Of 'Cocktail Parties' and Exoplanets

Data analysis in exoplanetary spectroscopy

Ingo Peter Waldmann

Submitted for the degree of Doctor of Philosophy

Department of Physics and Astronomy

University College London

March 2012

I, *Ingo Peter Waldmann*, confirm that the work presented in this thesis is my own.

Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Abstract

The field of transiting extrasolar planets and especially the study of their atmospheres is one of the youngest and most dynamic subjects in current astrophysics. To study the atmospheres of those foreign worlds, we typically require a 10^{-4} to 10^{-5} level of accuracy in flux. Currently available instruments were not designed with these precisions in mind. Calibrating an instrument without knowing its response function at the required level has become the central challenge of exoplanetary spectroscopy. A variety of parametric correction models are used in the literature. These show high degeneracies between the scientific result and the instrument correction used. Hence, an unbiased analysis of the data at the 10^{-4} level of accuracy is difficult and the cause of much controversy in the field.

In this thesis, I present three novel ways of de-trending exoplanetary data non-parametrically, i.e. without requiring auxiliary or prior information of the instrument or data. This removes correctional bias. These techniques are based on:

1) unsupervised machine-learning algorithms (Chapter 3) to de-convolve non-Gaussian signals, i.e. the systematic noise, from the desired astrophysical feature. Such a 'blind' signal de-mixing is commonly known as the 'Cocktail Party problem' in signal-processing. I demonstrate its capabilities using spectroscopic *Hubble*/NICMOS measurements of the hot-Jupiters HD189733b and XO1b and demonstrate the removal of stellar noise in *Kepler* photometry (Chapter 4).

2) Fourier/Wavelet based self-filtering algorithms based on the concepts of sparsity of the exoplanetary signal in the frequency domain (Chapter 5). The robustness of this method is demonstrated for very low signal-to-noise conditions using four nights of ground-based observations of the secondary eclipse of HD189733b in Chapter 6. Here I unambiguously confirm the detection of a strong non-LTE methane emission in the L-band and can test for residual telluric contamination using this method.

3) an Independent-component-analysis supported wavelet masking of multivariate data, which extends the non-parametric machine learning to Gaussian noise dominated data applications, Chapter 7.

In the light of ever increasing data analysis challenges, as we probe ever smaller signals and fainter targets, techniques such as the ones presented in this thesis are paramount to the success of exoplanetary characterisation in the future.

Acknowledgements

There are a great many people that deserve my gratitude and thanks. First and foremost I would like to thank my parents who have shown me the greatest measure of support and love that one could possibly ask for. Without your dedication and moral support I would have never made it this far. It was those many trips to the *Deutsches Museum* and the building of our small vehicles in our garage that have truly set the path to this thesis. To let me chose my future freely and to always support me with whatever I do is something extraordinary and I am forever thankful.

I also owe an immeasurable dept to my first supervisor, Giovanna Tinetti. I feel incredibly privileged to work with someone as talented as Dr Tinetti. The support you have given me is, simply put, enormous and the length of this thesis would not suffice to express my gratitude. It was, is and will be an absolute pleasure to work with you and I am looking forward to many years to come.

To Steve Fossey, my second PhD supervisor and Masters supervisor, I owe a special thanks. After some arduous and frustrating years as undergraduate, it was you with your dedication and passion for the subject, that has shown me the world beyond the quotidian tedium of undergraduate life. This thesis is a direct consequence of your continuous support and dedication.

I very much thank Mark Swain and Pieter Deroo for so patiently teaching me the tools of the trade and I am looking forward to working with you in the future.

I thank my two examiners, Prof. Richard Nelson and Prof. Mike Barlow for taking the time out of their busy schedules to read this work.

And last but not least, I would like to thank all my close friends, you know who you are, and the UCL PhD crowd for making these last years the most enjoyable of my life so far.

to my parents, my family

and friends

thank you.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

> *"Here and there they fancied they saw vast seas, scarcely kept together under so rarefied an atmosphere, and water-courses emptying the mountain tributaries. Leaning over the abyss, they hoped to catch some sounds from that orb forever mute in the solitude of space."*
>
> — **Around the Moon (Jules Verne)**

It is *Around the Moon* (orig. 'Autour de la Lune') by Jules Verne that I so vividly remember reading in my childhood. Written in 1870, it was certainly not the first work depicting the exploration of foreign worlds (for itself was a sequel to the *From the Earth to the Moon* novel written five years earlier), nor was it by any measures the last. Countless books, films and stories address the question of whether we are alone in the vastness of space and if not, which is always the conclusion of any credible science-fiction writer, we ask 'what may happen to us?'. Here, the ideas range from the pleasant encounter of Steven Spielberg's innocuous *E.T.* to the more frightening visions of H. G. Wells in his *War of the Worlds*.

One does not need to look far beyond the cheap thrills of popcorn cinema to find a profundity in the question 'are we alone?' that seems fundamental to the human condition. It not only plays to the curiosity of every child but also promises to provide an answer to age old questions of existentialism. Rather unsurprisingly the idea of habitable worlds besides our own is not a novelty.

Lacking modern instrumentation, early astronomers did not look too far for signs of life and it was commonly believed that our closest neighbours are as habitable as our own world. In fact, much of today's lunar geological nomenclature, introduced by Johannes Kepler, is based on this false assumption of habitability, where laval basins were readily mistaken as vast oceans or *maria*. Some 220 years after Kepler, it was the improved telescope design of Joseph von Frauenhofer

that allowed Johann Heinrich von Mädler in 1834 to be amongst the first to suggest that these *maria* were not at all filled with water, but that the Moon is more akin to a dry, barren place laking an atmosphere.

With the now quickly increasing resolving power of 19^{th} century telescopes, astronomers did not take long to look for life elsewhere. When the suggestion of vegetation was first offered to explain the canal like formations on Mars[1], it could not readily explain their perceived straight and narrow nature. The hypothesis of artificially built irrigation systems by a martian civilisation, based on these *canals*, was quickly accepted and later extended upon by the postulation of agricultural pumping systems to account for their lengths and shallow inclines (Lowell, 1903; Pickering, 1904).

Such an over-enthusiasm may be forgiven for it took little time for the technological revolution of the 20^{th} century to clean up the ideas of the past. Soon astronomers found themselves running short of places to look for life in our own solar system and with the detection of planets around other stars deemed technically impossible, the quest for an answer to one of the most important astronomical questions, had receded back to the works of science-fiction authors.

This remained true until November 1995 when Michael Mayor and Didier Queloz published a paper entitled 'A Jupiter-mass companion to a solar-type star' (Mayor and Queloz, 1995). The data was coherent and the detection significant to anyone's standard, yet their wording was cautious. It was clear that much was at stake and with a pathological reluctance to name the object 'a planet' did our worldview edge from a universe of one known solar-system, ours, to one of many.

Despite its importance, it was not the first extrasolar planet to be discovered, three years earlier Wolszczan and Frail (1992) detected an anomalous timing pattern of a nearby pulsar and concluded the existence of a planetary companion. However, it was largely ignored as an environment too strange to care for, and it was the discovery of 51Peg-b (Mayor and Queloz, 1995) that marked the beginning of the field of extrasolar planet studies.

[1] Astron. and Astro-Physics, 1892, XI, 670

1.1 From hot-Jupiters to Habitable Exo-Earths

With an orbital period of ∼4.23 days and a semi-major axis of 0.052 AU (astronomical units), the orbit of 51Peg-b is much closer to its host-star than anyone had deemed possible. As a comparison, Mercury's semi-major orbital axis lies at ∼0.387 AU, 7.4 times further away than 51Peg-b is from its host star. Given it is about half of the Jupiter mass orbiting its solar-type host, the scientific community began to refer to this object as a hot-Jupiter (see figure 1.1). Soon after the first discovery, systematic efforts yielded more of these curious objects and ∼17 years later, we can count 760 confirmed extrasolar planets (Schneider et al., 2011) and over 2000 candidates by the *Kepler* mission (Borucki et al., 2010, 2011; Tenenbaum et al., 2012). The vast majority of these objects discovered[2] in the first ten years were of the hot-Jupiter variety, generally characterised by their close in, tidally locked orbits, near Jupiter masses with highly inflated, hot and strongly irradiated atmospheres (Schneider et al., 2011). From figure 1.2 we can appreciate the almost exponentially increasing speed with which we discover these foreign worlds.

1.1.1 The search for exoplanets

Several techniques and observing strategies have been devised over the years to detect ever smaller extra-solar planets around ever fainter host stars. The first method was that of radial velocity or doppler shift measurements of extrasolar planets followed by transit surveys, microlensing detections and direct imaging of exoplanets. I will briefly outline these different techniques with a more in-depth treatment of transiting extrasolar planets in section 2.1.

Radial Velocity and Astrometry

51Peg-b and 699 other planets to date were discovered using the radial velocity method. This method is closely related to stellar astrometry technique and both methods depend on the fact that the extrasolar planet and its host star orbit a common barycentre. The orbit of the planet hence induces a measurable periodic motion in the star. Radial velocity and astrometry differ in the line of sight of the observer. In astrometry we observe the slight angular displacement of the host star when viewing the system at a normal to its planetary ecliptic, whereas in the case of a radial velocity measurement we view the system more head on and measure the spectroscopic doppler shift of the host star. Since the mid-1990's spectrographs have the required measurement accuracy of $\sim 20\text{ms}^{-1}$ (Mayor and Queloz, 1995; Marcy and Butler, 1996) to be able to detect Jupiter mass objects in close orbits around their hosts.

[2]statistics taken from: http://exoplanet.eu/

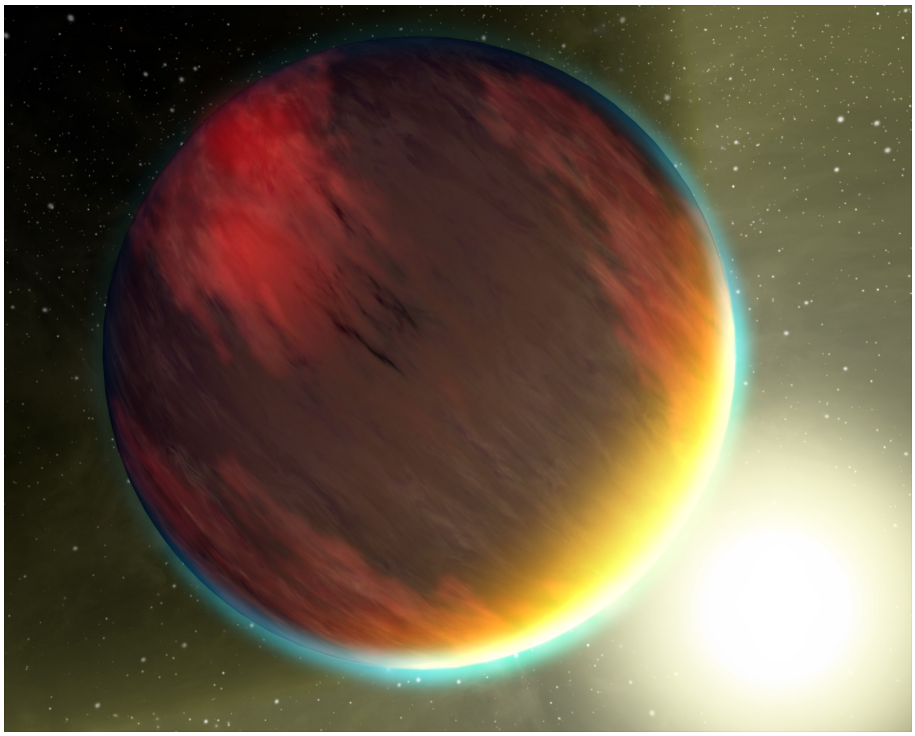


Figure 1.1: Artist impression of a hot Jupiter on its close in orbit around its host star. These objects are highly irradiated and feature strongly inflated atmospheres, typically several thousand Kelvin hot. When these objects transit their host in our line of sight, they make ideal candidates for transmission spectroscopy. *Source: NASA/JPL*

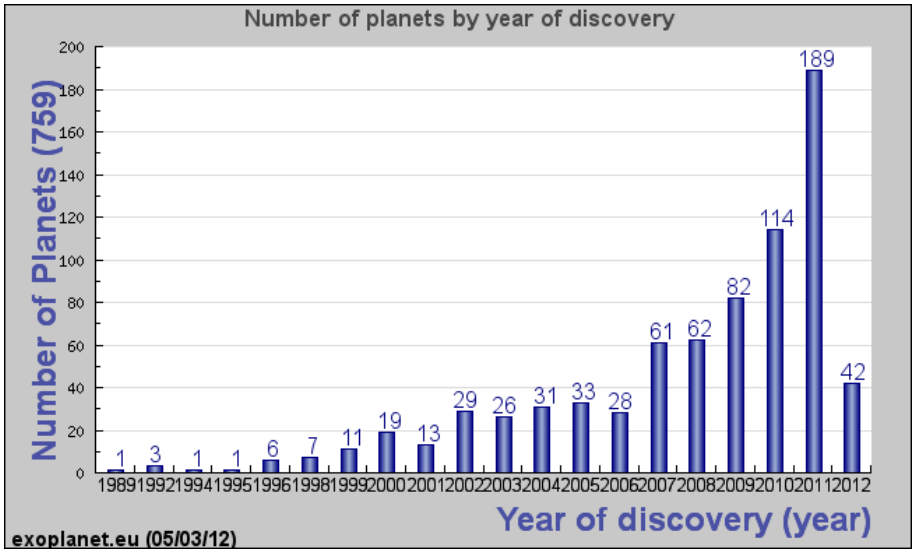


Figure 1.2: Bar chart of extrasolar planet detections published with the onset of dedicated planet search programmes in the early 2000's and the launch of exoplanet dedicated missions, CoRoT (in 2006; Baglin et al., 2006) and Kepler (in 2009; Borucki et al., 2010). *Source: http://exoplanet.eu/*

Figures 1.3 & 1.4 show two radial velocity measurements of HAT-P-7 (Pál, 2008; Winn et al., 2009) and HD80606 (Naef et al., 2001). The former shows a near circular orbit characterised by the sinusoidal nature of the radial velocity curve whilst the latter is one of the most eccentric ($e = 0.934$) orbits of a planet known. One can easily appreciate the very different natures of their orbital motions given the radial velocity curves at hand. We can now directly measure the orbital period, P, the radial velocity amplitude, K, and the eccentricity, e. From these observables we can, based on Kepler's third law of motion and the conservation of momentum, calculate the semi-major axis of the orbit, a,

$$a^3 = \frac{GM_*}{4\pi^2}P^2; \qquad M_P v_P = M_* v_{ast} \tag{1.1}$$

for a a circular orbit with G being the gravitational constant, M_* and M_P the mass of the star and planet respectively. The radial velocity amplitude, K, can no be related to the planetary mass via

$$K_* \simeq v_* \mathrm{sin}i = \frac{M_P \mathrm{sin}i}{M_*} v_P; \qquad v_P = \frac{2\pi a}{P} \tag{1.2}$$

where i the orbital inclination. The determination of M_p assumes a well known stellar mass which can often be obtained using spectroscopic observations and stellar models of the planet's host. It should be noted that with radial velocity alone we have a planetary mass and orbital inclination degeneracy and the mass can only be estimated to a factor of $\sin(i)$, i.e. $M_p \sin(i) < M_p$.

With the development of high precision spectrographs such as HARPS (Pepe et al., 2002; Mayor et al., 2003) on the 3.6m telescope in La Silla (Chile) and its proposed sister instrument on the TNG in La Palma (Spain), it has become possible to extend the discovery space down to Super-Earths, planets in size between Earth and Neptune, the detailed characterisation of multiple planet systems such as the four planets orbiting the M-dwarf GJ581 (Mayor et al., 2009; Gregory, 2011b; Forveille et al., 2011).

The detection of extrasolar planets using astrometry has so far been less successful but the European Space Agency successor to the *Hipparcos* mission, *Gaia*, is expected to yield on the order of 10^5 extrasolar planets with the astrometry technique (Sozzetti, 2010). The *Gaia* mission is currently in the implementation phase.

Transiting exoplanets

When a planet's orbital inclination is close enough to 90 degrees (measured from the normal of the observational plane) to have $(a/R_*)\mathrm{cos}i < 1$, where R_* is the stellar radius, we find the

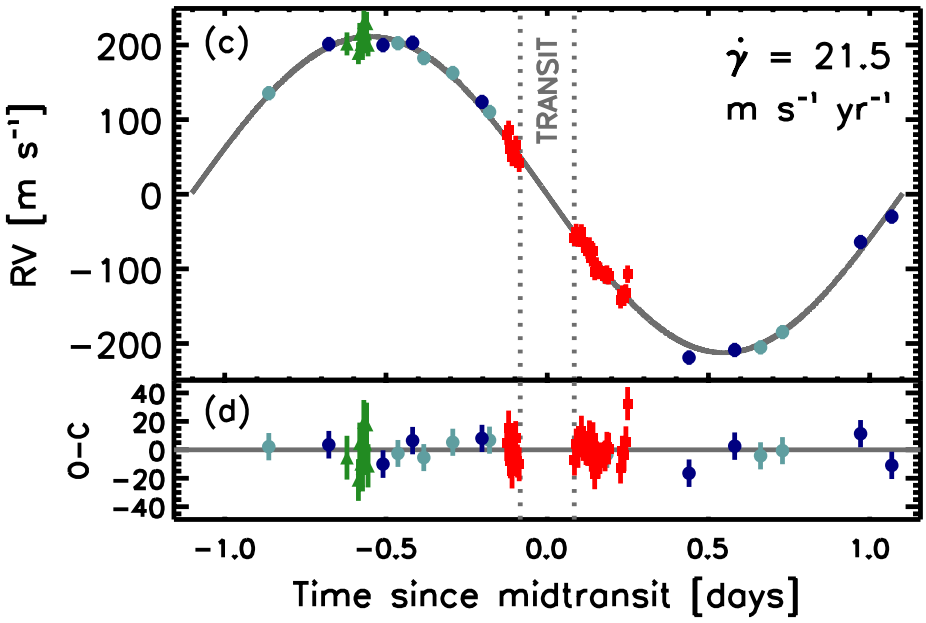


Figure 1.3: Radial velocity curve of HAT-P-7. The sinusoidal shape of the curve characterises a near-circular orbit of its companion planet (Winn et al., 2009).

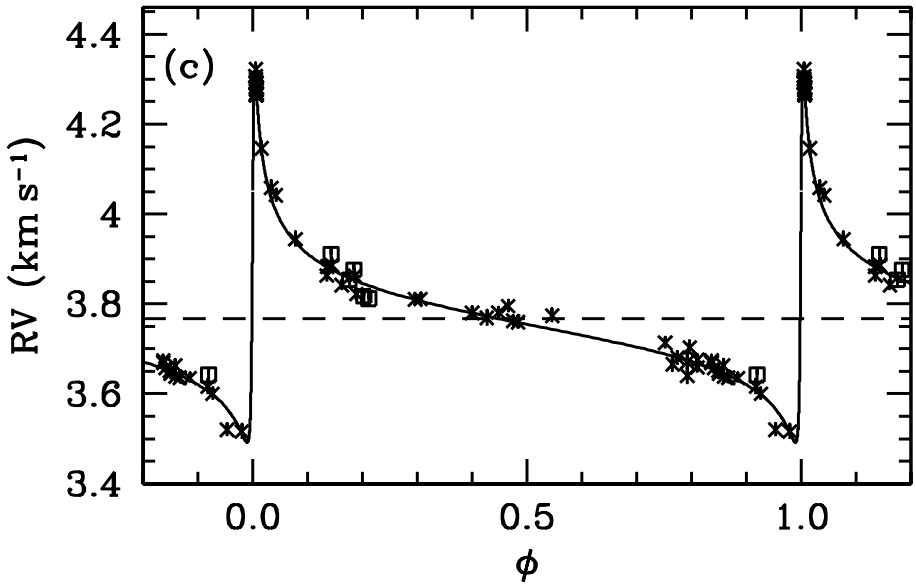


Figure 1.4: Radial velocity curve of HD80606 shows a highly eccentric orbital motion of its hot-Jupiter. Far from the sinusoidal shape of figure 1.3, HD80606b has one of the most eccentric orbits known (Naef et al., 2001).

exoplanet to transit in front of its host star. It was five years after the discovery of 51Pegb that Charbonneau et al. (2000) observed the first eclipse of an extrasolar planet, making it the third planet observed to transit in front of its host star. The other two planets being Mercury and Venus in our own solar system. The transit of HD209458b marked the second revolution in the still very young field of exoplanet astronomy. With the observation of a transit, we could not only solve the $M_p \sin(i)$ degeneracy but also determine the planetary size, the stellar mass and size, as well as the planet's oblateness, albedo and atmospheric absorption and emission signatures. For a detailed discussion of the transiting lightcurve morphology and derivable parameters, please refer to chapter 2.

Given this wealth of information obtainable with transit observations, it is unsurprising that much time and resources have been invested in the detection of these few that happen to transit from our earthly perspective. Despite early difficulties and a low detection rate, many international survey networks were set up to systematically monitor the sky for these rare transiting events. The first generation of these networks followed the approach of ultra-wide field surveys of bright stars using arrays of small and commercially available photographic lenses. Among the most successful of these projects were and are the *SuperWASP* (Pollacco et al., 2006) and the *HATNet* (Bakos et al., 2002, 2004) surveys. Both surveys used Canon 200mm f/1.8 lens for their surveying telescopes and were initially restricted to northern hemisphere observations but have recently expanded to cover the southern night-sky (Lister et al., 2007; Bakos, 2011). Other notable ground-based surveys are *MEarth* (Nutzman and Charbonneau, 2008; Charbonneau et al., 2009), *XO* (McCullough et al., 2005) and *TrES* (Dunham et al., 2004).

