$

Note that by the Cauchy-Schwarz inequality and this choice of m,

$$|\rho_m| \leq \frac{\|b_{k,m}\|_{\ell_2}}{\|a_{k,m}\|_{\ell_2}} \leq C \cdot N^{-(4s-1)/(2s+1)},$$

where the last inequality follows from Lemma 20 below.

Lemma 20. *Assume $m \leq N^{6/(4s+1)}$. Then*

$$\|b_{k,m}\|_{\ell_2}^2 \leq C \cdot N^5 m^{-4s-1},$$

for some fixed positive constant C.

Proof. Recall that $1 - \cos(x) \leq x^2/2$ for all x. Then

$$\begin{aligned} \|b_{k,m}\|_{L_2}^2 &= \int_0^1 (1 - \cos(N\varphi_{k,m}(t))^2 dt \leq \int_{t_k}^{t_{k+1}} \left(\frac{(Nm^{-s}\varphi(m(t-t_k)))^2}{2}\right)^2 dt \\ &= N^4 m^{-4s-1} \int_0^1 \frac{\varphi(t)^4}{4} dt = C \cdot N^4 m^{-4s-1}, \end{aligned}$$

where $C = \|\varphi(t)^2\|_{L_2}^2/2$. The bound for the ℓ_2 norm follows then from the accuracy of the trapezoidal rule which is negligible compared to the bound for the L_2 norm. □

The bound on ρ_m tells us that for N big enough, the real part of T_k contains hardly any information about the unknown ξ_k. ρ_m tends to zero as $N \to \infty$ and therefore $B_2 \to B$ as $N \to \infty$ so the Bayes risk of the real-valued and complex-valued problems are essentially as difficult. We finish establishing the bound in the case of complex-valued data by noting that,

$$\|g_{k,m}\|^2 = \|b_{k,m}\|^2 + \|a_{k,m}\|^2 \geq \|a_{k,m}\|^2,$$

and therefore (C.5) and the lower bound of the Bayes risk give us

$$R^*(\mathcal{F}_N) \geq C \cdot N^{-2(s-1)/(2s+1)},$$

for some positive constant C. $\square$

C.6 Proof of Theorem 7

Proof. We follow Donoho's and Johnstone's arguments in their paper on thresholding in the best orthobasis [34] almost exactly. See also Candès' survey paper [21] and Donoho's paper on the connection between the CART and Best-Ortho-Basis methodologies [31].

The empirical complexity for the estimator can be written as

$$\begin{aligned} K(\hat{f}, y) &= \|\hat{f} - y\|^2 + \Lambda(\hat{f}) = \|f - \hat{f} + z\|^2 + \Lambda(\hat{f}) \\ &= \|f - \hat{f}\|^2 + \Lambda(\hat{f}) + \|z\|^2 + 2Re(\langle f - \hat{f}, z\rangle) \\ &= \hat{K} + \|z\|^2 + 2Re(\langle f - \hat{f}, z\rangle), \end{aligned}$$

and similarly,

$$K(f_0, y) = K_0 + \|z\|^2 + 2Re(\langle f - f_0, z\rangle).$$

Since by the definition of $\hat{f}$, $K(\hat{f}, y) \leq K(f_0, y)$, we have

$$\begin{aligned} \hat{K} &\leq K_0 + 2Re(\langle f - f_0, z\rangle) - 2Re(\langle f - \hat{f}, z\rangle) \\ &= K_0 + 2Re(\langle \hat{f} - f_0, z\rangle). \end{aligned}$$

This provides the following bound on the square error of the estimator:

$$\|f - \hat{f}\|^2 \leq \hat{K} \leq K_0 + 2Re(\langle \hat{f} - f_0, z\rangle).$$

Define

$$\Delta(k) = \sup_{f_1, f_2 \in \mathcal{C}} \{Re(\langle f_1 - f_2, z\rangle : \|f_j - f\|^2 \leq k, \Lambda(f_j) \leq k, j = 1, 2\}.$$

Note that, by the definition of f_0, we have $K_0 \leq \hat{K}$. Then

$$\|f_0 - f\|^2 \leq K_0 \leq \hat{K}, \quad \Lambda(f_0) \leq K_0 \leq \hat{K},$$

and obviously

$$\|\hat{f} - f\|^2 \leq \hat{K}, \quad \Lambda(\hat{K}) \leq \hat{K}.$$

Therefore, $\hat{f}$ and f_0 are feasible solutions for the optimization problem for $\Delta(\hat{K})$, which gives

$$\hat{K} \leq K_0 + 2\Delta(\hat{K}) \leq \hat{K} + 2\Delta(\hat{K}).$$

The strategy is to show that $2\Delta(\hat{K})$ is small compared to $\hat{K}$ and therefore $\hat{K}$ and K_0 are of similar size.

Define $k_j = 2^j(1-8/\eta)^{-1}\max(K_0, \lambda^2)$ for $j \geq 0$ and remember that $\eta > 8$. Define the event

$$E_j := \left\{\Delta(k) \leq \frac{4k}{\eta}, \forall k \geq k_j\right\}.$$

Assume E_j holds. Then necessarily the event

$$B_j := \{\hat{K} \leq k_j\}$$

holds. Otherwise, if $\hat{K} > k_j$ we have $\Delta(\hat{K}) \leq 4\hat{K}/\eta$ by the definition of E_j. This would give

$$\hat{K} \leq K_0 + 2\Delta(\hat{K}) \leq K_0 + \frac{8}{\eta}\hat{K},$$

and $\hat{K} \leq (1-8/\eta)^{-1} K_0 \leq k_j$, which is a contradiction.

Since $B_j^c \subset E_j^c$, we have

$$\begin{aligned} E\hat{K} &\leq k_0 P(\hat{K} \leq k_0) + \sum_{j=0}^{\infty} k_{j+1} P(k_j \leq \hat{K} < k_{j+1}) \leq k_0 + \sum_{j=0}^{\infty} k_{j+1} P(k_j \leq \hat{K}) \\ &\leq k_0 + \sum_{j=0}^{\infty} k_{j+1} P(E_j^c). \end{aligned}$$

Next we use the bound from Lemma 21 to get

$$\begin{aligned} E\hat{K} &\leq k_0 + \sum_{j=0}^{\infty} k_0 \cdot 2^{j+1} \cdot \frac{1}{(2j)!} \\ &\leq k_0 + 2k_0 \cdot \sum_{j=0}^{\infty} \frac{2^j}{j!} = k_0 + 2(1 - 8/\eta)^{-1} \cdot \max(K_0, \lambda^2) \cdot e \\ &\leq 6.5(1 - 8/\eta)^{-1} \cdot \left(K_0 + \lambda^2\right). \end{aligned}$$

□

The following lemma is based on [34] and can also be found in [31]. The proof hardly needs to be changed for our situation of chirp estimation using chirplet paths:

Lemma 21.

$$P(E_j^c) \leq 1/(2^j)!.$$

Proof. Let $\mathcal{M}$ stand for our chirplet dictionary with the number of chirplets equal to $M_N = |\mathcal{M}|$. Fix $j \geq 0$ and take some $k \geq k_j$. Then $k \in [l\lambda^2, (l+1)\lambda^2)$ for some $l \in \mathbf{Z}^+$. Let f_1 and f_2 be feasible signals for the optimization problem $\Delta(k)$. Each of these signals has at most $l = \lfloor k/\lambda^2 \rfloor$ terms since $\Lambda(f_n) \leq k$, $n = 1, 2$, and therefore the difference $f_1 - f_2$ is a linear combination of at most $2l$ distinct terms from $\mathcal{M}$.

Let V be the linear space of dimension at most $2l$ spanned by those terms and let P_V be the orthogonal projection onto V. Using the triangular inequality and that f_1 and f_2 are feasible for $\Delta(k)$, we get $\|f_1 - f_2\| \leq \|f_1 - f\| + \|f - f_2\| \leq 2\sqrt{k}$. The Cauchy-Schwarz inequality gives

$$|Re(\langle f_1 - f_2, z\rangle)| \leq |\langle f_1 - f_2, z\rangle| \leq \|f_1 - f_2\| \cdot \|P_V z\|_2 \leq 2\sqrt{k}\|P_V z\|_2.$$

Consider the event

$$A_l = \{\|P_V z\|_2 \leq \sqrt{2l} \cdot \sqrt{2}(1 + \sqrt{2 \log M_N}),\ \text{for all } V \in \mathcal{M}\}$$

where $C(2l, M_N)$ is the collection of all subspaces spanned by $2l$ members out of M_N from

$\mathcal{M}$. On the event A_l we have,

$$\begin{aligned} |Re(\langle f_1 - f_2, z\rangle)| &\leq 2\sqrt{k} \cdot \sqrt{2l} \cdot \sqrt{2}(1 + \sqrt{2\log M_N}) \\ &\leq 4 \cdot \frac{k}{\lambda}(1 + \sqrt{2\log M_N}) = 4 \cdot \frac{k}{\eta}, \end{aligned}$$

since $l \leq k/\lambda^2$ and from the definition of λ. Thus, on the event $F_j := \cap_{l\geq 2^j} A_l$, $\Delta(k) \leq 4 \cdot \frac{k}{\eta}$ for all $k \in \cup_{l\geq 2^j}[l\lambda^2, (l+1)\lambda^2) = [2^j\lambda^2, \infty)$. Since $k_j \geq 2^j\lambda^2$, $\Delta(k) \leq 4 \cdot \frac{k}{\eta}$ for all $k \geq k_j$ on the event F_j. Therefore $F_j \subset E_j$, and we finish the proof by the help of Lemma 22 below,

$$\begin{aligned} P(E_j^c) &\leq P(F_j^c) \leq \sum_{l\geq 2^j} P(A_l^c) \leq \sum_{l\geq 2^j} 2M_N^{-1}/(2l)! \\ &\leq 2/M_N \sum_{l\geq 2^j} 1/l! \leq 1/(2^j)!, \end{aligned}$$

where we used the facts, $M_N > 6$ and $\sum_{l=n+1}^{\infty} 1/l! < 3/(n+1)!$, which can be derived from the remainder of the Taylor series for e^x (see, for example, [7]). □

The following bound is from Donoho and Johnstone [34] who gave it for real-valued vectors. For completeness we include their proof which basically holds unchanged for complex-valued white noise:

Lemma 22. *Let M vectors in C^N be given, and let $C(D, M)$ denote the collection of all subsets consisting of D out of those M vectors. Let $z = (z_k)$ be a random vector with i.i.d. entries such that $z_k = (z_k^1 + iz_k^2)/\sqrt{2}$ where z_k^1 and z_k^2 are i.i.d. $N(0, 1)$. Then for $\beta > 0$*

$$P\left(\sup_{V\in C(D,M)} \|P_V z\|_2 > \sqrt{D}(1 + \sqrt{2(1+\beta)\log(M)})\right) \leq 2M^{-\beta}/D!.$$

The result also holds if the vectors are in R^N and the random vector $z = (z_k)$ is real with i.i.d. standard Gaussian entries.

