Pontificia Universidad Católica de Chile
Facultad de Física
Departamento de Astronomía y Astrofísica

Component Analysis of astrophysical time series:

an application to transiting exoplanets.

by

Néstor Espinoza Perez

Practice report presented to the Physics Faculty
of Pontificia Universidad Católica de Chile,
as one of the requirements for the
Bachelor degree in Astronomy.

Supervisor	:	Dr. Andrés Jordán (PUC, Chile)
Correctors	:	Dr. Alejandro Clocchiatti (PUC, Chile)
		Dr. Márcio Catelan (PUC, Chile)

January, 2012

Santiago, Chile

Acknowledgments

First of all, I'd like to thank my supervisor, prof. Andrés Jordán, for giving me the oportunity and honor to work with him. I want to thank him not only as a supervisor but as a teacher: his ideas, support, trust and motivation towards my work has been an infinite source of motivation which much of the time made me feel more like a colleague than a student, which was an invaluable experience.

I want to thank my family for encouraging me and listening to me. Specially, I want to thank my mother, Gloria, for her infinite support and late night talks, which encouraged me to keep working hard. Thanks to my father who was also always interested in the new topics that I was learning and studying. I also want to specially thank my girlfriend, María Paz, for her infinite support, love and care. Much of her thoughtful critiques and my own attempts at explaining her the topics of this thesis motivated me to improve my own knowledge of certain topics and, therefore, the writing and re-writing of various passages of it. Thanks also goes for my friends Diego, Boris, Bárbara and Robert (and Iivari!). We didn't really had much time to see each other this year, but every time we did I had a really good time.

Last but not least, I want to thank all of the members of the Física Itinerante (FI) group. Thanks to Claudia for her infinite support not only as part of the administration of the group but also as a very good friend. Thanks to Naoto, Ignacio, Isabel, Tatiana, Jorge, Jaqueline, Gonzalo, Gabriela, Fabiola and José Miguel for their incredible work on all of our projects. Thanks also for all the talks and support regarding the administration of the group which, together with this thesis, where my two biggest projects for this year. Without all of you working with me in those projects, I wouldn't have been able to write this thesis: it has been an honor to work with all of you.

Contents

Introduction

Transiting extrasolar planets (a.k.a. transiting "exoplanets") is one of the "hot topics" in astrophysics today. An extrasolar planetary transit is defined as the passing of a planet in front of the disk of the star: if we are lucky enough to see this event, measuring the star's flux as a function of time (called a lightcurve) will reveal the "signature" of the exoplanet, as some of the starlight is blocked as the planet passes between us and the star. Since the pioneering work of Charbonneau et al. (2000), where the first detection of a transiting exoplanet was made, more than 130 transiting exoplanets have been discovered to date[1], leading to an important improvement of our knowledge of the physics behind planetary systems, including our own. One of the main advantages of exoplanetary transits is that they allow us to constrain the physical parameters of the systems given some radial velocity measurements and some knowledge of the parent star (Winn, 2010). In particular, the radius of the planet, R_p is a function of the depth of the transit, δ: given a stellar radius, the higher the depth, the higher the exoplanetary radius (i.e. more starlight is blocked by it).

Perhaps the most interesting feature of transiting exoplanets is that they allow us to investigate exoplanetary atmospheres. This can be done exploiting the fact that during a transit, some of the starlight is *transmitted* through the upper atmosphere of the exoplanet, leaving spectral signatures in the recieved stellar fluxes. Observing the transit at different wavelengths, then, one can search for differences in the measured radius of the exoplanet. If at a certain wavelength the upper atmosphere is opaque, then we should see that the radius of the planet increases (because the transit depth increases). This technique, which lets us measure the opacity and, therefore, strong spectral features of elements present in the upper atmosphere, is called *transmission*

[1] http://www.exoplanets.org

spectroscopy (Seager & Sasselov, 2000; Brown, 2001; Seager, 2010).

Despite its importance in order to constrain parameters of dynamic atmospheric models (e.g. Fortney et al., 2010) and detection of biomarkers of possible habitable planets (e.g. Kaltenegger & Traub, 2009) among other possible applications, measuring transmission spectra is not an easy task. Just measuring the planetary transit requires that we can measure variations on the stellar flux of the order of 3%; measuring flux variations on an area as small as the transparent part of the upper atmosphere of an exoplanet ($\sim 10^{-3}$ to 10^{-4} of the area of the star) requires even more precision. This precision is usually not constrained by independent and identically distributed sources of noise (e.g. Poisson statistics) but by the so-called **systematic errors** present in the measurements.

Systematic errors can be generated from a large number of sources: telescope guiding, changing airmass, atmospheric conditions, instrumental variations, etc. Models for estimating these systematic errors usually assume that this error is **correlated**, i.e., values at times t of a given lightcurve depend on past values not only as functions, but as random procceses. For example, in the context of transit light curves, Pont et al. (2006) suggested that the combination of some of these different sources can be modelled as "red noise", a stochastic process with a certain covariance structure. With this problem in mind, Mazeh & Tamuz (2007) proposed the SysRem Detrending Algorithm, which actually searches for linear correlations between the lightcurves and any information available (e.g. airmass), without really caring which control variables it de-correlates. Once they have obtained the structure of one systematic error, they simply substract it from the light curve, and search for a new systematic error, continuing this process iteratively until no more information can be obtained.

Until recently, the Sys-rem Detrending Algorithm was perhaps one of the most widely used algorithms to detect and substract systematic errors from transit data. However, the need for methods and models that could explain some residual signals that are not removed using these kind of algorithms (because of (a) lack of knowledge of more information or (b) simply because theses signals are not linearly correlated) in order to substract them from the transit light curves are needed in order to obtain credible distributions for the obtained parameters of the transits. This particular problem has been recently attracting the atention of the astrophysical community, because the removal of the sources that affect our measurements is what separates us from a new discovery. One powerful approach to the solution of this problem was introduced by Carter & Winn (2009), where one of the noise components present in the time series is assumed to be a stationary gaussian process called "flicker" noise, which has similar properties as the noise model proposed by Pont et al. (2006) (and that, in fact, could be seen as a generalization of the "red noise"), where the parameters of the noise can now be modelled easily thanks to a linear transformation of the time series to a wavelet basis. Trying to generalize this approach in order to use known information about systematic effects, Gibson et al. (2011) recently introduced the usage of general gaussian processes, which fit this stochastic structure, together with deterministic control variables (e.g. instrumental effects) to the data.

Another approach to the solution of the problem is to tackle it using Principal Component Analysis for time series (see, for example, Jolliffe, 2002; Ramsay and Silverman, 1997, 2002). Basically, PCA does the same as the Sys-rem algorithm, assuming that the measurement errors are negligible (i.e. assuming high S/N) but without any a-priori information about the sources that generate the trends: it searches for the "best uncorrelated time series" that can explain our data. This approach has been recently used for transmission spectra, yielding very good results (Thatte et al., 2010). However, thinking of the sources of noise as uncorrelated from each other rises doubts about the model, because linear uncorrelatedness is an entirely arbitrary assumption. With this in mind, Waldmann (2011) recently proposed to tackle the problem using Independent Component Analysis (ICA) for time series, where it is argued that if we could measure independent measurements of the transit light curves then we could, just by assuming that these measurements are composed of independent components, i.e., time series that are **independent** from each other, separate the transit from other noise sources. This could be seen like an improvement to the Principal Component Analysis approach, where now we are looking for the best independent time series that form our observed transits.

This last paper is the principal motivation of the present thesis. What we propose is to tackle the problem using Independent Component Analysis **before** the transit is obtained, using the comparison stars. The assumption here is that the observed flux of each comparison star and the objective star, in the absense of the transit, is formed from the same stochastic and deterministic procceses, with the only difference that these processes are weighted differently depending on the star. Therefore, if we could find the independent components that form our observed time series for each comparison star, then we could fit the proper weights for our objective star adding the transit model, thus solving the problem. This can be viewed as some kind of generalization of the Sys-rem algorithm, where the information comes from the assumption that the sources are independent.

One of the objectives of this thesis is to write routines that perform PCA and ICA in order to compare these techniques. Because of this, a deep understanding on the foundations of PCA and ICA is needed and, therefore, the theory and applications will be discussed in the present thesis. Another objective of the present work is to test if it is possible to measure with sufficient precision the planetary radius given a transit light curve with these methods. The motivation for this is that, as stated before, important applications such as transmission spectroscopy requiere high precision measurements of the planetary radius. In particular, most applications needs to find changes at an order of a 1% of the planetary radius.

This thesis is organized as follows. Chapter 1 presents the mathematical background needed to pose the above stated problems and methods in terms of probability theory, information theory and stochastic processes. Chapter 2 introduces the problem of Blind Source Separation and presents the two techniques to be studied and tested in the present thesis: Principal Component Analysis and Independent Component Analysis. Finally, Chapter 3 presents applications of the techniques presented in Chapter 2 to astrophysical time series in the context of exoplanet transits. In particular, we'll use spectroscopic measurements of the WASP-6 star which is known to have a transiting exoplanet (Gillon et al., 2009) in order to obtain it's white light transit curve, i.e., the observed transit when the measured flux in all wavelengths is summed. Our results will then be compared to previous measurements in order to test the precision of the applied techniques.

Chapter 1

Mathematical background

The aim of this chapter is to give the reader a mathematical background on the tools used on this practice report. First, some important concepts on statistics are reviewed in order to set the terminology to be used in the following sections and chapters. The references used are Shao (2003), Gregory (2005) and the first chapters in Hyvrinen, Karhunen & Oja (2000).

Next, time series analysis is described in order to introduce the reader to the typical analysis tools and definitions used in the related literature. The main references used in this section are the authoritative books on time series of Box & Jenkins (1976), Brockwell & Davis (2001) and Broersen (2006), and the reviews of Scargle (1981) and Koen & Lombard (1993).

1.1 Important concepts on Statistics

1.1.1 Random variables and probability distributions

In frequentist statistics, in much of the scientific literature, and in this work, the concept of a **random variable** is used. A random variable X is formally a measurable function from a *sample space* (the space of all possible outcomes of an experiment) and the collection of all its *events* (subsets of a sample space[1], i.e., subsets of the possible outcomes) to the real line and all its subsets (Shao, 2003). Intuitively, a random variable X transforms the possible outcomes of an experiment

[1] In fact, this subset needs certain properties to be matched in order to be called a subset of events. For a formal discussion on Probability Theory, see the references cited.

(i.e. the measurement operation) to real numbers.

In order to summarize the available information or the degree of certainty given the context of an experiment, we associate to each random variable a **probability distribution** which defines the probability of obtaining a certain set of values from the random variable X. Because the distributions define probabilities, they are always positive. In the discrete case, this distribution is called a probability mass function (PMF) and it describes the probability of obtainig a value $X = x$. In the continuous case this is called the probability density function (PDF) and it describes the probability density of obtaining a value between $x < X < x + dx$. We will focus on this last kind of probability distribution because most of the variables that are of interest for us are continuous.

If we denote the PDF by $f_X(X = x)$, or simply, $f_X(x)$, the probability $P(a < X < b)$ of obtaining a value of X between a and b is,

$$P(a < X < b) = \int_a^b f_X(x)dx$$

In the limits $a \to -\infty$ and $b \to \infty$ this integral is defined to be 1, i.e. there is complete certainty to obtain a value,

$$P(-\infty < X < \infty) = \int_{-\infty}^{\infty} f_X(x)dx = 1$$

The importance of this definition is that, by definition of a PDF, $f_X(x)$ is an integrable function, i.e. $f_X(x) \in \mathbf{L^1}(\mathbb{R})$. We'll make use of this property in the latter sections.

1.1.2 Moments, Characteristic Functions and Cumulants

In order to get a feeling of how the PDF looks like, we first define the **first moment of a random variable**, also known as the mean. It is defined as:

$$\mu = E\left[X\right] = \int_{-\infty}^{\infty} x f_X(x)dx$$

This is a measure of **location** of the random variable, i.e., a measure of where X is located with respect to the probability distribution that defines it. In physical terms it is analogous to the center of mass. Note that the expected value, $E[\cdot]$, can also be viewed as a linear operator.

In general, the expectation value of any function $g(X)$ of a given random variable is defined to be:

$$E\left[g(X)\right] = \int_{-\infty}^{\infty} g(x) f_X(x) dx$$

In this sense, the mean of a random variable would be a special case of this definition, setting $g(X) = X$.

Another important function frequently used in statistics is defined as the **n-th central moment**, which is defined by the expected value of the function $g_n(X) = (X-\mu)^n$:

$$\mu_n = E\left[(X-\mu)^n\right] = \int_{-\infty}^{\infty} (x-\mu)^n f_X(x) dx$$

The n moments usually serve to partially characterize a given probability distribution with respect to its mean μ. For example, the 2nd central moment is also known as the variance and, if finite, it can be viewed as a measure of the spread of the distribution around the mean: the longer the variance the longer the spread. Note that the n-th central moment might not be defined so one must be careful with the interpretations[2]: the integral needs to converge for the moment to exist. In practice, we can almost always find the mean of a given distribution and can even set $\mu = 0$ by substracting a sample of it to the data. In this case, the central moment is:

$$\mu_n = E\left[X^n\right] = \int_{-\infty}^{\infty} x^n f_X(x) dx$$

The problem with the central moments is their "lack of existence", i.e., there is no reason to believe that they actually exist for a given distribution. Therefore, a somewhat more intuitive and general form of characterization of a random variable has to be defined. This form is given by the **characteristic function** (CF), $\hat{f}_X(\omega)$, which is usually introduced by the **moment generating function** (MGF), $M_X(t)$, of the PDF. The latter function is defined to be the expected value of the function

[2] A classical example is that of the Cauchy-Lorentz distribution, where the first moment does not exist and, therefore, the nth central moment doesn't either.

$g(X) = e^{Xt}$:

$$M_X(t) = E\left[e^{Xt}\right] = \int_{-\infty}^{\infty} e^{xt} f_X(x)dx \tag{1.1}$$

It is called a "moment generating" function because if we differentiate this equation n times with respect to t we obtain:

$$\frac{d^n}{dt^n} E\left[e^{Xt}\right] = \int_{-\infty}^{\infty} x^n e^{xt} f_X(x)dx$$

Which evaluated at $t = 0$, gives the nth central moment. The CF, on the other hand, can be defined by letting t be a pure imaginary number, say $t = -i\omega$,

$$\hat{f}_X(\omega) = M_X(-i\omega) = E\left[e^{-i\omega X}\right] = \int_{-\infty}^{\infty} e^{-i\omega x} f_X(x)dx \tag{1.2}$$

Which is the ($\sqrt{2\pi}$ multiplied) Fourier Transform of the PDF. Because $f_X(x) \in \mathbf{L^1}(\mathbb{R})$ (by the definition of a Probability Density Function), the characteristic function $\hat{f}_X(\omega)$ always exists and is necesarily a continuous function of ω. This, of course, is not true in general for the moment generating function (or even for the moments, as exposed earlier). This function also makes it easier to interpret the moments of a PDF (if they exist). We can expand the exponential term in equation (1.2) in its Taylor series expansion around $\omega = 0$ to get:

$$\hat{f}_X(\omega) = \int_{-\infty}^{\infty} \sum_{n=0}^{\infty} \frac{(-i\omega x)^n}{n!} f_X(x)dx = \sum_{n=0}^{\infty} \frac{(-i\omega)^n}{n!} E\left[X^n\right] \tag{1.3}$$

Thus, the moments are part of the coefficients of the power series expansion of $\hat{f}_X(\omega)$. If the moments exist one may be tempted to argue that they also characterize completely the random variable but, as was shown by Heyde (1963), this is not necesarly true. What can be argued instead is as follows: if the radius of convergence of equation (1.3) is finite, the CF is unique, and the characteristic function can be used as a full characterization of a random variable (this is actually why the characteristic function has its name, although it can be misleading). Because we can usually generate estimates of the moments, this may seem everything we need to characterize a given sample of a random variable. The problem with moments is that, in general, they have complicated forms when it comes to characterize important random variables like a normally distributed one or when it comes to represent moments of random variables that are themselves functions of other random variables. The details on these complications will be given on future subsections, but these facts motivates the search of a similar function that, at the very least, present simplifications to those problems: this introduces the theory of **cumulants**.

To define a cumulant, first we have to introduce the **second characteristic function**, $\phi(\omega)$, which is defined as:

$$\phi(\omega) = \ln\left(\hat{f}_X(\omega)\right)$$

Why the natural logarithm of the CF is taken to define this function will be made clear in short. The basic idea is to first expand the second characteristic function in its Taylor series expansion and assume that we can write that expansion in the same way as we did in equation (1.3) for the characteristic function, but replacing the nth moment by what is called the **nth cumulant**, κ_n:

$$\phi(\omega) = \ln\left(\hat{f}_X(\omega)\right) = \sum_{n=0}^{\infty} \frac{(-i\omega)^n}{n!}\kappa_n \tag{1.4}$$

By the definition of a Taylor series expansion, the nth cumulant can be obtained by diferentiating n times the second characteristic function:

$$\kappa_n = i^n \left.\frac{d^n\phi}{d\omega^n}\right|_{\omega=0}$$

The first three cummulants, κ_0, κ_1 and κ_2, can be easily shown to be 0, the mean and the variance of the distribution, respectively. On the other hand, the fourth and fifth cumulants can be shown to be given by:

$$\begin{aligned}
\kappa_3 &= E[X^3] - 3E[X^2]E[X] + 2E[X]^3 \\
\kappa_4 &= E[X^4] - 3E[X^2]^2 - 4E[X^3]E[X] + 12E[X^2]E[X]^2 - 6E[X]^4
\end{aligned}$$

One of the most important properties of cumulants to be used in the present work is the fact that a normally distributed random variable, i.e. one with a PDF of the form $f_X(x) = \exp\left(-(x-\mu)^2/2\sigma^2\right)/\sqrt{2\pi\sigma^2}$ has a CF:

$$\hat{f}_X(\omega) = \exp\left(\frac{-\omega(\omega\sigma^2 + i2\mu)}{2}\right)$$

and, therefore, according to equation (1.4), its second CF is:

$$\phi(\omega) = \omega\left(-i\mu\right) + \omega^2\left(-\frac{\sigma^2}{2}\right)$$

What this result shows is that **only two of all the cummulants of a normally distributed random variable (the second and the third) are non-zero**. Furthermore, according to Marcinkiewicz's theorem (Rajagopal & Sudarshan, 1974), the normal distribution is the only one that has the property of having a finite number of non-zero cummulants, i.e., calculating higher order cummulants is some kind of measure of the degree of gaussianity (or not) of a random variable.

1.1.3 Extensions to higher dimensions, pt. 1

The concepts treated in past subsections have a natural extension to higher dimensions. Consider an experiment where m random variables are obtained. To characterize this set of random variables we now speak of a **joint probability distribution**, with distinctions similar as before but extended to an m-dimensional space. If we denote this set of m continuous random variables by $\vec{X} = (X_1, X_2, ..., X_m)^T$, also called a **random vector**, the joint probability density function $f_X(\vec{X} = \vec{x})$ or simply $f_X(\vec{x})$ asociated to this set, where $\vec{x} = (x_1, x_2, ..., x_m)^T$, describes the probability of obtaining values in the set $D = \{x_j < X_j < x_j + dx_j \mid j = 1, 2, ..., m\}$ of the m real lines. In particular, the probability of obtaining values in the set $A = \{a_j < X_j < b_j \mid j = 1, 2, ..., m\}$, $P(A)$, is

$$P(A) = \int_{a_1}^{b_1}\int_{a_2}^{b_2}\dots\int_{a_m}^{b_m} f_X(\vec{x})dx_1dx_2...dx_m = \int_A f_X(\vec{x})d^mx$$

Here again, by definition of a joint PDF, in the limit where every $a_j \to -\infty$ and $b_j \to \infty$ for $j = 1, 2...m$, also called the sample space Ω, the integral is 1,

$$P(\Omega) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\dots\int_{-\infty}^{\infty} f_X(\vec{x})dx_1dx_2...dx_m = \int_\Omega f_X(\vec{x})d^mx = 1$$

If we wish to find the PDF of a particular random variable, we can **marginalize** that PDF by integrating from $-\infty$ to ∞ for all but that particular random variable. For example, suppose we wanted to find the PDF for the p-th random variable X_p.

Then,

$$f_{X_p}(X_p = x_p) = \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} f_X(\vec{x}) dx_1 dx_2 ... dx_{p-1} dx_{p+1} ... dx_m$$

This is called the *p-th marginal* PDF.

What if we now consider **a set of random vectors?** Consider now n random vectors $\vec{Y_i}$, where $i = 1, 2, ..., n$. In general, because each random vector can have different lengths m_i, we associate them as an array of vectors $\mathbf{Y} = \left(\vec{Y_1}, \vec{Y_2}, ..., \vec{Y_n}\right)$. With this in mind, the PDF of this set is denoted by $f_{\mathbf{Y}}(\mathbf{Y} = \mathbf{y})$, where $\mathbf{y} = (\vec{y_1}, \vec{y_2}, ..., \vec{y_n})$ or simply $f_{\mathbf{Y}}(\vec{y_1}, \vec{y_2}, ..., \vec{y_n})$. All of the probabilities exposed earlier for a single vector can be naturally expanded and, in particular, the normalization condition of the PDF reads:

$$P(\Omega) = \int_{\Omega} f_{\mathbf{Y}}(\vec{y_1}, \vec{y_2}, ...) d^{m_1} y_1 d^{m_2} y_2 ... d^{m_n} y_n = 1$$

Also, the marginalization procedure explained before can be applied in order to obtain the marginal PDFs of the different random vectors present in our set.

What about moments in higher dimensions? The first moment of a random vector is easy to extend, and can be defined as the vector:

$$\vec{\mu} = E\left[\vec{X}\right] = \int_{\Omega} \vec{x} f_X(\vec{x}) d^m x$$

Note that each component of the $\vec{\mu}$ vector is a weighted marginalization of the random variables of the random vector. In general, the expectation value of any vector function $\vec{g}(\vec{X})$ can be defined in a similar way:

$$E\left[\vec{g}\left(\vec{X}\right)\right] = \int_{\Omega} \vec{g}\left(\vec{x}\right) f_X(\vec{x}) d^m x$$

According to this definition, the first moment can be viewed as a special case where $\vec{g}\left(\vec{X}\right) = \vec{X}$. The case where we use the array of random vectors $\mathbf{Y}$ is analogous. The second moment, on the other hand, can't be extended as just a vector because, in analogy to the one dimensional second moment, the multi-dimensional counterpart has to give us information about the m-dimensional spread of our distribution. Furthermore, what happens if one or more random variables of the random vector actually gives information about other random variables? This would clearly alter the shape of the distribution. This discussion motivates the next subsection which treats the concept of independance and correlation.

1.1.4 Independance and correlation

The concept of marginalization rises a question: we can see that the joint PDF determines the marginal PDF's, but, do the marginal PDF's determine the joint PDF alone? We may rise this question in a more intuitive form: does the information on one random variable (i.e. the marginal PDF) give us information about another random variable? If the answer is yes, then the joint PDF can't be formed by just knowing the marginal PDF's of each random variable, because these two random variables are **dependant of each other** and therefore the information is redundant (i.e. knowing the PDF of one of these random variables gives information about the other). With this idea in mind, we now define the concept of **independance**.

Definition

*Consider the random vector $\vec{X} = (X_1, X_2, ..., X_m)$ with joint PDF $f_X(\vec{x})$. We say that the random variables are **independent** if, and only if,*

$$f_X(X_1, X_2, ..., X_m) = f_{X_1}(X_1) f_{X_2}(X_2) ... f_{X_m}(X_m)$$

The definition given above can be extended to an array of random vectors as well. If we consider again the array of the n random vectors $\mathbf{Y} = (\vec{Y_1}, \vec{Y_2}, ..., \vec{Y_n})$, we say that the random vectors are **independent** if, and only if,

$$f_{\mathbf{Y}}(\vec{y_1}, \vec{y_2}, ..., \vec{y_n}) = f_{Y_1}(\vec{y_1}) f_{Y_2}(\vec{y_2}) ... f_{Y_n}(\vec{y_n})$$

Consider now a definition that is some sort of "weaker form of independance" called the **linear correlation** between random variables. It is defined as follows.

Definition

Consider the random vector $\vec{X} = (X_1, X_2, ..., X_m)$ with joint PDF $f_X(\vec{X})$. We say *that the random variables are* ***linearly uncorrelated*** *if, and only if*

$$E\left[X_1X_2...X_m\right] = E\left[X_1\right]E\left[X_2\right]...E\left[X_m\right]$$

This is the most used form of defining uncorrelatedness and is usually just said that if a set of random variables have this property they are simply uncorrelated. The problem with this definition is that some random variables may be non-linear functions of other random variables and the definition may still apply, misleading the conclusion of "correlation". For example, suppose we have two random variables X and Y, and suppose that $Y = X^2$ (i.e. they are dependant of each other). Furthermore, suppose that $E\left[X\right] = E\left[X^3\right] = 0$ (think of any symmetric distribution with zero mean, such as a normal random variable). Then, it is easy to show that

$$E\left[XY\right] = E\left[X\right]E\left[Y\right] = 0$$

And the random variables are linearly uncorrelated. On the other hand, note that in general:

$$E\left[h(X)g(Y)\right] \neq E\left[h(X)\right]E\left[g(Y)\right]$$

This more general property, called **non-linear uncorrelation**, is a much stronger form of independance (note that in fact this property would be true if the random variables where independent). If we again consider the random vector $\vec{X} = (X_1, X_2, ..., X_m)$ with joint PDF $f_X(\vec{X})$ and a set of functions $g_1(X), g_2(X)..., g_m(X)$, we say that the random variables are non-linearly uncorrelated if, and only if,

$$E\left[g_1(X_1)g_2(X_2)...g_m(X_m)\right] = E\left[g_1(X_1)\right]E\left[g_2(X_2)\right]...E\left[g_m(X_m)\right]$$

1.1.5 Extensions to higher dimensions, pt. 2

Having defined the concepts of independance and uncorrelation, we are ready to define the concepts of higher order moments. In order to define the second moment, we first have to introduce the concept of **covariance**, which measures how (linearly) correlated two random variables are:

$$\text{Cov}\left(X_i, X_j\right) = E\left[\left(X_i - E[X_i]\right)\left(X_j - E[X_j]\right)\right]$$

The special case $i = j$ defines the variance of the i-th random variable, σ_i^2. Also note that if the covariance is 0, then the random variables are (linearly) uncorrelated. To standarize this measurement, we define the **correlation coefficient** between two random variables as:

$$\rho(X_i, X_j) = \frac{\text{Cov}(X_i, X_j)}{\sigma_i \sigma_j}$$

This coefficient equals one when the random variables are completely (linearly) correlated (e.g., $X_i = X_j$) and equals zero when the variables are (linearly) uncorrelated. In this sense, the correlation (and hence the covariance) between two random variables also serves to characterize how the two random variables actually grow together. If the correlation is positive, then the two random variables increase together. If the value is negative, then as one variable increases the other decreases.

With these definitions in mind, the second moment is defined entirely in a very compact and useful form: the **covariance matrix $\mathbf{\Sigma}$**. The (i, j) element of this matrix is defined as the covariance between the ith and the jth random variable in our random vector, i.e.,

$$\mathbf{\Sigma} = E\left[(\vec{X} - \vec{\mu})(\vec{X} - \vec{\mu})^T\right] = \begin{bmatrix} \sigma_1^2 & \text{Cov}(X_1, X_2) & \dots & \text{Cov}(X_1, X_m) \\ \text{Cov}(X_2, X_1) & \sigma_2^2 & \dots & \text{Cov}(X_2, X_m) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(X_m, X_1) & \text{Cov}(X_m, X_2) & \dots & \sigma_m^2 \end{bmatrix}$$

Note that the covariance matrix is symmetric (because the covariance is invariant to permutations) and positive semidefinite, properties that will prove to be useful in future sections. On the other hand, if all the random variables are uncorrelated with each other, this transforms into a purely diagonal matrix.