The high success of these surveys, 38 and 70 exoplanets discovered by HATNet and SuperWasp respectively (Schneider et al., 2011), showcased an unexpected truth about exoplanet science: one can detect large exoplanets and retrieve their basic parameters with very little in terms of instrumentation. This is true as long as one understands the systematics of their instrument and in fact, the control of systematics is fundamental to all exoplanetary observations. An example of this is the successful detection of the transit of the highly eccentric HD80606b from the University College London Observatory (ULO) in the north London suburbs, using a 14 inch and a 10 inch refractor (Fossey et al., 2009), figure 1.5.

Following the success of these ground-based surveys, the next step was to take the search to the sky with dedicated space missions overcoming the observational limitations imposed by Earth's atmosphere. Launched in late 2006, *CoRoT* (Bordé et al., 2003; Barge et al., 2008) was the first space mission of its kind opening the door to the first detections of transiting

Super-Earths such as CoRoT-7b (Léger et al., 2009; Bruntt et al., 2010). To date, *CoRoT* has discovered 23 transiting exoplanets.

In March 2009, the *Kepler* observatory (Borucki et al., 2010; Caldwell et al., 2010; Jenkins et al., 2010; Koch et al., 2010) was launched into an Earth trailing orbit. Unlike *CoRoT*, *Kepler* 's observing strategy involves the uninterrupted observation of a 115 square-degree patch of the sky near the constellation of Cygnus. Three years into its mission lifetime, *Kepler* has confirmed 72 discoveries and detected over 2300 planetary candidates[3]. Amongst the smallest targets confirmed to date are the three planets of the KOI-961system with KOI-961d being a mere 0.57× Earth-radii (Muirhead et al., 2012) and the Kepler20 system with its five planets ranging from 0.868 (Kepler20f; Fressin et al., 2012) to 3.07 (Kepler20c; Gautier et al., 2011) Earth-radii. A powerful demonstration of the data quality of the *Kepler* mission is given in figure 1.6, an observation of the previously known hot-Jupiter HAT-P-7b. Here, subplot A) shows two primary transits with exquisite signal-to-noise. It is upon zooming into the data (subplot B) that we can make out the much smaller secondary eclipse, when the planet disappears behind the star, half phase between both primaries and the planetary phase curve as the planets day-side rotates into the view of the observer.

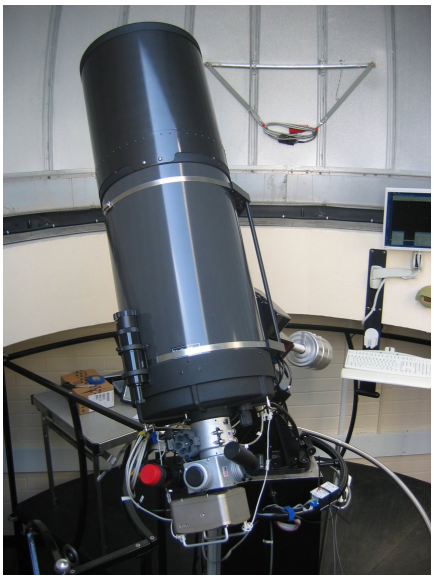
It is astonishing that within one decade, we have progressed from the detection of the first exoplanetary transit to dedicated space missions detecting Earth and Mars sized planets around distant stars in the dozens.

Other detection methods

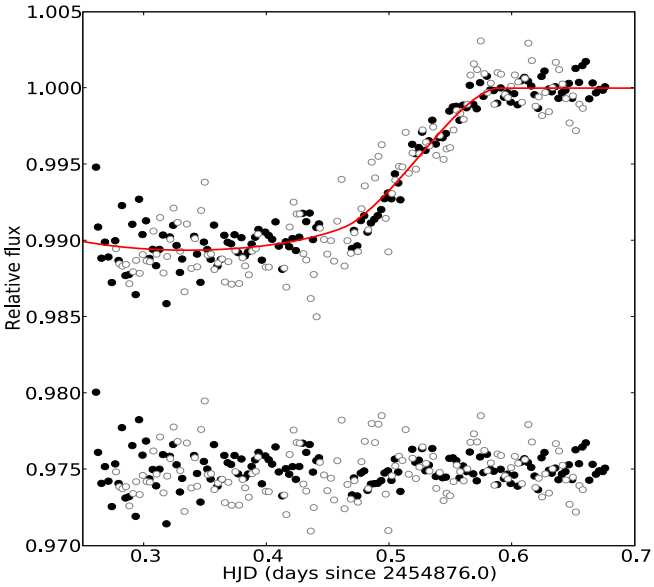
Besides the two very successful planet hunting techniques described above, it is important to mention two more: 1) gravitational microlensing and 2) direct imaging.

Gravitational lenses are a rather common occurrence in astronomy. First postulated by Albert Einstein, they appear whenever the gravitational field of a foreground object distorts the light of a further away luminous object in a way that focuses the distant light on the earthly observer and makes it appear brighter. This effect can be seen for galaxies, stellar clusters or even individual stars. In the case of a perfect super-position of two stars in the line of sight of the observer, we observe a perfect Einstein-Ring. This lensing effect can become perturbed by the gravitational potential of a planetary companion orbiting the foreground star. This perturbation can be observed by monitoring the lensing event over time. The first successful detection of an exoplanet using this method was made by Bond et al. (2004). Despite the transient nature of these events, microlensing finds its importance in exoplanet research by providing important

[3]http://kepler.nasa.gov/Mission/discoveries/



(a) Celestron-14in telescope at the University College London Observatory.



(b) Photometric lightcurve of the egress signal of HD80606b as observed by the C14 telescope (black dots) and a smaller 10-inch Meade telescope (circles)

Figure 1.5: The in-depth study of its instrument systematics allowed us to achieve a photometric accuracy of 2 parts in 1000. This allowed us to successfully detect the egress of the eccentric hot-Jupiter HD80606b (Fossey et al., 2009).

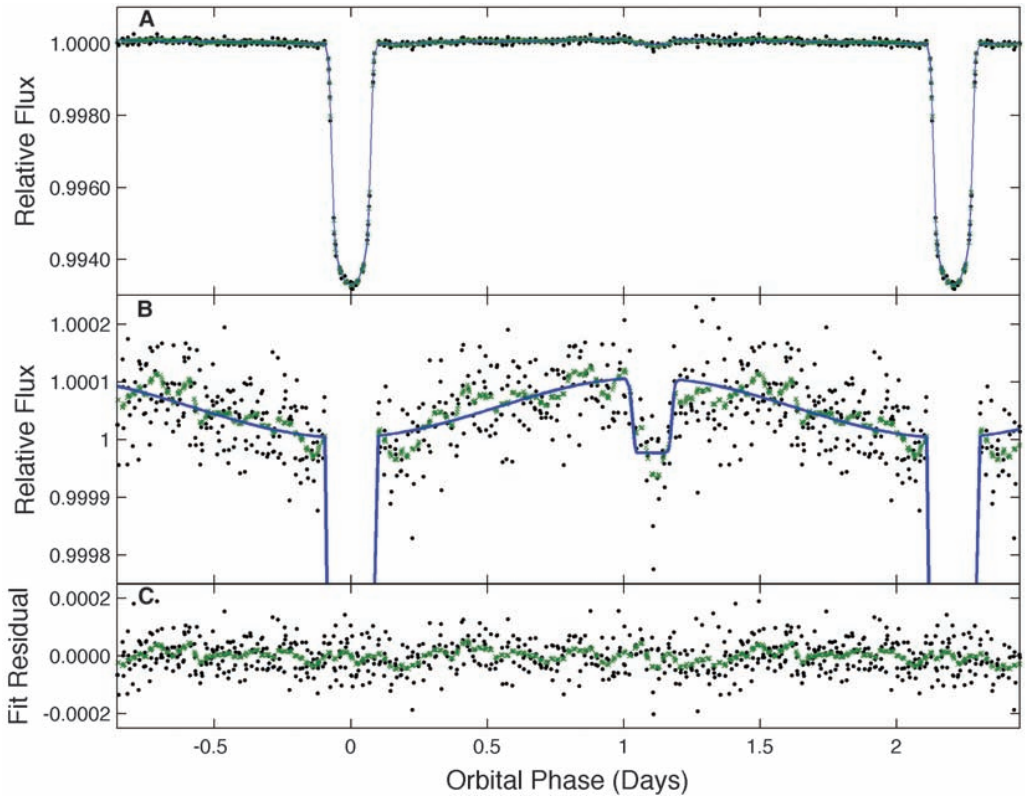


Figure 1.6: A) Lightcurve of two primary eclipses of HAT-P-7b observed by the *Kepler* spacecraft. B) zoomed in version of A), the secondary eclipse as well as the planetary phase curve as its day-side rotates into view are clearly visible. C) the model fit residuals (Borucki et al., 2009).

statistics on frequencies of small mass exoplanets in our galaxy (Beaulieu et al., 2006; Cassan et al., 2012).

Direct imaging of an exoplanet may conceptually be the easiest approach but is likely to be the hardest in practise. Here the challenge is the detection of the reflected and/or emitted light of a large extrasolar planet close to its much brighter parent star. Often the flux differences, or contrast ratios, between both objects are at the order of 10^4 to 10^5 making it difficult to disentangle the stellar light from the planetary flux on an absolute scale. This imaging is often done using sophisticated coronographs. The first successfully imaged planet was presented by Chauvin et al. (2004), with the most famous of these detections being the detection of Formalhaut-b (Stapelfeldt et al., 2004; Kalas et al., 2008).

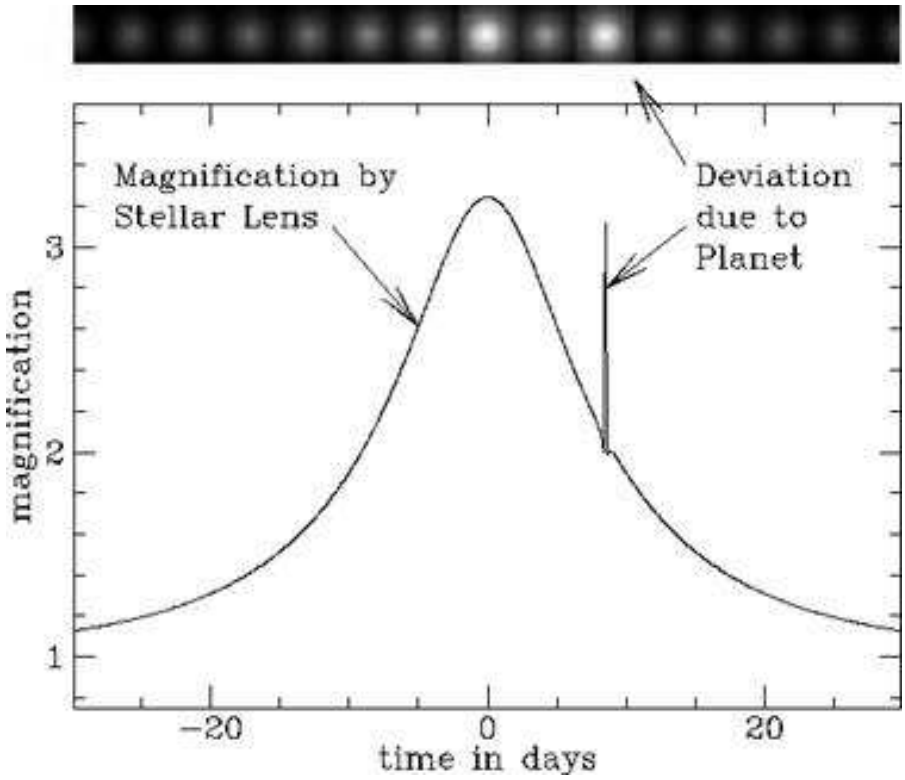


Figure 1.7: A microlensing lightcurve of a foreground star traversing a background source and gravitationally lensing the background star's light in the process. If the foreground star is orbited by a planet, the planet's gravitational field will disturb the lensing effect which can be detected by a sharp and brief flux increase.

1.2 From the detection to the characterisation

In less than two decades from the first detection, we saw the field of extrasolar planets evolve at a pace unmatched by any other in astronomy. A constant sense of urgency seems to permeate the field with today's superlatives being dwarfed by those of tomorrow. With well thought out and organised planet detection systems and networks, we know the orbital parameters, masses and radii of ever more exoplanets allowing us to compare their bulk properties to those planets found in our own solar system. Figure 1.8 shows the bulk compositions of the small mass/radius exoplanets currently known in relation to solar system objects and ice, rock and iron isochrones. From such diagrams, we can to a first order, study the nature of these distant worlds and infer their internal compositions by examining their average densities.

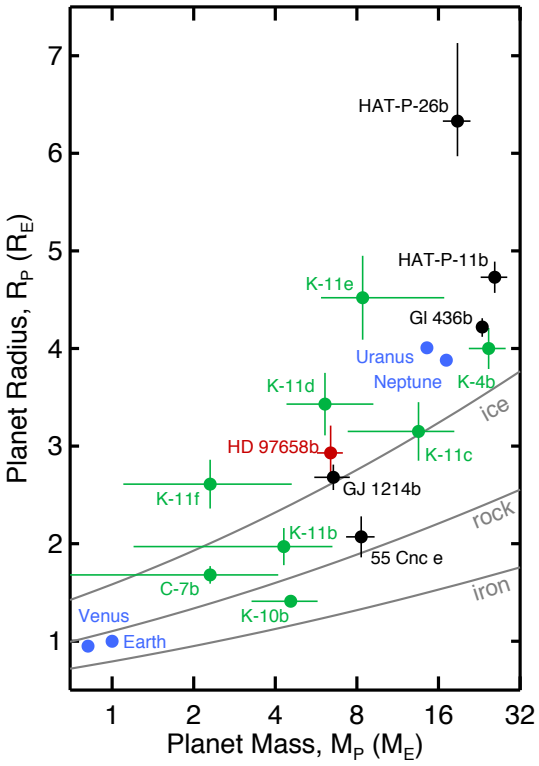


Figure 1.8: *Currently known Earth to Neptune sized extrasolar planets. Black/red: planets detected from the ground; green: detections by Kepler /CoRoT; blue: solar system objects; grey: ice, rock and iron density isochrones (Henry et al., 2011)*

However, this is often the extent of any such study and we know from examples of our own solar system, take Earth and Venus, that their bulk compositions may be near identical but both could not be further apart in terms of providing a habitable environment. In short, there is more to the characterisation of a planet than is to be appreciated by a measure of its mass and radius. One obvious way to further characterise the extrasolar planet at hand and to assert claims of habitability for the smaller Earth and Super-Earth type objects is to study their atmospheres. Here, transiting planets are of particular interest.

Whenever a planet passes in front of its host star from our line of sight, we see the typical diminishing in flux due to the obscuration of the stellar light by the planetary disk. Now let us assume the extrasolar planet features an atmosphere, we expect a fraction of the star's light to pass through the annulus created by the atmosphere. Let us further assume that these atmospheres contain a mix of atomic and molecular species that absorb the stellar radiation at given characteristic wavelengths of light. If we now observe the transiting extrasolar planet

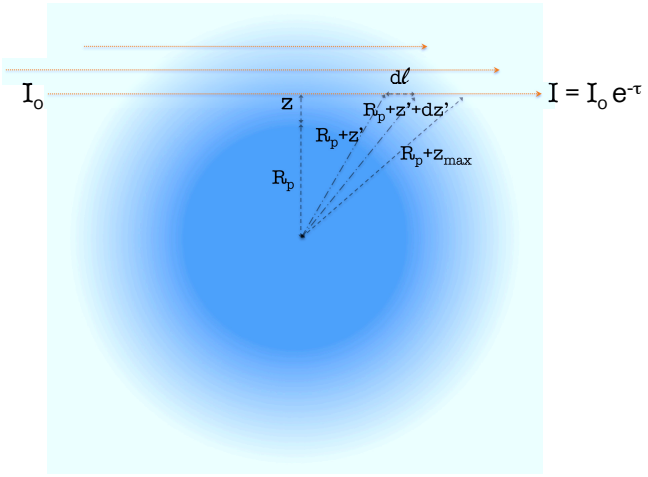


Figure 1.9: *Geometry of a transit observation: the stellar photons are filtered through the planetary atmosphere (Tinetti et al., 2012).*

at one of these given wavelengths, the planet radius appears slightly bigger than it otherwise would due to the absorption of the stellar light in the atmosphere. Such an increase in perceived planetary radius is reflected in a slightly increased depth of the transit lightcurve. If we now observe the transit event at different wavelengths, we can measure the fluctuations on the mean transit depth and with such measure the atmospheric opacity as function of wavelength, λ. For a hot-Jupiter featuring a highly inflated atmosphere, the amplitude of these spectral features are typically 1 part in 10^3 to 10^4, with the mean amplitude being dominated by the scale height of the atmosphere, H

$$H = \frac{kT_{eff}}{\mu g} \tag{1.3}$$

where k is the Boltzmann constant, T_{eff} the effective temperature of the planet, μ the mean molecular weight of the atmosphere and g the surface gravity of the planet. For a given wavelength of light, we can calculate the perceived transit depth as a function of the stellar and planetary radii plus the opacity of the absorbing species in the exoplanetary atmosphere. Here we can state that the flux from the planet is a function of the incident stellar flux I_0

$$I(\lambda, z) = I_0 e^{-\tau(\lambda,z)} \tag{1.4}$$

where τ is the total optical depth per wavelength and atmospheric depth, z, and is the sum of the optical depths of the different absorbing species, i, $\tau(\lambda, z) = \sum_i \tau_i(\lambda, z)$. Given the ideal gas law, the equation of hydrostatic equilibrium and the atmospheric density, ρ:

$$\frac{p}{\rho} = kT; \qquad \mathrm{d}p = \mu\rho g \mathrm{d}z; \qquad \rho = \frac{N}{V} \tag{1.5}$$

we can to first-order convert atmospheric pressure, p, into height in the atmosphere, z. We can then calculate the optical depth per absorbing species, i, by integrating over the atmospheric density, ρ, mixing-ratio of the ith absorber, χ_i, at height z and corresponding absorption coefficient σ_i

$$\tau_i(\lambda, z) = 2\int_0^{l(z)} \rho(z')\chi_i(z')\sigma_i(\lambda, z)\mathrm{d}l. \tag{1.6}$$

The path traveled through the atmosphere, $l(z)$, is simply given by

$$l(z) = \int \mathrm{d}l = \sqrt{(R_p + z_{max})^2 - (R_p + z)^2} \tag{1.7}$$

where R_p is the planetary radius and the geometry of the observation is illustrated in figure 1.9. We can now calculate the transit depth of the lightcurve due to the planet and star radii and the absorption of the atmosphere as a function of wavelength by

$$\kappa(\lambda) = \frac{R_p^2 + 2\int_0^{z_{max}} (R_p + z)(1 - e^{-\tau(\lambda,z)})\mathrm{d}z}{R_*^2} \tag{1.8}$$

where R_* is the stellar radius.

The above is true for primary eclipses, also known as transmission spectroscopy, and is most sensitive to the absorption of atmospheric species in the atmosphere of extrasolar planets. By observing the secondary eclipse, when the planet passes behind its parent star, we are sensitive to the thermal emission of the extrasolar planet as in the case of the secondary eclipse lightcurve since we loose the thermal emission of the planet as the star occults it. We define our transit depth as the contrast ratio F_p/F_*

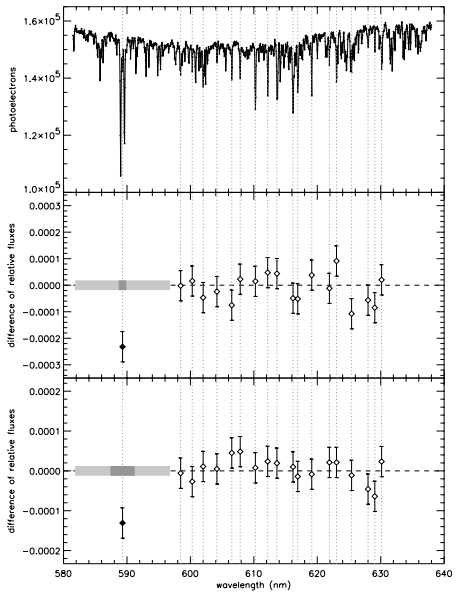
$$\frac{F_p}{F_*} = \frac{R_p^2}{R_*^2} \times \frac{I_p(\lambda, T_p)}{I_*(\lambda, T_*)} \tag{1.9}$$

where F_p and F_* are the planetary and stellar fluxes respectively. Similarly to the transmission spectroscopy case, we can measure a varying transit depth over a given wavelength range and from this calculate the emission of individual atmospheric constituents. This analysis, however, is made difficult by the need for information on the temperature-pressure profile of the extrasolar atmosphere.