Proof. The argument is based on Borell's inequality (see [56]). Take $V \in C(D, M)$. Then the dimensionality of V is at most D. This gives, $E\|P_V z\|_2^2 = \dim(V) \leq D$. Next, note that $\|P_V z\|_2$ is a Lipschitz functional on a Gaussian field with Lipschitz constant 1. Namely, since for any two vectors u and w in C^N,

$$|\|P_V u\|_2 - \|P_V w\|_2| \leq \|P_V u - P_V w\|_2 = \|P_V(u - w)\| \leq \|u - w\|,$$

where we have used the triangle inequality and the fact that P_V is a projector. Then for every V, we have (according to Lemma 28) that

$$P\left(\|P_V z\|_2 > \sqrt{D}+t\right) \leq P\left(\|P_V z\|_2 > \sqrt{E\|P_V z\|_2^2}+t\right) \leq 2e^{-t^2/2}, \quad t>0.$$

Setting $t = \sqrt{D}\cdot\sqrt{2(1+\beta)\log(M)}$, bounding $\begin{pmatrix} M \\ D \end{pmatrix}$ by $M^D/D!$, and using a union bound gives the inequality:

$$P\left(\sup_{V\in C(D,M)} \|P_V z\|_2 > \sqrt{D}(1+\sqrt{2(1+\beta)\log(M)})\right) \leq \begin{pmatrix} M \\ D \end{pmatrix} 2M^{-D(1+\beta)} \leq 2M^{-\beta}/D!.$$

□

C.7 Proof of Theorem 9

Proof. The proof is based on similar arguments as in Appendix C.6. The residual sum of squares can be written as

$$\|y-\hat{f}\|^2 = \|f-\hat{f}\|^2 + \|z\|^2 + 2Re(\langle f-\hat{f}, z\rangle),$$

and similarly,

$$\|y-f_0\|^2 = \|f-f_0\|^2 + \|z\|^2 + 2Re(\langle f-f_0, z\rangle).$$

By the definition of $\hat{f}$, $\|y-\hat{f}\|^2 \leq \|y-f_0\|^2$, and therefore

$$\|f-\hat{f}\|^2 \leq \|f-f_0\|^2 + 2Re(\langle \hat{f}-f_0, z\rangle).$$

Define

$$\Delta(k) = \sup_{f_1,f_2\in\mathcal{C}} \{Re(\langle f_1-f_2, z\rangle : \|f_j-f\|^2 \leq k, j=1,2\},$$

and write $\hat{K} := \|f-\hat{f}\|^2$, $K_0 := \|f-f_0\|^2$. Note that, by the definition of f_0, we have $K_0 \leq \hat{K}$ and therefore, $\hat{f}$ and f_0 are feasible solutions for the optimization problem for

$\Delta(\hat{K})$. This gives

$$\hat{K} \leq K_0 + 2\Delta(\hat{K}) \leq \hat{K} + 2\Delta(\hat{K}).$$

As in the previous proof, we want to show that $2\Delta(\hat{K})$ is small compared to $\hat{K}$.

Due to our choice of chirplet graph, there exists a constant $C(R) > 1$ such that

$$K_0 \leq C(R)L.$$

Recall that the number of chirplet paths in the graph is less than $Ae^{\gamma L}$, for some constant $A > 0$. Let

$$\xi^2 = \frac{3}{2}\gamma + \log(A) + \log(2)/2,$$

and

$$\lambda^2 = \eta^2(\sqrt{\max(C(R), \xi^2)} + 1)^2,$$

where $\eta > 8$. Note that $\lambda^2 L \geq K_0$. Define

$$k_j = (j+1)(1 - 8/\eta)^{-1}\lambda^2 L, \quad j \geq 0,$$

and the event

$$E_j := \left\{\Delta(k) \leq \frac{4k}{\eta}, \forall k \geq k_j\right\}.$$

As in the proof in Appendix C.6, the event E_j is a subset of

$$B_j := \{\hat{K} \leq k_j\}.$$

Then, since $B_j^c \subset E_j^c$, we have

$$\begin{aligned} E\hat{K} &\leq k_0 P(\hat{K} \leq k_0) + \sum_{j=0}^{\infty} k_{j+1} P(k_j \leq \hat{K} < k_{j+1}) \leq k_0 + \sum_{j=0}^{\infty} k_{j+1} P(k_j \leq \hat{K}) \\ &\leq k_0 + \sum_{j=0}^{\infty} k_{j+1} P(E_j^c). \end{aligned}$$

Next we use the bound from Lemma 21 to get

$$\begin{aligned} E\hat{K} &\leq k_0 + \sum_{j=0}^{\infty} k_0 \cdot (j+1) \cdot \frac{\exp(-(j+1)\gamma L)}{1-\exp(-\gamma L)} \\ &= k_0 + \frac{k_0}{1-\exp(-\gamma L)} \sum_{j=0}^{\infty} j \cdot \exp(-j\gamma L) \\ &= \left(1 + \frac{\exp(-\gamma L)}{(1-\exp(-\gamma L))^3}\right) \cdot k_0 \\ &= \left(1 + \frac{\exp(-\gamma L)}{(1-\exp(-\gamma L))^3}\right) (1-8/\eta)^{-1} \lambda^2 L, \end{aligned}$$

since $\sum_{j \geq 0} j a^j = a/(1-a)^2$, for $0 \leq a < 1$. □

The following lemma gives a bound on the probability of the event E_j^c:

Lemma 23.

$$P(E_j^c) \leq \frac{\exp(-(j+1)\gamma L)}{1-\exp(-\gamma L)}.$$

Proof. Fix $j \geq 0$ and take some $k \geq k_j$. Then $k \in [l\lambda^2 L, (l+1)\lambda^2 L)$ for some $l \in \mathbf{Z}^+$. Let f_1 and f_2 be feasible signals for the optimization problem $\Delta(k)$. Each of these signals is a sum of L chirplets. Therefore the difference $f_1 - f_2$ is a linear combination of at most $2L$ chirplets. Let V be the linear space of dimension at most $2L$ spanned by those chirplets and let P_V be the orthogonal projection onto V. Let $\mathcal{M}$ be the collection of all possible spaces of this form. Then, as the number of chirplet paths in the graph is less than $Ae^{\gamma L}$, we have

$$|\mathcal{M}| \leq A^2 e^{2\gamma L}.$$

Following the same arguments as in the proof of Lemma 21, we have

$$|Re(\langle f_1 - f_2, z \rangle)| \leq |\langle f_1 - f_2, z \rangle| \leq \|f_1 - f_2\| \cdot \|P_V z\|_2 \leq 2\sqrt{k} \|P_V z\|_2.$$

On the event

$$A_l = \{\|P_V z\|_2 \leq \sqrt{2l} \cdot \lambda/\eta \cdot \sqrt{2L}, \text{ for all } V \in \mathcal{M}\},$$

we have,

$$\begin{aligned} |Re(\langle f_1 - f_2, z\rangle)| &\leq 2\sqrt{k}\cdot\sqrt{2l}\cdot\lambda/\eta\cdot\sqrt{2L} \\ &\leq 4\cdot\frac{k}{\lambda\sqrt{L}}\cdot\lambda/\eta\cdot\sqrt{L} = 4\cdot\frac{k}{\eta}, \end{aligned}$$

since $l \leq k/(\lambda^2 L)$. Thus, on the event

$$F_j := \cap_{l\geq(j+1)} A_l,$$

$\Delta(k) \leq 4\cdot\frac{k}{\eta}$ for all $k \in \cup_{l\geq(j+1)}[l\lambda^2 L, (l+1)\lambda^2 L) = [(j+1)\lambda^2 L, \infty)$. Since $k_j \geq (j+1)\lambda^2 L$,

$$\Delta(k) \leq 4\cdot\frac{k}{\eta}, \quad \forall k \geq k_j,$$

on the event F_j. Therefore $F_j \subset E_j$, and we finish the proof by the help of Lemma 22 below,

$$\begin{aligned} P(E_j^c) &\leq P(F_j^c) \leq \sum_{l\geq j+1} P(A_l^c) \leq \sum_{l\geq j+1}\exp(-l\gamma L) \\ &= \frac{\exp(-(j+1)\gamma L)}{1-\exp(-\gamma L)}, \end{aligned}$$

using the formula for the geometric series which holds since $\exp(-\gamma L) < 1$. This finishes the proof. □

Lemma 24.

$$P(A_l^c) \leq \exp(-l\gamma L).$$

Proof. We have the concentration inequality

$$P\left(\|P_V z\|_2 > \sqrt{2L} + t\right) \leq 2e^{-t^2/2},$$

for all $t > 0$ (see the proof of Lemma 22). Set $t = \sqrt{2l\cdot\max(C(R),\xi^2)}\cdot\sqrt{2L}$ and note that

$$\begin{aligned} \sqrt{2l}\lambda/\eta\sqrt{2L} &= \sqrt{2l\cdot\max(C(R),\xi^2)}\sqrt{2L} + \sqrt{2l}\cdot\sqrt{2L} \\ &\geq \sqrt{2l\cdot\max(C(R),\xi^2)}\sqrt{2L} + \sqrt{2L}. \end{aligned}$$

Since $l \geq 1$ and $L \geq 1$,

$$\begin{aligned} 2l \cdot \max(C(R), \xi^2) \cdot L &\geq 2l \cdot \xi^2 \cdot L = 2l \cdot \frac{3}{2}\gamma L + 2l \log(A) L + l \log(2) L \\ &\geq 2\gamma L + l\gamma L + 2\log(A) + \log(2). \end{aligned}$$

Then from the concentration inequality and a union bound, we get

$$\begin{aligned} P(A_l^c) &\leq P\left(\sup_{V \in \mathcal{M}} \|P_V z\|_2 > \sqrt{2L} + \sqrt{2l \cdot \max(C(R), \xi^2)}\sqrt{2L}\right) \\ &\leq 2A^2 \exp(2\gamma L) \exp(-2l \cdot \max(C(R), \xi^2) \cdot L) \\ &\leq 2A^2 \exp(2\gamma L) \exp(-2\gamma L - l\gamma L - 2\log(A) - \log(2)) \\ &= \exp(-l\gamma L). \quad \square \end{aligned}$$

C.8 Proof of Lemma 10

We use similar arguments as in Chapter 6. Recall that $\cos(x) \geq 1 - x^2/2$ for all $x \in \mathbb{R}$. Assume $I = [t_0, t_1]$. From the fundamental theorem of calculus we get

$$|\varphi(t) - bt - \varphi(t_0)| \leq \Delta\omega |I|.$$

Then

$$\begin{aligned} |\langle f, c \rangle| &= |\langle f e^{-iN\varphi(t_0)}, c \rangle| \geq Re(\langle e^{iN(\varphi(t) - bt - \varphi(t_0))}, 1_I \rangle) \\ &= \langle \cos(N(\varphi(t) - bt - \varphi(t_0))), 1_I \rangle \\ &\geq \left(1 - (N\Delta\omega|I|)^2/2\right) \cdot \|I\|, \end{aligned}$$

since $(N\Delta\omega|I|)^2/2 \leq 1$. $\square$

C.9 Proof of Lemma 11

The bound can be established using integration by parts. First we have

$$\begin{aligned}\langle f, c\rangle &= \int_I e^{iN(\varphi(t)-bt)}dt = \int_I N(\varphi'(t)-b)e^{iN(\varphi(t)-bt)}\frac{1}{N(\varphi'(t)-b)}dt \\ &= \frac{e^{iN(\varphi(t_1)-bt_1)}}{N(\varphi'(t_1)-b)} - \frac{e^{iN(\varphi(t_1)-bt_0)}}{N(\varphi'(t_0)-b)} + \int_I e^{iN(\varphi(t)-bt)}\frac{\varphi''(t)}{N(\varphi'(t)-b)^2}dt.\end{aligned}$$

Then the triangular inequality gives

$$\begin{aligned}|\langle f, c\rangle| &\leq \frac{1}{N|\varphi'(t_1)-b|} + \frac{1}{N|\varphi'(t_0)-b)|} + \int_I \frac{|\varphi''(t)|}{N|\varphi'(t)-b|^2}dt \\ &\leq \frac{2}{N\Delta\omega} + \left(\sup_{t\in I}|\varphi''(t)|\right)\int_I \frac{1}{N(\Delta\omega)^2}dt \\ &= \frac{2}{N\Delta\omega} + \left(\sup_{t\in I}|\varphi''(t)|\right)\frac{|I|}{N(\Delta\omega)^2}.\end{aligned}$$

The result follows since $\sup_{t\in I}|\varphi''(t)| \leq R$. □

C.10 Proofs of Lemma 13 and Lemma 14

We start by analyzing $(P+\lambda QP)^n(\mathbf{1}+\lambda\mathbf{q})$. The goal is to find out what is the slowest rate that λ can go to zero such that the mean of the entries of

$$(P+\lambda QP)^n(\mathbf{1}+\lambda\mathbf{q}) = (P+\lambda QP)^n\mathbf{1} + \lambda(P+\lambda QP)^n\mathbf{q}$$

goes to 1 as $n\to\infty$. The sum over the entries of $(P+\lambda QP)^n\mathbf{1}$ is equal to the sum of the entries of the matrix $(P+\lambda QP)^n$, while the sum of the entries of $(P+\lambda QP)^n\mathbf{q}$ is the sum of the first column of that same matrix. Thus, it is sufficient to consider the first term in the above.