We end this section with a multidimensional re-definition of the characteristic function, which in turn will define all the remaining higher order cummulants and hence the moments. The ($(2\pi)^{m/2}$ multiplied) multidimensional Fourier Transform of a function $f_X(\vec{x})$, and hence, the multidimensional characteristic function $\hat{f_X}(\vec{\omega})$, is defined as:

$$\hat{f_X}(\vec{\omega}) = \int_\Omega e^{-i\vec{\omega}\cdot\vec{x}} f_X(\vec{x}) d^m x$$

Expanding the exponential function in its multi-dimensional Taylor series expansion around $\vec{\omega} = \vec{0}$, we get:

$$\hat{f}_X(\vec{\omega}) = \int_\Omega \sum_{\sum \alpha_k \geq 0} \prod_{k=1}^{m} \frac{(-i\omega_k x_k)^{\alpha_k}}{\alpha_k!} f_X(\vec{x}) d^m x = \sum_{\sum \alpha_k \geq 0} \prod_{l=1}^{m} \frac{(-i\omega_l)^{\alpha_l}}{\alpha_l!} E\left[\prod_{k=1}^{m} X_k^{\alpha_k}\right],$$

where $\alpha_k \in \mathbb{N}_0$, with $k = 1, 2, .., m$. Note that this definition includes the one-dimensional version of the characteristic function for each of the variables in $\vec{\omega}$ and hence includes the definition of moments in that case, which can be obtained if we set $\alpha_k = 0$ for $k \neq 1$. It also gives an interpretation to higher order moments, such as the covariance, which for zero-mean random variables are part of this expansion. With this in mind, the covariance matrix sumarizes "second order information" of the above presented expansion. Analogously to the one-dimensional case, the second characteristic function can be expressed as

$$\phi(\vec{\omega}) = \ln\left(\hat{f}_X(\vec{\omega})\right) = \sum_{\sum \alpha_k \geq 0} \prod_{k=1}^{m} \frac{(-i\omega_k)^{\alpha_k}}{\alpha_k!} \kappa(\alpha_1, \alpha_2, ..., \alpha_m),$$

where the higher-order cummulant is now a function of the α_k, and can be obtained according to the definition of the multi-dimensional Taylor series:

$$\kappa(\alpha_1, \alpha_2, ..., \alpha_m) = i^{\sum \alpha_k} \frac{\partial^{\alpha_1}}{\partial \omega_1^{\alpha_1}} \frac{\partial^{\alpha_2}}{\partial \omega_2^{\alpha_2}} ... \frac{\partial^{\alpha_m}}{\partial \omega_m^{\alpha_m}} \phi(\vec{\omega})\bigg|_{\vec{\omega}=\vec{0}}$$

This, of course, also reduces to the one dimensional case if we set set $\alpha_k = 0$ for $k \neq 1$. Accordingly, Marcinkiewicz's theorem has also its higher order version (for a multi-dimensional gaussian distribution), which is proved in Rajagopal & Sudarshan (1974).

1.2 Entropy

One of the most important concepts that we'll use in the present work is the concept of the entropy of a distribution. The motivation of this definition is that, in practice, we can't always exactly decide which distribution actually explains our data, because we only have little information available. This information comes as estimators of some set of functions $F = \{F_j(X) \mid j = 1, ..., m\}$ of the random variables, which can be expressed theoretically as

$$E\left[F_j(X)\right] = \int_{\Omega} f_X(x) F_j(x) dx = c_j \qquad \text{(continuous case) and} \tag{1.5}$$

$$E\left[F_j(X)\right] = \sum_{i=1}^{n} f_X(x_i) F_j(x_i) = c_j \qquad \text{(discrete case),} \tag{1.6}$$

where, for the discrete case, $f_X(x_i)$ is the probability of obtaining the value x_i, which we'll call p_i to simplify the notation. For example, the minimum information that we can have for a PDF or a PMF is its normalization, which can be interpreted in our notation as $F_1(X) = 1$, where $c_1 = 1$. The question is: what distribution makes the minimum number of assumptions given this information? Or, in other words, what's the distribution with the highest uncertainty that can explain our observed expected values, equations (1.5) or (1.6)? This last question rises a more fundamental question: is there a measure of the uncertainty of a probability distribution?

1.2.1 Shannon's entropy

Shannon (1948), assuming that the measure of the uncertainty of a probability distribution exists, derived it for the discrete case. This can be obtained if one makes the following assumptions:

1. This measure should be continuous in the p_i.

2. If all the p_i are equal, $p_i = 1/n$, then this measure should be a monotonic increasing function of n. With equally likely events there is more choice, or uncertainty, when there are more possible events.

3. If a choice is broken down into two succesive choices, the original measure should be the weighted sum of the individual values of this measure.

In his work, Shannon shown this measure, which he called $H(p_1, p_2, ..., p_n)$, the entropy (usually called the Shannon entropy), to be[3]

$$H(p_1, p_2, ..., p_n) = -k \sum_{i=1}^{n} p_i \ln(p_i), \tag{1.7}$$

[3] In his original paper, Shannon actually wrote this definition using the base 2 logarithm. In the present work, the natural logarithm will be used because it simplifies calculations.

where k is a positive constant that depends on the measurement units. This amazing result, as will be noted in short, can be maximized subject to the constraints given in equation (1.6) in order to find the p_i using the classical method of Lagrange multipliers: this is called the **Maximum Entropy Method** (MEM or MaxEnt).

Shannon's entropy can also be interpreted in physical therms by the entropy that is defined in statistical mechanics where, depending on the ensemble, i.e., the way in which we define the probabilites for each state, different results arise. For the microcanonical ensemble, for example, the probabilty of each state x_i is $p_i = 1/\Omega$, where Ω is the number of possible microstates of the system. Replacing this in equation (1.7), with $n = \Omega$, we get the well-known relation for the entropy $H = k\ln(\Omega)$, where we identify k as Boltzmann's constant. In fact, Jaynes (1957) shows that Shannon's entropy can be taken as the starting point in deriving the probability distributions used in statistical mechanics, as opossed to the classical method of starting from the equations of motion and additional hypotheses to finally identify the entropy of the system, by just maximizing equation (1.7) subject to certain constraints (e.g. using equation (1.6) with macroscopic measurements of physical quantities). Of course, this can be applied to an uncountable number of problems as we'll see in later sections and chapters.

We are now ready to solve the problem that we proposed: given $E[1] = \sum_{i=1}^{n} p_i = 1$, what distribution makes the minimum number of assumptions given this information? Our task is to maximize equation (1.7) given the constraint $g(p_1, p_2, ..., p_n) = E[1] = \sum_{i=1}^{n} p_i = 1$. Using the method of Lagrange multipliers, we have

$$\vec{\nabla} H = \sum_{i=1}^{n} \left(-\ln(p_i) - 1\right) \hat{x}_i = \lambda \vec{\nabla} g = \sum_{i=1}^{n} \lambda \hat{x}_i,$$

where the $\hat{x}_i$ is the orthonormal basis that defines our n-dimensional space. This reduces to the n equations $\ln(p_i) + 1 = \lambda$, which implies that $p_1 = p_2 = ... = p_n$. Replacing this in our constraint gives what we expected: $p_i = 1/n$. This is the well known *principle of insufficient reason*, which proposes that if we don't have any information available for our probabilities, we should assign equal probabilities to all our possible outcomes. It is possible to generalize this result to obtain the distribution that results if an arbitrary number of constraints are given in the form of equation (1.6). The result is (Papoulis, 1991)

$$p_i = A \exp\left(-\lambda_1 F_1(x_1) - ... - \lambda_m F_m(x_i)\right),$$

where A can be obtained from the normalization condition ($\sum p_i = 1$), which could be identified as the inverse of a function Z, the partition function, in analogy to the partition function in statistical mechanics:

$$A = \frac{1}{Z}, \text{ with } Z = \sum_{i=1}^{n} \exp\left(-\lambda_1 F_1(x_1) - ... - \lambda_m F_m(x_i)\right)$$

Here, the m constants λ_j (the Lagrange multipliers) can be obtained by replacing the p_i in the m equations (1.6), which can be written as:

$$-\frac{\partial \ln(Z)}{\partial \lambda_j} = c_j$$

Of course, posing the equations is easy but solving them usually isn't and numerical methods have to be used in order to solve for the p_i.

1.2.2 Extension to the continuous case: the differential entropy

Jaynes (1968) extended (in a rather intuitive way) Shannon's entropy to the continuous case as[4]:

$$H(X) = -\int f_X(x) \ln \frac{f_X(x)}{m(x)} dx \tag{1.8}$$

This is usually called the "differential" entropy, where $m(x)$ is a suitable function that ensures the entropy to be invariant under a change of variables. A somewhat more rigorous proof for this definition can be found on Papoulis (1991), where $m(x) = 1$ is used. We'll use this same definition for the function $m(x)$ because, as we'll see on the next chapter, we'll define a similar quantity, the negentropy, that will ensure invariance under linear transformations which is all what we really want. Note that now the entropy can be negative, because we are now measuring the uncertainty on the PDF, which is not to constrined to be less than one.

[4] Note that if we were to be extremely rigorous with our notation, we would have to write the entropy as $H(f_X(x))$, because we are measuring the entropy of the PDF. However, we write $H(X)$ for simplicity.

Analogously to the discrete case, it can be shown that the maximum entropy distribution for the continuous case is given by (Papoulis, 1991)

$$f_X(x) = A\exp\left(-\lambda_1 F_1(x) - ... - \lambda_m F_m(x)\right), \tag{1.9}$$

where, again, A can be obtained from the normalization condition ($\int f_X(x)dx = 1$) and be identified as the inverse of the partition function (which is now the integral of the exponential function)

$$\frac{1}{A} = Z = \int_\Omega \exp\left(-\lambda_1 F_1(x) - ... - \lambda_m F_m(x)\right)dx,$$

and the m constants can be found by replacing $f_X(x)$ in the m equations (1.5), which again leads to:

$$-\frac{\partial \ln(Z)}{\partial \lambda_j} = c_j$$

In the present work, we'll almost exclusively work with samples of random variables with zero mean and unit variance. We may then ask: what's the maximum entropy distribution given this information? Note that our distribution, according to equation (1.9) will be given by:

$$f_X(x) = A\exp\left(-\lambda_1 x - \lambda_2 x^2\right)$$

The first constraint of a zero mean random variable gives:

$$A\int_\Omega x\exp\left(-\lambda_1 x - \lambda_2 x^2\right)dx = -\frac{\sqrt{\pi}\lambda_1 e^{\lambda_1^2/4\lambda_2}}{2\lambda_2^{3/2}} = 0$$

Which implies $\lambda_1 = 0$ (A can't be zero). On the other hand, the constraint of unit variance ($\sigma^2 = 1$) gives:

$$A\int_\Omega x^2\exp\left(-\lambda_2 x^2\right)dx = \frac{A}{2}\sqrt{\frac{\pi}{\lambda_2^3}} = \sigma^2 \Rightarrow \lambda_2 = \left(\frac{\pi A^2}{4\sigma^2}\right)^{1/3} \tag{1.10}$$

Finally, the normalization condition gives:

$$A\int_\Omega \exp\left(-\lambda_2 x^2\right)dx = A\sqrt{\frac{\pi}{\lambda_2}} = 1 \Rightarrow \lambda_2 = \pi A^2 \tag{1.11}$$

Combining the results given in equations (1.10) and (1.11), we obtain $A = 1/\sqrt{2\pi\sigma}$ and $\lambda_2 = 1/2\sigma$. Accordingly, our maximum entropy distribution is:

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-x^2/2\sigma\right) \overset{(\sigma^2=1)}{=} \frac{1}{\sqrt{2\pi}} \exp\left(-x^2/2\right)$$

We recognize this distribution as a gaussian. This amazing result is stated as follows: from all the distributions with zero mean and unit variance, the one that makes the least number of assumptions or, as we stated earlier, the distribution that is "most uncertain" is the gaussian. This in fact is the reason why the gaussian distribution is so widely used: the gaussian distribution reflects our ignorance on the sampling distribution.

Despite this amazing result, be aware that there exists an upper limit for the maximum achievable entropy given by (Cover & Thomas, 1991)

$$sup(H) = H(X \sim N(0, \sigma^2 - \mu^2)) = \frac{1}{2}\ln(2\pi(\sigma^2 - \mu^2)e),$$

where μ is the first moment given by the constraint $F_1(X) = X$ and σ^2 the second moment given by the constraint $F_2(X) = X^2$. This may usually prevent us from obtaining the desired probability density function analitically if some functions are given as constraints in equation (1.5). The typical case is when we are given the above mentioned moments plus the third, $F_3(X) = X^3$. This constraint leads to the result that the normalization is undefined (i.e. $\int_\Omega f_X(x)dx = \infty$). However, this problem can be solved numerically, for example, by perturbing a zero mean normal random variable with variance $\sigma^2 - \mu^2$.

The idea of measuring entropies of PDFs can be further extended to random vectors (Hyvrinen, Karhunen & Oja, 2000) and we can calculate the differential entropy of this vector's PDF as

$$H(\vec{X}) = -\int_\Omega f_X(\vec{x}) \ln f_X(\vec{x}) d^m x,$$

where the maximum entropy distribution is now given by (Cover & Thomas, 1991)

$$f_X(\vec{x}) = A \exp\left(\sum_{i,j} \lambda_{ij} x_i x_j\right),$$

and the analogous constraints in the multidimensional case for equation (1.5) (for example, they could be given by measured values of the multidimensional moments derived for the multidimensional characteristic function in section 1.1.5). With this definition of the entropy for a random vector at hand and motivated by our previous results, we may then ask the following: given a zero mean vector and a covariance matrix $\mathbf{\Sigma}$, which is the distribution that maximizes the entropy? Because the distribution has an exponent with a quadratic form, it'll probably be of no surprise to know that the distribution that maximizes the entropy given those two constraints is given by:

$$f_X(\vec{x}) = \frac{1}{(2\pi)^{n/2}(\det \mathbf{\Sigma})^{1/2}} \exp\left(-\frac{1}{2}\vec{x}^T \mathbf{\Sigma}^{-1} \vec{x}\right)$$

i.e. an n-dimensional gaussian distribution.

1.3 Time series analysis

The statistical tools and definitions given in the past sections have their own counterparts when it comes to time series analysis. This is partly because in this context, the time structure gives us extra information about the random variables present in the experiment and therefore we expect some coherence in our time series, e.g., maybe past values can give us information about future values.

In order to state the standard framework of time series analysis, we first define two kind of procceses (also refered as "systems"). A **deterministic** process is any process where a realization $z(t)$ of it can be exactly predicted by a mathematical function such as

$$z(t) = ae^{bt},$$

where a and b are constants. On the other hand, we define a **non-deterministic** process to be any process where the realization $z(t)$ of it depends on a family of underlying random variables $Z(t)$, which may be mathematically described in terms of probability distributions. As can be seen from these definitions, the main difference between a deterministic and a non-deterministic process is that, given an index t, the former has fixed values for every realization of it, while the latter can give raise to different values in different realizations.

A time series that evolves in time according to probabilistic laws (i.e. not deterministic) is called a **stochastic process** (Box & Jenkins, 1976). Note that even if a realization of the non-deterministic process where dominated by a known probability distribution, the values of it cannot be exactly predicted. In contrast, however, our state of knowledge of the process is improved (e.g., we'll have better estimates for the *error* in our prediction). An example of such time-series would be a realization at M different times of a standard normal random variable[5] $Z(t) \sim N(0,1)$, $z(t)$. Knowing this makes us highly confident that the values of these realizations will probably be on the interval $[-3,3]$ (i.e., $[-3\sigma, 3\sigma]$), but we only know that with a determined degree of certainty (99.7% in this case). Such a realization for $M = 100$ is depicted in Figure 1.1.

In general, a real astrophysical time series (called "the signal", from now on) may be a sum of both, deterministic and non-deterministic processes. Because the underlying deterministic nature of the signal is based on the physical model, we'll focus the discussion on the non-deterministic part of it first. Chapter 3 will deal with the mixed model of deterministic and non-deterministic processes.

1.3.1 Characterization of stochastic processes

We can think of a stochastic process as a process Z that gives a realization of an M dimensional random vector $\vec{Z} = (Z(t_1), Z(t_2), ..., Z(t_M))$ with a joint PDF $f_Z(\vec{Z} = \vec{z})$. Furthermore, each random variable $Z(t)$ has its own PDF $f_t(Z(t) = z)$ which defines the sampling mechanism[6] of the individual realization $z(t)$. In order to characterize our stochastic process, we define the first two **moments** of the process, which are analogous to the ones defined in section 1.1.2 but with subtle differences regarding notation:

[5] Recall that a standard normal random variable is such that the probability distribution of the random variable is a gaussian function with mean $\mu = 0$ and variance $\sigma^2 = 1$. We denote this by $\sim N(\mu, \sigma^2)$.

[6] Note that it is not always true that $f(Z(t_1), Z(t_2), ..., Z(t_M)) = f_{t_1}(Z(t_1))f_{t_2}(Z(t_2))...f_{t_M}(Z(t_M))$, as explained in section 1.1.4.

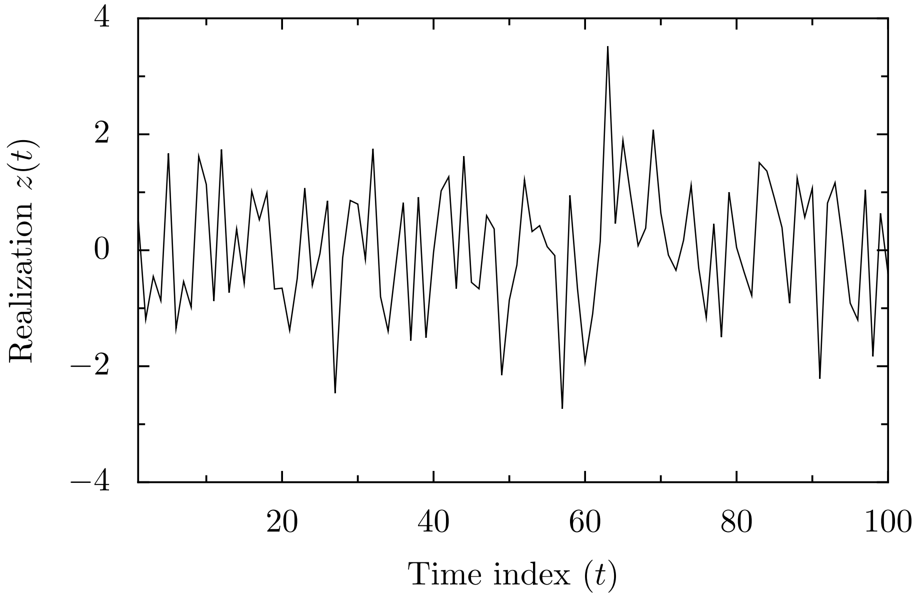

Figure 1.1: Realization $z(t)$ of a stochastic process $Z(t) \sim N(0,1)$. Although we knew by the probability distribution of the underlying random variables that the probability of exceeding the interval [-3,3] was low it does not mean it can't exceed it, as shown in the figure for $t \sim 60$.

- The **first moment (or mean)** of the process is defined as:

$$\mu(t) = E\left[Z(t)\right] = \int_{-\infty}^{\infty} z f_t(Z(t) = z) dz$$

Note that the time dependance arises because each time t may have a different random variable $Z(t)$ asociated to it and hence a different PDF $f_t(Z(t))$.

- The **second moment** is the covariance between the random variables $Z(t_n)$ and $Z(t_m)$, defined as:

$$\text{Cov}[Z(t_n), Z(t_m)] = E\left[\left(Z(t_n) - \mu(t_n)\right)\left(Z(t_m) - \mu(t_m)\right)\right]$$

Note that the special case $t = t_n = t_m$ defines the variance σ_t^2 of the random variable $Z(t)$:

$$\text{Cov}[Z(t), Z(t)] = E\left[(Z(t) - \mu(t))^2\right] = \sigma_t^2$$

These moments where not defined arbitrarly. If the above defined moments change or not with time (and how they change with it) will lead us to the concepts of stationarity and non-stationarity of stochastic processes used in current literature and particularly in this thesis. A **stationary process** is a special class of stochastic process which assumes that the underlying probabilistic nature of a realization $z(t)$ of it i.e., the distribution of the random variables $Z(t)$, do not change in time (i.e., the statistical properties of the process are independent of time). This in fact defines the concept of **strict stationarity** (Brockwell & Davis, 2001), which states that the marginal PDF of any subset of elements of the random vector $\vec{Z}$ is time-independent. This is stated as:

$$(Z(1), ..., Z(n)) \overset{P}{=} (Z(1+h), ..., Z(n+h))$$

Where $h \in \mathbb{N}_0$, $n \in \mathbb{N}$ and $\overset{P}{=}$ means an equality in the probabilistic sense (i.e. an equality between the joint distributions). This is an important theoretical definition but a rather impractical one because, according to our study of the multi-dimensional characteristic function in subsection 1.1.5, in order to prove and/or assume strict stationarity we would have to prove and/or assume that every cummulant of a given joint PDF is equal to the cummulant of the time-lagged version of it. On the other hand, strict stationarity seems to be a very limiting assumption when we think in physical processes. Because of these reasons in the present work, and unless otherwise stated, when we refer to stationary processes we'll use the definition of **wide sense or up to order two stationarity (WSS)**, which is defined below.

Definition

A process is called ***wide sense or up to order two stationary (WSS)*** *if:*

1. *The first moment is time-invariant.*

$$\mu_Z \equiv \mu(t) = E\left[Z(t)\right] = \int_{-\infty}^{\infty} z f_t(Z(t) = z) dz \ \forall \ t$$

2. *The second moment (covariance) is a function of the lag* $\tau = t_k - t_l$ *only,* $\forall \ k, l \in [0, 1, 2...M]$.

$$Cov[Z(t_k), Z(t_l)] \equiv R_Z(\tau)$$

   The function $R_Z(\tau)$ *is called the* ***autocovariance*** *of the process.*

Note that, by the above definition, the variance of the process is also a constant and is given by the autocovariance function evaluated at lag $\tau = 0$, $R(0) = \sigma_z^2$. Also note that, by the definition of this function, it follows that $R(\tau) = R(-\tau)$.

As a final note on the characterization of stationary stochastic procceses, note that stationarity does not mean that this kind of processes do not have some dependance on past values. Because the autocovariance function can take different values for differents lags τ, the process has some kind of "memory" of past values. If the autocovariance fades out quickly as the lag τ increases the process is called a "short memory" process, because only random variables that are close together are actually correlated. On the other hand, "long memory" processes can have a slowly decaying autocovariance function, meaning that the process actually has an important number of correlated random variables.

As an example, Figure 1.2 shows a realization $z(t)$ of a WSS stochastic process known as "flicker noise", which shows a clear time-dependance. This kind of noise is widely observed not only on astrophysics but on a wide variety of areas (Press, 1978) and will be treated implicitly in the present work. A strictly stationary process, on the other hand, could be the process shown in Figure 1.1, where the underlying random variable is $Z(t) \sim N(0, 1)$. Finally, an example of a non-stationary process would be, for example, $Z(t) \sim N(0, e^{-t/1000})$.

1.3.2 Power Spectral Density and Autocovariance Matrix

The discussion at the end of the past subsection motivates the study of the time structure of the autocovariance function. In particular, the ($\sqrt{2\pi}$ multiplied) Fourier Transform of the autocovariance is defined as the Power Spectral Density (PSD) $S_Z(f)$ of the process:

$$S_Z(f) = \int_{-\infty}^{\infty} R_Z(\tau) e^{-i2\pi f \tau} d\tau \tag{1.12}$$

Note that, for a general stochastic process, $S_Z(f)$ may not even exist. However, for a WSS process, it is ensured to exist and this relation is called the Wiener-Khinchin theorem for WSS processes. The reason why it is called the "Power Spectral Density" comes from the derivation of $S_Z(f)$. It is easy to show that, given the

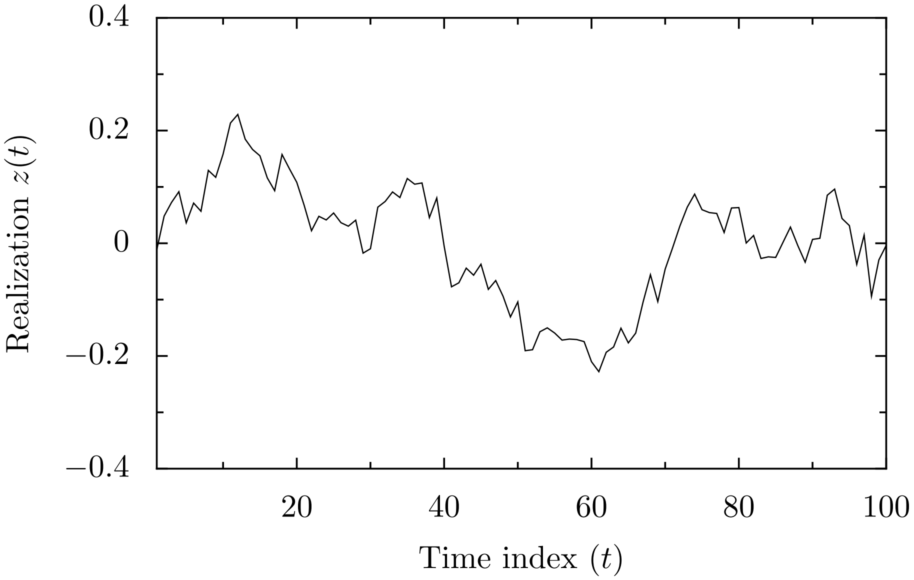

Figure 1.2: $1/f$ noise or "flicker" noise is a stationary process whose power spectrum (defined in section 1.3.2) is $S_Z(f) \propto 1/f^{\gamma}$, where $\gamma = 2H + 1 > 0$ and H is called the Hurst parameter. For this simulation, we have used $H = 0.5$.

($\sqrt{2\pi}$ multiplied) Fourier Transform $\hat{Z}_T(f)$ of the random variable truncated to[7] $0 \leq |t| \leq T/2$, $Z_T(t)$, the PSD can be defined as:

$$S_Z(f) = \lim_{T\to\infty} \frac{1}{T} E[|\hat{Z}_T(f)|^2]$$

In practice, the PSD is always limited to the sample-to-sample distance. If we denote this distance by Δt, then the minimum frecuency that we can achieve to see in the PSD, also called the Nyquist or critical frequency (Brault & White, 1971), is $0.5/\Delta t$. For simplicity, we'll normalize the sampling distance to 1, i.e., $\Delta t = 1$. With this in mind, the lag τ becomes $\tau = t_k - t_l = k - l$ and the integral relation of equation (1.12) becomes discrete:

$$S_Z(f) = \sum_{\tau=-\infty}^{\infty} R_Z(\tau) e^{-i2\pi f \tau} \tag{1.13}$$

Although this expresion is not always ensured to exist, it can always be found an analogous representation in terms of generalized distributions. However, these cases

[7] The reason why the random variable needs to be truncated is that, in general, the fourier transform of $Z(t)$ is not ensured to exist.

will not be treated in this work and we refer the interested reader to the discussion on Spectral Analysis of time series presented by Brockwell & Davis (2001).