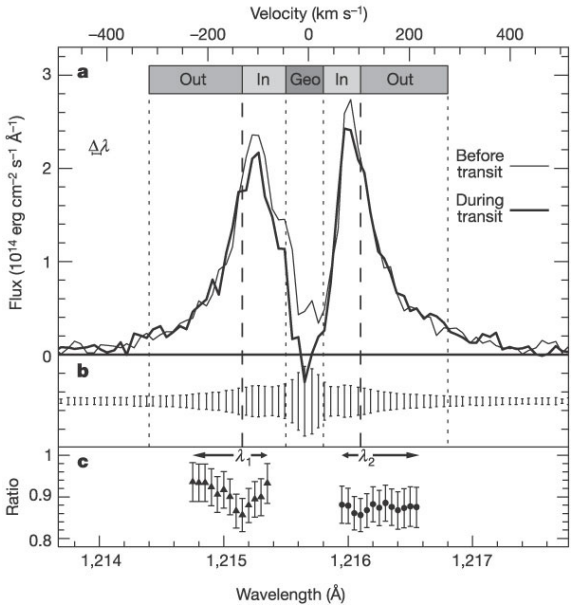
1.3 A brief history of exoplanetary spectroscopy

The idea of exoplanetary spectroscopy is not a new one and existed long before the first exoplanets were discovered (Angel et al., 1986). Prior to the detection of the hot-Jupiter 51Peg-b in 1995, it was commonly assumed that large scale, space-based interferometers and coronographs were needed to take a glimpse at planets beyond our own solar system. Missions such as *Darwin* (Léger et al., 1996) and the *Terrestrial Planet Finder* (Coulter, 2003) were highly complex and expensive endeavours and clearly children of their time. It was the first detection of a transiting exoplanet (Charbonneau et al., 2000) that lead to a paradigm shift in the understanding of observational techniques required to study these distant atmospheres.

Given the obvious limitations imposed by our own telluric atmosphere, the exoplanetary spectroscopy of transiting planets began with already existing space-based instruments. Using the STIS spectrograph on the *Hubble* telescope, Charbonneau et al. (2002) observed the primary transit of HD209458b in a wavelength region near 600nm for signs of a sodium absorption feature, figure 1.10a. These observations were only the beginning and using *Hubble*/STIS, other groups observed sharp metallic lines in the UV to visible bands of the spectrum. Amongst others but most notable are the observations of hydrogen (in particular Lyman α) by Vidal-Madjar et al. (2003); Ballester et al. (2007); Ben-Jaffel (2008); Jensen et al. (2012), and the

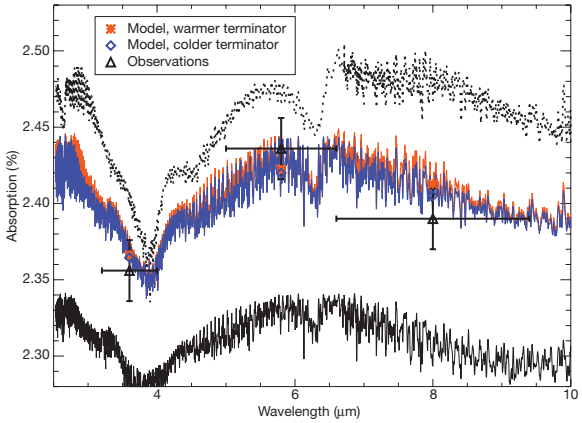


(a) Planetary absorption feature corresponding to atmospheric sodium. (Charbonneau et al., 2002)

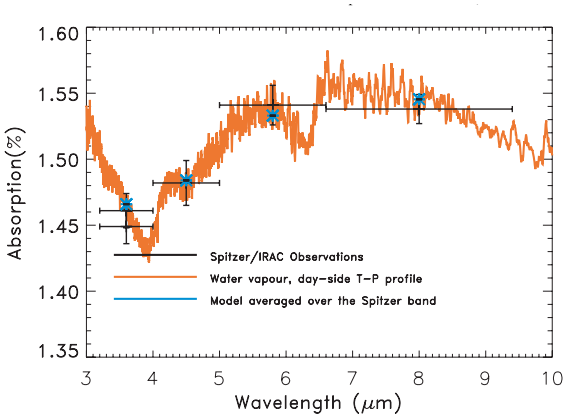


(b) Planetary absorption feature corresponding to atomic hydrogen, here visible is the Lyα line (Vidal-Madjar et al., 2003)

Figure 1.10: *Hubble/STIS observations of metallic line absorptions of HD209458 in the visible/near-UV. These absorption lines can be very prominent, up to ∼15% of the stellar light in the case of hydrogen.*



(a) Observations of HD189733b at 3.6, 5.4 and 8.0 μm overplotting a radiative transfer model of water (Tinetti et al., 2007)



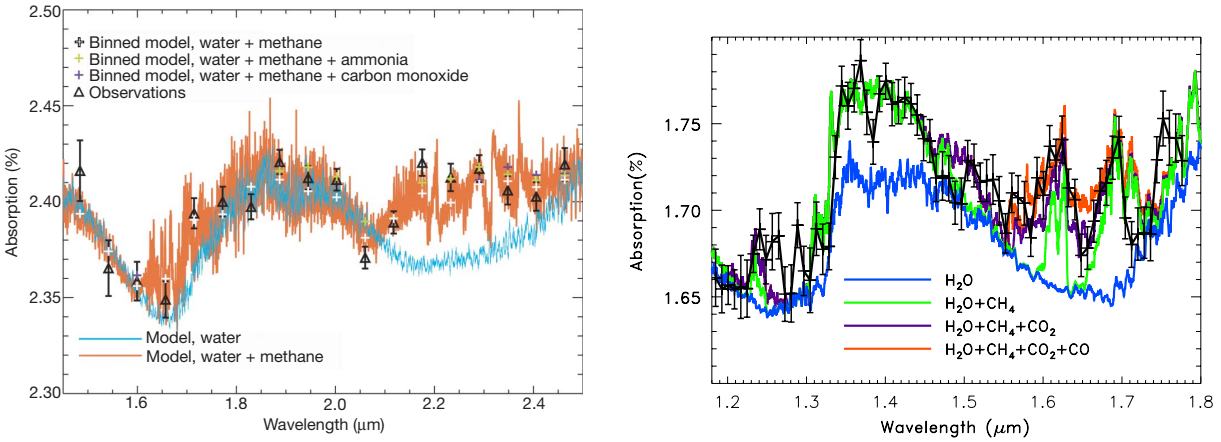
(b) Observations of HD209458b at 3.6, 4.5, 5.4 and 8.0 μ showing a good fit to a water radiative transfer model (Beaulieu et al., 2010)

Figure 1.11: *Spitzer/IRAC photometric observations of the primary eclipse of two hot-Jupiters HD189733b and HD209458b. Both planets show clear signs of atmospheric water.*

observations of CII, and SiIII by Linsky et al. (2010) on the hot-Jupiter HD209458b. The observations of atomic species in the atmospheres of hot-Jupiters can illuminate the upper-atmosphere ionisation and molecular dissociation processes and may indicate atmospheric escape processes. Observationally, these metal lines were a good point to start as their very strong absorptions, up to $\sim 15\%$ of the stellar flux in the case Lyα on HD209458b (Vidal-Madjar et al., 2003), make these easy to detect with the instrumentation at hand.

Whereas the UV and the visible are dominated by electronic transitions and their associated absorption lines, we find a very different picture in the infra-red. Here, due to the lower energy carried by the IR photons, the spectrum is less dominated by sharply peaked electronic transitions of atomic species but by roto-vibrational transitions of complex molecules. These individual molecular features are often smooth modulations over several microns and it is here that we find the signatures of water, methane and other important molecules.

In 2007, several teams published independently and nearly simultaneously tentative observations of water in the atmospheres of HD189733b and HD209458 using the mid-IR IRS spectrograph on the *Spitzer* telescope (Richardson et al., 2007; Swain et al., 2008a; Grillmair et al., 2007). Plagued by the poor signal-to-noise of their spectra, the detection of water was tentative at best. Another tentative observation of water was made by Barman (2007) in the visible wavelength spectrum of Knutson et al. (2007b). Using the IRAC photometry camera on the *Spitzer* space-telescope with its four broadband channels at 3.6, 4.5, 5.8 and 8.0μm, Tinetti et al. (2007) observed the transit of HD189733b. The distribution of the broadband photometric points could be explained by the broad roto-vibrational bands of water in this hot-Jupiter, figure 1.11a. This first cruder measurement was later confirmed by additional observations of



(a) Observations of HD189733b overlaid on a pure water model (cyan) and a water+methane model (Swain et al., 2008c)

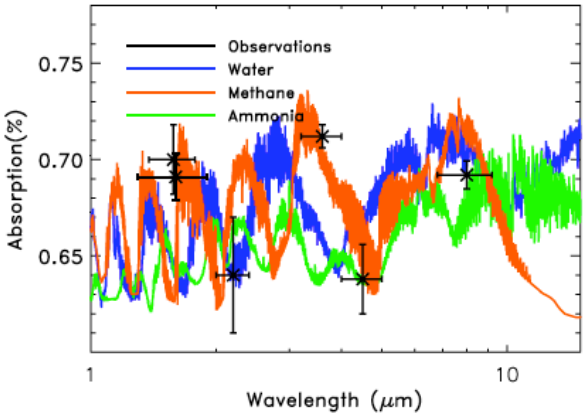
(b) Observations of XO1b overlaid on models showing various potential atmospheric absorbers. Water and methane are clearly visible in the data (Tinetti et al., 2010)

Figure 1.12: *Hubble/NICMOS observations of the primary eclipses of hot-Jupiters HD189733b and XO1b. Both planets show clear signs of atmospheric water and methane.*

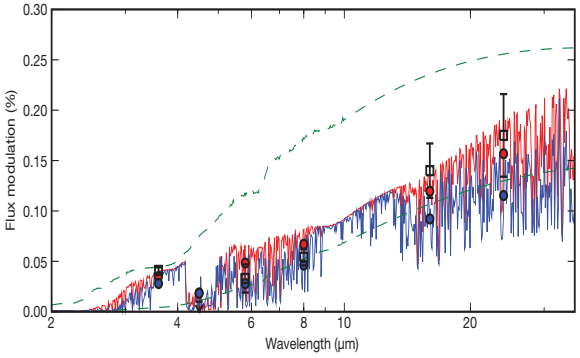
HD189733b (Charbonneau et al., 2008; Grillmair et al., 2008; Swain et al., 2008c, 2009b) and HD209458b (Knutson et al., 2008; Swain et al., 2009a; Beaulieu et al., 2010; Burrows et al., 2010).

One year later, observations of the primary eclipse of HD189733b using the *Hubble*/NICMOS spectrograph in the near infrared, resulted in a spectrum that could not be explained by the presence of water alone. The models based on Tinetti et al. (2007) and later confirmed by Madhusudhan and Seager (2009) could only describe the observations with the addition of methane as atmospheric constituent, figure 1.12a. Whilst the observations of Swain et al. (2008c) are controversial in the field and are extensively addressed in chapters 3 and 4, the detection of methane seemed solid and was independently confirmed by various groups (Agol et al., 2010; Désert et al., 2009; Beaulieu et al., 2008). Methane was furthermore found in the hot-Jupiter XO1b (Tinetti et al., 2010) and possibly in the warm-Neptune GJ436b (Beaulieu et al., 2010, 2011). However the detection in GJ436b remains inconclusive because of stellar variability and non-simultaneous observations. Additionally there is a controversy in the analysis of GJ246 *Spitzer* data (Stevenson et al., 2010; Knutson et al., 2011). Such discrepancies are related to the difficulty of calibrating one's data and showcase the degeneracy of data calibration with the scientific result. It should be noted that carbon-dioxide was quite unambiguously detected in another planet, HD209458b, using secondary eclipse measurements with a combination of instruments allowing us to build up a clear picture of the emission profile of HD209458b from the near infra-red, *Hubble*/NICMOS , to $\sim$15μm, *Spitzer*/IRAC , (Knutson et al., 2007b, 2008; Swain et al., 2008b, 2009a).

Initially, most atmospheric studies focused on the very few brightest hot-Jupiters. With the



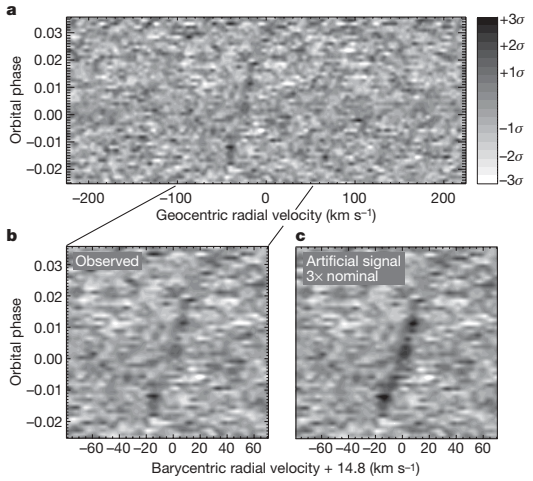
(a) Primary eclipse *Spitzer*/IRAC observations of GJ436b suggesting a methane rich atmosphere of the warm-Neptune (Beaulieu et al., 2011).



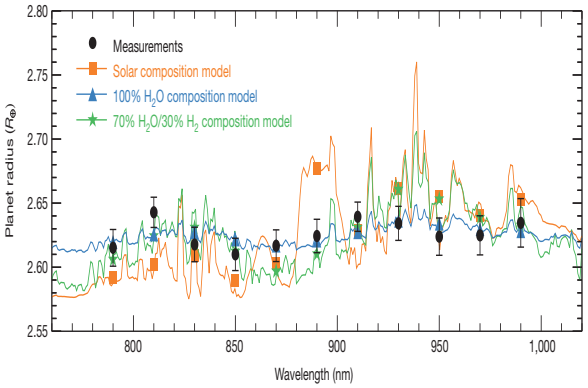
(b) Secondary eclipse *Spitzer*/IRAC measurements of GJ436b suggesting a strongly methane depleted atmosphere (Stevenson et al., 2010).

Figure 1.13: *Two measurements of the same warm-Neptune, GJ436b, using the same instrument. One measurement suggests a methane rich atmosphere whilst the other indicates a strongly methane depleted atmosphere. Degeneracies in data de-trending and radiative modelling often lead to conflicting results in the literature.*

increasing maturity of the techniques, we are now entering an era of atmospheric characterisations of Super-Earths such as GJ1214b (Charbonneau et al., 2009; Bean et al., 2010; Bean, 2011; Bean et al., 2011; Désert et al., 2011; Croll et al., 2011; Berta et al., 2012), figure 1.14b. Whilst some observations are still taken from space, in particular with the new WFC3 camera on *Hubble*, the times are clearly changing and observations from the ground become increasingly attractive and necessary (Redfield et al., 2008; Snellen et al., 2008; Jensen et al., 2012). Long frustrated by the absorptivity and variability of our own telluric atmosphere, the first successes in the infra-red were only achieved in 2010 using very different observational approaches (e.g. Swain et al., 2010; Snellen et al., 2010a; Bean et al., 2010), see introductions to chapters 5 & 6. Despite the obvious difficulties of observing the atmosphere of a distant exoplanet through that of our own, ground-based observations have their advantages as it allows a much more frequent access to a wide variety of telescopes and instruments, which lead to the first detections of carbon-monoxide on HD209458b (figure 1.14a; Snellen et al., 2010a) and non-LTE emission processes on the hotter HD189733b (figure 1.15; Swain et al., 2010).



(a) Ground-based detection of CO in hot-Jupiter HD209458b using the *VLT*/CRIRES spectrograph (Snellen et al., 2010a).



(b) Ground-based spectroscopy of the Super-Earth GJ1214b using the *VLT*/FORS spectrograph (Bean et al., 2010).

Figure 1.14: *Ground-based detections, despite the difficulties imposed by observing through our own telluric atmosphere, are becoming more and more common employing a variety of observing strategies.*

1.3.1 Heated debates and the pitfalls of data analysis

Despite the overwhelming successes of exoplanetary spectroscopy in the past years, it is fair to say that most, if not all of these results, are to some degree viewed with a critical eye by at least one other group in the field. As much as this is testament to the scientific morale and rigour of those involved, it poses the question of why a new result is not always universally accepted at the time of its publication. The answer is simple but the solution complicated: To extract a spectrum of an exoplanetary atmosphere, one typically needs a photometric precision of 1 part in 10000, over time spans that range between a few to tens of hours. Such a level of instrumental stability is difficult to attain, especially with instruments that were not calibrated nor intended to reach these levels of accuracy. In these cases, the occurrence of instrumental systematic noise at the level of the desired spectroscopic feature is commonplace and the data-analysis rapidly becomes non-trivial. Various publications discuss the difficulties of such an endeavour with often contradicting results based on the instrument corrections used (e.g. Beaulieu et al., 2011; Stevenson et al., 2010; Swain et al., 2008c; Gibson et al., 2011; Swain et al., 2011; Knutson et al., 2011).

Here, the solution is two-fold: 1) In the short and medium-term, we need to the develop more powerful and unbiased data-analysis routines and algorithms and this thesis is a contribution to that goal; 2) in the long-term we need a dedicated space-based instrument for exoplanetary spectroscopy, such as the proposed *FINESSE* and *EChO* missions (Tinetti et al., 2011a,b; Swain, 2010; Tessenyi et al., 2012).

1.4 Thesis outline

At the heart of exoplanetary spectroscopy lies the data. At a time of characterising ever smaller and fainter targets, the accuracy of data analysis becomes more and more important and has become of critical importance to the progress of this field and constitutes the core theme of this thesis.

In **Chapter 2**, I will briefly introduce the fundamental techniques that are implicitly applied to all subsequent chapters but not further discussed. These are the extraction of stellar spectra from the observed raw-data and a general introduction to instrumental noise sources with *Hubble*/NICMOS as a specific example. I will also further introduce the properties of the transiting lightcurve and how to model it.

Chapter 3 will discuss the inadequacy of many parametric noise correction methods in the literature, introduce the 'Cocktail Party problem' and discusses its applications to spectroscopic data of extrasolar planets. I will then proceed to defining an unsupervised machine learning algorithm for the iterative learning and consequent removal of instrument noise. The chapter will conclude with a study of the algorithm proposed using simulated data.

Following from the previous chapter, I discuss the application of the proposed algorithm to real data in **Chapter 4**. To begin I analyse two *Hubble*/NICMOS data sets of HD189733b and XO1b, both observed with different grisms and proceed to compute the fully de-correlated spectrum of HD189733b in a non-parametric way. These results are compared and contrasted to results in the literature and the adequacy of the proposed algorithm is discussed.

In **Chapter 5** I introduce the difficulties in analysing ground-based spectroscopic data of extrasolar planets and outline the need for a second analysis algorithm with a high degree of robustness in very low, <2, signal to noise (SNR) conditions. I introduce the concept of sparsity of lightcurves in the frequency domain and discuss a Fourier based convolution and test its properties using very low-SNR simulated data. At the end of this chapter I will also introduce the concept of wavelet demonising using a multi-resolution analysis of individual time series and Gaussian noise dampening by soft-thresholding.

Chapter 6 sees the application of the Fourier based techniques to four nights of ground-based secondary eclipse data of HD189733b observed in the K and L-bands using the *IRTF*/SpeX instrument on Mauna Kea. I apply the routines developed in the previous chapter to extract the emission spectrum of HD189733b's dayside. I extensively discuss our findings and test for residual telluric contamination in the reported spectra.

Chapter 7 discusses the complementary nature of both the algorithms discussed in Chapters 3 & 5 and proposes a novel way of combining independent component analysis and multi-

resolution wavelet analysis in one multi-purpose algorithm.

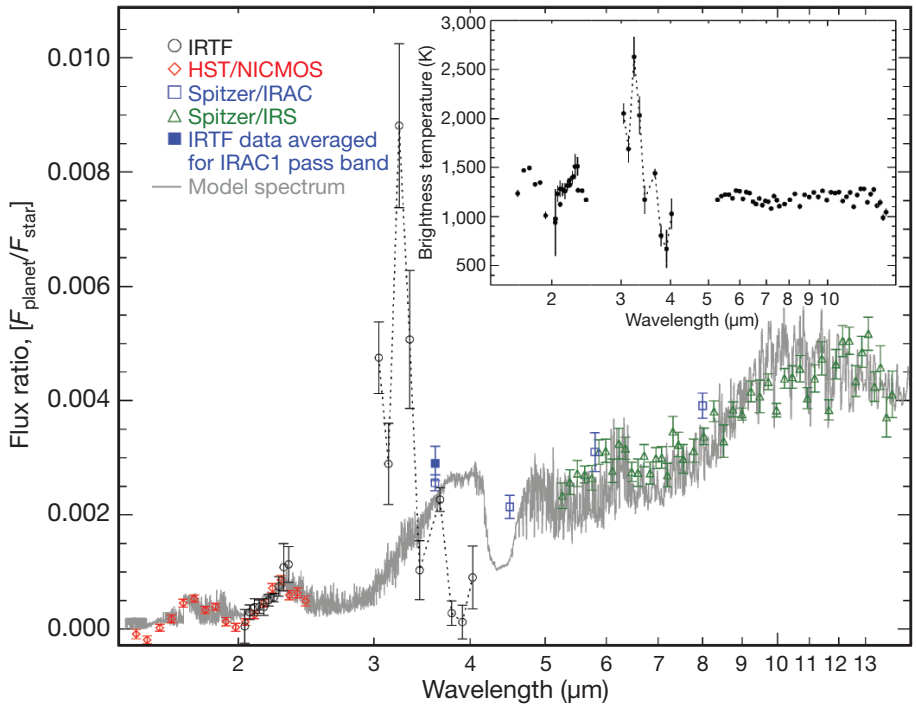


Figure 1.15: *Unexpectedly strong 3.25-μm emission present in the dayside spectrum. The brightness temperature of the 3.25-μm emission feature indicates the likely presence of a non-LTE emission mechanism. The dayside emission spectrum is based on the measurements taken by the IRTF/SpeX instrument, together with previous results from Hubble spectroscopy (red), Spitzer spectroscopy (green), and Spitzer photometry (blue); all data are shown 1σ errors. A radiative transfer model (grey) assuming LTE conditions and consistent with the measurements made with the Spitzer and Hubble telescopes fails to describe the emission structure at 3.14.1 μm. (Swain et al., 2010).*

Chapter 2

Introduction - Technical Details

Here I will briefly introduce and discuss some fundamental concepts for the analysis of exoplanet lightcurves, as well as for the extraction of spectra from the raw, observed data. Whereas in other fields of astronomy the data reduction ends with the extraction of the stellar spectrum from the observed data, this is not the case for exoplanetary spectra for which the 'standard' data reduction is a pre-requisite to further data de-trending algorithms as described in the following chapters. Hence, in the subsequent chapters, these technical fundamentals will be assumed and all mentions of 'raw' data will refer to the fully extracted spectra unless specifically stated otherwise.