Note that P is a stochastic matrix, since all of its rows sum to 1, which gives us the second of the following two indentities:

$$Q\mathbf{1} = \mathbf{q}, \qquad P^k\mathbf{1} = \mathbf{1},\ k\geq 0.$$

To bound the mean, we look at the binomial expansion of $(P+\lambda QP)^n$ and investigate each term when multiplying it by the unit vector $\mathbf{1}$ from the right. The terms are as follows:

- $P^n\mathbf{1} = \mathbf{1}$, whose sum is n.

- $P \cdots PQP \cdots P\mathbf{1} = \underbrace{P \cdots P}_{l} Q\mathbf{1} = P^l\mathbf{q} =$ first column of P^l. Since P is doubly stochastic, so is P^l and all of its columns sum to 1. There are n such terms, all with factor λ in front, so averaging them gives us λ.

- $P \cdots PQP \cdots P \underbrace{QP \cdots P\mathbf{1}}_{=\mathbf{q}} = P \cdots PQ \underbrace{P \cdots P}_{l} \mathbf{q} = \underbrace{P \cdots P}_{k}(p_{11}^{(l)}\mathbf{q}) = p_{11}^{(l)}P^k\mathbf{q}$, where $p_{11}^{(l)}$ is the upper left-most entry in P^l. The entries in $P^k\mathbf{q}$, the first column of P^k, sum to 1, so each term of this form sums to $p_{11}^{(l)}$. The number these kind of terms is $\begin{pmatrix} n \\ 2 \end{pmatrix} = \frac{n(n-1)}{2}$.

- In general, when λQP appears exactly $k > 1$ times in a term, its resulting sum is equal to
$$\lambda^k p_{11}^{(l_1)} p_{11}^{(l_2)} \cdots p_{11}^{(l_{k-1})},$$
where $p_{11}^{(l_m)}$ is the upper left-most entry in a matrix of the form P^{l_m}. There are $\begin{pmatrix} n \\ k \end{pmatrix}$ terms of this form.

The following bounds for p_{11}^l show that $p_{11}^{(l)}$ is roughly equal to $1/\sqrt{l}$:

Lemma 25. *For all integers $l \geq 1$,*
$$C_1 \frac{1}{\sqrt{l}} \leq p_{11}^{(l)} \leq C_2 \frac{1}{\sqrt{l}},$$
where C_1 and C_2 are constants which can be taken to be $C_1 = 2^{-3/2}$ and $C_2 = 2$.

We postpone the proof of this lemma until Appendix C.10.3.

C.10.1 Proof of Lemma 13

Proof. In view of the necessary condition of the rate that we have achieved, this sufficient condition is pretty tight, since it differs only by a log-factor. To prove this lemma we investigate the terms in $(P + \lambda QP)^n\mathbf{1}$. A term where QP appears exactly $k > 1$ times is of the form

$$\begin{aligned} P^{l_1}(QP)P^{l_2}(QP)\cdots(QP)P^{l_k}(QP)P^{n-l_1-l_2-\cdots-k}\mathbf{1} &= P^{l_1}QP^{l_2+1}\cdots QP^{l_k+1}Q\mathbf{1} \\ &= P^{l_1}QP^{l_2+1}\cdots p_{11}^{(l_k+1)}\mathbf{q} \\ &= p_{11}^{(l_2+1)}p_{11}^{(l_3+1)}\cdots p_{11}^{(l_k+1)}P^{l_1}\mathbf{q} \end{aligned}$$

where

$$\begin{aligned} l_1 &= 0,\ldots,n-k \\ l_2 &= 0,\ldots,n-k-l_1 \\ &\vdots \\ l_k &= 0,\ldots,n-k-l_1-\cdots-l_{k-1}. \end{aligned}$$

Summing over the entries in this vector gives

$$p_{11}^{(l_2+1)}p_{11}^{(l_3+1)}\cdots p_{11}^{(l_k+1)},$$

and if we collect all the terms of where QP appears exactly k times we get

$$B_{k,n} := \sum_{l_1=0}^{n-k}\sum_{l_2=0}^{n-k-l_1} p_{11}^{(l_2+1)} \sum_{l_3=0}^{n-k-l_1-l_2} p_{11}^{(l_3+1)} \cdots \sum_{l_k=0}^{n-k-l_1-l_2-\cdots-l_{k-1}} p_{11}^{(l_k+1)}.$$

By letting the index of all the sums range from 0 to n and using Lemma 25 we get the following upper bound:

$$\begin{aligned} B_{k,n} &\leq \sum_{l_1=0}^{n}\sum_{l_2=0}^{n} p_{11}^{(l_2+1)} \sum_{l_3=0}^{n} p_{11}^{(l_3+1)} \cdots \sum_{l_k=0}^{n} p_{11}^{(l_k+1)} \\ &\leq \sum_{l_1=0}^{n}\sum_{l_2=0}^{n} \frac{C}{\sqrt{l_2+1}} \sum_{l_3=0}^{n} \frac{C}{\sqrt{l_3+1}} \cdots \sum_{l_k=0}^{n} \frac{C}{\sqrt{l_k+1}} \\ &\leq nC^{k-1}(n^{1/2})^{k-1} = C^{k-1}n^{\frac{k+1}{2}} \end{aligned}$$

where C is a positive constant.

Then

$$\begin{aligned} Ave\left((P+\lambda QP)^n\mathbf{1}\right) &\le \frac{1}{n}\left(n+\sum_{k=1}^{n}\lambda^k C^{k-1} n^{\frac{k+1}{2}}\right) \\ &= \frac{1}{n}\left(n+\sqrt{n}/C\sum_{k=1}^{n}(C\lambda\sqrt{n})^k\right) \\ &= 1+\frac{1}{C}\frac{1}{\sqrt{n}}\sum_{k=1}^{n}(C\lambda\sqrt{n})^k. \end{aligned}$$

If $C\lambda\sqrt{n}\le a<1$ where a is some constant, we get that

$$\sum_{k=1}^{n}(C\lambda\sqrt{n})^k \le \sum_{k=0}^{\infty}a^k = \frac{1}{1-a}$$

and

$$1\le Ave\left((P+\lambda QP)^n\mathbf{1}\right)\le 1+\frac{1}{C}\frac{1}{\sqrt{n}}\frac{1}{1-a}\to 1 \qquad \text{as } n\to\infty.$$

□

C.10.2 Proof of Lemma 14

Proof. Lemma 25 gives us the following lower bound on the sum of the entries of $(P+\lambda QP)^n\mathbf{1}$:

$$\begin{aligned} Ave\left((P+\lambda QP)^n\mathbf{1}\right) &\ge \frac{1}{n}\left(n+\sum_{k=1}^{n}\begin{pmatrix} n \\ k \end{pmatrix}\lambda^k\left(2^{-3/2}\frac{1}{\sqrt{n}}\right)^{k-1}\right) \\ &= 1+2^{3/2}\frac{1}{\sqrt{n}}\sum_{k=1}^{n}\begin{pmatrix} n \\ k \end{pmatrix}\left(\frac{\lambda'}{\sqrt{n}}\right)^k \\ &= 1-2^{3/2}\frac{1}{\sqrt{n}}+2^{3/2}\frac{1}{\sqrt{n}}\left(1+\frac{\lambda'}{\sqrt{n}}\right)^n \end{aligned}$$

where $\lambda'=2^{-3/2}\lambda$. If the last term goes to ∞ as $n\to\infty$, then the average goes to ∞. This gives us a necessary condition for the rate of λ.

$$\frac{1}{\sqrt{n}}\left(1+\frac{\lambda'}{\sqrt{n}}\right)^n = \frac{1}{\sqrt{n}}\exp\left(n\log\left(1+\frac{\lambda'}{\sqrt{n}}\right)\right)\sim\frac{1}{\sqrt{n}}\exp\left(\sqrt{n}\lambda'\right) \qquad \text{as } n\to\infty.$$

This gives us the necessary condition

$$\lambda' = \lambda'_n < \frac{1}{2}\frac{\log(n)}{\sqrt{n}} \text{ as } n \to \infty,$$

since if $\lambda' = \lambda'_n \geq \frac{1}{2}\frac{\log(n)}{\sqrt{n}}$, we have that $\lim_{n\to\infty} \frac{1}{\sqrt{n}} \exp\left(\sqrt{n}\lambda'\right) \geq 1$ so

$$\lim_{n\to\infty} Ave\Big((P+\lambda QP)^n(\mathbf{1}+\lambda\mathbf{q})\Big) \geq \lim_{n\to\infty} Ave\Big((P+\lambda QP)^n\mathbf{1}\Big) \geq 1 + 2^{3/2} > 1.$$

□

C.10.3 Proof of Lemma 25

C.10.3.1 Lower bound for $p_{11}^{(l)}$

Notice that $p_{11}^{(l)}$ is simply the probability of being in state 1 after l steps, given that one started in state 1. By considering a random walk on a circle with n states, $l \leq n$, we can write down this probability explicitly. Ending up at the same place as one started in l steps amounts to taking k steps clockwise, k steps counterclockwise and $l-2k$ steps without moving. These movements happen with probability 1/4,1/4 and 1/2 respectively and we notice that with the restriction $l \leq n$ one cannot go all around the circle. We will consider the cases l even and odd seperately:

1. l **even,** $l = 2m$: In this case we have

$$p_{11}^{(2m)} = \sum_{k=0}^{m} \frac{(2m)!}{k!k!(2m-2k)!}\left(\frac{1}{4}\right)^k\left(\frac{1}{4}\right)^k\left(\frac{1}{2}\right)^{2m-2k}.$$

If we write

$$c_k := 2^{-2k}\frac{(2k)!}{(k!)^2}$$

we have

$$\begin{aligned} p_{11}^{(2m)} &= \sum_{k=0}^{m} 2^{-2k}\frac{(2k)!}{k!k!}\frac{(2m)!}{(2k)!(2n-2k)!}\left(\frac{1}{2}\right)^{2k}\left(\frac{1}{2}\right)^{2m-2k} \\ &= \sum_{k=0}^{m} c_k \left(\begin{array}{c} 2m \\ 2k \end{array}\right)\left(\frac{1}{2}\right)^{2k}\left(\frac{1}{2}\right)^{2m-2k}. \end{aligned}$$

The sequence (c_k) is decreasing, since

$$\frac{c_{k+1}}{c_k} = \frac{(2k+2)(2k+1)}{(k+1)^2}\frac{1}{2^2} = \frac{2k+1}{k+1}\frac{1}{2} = \frac{k+1/2}{k+1} < 1.$$

Therefore

$$p_{11}^{(2m)} \geq c_m \sum_{k=0}^{m} \left(\begin{array}{c} 2m \\ 2k \end{array} \right) \left(\frac{1}{2}\right)^{2m}.$$