The power spectral density can be thought as a measure of the frequency range of the process and is another way of looking at the concept of the "memory" of the process, which was introduced at the end of the past subsection. Large values of the PSD for the higher frequencies means that the process has short-memory correlations (the process "remembers" recent values only), while large values of the PSD for lower frequencies mean long-memory correlations (the process "remembers" values that are far in the past relative to the process time-scale). This interpretation motivates the definition of the "colors" of the signals, in analogy to spectroscopic measurements of light. Signals with large values of the PSD in the lower frequencies are called "red noise" signals, while signals with large values of the PSD on the upper frequency range are called "blue noise" signals. Of these "colored signals", the most widely known and used is "white noise", which is derived if one considers a zero-mean, uncorrelated stochastic proccess with variance σ_W^2, $W(t)$ (note that with these considerations, this process is automatically WSS)

$$\begin{aligned} E[W(t)] &= 0, \\ E[W(t_1)W(t_2)] &= R_W(\tau) = \sigma_W^2 \delta(\tau), \end{aligned}$$

where $\tau \geq 0$ and $\delta(\tau)$ stands for Kronecker's delta. According to the (discrete) definition of the PSD, equation (1.13), we have, noting that $R(\tau) = R(-\tau)$:

$$S_W(f) = \sum_{\tau=-\infty}^{\infty} R_W(\tau) e^{-i2\pi f \tau} = \sigma_W^2$$

Hence, the signal has a PSD that is a constant over frequency and, in analogy to white light which in theory has this same spectral form, it is called "white noise". A special case of white noise is that of a stochastic process where each random variable is distributed as a normal one with equal variance, as the one shown in Figure 1.1.

The time structure of the autocovariance function can be expressed in a very compact form using the well known covariance matrix. Given our normalization of the sampling distance, as noted earlier the autocorrelation function is now a function of $\tau = t_k - t_l = k - l$, and the covariance matrix can now be expressed as:

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_Z^2 & R_Z(1) & R_Z(2) & \dots & R_Z(M-1) \\ R_Z(1) & \sigma_Z^2 & R_Z(1) & \dots & R_Z(M-2) \\ R_Z(2) & R_Z(1) & \sigma_Z^2 & \dots & R_Z(M-3) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ R_Z(M-1) & R_Z(M-2) & R_Z(M-3) & \dots & \sigma_Z^2 \end{bmatrix}$$

This form of the covariance matrix where the elements in every diagonal are constants is called the **autocovariance matrix**. Analogously to the autocovariance, the **autocorrelation** is defined as:

$$\rho(\tau) = \frac{R(\tau)}{\sigma_Z^2}$$

(note that $\rho(0) = 1$, as expected). The corresponding **autocorrelation matrix** is defined as:

$$\mathbf{P} = \frac{1}{\sigma_Z^2}\mathbf{\Sigma}$$

Note that the autocorrelation and autocovariance matrices also have the important properties of covariance matrices, that is, symmetry and positive semidefiniteness.

Now we posess all the necesary tools to understand some stochastic models. In general, we'll want to model a given stochastic process. However, because the nature of these processes is probabilistic, the model also needs to be probabilistic. In the next subsections we treat one of the most widely used stochastics models that will become useful in our analysis of time series.

1.4 Stochastic modeling

In order to understand the nature of stochastic signals, we first have to find ways to model them. However, models for stochastic procceses can be hard to realize given the probabilistic nature of them and, in particular, chosing a model for our data is not a strightforward task: we need to choose a model that can take everything we know about the process and nothing more. Because of this, we need to expand the idea of entropy introduced in section 1.2 to stochastic processes, so we can be sure that the model we are chosing is one that makes the minimum number of assumptions using only the information at hand. The problem is that the previously defined entropy doesn't take into account the fact that our PDF may change with time and, therefore, our entropy may vary with time as well. This problem leads us to define an analogous entropy measure for stochastic processes.

1.4.1 Differential entropy rate

We've already defined the (differential) entropy for a random vector $\vec{X}$ in section 1.2. This was a measure of the uncertainty of the PDF $f_X(\vec{X})$ that defines our random vector. However, as we mentioned earlier, in a stochastic process and in general physical procceses, the fact is that the new random variables that appear with time may be perturbed by previous values. Given this fact, we'd like to know how the entropy of this (increasing) random vector grows with time and study it asymptotic behavior. This measure for a stochastic process is given by the **differential entropy rate**, defined as

$$h = \lim_{t \to \infty} \frac{H\left(Z(1), Z(2), ..., Z(t)\right)}{t}, \tag{1.14}$$

if the limit exists. Note that **this is not** the maximum possible absolute value of the rate of variation of the differential entropy: this is just the asymptotic value of this rate. For example, if the PDF of our process $f_Z(\vec{z})$ is composed of independent and identically distributed (i.i.d.) random variables, i.e. $f_Z(\vec{z}) = f_1(Z(1) = z_1)...f_t(Z(t) = z_t)$, then the differential entropy is

$$H\left(Z(1), Z(2), ..., Z(t)\right) = \int_\Omega f_Z(\vec{z}) \ln(f_Z(\vec{z})) d^t z = tH_Z,$$

where H_Z is a constant and we have used the fact that the random variables are identically distributed, i.e., the entropies of the PDFs of each random variable are equal to this constant ($H(Z(1)) = ... = H(Z(t)) = H_Z$). With this result at hand, the differential entropy rate, equation (1.14), is:

$$h = \lim_{t \to \infty} \frac{tH_Z}{t} = H_Z$$

This result makes a lot of sense. Being the entropy a measure of how uncertain is the PDF of our process ($f_Z(\vec{z})$), it obviously has to increase if the random variables are i.i.d. However, the asymptotic rate of change of the entropy is constant and equal to H_Z, as the differential entropy of the PDF suggested.

The definition above makes clear the usefulness of the entropy rate of a stochastic process. For a stochastic process with a series of constraints, we can find the model that maximizes the entropy rate: this will be the most uncertain of all possible processes or, in other words, the one that makes the minimum number of assumptions, in analogy to the maximum entropy method presented on Section 1.2. This is because the stochastic process that maximizes the entropy rate will maximize the asymptotic rate of change of the entropy, leading to an overall maximum entropy.

1.4.2 A maximum entropy rate process: the AR Model

If the stochastic process that maximizes the entropy rate is the most uncertain we may ask: given a covariance matrix, what's the most uncertain stochastic process that we can form? Choi & Cover (1984) showed that, a process with maximum entropy rate given a covariance matrix is the linear stationary **Autoregressive, AR(p) process**, defined as

$$\tilde{Z}(t) = \sum_{k=1}^{p} \alpha_k \tilde{Z}(t-k) + W(t), \tag{1.15}$$

where $\tilde{Z}(t)$ are zero-mean random variables and $W(t)$ is white noise of constant variance σ_W^2. Here the α_k, $k = 1, 2, ..., p$ are the parameters that define the regression, where in order to ensure the stationarity of the process the characteristic equation

$$F(y) = y^p - \alpha_1 y^{p-1} - ... - \alpha_p$$

must have roots outside the unit circle, i.e., $F(y) \neq 0$ for $|y| \leq 1$ (Box & Jenkins, 1976; Brockwell & Davis, 2001). The process is called autoregressive because, in analogy to a linear regression, this one is regressed with itself (i.e. past values are regressed to estimate future ones). Multiplying equation (1.15) by $\tilde{Z}(t+\tau)$, for $\tau \geq 0$, and taking expected values at both sides, it is easy to show that the autocovariance function of this process is given by:

$$E[\tilde{Z}(t)\tilde{Z}(t+\tau)] = R_{\tilde{Z}}(\tau) = \sum_{k=1}^{p} \alpha_k R_{\tilde{Z}}(\tau - k) + \sigma_W^2 \delta(\tau) \tag{1.16}$$

And the autocorrelation function of the process is given by:

$$\rho_{\tilde{Z}}(\tau) = \sum_{k=1}^{p} \alpha_k \rho_{\tilde{Z}}(\tau - k) + \frac{\sigma_W^2}{\sigma_{\tilde{Z}}^2}\delta(\tau) \tag{1.17}$$

Collecting these equations for $\tau = 1, 2...p$, we form the famous **Yule-Walker** equations, which serve to characterize a given AR process. This system of equations can be written in matrix form as:

$$\begin{bmatrix} 1 & \rho_{\tilde{Z}}(1) & \rho_{\tilde{Z}}(2) & \dots & \rho_{\tilde{Z}}(p-1) \\ \rho_{\tilde{Z}}(1) & 1 & \rho_{\tilde{Z}}(1) & \dots & \rho_{\tilde{Z}}(p-2) \\ \rho_{\tilde{Z}}(2) & \rho_{\tilde{Z}}(1) & 1 & \dots & \rho_{\tilde{Z}}(p-3) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_{\tilde{Z}}(p-1) & \rho_{\tilde{Z}}(p-2) & \rho_{\tilde{Z}}(p-3) & \dots & 1 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \vdots \\ \alpha_p \end{bmatrix} = \begin{bmatrix} \rho_{\tilde{Z}}(1) \\ \rho_{\tilde{Z}}(2) \\ \rho_{\tilde{Z}(3)} \\ \vdots \\ \rho_{\tilde{Z}}(p) \end{bmatrix}$$

If one obtain estimates of the autocorrelation functions at each lag τ, then one has a set of p linear equations and can therefore solve for the α_k. Finally, given the autocovariance function, equation (1.16), we can also derive the PSD of the AR process. For an autoregressive model of order p, the PSD is given by (Brockwell & Davis, 2001):

$$S_{\tilde{Z}}(f) = \frac{\sigma_W^2}{|1 - \sum_{k=1}^{p} \alpha_k e^{-i2\pi f k}|^2}$$

As a final note on the properties of AR processes, we'd like to note that they are not the only processes which have maximum entropy rate. Boshnakov & Iqelan (2010) showed that, in fact, there exist many maximum entropy processes. Furthermore, they show that these processes need not to be gaussian-driven nor stationary. However, these processes are out of the scope of the present thesis for reasons of extension and we'll limit the discussion and analysis to stationary time series.

1.4.3 An example of an AR process

As an example, and in order to discuss some of the considerations to be taken when sampling stochastic procceses (i.e. when obtaining a realization) consider the first order AR(1) process

$$\tilde{Z}(t) = \alpha_1 \tilde{Z}(t-1) + W(t),$$

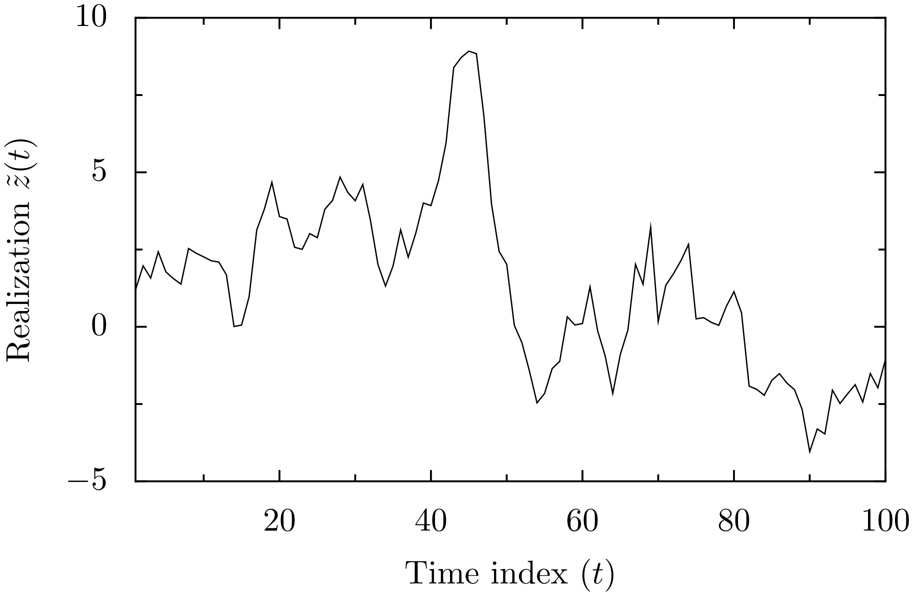

Figure 1.3: Realization of an AR(1) process $\tilde{Z}(t) = 0.9\tilde{Z}(t-1) + N(t)$, where $W(t) \sim N(0, \sigma_W^2 = 1)$.

with $\alpha_1 = 0.9$ and $W(t) \sim N(0,1)$. Note that this process is stationary but is very close to the limit of non-stationarity (i.e. $\alpha_1 \geq 1$). A realization $\tilde{z}(t)$ of this process is plotted in Figure 1.3. The autocovariance of this process can be derived by noting that:

$$R_{\tilde{Z}}(\tau) = \alpha_1 R_{\tilde{Z}}(\tau - 1) = \alpha_1(\alpha_1 R_{\tilde{Z}}(\tau - 2)) = ... = \alpha_1^\tau R_{\tilde{Z}}(0)$$

Because $R(0) = \alpha_1 R_{\tilde{Z}}(1) + \sigma_W^2$, and by the above result $R(1) = \alpha_1 R(0)$, it follows that $R(0) = \sigma_{\tilde{Z}}^2 = \sigma_W^2/(1-\alpha_1^2)$. Finally, then, the autocovariance is given by:

$$R_{\tilde{Z}}(\tau) = \alpha_1^\tau \sigma_{\tilde{Z}}^2 = \alpha_1^\tau \frac{\sigma_W^2}{1-\alpha_1^2}$$

Note that because $\alpha_1 < 1$, the autocovariance decays exponentially. In time series analysis, however, it is customary to analyse the autocorrelation function rather than the autocovariance which, according to equation (1.17) is given, for $\tau > 0$, by:

$$\rho_{\tilde{Z}}(\tau) = \alpha_1^\tau$$

But, how can we estimate this value from our data? As we'll see, in order to obtain estimations of the autocorrelation function given our realization of the process, some considerations must be taken into account. The next subsection deals with estimators for the autocorrelation and autocovariance of stationary procceses.

1.4.4 Estimating autocorrelations and autocovariances

Let $\hat{\rho}(\tau)$ denote the estimator of the autocorrelation function and $\hat{R}(\tau)$ the estimator for the autocovariance function. In order to obtain an estimate for the autocorrelation we need to find an estimate for the autocovariance, because $\hat{\rho}(\tau) = \hat{R}(\tau)/\hat{R}(0)$. From the definition of the autocovariance, recall that

$$R(\tau) = E[(Z(t) - \mu_Z)(Z(t+\tau) - \mu_Z)],$$

where $\mu_Z = E[Z(t)]$ is the mean of the process. In order to estimate this expected value, let's first consider that the mean value is known. With this in mind, consider the statistic[8]

$$S'(\tau) = \frac{1}{M-\tau}\sum_{t=1}^{M-\tau}(Z(t) - \mu_Z)(Z(t+\tau) - \mu_Z),$$

where M is the number of random variables present on our process and $\tau > 0$. Taking the expected value of this statistic, we get

$$\begin{aligned} E[S'(\tau)] &= \frac{1}{M-\tau}\sum_{t=1}^{M-\tau} E[(Z(t) - \mu_Z)(Z(t+\tau) - \mu_Z)] \\ &= \frac{1}{M-\tau}\sum_{t=1}^{M-\tau} R(\tau) \\ &= \frac{M-\tau}{M-\tau} R(\tau) \\ &= R(\tau), \end{aligned}$$

and, therefore, $S'(\tau)$ is an unbiased[9] estimator of the autocovariance. However, a sometimes overlooked problem is that, in practice, μ_Z is unknown and has to be estimated from the sample mean statistic $\bar{Z} = \frac{1}{M}\sum_{t=1}^{M} Z(t)$. With this in mind, we

[8] Recall that a statistic is defined as "any function of the observed random variables in a sample such that the function does not contain any unkown quantities" (Gregory, 2005).

[9] A biased estimator $\hat{\theta}$ is defined as an estimator of the parameter θ for which $E[\theta - \hat{\theta}] = B$, where B is called **the bias** of the estimator. If the bias is 0, then the estimator is unbiased

may consider the statistic

$$S^{\dagger}(\tau) = \frac{1}{M-\tau}\sum_{t=1}^{M-\tau}(Z(t)-\bar{Z})(Z(t+\tau)-\bar{Z})$$

as an estimator of the autocovariance. However, it can be seen that this estimator is biased. Taking expected values at both sides we get

$$\begin{aligned}
E[S^{\dagger}(\tau)] &= \frac{1}{M-\tau}\sum_{t=1}^{M-\tau} E[(Z(t)-\bar{Z})(Z(t+\tau)-\bar{Z})] \\
&= \frac{1}{M-\tau}\sum_{t=1}^{M-\tau} E\left[\big((Z(t)-\mu_Z)-(\bar{Z}-\mu_Z)\big)\big((Z(t+\tau)-\mu_Z)-(\bar{Z}-\mu_Z)\big)\right] \\
&= \frac{1}{M-\tau}\sum_{t=1}^{M-\tau} R(\tau) - \frac{1}{M}\left[\sum_{j=1}^{M}(R(t-j)+R(t+\tau-j)) - \frac{1}{M}\sum_{j,k=1}^{M} R(j-k)\right] \\
&= R(\tau) - B_{S^{\dagger}}(\tau),
\end{aligned}$$

where

$$B_{S^{\dagger}}(\tau) = \frac{1}{M-\tau}\sum_{t=1}^{M-\tau}\frac{1}{M}\left[\sum_{j=1}^{M}(R(t-j)+R(t+\tau-j)) - \frac{1}{M}\sum_{j,k=1}^{M} R(j-k)\right]$$

is the (negative) bias of the estimator. Note, however, that as $M \to \infty$, $B(\tau) \to 0$ as long as the series for the autocovariance converges (which is ensured for stationary processes). As noted by various authors (e.g. Box & Jenkins, 1976; Fuller, 1996; Percival, 1993; Brockwell & Davis, 2001), with $S^{\dagger}(\tau)$ as an estimator of the autocovariance, the autocovariance matrix is no longer positive definite, which as we'll see is of fundamental importance. Therefore, a classical biased but recomended estimator for the autocovariance is given by

$$\hat{S}(\tau) = \frac{1}{M}\sum_{t=1}^{M-\tau}(Z(t)-\bar{Z})(Z(t+\tau)-\bar{Z})$$

If this statistic is used, then the positive definiteness of the sample covariance matrix is guaranteed. Furthermore, in a similar way as for the $S^{\dagger}(\tau)$ statistic, the bias of the $\hat{S}(\tau)$ estimator can be shown to be:

$$B_{\hat{S}}(\tau) = R(\tau)\frac{\tau}{M} + \frac{1}{M}\sum_{t=1}^{M-\tau}\frac{1}{M}\left[\sum_{j=1}^{M}(R(t-j)+R(t+\tau-j)) - \frac{1}{M}\sum_{j,k=1}^{M}R(j-k)\right]$$

As noted by Percival (1993), the estimator $S^{\dagger}(\tau)$ can be in fact "more biased" than the estimator $\hat{S}(\tau)$ for cases of interest. For example, if we consider a white noise process ($R(\tau) = 0 \ \forall \ \tau \neq 0$), it is easy to see that the biases are given by

$$B_{S^{\dagger}}(\tau) = \frac{\sigma_W^2}{M} \text{ and } B_{\hat{S}}(\tau) = \frac{\sigma_W^2}{M}\left(1 - \frac{\tau}{M}\right)$$

For $\tau > 0$, then:

$$\frac{B_{\hat{S}}(\tau)}{B_{S^{\dagger}}(\tau)} = \left(1 - \frac{\tau}{M}\right) < 1 \Longrightarrow B_{\hat{S}}(\tau) < B_{S^{\dagger}}(\tau)$$

For all the above given reasons, this will be the estimator used for the autocovariance in the present thesis. Given this estimator and a realization $z(t)$ of the process, then, the sample autocovariance can be calculated as

$$\hat{R}(\tau) = \frac{1}{M}\sum_{t=1}^{M-\tau}(z(t) - \hat{\mu}_Z)(z(t+\tau) - \hat{\mu}_Z),$$

where $z(t)$ is the t-th sample and $\hat{\mu}_Z$ is the estimated mean via the sample mean statistic. Accordingly, the sample autocorrelation function can be calculated as:

$$\hat{\rho}(\tau) = \frac{\hat{R}(\tau)}{\hat{R}(0)}$$

The problem now resides in estimating a measure of the standard error of our estimator. Note that to obtain this, we have to obtain $\text{Var}[\hat{S}(\tau)]$ but, in the general case, there's no explicit analytical solution. The classical aproximation for the standard error on the autocorrelation function is given by Bartlett's aproximation, $\hat{\sigma}_\rho = 1/\sqrt{M}$ (Box & Jenkins, 1976). However, this aproximation is (1) only valid for linear time series (which is the case of AR procceses) and (2) they usually only serve as a test for zero autocorrelation for the first τ terms of $\hat{\rho}(\tau)$. This is a problem for us because what we want to do is to actually see the shape of the autocorrelation function by

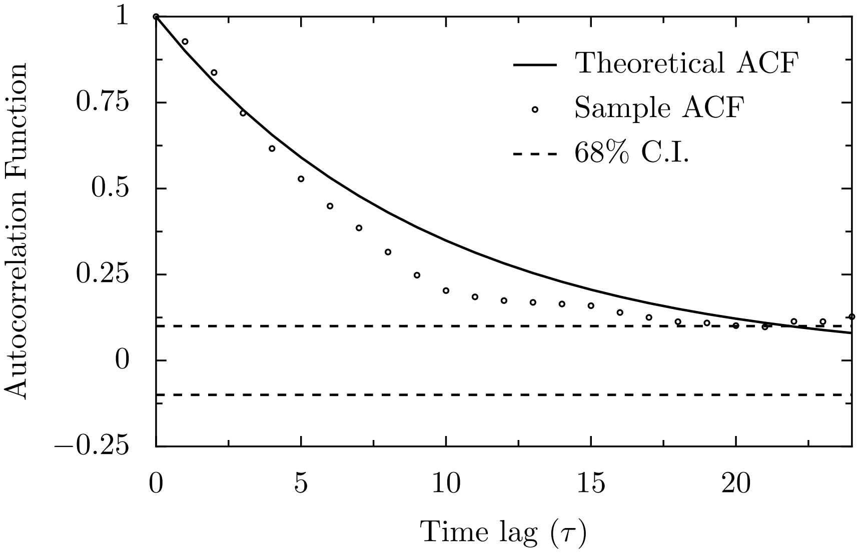

Figure 1.4: Theoretical (thick line) and sample (dots) autocorrelation function for the AR(1) process. The confidence intervals (dashed lines) where obtained using Bartlett's approximation $1/\sqrt{M}$ as the standard error (68% confidence interval).

assuming as little as possible about the underlying model: we actually want to see if the models that we propose for the data apply. Figure 1.4 shows the sample autocorrelations found for our realization of the AR(1) process with $\alpha_1 = 0.9$ along with the theoretical autocorrelation function using this standard measure for the standard error. Note that the sample autocorrelation is clearly biased, and the standard error doesn't really help in characterizing the process. Because of these problems, it is better to use a relative measure of autocorrelation which will be introduced in the next subsection: the partial autocorrelation.

1.4.5 Finding the order of an AR process

We end this section solving one of the problems that we may encounter in the analysis of time series: given a realization $z(t)$ of an AR process, how can we find the order (p) of the process? One first attempt would be to look at the autocovariance or autocorrelation functions of the process and see if it follows the expected "shape" of an AR process. However, this may not be a good idea since high order AR processes have propagated dependancies (i.e. the value of these functions at a time t depends on the value at a time $t-1$, which in turn depends on the value at a time $t-2$, etc.). Furthermore, as we saw one the past subsection, the sample autocorrelation

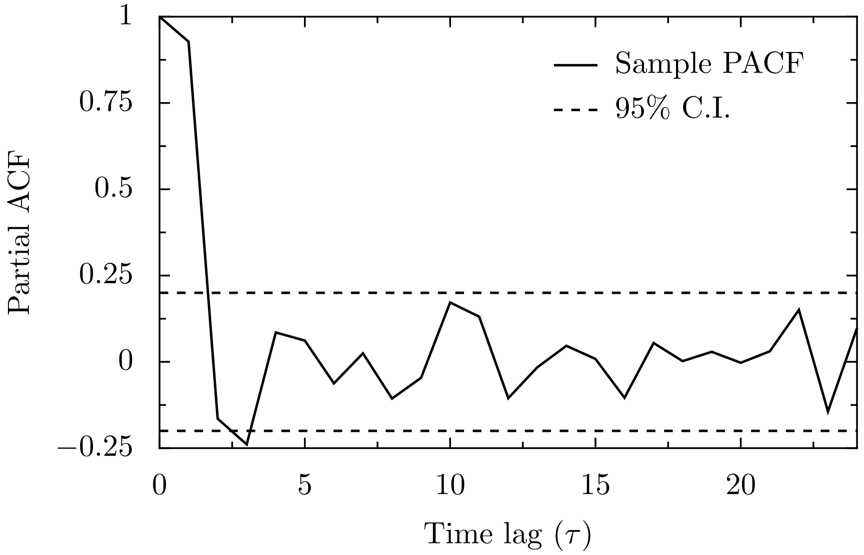

Figure 1.5: Sample partial autocorrelation function (PACF) for the AR(1) process using $1/\sqrt{M}$ as the standard error, which was multiplied by 2 to obtain the 95% confidence interval (dashed lines). The PACF is efectively zero for lags $\tau > 1$, as expected.

is an interesting tool for qualitative analysis, but a poor quantitative one because it is biased.

The classic solution to the problem is to use the **partial autocorrelation function** (Box & Jenkins, 1976; Brockwell & Davis, 2001), which aims at the obtention of the autocorrelation without the mentioned propagated dependencies, i.e., the correlation between two random variables given that the propagated correlation with the other random variables has been substracted. The partial autocorrelation $\tilde{\rho}(\tau)$ at lag τ can be shown to be equal to the last parameter α_τ obtained when fitting an AR(τ) process to a given realization. This means that for an AR process the importance is double: the partial autocorrelation will not only help us find the "true" autocorrelation between a time-lagged pair of random variables, but will also help us to find the order of our AR process. This is because for an AR process of order p, $\tilde{\rho}(\tau) = 0 \; \forall \; \tau > p$.

The classical way of "fitting" an AR model and, therefore, finding the partial autocorrelation of the process is to solve the Yule-Walker equations and let $\tilde{\rho}(\tau) = \hat{\alpha}_\tau$, where $\hat{\alpha}_\tau$ is the estimated τ-th parameter. If the underlying model is AR, it can be shown that for large samples this coefficient has a distribution of zero mean and variance equal to $1/M$, thus providing a standard error of $1/\sqrt{M}$ for it. We performed this analysis to our realization of the AR(1) process, and found $\tilde{\rho}(1) = 0.927$, which is actually what we expected. A plot of this partial autocorrelation function along with the 95% confidence interval is shown in Figure 1.5. Note, however, that the standard error assumed here is $1/\sqrt{M} = 0.01$ (where we have used $M = 100$) and, therefore, the obtained value for $\tilde{\rho}(1)$ fits at the very tail of the 95% confidence interval. This is a clear sign of a bias (which was expected because the autocorrelation itself is biased), which has been solved in the literature under some assumptions (see, e.g., Shaman & Stime, 1988). However, in our applications this bias is small in comparison with other effects that will be discussed in future chapters, and therefore we'll not correct for it.

Chapter 2

Blind Signal Separation

Blind Signal Separation (BSS) is a now widely used technique for the identification and separation of mixed signals. It is called blind because usually we don't have much information about the forms in which these signals are actually mixed, although we usually might identify and interpret the different signals if they could be separated.