2.1 The Lightcurve morphology

Figure 2.1 illustrates the rationale of transiting exoplanet lightcurves. When the extrasolar planet passes in front of its host star, in the line of sight of the observer, it will obscure some of the host star's flux and we will observe a temporary decrease in the flux intensity. A similar effect can be observed when the exoplanet, emitting thermal radiation and reflecting some of the light of the host star, is occulted by the star. As described in chapter 1 we can appreciate that should the exoplanet feature an atmosphere, which is always the case for gas giants, some of the stellar light will filter through the small atmospheric annulus (the limb) when the planet is passing in front of its host star. This atmospheric annulus will be more or less opaque depending on whether the atmospheric constituents absorb the stellar radiation at a given wavelength. The result is a change in the observed planetary radius and hence transit depth of the observed lightcurve. If the transit is now observed at multiple wavelengths, e.g. with a spectrograph, one obtains a time resolved lightcurve per spectral resolution element, see figure 2.2 for an illustration.

Although in this work I am mainly concerned with the transit depth, δ, as the fluctua-

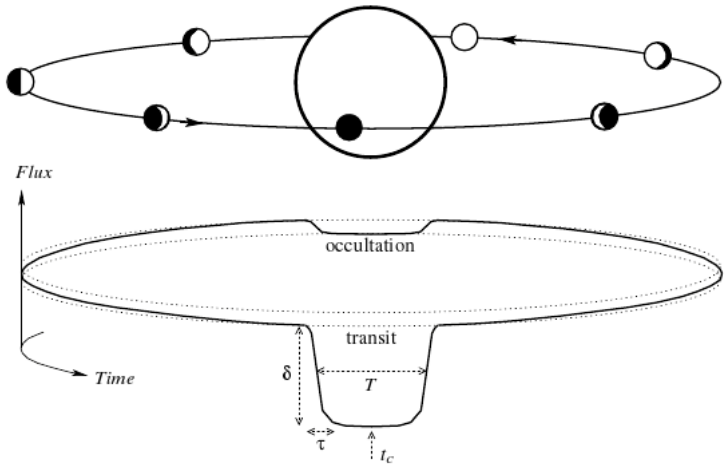


Figure 2.1: Illustration of transits and occultations. During a transit the planet blocks part of the host-star's light. As the planet orbits around its host star, the commonly tidally locked exoplanet's day-side comes into view and a slight increase in flux is observed due to the added thermal emissions of the exoplanet. This thermal emission is lost as the star eclipses the planet in the secondary eclipse, resulting in the secondary eclipse lightcurve feature. (Winn, 2008).

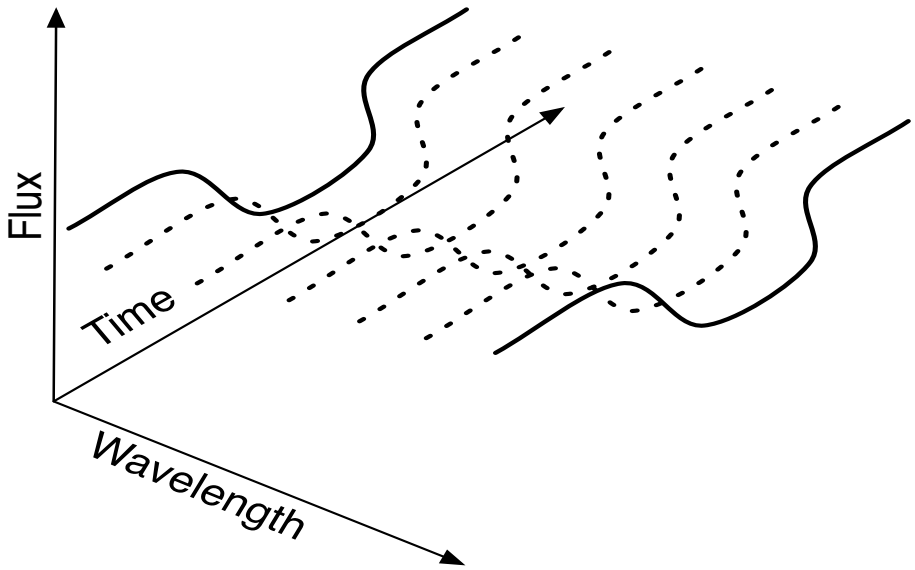


Figure 2.2: Illustration of a time-resolved spectroscopic data-set. Each point in time constitutes a stellar spectrum which is attenuated by the dimming flux due to the primary or secondary eclipse of the transiting exoplanet. When the eclipse is observed using a spectrograph, we obtain a lightcurve per spectral resolution element.

tion about its mean constitutes the exoplanetary spectrum, it is still important to discuss the equations governing the transit lightcurve.

Since the first confirmation of a detected transit (Charbonneau et al., 2000; Henry et al., 2000), the analytical derivation of a lightcurve feature, created by an occultation event, was pioneered by Mandel and Agol (2002) and Seager and Mallén-Ornelas (2003). Since then, much work has been undertaken to improve these analytical solutions to break degeneracies and to adapt them to specific cases, such as the modelling of exoplanets hosting exomoons (e.g. Giménez, 2006; Udry and Santos, 2007; Southworth et al., 2007; Bakos et al., 2007; Pál, 2008; Carter et al., 2008; Torres et al., 2008; Winn, 2008; Southworth, 2008; Kipping, 2008, 2009, 2010; Kipping and Tinetti, 2010; Carter and Winn, 2010).

Simple geometric arguments and Kepler's laws were used by Seager and Mallén-Ornelas (2003) to obtain the following six parameters of the orbiting system: the radius of the star, R_*, the stellar mass, M_*, the orbital inclination, i, the orbital semi-major axis, a, the stellar density, ρ_*, and the planetary radius, R_p. Out of the four observable parameters, the period, P, the transit depth, δ, the full time of transit (between first and last contact), t_T, and the transit time between ingress and egress, t_F, one can obtain the following analytical relationships:

$$a = \left(\frac{P^2 G M_*}{4\pi^2} \right)^{1/3} \tag{2.1}$$

$$b = \left[\frac{\left(1 - \sqrt{\delta}\right)^2 - (t_F/t_T)^2 \left(1 + \sqrt{\delta}\right)^2}{1 - (t_F/t_T)^2} \right]^{1/2} \tag{2.2}$$

where b is the impact parameter, defined as the projected distance between the star and planet centres during mid transit in units of R_*.

$$i = cos^{-1} \left(b \frac{R_*}{a} \right) \tag{2.3}$$

$$\frac{a}{R_*} = \frac{2P}{\pi} \frac{\delta^{1/4}}{\left(t_T^2 - t_F^2\right)^{3/2}} \tag{2.4}$$

$$\rho_* = \frac{32}{G\pi} P \frac{\delta^{3/4}}{(t_T^2 - t_F^2)^{3/2}} \tag{2.5}$$

and

$$p = \frac{R_p}{R_*} = \sqrt{\delta} \tag{2.6}$$

Furthermore, Southworth et al. (2007) identified the surface gravity of the planet, g_p, to be directly related to observable transiting lightcurve and radial velocity parameters.

$$g_p = \frac{2\pi}{P} \frac{(1-e^2)^{\frac{1}{2}} K_*}{r_p^2 \sin i} \tag{2.7}$$

where K_* is the velocity amplitude of the star, e is the orbital eccentricity, both obtained from radial velocity data, and $r_p = R_p/a$.

Kipping and Tinetti (2010) considered the effect of a pollution of the transit-depth, p, due to thermal emission of the night-side of tidally locked hot-Jupiters. This is a relevant study as this pollution effect can be shown to contribute at the 10^{-4} level of flux, which is the measurement precision we seek for the analysis of exoplanetary spectra. In their study, they introduced a night-side emission blend factor, $B = (F_* + F_{P,night})/F_*$, where F_* is the flux coming from the star and $F_{P,night}$ that of the planet's night side. Following from equations 2.2 & 2.4 they derive the night side polluted forms

$$b^2_{polluted} = \frac{p^2 - (1 + p^2 - b^2)\sqrt{B} + B}{B} \tag{2.8}$$

$$b^2_{polluted} = b^2 + \frac{1}{2}(1 - b^2 - p^2)(B-1) + \mathcal{O}[(B-1)^2] \tag{2.9}$$

$$(a/R_*)^2_{polluted} = (a/R_*)^2 - \frac{(1-b^2-p^2)[(a/R_*)^2 - (1+p)^2]}{2[(1+p)^2 - b^2]}(B-1) + \mathcal{O}[(B-1)^2] \tag{2.10}$$

The above equations suggest a systematic overestimation of the impact parameter, b, as well

as a systematic underestimation of the transit-depth. Such an effect would be most pronounced for hot-Jupiters observed at the wavelengths corresponding to the peak of their respective black-body curves. However, these effects are only an observational reality for a very quiet and stable planet-star system. Czesla et al. (2009), for example, has shown that the effect of stellar spots in CoRot-2b can amount up to 3%, 5-10 times that of the night-side pollution.

In the equations above I have so far ignored a very important contribution: that of limb-darkening of the star. This effect is critical in the UV, visible and near-IR, where a poor treatment of the stellar limb-darkening can offset the fitted transit depth by several percent of the total eclipse depth. Seager and Mallén-Ornelas (2003) assumed the limb-darkening laws to be linear, which is clearly insufficient. Mandel and Agol (2002) presented a more elaborate approach that allows for quadratic and non-linear limb darkening laws to be used. The computational code, they provide on their website[1], allows the iterative fitting of the observable parameters using limb-darkening coefficients, usually obtained from Claret (2000).

For the purpose of this work, I am mainly interested in the planet-star ratio, p, since the fluctuations about its mean constitute the spectral signatures. For most observations I assume all other parameters to be fixed to their established values in the literature. I hence use the standard modelling code by Mandel and Agol (2002). Here the non-linear transit equation for the lightcurve of a uniform source, $F(p,z)$, is given by

$$F(p,z) = \left[\int_0^1 \mathrm{d}r 2rI(r)\right]^{-1} \int_0^1 \mathrm{d}rI(r)\frac{\mathrm{d}[F(p/r,z/r)r^2]}{\mathrm{d}r}, \tag{2.11}$$

where $I(r) = 1-\sum_{n=1}^{4} c_n(1-\mu^{n/2})$, $\mu = \cos(\theta) = \sqrt{1-r^2}$, c are the limb-darkening coefficients, θ is the orbital phase in steradian and z is the projected distance of the planet-star centres and is given by

$$z = \sqrt{1-\sin^2(\theta)\cos^2(i)} \tag{2.12}$$

for a circular orbit. Equation 2.11 is not usually computed in its non-linear form, but approximated by a version assuming quadratic limb-darkening which permits a very fast and compact code. For the secondary eclipse case, we do not require limb-darkening of the star and equation 2.11 simplifies to

$$F(p,z) = 1 - \frac{I(z)}{4\Omega}\left[p^2\cos^{-1}\left(\frac{z-1}{p}\right) - (z-1)\sqrt{p^2-(z-1)^2}\right] \tag{2.13}$$

Since in our fitting procedure, all the parameters but p are fixed, there is no need to use too

[1] http://www.astro.washington.edu/agol

complex minimisation algorithms to solve for p. A standard downhill-simplex algorithm (Press et al., 2007) is more than appropriate. This said, much work has been undertaken in the field of parameter estimation, using predominantly Markov-Chain Monte-Carlo (MCMC) style codes (e.g. Bakos et al., 2007; Collier Cameron et al., 2007; Gregory, 2011a; Kipping, 2011). These procedures are important in the context of detecting new planets or searching for exomoons where the number of free parameters to be minimised is much greater.

2.2 Signal to noise calculations of Exoplanet spectra

When planning the observation of extrasolar planet spectra we search for temporal fluctuations on the stellar spectrum at the level of 10^{-3} to 10^{-5} of the total stellar flux. It is hence important to estimate the expected signal-to-noise ratio (SNR) for a given star, planet and instrument before attempting the observations. I here summarise the necessary calculations to estimate the mean planetary signal to noise.

Let us start with the typical CCD SNR equation for a stellar point source, assuming photometric (not spectroscopic) observations:

$$SNR_{star} = \frac{N_*}{\sqrt{N_* + n_{pix}(N_S + N_D + N_R^2)}} \tag{2.14}$$

where N_* are the total number of photons collected from the target star, n_{pix} the number of pixels considered in the SNR (also the resolution element for spectroscopy), N_s total number of sky background counts per pixel, N_D total number of dark current electrons per pixel and N_R is the square of the total number of electrons per pixel resulting from read noise. We define a photon limited case when $N_s, N_D, N_R \to 0$, which reduces the above equation to:

$$SNR_{star} = \frac{N_*}{\sqrt{N_*}} = \sqrt{N_*} \tag{2.15}$$

Equation 2.15 shows the SNR for a single frame. For a spectroscopic signature we have

$$SNR_{star} = \frac{N_*}{\sqrt{N_*}} \times \sqrt{N_F/2} \times \sqrt{\Delta\lambda} \tag{2.16}$$

Where N_F is the number of frames per observing run and $\Delta\lambda$ is the spectral range per resolution element, which is given by $\Delta\lambda = \lambda_c/R$, with λ_c being the central wavelength of the resolution element and R the resolving power of the spectrograph.

In the case of a photometric observation where we integrate over the whole stellar black-body

emission, we require the following set of equations to calculate SNR_*:

$$L_* = 4\pi R_*^2 \sigma_{SB} T_*^4 \quad \text{(W)} \tag{2.17}$$

where L_* is the bolometric luminosity of the star and equation 2.17 is the Stefan-Boltzmann law with σ_{SB} being the Stefan-Boltzmann constant. Alternatively, one can also calculate the luminosity as integral of the stellar black-body:

$$\begin{aligned} L_* &= 4\pi R_*^2 \int_0^\infty I(\nu, T_*) \mathrm{d}\nu \int \mathrm{d}\Omega \\ &= 4\pi^2 R_*^2 \int_0^\infty I(\nu, T_*) \mathrm{d}\nu \quad \text{(W)} \end{aligned} \tag{2.18}$$

$$I(\nu, T_{eff}) = \frac{2h\nu^3}{c^2} \frac{1}{e^{(h\nu/kT_{eff})} - 1} \quad (\mathrm{Wm^{-2}\Omega^{-1}}) \tag{2.19}$$

where $I(\nu, T_{eff})$ is the black body emission of the star for a given frequency ν, $\nu = c/\lambda$, c is the speed of light and λ is the wavelength of the photons, h is the Planck constant, k is the Boltzmann constant and Ω is the solid angle in steradians. The stellar flux density, F_*, at the observer's distance is now given by:

$$F_* = \frac{L_*}{4\pi D^2} \quad (\mathrm{Wm^{-2}}) \tag{2.20}$$

where D is the distance of the star from the observer. The number of photons collected by the the telescope per second integration time is hence given by

$$N_* = \frac{F_* \pi R_{tele}^2}{h\nu} = \frac{R_*^2 R_{tele}^2 \pi \sigma_{SB} T_*^4}{D^2 h\nu} \quad \text{(No. photons / s)} \tag{2.21}$$

where R_{tele} is the radius of the telescope. For a photon-limited case we now get the following SNR_{star}:

$$SNR_{star} = \frac{R_{tele} R_* T_*^2}{D} \sqrt{\frac{\pi \sigma_{SB}}{h\nu}} \times \sqrt{N_F/2} \times \sqrt{\Delta\lambda} \tag{2.22}$$

Now this is for the star only. For the primary eclipse case, we need to calculate the fraction of photons filtering through the exoplanetary atmosphere at the limb, we calculate the area of the atmospheric annulus, A_{atm}, which depends on the atmospheric scale height, H. The ratio

of the annulus area and the area of the star, A_{atm}/A_*, is

$$\frac{A_{atm}}{A_*} = \frac{\pi(R_p + nH)^2 - \pi R_p^2}{\pi R_*^2} = \frac{2R_p nH + (nH)^2}{R_*^2} \tag{2.23}$$

for an atmosphere that is n scale-heights thick, where $H = kT_p/Mg$, and k is the Boltzmann constant, T_p the planetary temperature, μ the mean molecular mass of the atmosphere and g is the planetary surface gravity.

We find the SNR of the atmospheric contribution for the primary transit to be the fraction of the stellar photons, N_*, that travels through the atmosphere divided by all the noise sources in the denominator of equation 2.14.

$$SNR_{primary} = \frac{N_*(2nR_pH + (nH)^2)\sqrt{N_F\Delta\lambda/2}}{R_*^2\sqrt{N_* + n_{pix}(N_S + N_D + N_R^2)}} = \frac{A_{atm}}{A_*} \times SNR_{star} \tag{2.24}$$

For the secondary eclipse case we have a similar calculation. Here one needs to calculate the ratio of the planetary flux, F_p over the total stellar flux, F_*, at the wavelength required. The planet-star contrast ratio is given by:

$$\frac{F_p}{F_*} = \frac{R_p^2}{R_*^2} \times \frac{I_p(\nu, T_p)}{I_*(\nu, T_*)} \tag{2.25}$$

Hence, we multiply this ratio (as for the primary transit) by the total number of photons recorded from the star, N_*, in equation 2.21. This yields

$$SNR_{secondary} = \frac{F_p \times N_*\sqrt{N_F\Delta\lambda/2}}{F_* \times \sqrt{N_* + n_{pix}(N_S + N_D + N_R^2)}} = \frac{F_p}{F_*} \times SNR_{star} \tag{2.26}$$

2.3 Reducing the raw data

After the observations were planned and executed, the next step is the data reduction and the extraction of the stellar spectra from the raw data. In the following chapters the term 'raw' data refers to the fully reduced and extracted spectra unless specified otherwise.