Using Stirling's approximation for the factorials, we get that

$$c_m \sim \frac{1}{\sqrt{\pi m}}$$

as $m \to \infty$. In fact we can prove by induction that

$$c_m \geq \frac{1}{2\sqrt{m}}.$$

This holds for $m = 1$ since $c_1 = 1/2$. Assume this is true for c_m. Now $(2m+1)^2 = 4m^2 + 4n + 1 > 4m(m+1)$ and therefore

$$\frac{(2m+1)^2}{(m+1)^2}\frac{1}{4}\frac{1}{m} > \frac{1}{m+1}$$

which gives us that

$$c_{m+1} = \frac{2m+1}{m+1}\frac{1}{2}c_m \geq \frac{2m+1}{m+1}\frac{1}{2}\frac{1/2}{\sqrt{m}} > \frac{1}{2\sqrt{m+1}}.$$

Now all that is left is to handle the term,

$$\sum_{k=0}^{m} \left(\begin{array}{c} 2m \\ 2k \end{array} \right) \left(\frac{1}{2}\right)^{2m}$$

which is identically equal to $1/2$. This follows from the well-known identity

$$\left(\begin{array}{c} n \\ k \end{array} \right) = \left(\begin{array}{c} n-1 \\ k-1 \end{array} \right) + \left(\begin{array}{c} n-1 \\ k \end{array} \right), \quad \text{for all } 1 \leq k \leq n-1$$

and the binomial formula:

$$\begin{aligned}\sum_{k=0}^{m}\left(\begin{array}{c}2m\\2k\end{array}\right)\left(\frac{1}{2}\right)^{2m} &= 2^{-2m}\left(\sum_{k=1}^{m-1}\left(\begin{array}{c}2m-1\\2k-1\end{array}\right)+\sum_{k=1}^{m-1}\left(\begin{array}{c}2m-1\\2k\end{array}\right)+2\right)\\ &= 2^{-2m}\sum_{k=0}^{2m-1}\left(\begin{array}{c}2m-1\\k\end{array}\right)=2^{-1}.\end{aligned}$$

Thus we have that

$$p_{11}^{(2n)} \geq 2^{-3/2}\frac{1}{\sqrt{2n}}.$$

2. l **odd,** $l=2m+1$: This case can be tackled in a similar way as before. Here we have

$$\begin{aligned}p_{11}^{(2m+1)} &= \sum_{k=0}^{m}c_k\left(\begin{array}{c}2m+1\\2k\end{array}\right)\left(\frac{1}{2}\right)^{2m+1}\\ &\geq c_m\sum_{k=0}^{m}\left(\begin{array}{c}2m+1\\2k\end{array}\right)\left(\frac{1}{2}\right)^{2m+1}\\ &= c_m\cdot\frac{1}{2}\\ &\geq \frac{1}{4\sqrt{m}}\end{aligned}$$

since

$$\begin{aligned}\sum_{k=0}^{m}\left(\begin{array}{c}2m+1\\2k\end{array}\right)\left(\frac{1}{2}\right)^{2m+1} &= 2^{-2m-1}\sum_{k=0}^{m}\frac{1}{2}\left(\left(\begin{array}{c}2m+1\\2k\end{array}\right)+\left(\begin{array}{c}2m+1\\2m+1-2k\end{array}\right)\right)\\ &= 2^{-2m-2}\left(\sum_{k=0}^{m}\left(\begin{array}{c}2m+1\\2k\end{array}\right)+\sum_{k=0}^{m}\left(\begin{array}{c}2m+1\\2m+1-2k\end{array}\right)\right)\\ &= 2^{-2m-2}\left(\sum_{k=0}^{m}\left(\begin{array}{c}2m+1\\2k\end{array}\right)+\sum_{k=0}^{m}\left(\begin{array}{c}2m+1\\2k+1\end{array}\right)\right)\\ &= 2^{-2m-2}\,2^{2m+1}=2^{-1}.\end{aligned}$$

Since $2^3(2m+1)>16m$, we have

$$p_{11}^{(2m+1)} \geq \frac{1}{4\sqrt{m}} > 2^{-3/2}\frac{1}{\sqrt{2m+1}}.$$

Hence, in general,

$$p_{11}^{(l)} \geq 2^{-3/2} \frac{1}{\sqrt{l}},$$

which proves the first inequality in Lemma 25.

C.10.3.2 Upper bound for $p_{11}^{(l)}$

We first consider $l = 2m$ and write

$$p_{11}^{(2m)} = \sum_{k=0}^{m} a_k, \quad \text{where } a_k = \frac{2^{-2k}}{(k!)^2(2m-2k)!} \cdot 2^{-2m}(2m)!.$$

Note that

$$\begin{aligned} \frac{a_{n+l}}{a_{n-l-1}} &= \frac{1}{2^{4l+2}} \frac{\prod_{k=-2l+1}^{2l+2}(2m+k-2n)}{\prod_{k=-l}^{l}(n+k)^2} \\ &= \frac{\prod_{k=-l}^{l}(2m+2k+1-2n)(2m+2k+2-2n)}{\prod_{k=-l}^{l}(2n+2k)^2}. \end{aligned}$$

If m is even, we get

$$\frac{a_{m/2+l}}{a_{m/2-l-1}} = \prod_{k=-l}^{l} \frac{(m+2k+1)}{(m+2k)} \cdot \frac{(m+2k+2)}{(m+2k)} > 1,$$

and therefore $a_{m/2+l} > a_{m/2-l-1}$, $l = 0, \ldots, m/2-1$. This gives

$$\begin{aligned} p_{11}^{(2m)} &= \sum_{k=0}^{m/2-1} a_k + \sum_{k=m/2}^{m} a_k \\ &\leq \sum_{k=0}^{m/2-1} a_{m/2+k} + \sum_{k=m/2}^{m} a_k \\ &= 2 \sum_{k=m/2}^{m-1} a_k + a_m \leq 2 \sum_{k=m/2}^{m} a_k. \end{aligned}$$

Since $a_k = c_k \left(\begin{array}{c} 2m \\ 2k \end{array} \right) \left(\frac{1}{2}\right)^{2m}$ and c_k is decreasing, we get

$$\begin{aligned} p_{11}^{(2m)} &\leq 2c_{m/2} \sum_{k=m/2}^{m} \left(\begin{array}{c} 2m \\ 2k \end{array} \right) \left(\frac{1}{2}\right)^{2m} \\ &\leq 2c_{m/2}1/2 = c_{m/2} \leq \frac{1}{\sqrt{m/2+1}} \\ &\leq \frac{2}{\sqrt{2m}} \end{aligned}$$

by using Lemma 26 below. If m is odd, we get in a similar way that

$$p_{11}^{(2m)} \leq c_{(m-1)/2} \leq \frac{1}{\sqrt{(m+1)/2}} \leq \frac{2}{\sqrt{2m}}.$$

A similar argument works for l odd.

Lemma 26. *For $k \geq 1$*

$$c_k \leq \frac{1}{\sqrt{k+1}}$$

Proof. This lemma can be proven by induction: First note that $c_1 = 1/2$ so the lemma is true for $k = 1$. Assume for some particular $k \geq 1$ that $c_k \leq \frac{1}{\sqrt{k+1}}$. Then

$$\begin{aligned} c_{k+1}^2 - \frac{1}{k+2} &= \left(\frac{2k+1}{2(k+1)}c_k\right)^2 - \frac{1}{k+2} \leq \left(\frac{2k+1}{2(k+1)}\right)^2 \frac{1}{k+1} - \frac{1}{k+2} \\ &= \frac{-3k-4}{4(k+1)^3(k+2)} \leq 0 \end{aligned}$$

so $c_{k+1} \leq \frac{1}{\sqrt{k+2}}$. □

Appendix D

Configurations for Two-Stage BP Test in Section 4.6

This appendix describes the configurations for the chirplet graphs used in the simulations for the two-stage BP test in Section 4.6. It is in the format of a `Matlab`-code readable by *ChirpLab* (see Section 10.1). This software package accompanies the ETD copy of this thesis and includes documentation for the functions in the code below (see also the file `TwoStageConfiguration.pdf` included in the ETD copy).

D.1 Configuration for the First Stage

The following lines of code configure the first stage of the test.

```
% Note Fs: sampling frequency in Hz
J = 8
N = 2^J;
sldf = 8;     % slope discretization factor
fmin = floor(40/Fs*N);    % minimum frequency in chirplet graph
fmax = N/2-1; % maximum frequency in chirplet graph
csc = 1;      % coarsest scale in chirplet graph
fsc = 1;      % finest scale in chirplet graph
graphParamStepI = GetChirpletGraphParam(N,csc,fsc,sldf,[],fmin,fmax,'COLOREDNOISE');
sbase = csc;      % scale index for the base time interval in search
smin = 7;
smax = 9;
```

D.2 Configuration for the Second Stage

The Matlab-function below is used to configure the chirplet graphs for the second stage of the test. The input P is the power spectrum of the noise. A data file storing this power spectrum can be found in the software package which comes with the ETD copy of the thesis.

```
function [graphParamExtInt, CnormExtInt, maxLength] = InitGraphParam(P)
% InitGraphParam -- a utility script for configuring chirplet graph parameters
% and precalculating chirplet norms to use for extended intervals
%  Input
%    P      noise spectrum
%  Output
%    graphParamExtInt  graph parameters
%    CnormExtInt       chirplet norms
%    maxLength         maximum number of chirplets to use in the BP algorithm

smin = 7;
smax = 10;
nScales = smax-smin+1;
graphParamExtInt = cell(1,nScales);
CnormExtInt = cell(1,nScales);
maxLength = zeros(1,nScales);

% Different chirplet graph parameters are used for different
% lengths of base intervals that are extended
for s=smin:smax,
  sindex=s-smin+1;

  steepSlopeParam = 1;
  J = s;
  % CONFIGURATIONS FOR THE STATISTICS
  Fs = 2048;
```

```
N = 2^J;
sldf = 8;     % slope discretization factor
fmin = floor(40/Fs*N);    % minimum frequency in chirplet graph
fmax = N/2-1; % maximum frequency in chirplet graph
csc = 0;      % coarsest scale in chirplet graph
if (J==10),
  fsc = 6;      % finest scale in chirplet graph
  ml = 16;
elseif(J==9),
  fsc = 5;      % finest scale in chirplet graph
  ml = 16;
elseif(J==8),
  fsc = 4;      % finest scale in chirplet graph
  ml = 8;
elseif(J==7),
  fsc = 3;
  ml = 8;
end
gp = GetChirpletGraphParam(N,csc,fsc,sldf,[],fmin,fmax,'COLOREDNOISE');

if (steepSlopeParam>0),
  % use steeper slopes at finest scales
  gp{3}{fsc+1} = [-0.5:sldf*2^((fsc)-J):8];
  if (steepSlopeParam>1),
    gp{3}{fsc} = [-0.5:sldf*2^(fsc-1-J):8];
  end
end

graphParamExtInt{sindex} = gp;
maxLength(sindex) = ml;

% calculate chirp norms
```

```
  CnormExtInt{sindex} = CTNormsDataStream(P,graphParamExtInt{sindex});
end
```

Appendix E

Signal Model For Binary Black Hole Coalescence

This section describes some of the test signals that were used in the numerical simulations in Chapter 4. The signal model was provided to me by Philip Charlton, who also wrote the remainder of this section. Since at the present time, this description has not been published, we choose to include it in the thesis for completeness, with Philip's permission.

Since the object of the exercise is to detect "real" gravitational waves, we will use as our test signals a collection of physically realistic waveforms for binary black hole coalescence. We use a modification of the method suggested by Anderson and Balasubramanian to model a complete coalescence waveform [6]. The signal consists of an inspiral phase, a merger phase, and a ringdown phase. While the inspiral and ringdown phase models are reasonable, the simulated merger phase should not be taken to be physically realistic. Instead, it is meant to approximate the gross time, frequency, and energy characteristics of a real merger.