In this chapter we present two techniques that will be used in the present work: Principal Component Analysis, which is a way to decorrelate linearly a set of random variables and Independent Component Analysis, which is based in the assumption that the signals that where mixed are not only uncorrelated, but also independent. As we will see, both techniques have a wide range of application in time-series data, but the latter can be viewed as an extension of the former.

2.1 Principal Component Analysis

Principal Component Analysis (PCA) is a classical multivariate analysis method that has two basic ideas. The first one is to find a linear transformation $\mathbf{V}$ that transform a given zero-mean $N-$dimensional random vector[1] $\vec{X}$, which may have correlated random variables, into a new $N-$dimensional random vector $\mathbf{V}\vec{X} = \vec{Z}$ that has uncorrelated random variables. On the other hand, the second idea is to maximize the variance of each linear transformation of the elements of $\vec{X}$, i.e.,

[1] This asumption is made in order to simplify the notation. However, in practice we can always take a sample of this random vector and substract the empirical mean in order to simplify the problem.

maximize the variance of each element of the random vector $\vec{Z}$, which are called the principal components. As we will see, the linear transformation that projects the random vector $\vec{X}$ to an uncorrelated random vector $\vec{Z}$ is not unique, and we'll search for the optimal one in the context of time series analysis.

2.1.1 A derivation of the principal components

The question now is: how do we find this transformation? The classic derivation (Jolliffe, 2002) uses the fact that we want to first maximize the variance of each linear combination of $\vec{X}$, i.e., maximize the variance of the principal components, obtaining the desired transformation componentwise. The idea is that the desired transformation matrix, $\mathbf{V}$, contains the coefficients of these linear transformations. Let's denote the elements of this matrix by $v_{i,j}$. Then, the i-th principal component (the i-th element of the vector $\vec{Z}$) is given by:

$$Z_i = v_{1,i}X_1 + v_{2,i}X_2 + ... + v_{N,i}X_N = \vec{v}_i^T\vec{X}$$

Where the elements of the vectors $\vec{v}_i$ are the coefficients of the linear combination of the elements of $\vec{X}$ corresponding to this i-th principal component. Our first task is to maximize the variance of the first principal component, $E\left[Z_1^2\right] = \vec{v}_1^T\mathbf{\Sigma}_X\vec{v}_1$. Note that, however, we need to constrain the values of the vectors $\vec{v}_i$ in order to do this. The constraint that we'll impose is $\vec{v}_i^T\vec{v}_i = 1$. In summary, the problem is stated as follows: maximize the function $f(\vec{v}_1) = \vec{v}_1^T\mathbf{\Sigma}_X\vec{v}_1$ given the constraint $g(\vec{v}_1) = \vec{v}_i^T\vec{v}_i = 1$. Using the method of Lagrange multipliers (and remembering that the covariance matrix is symmetric), we have:

$$\vec{\nabla}_1 f = 2\mathbf{\Sigma}_X\vec{v}_1 = \lambda_1\vec{\nabla}_1 g = 2\lambda\vec{v}_1 \Longrightarrow \mathbf{\Sigma}_X\vec{v}_1 = \lambda\vec{v}_1$$

Where the operator $\vec{\nabla}_1$ represents the gradient with respect to the elements of $\vec{v}_1$. Here we see that $\vec{v}_1$ is an eigenvector of the covariance matrix of $\vec{X}$, where the corresponding eigenvalue is the Lagrange multiplier, λ_1. To obtain the second principal component, we repeat the maximization problem that we made for the first one but now we add one more constraint: we want Z_1 and Z_2 to be uncorrelated, i.e. $h(\vec{v}_1, \vec{v}_2) = \mathrm{Cov}(Z_1, Z_2) = E[Z_1^T Z_2] = \vec{v}_1^T\vec{v}_2\lambda_1 = 0$. In other words, the vectors of coefficients are orthogonal. Using again the method of Lagrange multipliers, but now with two constraints (normality of $\vec{v}_1$ and orthogonality between $\vec{v}_1$ and $\vec{v}_2$) and, therefore, two multipliers λ_2 and λ_3 we have:

$$\vec{\nabla}_2 f = 2\mathbf{\Sigma}_X \vec{v}_2 = \lambda_2 \vec{\nabla}_2 g + \lambda_3 \vec{\nabla}_2 h = 2\lambda_2 \vec{v}_2 + \lambda_3 \vec{v}_1$$

Taking the dot product with respect to $\vec{v}_1^T$ from the left, we can see that $\lambda_3 = 0$. This implies that:

$$\mathbf{\Sigma}_X \vec{v}_2 = \lambda_2 \vec{v}_2$$

And, again, $\vec{v}_2$ is an eigenvector of $\mathbf{\Sigma}_X$, where λ_2 is the corresponding eigenvalue. We can repeat this process N times to find that the i-th coefficient vector is given by the i-th eigenvector of the covariance matrix $\mathbf{\Sigma}_X$. Because of this, one way to obtain the desired linear transformation $\mathbf{V}$ that takes the vector $\vec{X}$ and transforms it to the new random vector $\vec{Z}$ of uncorrelated random variables can be obtained by finding the eigenvectors of the covariance matrix of $\vec{X}$, and letting each eigenvector be a row of $\mathbf{V}$ (note that, because of this, $\mathbf{V}$ is orthogonal). This can be efficiently done for any sample covariance matrix via a Singular Value Decomposition (SVD) algorithm, which, for our case (real signals) will lead to the SVD decomposition $\mathbf{\Sigma}_X = \mathbf{EDE}^T$, where $\mathbf{E}$ is a matrix that contains the eigenvectors in the columns and $\mathbf{D}$ is a diagonal matrix with the corresponding eigenvalues. Note that the transformation that we found, $\mathbf{V} = \mathbf{E}^T$, gives the following covariance matrix for $\vec{Z}$:

$$E\left[\vec{Z}\vec{Z}^T\right] = \mathbf{E}^T E\left[\vec{X}\vec{X}^T\right]\mathbf{E} = \mathbf{E}^T\mathbf{EDE}^T\mathbf{E} = \mathbf{D} \tag{2.1}$$

2.1.2 Interpretation of the Principal Components

Perhaps the most important feature of the principal components is that their respective eigenvalues give information about which principal component has the largest variance. The higher the eigenvalue, the higher the variance. This can be observed from equation (2.1), where for a given principal component, the variance is given by:

$$E\left[Z_i Z_i^T\right] = E\left[\vec{v}_i^T \vec{X}\vec{X}^T \vec{v}_i\right] = \vec{v}_i^T \mathbf{\Sigma_X} \vec{v}_i = \lambda_i$$

The importance of this is given because the direction of the vector of coefficients that define each principal component, $\vec{v}_i$, define directions of maximum dispersion.

Consider, for example, the random variables $X \sim N(0,1)$, $Z \sim N(0,1/2)$ and $Y = X + Z$. For the random vector $\vec{X} = (X, Y)^T$, it is straightforward to check

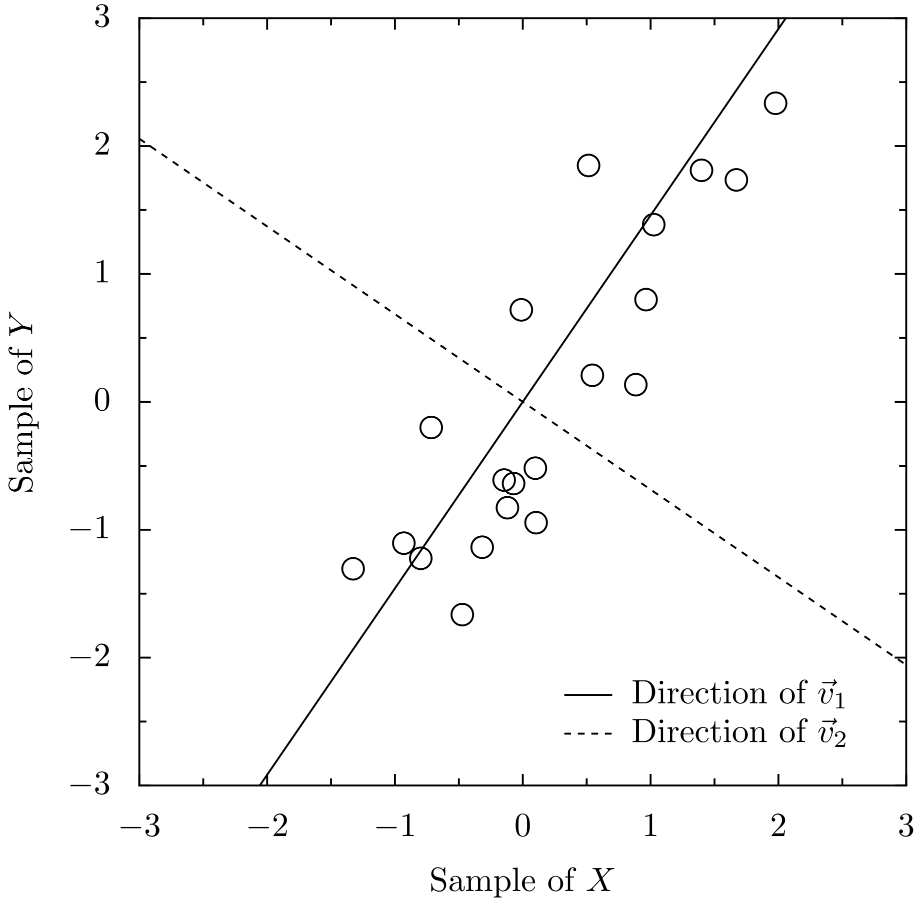

Figure 2.1: Samples of the random variables X and Y in the example.

that the eigenvalues of the covariance matrix are $\lambda_1 = 2.3$ and $\lambda_2 = 0.2$, where the corresponding eigenvectors are $\vec{v}_1 = (-0.6, -0.8)^T$ and $\vec{v}_2 = (-0.8, 0.6)^T$. 20 samples of the random variables X and Y where taken and the values obtained are shown in Figure 2.1. The direction of the eigenvectors is also plotted.

As we derived, the direction of $\vec{v}_1$ has maximum dispersion. However, note that the direction of the eigenvectors seem to form the “principal axes of the data”, which is actually a property of the coefficient vectors. Jolliffe (2002) sumarizes a large number of properties and ways in which the principal components and their corresponding vectors of coefficients (the eigenvectors of the covariance matrix) can be interpreted. Perhaps the most important property is dimensionality reduction: the fact that the N principal components, Z_i, can be reduced to $q < N$ principal components in order to minimize the sum of the squared perpendicular distances of the samples measured from this subspace (the sum of the squared perpendicular distance from the lines formed by the eigenvectors in Figure 2.1 to the samples). This is extremely useful when we are dealing with high dimensional random vectors: it means that we can apply our derived linear transformation to the random vector $\vec{X}$, obtain the principal components, select the ones that explain most of the variation on our data and analyse that sub-set of the data with minimum loss of information.

How can we interpret all this in the time series context? As we showed in Chapter 1, time series analysis is way more complex than just talking about random variables, because a process is a collection of random variables and, therefore, the probabilistic nature of it is changing with time. Unfortunately, PCA can't take into account this fact because we don't have enough information about the different random variables at each time index. This may seem rather dissapointing but, as will be shown, it is a good starting point in the analysis of time series.

The idea in applying PCA to indexed series is that, in practice, we may collapse a given process into a single random variable and at the same time see this random variable as a sum of other different random variables (which in the context of time series where also different procceses). For example, the flux of a star as measured from an instrument may be thought as a random variable which is the sum of different atmospheric and instrumental effects. This makes sense if we think in the distributions of these random variables: different procceses may produce different distributions when collapsed in a single random variable. To illustrate this concept, consider the flux measurements of a star as a function of time, $z_1(t)$, plotted in Figure 2.2. Here, the different values can be thought as being realizations of different random variables $Z_1(t)$. On the other hand, observing the frequency distribution of the star's flux, it can also be thought as realizations of a single random variable. Therefore, this frequency distribution can also be thought as our measurement of the distribution of the collapsed process, thinking of it as the measurement of the distribution of a single random variable.

A related subject of this kind of analysis is Functional Data Analysis, a concept that was introduced by Ramsay and Silverman (1997), which uses PCA in the context of continuous deterministic series, i.e., they assume that the data is a sample of some continuous function of some index (e.g. time). In this context, they ar-

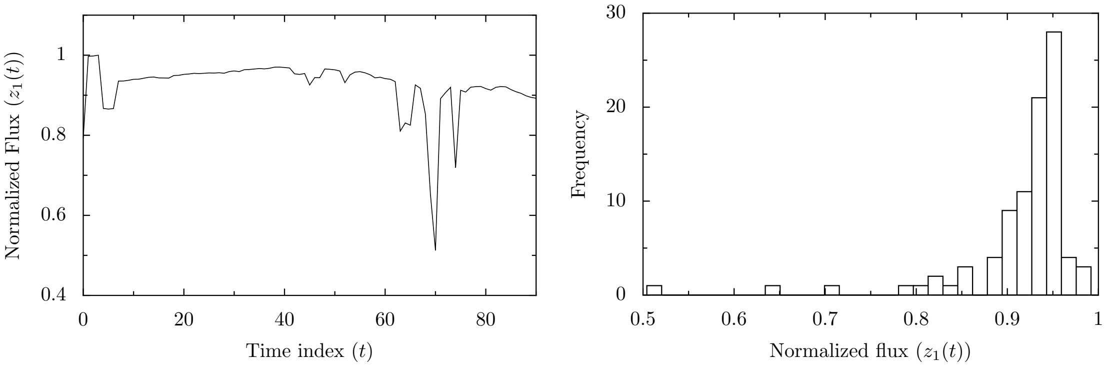

Figure 2.2: (Left) Light curve of a star (measuriement of the (normalized) flux as a function of time). (Right) Frequency distribution of the values of the star's flux.

gue that if we take independent measures of some variables (e.g. curves of human growth, econometric time series, etc.) what PCA actually does is to decompose each observation into a suitable orthonormal basis (the coefficients) whose (uncorrelated) coefficients are the principal components. In this sense, the principal components show special features of the data, which has very good results in a wide variety of areas (Ramsay and Silverman, 2002).

To see how this applies to our collapsed procceses, consider the random vector $\vec{X} = (X_1, X_2, ..., X_N)^T$, where this time the random variables may represent the different outcomes of N stochastic processes $X_i(t)$, $i = 1, 2, ..., N$, collapsed in them, e.g., the fluxes of different stars measured from an instrument (where now the PDF of each random variable has to be thought of as "the probability density of obtaning a given value for the flux"). Applying the linear transformation $\mathbf{V}$, recall that the i-th principal component is given by:

$$Z_i = \sum_{j=1}^{N} v_{i,j} X_j$$

Because the linear transformation $\mathbf{V}$ is orthonormal, $\mathbf{V}^{-1} = \mathbf{V}^T$ and the i-th random variable can be written as:

$$X_i = \sum_{j=1}^{N} v_{j,i} Z_j$$

This is almost what we where searching for! The above expresion can be thought as an expansion of the random variable X_i in terms of an orthonormal basis (the vectors of coefficients) and coefficients given by another (uncorrelated) random variable (the principal components). In the limit $N \to \infty$, this is known as the Karhunen-Loève theorem or expansion, which states that a random variable can be represented as an infinite linear combination of orthogonoal functions, whose coefficients are uncorrelated random variables. Because in our case we have a limited set of samples for each random variable, PCA may be seen as a truncated form of this expansion, which is known in the signal proccesing jargon as the Karhunen-Loève transform. Note that this transform is somewhat different from the usual transforms: here **the coefficients are the random variables** and the deterministic vectors contained in the linear transformation $\mathbf{V}$ are the functions.

In practice what we actually have are samples of the random variables, $x_i(t)$ (the different time series for each star), which are collected in order to create a data matrix, $\mathbf{X}$, where each row represents a different star and each column is a time index. Then, after substracting the mean from each row we create the **sample covariance matrix** where the element (i, j) of that matrix is:

$$\hat{\mathbf{\Sigma}}_X(i,j) = \frac{1}{M}\sum_{t=1}^{M} x_i(t)x_j(t)$$

Where $x_i(t)$ is the i-th star's time series and M is the number of samples. Once we obtain the sample covariance matrix $\hat{\mathbf{\Sigma}}_X$, we obtain its eigenvalues and eigenvectors and obtain the linear transformation $\mathbf{V}$. Finally, we apply this transformation to our data matrix $\mathbf{X}$ to obtain:

$$\mathbf{VX} = \mathbf{Z}$$

Where the i-th row of the matrix $\mathbf{Z}$ is the corresponding time series for the i-th principal component. In summary, what this really means is that we can find projections of the samples where the resulting time series are uncorrelated from each other.

The interesting interpretation about PCA in the time series context is, then, that the principal components define a set of uncorrelated signals that best explain our data when properly weighted by the coefficient vector. Furthermore, the eigenvalues asociated with each principal component are a measure of "how important" is a given principal component time series in order to explain our observed data. However, note that the obtained expansion is given in terms of uncorrelated random variables... is it possible to make an analogous expansion using **independent random variables** (which was what we wanted in the first place)? This question will be answered in the next section.

2.1.3 PCA and Whitening

We end this section with a discussion in a subject that will be of fundamental importance on the next section, which is the "whitening" of a random vector. As was stated in equation 2.1, given our linear transformation $\mathbf{V}$, the new random vector $\vec{Z}$ is uncorrelated but it is not white (i.e. their random variables have different variances). This is desirable because, as we'll see in future sections, it simplifies a lot of calculations and interpretations of the random variables. In this sense, a white random vector is "better" than an uncorrelated random vector.

It easy to show that a transformation that can project the initial random vector $\vec{X}$ to a white random vector $\vec{Z}$ is given by $\mathbf{V} = \mathbf{D}^{-1/2}\mathbf{E}^T$, because:

$$E\left[\vec{Z}\vec{Z}^T\right] = \mathbf{D}^{-1/2}\mathbf{E}^T E\left[\vec{X}\vec{X}^T\right]\mathbf{E}\mathbf{D}^{-1/2} = \mathbf{I}$$

This is called a **whitening transform**, for obvious reasons. It is interesting to note, however, that in fact any transformation of the form $\mathbf{V} = \mathbf{P}\mathbf{D}^{-1/2}\mathbf{E}^T$, where $\mathbf{P}$ is an orthogonal matrix will make a whitening transform. In order to simplify the notation, we'll use the transform $\mathbf{\Sigma}_X^{-1/2} = \mathbf{E}\mathbf{D}^{-1/2}\mathbf{E}^T$, which is widely used in the signal processing literature.

2.2 Independent Component Analysis

In the past section we saw that PCA is a powerful analysis tool, because it can decompose a given random variable into a series of uncorrelated random variables, properly weighted by an orthonormal basis. However, recall that, as we showed on Chapter 1, uncorrelated does not mean independent, so the components obtained with PCA may still be dependant of each other. With this in mind we posed the following question: if there exists a decomposition of a random variable in terms of uncorrelated random variables, is it possible to decompose a given random variable in terms of **independent** random variables? A classical way to introduce this problem is by the *cocktail-party problem* (Hyvrinen & Oja, 2000), a mathematical description of the problem first posed by Cherry (1953).

Imagine a room with N persons talking, which can be thought as N source signals which we'll identify by the random variable S_i and n microphones placed in different locations of the room which will sample the signals. Because the microphones are placed in different locations, they will scale differently the corresponding signals. Let the random variable that identifies the signal recieved by the i-th microphone be X_i. With this in mind, we can write the signals recieved by the microphones as:

$$X_i = a_{i,1}S_1 + a_{i,2}S_2 + ... + a_{i,N}S_N$$

Where the weights $a_{i,j}$ are real numbers that identify the weighting of each signal. Let the matrix $\mathbf{A}$ have the element $a_{i,j}$ at the ith row and jth column. The problem can then be stated as:

$$\vec{X} = \mathbf{A}\vec{S} \tag{2.2}$$

Where $\vec{X}$ is the random vector containing the microphone signals (the mixtures), $\vec{S}$ is the random vector containing the source signals (also called the independent components) and $\mathbf{A}$ is called the mixing matrix. The problem could be solved if the mixing matrix is invertible, because then:

$$\mathbf{A}^{-1}\vec{X} = \vec{S}$$

The question is: by assuming that the source signals $\vec{S}$ are independent, can we find an invertible matrix $\mathbf{A}$ and solve the problem? This is the problem that Independent Component Analysis (ICA) focuses to solve.

2.2.1 Complexity Reduction of the ICA problem

In order to solve the ICA problem, we'll simplify it further by assuming, without loss of generality, that the expected value of the random variables present on the random vector $\vec{X}$ are zero (note that this can always be made true by substracting its mean vector $\vec{\mu}$). On the other hand, note that there's a scale ambiguity for the ICA problem, equation (2.2). Because the mixing matrix and the independent components are unknown, any scalar multiplier present in one of them could be cancelled by dividing by this multiplier on the other. Because of this, we'll set the variance of the source signals to 1, i.e., $E[\vec{S}\vec{S}^T] = \mathbf{\Sigma}_S = \mathbf{I}$. Note that this still leaves a sign ambiguity for the independent components.

The starting point to solve the ICA problem is to apply a whitening transformation to our vector of mixtures. This is because, as we'll see in short, it'll reduce the complexity of the problem. Recall that one such transformation is $\mathbf{\Sigma}_X^{-1/2} = \mathbf{E}\mathbf{D}^{-1/2}\mathbf{E}^T$. Applying this transformation on both sides of equation (2.2), we obtain:

$$\vec{Z} = \tilde{\mathbf{A}}\vec{S} \tag{2.3}$$

Where $\vec{Z} = \mathbf{\Sigma}_X^{-1/2}\vec{X}$ is the new whitened vector and $\tilde{\mathbf{A}} = \mathbf{\Sigma}_X^{-1/2}\mathbf{A}$ is the new mixing matrix. Note that this is of the same form of the original ICA problem, with the subtle difference that our problem has been reduced by half. To see this, note that the new mixing matrix, $\tilde{\mathbf{A}}$, is orthogonal because:

$$E\left[\vec{Z}\vec{Z}^T\right] = \tilde{\mathbf{A}}\left[\vec{S}\vec{S}^T\right]\tilde{\mathbf{A}}^T = \tilde{\mathbf{A}}\tilde{\mathbf{A}}^T$$

And because the random vector $\vec{Z}$ is white, we have:

$$\tilde{\mathbf{A}}\tilde{\mathbf{A}}^T = \mathbf{I}$$

The important feature of $\tilde{\mathbf{A}}$ being orthogonal is the fact that an orthogonal matrix has only $N(N-1)/2$ free parameters. Thus, by applying a whitening transform, we reduced the complexity of the problem considerably.

The whitening transformation that we just performed justifies one last assumption that will be fundamental in the solution of the ICA problem: **we'll assume that the distribution of the independent components are not gaussian**. The practical reason for this is that when the distributions of the independent components are gaussians, then ICA can go no further than a whitening transformation, which can be shown as follows. Consider the generalized gaussian distribution for the random vector $\vec{S}$:

$$f_S(\vec{s}) = \frac{1}{(2\pi)^{n/2}}\exp\left(-\frac{1}{2}\vec{s}^T\vec{s}\right) = \frac{1}{(2\pi)^{n/2}}\exp\left(-\frac{1}{2}||\vec{s}||^2\right)$$

If the mixing matrix is orthogonal (which is the case if we whitened the mixtures vector), then $\tilde{\mathbf{A}}^T\vec{Z} = \vec{S}$ and the corresponding PDF for $\vec{Z}$ is then:

$$f_Z(\vec{z}) = \frac{1}{(2\pi)^{n/2}}\exp\left(-\frac{1}{2}||\tilde{\mathbf{A}}^T\vec{z}||^2\right)|\det\tilde{\mathbf{A}}^T| = \frac{1}{(2\pi)^{n/2}}\exp\left(-\frac{1}{2}||\vec{z}||^2\right)$$

And the distributions are invariant under orthogonal transformations: we can't unmix the mixtures in $\vec{X}$. This could seem very dissapointing on the time series context but, as we'll see in future subsections, there are algorithms that can use the time structure of the mixtures in order to extract independent components whose stochastic nature is gaussian (like AR procceses, for example). The first algorithms that we'll present are the FastICA and the EFICA algorithms, the former being the classic algorithm for the solution of the ICA problem and the last one being an optimization of the former.

2.2.2 Solving the ICA problem: the FastICA algorithm

As we introduced at the end of the past sub-section, one of the assumptions of the ICA model is that the independent components are non-gaussian. Thus, considering the central limit theorem (which can be generalized to the sum of finite-variance independent random variables by the Lindeberg-Feller Central Limit Theorem proven in a probabilistic abordable way in Goldstein (2009)), a linear combination of these components should be always more gaussian than any of the independent components alone. This fact will be used in order to solve the ICA problem.

Consider a that we have performed the whitening transformation proposed on the last subsection, and form a linear combination of the random variables present on the whitened mixture vector $\vec{Z}$:

$$Y_1 = \sum_{i=1}^{N} w_i Z_i = \vec{w}^T\vec{Z},$$

where the vector $\vec{w}$ contains the coefficients w_i, $i = 1, .., n$. We can use equation (2.3) to write $\vec{Z} = \tilde{\mathbf{A}}\vec{S}$. With this we get:

$$Y_1 = \vec{w}^T\tilde{\mathbf{A}}\vec{S} = \vec{\alpha}_1^T\vec{S},$$

where $\vec{\alpha}_1 = \vec{w}^T\tilde{\mathbf{A}}$. What we have just proven is that Y_1 is not only a linear combination of the random variables present in the mixture vector, but also a linear combination of the independent components. Because this linear combination of the independent components will be more gaussian than any of the components, if we could find a measure of non-gaussianity of Y_1, maximization of that criterion is equivalent to the search of an independent component.

The work of Hyvrinen & Oja (1997) implements this method considering kurtosis, which is the name for the fourth cummulant of a random variable, as a measure of non-gaussianity. This is because, as we showed at the end of section 1.1.2, for a gaussian random variable kurtosis is zero. Therefore, because depending on the distribution kurtosis can be positive (for PDFs peaked at zero) or negative (for flatter PDFs), maximization or minimization of kurtosis is equally valid on the search for the independent components. In order to form a well-defined problem, they also assume that $||\vec{w}|| = 1$ and therefore minimize or maximize the criterion:

$$\kappa_4(\vec{w}^T\vec{Z}) = E\left[(\vec{w}^T\vec{Z})^4\right] - 3||\vec{w}||^4$$

Which can be done by a gradient descent or ascent algorithm, depending if what is wanted is maximization or minimization of kurtosis.

One of the problems with the explained ICA algorithm is that kurtosis is very sensitive to outliers. This is because it has to be estimated from samples of the random variables and values on the tails of a PDF can affect considerably our measurements, such as spikes or simply bad values. This motivates the usage of a different approach for the search of non-gaussianity and also the search for a better justification of this solution to the ICA problem.

2.2.3 The robust FastICA and EFICA algorithms

Comon (1994) had already introduced the different views that the ICA problem can have and, specially, dedicated to its solution as an extension of PCA. He proposed the minimization of a different quantity that could measure the independance between the independent components: mutual information.