2.3.1 Origins of instrument systematics

Before I explain the extraction of the stellar spectra, it is important to discuss the origins of instrument systematics. Figure 2.3 illustrates the nature of systematic noise in observations obtained with the *Hubble*/NICMOS instrument, however it should be noted that these systematics are commonplace in other instruments too. *Hubble*/NICMOS is predominantly a photometer with two grisms attached to its filter wheel. This makes it a low resolution, slit-less spectrograph when either grism is selected. The blue dot in figure 2.3 marks the position of the target star in the field of view (FOV) of the detector and the green, horizontal line indicates the dispersion axis of the grism. Two elongated illuminated patches are visible. The right-hand patch marked by the red box is the first order of the grism dispersion and constitutes the science data and ∼80% of the flux collected. Here, every 5 pixels in the dispersion direction constitute a spectral channel and the stellar spectrum is extracted from this first order only. The second order contains no additional information. Besides the first and second orders of the grism, we can see several other features in the exposure: these are predominantly systematics. Features marked with orange (A)s are ghosting spectra of the grism or spectra of faint background sources which can overlay the science data. The most important source of instrument systematics are the changing quantum efficiencies of the detector. This effect may either cause a gradual change or sudden jumps in the number of counts recorded for a uniform source. Most of these effects are attenuated by the use of flat-fields. During flat-fielding, the detector is uniformly illuminated allowing us to map out the varying sensitivities and to subsequently subtract these of the science exposure. Unfortunately, for spectroscopic measurements, the flat-field is also wavelength dependent and in the case of *Hubble*/NICMOS no such calibration is readily available. These varying quantum efficiencies of the detector in conjunction with the positional instability of the spectrum on the detector are the main causes for most systematics observed. Such variations are typically at the order of 10^{-3} in flux whilst the exoplanetary spectrum has a typical contribution of 10^{-4}. Additionally, we find systematics due to: 1) grism angle instabilities, 2) detector temperature fluctuations, 3) varying sunshield illumination angles and 4) astrophysical noise due to stellar variability and spots. *All these effects are at the order of magnitude or bigger than the exoplanetary spectrum.*

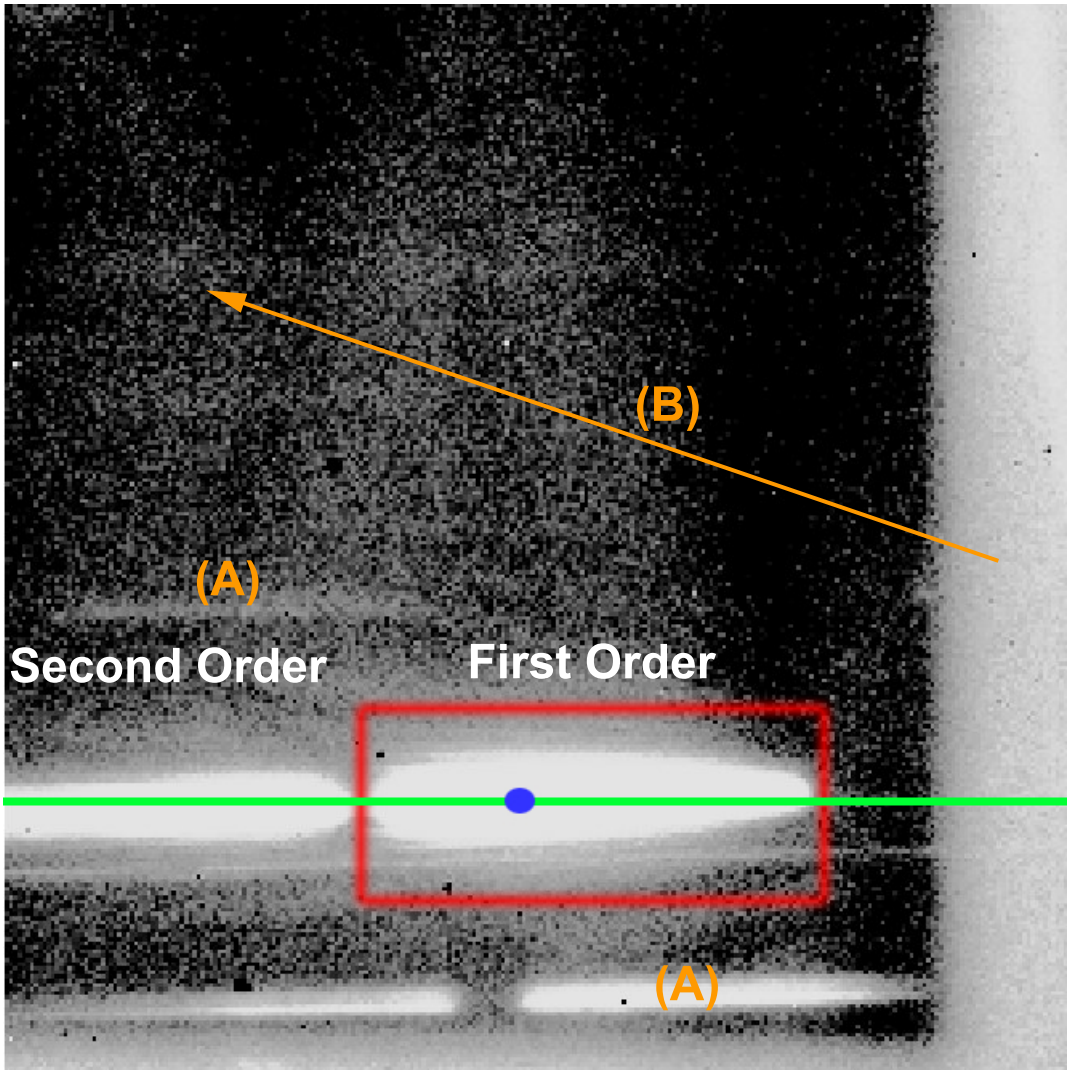


Figure 2.3: Single frame of HD189733b obtained by *Hubble*/NICMOS with the G206 grism in place. The red box marks the position of the desired first order spectrum with the green line marking the dispersion axis and the blue dot the approximate position of the target star in the field. Other features are present, namely the second order spectrum on the left, 'ghost' and background source spectra above and below the marked by (A) as well as various irregular flat field features across the detector (B).

2.3.2 Extracting a spectrum

Several excellent astronomical software packages exist for the reduction of spectroscopic data. The most versatile is IRAF[2], and its python interface language PyRAF[3], which is extensively used for *Hubble* and *Spitzer* data analysis in most fields of astronomy. These semi to fully automated tools are sufficient in accuracy for most spectroscopic applications. However, despite providing an important consistency check, I often find custom-built routines to be more precise at the 10^{-4} level of flux. This is particularly true for slit-less spectra as in the case of *Hubble*, where flat-fielding is not a viable option and the data are processed through the *Hubble* calibration pipeline. Here, no further bias subtraction or other pre-processing is required and I directly proceed to the extraction and wavelength calibration of the data. More specifically:

1. I crop all frames to only include the first order (red square in figure 2.3)

2. One-dimensional Gaussians are fitted along the vertical axis for every pixel-row in the dispersion direction.

3. I sum the counts of the pixels enclosed in four times the full-width-half-maximum (FWHM) of the fitted Gaussians and bin for 5 pixel rows in the dispersion direction.

4. From the Gaussian centroids, I can obtain the mean y-positional drift of the spectrum on the detector, fit for the angle of the dispersion axis to a horizontal. I also obtain the FWHM of the spectrum.

5. I cross-correlate the one dimensional extracted spectra to obtain the x-positional drift of the spectrum on the detector. These drifts, as well as the parameters derived in the previous point, constitute the optical state vectors. For a more in-depth discussion of these state vectors, please refer to section 4.2.

6. I wavelength calibrate the spectrum using a photometric exposure of the target star, taken before the grism was rotated in place, as reference. The wavelength solution for the G206 grism for each resolution element, i is then given by

$$\lambda_i = 2.045 - 0.01152(x_i - x_0) \tag{2.27}$$

   where x_i is the x-position on the detector for each resolution element and x_0 is the reference position of the target star.

[2] $http://iraf.noao.edu/$
[3] $http://www.stsci.edu/institute/software\_hardware/pyraf$

7. Finally, the exposure time-stamp is converted to terms of orbital phase, ϕ, by

$$\phi = \frac{T_{\mathrm{H}JD} - T_{\mathrm{E}ph}}{P} - \left\lceil \frac{T_{HJD} - T_{Eph}}{P} \right\rceil \tag{2.28}$$

where $T_{\mathrm{H}JD}$ is the time-stamp of the exposure in heliocentric julian date, $T_{\mathrm{E}ph}$ the ephemeris of the orbit and P the period of the orbit.

Two extracted spectra are shown in figure 2.4: One out-of-transit spectrum in blue and one in-transit spectrum in red. The flux difference is a combination of flux loss due to the transiting exoplanet as well as time variable systematic noise. The rest of this thesis is concerned with the disentangling of the astrophysical signal from these instrument systematics in order to reach the required 10^{-4} precision in flux.

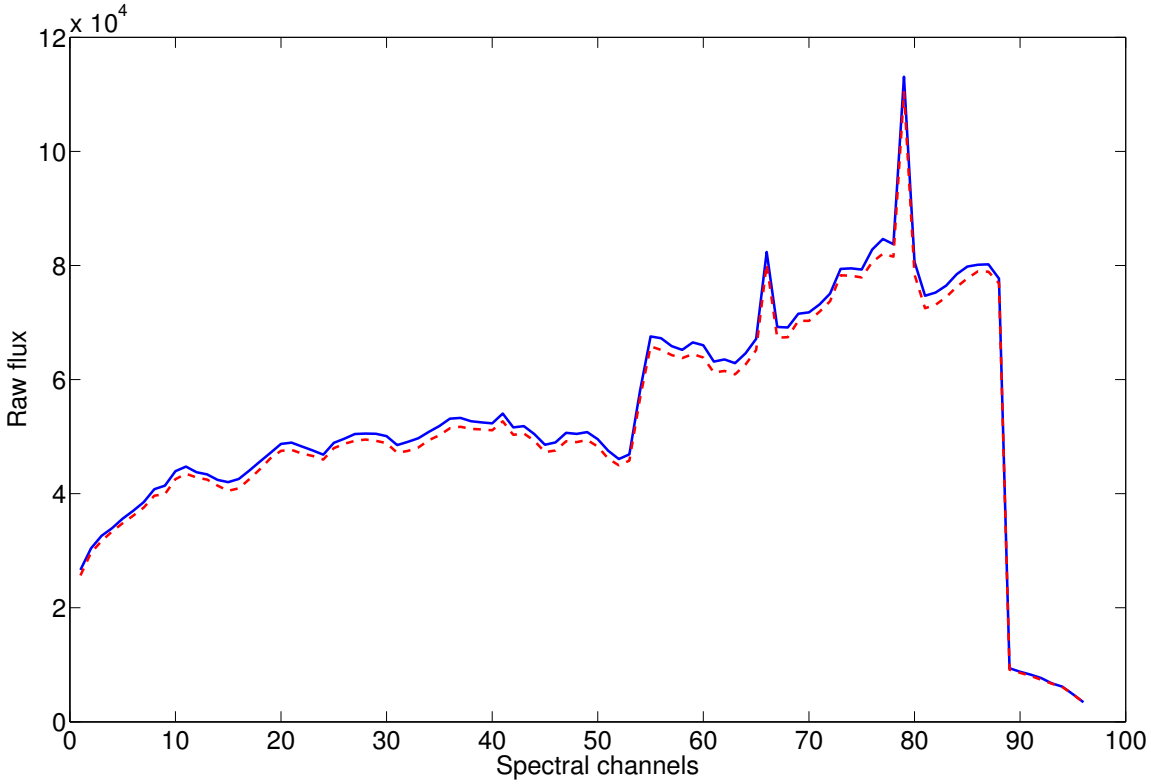


Figure 2.4: Two extracted stellar spectra of HD189733b primary transit. The blue continuous spectrum is a randomly chosen out-of-transit spectrum in the second orbit and the red, discontinuous spectrum is in the middle of the transiting event. We can see a slight flux difference due to the planetary transit.

Chapter 3

The Cocktail Party Problem

"The bar is in full swing, and floating rounds of cocktails permeate the garden outside, until the air is alive with chatter and laughter, and casual innuendo and introductions forgotten on the spot, and enthusiastic meetings between women who never knew each others names."

— **The Great Gatsby (F. Scott Fitzgerald)**

3.1 Introduction

It has become clear from the previous chapter that the aim of characterising ever smaller targets is equally a quest for constantly improving accuracy of our data. Permanently at the edge of technical feasibility, we often find the limiting factor of our analysis to be the stability of our instrument. Dedicated missions such as the *Kepler* space-craft (Borucki et al., 1996, 2010; Jenkins et al., 2010; Koch et al., 2010) feature an *a priori* calibration plan that allows for the detection of the faintest time-series deviations at a photometric accuracy of typically 10^{-5}. However, due to the young age of the field, such specifically designed instruments are rare and only at the proposal stage for spectroscopic measurements (e.g. Swain, 2010; Tinetti et al., 2011a,b; Tessenyi et al., 2012). Current spectroscopic data obtained by instruments such as *Hubble*/NICMOS and *Spitzer*/IRS feature a typical, native stability of $\sim 10^{-3}$ over the course of a single observing run. The required precision for the detection of an atmospheric feature is typically ten to hundred times higher. For these instruments we often find the astrophysical signal not to be limited by the photon-noise floor but by additional wavelength and/or time dependent systematics of the instrument.

To minimise the impact of these systematic noise components, different approaches have been proposed in the past. For space and ground-based observations, eg. Spitzer and Hubble (eg.

Agol et al., 2010; Beaulieu et al., 2008, 2011; Charbonneau et al., 2002, 2005, 2008; Deming et al., 2007; Gillon et al., 2010; Grillmair et al., 2008; Knutson et al., 2007a,b; Sing et al., 2011; Snellen et al., 2010a; Bean et al., 2011; Swain et al., 2008c, 2009b,a; Tinetti et al., 2007, 2010), instrumental systematic noise has been approximated using parametric models, often based on auxiliary information (optical state vectors) such as instrumental temperature, orbital inclination, inter and intra-pixel positions of the point-spread-function (see sections 2.3 & 4.2). Using optical state vectors to de-correlate one's data is an effective technique (Swain et al., 2008c). However for instruments that lack a calibration plan at the precision of 10^{-4}, the accuracy of the retrieved optical state vectors (e.g. sensor sampling rates) and the adequacy of the instrument model's definition itself become difficult to determine. Some of the recent controversy over results reported by various teams can be attributed to this circumstance (Knutson et al., 2011; Stevenson et al., 2010; Beaulieu et al., 2011; Swain et al., 2008c; Gibson et al., 2011; Pont et al., 2011; Hatzes et al., 2011; Bruntt et al., 2010). The situation is even further complicated by systematic noise emanating from the star in the form of stellar pulsations or spots as described in the following chapter.

In many cases, we can conclude that very little is known about the exact nature of the instrument response function (IRF) or the stellar activity of the system to make a parametric correction free of the risks of over-correcting or biasing the data. It is easy to see why it becomes important to work towards an alternative route to quantify or remove systematic noise using non-parametric models.

Carter and Winn (2009), Gibson et al. (2012) and Thatte et al. (2010) have recognised this need and developed systematic noise removal/corrections algorithms using adapted wavelet filters, gaussian-processes (GP) and principal component analysis (PCA) respectively. Whilst they are significant steps forward in the non-parametric de-trending of observational data, these techniques are inherently limited by their underlying assumptions or by the semi-parametric nature as will be discussed later in the text. In Waldmann (2012) I took the novel approach of 'blind-source-separation' using unsupervised machine-learning algorithms based on the concepts of independent component analysis (ICA; Hyv̈arinen, 1999).

The topic of non-parametric noise correction is split between this chapter and chapter 4. Whilst in this chapter I will discuss the statistical background of blind-source separation algorithms, namely ICA. In section 3.2 I will introduce the concepts of 'blind' deconvolution followed by a more in-depth discussion of the algorithm proposed in section 3.3 and demonstrate its properties using simulated data in section 3.4. Chapter 4 will be concerned with a discussion of parametric reduction routines and the applications of non-parametric methods to real data.

3.2 Observations of a Cocktail Party

The classic analogy of blind-source separation is known as the 'Cocktail Party Problem'. Let us assume three people are talking simultaneously in one room. The speech signals of these people are denoted by $s_1(t)$, $s_2(t)$ and $s_3(t)$. In the same room are three microphones recording the observed signals $x_1(t)$, $x_2(t)$ and $x_3(t)$. The observed signals can be expressed in terms of the original speech signals:

$$\begin{aligned} x_1(t) &= a_{11}s_1(t) + a_{12}s_2(t) + a_{13}s_3(t) \\ x_2(t) &= a_{21}s_1(t) + a_{22}s_2(t) + a_{23}s_3(t) \\ x_3(t) &= a_{31}s_1(t) + a_{32}s_2(t) + a_{33}s_3(t). \end{aligned} \tag{3.1}$$

Instead of assuming $x(t)$ and $s(t)$ to be proper time signals, we drop the time dependence and assume them to be random variables

$$x_k = a_{k1}s_1 + a_{k2}s_2 + ... + a_{kN}s_N, \quad \text{for all } k = 1, ..., N \tag{3.2}$$

where a_{kl} is a weighting factor (in this case the square of the distance of the speakers to the microphone) and $k, l = 1, ..., N$ are some real coefficients with N being the maximum number of observed signals. The individual time series can also be expressed in terms of vectors where bold lower-case letters denote column vectors and upper-case letters denote matrices:

$$\mathbf{x} = \mathbf{A}\mathbf{s} \tag{3.3}$$

where the rows of $\mathbf{x}$ comprise the individual time series, x_k, and similarly $\mathbf{s}$ is the signal matrix of the individual source signals s_l. $\mathbf{A}$ is the 'mixing matrix' consisting of the weights a_{kl}. Equation 3.3 is also known as the instantaneous mixing model and often referred to as the classical 'Cocktail Party Problem' (Hyvärinen et al., 2001; Hyvärinen, 1999).

The challenge is to estimate the mixing matrix, $\mathbf{A}$ and its (pseudo)inverse the de-mixing matrix, $\mathbf{W}$,

$$\mathbf{W} = \mathbf{A}^{-1} \tag{3.4}$$

given the observations contained in $\mathbf{x}$ without any additional prior knowledge of either $\mathbf{A}$ or $\mathbf{s}$,

or for some methods without restrictions of $\mathbf{A}$ & $\mathbf{s}$.

Several algorithms have been proposed to find the linear transformation of equation 3.3. Amongst these are principal component analysis (PCA) (Pearson, 1901; Manly, 1994; Jolliffe, 2002; Press et al., 2007; Oja, 1992), factor analysis (FA) (Jolliffe, 2002; Harman, 1967), projection pursuit (PP) (Friedman, 1987; Huber, 1985) and the more recently developed independent component analysis (ICA) (Comon, 1994; Hyv̈arinen, 1999; Hyvärinen and Pajunen, 1999; Hyvärinen and Oja, 2000; Hyvärinen et al., 2001; Comon and Jutten, 2010; Stone, 2004).

The underlying differences between PCA and FA on one hand and ICA and PP on the other are the underling assumptions on the probability distribution functions (pdfs) of the signals comprising $\mathbf{x}$. The former group assumes the signals to follow: 1) a Gaussian distribution whilst the latter assume the signals to be, 2) predominantly non-Gaussian or sparse with specific signatures in the spectral domain (e.g. SMICA Delabrouille et al., 2003). This results in significant differences in the way we estimate our signal components.

3.2.1 Independence and Uncorrelatedness

The definitions of uncorrelatedness, orthogonality and independence are most times assumed to be equivalent and little distinction is made between these terms in most texts. For cases where the underlying pdfs are predominantly Gaussian, we can in fact prove that in the limit of the Central Limit Theorem (CLT), these terms are equivalent.

In Gaussian statistics, our probability densities are fully defined by the first and second statistical moments, i.e. their means and covariances. Two random vectors, $\mathbf{s_l}$ and $\mathbf{s_{l+1}}$, are said to be uncorrelated when their covariance ($\mathbf{C}_{\mathbf{s_l},\mathbf{s_{l+1}}}$) is zero:

$$\begin{aligned} \mathbf{C}_{\mathbf{s_l},\mathbf{s_{l+1}}} &= E[(s_l - E[s_l])(s_{l+1} - E[s_{l+1}])] \\ &= E[s_l, s_{l+1}] - E[s_l]E[s_{l+1}] = 0 \end{aligned} \tag{3.5}$$

where $E[s_l]$ is the expectation value of s_l which can be approximated by the mean in this case by

$$E[s_l] \approx \frac{1}{M}\sum_{t=1}^{M} s_l(t) \tag{3.6}$$

with M being the number of data points in the time series.

Furthermore, we define two random variables (s_l and s_{l+1}) to be orthogonal, when both their expectation values, in addition to their covariance are zero:

$$\mathrm{E}[s_l] = \mathrm{E}[s_{l+1}] = \mathbf{C}_{s_l, s_{l+1}} = 0 \tag{3.7}$$

We can always find an affine, linear transformation from a correlated set of variables to an orthogonal one.

Finally, our two random vectors s_l and s_{l+1} are independent from one another if and only if the joined probability distribution $P(s_l, s_{l+1})$ of both signals are factorisable into the product of their marginal pdfs, $P(s_l)$ and $P(s_{l+1})$:

$$P(s_l, s_{l+1}) = P(s_l)P(s_{l+1}) \tag{3.8}$$

and satisfy the property

$$\mathrm{E}[g(s_l)h(s_{l+1})] = \mathrm{E}[g(s_l)]\mathrm{E}[h(s_{l+1})] \tag{3.9}$$

where $g(s_l)$ and $h(s_{l+1})$ are absolutely integrable functions of s_l and s_{l+1} respectively. From the definition of independence in equation 3.9, we obtain the definition of uncorrelatedness (equation 3.7) in the special case where both s_l and s_{l+1} are linear and are only defined by their covariances, i.e. no higher order statistical moments (Hyvärinen and Oja, 2000; Hyvärinen et al., 2001; Riley et al., 2002). In other words, uncorrelatedness is a special case of independence. Uncorrelated Gaussian random variables are always also independent and the definitions of uncorrelatedness, orthogonality (for zero mean) and statistical independence become equivalent.

Returning to the previous example, if the observed signals comprising $\mathbf{x}$ follow Gaussian distributions, we can define their statistics by the first and second statistical moments (mean and covariance) only. Algorithms such as PCA and FA find a linear transformation from the correlated observed signals, $\mathbf{x}$, to a set of uncorrelated signals, $\mathbf{s}$. Such a linear transformation is always possible and easily achieved using, for example, single value decompositions (see below). An application of PCA to exoplanetary light curve de-trending is given by Thatte et al. (2010).

In the case of the observed signals following non-Gaussian distributions, significant information is contained in the higher statistical moments (skew & kurtosis) and it can be shown that uncorrelated signals (as produced by PCA & FA) are not necessarily mutually independent and hence not optimally separated from one another. Here uncorrelatedness becomes a special case of independence and is a weaker constraint. It can be said that independent signals are always uncorrelated but not vice versa. As a consequence using PCA or FA algorithms may only yield a partially separated result for non-Gaussian sources.

Whitening the data

For non-Gaussian signals, we can now understand uncorrelatedness and orthogonality to be prerequisites to statistical independence. We can hence understand the process of deconvolution of a non-Gaussian process as a two step process: Find an uncorrelated set of vectors using the first two statistical moments, followed by a further deconvolution of the uncorrelated vector space using the information contained in the higher statistical moments. Such a de-correlating 'preprocessing' step is known as *whitening* or *sphering*.