The test signals are parametrised by the masses of the two bodies m_1 and m_2 in units of solar mass $M_{\odot}$. It is convenient to write the waveforms in terms of the total mass $M = m_1 + m_2$ and the symmetric mass ratio $\eta = m_1 m_2 / M^2$. The final waveform consists of three components combined in such a way that the result is continuous up to first derivatives:

$$h(t) \quad = \quad \begin{cases} h^{\mathrm{insp}}(t) & t \leq 0 \\ h^{\mathrm{merge}}(t) & 0 < t \leq t_m \\ h^{\mathrm{ring}}(t) & t > t_m \end{cases} \tag{E.1}$$

where we have arranged for the inspiral phase to end at $t = 0$ and the merger phase to end at $t = t_m$.

In the literature it is common to express the strain at the detector in terms of the + and × polarisations, with factors depending on the orientation of the binary system and its position in the sky relative to the axes of detector (see, for example, [8]). For simplicity we will replace these factors by their values averaged over orientations and positions. We then obtain an expression which is the sum of a cosine and a sine term, each with the same amplitude $A(t)$, which we combine into a single sinusoidal term of the form

$$h(t) \quad = \quad \sqrt{2}\,A(t)\cos\left[\phi(t)-\pi/4\right]. \tag{E.2}$$

For consistency with definitions of $A(t)$ and $\phi(t)$ in the literature we retain the factor of $\sqrt{2}$ and the phase offset $-\pi/4$.

For the inspiral component of the signal we use the non-spinning post^2-Newtonian approximation for the phase in the form given by [77, eqn. 15.24],

$$\phi^{\text{insp}}(t) = \phi_c - \frac{2}{\eta}\Big[\Theta^{5/8} + \left(\frac{3715}{8064}+\frac{55}{96}\eta\right)\Theta^{3/8} - \frac{3\pi}{4}\Theta^{1/4} \tag{E.3}$$

$$+\left(\frac{9275495}{14450688}+\frac{284875}{258048}\eta+\frac{1855}{2048}\eta^2\right)\Theta^{1/8}\Big] \tag{E.4}$$

where

$$\Theta(t) \quad = \quad \frac{\eta M_\odot}{5T_\odot M}(t_c - t) \tag{E.5}$$

and $T_\odot = GM_\odot/c^3 = 4.925491\times10^{-6}$s is the mass of the sun in geometrised units. The parameters ϕ_c and t_c are the phase and time at which coalescence occurs. This gives an instantaneous frequency

$$f^{\text{insp}}(t) = \frac{M_\odot}{8\pi T_\odot M}\Big[\Theta^{-3/8} + \left(\frac{743}{2688}+\frac{11}{32}\eta\right)\Theta^{-5/8} - \frac{3\pi}{10}\Theta^{-3/4} \tag{E.6}$$

$$+\left(\frac{1855099}{14450688}+\frac{56975}{258048}\eta+\frac{371}{2048}\eta^2\right)\Theta^{-7/8}\Big]. \tag{E.7}$$

For the amplitude of the gravitational wave strain we use the leading order (i.e., Newtonian) expression given in [77, eqn. 15.27–28]. Averaging over orientations (ι, β) allows us to replace the terms involving ι with $4\sqrt{5}$. Further averaging over sky position gives an additional reduction factor of $1/\sqrt{5}$, so the final amplitude of both the + and × components is

$$A^{\text{insp}}(t) = \frac{8}{5}\frac{T_\odot c}{D}\frac{\eta M}{M_\odot}\left[\frac{\pi T_\odot M f^{\text{insp}}(t)}{M_\odot}\right]^{2/3} \tag{E.8}$$

where D is the distance to the source.

A real inspiral would extend far into the past, with the instantaneous frequency approaching 0 as $t \longrightarrow -\infty$. We will only model the part of the inspiral detectable by a ground-based interferometer, that is, from the time the instantaneous frequency enters the sensitive band of the detector above f_s up to the commencement of the merger phase. Deciding where the boundary between inspiral and merger lies is somewhat arbitrary. We follow [38] in making the transition at the point where post-Newtonian approximations begin to break down. Assuming that the mass parameters m_1 and m_2 are decided upon beforehand, it is convenient to fix the time of transition from inspiral to merger occur at $t = 0$. A conservative estimate is that errors in the 2PN approximation become significant when the instantaneous frequency reaches

$$f_0 = \frac{M_\odot}{M} \times 4100 \text{ Hz}, \tag{E.9}$$

so we can set t_c by finding a suitable solution to $f^{\text{insp}}(0) = f_0$ [38]. Once this is known we can calculate the time at which the inspiral enters the sensitive band by solving $f^{\text{insp}}(t) = f_s$. Finally, we choose ϕ_c so that $\phi^{\text{insp}}(0) = \pi/4$ and the total phase in (E.2) is zero at $t = 0$.

The ringdown component is assumed to be an exponentially damped sinusoid with constant frequency f^{ring}, estimated to be

$$f^{\text{ring}} = \left[1 - 0.63(1-a)^{0.3})\right] \frac{M_\odot}{M} \times 32000 \text{ Hz}, \tag{E.10}$$

where a is a dimensionless spin parameter which is 0 for a Schwarzschild black hole and 1 for an extreme Kerr black hole [77, eqn. 18.3]. The ringdown phase is then

$$\phi^{\text{ring}}(t) = 2\pi f^{\text{ring}}(t - t_m) + \phi_m, \tag{E.11}$$

where we have introduced the ringdown phase ϕ_m at t_m. This will be set later to match the phase at the end of the merger.

The amplitude model we use is adapted from [80],

$$A^{\rm ring}(t) = \frac{\mathcal{A}}{\sqrt{20\pi}} \frac{T_\odot c}{D} \frac{M}{M_\odot} e^{-\pi f^{\rm ring}(t-t_m)/Q} \tag{E.12}$$

where $Q = 2(1-a)^{-0.45}$ is the quality factor,

$$\mathcal{A} = 4 \left[\frac{\pi \epsilon}{Q\left[1 - 0.63(1-a)^{0.3}\right]} \right]^{1/2} \tag{E.13}$$

and ϵ is the fraction of M radiated as gravitational waves during the ringdown. The factor of $1/\sqrt{20\pi}$ in (E.12) comes from averaging over orientations (giving a factor of $1/\sqrt{4\pi}$) and sky positions (giving a factor of $1/\sqrt{5}$). When written in the form (E.2) the overall amplitude is the same as that given in [77, eqn. 18.5]. In simulations we cut off the ringdown at $t = t_m + 5Q/\pi f^{\rm ring}$, that is, when the amplitude has been reduced by a factor of $e^{-5} \sim 1/148$.

Our inspiral component has been arranged to terminate at $t = 0$, with ringdown commencing at $t = t_m$. Since no analytic models exist for the merger part of the coalescence, we fit the amplitude and phase functions to bridge the gap between inspiral and ringdown. Flanagan and Hughes have estimated the merger duration to be $t_m \sim 50M/M_\odot \times T_\odot$ [38]. Assuming that the merger waveform is chirp-like of the form (E.2), a simple way to connect the inspiral and ringdown waveforms is to require that the instantaneous frequency and amplitude functions be continuous up to first derivatives at the boundaries between the merger component and each of the other two components. This gives four conditions that must be satisfied by $f^{\rm merge}(t)$ and $A^{\rm merge}(t)$ at $t = 0$ and $t = t_m$, so we will model $f^{\rm merge}(t)$ and $A^{\rm merge}(t)$ by cubic polynomials,

$$f^{\rm merge}(t) = \sum_{k=0}^{3} f_k t^k \tag{E.14}$$

$$A^{\rm merge}(t) = \sum_{k=0}^{3} A_k t^k \, . \tag{E.15}$$

For the instantaneous frequency, the continuity condition $f^{\rm merge}(0) = f^{\rm insp}(0)$ immediately tells us that f_0 is the value given by (E.9), while continuity in the first derivative tells us that $f_1 = \dot{f}^{\rm insp}(0)$. The remaining coefficients f_2 and f_3 can be found by solving

$$f^{\rm merge}(t_m) = f^{\rm ring} \tag{E.16}$$

$$\dot{f}^{\rm merge}(t_m) = 0 \tag{E.17}$$

noting that the right-hand side of (E.17) vanishes since $f^{\rm ring}$ is constant. We also want the phase to be continuous at $t = 0$, so we simply use the anti-derivative of $f^{\rm merge}$ with appropriate constant of integration,

$$\phi^{\rm merge}(t) = \pi/4 + 2\pi \int f^{\rm merge}(t) dt\,. \tag{E.18}$$

From $\phi^{\rm merge}(t)$ we can set $\phi_m = \phi^{\rm merge}(t_m)$.

For the merger amplitude coefficients we have $A_0 = A^{\rm insp}(0)$ and $A_1 = \dot{A}^{\rm insp}(0)$. The remaining continuity conditions

$$A^{\rm merge}(t_m) = A^{\rm ring}(t_m) \tag{E.19}$$

$$\dot{A}^{\rm merge}(t_m) = \dot{A}^{\rm ring}(t_m), \tag{E.20}$$

can be solved to find A_2 and A_3. Note that this is slightly different from the procedure used by Anderson and Balasubramanian in that they modeled $A^{\rm merge}(t)$ as a quartic polynomial. The extra coefficient was fixed by requiring that the merger waveform satisfy an energy condition, namely that the energy of the merger waveform be approximately three times that of the ringdown waveform, as estimated by Flanagan and Hughes assuming a near-maximal spin parameter $a = 0.98$. Justified by recent results from numerical experiments, we instead used the value $a = 0.7$. With this choice of spin parameter we found that imposing the energy condition produced a merger amplitude which looked artificially small compared with the adjacent inspiral and ringdown amplitudes. A more satisfactory waveform was obtained by simply fitting the amplitude with a cubic.

Appendix F

Detection of Sinusoids of Unknown Support

F.1 Statistical Model

We assume that we have given a long stream of M uniformly sampled data points where M is a dyadic number:

$$y_n = \alpha_M s_n + z_n, \quad n = 0, 1, \ldots, M-1,$$

where (s_k) is a vector of equispaced time samples of a complex-valued local sinusoid, and where $z = (z_k)$ is a complex-valued sequence of white noise where $z = z^1 + iz^2$ and z^1 and z^2 are two independent vectors of i.i.d. $N(0, 1/2)$ variables. α_M is assumed to be an unknown real scalar. Let $\mathbf{I} = \mathbf{I}_M$ be some collection of subintervals in $\{0, \ldots, M-1\}$. The signal is sampled at a fixed rate of N_s samples per second so that $s_n = s(n/N_s)$ where $s(t)$ is of the form,

$$s(t) = e^{i(N_s \omega t + \theta)} 1_I(t)/\sqrt{|I|},$$

where the support of the signal, $I \in \mathbf{I}$, the frequency $\omega \in [0, 2\pi)$ and the phase offset $\theta \in \mathbf{R}$ are assumed to be unknown *a priori*. Denote the class of these signals by $\mathcal{G}(\mathbf{I})$. Observing y_k, the goal is to decide between,

$$H_{0,M} : \alpha = 0, \quad \text{i.e., the data is only noise,}$$

and

$$H_{1,M} : \alpha \neq 0, \quad \text{i.e., there is a sinusoidal signal somewhere in the data.}$$

As we let $M \to \infty$, it is natural to ask whether there is a *threshold phenomenon*; i.e., a rate ρ_M such that if α_M grows slightly slower than ρ_M, every sequence of tests is asymptotically powerless and if α_M grows faster than ρ_M there exists a sequence of tests that is asymptotically powerful.

Remark: In this statistical model, the maximum frequency does not increase with the number of samples M, but stays fixed and determined by the sampling frequency. This is similar to common practical situations where data is collected at a fixed sampling rate over a period of time. Letting M grow is analogous to gathering more data.