In order to define mutual information we first define the concept of **negen-**

tropy. We've already introduced the concept of differential entropy in section 1.2, and showed that given a covariance, the maximum entropy distribution is that of a gaussian distribution. However, we had certain problems: we defined the differential entropy in such a way that it is not invariant under transformations and, furthermore, it can be negative. Negentropy, then, is defined as:

$$J(\vec{Y}) = H(\vec{Y_g}) - H(\vec{Y})$$

Where $\vec{Y_g}$ is a gaussian random variable with the same mean and covariance as $\vec{Y}$. This definition solves our worries: it is always positive (because the gaussian distribution has maximum entropy), invariant under linear transformations and is zero only if $\vec{Y}$ has a gaussian distribution. With this at hand, the mutual information is defined as:

$$I(\vec{Y}) = J(\vec{Y}) - \sum_{i}^{N} J(Y_i)$$

As we stated earlier, this is a measure of independance of the random variables present in the random vector $\vec{Y}$ (note that this measure is zero only if the random variables are independent). The important interpretation of the mutual information is that it measures the redundancy of the random variables.

Given the above definitions, by maximization of the negentropy of a random variable we measure its non-gaussianity. Therefore, maximization of the negentropy of the random variable Y_i, introduced in the past subsection, will lead us to find an independent component. On the other hand, in order to minimize the mutual information we also have to maximize the negentropy of the individual random variables (the sum in the definition of mutual information). Both interpretations then help us to solve the problem.

In Hyvrinen & Oja (1999), this method is implemented. However, despite the amazing interpretations considered earlier, measuring negentropy is not an easy task. Comon (1994) showed that an approximation of negentropy could be made in terms of the cummulants of the PDF but, as we stated earlier, this leads to estimation problems. Using the maximum entropy principle, however, Hyvrinen (1998) found that a suitable aproximation for negentropy is:

$$J(Y_i) \approx c\left(E[G(Y_i)] - E[G(V)]\right)^2$$

Where c is a constant, V is a zero-mean and unit variance random variable and the function G is a suitable function such that: (a) its estimation from sample values is not statistically difficult (and is robust in the sense that is not too sensitive to outliers), (b) it doesn't grow faster than quadratically and (c) it is at least a two times differentiable function. The classical functions (Hyvrinen, Karhunen & Oja, 2000) used in ICA algorithms are the logarithm of the hyperbolic cosine, $G(x) = k\log(\cosh(kx))$, where $1 < k < 2$ is a constant and the negative version of the gaussian function, $G(x) = -\exp(-x^2/2)$.

The mentioned FastICA algorithms have two versions: the one-unit and symmetric approaches. The first approach finds the independent components one by one by maximization of negentropy on each of the Y_i, while the second approach finds these components simultaneously by paralell one-unit iterations. Because the latter can use different functions G in order to maximize negentropy for each unit, the functions can be smartly chosen in order to minimize the covariance matrix of the error in the estimation of the independent components. This covariance matrix has a theoretical lower bound, called the **Cramer-Rao lower bound**, which is desirable to be obtained at least asymptotically. This idea is implemented in the work of Koldovský, Tichavský & Oja (2006), where an adaptive choice of the functions is made for each source after determining (with the original symmetric FastICA with any function) what the source signals "look like". This efficient form of the FastICA algorithm is called EFICA (Efficient Fast ICA) and will be tested in the present work for astrophysical time series.

2.2.4 Using the time structure: the SOBI algorithm

Our analysis so far has been treating our process as if it where collapsed into a single random variable. In practice, this means that if we shuffle our time series data we should obtain the same results as the unshuffled version of it. If our data where composed of independent not time-correlated procceses, this would clearly make sense but in real applications this seems like a waste of information because we know that there are time correlated procceses present in our time series. Let's redefine, then, the ICA problem in terms of stochastic processes.

Consider the "stochastic" ICA problem now defined as:

$$\vec{X}(t) = \mathbf{A}\vec{S}(t) \tag{2.4}$$

Where now the random vectors $\vec{X}(t) = (X_1(t), ..., X_n(t))^T$ and $\vec{S}(t) = (S_1(t), ..., S_n(t))^T$ are vectors that carry, in principle, a different process in each one of their elements. In other words, they are allowed to change in time (i.e. at each time t there's a different random variable). The starting point is to whiten the random vector $\vec{X}$, just as we did in the original solution of the ICA problem. With this we get:

$$\vec{Z}(t) = \tilde{\mathbf{A}}\vec{S}(t)$$

Recall that the problem we want to solve is to find the separated signals $\tilde{\mathbf{A}}^T\vec{Z}(t) = \vec{S}(t)$ (recall that $\tilde{\mathbf{A}}$ is orthogonal so $\tilde{\mathbf{A}}^{-1} = \tilde{\mathbf{A}}^T$), subject to the independance of the $S_i(t)$. This means that not only the random variables present at a given time in our source signals (now source "procceses") are independent from each other, but also their time-lagged versions. In order to found a theoretical ground to work on, we'll assume that the sources are made up of (wide sense) stationary processes. To state all this in a mathematical formalism, we want the different time lagged cross-covariances of the procceses to be zero:

$$E\left[S_i(t)S_k(t-\tau)\right] = 0, \ \forall \ i, k, \tau, \ i \neq k \tag{2.5}$$

We can summarize this information into the time-lagged covariance matrix, which is defined as:

$$\mathbf{\Sigma}_S^\tau = E\left[\vec{S}(t)\vec{S}(t-\tau)^T\right]$$

Where, according to the constraint in equation (2.5), $\mathbf{\Sigma}_S^\tau$ is a diagonal matrix for all values of τ. Consider now a modified version of the time-lagged covariance matrix of the random vector $\vec{Z}(t)$, $\mathbf{\Sigma}_Z^\tau$, for reasons that will become clear in short, given by:

$$\bar{\mathbf{\Sigma}}_Z^\tau = \frac{1}{2}\left[\mathbf{\Sigma}_Z^\tau + (\mathbf{\Sigma}_Z^\tau)^T\right] \tag{2.6}$$

Note that:

$$\mathbf{\Sigma}_Z^\tau = E[\vec{Z}(t)\vec{Z}(t-\tau)^T] = \tilde{\mathbf{A}}E[\vec{S}(t)\vec{S}(t-\tau)^T]\tilde{\mathbf{A}}^T = \tilde{\mathbf{A}}\mathbf{\Sigma}_S^\tau\tilde{\mathbf{A}}^T$$

Replacing this result in equation (2.6) we get:

$$\bar{\mathbf{\Sigma}}_Z^\tau = \frac{1}{2}\left[\tilde{\mathbf{A}}\mathbf{\Sigma}_S^\tau\tilde{\mathbf{A}}^T + \tilde{\mathbf{A}}\left(\mathbf{\Sigma}_S^\tau\right)^T\tilde{\mathbf{A}}^T\right] = \tilde{\mathbf{A}}\mathbf{\Sigma}_S^\tau\tilde{\mathbf{A}}^T$$

What this result states is that the mixing matrix and the lagged (diagonal) covariance matrix of $\vec{S}(t)$ are given by the eigenvalue decomposition of the matrix $\bar{\mathbf{\Sigma}}_Z^\tau$! Note that this decomposition exists because this matrix is symmetric (that's actually the reason why we defined such a matrix). In theory, if our assumptions of stationarity hold, we only need to compute one lagged covariance matrix of $\vec{Z}(t)$, find the eigenvalue decomposition of $\bar{\mathbf{\Sigma}}_Z^\tau$ and the problem is solved. However, in real applications, we can only **estimate** the covariance matrix and, therefore, can only estimate its eigenvalue decomposition. This is a huge problem because the choice of the lag τ is fundamental and could even lead to an indetermination of the problem due to the possible degeneracy of eigenvalues for matrix $\bar{\mathbf{\Sigma}}_Z^\tau$.

Belouchrani et al. (1997) proposed the Second Order Blind Identification (SOBI) algorithm to solve this problem. This algorithm proposes the *joint diagonalization* of the matrix:

$$\tilde{\mathbf{A}}^T\bar{\mathbf{\Sigma}}_Z^\tau\tilde{\mathbf{A}} = \mathbf{\Sigma}_S^\tau$$

For several lags τ. This consists in finding an orthogonal matrix $\hat{\mathbf{A}}$ (the estimate of $\tilde{\mathbf{A}}$) such that the matrices $\hat{\mathbf{A}}^T\bar{\mathbf{\Sigma}}_Z^\tau\hat{\mathbf{A}}$, where $\tau \in \{\tau_j | j = 1, ..., K\}$, are as diagonal as possible. This criterion, defined in the work of Belouchrani et al. (1997) as the "joint diagonalization criterion" (JD criterion) is defined as:

$$C(\bar{\mathbf{\Sigma}}_Z^\tau, \hat{\mathbf{A}}) = \sum_{k=1}^{K} \text{off}\left(\hat{\mathbf{A}}^T\bar{\mathbf{\Sigma}}_Z^{\tau_k}\hat{\mathbf{A}}\right) \tag{2.7}$$

Where off($\mathbf{M}$) is a scalar-valued function of the matrix $\mathbf{M}$ which returns the absolute value of the squared sum of the off-diagonal terms $M_{i,j}$ of this matrix, i.e.:

$$\text{off}(\mathbf{M}) = \sum_{i,j,i\neq j}^{N} |M_{i,j}|^2$$

The minimization, then, of the JD criterion, equation (2.7), gives the desired mixing matrix estimate $\hat{\mathbf{A}}$. This can be easily done by efficient joint diaginalization algorithms, such as FFDIAG (Ziehe et al., 2004). Another interesting part of this criterion is that its minimization doesn't need that the individual matrices are actually diagonalizable by an orthogonal matrix: the joint diagonalizer matrix $\hat{\mathbf{A}}$ is just the closer we can get to diagonalize matrix $\bar{\mathbf{\Sigma}}^{\tau}_{Z}$. This is just what we wanted in statistical terms, because as we stated before, we actually estimate this matrix from samples. This is just like fitting a curve: we don't actually expect the curve to fit *literally* the sampled data, but that it minimizes the distance between it and the samples.

2.2.5 The robust version of SOBI: the WASOBI algorithm

The view of the SOBI algorithm as a "matrix fitting" algorithm, i.e., a least-squares fit of the matrix $\hat{\mathbf{A}}$ with respect to the matrices $\bar{\mathbf{\Sigma}}^{\tau}_{Z}$ showed that the SOBI algorithm is not optimal. Yeredor (2000) exposed the problem that, because the application of the whitening transformation $\mathbf{V}$ to the initial stochastic ICA problem, equation (3.2), is obtained via the covariance matrix at zero lag, i.e., it focuses in the diagonalization of the covariance matrix $\mathbf{\Sigma}_X = \mathbf{\Sigma}^{\tau=0}_{X}$, it may produce a poor diagonalization at different lags. Another problem also exposed is that the errors in estimating the covariances are strongly correlated.

In order to solve these problems, Yeredor (2000) proposed the formulation of a weighted least-squares problem, called the Weight-Adjusted Second Order Blind Identification (WASOBI) algorithm, starting from the initial stochastic ICA problem, equation (3.2). Following the same arguments as for the derivation of the SOBI algorithm, but now using the original lagged covariance matrix of the random vector $\vec{X}(t)$, it is easy to show that:

$$\bar{\mathbf{\Sigma}}^{\tau}_{X} = \mathbf{A}^{-1}\mathbf{\Sigma}^{\tau}_{S}\left(\mathbf{A}^{-1}\right)^{T}$$

With this in mind, Yeredor focuses on obtaining estimates of the matrix $\mathbf{A}^{-1}$ and of the diagonal elements of $\mathbf{\Sigma}_S^\tau$ directly, such that they "optimally fit" the lagged covariance matrices $\bar{\mathbf{\Sigma}}_X^\tau$. In order to pose this problem, consider the array $\mathbf{Y}$, composed of τ_{max} rows (here τ_{max} represents the number of lagged matrices to be estimated), where each row is a row vector with the estimates of the upper triangular matrix of $\bar{\mathbf{\Sigma}}_X^\tau$ (because considering the whole matrix would be redundant) and the array $\mathbf{G}$, composed of the corresponding elements of the unknown matrix $\mathbf{A}^{-1}\mathbf{\Sigma}_S^\tau\left(\mathbf{A}^{-1}\right)^T$. The proposed weighted least-squares criterion to be minimized is:

$$C_{WLS} = [\mathbf{Y} - \mathbf{G}]^T\mathbf{W}[\mathbf{Y} - \mathbf{G}]$$

In order to find the weights, however, some assumptions on the stochastic models of the sources must be made. On Yeredor's original paper, he derived the corresponding weights for a moving average process. However, as we saw on Chapter 1, it would be desirable to obtain the weights for a maximum entropy process such as an autoregressive (AR) process. Tichavský et al. (2006) made this derivation, where he assumes that the underlying model is composed of a weighted sum of independent (and different) AR procceses. A WASOBI implementation using this model will also be tested on the present work for astrophysical signals.

2.2.6 Combining strengths: the MULTI-COMBI algorithm

As it may be suggested from our previous presentations of the different ICA algorithms, real time series are composed of both, independently distributed sources (where, as long as they are non-gaussian, EFICA should outperform) and time-correlated sources (where, as long as the signals are stationary and have different spectra, WASOBI should outperform). Therefore, the ideal would be to combine both algorithms, letting WASOBI take care of the time-correlated signals and EFICA of the independently distributed sources. This is idea was first implemented in an algorithm called COMBI, which has been upgraded to form the MULTI-COMBI algorithm Tichavský et al. (2008).

In order to present how this algorithm works, we may first ask how to measure "how well" an algorithm decomposed a given independent component. In order to define such a measure, it is important to define the so-called gain matrix:

$$\mathbf{G} = \mathbf{A}\mathbf{A}^{-1}$$

Where $\mathbf{A}$ is the mixing matrix of the ICA problem, present in equations (2.2) and/or (3.2). Of course, in the theoretical case $\mathbf{G}$ should be the identity matrix, but in real applications this is not the case. In terms of this matrix, the standard measure of "how well" an algorithm performed blind separation is defined the **interference-to-signal ratio** (ISR) as the matrix **rISR**, whose elements $\text{rISR}_{i,j}$ are given by:

$$\text{rISR}_{i,j} = G_{i,j}^2 / G_{i,i}^2$$

Note that in the presented algorithms one usually obtains the inverse of the mixing matrix *from* the estimated mixing matrix and viceversa, so this definition of the ISR may seem rather pointless. However, given some suitable models for the sources for which the algorithms are optimized, one can obtain the expected value (in practice the asymptotic expected value) of the **rISR** matrix, $E[\mathbf{rISR}] = \mathbf{ISR}$, which in turn defines the ISR vector $\vec{isr}_i = \sum_{j=1, j\neq i} ISR_{i,j}$. For example, the asymptotic ISR for the EFICA algorithm can be estimated taking different combinations of the expected values of the functions used in the algorthm to obtain the aproximations for negentropy, evaluated at each obtained independent component (Koldovský, Tichavský & Oja, 2006). On the other hand, the asymptotic ISR of the WASOBI algorithm can be obtained from the AR parameters for each source, along with their autocorrelation estimates (Tichavský et al., 2006). With this at hand, it is possible then to measure, at least asymptotically, how well did each algorithm perform.

On one side, the COMBI algorithm works by finding those signals that have the lowest ISR vectors. First, a run is made with both EFICA and WASOBI and the signal with the lowest ISR vector is separated from the rest, and the problem is redefined as separating the remaining signals. On the other hand, the upgraded version of the COMBI algorithm, the MULTI-COMBI algorithm, uses the idea of "multidimensional components", which are clusters of signals that can be well-separated from the mixture but are hard to separate from each other. In the case of EFICA, components that have nearly gaussian distributions may form such clusters, while in the case of WASOBI components with similar spectra may form such components. With this in mind, the MULTI-COMBI algorithm tries to find clusters of signals by observing the structure of the ISR matrix forming clusters (groups) of signals. The idea is that if the ISR matrix for one of the algorithms (EFICA or WASOBI) has, for example, a block diagonal structure, then clusters may exist and the other algorithm may be capable of separating those clusters. This algorithm will also be tested in the present work for astrophysical signals.

Chapter 3

Applications of PCA and ICA to astrophysical time series

In the present chapter, we'll apply the PCA and ICA algorithms presented in Chapter 2 to astrophysical time series. In particular, we'll use those algorithms in the context of extrasolar planets, where we'll test the algorithms in order to obtain the transit curve of the exoplanet WASP-6b (Gillon et al., 2009). Implementations of algorithms performing PCA and ICA where made by the author. For PCA, routines that perform standard PCA (non-whitening transform) and whitening transforms where implemented for the PYTHON programming language. For ICA, algorithms for the EFICA, WASOBI and MULTI-COMBI algorithms where also implemented for the PYTHON programming language. A total of 1500 lines of code summarize these implementations.

This chapter is divided in two parts. On the first part we analyse different methods for determining the principal sources of multiplicative noise present in our measurements of the flux of different stars, along with a discussion based on the results. On the second part, we use these sources to obtain the transit curve for WASP-6b. This will aid us in comparing how well our algorithms perform in the light of the published parameters of this exoplanet.

3.1 Transit light curve model

The classical model for planetary transit light curves has been analitically solved by Mandel & Agol (2002). Basically, they model the flux of the planet-star system as

$$F(t) = F_*(t) + F_p(t) - I(t),$$

where $F_*(t)$ is the flux of the star, $F_p(t)$ is the flux (emmited and/or reflected) of the planet and $I(t)$ in our case of interest is the decrement in flux given by the passage of the planet in front of the star. In their original paper, Mandel & Agol assumed implicitly that the "unobscured flux", $F_*(t) + F_p(t)$, is constant, and modelled the normalized transit light curve with respect to this flux. This can be obtained dividing by $F_*(t) + F_p(t)$ in this last equation, giving

$$\frac{F(t)}{F_*(t) + F_p(t)} = f(\vec{\theta}, t), \tag{3.1}$$

where the deterministic function $f(\vec{\theta}, t)$ represents the normalized transit light curve, which will be refered from now on as the "transit light curve" and $\vec{\theta}$ is a vector of parameters that define the light curve, such as the planet-to-star radius ratio, the inclination of the orbit of the planet, the period, etc. It is interesting to note that this parameter vector can also take into account stellar atmospheric effects such as limb darkening, which affects the shape of the observed transit. For the sake of brevity we'll not discuss the details of the functional form of this function in the present thesis, but the interested reader is refered to the excellent review of transit light curves of Winn (2010).

3.2 The Data

The data to be used in our applications of the PCA and ICA algorithms is from a transmission spectra of the extrasolar planet WASP-6b (Gillon et al., 2009). The data consists in spectra of different stars taken with the Inamori Magellan Aeral Camera and Spectrograph (IMACS) at the Magellan Baade 6.5m telescope, located at Las Campanas Observatory, Chile. These spectra where measured as a function of time, for $M = 91$ samples taken from 55473.0381 to 55473.2779 MHJD. However, for our calculations we'll only make use of the white light curve of our stars, namely, the resulting light curve after summing the light present in our stars in all wavelengths.

In the present chater we'll use 10 comparison stars in order to identify the different components that form the time series. To do this, we'll proceed as follows:

1. Obtain the light curves for the 10 comparison stars and for WASP-6 by summing the flux present in all wavelengths of each spectra and taking the logarithm (base 10) of this flux.

2. Preprocess the 10 comparison stars with PCA and select the principal components that will be used both for the ICA algorithms and for a comparison fitting using these PCs (i.e. perform dimensionality reduction).

3. Perform ICA using EFICA, WASOBI and MULTI-COMBI separately, in order to obtain the independent components by three different methods.

4. Perform a Markov Chain Monte Carlo (Ford, 2005; Gregory, 2005) fit to a transit light curve of WASP-6 in four different ways: using the Independent Components obtained by the EFICA, WASOBI and MULTI-COMBI algorithms and finally using the Principal Components as fitting time-series. All of this, of course, along with a transit light curve model (Mandel & Agol, 2002) in order to extract the parameters of the transit.

The idea of the 4th step is that we assume that the measured flux from our objective star (WASP-6 in our case) is composed of independent components (systematic effects) which affect it. In other words, mathematically we model the measured flux of the star-planet system, $X(t)$, as

$$X(t) = Y(t)F(t),$$

where $Y(t) = Y_1(t)Y_2(t)...Y_{n-1}(t)$ are the $n-1$ multiplicative sources of noise (atmospheric effects, instrumental systematics, etc.) and $F(t) = (F_*(t) + F_p(t))f(\vec{\theta}, t)$ (see equation 3.1). Taking the logarithm of this measured flux $X(t)$ and replacing the given expansions of the $F(t)$ and $Y(t)$ terms we obtain

$$\log_{10}[X(t)] = \sum_{i=1}^{n-1} \log_{10}[y_i(t)] + \log_{10}[F_*(t) + F_p(t)] + \log_{10}[f(\vec{\theta}, t)],$$

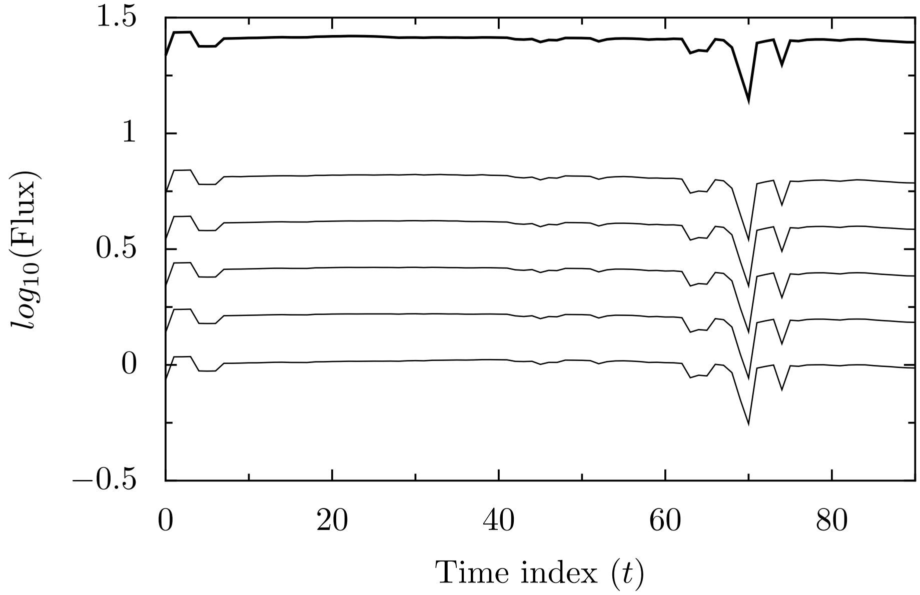

Figure 3.1: Some (white) light curves of comparison stars (bottom, thin lines) and of WASP-6 (top, thick line) obtained from the data. They where shifted in log-flux by an arbitrary constant in order to show the trends present in them.

where, in our terminology of the ICA problem, we could formulate the problem as

$$\log_{10}[X(t)] = \sum_{i=1}^{n} \alpha_i S_i(t) + \log_{10}[f(\vec{\theta}, t)],$$

where the coefficients α_i would be the mixing matrix elements from equation (3.2), each source signal S_i represents a different independent random variable that define the multiplicative noise along with the flux of the star if the transit where not present and $f(\vec{\theta}, t)$ is the transit light curve model (with parameter vector $\vec{\theta}$). Five white light curves obtained for our comparison stars along with the light curve for WASP-6 (thick line) are plotted in Figure 3.1. All the light curves are stored in a data matrix $\mathbf{X}$ where each row is a different light curve that has been mean substracted.

3.3 Pre-proccesing of the data

3.3.1 PCA of the comparison stars: how many PCs?

It may be clear by now that considering to keep all of the principal components that can be obtained from our set of 10 comparison stars would not be a good idea. This is because we know that it is very unlikely that *exactly* 10 different uncorrelated random variables form the basis of our time-series. Clearly, the first , most significant, principal components have to be the ones to keep, but, what do the remaining principal components tells us about the process? Consider for a moment the noisy ICA model (Hyvrinen, Karhunen & Oja, 2000)

$$\vec{X}(t) = \mathbf{A}\vec{S}(t) + \vec{N}(t),$$

where $\vec{X}(t)$ is our mixture vector, $\vec{S}(t)$ our source vector, $\mathbf{A}$ our mixing matrix and $\vec{N}(t)$ is a zero-mean white noise process. Assume also that there's white noise of equal variance added to our lightcurves. This means that the covariance matrix of our mixture vector is given by

$$\mathbf{\Sigma}_X^t = \mathbf{A}\mathbf{A}^T + \sigma^2\mathbf{I},$$

where we have used the fact that $\mathbf{\Sigma}_S^t = \mathbf{I}$, i.e., the sources have unit variance. Let's make our assumption, then, that in the case where $\sigma^2 = 0$, the SVD decomposition of $\mathbf{A}\mathbf{A}^T$ only has $n < N$ uncorrelated random variables (i.e. at least n independent components) that define our PCA expansion, giving then that the SVD decompositon is given by $\mathbf{L}\mathbf{D}\mathbf{R}^T$, where the diagonal matrix $\mathbf{D}$ has only n non-zero diagonal elements. Then, the SVD decomposition of the covariance matrix of our mixture vector is

$$\mathbf{\Sigma}_X^t = \mathbf{L}\mathbf{D}\mathbf{R}^T + \sigma^2\mathbf{I} = \mathbf{E}\mathbf{D}\mathbf{E}^T,$$

where the $N-n$ last eigenvalues λ_i of the decomposition $\mathbf{E}\mathbf{D}\mathbf{E}^T$ are constants equal to σ^2. What this tells us is that, in the presence of white noise, an approach to find how many principal components we should leave is to determine where the eigenvalues are aproximately equal. To illustrate this method, we applied standard PCA to the data matrix $\mathbf{X}$. The resulting eigenvalues not taking into account the first and second one (for illustration) are plotted in Figure 3.2.

As we can see from Figure 3.2, the suggestion is to retain the first 7 principal components, because the eigenvalues of the last 3 principal components are almost equal and constant (which is even clearer after seeing the close up of Figure 3.2). However, we may still ask if the source of the remaining principal components is really the noise proposed by the noisy ICA model. Another way of looking at the problem is to

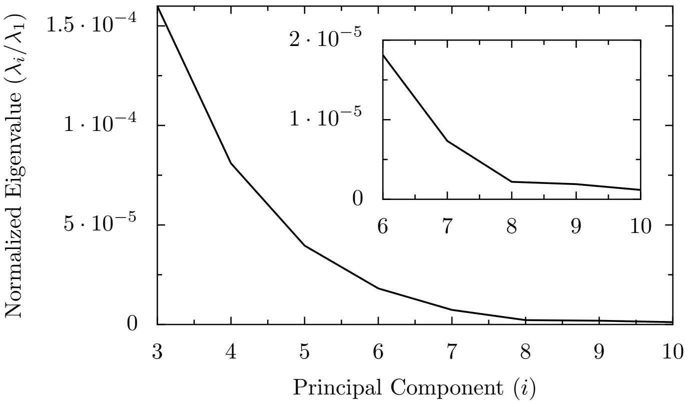

Figure 3.2: Normalized eigenvalues λ_i/λ_1 for $i = 3, ..., 10$ with a close up for $i = 6...10$.

think of the extra principal components as redundant information: time series that are created in order to "match" the constrained number of principal component that best explain our data. Given these two interpretations of the "break-point" on the eigenvalue v/s index plot, there are several methods to estimate which components we should retain (see, e.g., Chapter 6 in Jolliffe, 2002), but none of the standard measures seem to suit our needs in the time series analysis context (see Cangelosi & Goriely, 2007, for a survey on the current methods of estimation of the number of PCs to retain). Because of this, we'll use a very simple method that has been widely used in classification of morphological patterns: the "leave-one-out" cross-validation (LOOCV) (Diana & Tommasi, 2002).