The process of identifying statistical independent components is greatly simplified if the input signals to any ICA algorithm have previously been whitened. Whitening is essentially a transformation of our input data matrix $\mathbf{x}$ into a mean subtracted, $(\mathbf{x} - \bar{\mathbf{x}})$, orthogonal matrix $\tilde{\mathbf{x}}$, where its auto-covariance matrix, $\mathbf{C}_{\tilde{\mathbf{x}}}$, equals the identity matrix, $\mathbf{C}_{\tilde{\mathbf{x}}} = \mathbf{E}[\tilde{\mathbf{x}}\tilde{\mathbf{x}}^{\mathbf{T}}] = \mathbf{I}$. The instantaneous mixing model for the whitened data is now given by

$$\tilde{\mathbf{x}} = \mathbf{C}_{\mathbf{x}}^{-\mathbf{1/2}}(\mathbf{x} - \bar{\mathbf{x}}) = \tilde{\mathbf{A}}\mathbf{s} \tag{3.10}$$

where $\mathbf{C}_{\mathbf{x}}^{-\mathbf{1/2}}$ is the inverse square root of $\mathbf{C}_{\mathbf{x}}$ and $\tilde{\mathbf{A}}$ the corresponding mixing matrix of $\tilde{\mathbf{x}}$. This will be further elaborated in section 3.3.1. The whitening step is not strictly necessary but it can be shown to have two major advantages: 1) The complexity of the mixing matrix $\mathbf{A}$ can be significantly reduced from n^2 parameters, to $n(n-1)/2$ for a whitened, orthogonal matrix $\tilde{\mathbf{A}}$ (Hyvärinen et al., 2001) and 2) using whitening by principal components, we can reduce the dimensionality of the data-set by only maintaining a sub-set of eigenvectors. This reduces possible redundancies of the components comprising the data and prevents the later to be employed ICA algorithm from over-learning for over-complete sets.

Dimensionality reduction using PCA

Whilst the whitening procedure of equation 3.10 usually suffices for smaller data sets, we can easily reduce the dimensions of a data set using PCA. There exist several ways to calculate the PCA of a multivariate data-set. These range from online learning rules of neural networks to their calculation through the covariance of the data (as in equation 3.10). For small to medium sized data sets (typically under 1000 measurements), the single value decomposition (SVD; Pearson, 1901; Manly, 1994; Jolliffe, 2002; Press et al., 2007) is a convenient way to compute PCs. For a given observed data set, $\mathbf{x}$, with dimensions of $N \times k$, we can always obtain the following decomposition

$$\mathbf{x} = \mathbf{ULA}' \qquad \mathbf{U}'\mathbf{U} = \mathbf{I_k} \qquad \mathbf{A}'\mathbf{A} = \mathbf{I_k} \tag{3.11}$$

where **U** and **A** are orthonormal matrices with dimensions of $N \times N$ and $k \times N$ respectively and **L** is a $k \times k$ dimensional diagonal matrix. **A** and **L** describe the eigenvectors and the square rooted eigenvalues respectively whilst **U** contains scaled versions of the principal components. We can now multiply equation 3.11 to the right by **A** to get

$$\mathbf{xA} = \mathbf{ULA}'\mathbf{A} = \mathbf{UL} = \mathbf{z} \tag{3.12}$$

where **z** are the correctly scaled principal components, also referred to as *scores*. The fraction of total variance contained in each principal component by $\Delta_{\sigma^2} = l_k / \sum l_k$, where Δ_{σ^2} is the fractional variance and l_k is the kth eigenvalue in matrix **L**. We can now reconstruct the data by excluding the principal components containing the least variance in the data and hence achieve a dimensionality reduction with minimal loss of information.

3.2.2 Demixing signals using ICA

The basic assumptions of ICA are that the elements comprising **s**, s_l, are mutually independent random variables with probability densities, $p_l(s_l)$. We further assume that all (or at least one) of the probability densities, $p_l(\cdot)$, are non-Gaussian. This non-Gaussianity is key since it allows the de-mixing matrix, **W**, to be estimated. From the central limit theorem, we know that a convolution of any arbitrary probability distribution functions (pdfs) that feature a formal mean and variance, asymptotically approaches a Gaussian distribution in the limit of large N (Riley et al., 2002). In other words, the sum of any two non-Gaussian pdfs (ie. $p_l(\cdot)$ and $p_{l+1}(\cdot)$) is more Gaussian than the respective original pdfs. Therefore by maximising the non-Gaussianity of the individual signals, we maximise their statistical independence. (Comon, 1994; Hyv̈arinen, 1999; Hyvärinen and Oja, 2000; Koldovsky et al., 2006; Hyvärinen et al., 2001; Comon and Jutten, 2010; Stone, 2004).

It is important to note that most observed signals, astrophysical or stellar/instrumental noise, are predominantly non-Gaussian by nature and we can also state that most of these signals should be statistically independent from one another (e.g. an exoplanet light curve signal is independent of the systematic noise of the instrument with which it was recorded). These properties have led to a surge in ICA based analysis methods in current cosmology and extra-galactic astronomy. Here ICA is used to separate the cosmic microwave background (CMB) or signatures of distant

galaxies from their Galactic foregrounds (e.g. Stivoli et al., 2006; Maino et al., 2002, 2007; Wang et al., 2010). Aumont and Macías-Pérez (2007) furthermore separates instrumental noise from the desired astrophysical signal. Other applications include data-compression of sparse, large data sets to improve model fitting efficiencies (e.g. Lu et al., 2006; Delabrouille et al., 2003).

Information Entropy

Although several measures of non-Gaussianity exist (I refer the reader to Cichocki and Amari (2002), Hyvärinen and Oja (2000), Hyvärinen et al. (2001) and Comon and Jutten (2010) for detailed summaries), we here use the concept of 'negentropy' (Brillouin, 1953). Negentropy is derived from the basic information-theoretical concept of differential entropy. In information theory, in close analogy to thermodynamics, the entropy of a system is at its maximum when all data points are at their most random. In thermodynamics we measure the distribution of particles, in information theory it is the probability distribution of a random variable. From information theory we can derive the fundamental result that a Gaussian distribution has the largest entropy among all random variables of equal variance and known mean (see proof in appendix 3.7.1). Hence, by minimising the entropy of a variable, we maximise its non-Gaussianity. For a random vector $\mathbf{y}$, with random variables y_i, $i = 1, ..., n$, the entropy is given by

$$\mathrm{H}(\mathbf{y}) = -\int p(\mathbf{y}) \log_2 p(\mathbf{y}) \mathrm{d}\mathbf{y} \tag{3.13}$$

where $\mathrm{H}(\mathbf{y})$ is the differential or Shannon entropy (Shannon, 1948) and $p(\mathbf{y})$ is the pdf of the random vector $\mathbf{y}$. $\mathrm{H}(\mathbf{y})$ is at its minimum when $p(\mathbf{y})$ is at its most non-Gaussian. We can now normalise equation 3.13 to yield the definition of negentropy:

$$\mathrm{J}(\mathbf{y}) = \mathrm{H}(\mathbf{y}_{gauss}) - \mathrm{H}(\mathbf{y}) \tag{3.14}$$

where $\mathbf{y}_{gauss}$ is a random Gaussian vector with the same covariance matrix as $\mathbf{y}$. Now $\mathbf{y}$ is at its most non-Gaussian when $\mathrm{J}(\mathbf{y})$ is at its maximum. It is important to note that negentropy is in-sensitive to a multiplication by a scalar constant (see appendix 3.7.2 for proof). However, the normalisation process required by the ICA iteration scheme has the important consequence of introducing a sign and scaling ambiguity into the retrieved signal components of $\mathbf{s}$. I will discuss the implications of this limitation in section 3.3.3.

Contrast functions

In practice it is very difficult to calculate the negentropy of a system and various methods were devised to approximate J($\mathbf{y}$). The classic method is to measure the kurtosis of mean-subtracted $\mathbf{y}$ with unit variance. However, kurtosis is very prone to distortions by outliers in the data and hence lacks the robustness required as measure of negentropy (Hyvärinen et al., 2001). To overcome this limitation, more robust measures of negentropy have been devised. One can approximate negentropy by equation 3.15 (Hyvärinen et al., 2001; Hyv̈arinen, 1999; Comon and Jutten, 2010; Stone, 2004)

$$\mathrm{J}(\mathbf{y}) \propto (\mathrm{E}[G(\mathbf{y})] - \mathrm{E}[G(\nu)])^2 \tag{3.15}$$

where ν is a random Gaussian variable with zero mean and unit variance and G is a non-quadratic contrast function chosen to approximate the underlying probability distribution. There are usually three types of contrast functions: G_1 as general purpose function, G_2 optimised for super-Gaussian (leptokurtic) distributions and G_3 optimised for sub-Gaussian (platykurtic) distributions (Hyv̈arinen, 1999; Hyvärinen et al., 2001; Comon and Jutten, 2010):

$$\begin{aligned} G_1(y) &= \frac{1}{a_1} log[\cosh(a_1 y)] \\ G_2(y) &= -\exp(-y^2/2) \\ G_3(y) &= \frac{1}{4} y^4 \end{aligned} \tag{3.16}$$

The choice of contrast function is only important if one wants to optimise the performance of the ICA algorithm as it is done for the EFICA (Koldovsky et al., 2006) algorithm where choices of contrast functions are tried iteratively.

Deflationary FastICA

Finally, we can maximise the negentropy given in equation 3.15 by finding a unit vector $\mathbf{w}$ that maximises the non-Gaussianity of the projection $y_i = \mathbf{w^T}\tilde{\mathbf{x}}$, so that J($\mathbf{w^T}\tilde{\mathbf{x}}$) is at its maximum. For the fixed-point FastICA algorithm, this can be achieved by the simple iteration scheme (Hyv̈arinen, 1999; Hyvärinen and Oja, 2000):

1. Choose initial (i.e. random) weight vector $\mathbf{w}$
2. Increment: $\mathbf{w}^+ = \mathrm{E}[\tilde{\mathbf{x}} g(\mathbf{w}^T\tilde{\mathbf{x}})] - \mathrm{E}[g^{'}(\mathbf{w}^T\tilde{\mathbf{x}})]\mathbf{w}$

3. Normalise: $\mathbf{w} = \mathbf{w}^{+}/||\mathbf{w}^{+}||$

4. Repeat steps 2 & 3 until converged

where g and $g^{'}$ are the first and second derivatives of the contrast function $G(\cdot)$:

$$\begin{aligned} g_1(y) &= \tanh(a_1 y) \\ g_2(y) &= y\exp(-y^2/2) \\ g_3(y) &= y^3 \end{aligned} \tag{3.17}$$

The iteration scheme above estimates only one weight vector at a time. This is called a deflationary algorithm where the computed IC is subtracted from the data before the second IC is computed. This serial scheme has the advantages of not requiring knowledge of how many non-Gaussian sources are present in the data. Estimating one source at a time is very closely related to the concept of Projection Pursuit which does not require an underling data model. The disadvantage of this serial approach is propagation of estimation errors. Since the individual IC is subtracted from the data before the next is computed, any errors occurring in the estimation process will be passed down the line to the next estimation via the orthogonalisation step in the iteration scheme. This is a cumulative effect and can severely corrupt the estimation accuracy of weaker signals (Hyvärinen et al., 2001; Comon and Jutten, 2010).

Projection Pursuit and Deflationary ICA

In the mid-1980's, Huber (1985) and Friedman (1987) realised that in most cases, the non-Gaussian signals in any data set are the most "interesting". From an information entropy point of view, as discussed above, a high entropy indicates a lack of structure and it is hence more interesting to project multidimensional data along the vectors of lowest entropy and highest amount of structure. This realisation came significantly before the development of ICA and we can understand the deflationary ICA algorithm to be equivalent to a numerical implementation of PP. As opposed to ICA, PP does not need an underlying data model and no assumptions about independent components are made. If the ICA model holds, optimising the ICA non-Gaussianity measures produces independent components; if the model does not hold, then what we get are the projection pursuit directions (Hyvärinen and Oja, 2000; Stone, 2004). This is an important point to make with regards to time-consecutive transit observations, which break the underlying ICA assumptions otherwise and will be discussed in section 4.5.

Parallel-FastICA and Efficient ICA

The problem of the cumulative error build up in the deflationary FastICA algorithm can be circumvented by estimating all ICs simultaneously. This parallel FastICA is similar to the single unit iteration, the whitened demixing matrix $\tilde{\mathbf{W}}$ is at its most mutually independent when the projection $\mathbf{y} = \tilde{\mathbf{W}}^{\mathbf{T}}\tilde{\mathbf{x}}$ is at its most non-Gaussian. The FastICA fixed point iteration step is then given by

$$\tilde{\mathbf{W}}^{+} \leftarrow \mathbf{g}(\tilde{\mathbf{W}}\tilde{\mathbf{x}})\tilde{\mathbf{x}}^{\mathbf{T}} - \mathrm{diag}[\mathbf{g}'(\tilde{\mathbf{W}}\tilde{\mathbf{x}})\mathbf{1_N}]\tilde{\mathbf{W}} \tag{3.18}$$

where $\tilde{\mathbf{W}}^{+}$ is the unnormalised next iteration of $\tilde{\mathbf{W}}$, $\mathbf{1_N}$ is an N x 1 vector of 1's and $g(.)$ and $g'(.)$ are the first and second order derivatives of the nonlinear function $G(.)$, equation 3.17. This is followed by a symmetric orthogonalisation step:

$$\tilde{\mathbf{W}} \leftarrow (\tilde{\mathbf{W}}^{+}\tilde{\mathbf{W}}^{+\mathbf{T}})^{-\mathbf{1/2}}\tilde{\mathbf{W}}^{+} \tag{3.19}$$

Equations 3.18 & 3.19 are iterated until the result has converged.

For a full derivation we refer you to Hyvärinen (1999) and Hyvärinen et al. (2001). Whereas the convergence of the FastICA algorithm is often dependent on the non-linearity chosen by the user, the EFICA (Koldovsky et al., 2006) algorithm employed here is a variant of the above iteration scheme and allows for different non-linearities to be assigned adaptively to different sources. Koldovsky et al. (2006) showed that EFICA is asymptotically efficient, i.e. reaches the Cramer-Rao Lower Bound (CRLB) in an ideal case where the nonlinearity $G(.)$ equals the score function. In other words, the algorithms employed here can be shown to converge to the correct solution given the original source signals and in the limit of $N \to \infty$ iterations.

In reality the number of iterations is finite and and imperfect convergence results in traces of other sources to remain in the individual signals comprising $\mathbf{s}$. We can hence state that equation 3.4 becomes

$$\mathbf{W} \simeq \mathbf{A}^{-\mathbf{1}} \tag{3.20}$$

A measure of this error is the deviation of $\mathbf{WA}$ (or $\tilde{\mathbf{W}}\tilde{\mathbf{A}}$ for the whitened case) from the unity matrix by inspecting the variance of its elements (Koldovsky et al., 2006; Hyvärinen et al., 2001).

To assert a good degree of separation, we can define $\mathbf{G}$ as the gain matrix. For a perfectly estimated de-mixing matrix, $\mathbf{W}$, the gain matrix is equal to its identity matrix

$$\mathbf{G} = \mathbf{WA} = \mathbf{I} \tag{3.21}$$

In signal processing, the performance of blind-source separation algorithms is usually measured by the interference over signal ratio matrix, **ISR**. The ISR is the standard measure in signal processing of how well a given signal has been transmitted or de-convolved from a mixture of signals. It can be understood as the inverse of the signal-to-noise ratio (SNR). The higher the ISR for a specific signal, the less well has it been separated from the original mixture. For a real case example, **A** is unknown and the ISR needs to be estimated.

$$\mathbf{ISR}_{kl} = \frac{\mathbf{G}_{kl}^2}{\mathbf{G}_{kk}^2},\ k, l = 1, 2, ..., d \tag{3.22}$$

where k and l denote the observed and estimated sources. The ISR for and individual observed signal k is given by

$$\mathbf{isr}_k = \frac{\sum_{l=1, l\neq k}^{d} \mathbf{G}_{kl}^2}{\mathbf{G}_{kk}^2},\ k = 1, 2, ..., d \tag{3.23}$$

However, the original mixing matrix, **A**, is not generally known for real data sets and equations 3.22 & 3.23 are only useful in the case of simulations. Tichavsky et al. (2006) have shown that the whole **ISR** matrix for the EFICA algorithm can be approximated by

$$\mathbf{ISR}_{kl}^{EF} \simeq \frac{1}{N} \frac{\gamma_k(\gamma_l + \tau_l^2)}{\tau_l^2\gamma_k + \tau_k^2(\gamma_l + \tau_l^2)} \tag{3.24}$$

$$\begin{aligned} \gamma_k &= \beta_k - \mu_k^2 \\ \mu_k &= E[\hat{s}_k g_k(\hat{s}_k)] \\ \tau_k &= |\mu_k - \rho_k| \\ \rho_k &= E[g_k'(\hat{s}_k)] \\ \beta_k &= E[g_k^2(\hat{s}_k)] \end{aligned} \tag{3.25}$$

where $\hat{\mathbf{s}}_k$ and $\hat{\mathbf{s}}_l$ are the kth and lth observed and estimated signals of **s** in equation 3.3, $g_k(.)$ and $g_k'(.)$ the first and second derivative of $G(.)$ for signal k and N is the number of signals estimated. Here it should be mentioned that, of course, the true realisation of each **ISR** component is unknown and a mean-**ISR** is computed leading to the best 'on average' separation of the signals.

WASOBI

Whilst EFICA is optimised for the separation of instantaneously mixed, non-Gaussian sources, second-order statistics blind-source-separation algorithms rely on time-structure in the sources' correlation function to estimate $\tilde{\mathbf{W}}$. A variety of algorithms exist in the literature, here we use a derivative of the popular SOBI algorithm (Belouchrani et al., 1997), WASOBI (Yeredor, 2000; Tichavský et al., 2006) to separate Gaussian auto-regressive (AR) sources in the input data $\tilde{\mathbf{x}}$. WASOBI is an asymptotically efficient version of the SOBI algorithm (Belouchrani et al., 1997), and is geared towards separating Gaussian auto-regressive (AR) and time-correlated components. It uses second-order statistics and can be understood to be similar to principal component analysis. The use of both, EFICA and WASOBI, algorithms is necessary since a real life data set will always contain a mixture of both, non-Gaussian and Gaussian AR processes.

In the case of WASOBI, the blind source separation follows the same linear model as in equation 3.3 and the mixing matrix $\tilde{\mathbf{A}}$ is estimated by a joint diagonalisation of the signals' autocorrelation matrices. The unknown correlation matrices of the observed signals for a given lag τ, $\mathbf{R}_x[\tau]$

$$\mathbf{R}_x[\tau] \stackrel{\triangle}{=} \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}[n]\mathbf{x}^T[n+\tau], \quad \tau = 0, ..., M-1 \tag{3.26}$$

satisfies the relation

$$\mathbf{R}_x[\tau] = \tilde{\mathbf{A}}\mathbf{R}_\mathbf{s}[\tau]\tilde{\mathbf{A}}^\mathbf{T}, \quad \forall\tau \tag{3.27}$$

where $\mathbf{R}_s[\tau] \stackrel{\triangle}{=} E[\mathbf{s}[n]\mathbf{s}^T[n+\tau]]$ are the source signals' diagonalised correlation matrices (Yeredor, 2000). Hence, if the correlation matrices are diagonal, ie. the off-diagonal components are zero, the separated signals can be said to be independent from each other. The SOBI & WASOBI algorithms estimate $\tilde{\mathbf{A}}$ as the joint diagonaliser of a set of correlation matrices. Similar to the EFICA code, we can define an asymptotic estimate of the **ISR** matrix

$$\mathbf{ISR}_{kl}^{WA} \simeq \frac{1}{N}\frac{\phi_{kl}}{1-\phi_{kl}\phi_{lk}}\frac{\sigma_k^2 R_l[0]}{\sigma_l^2 R_k[0]} \tag{3.28}$$

$$\phi_{kl} \stackrel{\triangle}{=} \frac{1}{\sigma_k^2}\sum_{i,j=0}^{M-1} a_{il}a_{jl}R_k[i-j] \tag{3.29}$$

where k and l denote the observed and the estimated sources, $\{R_k[\tau]\}_{\tau=0}^{M-1}$ is the covariance sequence of the k-th source, σ_k^2 is the variance of the source and $\{a_{il}\}_{i=0}^{M-1}$ are the auto-regression coefficients of the l-th source (Tichavský et al., 2006).

3.2.3 ICA in the context of exoplanetary lightcurves

It now remains to bring the previous sections into the context of exoplanetary lightcurve analysis, which requires an adapted definition of the instantaneous mixing model (equation 3.2) to our purposes. Consider multiple time series observations of the same exoplanetary eclipse signal either in parallel, by performing spectrophotometry with a spectrograph, or consecutive in time (as explained in section 4.5).

Without excluding the most general case, let us focus on a time-resolved spectroscopic measurement of an exoplanetary eclipse. For most observations, the signal recorded is a mixture of astrophysical signal, Gaussian (white) noise and systematic noise components originating from instrumental defects and other sources such as stellar activity and telluric fluctuations. We can therefore write the individual time series from equation 3.2, as sum of the desired astrophysical signal, s_a, systematic (non-Gaussian) noise components, s_{sn}, and Gaussian noise, s_{wn}. We can now define the underlying linear model of our time series data to be

$$x(t) = a_1 s_a(t) + a_2 s_{sn1}(t) + a_3 s_{sn2}(t) + ... + s_{wn}(t) \tag{3.30}$$

or

$$x_k = a_{k1}\mathbf{s}_a + \sum_{l_2=1}^{N_{sn}} a_{kl_2}\mathbf{s}_{sn,l_2} + \sum_{l_3=1}^{N_{wn}} a_{kl_3}\mathbf{s}_{wn,l_3} \tag{3.31}$$

where N_{sn} and N_{wn} are the number of systematic and white noise components respectively and $N = N_{sn} + N_{wn} + 1$ assuming only one component is astrophysical.