F.1.1 Three different cases for $\mathcal{G}(\mathbf{I})$

We will consider three different cases for the set of signals in the alternative, differing only in the set of possible supports $\mathbf{I}$:

- Case 1: All the intervals in $\mathbf{I}$ have the same length, say, dyadic length 2^j for some $j \in \{0, \ldots, \log_2 M\}$
- Case 2: The intervals in $\mathbf{I}$ have dyadic lengths.
- Case 3: The intervals in $\mathbf{I}$ are the collection of all subintervals in $\{0, \ldots, M-1\}$.

F.1.2 The near-optimal tests

Our approach will be based on the GLRT paradigm. We construct a dictionary of functions of the form

$$s_{k,J} = e^{i\omega_k n} 1_J(n)/\sqrt{|J|}, \quad n = 0, \ldots, M-1$$

where J belongs to a set of intervals $\mathbf{J}$ and the frequency ω_k belongs to a set of discrete frequencies $\Omega(\mathbf{J})$. Call this dictionary $\mathcal{F} = \mathcal{F}(\mathbf{J})$. Before we give concrete choices for $\mathbf{J}$, we introduce a particular set of intervals. Let the set $\mathbf{J}_{d,j}$ of intervals of length 2^j and *overlapping degree* $d = 0, 1, \ldots$ be defined by

$$\mathbf{J}_{d,j} = \left\{ J : J = \{l2^{j-d}, \ldots, l2^{j-d} + 2^j - 1\}, l = 0, 1, \ldots, M2^{d-j} - 2^d \right\}.$$

Define the measure of *affinity* between intervals by

$$\rho(I, J) = \frac{|I \cap J|}{\sqrt{|I|}\sqrt{|J|}}.$$

We have the following lemma:

Lemma 27. *For any interval I of length $N = 2^j$, there is an interval $J \in \mathbf{J}_{d,j}$ such that*

$$\rho(I, J) \geq 1 - 2^{-d-1}.$$

Proof. Let $I = [k, k + N - 1]$ and choose from $\mathbf{J}_{d,j}$ the interval $J = [lN2^{-d}, lN2^{-d} + N - 1]$ such that $lN2^{-d}$ is as close to k as possible. Then obviously

$$|lN2^{-d} - k| \leq N2^{-d-1}.$$

Now $I \cap J$ is either $[k, lN2^{-d} + N - 1]$ or $[lN2^{-d}, k + N - 1]$, and therefore, $|I \cap J| \geq lN2^{-d} + N - k$ or $|I \cap J| \geq k + N - lN2^{-d}$. In either case,

$$|I \cap J| \geq N - |lN2^{-d} - k| \geq N - N2^{-d-1}.$$

This gives

$$\rho(I, J) \geq N(1 - 2^{-d-1})/(\sqrt{N}\sqrt{N}) = 1 - 2^{-d-1}.$$

□

Now we choose $\mathbf{J} = \mathbf{J}_M$ for the three cases:

- Case 1: We approximate $\mathbf{I}$ by a set of overlapping intervals. All these intervals have the same length, 2^j, as the intervals in $\mathbf{I}$. Denote this set by $\mathbf{J}_{d,j}$.
- Case 2: The approximating set of intervals is a set of overlapping intervals of every possible dyadic length. Denote this set by J_d.
- Case 3: $\mathbf{I}$ is approximated by a set of extended intervals, $\mathbf{J}_l$, as defined in Definition 4.

The set of discrete frequencies $\Omega(\mathbf{J})$ is defined by

$$\Omega(\mathbf{J}) = \bigcup_{J \in \mathbf{J}} \Omega_J$$

where

$$\Omega_J = \left\{ \omega_k = \frac{2\pi k}{|J| \log M}, \quad k = 0, \ldots, (|J| \log(M) - 1) \right\}.$$

Given a collection of intervals $\mathbf{J}_M$ we define the test statistic:

$$T_M^* = \max_{f \in \mathcal{F}(J_M)} |\langle y, f \rangle|. \tag{F.1}$$

F.2 Results

The results in this section hold for cases 1, 2, and 3 for the sequence of tests is given in the previous section. We have the following bound for our test statistics:

Theorem 15. *For each $\eta > 0$,*

$$P_{H_{0,M}} \left(T^*_{M,\mathbf{J}_M} > \sqrt{2(1+\eta) \log M} \right) \to 0, \quad M \to \infty.$$

We have the following lower bound:

Theorem 16. *Let $\alpha_M = \sqrt{2(1-\eta)\log M}$ be a sequence of signal amplitudes with $\eta > 0$. Then there is a sequence of distributions on local sinusoids in $\{0, \ldots, M-1\}$ such that every sequence of tests (T_M) is asymptotically powerless.*

The upper bound:

Theorem 17. *Let $\alpha_M = \sqrt{2(1+\eta)\log M}$ be a sequence of signal amplitudes with $\eta > 0$. Then the sequence of tests (T_M^*) is asymptotically powerful.*

F.3 Proof of Theorem 16:

Consider the following Bayes problem: Split the M data points into disjoint intervals of length N each. This gives us $K = M/N$ segments. On each segment, pick N orthonormal sinusoids that have their support entirely in that segment. This results in $N \times K = M$ orthonormal signals. Putting a uniform prior on these signals results in a Bayes problem which is equivalent, by sufficiency, to the well-known needle-in-a-haystack problem. Indeed, the inner products of the data and our M local sinusoids are Gaussian and all, but perhaps one, with mean zero. Therefore, if α_M grows slower than $\sqrt{2\log(M)}$, every sequence of test is asymptotically powerless. Note that this rate is independent of the signal support N.

F.4 Proof of Theorem 15:

We start by stating some well-known concentration inequalities. Let γ_N be the canonical Gaussian probability measure on $\mathbf{R}^N$ with density

$$\gamma_N(dx) = (2\pi)^{-N/2}\exp(-|x|^2/2)dx.$$

We say that a function $f : \mathbf{R}^N \to \mathbf{R}$ is a *Lipschitz*(C) function, or $f \in Lip(C)$ for short, if for $x, y \in \mathbf{R}^N$,

$$|f(x) - f(y)| \leq C\|x - y\|,$$

where $\|\cdot\|$ is the Euclidian norm. Then we have the following concentration of measure inequality (see for example [56]):

Theorem 18 (Concentration Bound). *If the function $f : \mathbf{R}^N \to \mathbf{R}$ is Lipschitz(C), then for any $t > 0$,*

$$\gamma_N\left(|f(X) - E[f(X)]| > t\right) \leq 2\exp\left(-t^2/2C^2\right).$$

Since $(E[f(X)])^2 \leq E[f(X)^2]$, this implies:

Lemma 28. *If $f : R^N \to R$ is a non-negative Lipschitz(C) function, then for any $t > 0$,*

$$\gamma_N\left(f(X) > t + \sqrt{E[f(X)^2]}\right) \leq 2\exp\left(-t^2/2C^2\right).$$

F.4.1 The distribution of $\langle Z, e^{i\omega t}1_I(t)/\sqrt{|I|}\rangle$

Consider the vector $Z = (Z_t, t \in \{0, \ldots, M-1\})$ of i.i.d. complex-valued random variables where $Z_t = Z_t^1 + iZ_t^2$ is a random variable with Z_1 and Z_2 i.i.d. $N(0, 1/2)$. We wish to find the distribution of the inner product[1] $\langle Z, g\rangle$, where $g = [g_0, \ldots, g_{M-1}]$ and $g_t = e^{i\omega t}1_I(t)/\sqrt{|I|}$.

Let's first look at $Y = Y_1 + iY_2 := e^{i\theta}Z = \cos\theta Z_1 - \sin\theta Z_2 + i(\sin\theta Z_1 + \cos\theta Z_2)$, whose distribution is easy to find since this is just a rotation of a vector of i.i.d. Gaussians. Namely, by using vector notation, we have

$$\mathbf{Y} := \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix},$$

[1] The inner-product is defined as $\langle a, b\rangle = \sum_k a_k b_k^*$ for complex vectors a and b.

so $\mathbf{Y}$ is a multivariate Gaussian vector of mean 0 and with the same covariance matrix as $[Z_1 \quad Z_2]^T$. That is, Y_1 and Y_2 are i.i.d. $N(0, 1/2)$. Then we can write

$$Z_t e^{-i\omega t} = Y_t^1 + iY_t^2$$

where Y_t^1, Y_t^2, $t \in I$ are i.i.d. $N(0, 1/2)$. We have that $\sum_{t\in I} Y_t^1$ and $\sum_{t\in I} Y_t^2$ are i.i.d $N(0, |I|/2)$ and therefore,

$$\begin{aligned} \langle Z, g\rangle &= \frac{1}{\sqrt{|I|}} \sum_{t\in I} Z_t e^{-i\omega t} = \frac{1}{\sqrt{|I|}} \left(\sum_{t\in I} Y_t^1 + i \sum_{t\in I} Y_t^2 \right) \\ &= W^1 + iW^2, \end{aligned}$$

where W^1 and W^2 are i.i.d. $N(0, 1/2)$. Then we have:

Lemma 29. *Let Z and g be as above. Then, $E|\langle Z, g\rangle|^2 = 1$. Also,*

$$\langle Z, g\rangle = W^1 + iW^2,$$

where W^1 and W^2 are i.i.d. $N(0, 1/2)$.

A simpler derivation of the lemma: We could also have simply used vector notation to derive this lemma. Indeed, since $\langle Z, g\rangle = g^*Z$ we have

$$E|\langle Z, g\rangle|^2 = Eg^*ZZ^*g = g^*Ig = \|g\|^2 = 1.$$

This shows that the lemma also holds in case the random vector Z is real standard normal, $Z \sim N(0, I)$. In the real-valued case we get that $\langle Z, g\rangle = W^1 + iW^2$ where W^1 and W^2 are dependent normal variables and not necessarily with variance 1/2.

F.4.2 Bounds on $|\langle Z, e^{i\omega t} 1_I(t)/\sqrt{|I|}\rangle|$

Let g be a vector of norm 1, i.e., $\|g\| = 1$. Note that $|\langle Z, g\rangle|$ is a Lipschitz(1) functional on a Gaussian field. Namely, for any vectors x and y of suitable sizes, we have

$$\Big| |\langle x, g\rangle| - |\langle y, g\rangle| \Big| \le |\langle x - y, g\rangle| \le \|x - y\| \cdot \|g\| = \|x - y\|,$$

where we have used the triangular inequality followed by the Cauchy-Schwartz inequality. As before, let $g = [g_0, \ldots, g_{M-1}]$ with $g_t = e^{i\omega t} 1_I(t)/\sqrt{|I|}$. From Lemma 29 we have that $E|\langle Z, g\rangle|^2 = 1$, and therefore by Lemma 28 we get for every $t > 0$ that

$$P\left(|\langle Z, g\rangle| > 1 + t\right) \leq 2e^{-t^2/2}. \tag{F.2}$$

Consider a set of vectors $\mathcal{F}$ of cardinality $|\mathcal{F}| \leq \mathcal{M}$. Pick some $\eta > 0$. Then by taking $t = \sqrt{2(1+\eta)\log \mathcal{M}}$ and using a union bound, we get

$$P\left(\sup_{g\in\mathcal{F}} |\langle Z, g\rangle| > 1 + \sqrt{2(1+\eta)\log \mathcal{M}}\right) \leq 2|\mathcal{F}|\frac{1}{\mathcal{M}^{1+\eta}} \leq \frac{2}{\mathcal{M}^{\eta}}.$$

F.4.3 Case 1: Size of support known

Assume that the size of the support of the unknown signal is N. We have,

$$\#\mathbf{J}_d \leq \frac{M}{N} 2^{d-1}$$

and on each interval we correlate the signal with complex sinusoids of $N \log M$ different equispaced frequencies. The size our dictionary $\mathcal{F}$ of test functions satisfies the bound

$$|\mathcal{F}| \leq 2^{d-1} M \log M.$$

Let $\eta > 0$. Taking $t = \sqrt{2(1+\eta)\log M}$ and using a union bound we get from the inequality (F.2) that

$$P\left(\sup_{g\in\mathcal{F}} |\langle Z, g\rangle| > 1 + \sqrt{2(1+\eta)\log M}\right) \leq 2 \cdot 2^{d-1} M \log M \frac{1}{M^{1+\eta}} = 2^d \frac{\log M}{M^{\eta}}.$$

As long as $d = d_M$ goes to infinity slowly enough, say $d/\log(M) \to 0$ as $M \to \infty$, we have $P\left(\sup_{g\in\mathcal{F}} |\langle Z, g\rangle| > \sqrt{2(1+\eta')\log M}\right) \to 0$ as $M \to \infty$, for any constant $\eta' > 0$.