The simplest form of LOOCV consists in leaving out one sample of our data matrix, estimate the corresponding covariance matrix without this value and then calculate the corresponding principal components. By calculating the PCs, we can formulate a prediction on the missing value and compare it with the real value by using just the first k principal components. We repeat this procedure for every sample on our data matrix and obtain the mean absolute difference (MAD) $e(k)$ between the predicted value using the first k principal components and the real value. The idea is then to compare in which number of k principal components, adding a new PC doesn't improve significantly the prediction with respect to the prediction made by the $k - 1$th PCs. Figure 3.3 shows the ratio $e(k + 1)/e(k)$.

One might be tempted to select the first two principal components, because there's a minimum for $k = 2$ (i.e., $e(k + 1)$, the error in the prediction using the

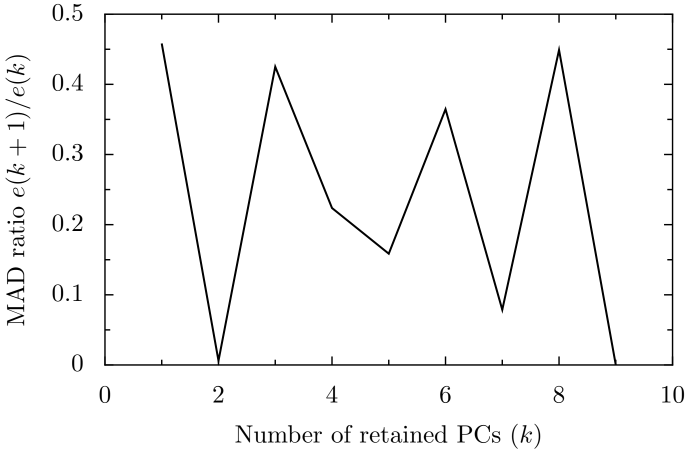

Figure 3.3: Ratio of mean absolute difference $e(k)$. The local minimum is clearly visible at $k = 7$.

first $k = 3$ components, is very small compared to $e(k)$, the error in the prediction for $k = 2$). However, just using two principal components, gives us an error in the prediction at an order of millimagnitudes, which is large compared to our needs (for transit data we need fractions of millimagnitudes in precision). If we add the fact that we need to predict a whole time series of 91 samples, taking just the two first principal components is not acceptable. The obvious next minimum is at $k = 9$ but, again, using all the principal components could introduce random errors given by random noise produced by white noise or by redundant information. The next minimum is at $k = 7$, which is consistent with our observations of the eigenvalue v/s index plot and therefore we decide to keep the first 7 components.

This space of the first 7 principal components is usually called the “signal subspace” in the signal proccesing jargon, because of the results shown concerning the noisy ICA model: these are *possibly*[1] the “real signals” that formed our mixture, whereas the remaning components are thought to be noise and/or artifacts from overlearning (obtention of more mixtures than sources). The first 7 PCs to be retained are shown in Figure 3.4. In order to observe how well we are separating the “real” components from noise, the last three (rejected) principal components are plotted in Figure 3.5. It is clear, at least graphically, that these signals are almost

[1] We enfatize the fact that what we just found are the best uncorrelated time series that best explain our data. They may be independent or not, so these seven series represent an upper bound to the number of independent sources of the ICA model.

pure noise. Furthermore, the apparent “lack of structure” increases as the principal component number increases.

From the plot of the retained PCs, it is clear that they still show strong correlations (for example, the spikes at $t \sim 70$ are present in almost every principal component). However, they seem to show some interesting features concerning some known sources of multiplicative noise. For example, the second PC has a strong correlation with the rotator off-set angle (93% of linear correlation), which is a well-known instrumental systematic effect addeded to our data. Measures of this off-set angle as a function of time are shown for comparison in Figure 3.6. On the other hand, a rather lower correlation is shown with the derivative of the gravity angle, $d\phi_g/dt$, as a function of time (79% of linear correlation). The rest of the PCs don’t show significant correlations with known control variables.

3.4 Obtaining the Independent Components

According to the results in the last section, we are now ready to extract the independent components using the presented ICA algorithms for the white light curves. In the present section, extractions are made by EFICA, WASOBI and MULTI-COMBI separately, in order to assess the performance of each algorithm.

As was stated also in the past subsection, only the first seven Principal Components where passed as inputs to the algorithms. To do this, we performed standard PCA to our matrix $\mathbf{X}$, and let only the first seven rows of the resulting matrix $\mathbf{Z} = \mathbf{V}\mathbf{X}$ be passed as inputs for the ICA algorithms (note that in each ICA algorithm a whitening transformation is also applied).

3.4.1 Obtaining the ICs with the EFICA algorithm

The resulting independent components obtained by the EFICA algorithm are shown in Figure 3.7, which are ordered in terms of the ISR vector: the first component corresponds to the corresponding element of the vector with the lowest value and the last component to the corresponding element with the largest value (i.e. they are ordered in terms of “how well they where separated”).

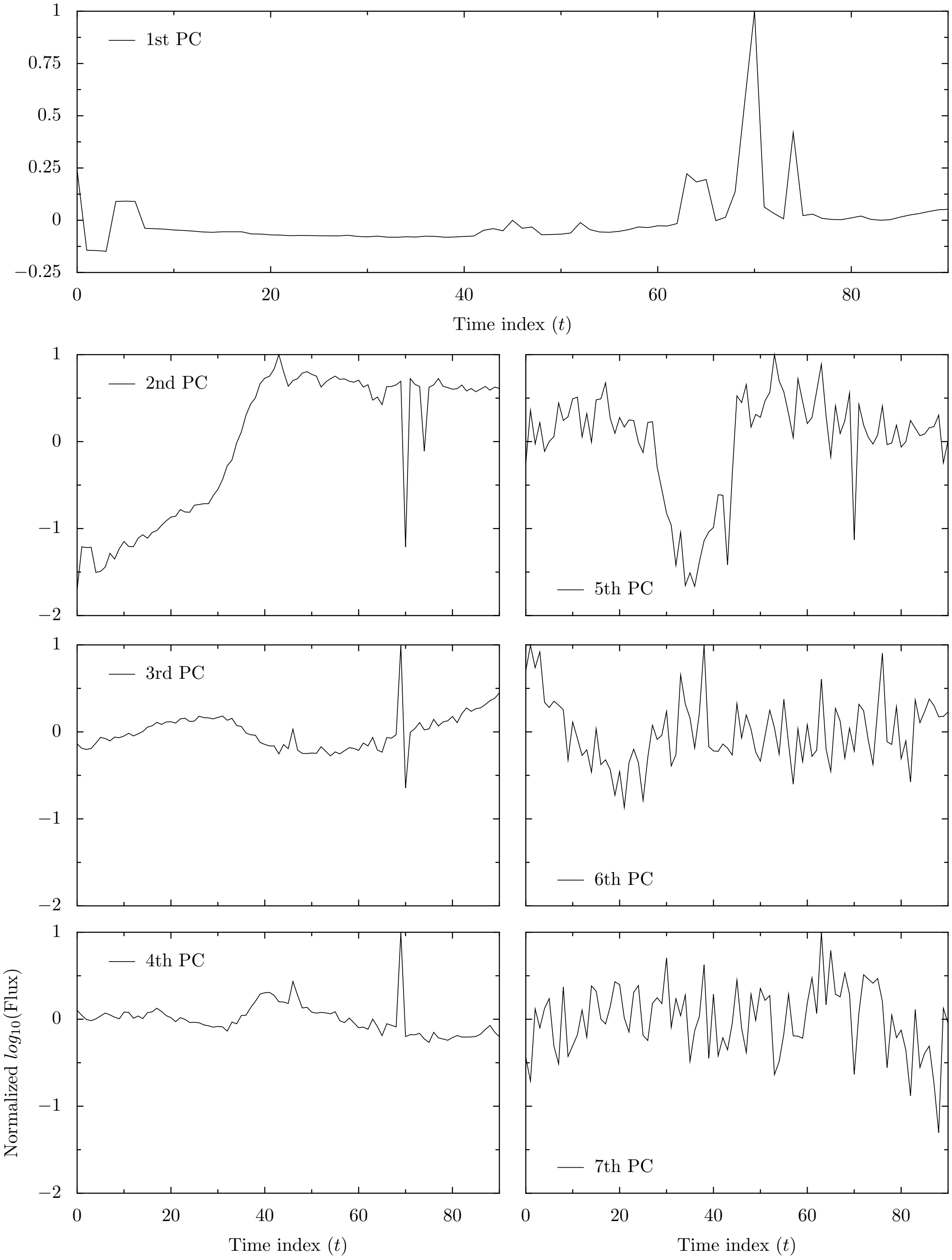

Figure 3.4: The first 7 principal components. Note that the first component has the overall structure of the signals (compare with Figure 3.1).

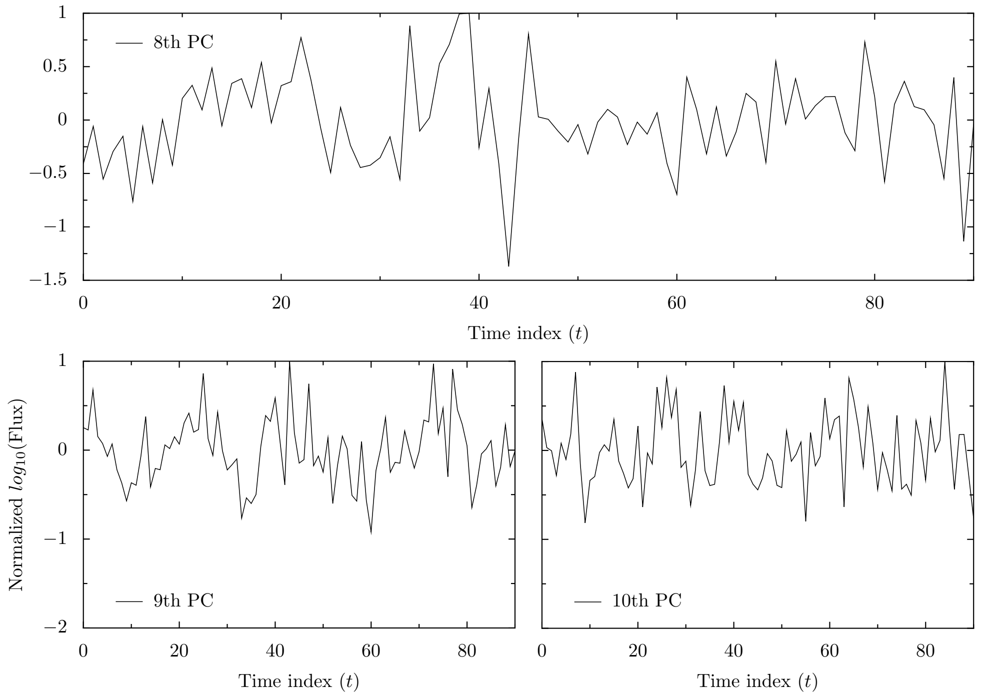

Figure 3.5: Last three rejected principal components. Note that the structure of the signals has no apparent time structure.

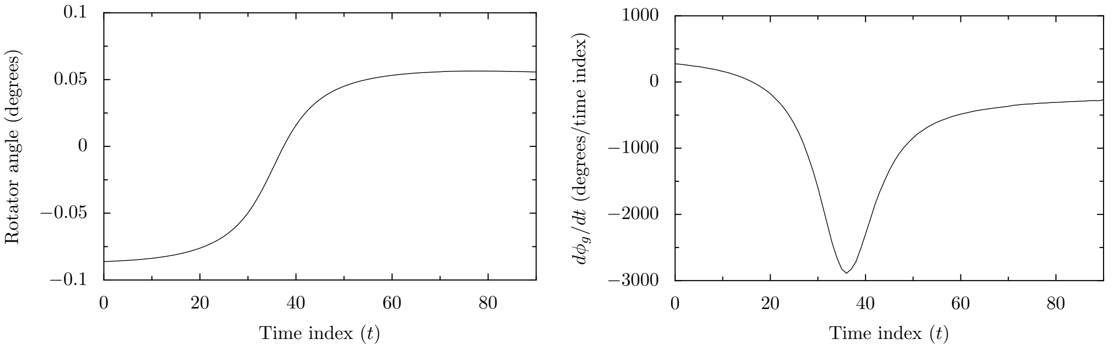

Figure 3.6: (Left) Rotator off-set angle as a function of the time index. (Right) Derivative of the gravity angle, $d\phi_g/dt$, as a function of time. This derivative for the rotator angle has a similar shape.

It is interesting to note that the first three components still have the sharp correlated peaks that we saw on the PCA decomposition of the signal. This is expected, because there isn't enough information to separate that peak from the data: its location in time is clearly visible, but EFICA doesn't use the time structure of the data to separate the components. Instead, it uses the time series as if it where collapsed

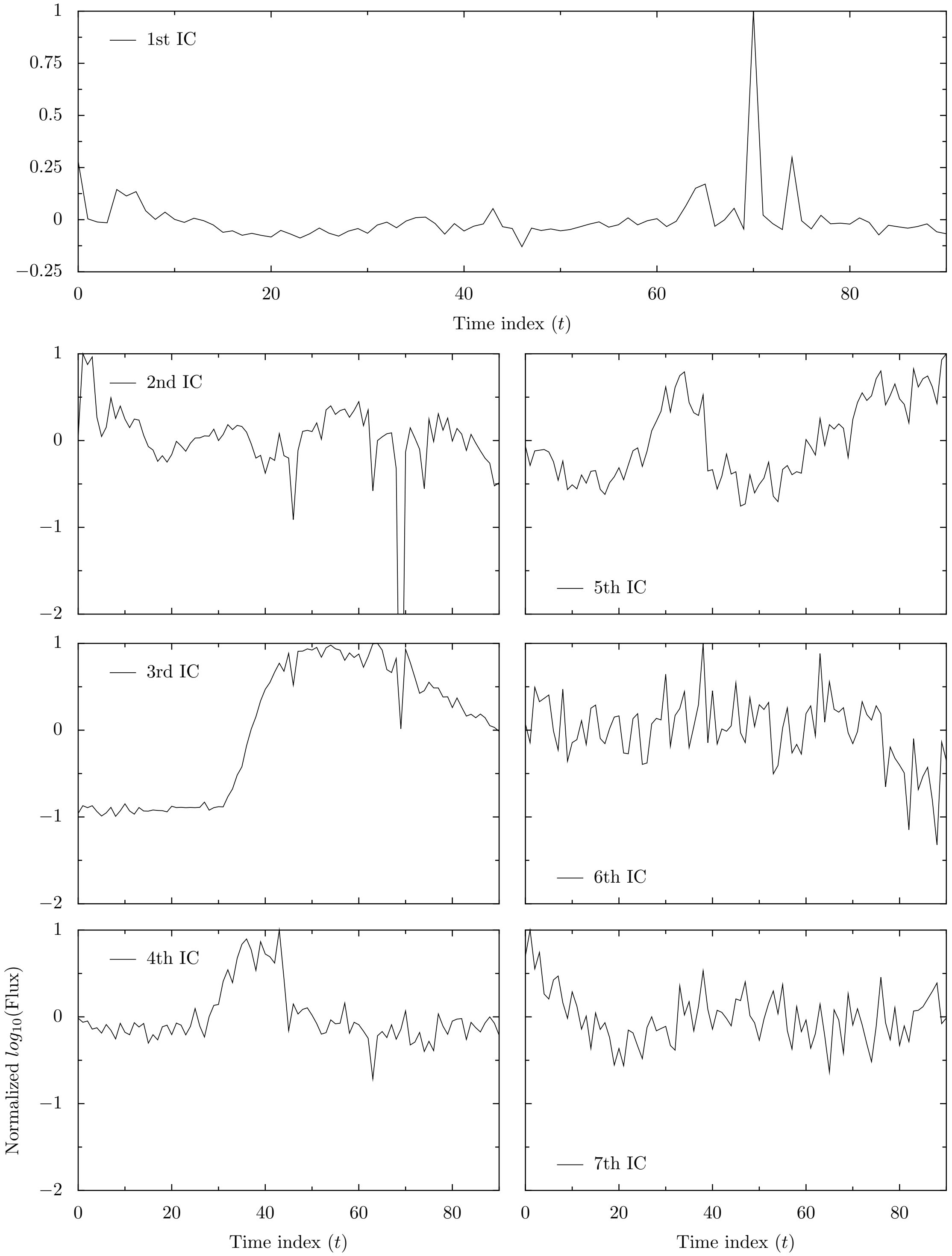

Figure 3.7: The seven independent components obtained by the EFICA algorithm.

in a single random variable. Considering this, intuitively we can't really tell if this spike corresponds to just one component, or if it is a weighted sum of spikes present in various components. Another interesting feature about the obtained components is the shape of the third one. This clearly resembles the rotator off-set angle again. This angle as a function of time was plotted in Figure 3.6. When compared to this third independent component, we find again a (linear) correlation coefficient of 93%.

The first and second components appear to get the overall features of the atmospheric variations (which are perhaps mixed with some non-obvious functions of other instrumental systematic errors), while the fourth component has a very suggestive shape. This component correlates (linearly) well with both the derivative of the rotator angle and the derivative of the gravity angle, $d\phi_g/dt$, as a function of time (correlation coefficients are 84% and 83% respectively). The derivative of the gravity angle as a function of time was also plotted in Figure 3.6. Despite the apparent similarity between the 4th and 5th independent components (they do not show significant correlation), the latter doesn't show any significant correlation. However, the former shows also a small correlation coefficient with airmass (61% of linear correlation coefficient).

Finally, the last two independent components are particularly hard to interpet. We haven't found linear correlations between known instrumental parameters, so we suspect that these represent different atmospheric features as the ones found on the second and third independent components, non-obvious correlations arising from instrumental systematic errors or simply correlated noise. In order to test these posibilities, we obtained the (sample) partial autocorrelation functions $\tilde{\rho}(\tau)$, hoping to find some illuminating pattern that may help us interpret these components. Figure 3.8 shows both, the sample partial autocorrelation function (PACF), $\tilde{\rho}(\tau)$, for the 6th and 7th independent components.

As can be seen from those plots, the PACF for both components show a decaying pattern, as expected. However, a closer look at those plots shows a diference between the 6th and the 7th component: the 6th seems to have a "better memory" than the 7th, whose PACF decays faster to zero (note that the 6th component has a long-range correlation at $\tau \sim 18$). The slowly decay rate of the PACF of the 7th component may indicate a good fit with an order 1 or 2 AR process. In contrast,

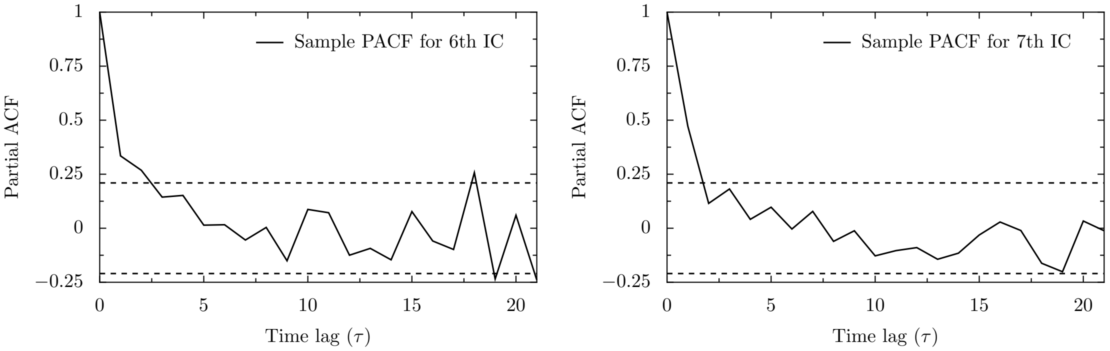

Figure 3.8: (Left) Sample partial autocorrelation function for the 6th independent component. (Right) Sample partial autocorrelation function for the 7th independent component.

the relatively slow decay of the 6th component and the corresponding long-range correlation make it hard to interpret.

The apparent stochastic nature (given by the PACF) of these last components present on the outputs of the EFICA algorithm, motivates then the usage of the WASOBI algorithm in order to characterize these (and possibly others) stochastic time series present on our data.

3.4.2 Obtaining the ICs with the WASOBI algorithm

In order to obtain the ICs with the AR-model version of WASOBI , further analyisis on the structure of the time series must be done. This is because as explained on Chapter 2, the AR-version of WASOBI (Tichavský et al., 2006) assumes that the underlying process is a sum of different and independent AR procceses and, therefore, we must estimate the maximum AR order that we think is present on the time series. This estimation is not crucial: it is used in order to set an upper bound to the AR orders present in our data. In order to estimate this upper bound, then, we computed the partial autocorrelation function for every uncorrelated component present on our matrix **Z**, and searched for the maximum lag that produced a deviation from zero taking in consideration the 95% confidence interval of the PACF. Figure 3.9 shows this plot for several time lags.

From the PACF plot it can be seen that at lag $\tau = 21$ one signal clearly deviates

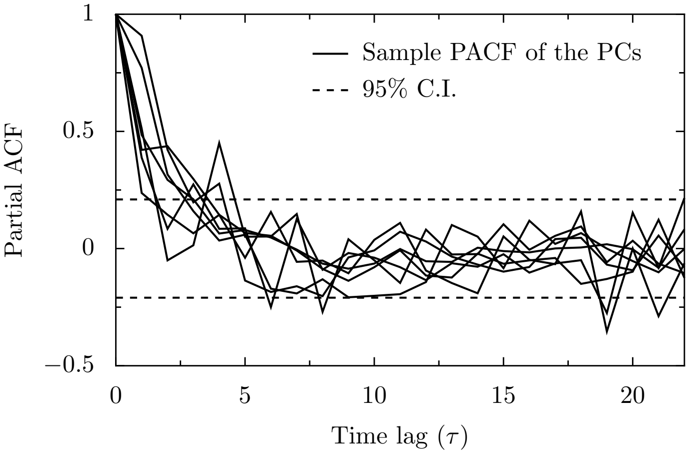

Figure 3.9: Sample partial autocorrelation function for the 7 principal components obtained after applying standard PCA to the ligh curves of our comparison stars.

from zero. Therefore, we'll assume that the largest AR order that is present in our data is 21. The resulting signals after performing the WASOBI algorithm on our principal components are shown in Figure 3.10, where the ICs are ordered again in terms of their ISR vector values.

By far, the most interesting improvement of WASOBI is its ability in removing clear time correlated signals such as the spikes at $t \sim 70$, which was a clear problem for EFICA. Now the major contribution for these spikes are given by the 1st component, and some remaining artifacts by the 5th and 6th components.

The rotator component is again present on the 3rd component (93% linear coefficient). However, this component is also apparently present on the 4th component (71% linear coefficient), which we suggest to be an artifact due to overlearning. On the other hand, the airmass appears to be slightly correlated with the 2nd component (74% linear coefficient). The airmass as a function of the time index is plotted in Figure 3.11.

The remaining ICs are hard to interpret. Despite the suggestive form of the 5th and 7th ICs, which may indicate a correlation with the derivative of the gravity angle they don't show significant correlation with this control variable. It is apparent then that all the suggestive information is bounded to a few ICs. We suggest that

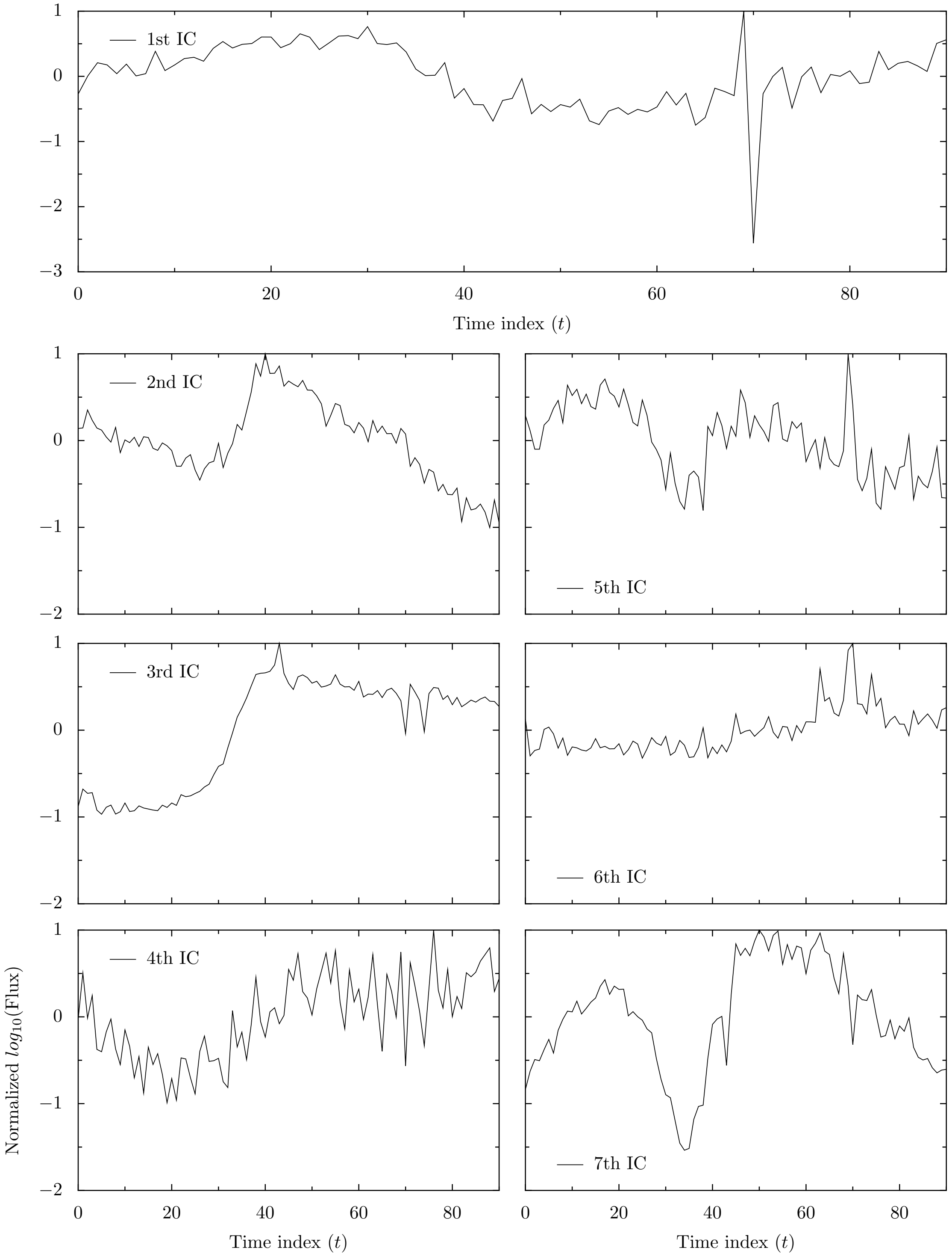

Figure 3.10: The seven independent components obtained by the WASOBI algorithm.

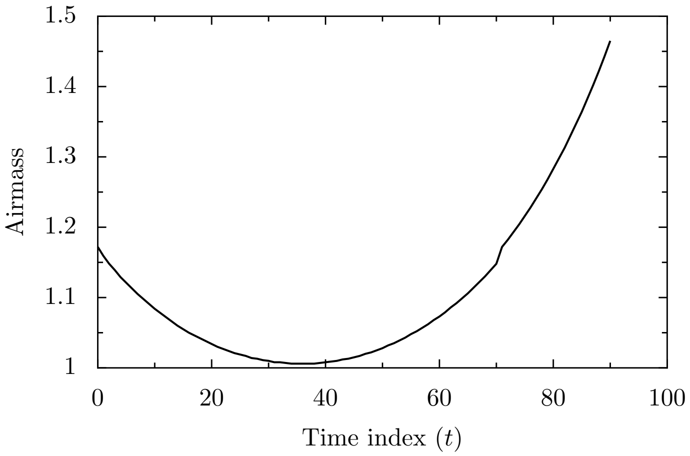

Figure 3.11: Airmass as a function of the time index (to be compared with the second IC in Figure 3.10).

this could, in principle, be a sign of overlearning (due to the compactness of known control variables to a few ICs) or maybe non-obvious correlations.