3.3 The algorithm

Following from the discussion above, we can understand the signal de-mixing to be a two step process: de-correlation of the Gaussian components in the observed data using PCA, followed by the de-correlation of non-Gaussian components using ICA and WASOBI algorithms. The de-correlation of Gaussian components to form a new uncorrelated vectors set can be understood as a pre-processing step to the ICA procedure. The algorithm proposed here consists of five main parts: 1) Pre-processing of the observed data, 2) Signal separation, 3) Signal reconstruction 4) Lightcurve fitting and 5) Post-analysis. Figure 3.1 lays out the individual processing steps of the algorithm. A more detailed description of the software design can be found in Appendix 3.7.4.

3.3.1 Signal pre-processing

In section 3.2.1, we introduced the concept of *whitening* or *sphering*. Since uncorrelatedness is a prerequisite of statistical independence, it is useful to decorrelate the data using the first two statistical moments before proceeding to the non-Gaussian signal separation. From equation 3.10 (repeated here for clarity as equation 3.32) we can see that the inverse square-root of the covariance matrix $\mathbf{C_x}$ needs to be calculated in order to de-correlate the observed signal vector $\mathbf{x}$ to the whitened vector $\tilde{\mathbf{x}}$.

$$\tilde{\mathbf{x}} = \mathbf{C_x^{-1/2}}(\mathbf{x} - \bar{\mathbf{x}}) = \tilde{\mathbf{A}}\mathbf{x} \tag{3.32}$$

The covariance matrix of $\mathbf{x}$, $\mathbf{C_x}$, is given by $\mathbf{C_x} = \mathbf{EDE^T}$, where $\mathbf{E}$ is the matrix of eigenvectors and $\mathbf{D}$ the diagonal matrix of eigenvalues, $\mathbf{D} = diag(d_1, d_2, ..., d_n)$. Using principal component analysis (PCA), we can compute $\mathbf{E}$ and $\mathbf{D}$ and hence the whitening matrix $\mathbf{C_x^{-1/2}}$ is given by equation 3.33 (Hyvärinen et al., 2001; Jolliffe, 2002).

$$\mathbf{C_x^{-1/2}} = \mathbf{ED^{-1/2}E^T} \tag{3.33}$$

where $\tilde{\mathbf{W}} \triangleq \tilde{\mathbf{A}}^{-1}$ and is the de-mixing matrix of the whitened observed signals $\tilde{\mathbf{x}}$.

It is useful to note that this transformation always exists and is quick to calculate. Several forms of PCA are available in the literature and are suited for different sized data sets. For data sets considered here and in the following chapters, PCA based on single value decomposition (SVD) is ideal due to the simplicity of its code. For larger data sets, online-learning and neural-network PCA methods are the preferred choice (Jolliffe, 2002). Using PCA to calculate the whitening matrix also allows us to determine the dimensions of the input data to subsequent steps of the algorithm. For larger data-sets that do contain fewer non-Gaussian signals than

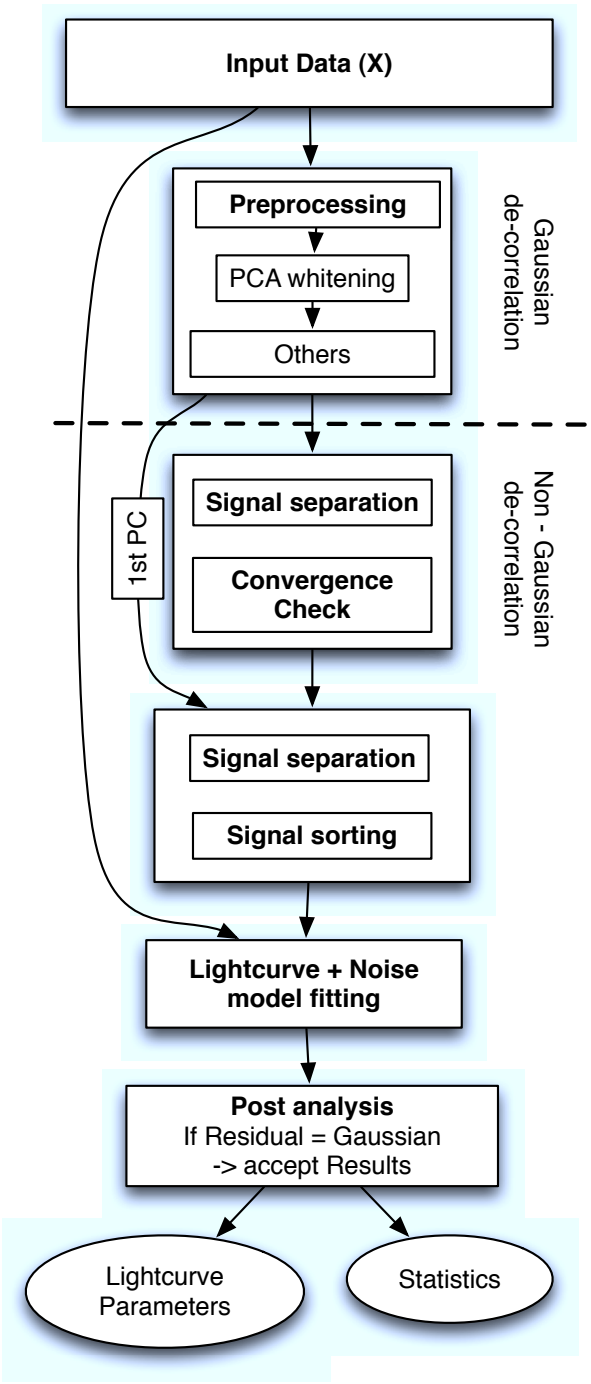


Figure 3.1: Flowchart illustrating the algorithm. The input data is first transformed into an orthogonal set using PCA. The latent signals comprising the input data are then separated using the MULTI-COMBI algorithm which is followed by a signal sorting step. The separated lightcurve and systematic noise components are then fitted to the original data.

observations in $\mathbf{x}$ it is often useful to reduce the dimensionality of the data-set to the number of non-Gaussian signals in the data, which prevents the ICA algorithm from a common problem of 'over-learning' (Hyvärinen et al., 2001). This dimensionality reduction is easily done by only retaining the eigenvectors with the highest eigenvalues associated to them.

We also like to note that any type of additional linear signal cleaning or pre-processing step, such as those described by Carter and Winn (2009); Waldmann et al. (2012b), are allowed. Linear data filtering or cleaning can be understood as multiplying equation 3.3 from the left with a linear transformation $\mathbf{B}$ to get: $\mathbf{Bx} = \mathbf{BAx}$. The underlying data model assumed in this paper is hence not affected.

3.3.2 Signal separation

After the observed signals have successfully been whitened ($\tilde{\mathbf{x}}$), we estimate the mixing matrix of the whitened signal, $\tilde{\mathbf{A}}$, using the MULTI-COMBI algorithm (Tichavsky et al., 2006). MULTI-COMBI is comprised of two algorithms, EFICA (Koldovsky et al., 2006) and WASOBI (Yeredor, 2000), which were introduced in section 3.2.2.

These algorithms are highly complementary to each other. Whilst EFICA has an asymptotically efficient performance in separating non-Gaussian instantaneous mixtures, WASOBI is asymptotically efficient in separating Gaussian time-correlated signals. Both these properties are necessary since a real data set will have both of the aforementioned properties and its components would hence not be optimally de-mixed if one would only employ one type of algorithm. MULTI-COMBI (Tichavsky et al., 2006) uses a clustering technique in which both algorithms are run on the set of unseparated sources $\tilde{\mathbf{x}}$ and their interference over signal matrices, $\mathbf{ISR}^{\mathbf{EF}}$ and $\mathbf{ISR}^{\mathbf{WA}}$, are estimated. The signals are then clustered depending on whether their specific $\mathbf{ISR_{kl}}$ is lower for the EFICA or WASOBI case. Then, the process is repeated until all clusters are singeltons, i.e. only contain one signal per cluster, and the signals are hence optimally separated.

Convergence check

From the MULTI-COMBI algorithm, we obtain the estimated signal matrix $\hat{\mathbf{s}}$, an overall $\mathbf{ISR}$ matrix as well as final $\mathbf{ISR}^{\mathbf{EF}}$ and $\mathbf{ISR}^{\mathbf{WA}}$. Since the algorithms used here use fixed-point convergence techniques, the problem of non-repeatability of the separation process is less than for neural network based approaches. However, it is common sense to check the stability of the result obtained and to estimate the error on $\hat{\mathbf{s}}$.

In order to estimate the stability of the convergence, we perturb the unknown mixing matrix

$\mathbf{A}$ with a random and known mixing matrix $\mathbf{P}$ to give a new mixing matrix $\mathbf{A_2} = \mathbf{PA}$ and equation 3.3 becomes: $\mathbf{x} = \mathbf{PAs} = \mathbf{A_2 s}$. This is equivalent to multiplying the whitened signal $\tilde{\mathbf{x}}$ with $\mathbf{P}$

$$\tilde{\mathbf{x}}_\mathbf{2} = \mathbf{P}\tilde{\mathbf{x}} = \mathbf{P}\mathbf{C}_\mathbf{x}^{-\mathbf{1/2}}(\mathbf{x} - \bar{\mathbf{x}}) = \tilde{\mathbf{A}}_\mathbf{2}\mathbf{s} \tag{3.34}$$

We re-run the separation step and estimate $\mathbf{A_2}$. Since $\mathbf{P}$ is known, we can reconstruct the original mixing-matrix and compare it with the new result. In the scope of an automated algorithm, the sum of all terms of $\mathbf{ISR_A}$ is compared to the sum of $\mathbf{ISR_{A2}}$ and the result is reported.

To identify the stochastic nature of the retrieval we furthermore re-run the separation step with the same whitened signal, $\tilde{\mathbf{x}}$, akin to a Monte Carlo simulation. We perform i realisations (where $i = 10 - 100$ typically) and use the de-mixing matrices $\tilde{\mathbf{W}}_\mathbf{i}$ to construct mean noise models later on. This way, we propagate the signal separation error to the model-fitting in a coherent manner.

3.3.3 Signal reconstruction

Once the mixing matrix, $\tilde{\mathbf{A}}$ is estimated, we need to identify which signals are astrophysical, which ones are white and which are systematic noise. This is done in a two step process:

1) We construct the estimated signal matrix, $\hat{\mathbf{s}}$, and for its individual components $\hat{\mathbf{s}}_\mathbf{l}$ compute the Pearson correlation coefficient between $\hat{\mathbf{s}}_\mathbf{l}$ and the first principal component of the PCA decomposition in section 3.3.1. For medium signal to noise (SNR) observations, the first principal component (PC), i.e. the one with the highest eigenvalue associated to it, will contain the predominant lightcurve shape. As previously discussed, the first PC is not perfectly separated from the systematic signals and hence cannot be used directly for further analysis but it is good enough to use it as lightcurve identification. The identified lightcurve signal is labeled $\hat{\mathbf{s}}_\mathbf{a}$.

2) Once the lightcurve signal is identified, we exclude this row from $\hat{\mathbf{s}}$ and proceed to classify the remaining signals with respect to their non-Gaussianity (ie. systematic noise sources). Here we use the Ljung-Box portmanteau test (see Appendix 3.7.3 and Brockwell and Davis, 2006) to test for the hypothesis that the time series is statistically white. This test was originally designed to check the residuals of auto-regressive moving-average (ARMA) models for significant departures from uncorrelatedness. It is hence ideally suited for our need to identify which estimated signal components are the desired non-Gaussian ones.

The identified non-Gaussian, systematic noise, signals are hence labeled $\hat{\mathbf{s}}_\mathbf{sn}$ and the remaining white noise signals $\hat{\mathbf{s}}_\mathbf{wn}$ to give

$$\hat{\mathbf{s}}_\mathbf{a} + \hat{\mathbf{s}}_\mathbf{sn} + \hat{\mathbf{s}}_\mathbf{wn} \triangleq \hat{\mathbf{s}} = \tilde{\mathbf{W}}\tilde{\mathbf{x}} \tag{3.35}$$

where the de-mixing matrix is given by $\tilde{\mathbf{W}} = \tilde{\mathbf{A}}^{-1}$.

As previously mentioned, the components of $\hat{\mathbf{s}}$ have ambiguities in scaling and sign and can be thought to be similar to the eigenvectors of a principal component analysis with missing eigenvalues. Fortunately there exist two approaches to resolving this degeneracy:

1. In the case of $\hat{\mathbf{s}}_\mathbf{a}$ being well separated as individual component, we can take $\hat{\mathbf{s}}_\mathbf{a}$ and the de-mixing matrix $\tilde{\mathbf{W}}$ and only retain the row containing the astrophysical signal component forming the row-vector $\tilde{\mathbf{w}}_\mathbf{a}$. We then reconstruct the original data $\tilde{\mathbf{x}}$ using only the separated signal component:

$$\tilde{\mathbf{x}}_\mathbf{a} = \tilde{\mathbf{w}}_\mathbf{a}^{-1}\hat{\mathbf{s}}_\mathbf{a} = \tilde{\mathbf{w}}_\mathbf{a}^{-1}\tilde{\mathbf{W}}\tilde{\mathbf{x}} \tag{3.36}$$

   where $\tilde{\mathbf{x}}_\mathbf{a}$ is the reconstructed whitened data with all but the astrophysical signal components removed. Using equation 3.10, we can now calculate the un-whitened matrix $\mathbf{x}_\mathbf{a}$.

$$\mathbf{x}_\mathbf{a} = \mathbf{z}(\mathbf{x} - \bar{\mathbf{x}}) + \bar{\mathbf{x}} \tag{3.37}$$

$$\mathbf{z} = \tilde{\mathbf{w}}_\mathbf{a}^{-1}\tilde{\mathbf{W}} \tag{3.38}$$

   Hence we can think of $\mathbf{z}$ as a linear, optimal filter for the signal component in $\mathbf{x}$. Please note that this linear filtering does not impair the scaling information as this is re-instated going from $\hat{\mathbf{s}}_\mathbf{a}$ to $\mathbf{x}_\mathbf{a}$.

2. In the case of $\hat{\mathbf{s}}_\mathbf{a}$ not being well separated but other systematic noise components are, a different, more indirect approach can be used. Here, the systematic noise components, $\hat{\mathbf{s}}_\mathbf{sn}$ which do not contain sign or scaling information, are simultaneously fitted to the time series data (preferably out-of-transit data), x_k. We therefore define the systematic noise model for an individual time series by m_k,

$$m_k = \mathbf{O}_\mathbf{k}\hat{\mathbf{s}}_\mathbf{sn} \tag{3.39}$$

   where $\mathbf{O}_\mathbf{k}$ is a $N_{sn} \times N_{sn}$ diagonal scaling matrix of $\hat{\mathbf{s}}_\mathbf{sn}$, which needs to be fitted iteratively as free parameters in the following section. We furthermore define $\mathbf{m}_\mathbf{sn}$ to be a column vectors composed of all individual noise models m_k.

3.3.4 Lightcurve fitting

Having either filtered the data to obtain $\mathbf{x_a}$ or constructed the noise model $\mathbf{m_{sn}}$, we can now fit the original time series, x_k using the standard analytical lightcurve models (Mandel and Agol, 2002; Seager and Mallén-Ornelas, 2003) in addition to the diagonal matrix $\mathbf{O}$, if necessary. For the purpose of this chapter, which focuses on blind-source-separation, we will restrict ourselves to demonstrating the feasibility of estimating $\mathbf{O}$ and only leave the transit depth as variable lightcurve parameter. We use the analytic lightcurve model by Mandel and Agol (2002) and a amoeba minimisation algorithm (Press et al., 2007). For real data applications, we advise the reader to use Markov Chain Monte Carlo methods, or similar, which have become standard in the field of exoplanets and allow the estimation of the posterior probability distributions and their associated errors (Bakos et al., 2007; Burke et al., 2007; Collier Cameron et al., 2007; Ford, 2006; Gregory, 2011a).

3.3.5 Post-analysis

Once the model fitting stage has been completed, we are left with fitting residual, r_k, i.e. $r_k = x_k - m_k$. Several tests are useful to determine how well the signals have been removed from the original time series, x_k. In the case of a full Markov Chain Monte Carlo fitting, the posterior distributions of the fitting parameters may be used to asses the impact of the remaining systematic noise in the data when compared to a simulated data set only containing white noise. Portmanteau tests on individual time series are useful to test for non-white noise signals and allow a measure of the overall performance of the algorithm (Brockwell and Davis, 2006). Additionally, we can determine the Kullback-Leibler divergence (Kullback and Leibler, 1951) of our residual's probability distribution function (pdf) to an idealised Gaussian case, see appendix 3.7.1.

For the simulations and real-data examples presented in the following section, we have merely plotted the autocorrelation functions (ACF) of the residuals obtained to determine whether for a given lag, these are within the 3σ confidence limit of the residual being dominated by white-noise (Brockwell and Davis, 2006; Davison, 2009). Here the ACF is given by:

$$ACF(k,\tau) = \frac{1}{m}\sum_{t=1}^{m-\tau}(r_{k,t} - \bar{r}_k)(r_{k,t+\tau} - \bar{r}_k) \tag{3.40}$$

$$\tau = 0, 1, 2, 3, ...m/2$$

where m is the number of data points in the time series, τ the specific lag and the confidence

intervals are given by $\pm\sigma/\sqrt{m}$.

3.4 Simulations

In order to test the behaviour and efficiency of the algorithm described above, we produced a toy model simulation with five observed signals: 1) a secondary eclipse Mandel and Agol (2002) lightcurve; 2) a sinusoidal signal; 3) a sawtooth function; 4) a fourth order auto-regressive signal to simulate time-correlated signals; 5) Gaussian noise with a full width half maximum (FWHM) of 0.01 magnitudes. The premixed signals are displayed in figure 3.2. This gives us our signal matrix, $\mathbf{s}$, which needs to be recovered later on. We have then proceeded to mix the signals in figure 3.2 using a random mixing matrix, $\mathbf{A}$, to obtain our 'observed signals', $\mathbf{x}$, in figure 3.3. For the sake of comparability we keep the mixing matrix $\mathbf{A}$ to be the same for all simulations.

We now subdivide the simulations to illustrate the two possible methods of the signal reconstruction. *Method* 1 computes $\mathbf{x_a}$ using equation 3.37 whilst *Method* 2 fits the noise model $\mathbf{m_{sn}}$ (equation 3.39) simultaneously with the Mandel and Agol (2002) lightcurve. These two examples demonstrate that both techniques work equally well for a well behaved data set.

3.4.1 Method 1: Filtering out the signal

In this example, we use the 'observed' signals in figure 3.3 as input to the algorithm. We however do not perform a dimensionality reduction using PCA since we are not dealing with an over-complete set in this example. The results of the separation are shown in figure 3.5. Here the top three, red lightcurves are the estimated systematic noise components as identified by the algorithm. The fourth component is Gaussian noise and the bottom is an inverse of the lightcurve signal. It should again be noted here that the blind-source separation does not preserve the scaling nor the signs of the signals in $\hat{\mathbf{s}}$. However, when the original data is reconstructed using only the signal component, $\hat{\mathbf{s}}_\mathbf{a}$, to obtain $\mathbf{x_a}$ (equation 3.37), the scaling and sign informations are re-instated. For a well behaved data set, i.e. one that obeys the instantaneous mixing model and has negligible Gaussian noise in their signal components, it is therefore possible to re-construct the lightcurve signal from the raw data as explained in section 3.3.3. Figure 3.4 shows the top lightcurve of figure 3.3 (blue circles) and overplotted the retrieved signal component (red crosses) and offset below the systematic noise component (black squares).

As a useful by-product of the algorithm, we obtain the interference over signal matrices (ISR, equations 3.24 & 3.28) for both the EFICA and WASOBI algorithms. These give us valuable information on the efficiency at which the signals have been separated. Figure 3.6 shows the

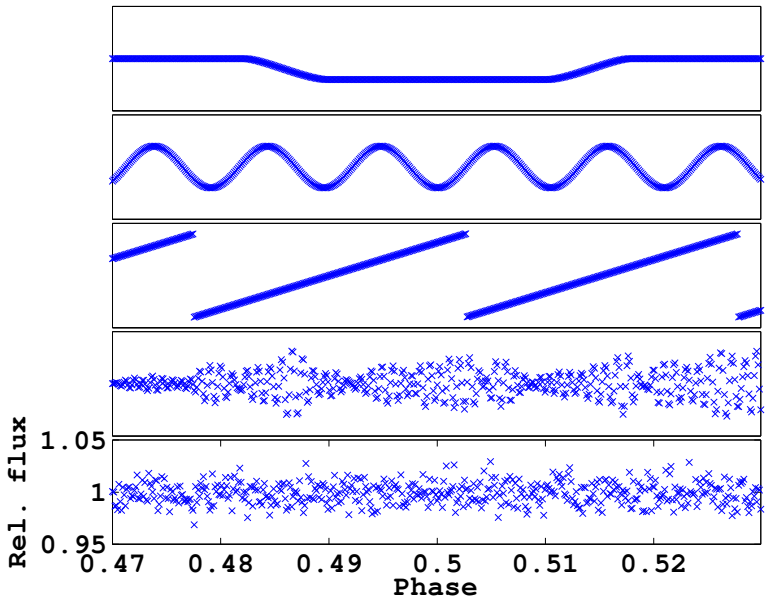


Figure 3.2: Simulated input signals before mixing. From top to bottom: 1) secondary eclipse Mandel and Agol (2002) curve, 2) sinusoidal function, 3) sawtooth function, 4) time-correlated auto-regressive function, 5) Gaussian noise. The scaling of the ordinate is identical for all subplots.