F.4.4 Case 2: Support is of an unknown dyadic length

Assume we consider dyadic intervals of length $N_j = 2^j$ where $j = 0, \ldots, \log_2 M$. Let $\mathbf{J}_{d,j}$ be the set of intervals of overlapping degree d and length 2^j. Then,

$$\#\mathbf{J}_{d,j} \leq \frac{M}{N_j} 2^{d-1}.$$

On each interval in $\mathbf{J}_{d,j}$ we correlate the signal with complex sinusoids of $N_j \log M$ different equispaced frequencies so the number of test functions corresponding to $\mathbf{J}_{d,j}$ is $2^{d-1} M \log M$. Then the size of our dictionary $\mathcal{F}$ satisfies

$$|\mathcal{F}| \leq 2^{d-1} M \log M (1 + \log M) \leq 2^d M (\log M)^2.$$

Similar to Case 1,

$$P\left(\sup_{g \in \mathcal{F}} |\langle Z, g \rangle| > \sqrt{2(1+\eta)\log M} \right) \to 0, \quad \text{as } M \to \infty,$$

provided that $d = d_M$ does not increase too fast ($d_M = o(\log M)$ is sufficient).

F.4.5 Case 3: Support is of unknown length

Recall Lemma 2 for the set of extended dyadic intervals. The number of dyadic intervals of length 2^s is $M/2^s$ and the number of l-level extensions per dyadic interval is $2 \cdot 4^l$. The maximum length of an interval J from $\mathbf{J}_l[I]$ is $4|I|$ and the maximum number of frequencies considered on J is $4|I| \log(M)$. Thus, the number of test functions resulting from $\mathbf{J}_l[I]$ is bounded by

$$M/|I| \times 2 \cdot 4^l \times 4|I| \log M = 8 \cdot 4^l \cdot M \log M.$$

There are $\log(M)+1$ to consider for the dyadic intervals, so the size of our dictionary satisfies

$$|\mathcal{F}| \leq 16 \cdot 4^l \cdot M (\log M)^2.$$

By similar arguments as for cases I and II, we get

$$P\left(\sup_{g \in \mathcal{F}} |\langle Z, g \rangle| > \sqrt{2(1+\eta)\log M} \right) \to 0, \quad \text{as } M \to \infty,$$

as long as $l = l_M$ does not increase too fast ($l_M = o(\log M)$ is sufficient).

F.4.6 Proof of Theorem 17:

Assume

$$y_n = \alpha_M e^{i(\omega n+\theta)} \frac{1_I(n)}{\sqrt{|I|}} + z_n$$

with $\alpha_M \geq \sqrt{2(1+\eta)\log M}$. We wish to show, using an appropriate sequence of thresholds, the test T_M^* rejects $H_{0,M}$ with overwhelming probability. Then,

$$T^* \geq \left| \langle y, e^{i\omega_k n} \frac{1_J(n)}{\sqrt{|J|}} \rangle \right|$$

for every frequency $\omega_k \in \Omega(\mathbf{J})$ and every interval $J \in \mathbf{J}$. In particular, this inequality holds for an interval J such that

$$\rho(I, J) \geq 1 - 2^{-b}$$

for some b depending on either l or d in such a way, that, as l or d increases without bounds as $M \to \infty$, so does b. Such a b exists by Lemma 27 and Lemma 2. On the interval J, there is a frequency $\omega_k \in \Omega$ such that

$$|\omega - \omega_k| \leq \frac{\pi}{|J| \log M}.$$

From previous results we can write

$$\begin{aligned}
\langle y, e^{i\omega_k n} \frac{1_J(n)}{\sqrt{|J|}} \rangle &= \alpha_M \langle e^{i(\omega n+\theta)} \frac{1_I(n)}{\sqrt{|I|}}, e^{i\omega_k n} \frac{1_J(n)}{\sqrt{|J|}} \rangle + \langle z, e^{i\omega_k n} \frac{1_J(n)}{\sqrt{|J|}} \rangle \\
&= \alpha_M \frac{e^{i\theta}}{\sqrt{|I|}\sqrt{|J|}} \sum_{n \in I \cap J} e^{i(\omega-\omega_k)n} + W^1 + iW^2 \\
&= \beta_M + W^1 + iW^2,
\end{aligned}$$

where W^1 and W^2 are i.i.d. $N(0, 1/2)$ and we have replaced the deterministic term by β_M. Using the inequality, $\cos(x) \geq 1 - x^2/2$, we get

$$\cos((\omega - \omega_k)t) \geq 1 - (\omega - \omega_k)^2 t^2 \geq 1 - \frac{\pi^2}{|J|^2 (\log M)^2} t^2.$$

Then we get,

$$\begin{aligned}\left|\sum_{n\in I\cap J} e^{i(\omega-\omega_k)n}\right| &= \left|\sum_{n=0}^{|I\cap J|} e^{i(\omega-\omega_k)n}\right| \\ &\geq Re\left\{\sum_{n=0}^{|I\cap J|} e^{i(\omega-\omega_k)n}\right\} = \sum_{n=0}^{|I\cap J|}\cos((\omega-\omega_k)n) \\ &\geq |I\cap J|\left(1-\frac{\pi^2}{|J|^2(\log M)^2}|I\cap J|\right) \\ &\geq |I\cap J|\left(1-\frac{\pi^2}{(\log M)^2}\right),\end{aligned}$$

since $|I\cap J|\leq |J|$. Therefore,

$$\begin{aligned}|\beta_M| &= \left|\alpha_M\frac{e^{i\theta}}{\sqrt{|I|}\sqrt{|J|}}\sum_{n\in I\cap J} e^{i(\omega-\omega_k)n}\right| \geq \alpha_M\cdot\rho(I,J)\left(1-\frac{\pi^2}{(\log M)^2}\right) \\ &\geq \alpha_M(1-2^{-b})\left(1-\frac{\pi^2}{(\log M)^2}\right).\end{aligned}$$

Since $E|W^1+iW^2|^2=1$ and $|\cdot|$ is Lipschitz(1), Lemma 28 gives us,

$$P(|W^1+iW^2|>1+t)\leq 2e^{-t^2/2} \tag{F.3}$$

for all $t>0$. Let $\tau_{M,\alpha}$ denote the $1-\alpha$ quantile of T^*_M. Let $\alpha_M\to 0$ slowly enough such that

$$\tau_{M,\alpha_M}\sim\sqrt{2\log M},\quad \text{as } M\to\infty.$$

Theorem 15 implies that for all sufficiently large M we have

$$\tau_{M,\alpha_M}<\sqrt{2(1+\eta')\log M}.$$

Then, using the previous results and the triangular inequality,

$$\begin{aligned}P\left(T^*_M<\tau_{M,\alpha_M}\right) &\leq P\left(\left|\langle y, e^{i\omega_k n}\frac{1_J(n)}{\sqrt{|J|}}\rangle\right|<\tau_{M,\alpha_M}\right) = P\left(\left|\beta_M+W^1+iW^2\right|<\tau_{M,\alpha_M}\right) \\ &\leq P\left(|\beta_M|-|W^1+iW^2|<\tau_{M,\alpha_M}\right).\end{aligned}$$

Picking $t = |\beta_M| - \tau_{M,\alpha_M}$ and taking M and d (or l, in Case 3), large enough so that $t \to \infty$ as $M \to \infty$, we get from the inequality (F.3) that $P\left(|\beta_M| - |W^1 + iW^2| < \tau_{M,\alpha_M}\right) \to 0$ as $M \to \infty$, so

$$P\left(T_M^* < \tau_{M,\alpha_M}\right) \to 0, \quad M \to \infty.$$

F.5 Detection of Linear Chirps

We expect to get similar results for the detection of linear chirps instead of monochromatic sinusoids. We would replace the local sinusoids by local linear chirps in the GLRT, and use the same kind of techniques to establish upper bounds. The lower bound we had before still holds in this case since the set of sinusoids is a subset of the set of linear chirps.

Bibliography

[1] A. Abramovici, W. E. Althouse, R. W. P. Drever, Y. Gursel, S. Kawamura, F. J. Raab, D. Shoemaker, L. Sievers, R. E. Spero, and K. S. Thorne. LIGO-The Laser Interferometer Gravitational-Wave Observatory. *Science*, 256:325–333, April 1992.

[2] R. K Ahuja, T. L. Magnanti, and J. B. Orlin. *Network flows. Theory, algorithms and applications.* Prentice-Hall, New York, 1993.

[3] H. Akaike. A new look at the statistical model identification. *IEEE Trans. Automatic Control*, AC-19:716–723, 1974. (System identification and time-series analysis).

[4] B. Allen. Gravitational radiation analysis and simulation package (GRASP). Technical report, Department of Physics, University of Wisconsin, Milwaukee, P.O. Box 413, Milwaukee, WI 53201, 2000. Available at `http://www.lsc-group.phys.uwm.edu/~ballen/grasp-distribution/index.html`.

[5] W. G. Anderson and R. Balasubramanian. Time-frequency detection of gravitational waves. *Physical Review D (Particles, Fields, Gravitation, and Cosmology)*, 60(10):102001, 1999.

[6] W. G. Anderson and R. Balasubramanian. Time-frequency detection of gravitational waves. *Phys. Rev. D*, 60:102001, 1999.

[7] T. M. Apostol. *Calculus. Vol. I: One-variable calculus, with an introduction to linear algebra.* 2nd ed. Blaisdell Publishing Co. Ginn and Co., Waltham, Mass., 1967.

[8] T. A. Apostolatos, C. Cutler, G. J. Sussman, and K. S. Thorne. Spin-induced orbital precession and its modulation of the gravitational waveforms from merging binaries. *Phys. Rev. D*, 49:6274, 1994.

[9] E. Arias-Castro, E. J. Candès, H. Helgason, and O. Zeitouni. Searching for a trail of evidence in a maze. *The Annals of Statistics*, (in press) 2007.

[10] E. Arias-Castro, D. L. Donoho, and X. Huo. Near-optimal detection of geometric objects by fast multiscale methods. *IEEE Trans. Inform. Theory*, 51(7):2402–2425, 2005.

[11] E. Arias-Castro, D. L. Donoho, and X. Huo. Adaptive multiscale detection of filamentary structures in a background of uniform random points. *Ann. Statist.*, 34(1), 2006.

[12] R. G. Baraniuk and D. L. Jones. Shear madness: New orthonormal bases and frames using chirp functions. *IEEE Transactions on Signal Processing*, 41(12):3543–3548, 1993.

[13] A. Barron, L. Birgé, and P. Massart. Risk bounds for model selection via penalization. *Probab. Theory Related Fields*, 113(3):301–413, 1999.

[14] A. R. Barron and T. M. Cover. Minimum complexity density estimation. *IEEE Trans. Inform. Theory*, 37(4):1034–1054, 1991.