Despite the apparent loss of information by WASOBI in comparison to EFICA for some instrumental artifacts, this should not be seen as a failure of the algorithm. In fact, as stated above WASOBI could be perfectly obtaning signals that would have been impossible to obtain even knowing a large number of instrumental parameters. Furthermore, as discussed above, the time separation of WASOBI is far superior than EFICA. These results motivate then the usage of the MULTICOMBI algorithm, which combines both of these algorithms as explained in Chapter 2.

3.4.3 Obtaining the ICs with the MULTICOMBI algorithm

The resulting independent components as obtained by the MULTICOMBI algorithm, sorted in the same manner as for the EFICA and WASOBI algorithms are presented in Figure 3.12.

As can be seen from the shapes of these signals, the MULTICOMBI algorithm selected “the best” of each algorithm, as was expected. For example, the first component is again the rotator, which is obviously the one obtained by the EFICA algorithm (again, the correlation coefficient is 93%). However, new components appear too. For example, the 2nd IC correlates slightly with airmass (where the linear

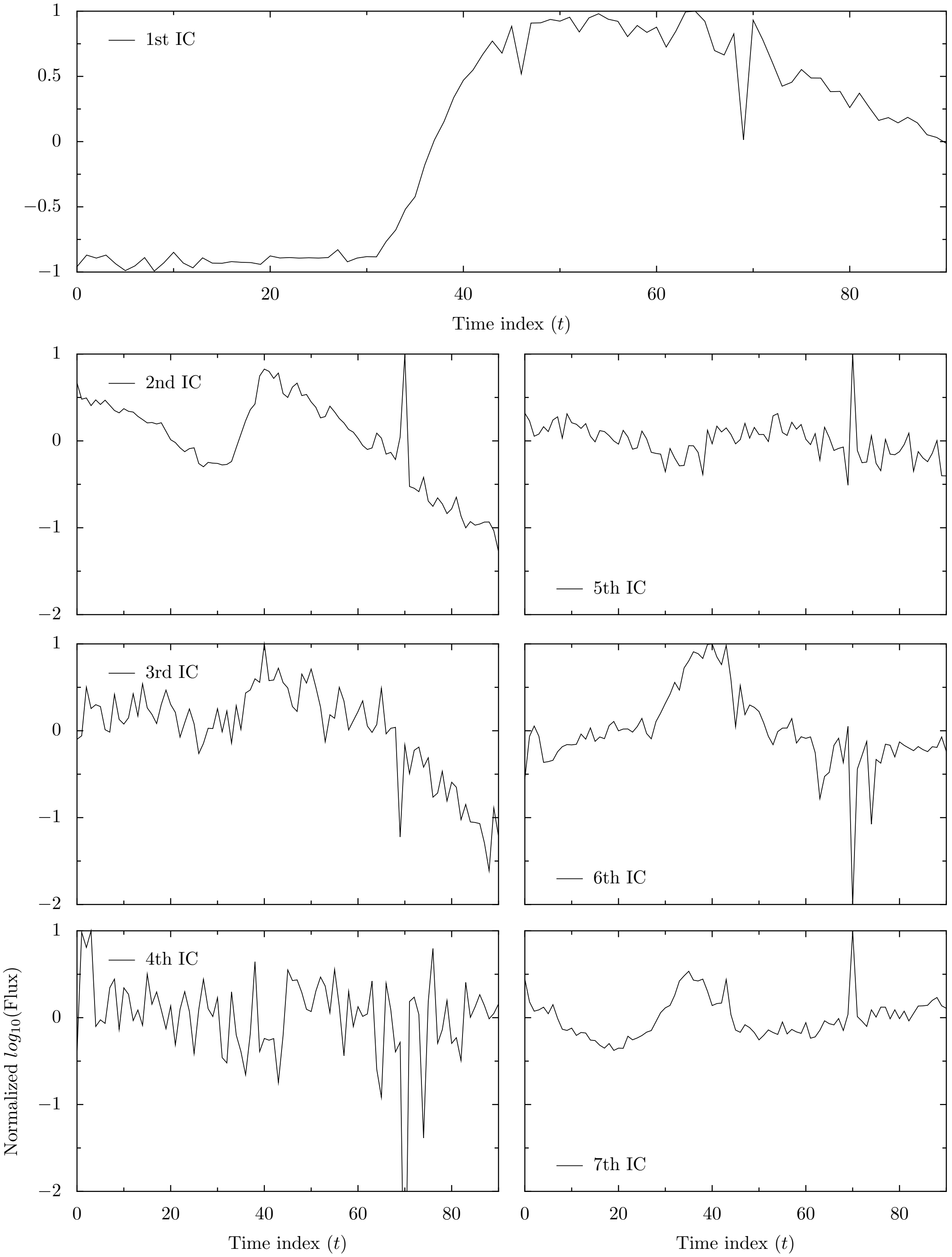

Figure 3.12: The seven independent components obtained by the MULTICOMBI algorithm.

correlation coefficient is 75%) while, on the other hand the 3rd component correlates strongly with it (correlation coefficient is 86%).

The fourth, fifth and seventh components have no obvious correlation with any of our measured variables. On the other hand, the sixth appears to be correlated with the derivative of the gravity angle (correlation coefficient is 75%).

3.5 MCMC fits to the data

In the following section we perform MCMC fits to our data in order to obtain the transit curve for WASP-6. In order to motivate this, we'll first see if we actually can obtain our original lightcurves for our comparison stars back from our estimated source signals.

3.5.1 Preliminaries

Recall that the source signals obtained in each of the ICA algorithms allows us to estimate the mixing matrix $\mathbf{A}$, which in turn has to give us back an estimate of the decorrelated signals present on the $\mathbf{Z}$ matrix:

$$\mathbf{Z} = \mathbf{AS} \tag{3.2}$$

On the other hand, recall that we performed dimensionalty reduction on the matrix $\mathbf{Z}$, where we eliminated the last 3 components. This is equivalent to the obtention of a new matrix, which we'll call $\mathbf{V}_r$, which is the original PCA matrix $\mathbf{V}$ with the last 3 rows removed. Recall that the rows of the PCA transformation matrix where actually orthonormal vectors (the eigenvectors of the covariance matrix), so performing dimensionality reduction is equivalent to the removal of three basis vectors, which ensures that the transformation $\mathbf{V}_r$ is still orthogonal. Therefore, replacing $\mathbf{Z} = \mathbf{V}_r\mathbf{X}$ on equation (3.2), we get:

$$\mathbf{V}_r\mathbf{X} = \mathbf{AS} \Longrightarrow \mathbf{X} = \mathbf{V}_r^T\mathbf{AS} = \mathbf{A}_f\mathbf{S}$$

Where $\mathbf{A}_f = \mathbf{V}_r^T\mathbf{A}$ is the final calculated mixing matrix. Note that each row of $\mathbf{A}_f$ is a (row) vector, $\vec{a}_i^T$, that contains the weight of each source signal (where the first element, $a_{i,1}$ is the weight for the first signal, the second element $a_{i,2}$ the weight for

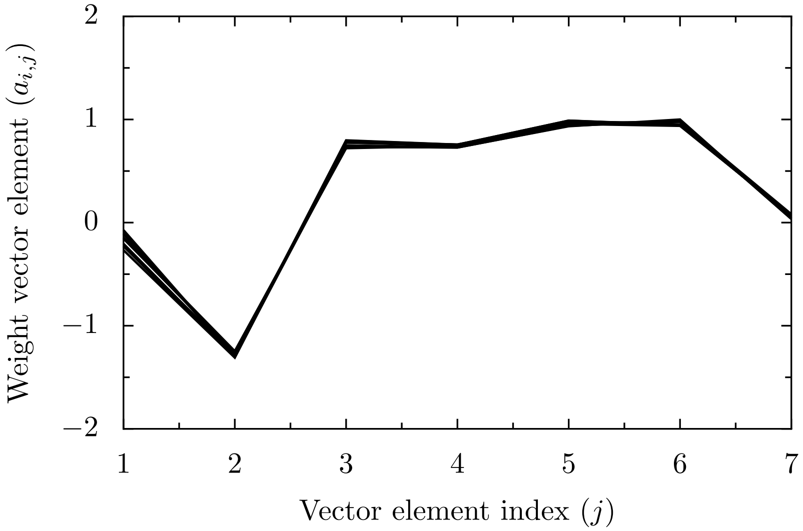

Figure 3.13: Components of the different vectors $\vec{a}_i^T$ that form the weights of the signals of the MULTICOMBI algorithm (Figure 3.12) in order to obtain the observed mixtures $x_i(t)$.

the second, and so on) in order to form the observed mixture $x_i(t)$. These weights for the case of the MULTICOMBI algorithm (whose source signals where shown in Figure 3.12) are shown in Figure 3.13, where the values have been normalized to the highest value of the matrix $\mathbf{A}_f$.

It is interesting to note that the first and seventh independent components are the lowest weighted signals. This suggests that much of the variation of the signal is accounted for by airmass, the gravity and rotator angles and other effects that could not be identified from the instrumental control variables, which we suggest are probably atmospheric fluctuations. On the other hand, despite the clear shape of the rotator angle off-set present in the first component, it appears that the corresponding weight for this signal is not as significant as the others, which suggests that the rotator is not as important as other systematic effects.

We are now ready to perform some MCMC fits in order to recover the transit parameters. However, we'll first need to obtain some uncertainty measures on the weights of our signals in order to constain them. This will be illustrated first by fitting our source signals with the corresponding weights to one of our comparison stars.

3.5.2 MCMC fits to a comparison star

In order to perform the MCMC fits to the comparison stars, we need to put some prior distributions on each of the coefficients of the corresponding weight vectors to be estimated. Assuming that the weights of the source signals for each star do not deviate too much from each other (which is clear in Figure 3.13), the information contained in our mixing matrix $\mathbf{A}_f$ can be used as an estimate of the distribution of the different weights. For each vector index j, then, we calculate the mean and variance of the observed weights by:

$$\begin{aligned}\bar{a}_j &= \sum_{i=1}^{n} a_{i,j}, \\ \hat{\sigma}_j^2 &= \sum_{i=1}^{n} \frac{(a_{i,j}-\bar{a}_j)^2}{n-1},\end{aligned}$$

and then put the prior $a_j \sim U(\bar{a}_j - 5\hat{\sigma}_j, \bar{a}_j + 5\hat{\sigma}_j)$, where $U(a,b)$ denotes the uniform distribution bounded to $a < X < b$, for each weight a_j of the j-th source signal. This prior is a way of letting the weights go "free".

Another technical definition that we need to do is that, in order to perform MCMC, we need to model the likelihood of our data, $\mathcal{L}(D|\vec{\theta}_0)$, where D is the set that contains our data and the vector $\vec{\theta}_0$ is the so-called "parameter vector", and it contains all the parameters to be estimated. The likelihood, then, is the probability density function of obtaining our data set D given the parameter vector $\vec{\theta}_0$. For our applications, if $x(t)$ represents the t-th data point and $\hat{x}(t)$ represents the t-th estimated data point by our model, then $r(t) = x(t) - \hat{x}(t)$ is the residual. With this in mind, assuming that each data point where drawn independently, we'll model the likelihood of each one of them as a normal random variable with mean $r(t)$ and constant variance σ_W^2 i.e.

$$\mathcal{L}(D|\vec{\theta}_0) = \frac{1}{(2\pi\sigma_W^2)^{M/2}} \exp\left(\frac{\sum_{t=1}^{M} r(t)}{2\sigma_W^2}\right),$$

where M is the number of data points (91 in our case). Therefore, our parameter vector $\vec{\theta}_0$ consists not only on the model parameters for our deterministic model of the data ($\vec{a}^T$) but also on the parameter that defines the distribution, σ_W. Therefore, one has to set a prior for this value. We estimate this prior by obtaining the mean and variance of the square root of the 3 last eigenvalues of the eigenvalue decomposition of the covariance matrix of $\mathbf{X}$, in an analogous way as was done for the weights. Then, the prior for the standard deviation is $\sigma_W \sim U(0, \bar{\lambda}^{1/2} + 5\hat{\sigma}_\lambda)$, where $\bar{\lambda}^{1/2}$ is the mean of the square root of the last 3 eigenvalues of the decomposition and $\hat{\sigma}_\lambda^2$ is the estimated variance of the square root of the last three eigenvalues of the eigenvalue decomposition. This is because we are assuming that an underlying white noise process generates the deviation from the residuals, which is given, as an upper bound, by the noisy ICA model. For our measurements, we found that $\bar{\lambda}^{1/2} = 1.61 \times 10^{-4}$ and $\hat{\sigma}_\lambda = 0.20 \times 10^{-4}$, giving then $\sigma_W \sim U(0, 2.61 \times 10^{-4})$.

In order to carry out our fit, as stated earlier, we decided to model the light curve as:

$$\hat{x}(t) = \vec{a}^T \vec{s} = \sum_{j=1}^{n} \hat{a}_j s_j(t),$$

where the $\hat{a}_j$ are the estimated weights (the mean of the distribution of each a_j) and $s_j(t)$ is the i-th source signal. We performed 20.000 iterations of the MCMC algorithm. The results for the signals obtained by the PCA (using matrix $\mathbf{V}_r$ as the analogous of the mixing matrices for the ICA algorithms), EFICA, WASOBI and MULTICOMBI algorithms had very interesting properties. In particular, the EFICA, WASOBI and PCA fits were almost equal (with a slight difference in the mean absolute deviation of the fit of 0.03×10^{-5} between PCA and WASOBI, while practically zero between PCA and EFICA). Because of this, only the results for PCA and MULTICOMBI are shown in Figure 3.14 for illustration purposes.

As can be seen from the Figures, both fits are excellent. However, there seems to be a high peak on the residuals of the signal obtained with MULTICOMBI, which shows a little offset given by the bumps of the original lighcurve at $t = 69$, and therefore the mean square deviation of this fit is higher. However, the partial autocorrelation of the residuals for each method, shown in Figure 3.15, suggests that the PCA residuals are more correlated that MULTICOMBI's (specially at lower lags, where for PCA the first coefficients of the PACF are in the line of the 95% confidence interval). This suggests that, for the purposes of this application, choosing between PCA, EFICA or WASOBI and MULTICOMBI is a trade-off between the absolute deviation of the fits and the ammount of temporal correlation remaining in our data.

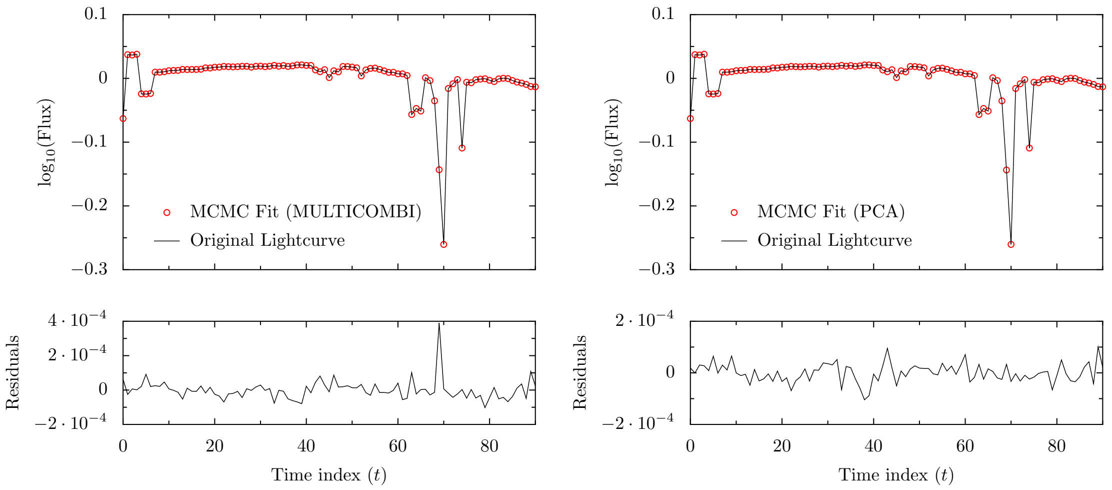

Figure 3.14: (Left) MCMC fit to the comparison star using the priors discussed on this subsection using MULTICOMBI (mean absolute deviation of the fit 3.48×10^{-5}). (Right) MCMC fit to the comparison star using the priors discussed on this subsection using PCA (mean absolute deviation of the fit 2.77×10^{-5}).

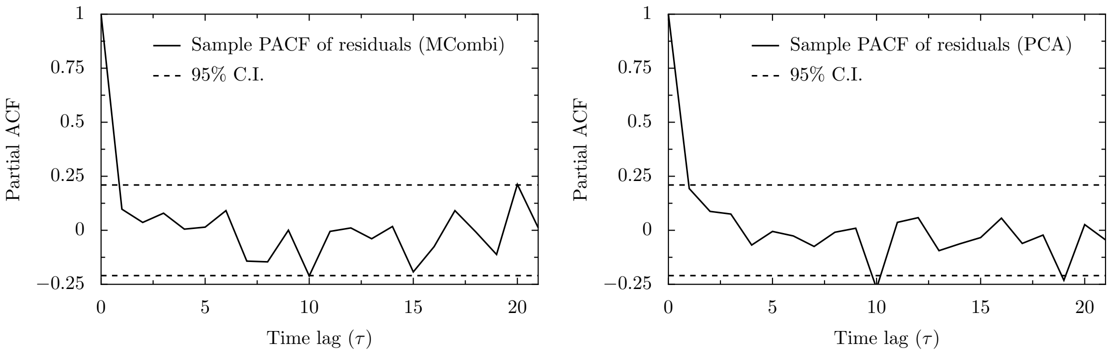

Figure 3.15: Partial autocorrelation functions for the residuals using MULTICOMBI (left) and PCA (right).

Despite the apparent clarity of the above results, it is important to note that we are fitting a light curve that was used to obtain the source signals, which can be seen as “stacking the deck”. In order to test the validity of our approach, then, we performed PCA and ICA on 9 of the 10 comparison stars in the exact same manner as was done here: only the 7 first principal components where retained. Then, we did the exact same MCMC fits in order to fit the left-out comparison star. We found that the residuals where bigger by a factor of 100, i.e., we attained the order of

millimagnitudes in our predictions of the lightcurves.

3.5.3 Finding the transit "white light" curve of WASP-6b

As stated in the introduction of this chapter, we'll now fit transit light curves to our data. As exposed earlier, the light curve model that we'll for WASP-6 is:

$$X(t) = \vec{a}^T\vec{s} + \log_{10}[f(\vec{\theta},t)] + C = \sum_{i=1}^{n} \hat{a}_i s_i(t) + \log_{10}[f(\vec{\theta},t)] + C, \qquad (3.3)$$

where the only difference is the addition of an aditional constant $-1 < C < 1$, which corrects for any bias made when substracting the mean of the objective star. For the priors, the same considerations that we applied to the model of the likelihood of our comparison star apply now to our objective star, with the exception that, in order to be conservative on our fits, we'll let $\sigma_W \sim U(0,1)$. The main reason for this change is because we found that the obtained values of σ_W when we modelled the comparison star in the past subsection were very close to the upper limit that we gave for the original prior.

For our fits, we'll assume that the orbital parameters of the system are known (e.g. from radial velocity measurements), along with a measure of the star's radius and limb darkening coefficients (which can be obtained from stellar models). Therefore, we'll fit only two parameters that characterize the transit light curve: the planet-to-star radius ratio, $p = R_p/R_*$ and the mid-transit time, T_0. The motivations for the fits of these parameters is that p and, therefore, R_p (because R_* is assumed to be known), is a crucial quantity for spectroscopic measurements of exoplanets. As stated in the introduction of this thesis, spectroscopic measurements rely on the precise measurement of this value in order to test for spectral content in the exoplanetary upper atmosphere. T_0, on the other hand, is a unique piece of information in each transit event (where it's variation from transit to transit may indicate, e.g., perturbations by other bodies) and, therefore, its measurement is valuable even if is not used in the study. Because the applications that we have in mind are for transmission spectra, we'll put some large priors on each of these parameters which, except from the limb darkening coefficients which are taken from Kurucz Atlas Stellar Models, are based on the parameters given by Gillon et al. (2009). The parameters and their corresponding prior distributions, along with the measured values for orbital param-

Table 3.1: Parameters to be used in the fits (obtained, except for μ_1 and μ_2, from Gillon et al. (2009)). The limb darkening coefficients were obtained from Kurucz Atlas Stellar Models, as explained in the text.

Parameter	Prior Distribution (or value)	Units
Period, P	3.361006 (fixed)	days
Inclination, i	1.544 (fixed)	radians.
Eccentricity, e	0.054 (fixed)	
Argument of Periapsis, ω	1.7 (fixed)	degrees.
Stellar radius, R_*	0.87 (fixed)	R_{Sun}.
$r_a = R_*/a$	0.0961436507 (fixed[1])	
First limb darkening coefficient, μ_1	0.5049 (fixed)	
Second limb darkening coefficient, μ_2	0.2273 (fixed)	
Mid-transit time T_0	$U(55473.145236, 55473.165236)$	MHJD
Planet-to-star radius ratio, R_p/R_*	$U(0.134637, 0.154637)$	

[1] Here, a stands for the semi-major axis of the planetary orbit.

eters and limb darkening coefficients are presented in Table 3.1.

The reasons why we choose the priors that are shown on Table 3.1 are the following. In their work, p^2 was measured to be $0.02092^{+0.00019}_{-0.00025}$, which implies that[2] $p \approx 0.144637^{+0.00065}_{-0.00086}$. However, as we mentioned, we want to be as conservative as possible in order to test the convergence of our methods. and therefore we let the prior for this value be $p \sim U(0.134637, 0.154637)$ (i.e., we let the parameter go free in a range of 0.1 around its measured mean). The mid-transit time, on the other hand, was measured to be (scaling and propagating the period P to the time of our measurements) $T_0 \approx 55473.155236^{+0.000593}_{-0.000919}$ (in MHJD). Again, however, we set a large prior on this value of $T_0 \sim U(55473.145236, 55473.165236)$ (i.e., we let it go free in a range of 0.1 MHJD about its mean).

The limb darkening coefficients for a quadratic law were obtained from Kurucz Atlas Stellar Models considering the stellar properties of WASP-6 (which where calculated by David Sing using $T_{eff} = 5500$, $log(g) = 4.5$ and M/H=-0.2, private communication). This initially gave us limb darkening coefficients as a function of wavelength, which where interpolated in order to obtain the coefficients for the

[2] Here we have used the delta method to approximate the variances, i.e., $\text{Var}[f(X_1, X_2, ..., X_n)] \approx \sum (\partial f/\partial X_i)^2 \, \text{Var}(X_i)$.

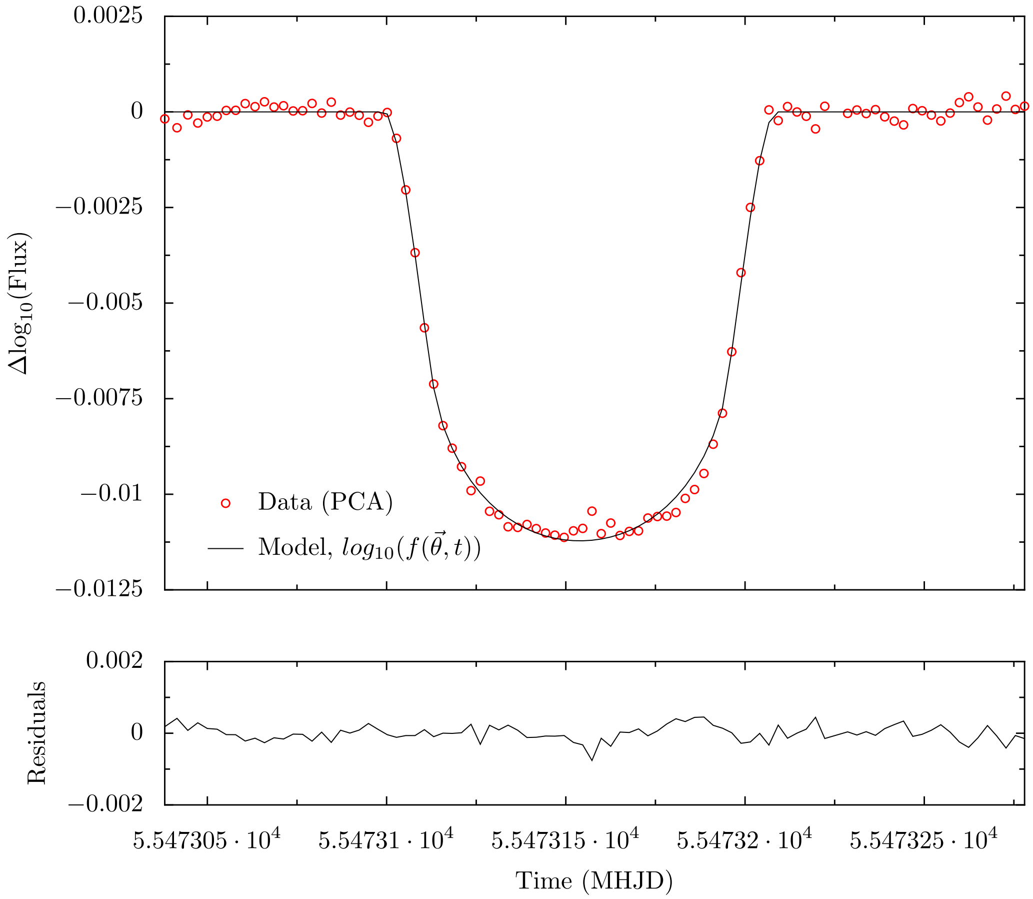

Figure 3.16: Transit white light curve for WASP-6b using PCA. The measured planet-to-star radius according to the MCMC fit was $p = 0.14381^{+0.00074}_{-0.00076}$, which gives us a planetary radius of $R_p = 1.245^{+0.036}_{-0.052}\ R_{Jup}$. Accordingly, the measured mid-transit time according to our fit was $T_0 = 55473.153950^{+0.000144}_{-0.000146}$ MHJD. The measured mean absolute deviation of the residuals is 1.64×10^{-4}, i.e., a precision of 0.410 millimagnitudes.

wavelengths obtained in the spectral measurements of the transit. Finally, a weighted average of them was taken, where each weight corresponds to the flux recieved in each spectral bin.

Figure 3.16 shows the MCMC fit for 100.000 iterations performed using the PCA signals along with the best-fit parameters obtained. Here, the modelled "baseline log-flux", i.e., the terms $\vec{a}^T\vec{s}+C$ in equation (3.3), were substracted from the original lightcurve of WASP-6. As can be seen from the plot, the fit is excellent. However, the residuals appear to have some kind of structure and we suspect they are correlated.

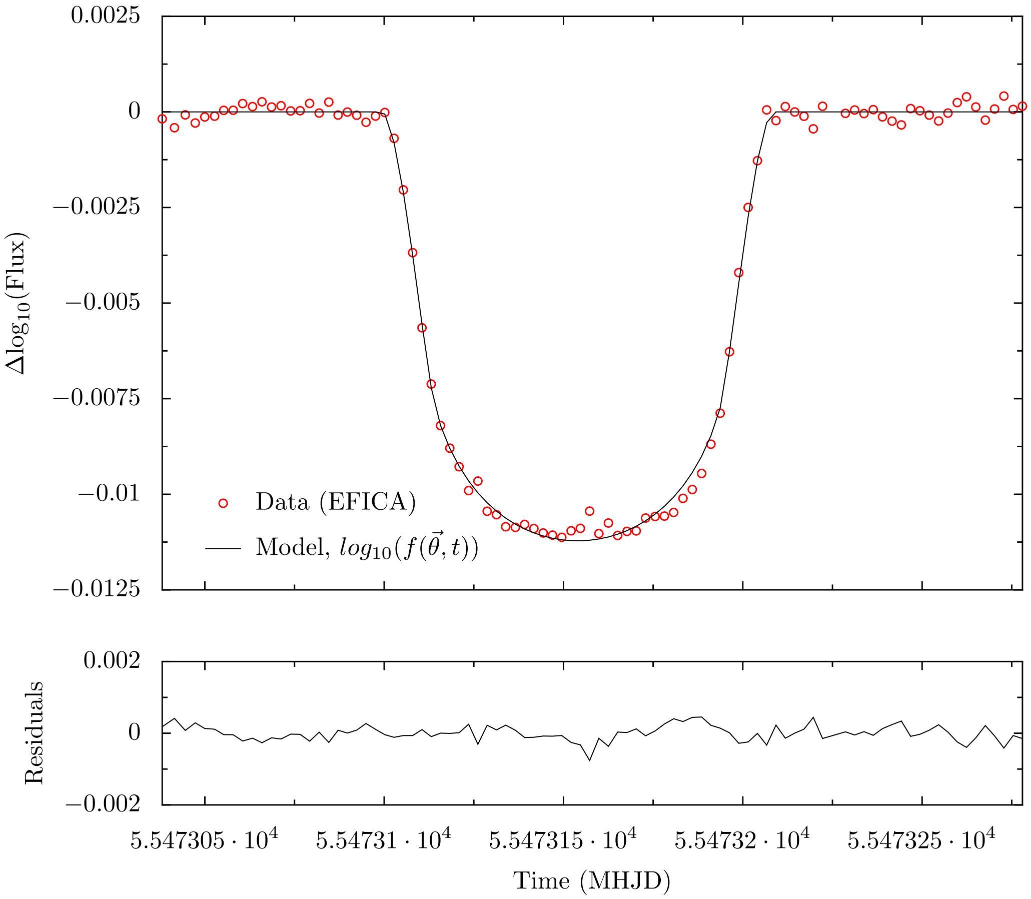

Figure 3.17: Transit white light curve for WASP-6b using EFICA. The measured planet-to-star radius according to the MCMC fit was $p = 0.14379^{+0.00074}_{-0.00071}$, which gives us a planetary radius of $R_p = 1.245^{+0.036}_{-0.052}\ R_{Jup}$. Accordingly, the measured mid-transit time according to our fit was $T_0 = 55473.153950^{+0.000147}_{-0.000139}$ MHJD. The measured mean absolute deviation of the residuals is 1.63×10^{-4}, i.e., a precision of 0.408 millimagnitudes.

For the ICA algorithms we found that the MCMC fits for EFICA, WASOBI and MULTICOMBI where escentially the same. Because of this, only the fit for EFICA is shown in Figure 3.17 for illustration purposes. The fits seem slightly better than PCA's if we take into account the mean absolute deviation but the difference is negligible: overall all the fits are equally good. This is very interesting: all methods converge to almost the same results. Note also that the apparently correlated residuals also appear for the ICA algorithms.

It is also interesting to note that the fits between PCA, EFICA, WASOBI and

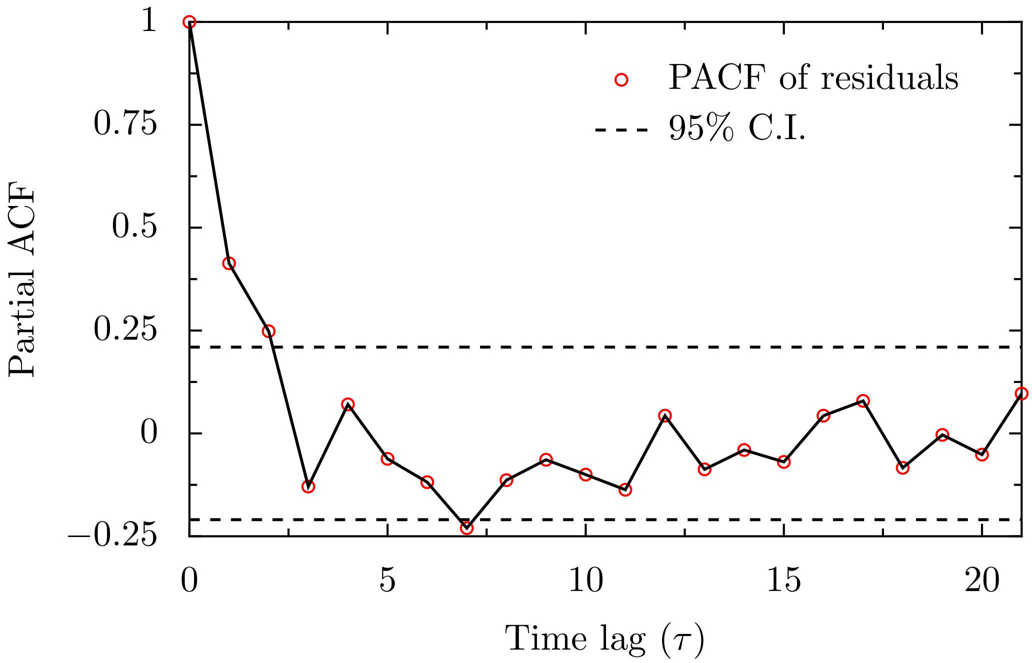

Figure 3.18: Sample PACF for the residuals of the MCMC fits using EFICA and PCA. Both residuals are clearly correlated, showing short-memory correlations.

MULTICOMBI are so similar, that the sample PACF of the residuals for all the used methods is escentially the same. This function is plotted in Figure 3.18 in order to investigate the sugestive correlation that appears in these residuals. The correlation is clearly observed in the PACF and it suggest an underlying short-memory proccess. We believe that the trend is created from residual signals that could not be identified by the algorithms and, therefore, fitted to our model.

Table 3.2 sumarizes our results for the MCMC fits for each algorithm, where the planetary radius can be obtained by propagating the errors obtained for p and for the star's radius, $R_* = 0.870^{+0.025}_{-0.036} R_{Sun}$ (Gillon et al., 2009). Using the delta method, this gives

$$\mathrm{Var}[R_p] \approx R_*^2 \mathrm{Var}[p] + p^2 \mathrm{Var}[R_*],$$

which, for all the algorithms used, gives us a radius for WASP-6b of $R_p = 1.245^{+0.036}_{-0.052}\, R_{Jup}$, which agrees with their measurements whithin the error bounds. As can be seen from our results, the mean value of of the parameters and the corresponding confidence

Table 3.2: Obtained parameters via MCMC fits using different algorithms.

Algorithm	$p = R_p/R_*$	Mid-transit time (T_0)	σ_W
PCA	$0.14381^{+0.00074}_{-0.00076}$	$55473.153950^{+0.000144}_{-0.000146}$	$0.0002293^{+0.0000202}_{-0.0000166}$
EFICA	$0.14379^{+0.00074}_{-0.00071}$	$55473.153950^{+0.000147}_{-0.000139}$	$0.0002289^{+0.0000202}_{-0.0000166}$
WASOBI	$0.14381^{+0.00077}_{-0.00073}$	$55473.153947^{+0.000144}_{-0.000145}$	$0.0002293^{+0.0000204}_{-0.0000162}$
MULTICOMBI	$0.14383^{+0.00073}_{-0.00077}$	$55473.153948^{+0.000143}_{-0.000146}$	$0.0002291^{+0.0000195}_{-0.0000168}$

intervals (68%) obtained by the different algorithms are practically the same. The difference between the algorithm used also changes slightly the confidence intervals for each parameter, but all of them are within the error bounds.

Although the obtained means and confidence intervals for both, p and T_0 are clearly different from the measurements of Gillon et al. (2009), our parameters agree within the error bounds at the 95% confidence interval. The posterior joint PDF for our parameters obtained from our MCMC chains (which for all algorithms are escentially equal) compared with the measurements made by Gillon et al. are plotted in Figure 3.19.

As a final note on this section, it is interesting to note that the estimations for the parameter σ_W with different algorithms are within the upper bounds that we obtained in our priors used for the comparison star in the past subsection ($\sigma_W \sim U(0, 0.000261)$). This suggests that the noisy ICA model that we introduced in order to perform dimensionality reduction on the principal components was indeed a relatively good aproximation.

3.6 Discussion of the results

According to our obtained results, both PCA and ICA performed very well obtaining the principal features of our comparison stars and, in particular, obtaining the transit light curve of the exoplanet WASP-6b. What we actually did, then, was to model the flux of our objective star based on certain “common features” observed

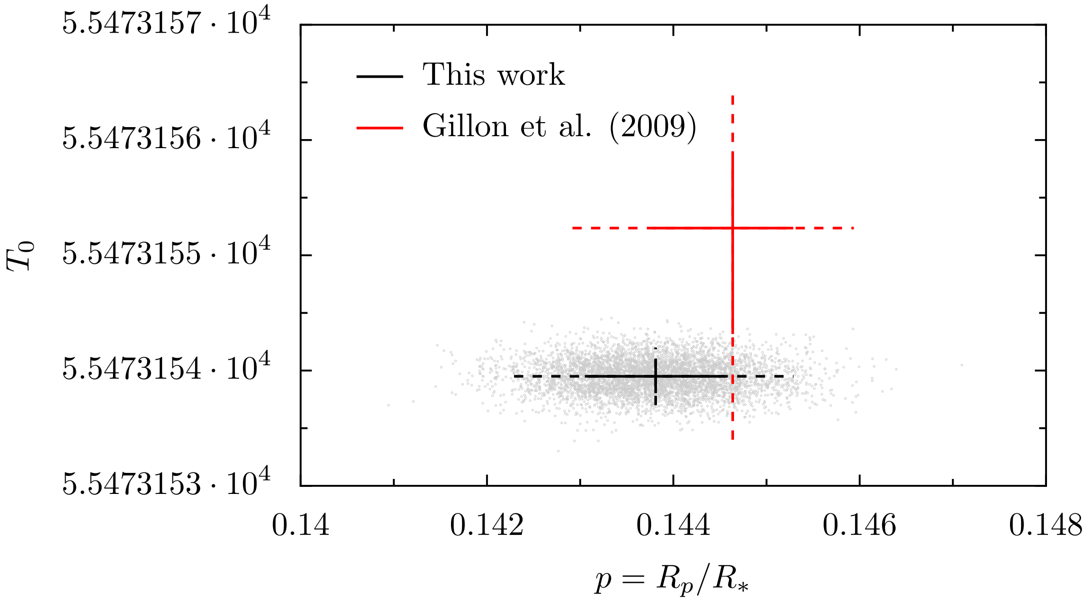

Figure 3.19: Joint probability density functions of the parameters p and T_0. The 68% confidence intervals (thick line) and 95% confidence intervals (dashed lines) are shown for both, this work and the work of Gillon et al. (2009). The grey points represent 5.000 samples of our MCMC chains.

in our comparison stars.

The most striking features of PCA and ICA was their ability to find certain physical signatures of different control variables in order to understand the sources of multiplicative noise addeded to our data. Furthermore, ICA excelled in this "feature extraction" part of the data analysis, finding strong correlations with variables that weren't apparent with the Principal Components.

As we saw, the observed transit light curves attain almost the same precision, with ICA apparently being slightly better than PCA. This means that, for this particular application, the assumption of uncorrelatedness of the underlying physical procceses seem as good as the assumption of independence. However, we experimented that this precision is asymptotically attained if a sufficiently large number of iterations on the MCMC chains are made. Figure 3.20 shows the ratio of the mean residuals obtained by fitting our transit light curve with PCA and MULTICOMBI ranging from 10.000 iterations to 50.000 iterations for the MCMC chains. This was repeated 10 times in order to obtain the mean of the mean absolute deviations of

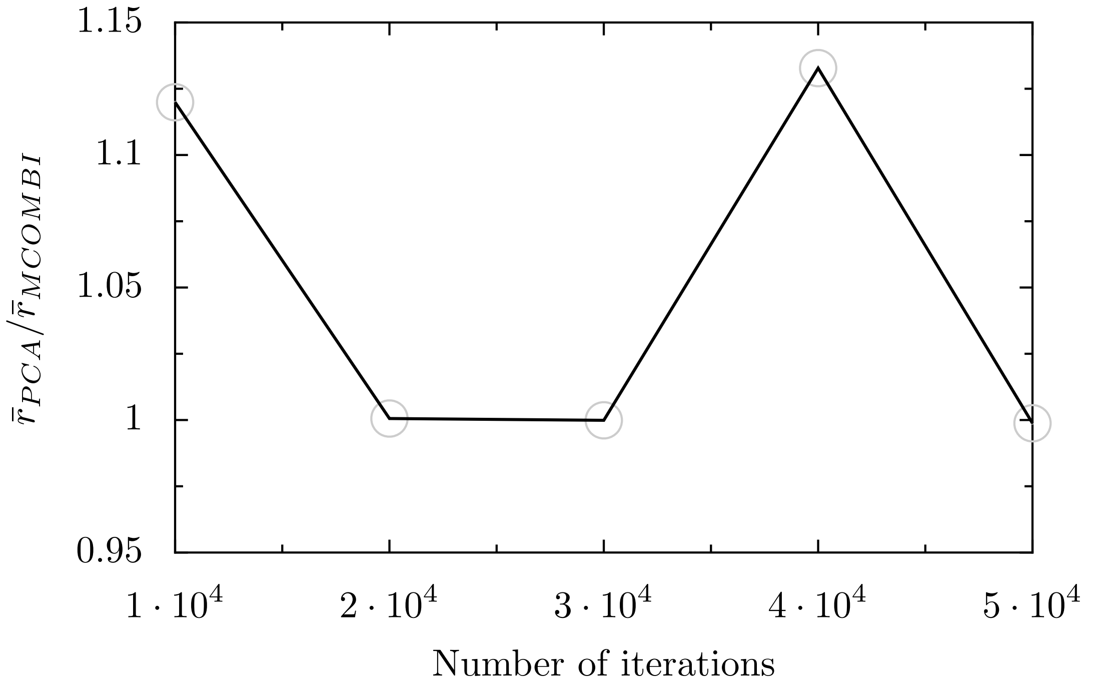

Figure 3.20: Ratio of the mean of the residuals (mean absolute deviation) of 10 iterations for 10.000, 20.000, 30.000, 40.000 and 50.000 iterations of an MCMC chain using separately the PCA and MULTICOMBI obtained signals, in order to fit our transit lightcurve.

the residuals for each number of iterations. The plot shows that, on average, ICA is always better than PCA. Our interpretation for this is that using independent components to model the flux of the objective star is "easier" than trying it with uncorrelated components. This can be seen from the fact that the ICA algorithms clearly "disentagle" important known sources of multiplicative noise such as the airmass, whereas for PCA this disentaglement was not obvious (probably because these sources were "buried" or mixed in various principal components).

Finally, we emphatize the fact that our measurements of the parameters p and T_0 agree with the work of Gillon et al. (2009) at the 95% confidence interval with the used techniques. With this, the obtained planetary radius for WASP-6b would be $R_p = 1.245^{+0.036}_{-0.052}\ R_{Jup}$, which'll make PCA and ICA excellent and promissing statistical analysis tools for transmission spectroscopy, allowing precise measurements of the planetary radius with a precision of a $0.8\% < 1\%$ at the 68% confidence interval. The main advantage of these algorithms is that they only make one physically plausible assumption: the sources of noise in our measurements are uncorrelated (PCA) and/or independent (ICA).

Conclusions

According to our results with the white light curves of the transiting exoplanet WASP-6b, the implemented algorithms that performed PCA and ICA worked very well. The comparison of the obtained physical parameters of WASP-6b with our analysis with known measurements by Gillon et al. (2009) are very good and whithin the 95% confidence interval. However, we also find differences in our mean values and confidence intervals, which are narrower than Gillon's work. This may suggest an improvement to the ephemeris of WASP-6b. Other improvements to this ephemeris has also been recently proposed by Dragomir et al. (2011), who obtained photometric measurements of the transit of WASP-6b and also found differences with the means and confidence intervals obtained by Gillon et al. Future work will study this issue.

The PCA and ICA algorithms were specially good in identifying the sources of systematic noise present in our measurements, and apparently obtained much (but not all) of the variation of our light curves. However, ICA was superior in this "feature extraction" procedure, showing signals that correlated very well with known control variables (such as airmass) that where not obviously obtained with PCA. If we assume that the results of our algorithms aren't biased, we could say that both algorithms where capable of obtaining high precision measurements of the radius of the exoplanet WASP-6b, which we measured to be $R_p = 1.245^{+0.036}_{-0.052}\ R_{Jup}$, attaining a precision of a $0.8\% < 1\%$ at the 68% confidence interval. This makes PCA and ICA very promising tools for measurements of transmission spectra, where this order of precision is needed.

Despite of the above stated results, ICA and PCA couldn't account for all the correlated variation in our transit lightcurves, as was shown on the PACF of the residuals of the fits. There are short-memory procceses that suggest that we are missing some components and, therefore, losing precision in our fits. We attribute this to two principal reasons. The first reason is that, just like in PCA we only retrieved some of the Principal Components, in ICA this also needs to be done in order to obtain better results. The reason is that even if we choose a number of PCs in order to obtain the sources that belong to the signal subspace, this is still a subspace of uncorrelated signals. In terms of information theory, some of the mixtures present on different principal components may have non-linear correlations that makes them redundant. When we apply ICA to these redundant signals, ICA will probably separate them from the "real" source signals, therefore leaving a subset of signals that appear to us as independent components but that in reality may be just noise arising from this redundancy. This kind of ICA dimensionality reduction can be done in analogous (although more complicated) forms to the cross-validations techniques applied for PCA (see, e.g., Westad & Kermit, 2003) and will be investigated in the near future.

The second reason, and perhaps the most important one which could aid in attaining a better precision is because we assumed (wide sense) **stationarity** in the signals for the development of the ICA algorithms, which is a hard assumption to work on given the conditions of ground based data. The spikes that emerge from (possibly) clouds like the ones in our light curves for $t \sim 70$ are hard to model by autoregressive models, because they need high orders to be modelled and, therefore, to be separated from the rest of the signals. A solution to this problem that has been widely used in econometrics to model bumps and spikes could be to use some of the recently developed ICA algorithms for non-stationary time series, such as the Time Varying ICA (TVICA, Chen et al., 2011). TVICA analyses time series in blocks which are detected in terms of different sections of the time series that behave differently. Therefore, this seems the perfect idea for the atmospheric problems that usually arise in astrophysical time series. Another possible solution would be to model the signals with different models that maybe don't have the property of being maximum entropy rate procceses, but that may have a better performance for our needs. The Autoregressive Moving Average (ARMA) models may seem good choices to predict the slow trends present in our lightcurves, while non-stationary time series models such as ARIMA, ARCH or GARCH models would be optimal in the prediction of the aparent volatilities present in our time series (e.g. clouds).

It should be noted that in order to use the algorithms and analysis tools presented in this thesis for medium to low signal-to-noise ratio measurements some initial time-filtering, such as wavelet shrinkage, has to be applied (see, e.g., Abramovich et al., 2000). This is because we have to deal with random fluctuations first in order to "recover" the real trends that form our time series. Another important fact to have in mind is that the kind of analysis to be made is strongly dependant on the application and the underlying physical procceses. For example, if we were working with periodic signals, a better choice would be to perform PCA or ICA in the Fourier space (note that the ICA model still applies because the fourier transform is a linear operation). This observation suggests that another solution to our observed problems regarding ground-based time series measurements is to use a basis that is both localized in time and frequency, in order to take care of fast and slow fluctuations at the same time. Perhaps, then, performing a wavelet transform and doing PCA or ICA on the wavelet domain would give better results. This topic will also be investigated in the near future.

Finally, we would like to note that the analysis tools presented in this thesis are by no means useful only in the area of exoplanetary transits nor only in astrophysics. The applications are very general and not only useful in the analysis of time series but in any indexed series of measurements (recall that, in fact, PCA and EFICA were not initially thought to be used for time series).

Bibliography

Abramovich, F., Bailey, T. & Sepatinas, T., (2000), The Statistician 49, Part 1, pp. 1-29.

Belouchrani, A., Abed-Meraim, K., Cardoso, J.F., Moulines, E., (1997), IEE Transactions on Signal Processing 45, 2, pp. 434-444.

Boshnakov, G., Iqelan, B., (2010), Research Report No. 4, Probability and Statistic Group School of Mathematics, The University of Manchester.

Box, G., Jenkins, G., (1976) *Time series analysis. Forecasting and control* (Revised Edition), Holden-Day.

Brault, J.W., White, O. R., (1971), Astron. & Astrophys., 13, pp. 169-189.

Brockwell, P. & Davis, R. (2001) *Introduction to Time Series and Forecasting*, Second Edition, Springer.

Broersen, P., (2006) *Automatic Autocorrelation and Spectral Analysis*, Springer-Verlag.

Brown, T., (2001), The Astrophysical Journal, 553, pp. 1006-1026.

Cangelosi, R. & Goriely, A., (2007), Biology Direct, 2, pp. 2-27.

Carter, J. & Winn, J., (2009), The Astrophysical Journal, 704, 1, pp. 51-67.

Charbonneau, D., Brown, T. M. Latham, D.W. & Mayor, M., (2000), The Astrophysical Journal, 529, pp. L45-L48.

Chen, R.B., Chen, Y. & Hrdle, W. (2011), SFB 649 Discussion Paper, 54, pp. 1-26.

Cherry, E., (1953), The Journal of The Acoustical Society of America, Vol. 25, N 2, pp. 975-979.

Choi, B.S. & Cover, T. M., (1984), Proc. of the IEEE, Vol. 72, 8, pp. 1094-1095.

Comon, P., (1994), Signal Processing, Vol. 36, pp. 287-314.

Cover, T. & Thomas, J., (1991) *Elements of Information Theory*, John Wiley & Sons, Inc.

Diana, G. & Tommasi, C., (2002), Statistical Methods & Applications, 11, pp. 71-82.

Dragomir, D., Kane, S., Pilyavsky, G., Mahadevan, S., Ciardi, D., Gazak, J. Z., Gelino, D., Payne, A., Rabus, M., Ramirez, S., von Braun, K., Wright, J. T., Wyatt, P., (2011), The Astronomical Journal, 142, 11, pp. 115-124.

Ford, E., (2005), The Astronomical Journal, 129, pp. 1706-1717.

Fortney, J. J., Shabram, M., Showman, A.P., Lian, Y., Freedman, R.S., Marley, M.S. & Lewis, N.K., (2010), The Astrophysical Journal, 709, pp. 1396-1406.

Fuller, W., (1996) *Introduction to statistical time series*, John Wiley & Sons.

Gibson, N., Aigrain, S., Roberts, S., Evans, M., Osborne, M, Pont, F., (2011), Mont. Not. R. Astron. Soc., 419, 2.

Gillon, M., Anderson, D. R., Triaud, A. H. M. J., Hellier, C., Maxted, P. F. L., Pollaco, D., Queloz, D., Smalley, B., West, R. G., Wilson, D. M., Bentley, S. J., Collier Cameron, A., Enoch, B., Hebb, L., Horne, K., Irwin, J., Joshi, Y. C., Lister, T. A., Mayor, M., Pepe, F., Parley, N., Segransan, D., Udry, S., Wheatley, P. J., (2009), Astronomy and Astrophysics, 501, 2, pp. 785-792.

Goldstein, L., (2009), American Mathematical Monthly 166, 1, pp. 45-60.

Gregory, P., (2005), *Bayesian Logical Data Analysis for the Physical Sciences*, Cambridge University Press

Heyde, C.C., (1963), Journal of the Royal Statistical Society, Series B, 29, 392393.

Hyvrinen, A., (1998), Advances in Neural Information Processing Systems, 10, pp. 273-279.

Hyvrinen, A. & Oja, E., (1997), Neural Computation, 9(7), pp. 1483-1492.

Hyvrinen, A. & Oja, E., (1999), Neural Networks, 10(3), pp. 626-634.

Hyvrinen, A. & Oja, E., (2000), Neural Networks, 13(4-5), pp. 411-430.

Hyvrinen, A., Karhunen, J. & Oja, E., (2000), *Independent Component Analysis*, John Wiley & Sons, Inc.

Jaynes, E.T., (1957), The Physical Review, Vol. 106, N 4, pp. 620-630.

Jaynes, E.T., (1968), IEEE Transactions On Systems Science and Cybernetics, Vol. sec-4, N 3, pp. 227-241.

Jolliffe, I. T., (2002) *Principal Component Analysis* (Second Edition), New York: Springer

Kaltenegger, L., Traub, W. A. (2009), The Astrophysical Journal, 698, 1, pp. 519-527.

Koen, C. & Lombard, F., (1993), Mon. Not. R. Astron. Soc., 263, pp. 287-308.

Koldovsky, Z., Tichavsky, P. & Oja, E. (2006), Neural Networks, Vol. 17(5), pp. 1265-1277.

Mandel, K., Agol, E. (2002), The Astrophysical Journal, 580, 2, pp. L171-L175.

Mazeh, T. & Tamuz, O. (2007), Transiting Extrasolar Planets Workshop ASP Conference Series, 366, pp. 119-126.

Papoulis, A., (1991) *Probability, Random Variables, and Stochastic Processes* (Third Edition), McGraw-Hill, Inc.

Percival, (1993), The American Statistician, 47, 4, pp. 274-276.

Pont, F., Zucker, S. & Queloz, D. (2006), Mont. Not. R. Astron. Soc., 373, pp. 231-242.

Press, W.H., (1978), Comments Astrophys., 7, 103.

Rajagopal, A.K. & Sudarshan, E.C.G., (1974), Phys. Review A., Vol. 10, 5, pp. 1852-1857.

Ramsay, J. O. & Silverman, B.W., (1997) *Functional Data Analysis*, New York: Springer

Ramsay, J. O. & Silverman, B.W., (2002) *Applied Functional Data Analysis: Methods and Case Studies*, New York: Springer

Scargle, J.D., (1981), The Astrophysical Journal Sup. S., 45, pp. 1-71.

Seager, S., (2010), *Exoplanet Atmospheres*, Princeton Series in Astrophysics.

Seager, S., Sasselov, D. D., (2000), The Astrophysical Journal, 537, pp. 916-921.

Shaman, P. & Stine, R., (1988), Journal of the American Statistical Association, 83, 403, pp. 842- 848.

Shannon, C.E., (1948), The Bell System Technical Journal, Vol. 27, pp. 379423, 623-656.

Shao, J., (2003) *Mathematical Statistics*, Springer.

Thatte, A., Deroo, P. & Swain, R., (2010), A&A, 523, A35.

Tichavsky, P., Doron, E., Yeredor, A., Nielsen, J., (2006), Proc. EUSIPCO-2006, Florence, Italy, 2006.

Tichavsky, P., Koldovsky, Z., Yeredor, A., Gmez-Herrero, G., Doron, E.,, (2008), IEE Transactions on Neural Networks, 19, 3, pp. 421-430.

Waldmann, I., (2011), eprint arXiv:1106.1989.

Westad, F. & Kermit, M., (2003), Analytica Chimica Acta 490, pp. 341-354.

Winn, J., (2010), eprint arXiv:1001.2010.

Yeredor, A., (2000), IEE Signal Processing Letters 7, 7, pp. 197-200.

Ziehe, A., Laskov, P., Nolte, G., Mller, K.R., (2004), Journal of Machine Learning Research, 5, pp. 777-800