[15] I. Benjamini, R. Pemantle, and Y. Peres. Unpredictable paths and percolation. *Ann. Probab.*, 26(3):1198–1211, 1998.

[16] J. Berger, R. Coifman, and M. Goldberg. Removing noise from music using local trigonometric bases and wavelet packets. *J. Audio Eng. Soci.*, 42(10):808–818, 1994.

[17] L. Birgé and P. Massart. From model selection to adaptive estimation. In *Festschrift for Lucien Le Cam*, pages 55–87. Springer, New York, 1997.

[18] L. Birgé and P. Massart. Gaussian model selection. *J. Eur. Math. Soc. (JEMS)*, 3(3):203–268, 2001.

[19] S. Boyd and L. Vandenberghe. *Convex optimization.* Cambridge University Press, Cambridge, 2004.

[20] E. J. Candès. Multiscale chirplets and near-optimal recovery of chirps. Technical report, Stanford University, 2002.

[21] E. J. Candès. Modern statistical estimation via oracle inequalities. *Acta Numer.*, 15:257–325, 2006.

[22] E. J Candès, P. Charlton, and H. Helgason. Chirplets: recovery and detection of chirps. 2004. Presentation at the Institute of Pure and Applied Mathematics.

[23] E. J. Candès, P. Charlton, and H. Helgason. Detecting highly oscillatory signals by chirplet path pursuit. *Appl. Comput. Harmon. Anal.*, 2007.

[24] E. J. Candés and T. Tao. The dantzig selector: statistical estimation when p is much larger than n. *Ann. Statist.*, 2005. To appear.

[25] E. Chassande-Mottin and P. Flandrin. On the time-frequency detection of chirps. *Appl. Comput. Harmon. Anal.*, 6(2):252–281, 1999.

[26] E. Chassande-Mottin and A. Pai. Best chirplet chain: near-optimal detection of gravitational wave chirps. *Phys. Rev. D*, 73(4):042003 — 1–25, 2006.

[27] S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM J. Sci. Comput.*, 20(1):33–61, 1998.

[28] R. R. Coifman and M. V. Wickerhauser. Entropy-based algorithms for best basis selection. *IEEE Trans. Info. Theory*, 38:713–718, 1992.

[29] G. B. Dantzig, W. O. Blattner, and M. R. Rao. Finding a cycle in a graph with minimum cost to time ratio with application to a ship routing problem. In *Theory of Graphs (Internat. Sympos., Rome, 1966)*, pages 77–83. Gordon and Breach, New York, 1967.

[30] I. Daubechies. *Ten lectures on wavelets.* Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 1992.

[31] D. L. Donoho. CART and best-ortho-basis: a connection. *Ann. Statist.*, 25(5):1870–1911, 1997.

[32] D. L. Donoho and X. Huo. Beamlets and multiscale image analysis. In *Multiscale and multiresolution methods*, volume 20 of *Lect. Notes Comput. Sci. Eng.*, pages 149–196. Springer, Berlin, 2002.

[33] D. L. Donoho and J. Jin. Higher criticism for detecting sparse heterogeneous mixtures. *Ann. Statist.*, 32(3):962–994, 2004.

[34] D. L. Donoho and I. M. Johnstone. Ideal denoising in an orthonormal basis chosen from a library of bases. *C. R. Acad. Sci. Paris Sér. I Math.*, 319(12):1317–1322, 1994.

[35] D. L. Donoho and I. M. Johnstone. Minimax estimation via wavelet shrinkage. *Ann. Statist.*, 26(3):879–921, 1998.

[36] D. L. Donoho, I. M. Johnstone, G. Kerkyacharian, and D. Picard. Wavelet shrinkage: asymptopia? *J. Roy. Statist. Soc. Ser. B*, 57(2):301–369, 1995. With discussion and a reply by the authors.

[37] W. Feller. *An introduction to probability theory and its applications. Vol. I.* Third edition. John Wiley & Sons Inc., New York, 1968.

[38] É. É. Flanagan and S. Hughes. Measuring gravitational waves from binary black hole coalescences. i. Signal to noise for inspiral, merger, and ringdown. *Phys. Rev. D*, 57:4535, 1998.

[39] P. Flandrin. *Time-frequency/time-scale analysis*, volume 10 of *Wavelet Analysis and its Applications.* Academic Press Inc., San Diego, CA, 1999. With a preface by Y. Meyer, Translated from the French by J. Stöckler.

[40] D. P. Foster and E. I. George. The risk inflation criterion for multiple regression. *Ann. Statist.*, 22(4):1947–1975, 1994.

[41] D. Gabor. Theory of communication. *J. IEE*, 93:429–457, 1946.

[42] P. J. Green and B. W. Silverman. *Nonparametric regression and generalized linear models*, volume 58 of *Monographs on Statistics and Applied Probability.* Chapman & Hall, London, 1994.

[43] L. P. Grishchuk, V. M. Lipunov, K. A. Postnov, M. E. Prokhorov, and B. S. Sathyaprakash. Gravitational wave astronomy: in anticipation of first sources to be detected. *Physics-Uspekhi*, 73, 2000.

[44] P. C. Hansen. *Rank-deficient and discrete ill-posed problems.* SIAM Monographs on Mathematical Modeling and Computation. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1998.

[45] P. C. Hansen and D. P. O'Leary. The use of the L-curve in the regularization of discrete ill-posed problems. *SIAM J. Sci. Comput.*, 14(6):1487–1503, 1993.

[46] S. Haykin, B. W. Currie, and V. Kezys. Surface-based radar: coherent. In S. Haykin, E. O. Lewis, R. K. Raney, and J. R. Rossiter, editors, *Remote Sensing of Sea Ice and Icebergs*, pages 443–504. John Wiley and Sons, New York, 1994.

[47] R. A. Hulse and J. H. Taylor. Discovery of a pulsar in a binary system. *Astrophysical Journal*, 195:L51–L53, January 1975.

[48] M. F. Hutchinson and F. R. de Hoog. Smoothing noisy data with spline functions. *Numer. Math.*, 47(1):99–106, 1985.

[49] I. A. Ibragimov and R. Z. Has′minskii. *Statistical estimation*, volume 16 of *Applications of Mathematics*. Springer-Verlag, New York, 1981. (Asymptotic theory, Translated from the Russian by Samuel Kotz).

[50] Y. I. Ingster and I. A. Suslina. *Nonparametric goodness-of-fit testing under Gaussian models*, volume 169 of *Lecture Notes in Statistics*. Springer-Verlag, New York, 2003.

[51] J. M. Innocent and B. Torresani. Wavelets and binary coalescences detection. *Appl. Comput. Harmon. Anal.*, 4:113–116, 1997.

[52] S. Jaffard, Y. Meyer, and R. D. Ryan. *Wavelets*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, revised edition, 2001. Tools for science & technology.

[53] I. M. Johnstone. Wavelets and the theory of non-parametric function estimation. *R. Soc. Lond. Philos. Trans. Ser. A Math. Phys. Eng. Sci.*, 357(1760):2475–2493, 1999.

[54] H. C. Joksch. The shortest route problem with constraints. *J. Math. Anal. Appl.*, 14:191–197, 1966.

[55] B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *Ann. Statist.*, 28(5):1302–1338, 2000.

[56] M. Ledoux and M. Talagrand. *Probability in Banach spaces*, volume 23 of *Ergebnisse der Mathematik und ihrer Grenzgebiete (3) [Results in Mathematics and Related Areas (3)]*. Springer-Verlag, Berlin, 1991. (Isoperimetry and processes).

[57] S. Mallat. *A wavelet tour of signal processing.* Academic Press Inc., San Diego, CA, 1999.

[58] S. Mallat and Z. Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, 41:3397–3415, 1993.

[59] C. L. Mallows. Some comments on C_p. *Technometrics*, 15:661–676, 1973.

[60] S. Mann and S. Haykin. The chirplet transform: A generalization of Gabor's logon transform. *Vision Interface '91*, pages 205–212, June 3-7 1991. ISSN 0843-803X.

[61] S. Mann and S. Haykin. The chirplet transform: Physical considerations. *IEEE Transactions on Signal Processing*, 43(11):2745–2761, 1995.

[62] D. Mitrović and D. Žubrinić. *Fundamentals of applied functional analysis*, volume 91 of *Pitman Monographs and Surveys in Pure and Applied Mathematics.* Longman, Harlow, 1998. (Distributions—Sobolev spaces—nonlinear elliptic equations).

[63] M. Morvidone and B. Torresani. Time scale approach for chirp detection. *Int. J. Wavelets Multiresolut. Inf. Process.*, 1(1):19–49, 2003.

[64] A. V. Oppenheim, R. W. Schafer, and J. R. Buck. *Discrete-time signal processing (2nd ed.).* Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1999.

[65] B. J. Owen and B. S. Sathyaprakash. Matched filtering of gravitational waves from inspiraling compact binaries: Computational cost and template placement. *Physical Review D (Particles, Fields, Gravitation, and Cosmology)*, 60(2):022002, 1999.

[66] C. H. Reinsch. Smoothing by spline functions. I, II. *Numer. Math.*, 10:177–183; ibid. 16 (1970/71), 451–454, 1967.

[67] J. E. Reynolds III and S. A. Rommel. *Biology of Marine Mammals.* Smithsonian Institution Press, 1999.

[68] J. A. Rice. *Mathematical Statistics and Data Analysis.* Duxbury Press, Belmont, California, 2nd edition, 1995.

[69] F. Santosa and W. W. Symes. Linear inversion of band-limited reflection seismograms. *SIAM Journal on Scientific and Statistical Computing*, 7(4):1307–1330, 1986.

[70] M. Schwartz. *Infomation transmission, modulation, and noise.* McGraw-Hill, 4th edition, 1990.

[71] G. Schwarz. Estimating the dimension of a model. *Ann. Statist.*, 6(2):461–464, 1978.

[72] A. Sha' ashua and S. Ullman. The detection of globally salient structures using a locally connected network. In *Proceedings of the Second International Conference on Computer Vision*, pages 321–327, 1988.

[73] J. Simmons. Echolocation in bats: signal processing of echoes for target range. *Science*, 171(974):925–928, 1971.

[74] J. Simmons. The resolution of target range by echolocating bats. *The Journal of the Acoustical Society of America*, 54:157–173, 1973.

[75] C. J. Stone. Consistent nonparametric regression. *Ann. Statist.*, 5(4):595–645, 1977. (With discussion and a reply by the author).

[76] J. H. Taylor, L. A. Fowler, and P. M. McCulloch. Measurements of general relativistic effects in the binary pulsar psr1913 + 16. *Nature*, 277:437–440, 1979.

[77] The LIGO Scientific Collaboration. LAL Software Documentation. http://www.lsc-group.phys.uwm.edu/lal/slug/nightly/doc/lsd-nightly.pdf.

[78] J. A. Thomas, C. F. Moss, and M. Vater. *Communication and Behaviour of Whales.* AAAS Selected Symposium, 1983.

[79] J. A. Thomas, C. F. Moss, and M. Vater. *Echolocation in bats and dolphins.* The University of Chicago Press, 2004.

[80] K. S. Thorne. Graviational radiation. In S. W. Hawking and W. Israel, editors, *300 Years of Gravitation*, pages 330–358. Cambridge University Press, Cambridge, England, 1987.

[81] R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58(1):267–288, 1996.

[82] J. Ville. Théorie et applications de la notion de signal analytique. *Cables et Transmissions*, 2A:61–74, 1948.

[83] L. F. Villemoes. Adapted bases of time-frequency local cosines. *Appl. Comput. Harmon. Anal.*, 10(2):139–162, 2001.

[84] P. Westfall and S. Young. *Resampling-based Multiple Testing: Examples and Methods for P-value Adjustment.* Wiley, New York, 1993.