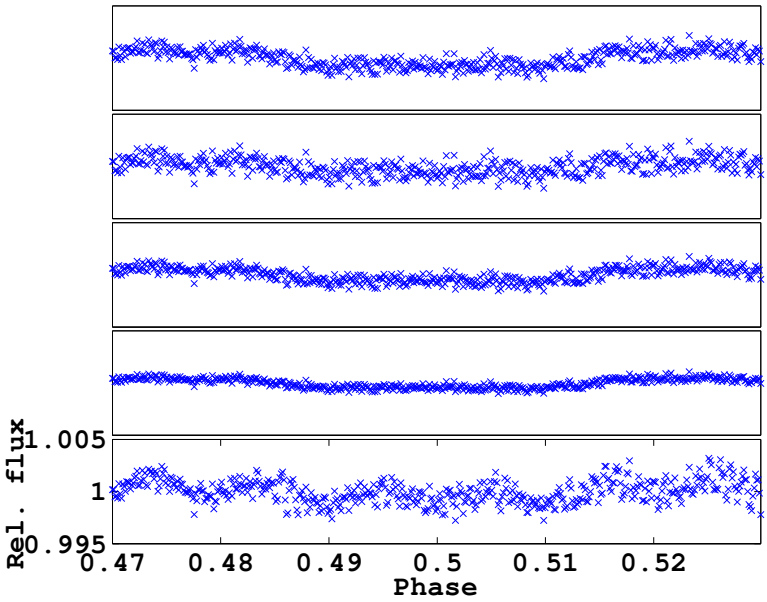


Figure 3.3: The signals, **s**, in figure 3.2 were mixed using a random mixing matrix **A** to obtain the 'observed signals', **x** normalised to unity, shown in this diagram. The algorithm takes the lightcurves in this diagram as starting values. No further input is provided or assumptions on the underlying signals made. The scaling of the ordinate is identical for all subplots.

Hinton diagrams of the EFICA and WASOBI **ISR** matrices. Here, the smaller the off-diagonal elements of the matrix, the better the signal separation. In this example, the EFICA algorithm outperforms the WASOBI one, which is to be expected since all signals but one are non-Gaussian.

3.4.2 Method 2: Fitting a noise model to the data

In the previous section, we have shown that in the case that the astrophysical component $\hat{\mathbf{s}}_\mathbf{a}$ is well separated as an individual signal, we can create a filter for the raw data that directly filters the lightcurve signal from the noise. However, in most real data applications, $\hat{\mathbf{s}}_\mathbf{a}$, is not perfectly separated but the components of $\hat{\mathbf{s}}_\mathbf{sn}$ may be more so. In this case we can construct the noise model $\mathbf{m}_\mathbf{sn}$ given by equation 3.39 and the diagonal elements of $\mathbf{O}$ are fitted as described in section 3.3.4. The starting position of the algorithm is the same as for the previous example (figure 3.3). The model fit of the first lightcurve in figure 3.3 and its residuals are shown in figure 3.7. The autocorrelation function for 100 lags is plotted in figure 3.8. All but two lags are within the 3σ confidence limit that the residual is white noise dominated, indicating that all signals have been removed effectively.

Finally we simulate the convergence properties of both EFICA and WASOBI under varying white noise conditions. Here we repeatedly run the algorithm until signal separation is completed and record the mean ISRs of the source separation. We performed this simulation 300 times for Gaussian noise FWHMs varying from 0.0 - 0.3 magnitudes (figure 3.7 has a $\mathrm{FWHM}_{gauss} = 0.01$) and every ISR measurement reported is the mean of 10 iterations. Figure 3.9 summarises the results. Here, the red circles represent the mean ISR of the EFICA algorithm and the blue crosses that of WASOBI. It can clearly be seen that for this example the EFICA algorithm outperforms WASOBI and on average reaches lower ISR values. We can further note that the blind source separation is not significantly affected by the magnitude of the white noise and performs well under difficult signal to noise conditions.

3.4.3 Breaking the instantaneous mixing model

In the previous examples we assumed that the instantaneous mixing model (equations 3.2 & 3.3) holds perfectly and all Gaussian noise is presented by a single component. This assumption is generally valid and the instantaneous mixing model can in fact always be set up this way. However, one can imagine instrumental noise (for example) to be both, systematic and Gaussian. In order to test the efficiency of the algorithm proposed here we therefore consider the case where the systematic noise sources themselves are contaminated with some degree of Gaussian noise. This leads to an underlying data model defined by equation 3.41.

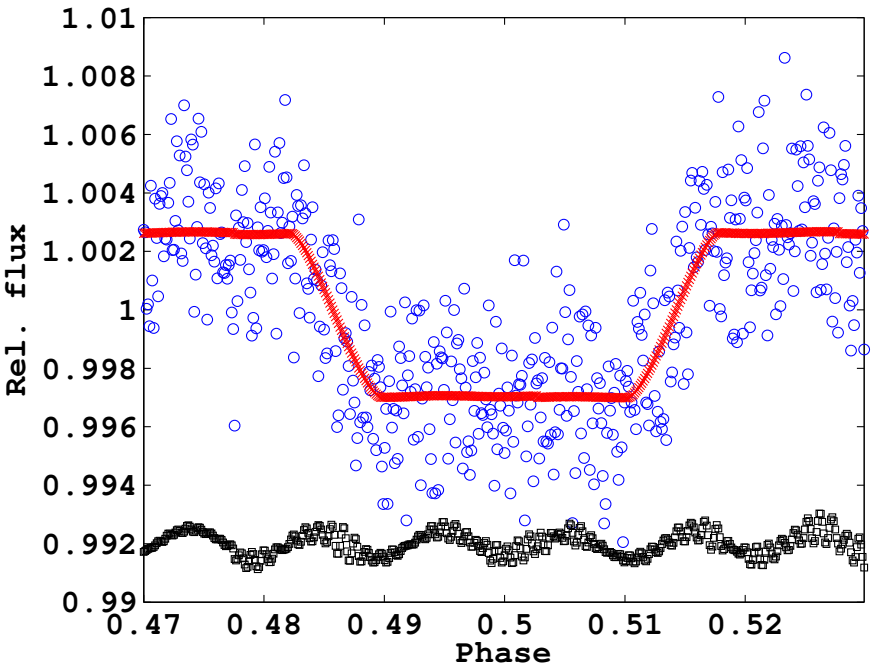


Figure 3.4: Results of the blind-source separation. The blue circles present the the first lightcurve of the raw data $\mathbf{x}$, the red crosses the retrieved signal component, $\mathbf{x_a}$, and the black squares the systematic noise component $\mathbf{x_{sn}}$.

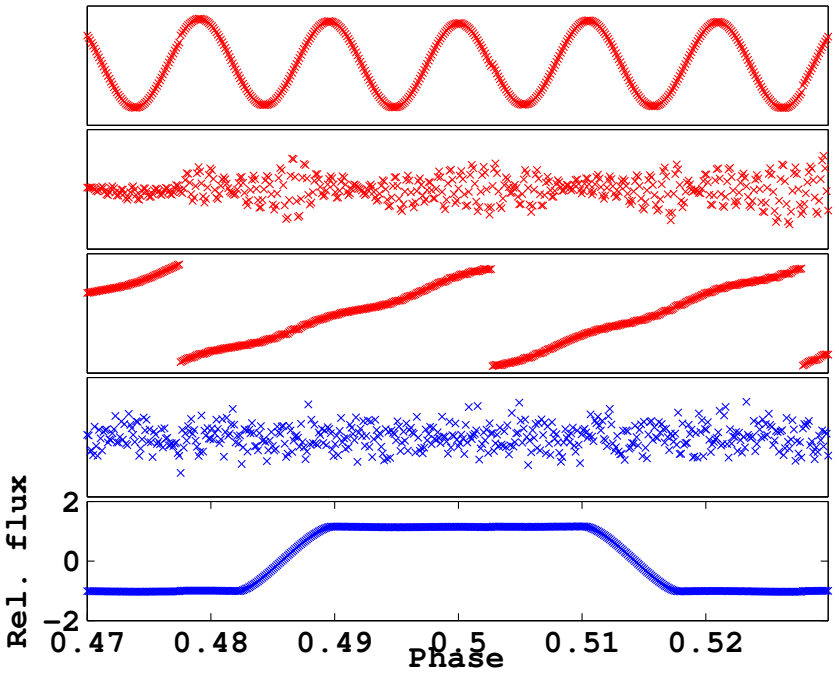


Figure 3.5: Results of the blind-source separation. The top three signals in red were identified by the algorithm to comprise the systematic noise model, $\hat{\mathbf{s}}_{\mathbf{sn}}$. The 4th signal was correctly identified to be Gaussian noise and the bottom to be the lightcurve signal. Note that the blind-source-separation does not preserve signs nor scaling of the estimated signals. The scaling of the ordinate is identical for all subplots.

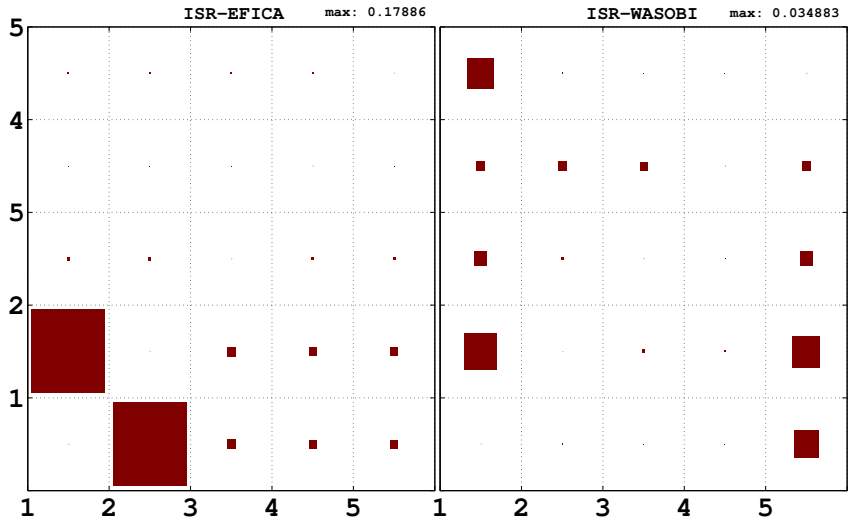


Figure 3.6: Hinton diagram of the EFICA and WASOBI interference-over-signal matrices for Example 1. The polygon areas are normalised to the highest value in the matrix (given in the bottom corners). The smaller the off-diagonal elements of the matrix, the higher the signal separation efficiency of the algorithm. In this case we can see the EFICA algorithm to perform better than the WASOBI one.

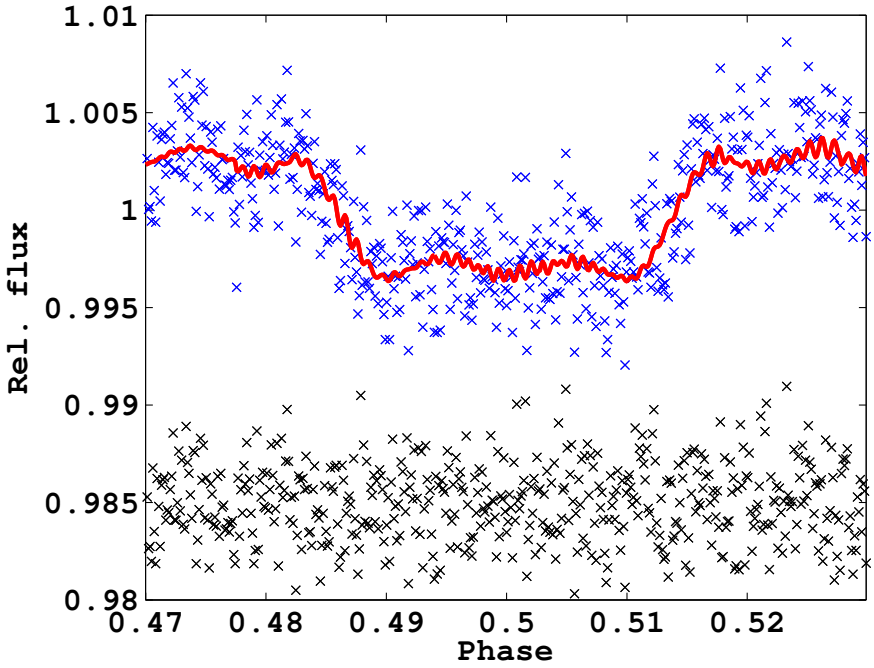


Figure 3.7: showing the raw lightcurve (first row in figure 3.3, blue) normalised to unity, with the model fit (red) overlaid and the fitting residuals plotted underneath (black).

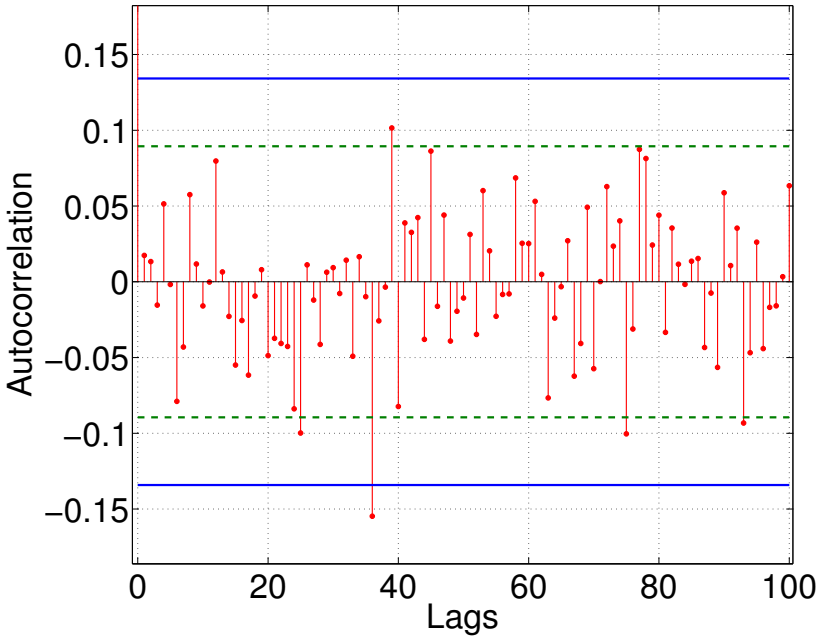


Figure 3.8: showing the auto-correlation function for 250 lags (red). The 3σ confidence limits that the observed residual is normally distributed are shown in blue. All but two lags are within the confidence limits, strongly suggesting that the residual is dominated by white noise and correlations were efficiently removed.

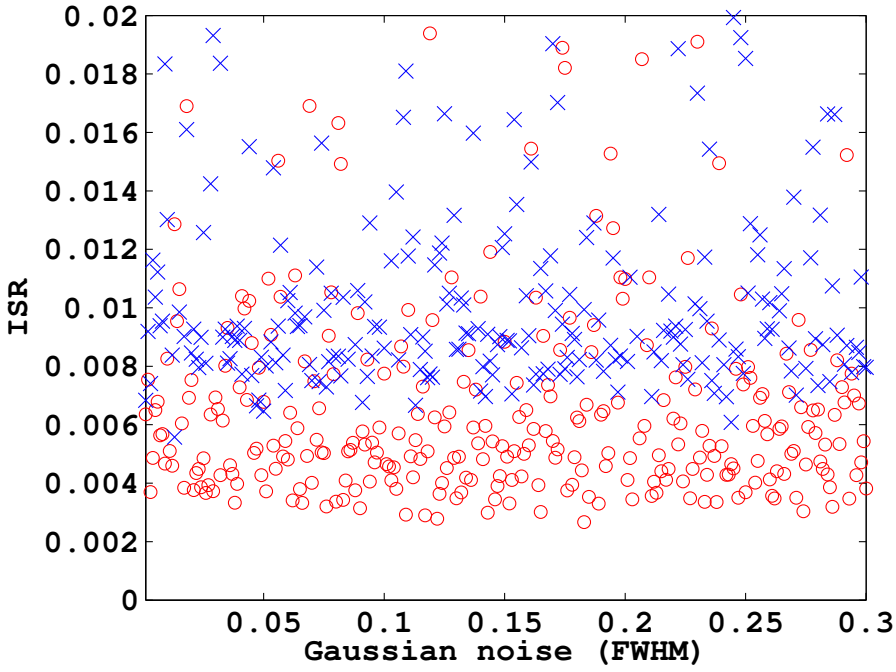


Figure 3.9: showing the mean interference over signal ratios (ISRs) for both the EFICA (red circles) and WASOBI (blue crosses) algorithms for Example 1. In this example, the EFICA algorithm clearly outperforms WASOBI by reaching lower ISR values. Both algorithms are stable even under low signal to noise conditions.

$$x_k = \sum_{l=1}^{N} a_{lk}(s_{lk} + s_{wn,k}) + s_{wn} \tag{3.41}$$

The pre-mix sources were taken from section 3.4.1 and a small Gaussian noise contribution with full-width-FWHM = 0.001 has been added to every channel, in addition to the fully Gaussian channel from the previous example, figure 3.10. The signals were then mixed in the same manner as in section 3.4.1 to produce figure 3.11. Figure 3.12 is a zoomed in version of the first row of figure 3.11 to illustrate the deteriorated data quality of the 'observed' signal in comparison to that in figure 3.7.

Using no pre-filtering

The algorithm was now run on the mixed sources without pre-filtering the data and the outcome reported in figure 3.13. It can clearly be seen that the signal separation was suboptimal this time and neither the systematic noise sources nor the lightcurve signal could fully be recovered. Figures 3.14 & 3.15 indicate that WASOBI has outperformed EFICA in signal separation which is little surprising considering that WASOBI was designed with Gaussian mixtures in mind and EFICA for purely non-Gaussian cases. We did not attempt to construct a noise model or fit the data in the light of the poor separation.

Using kernel regression pre-filtering

We have seen in the previous example that additional Gaussian noise added to all components severely impairs the signal separation process. In this final example we make use of a kernel regression pre-processing step. For a well sampled time series we can optionally use additional filters to decrease the variance due to Gaussian noise, σ_k^{Gauss} where k is the kth observed signal of $\tilde{\mathbf{x}}$, see section 3.3.1. Here we use a non-parametric Gaussian kernel regression with a Nadaraya-Watson estimator (Nadaraya, 1964; Watson, 1964) and use the kernel bandwidth, h, as smoothing factor. The choice of h is important since a too high value will impair the underlying astrophysical signal and we limit the bandwidth, using simulations, to be $h \leq 10^{-4}$ for N $\sim$ 500 points per time series x_k.

Assuming a Gaussian kernel, we can specifically filter Gaussian noise and leave the underlying signals largely unharmed. Beginning from figure 3.11, as in section 3.4.3, we can now separate the sources much more effectively, figure 3.16. The raw data has been fitted as in section 3.4.2, figure 3.19, and the autocorrelation function plotted in figure 3.20. We can see that with the use of the kernel regression preprocessing step the signal can again be removed in its entirety from the observed time series and the autocorrelation function of the residual does not indicate

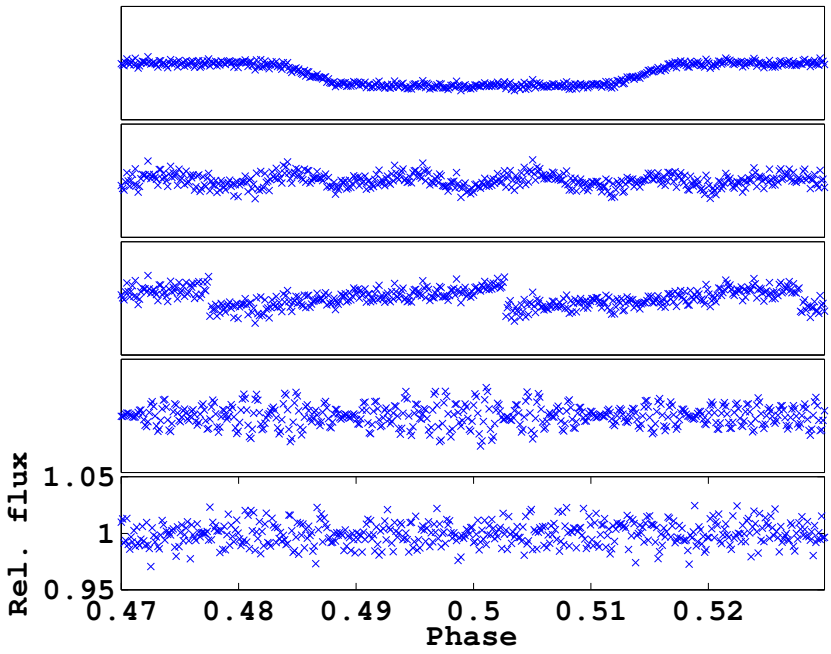


Figure 3.10: shows the same model as in figure 3.2, but with added Gaussian noise (FWHM = 0.001) to each component. The scaling of the ordinate is identical for all subplots.

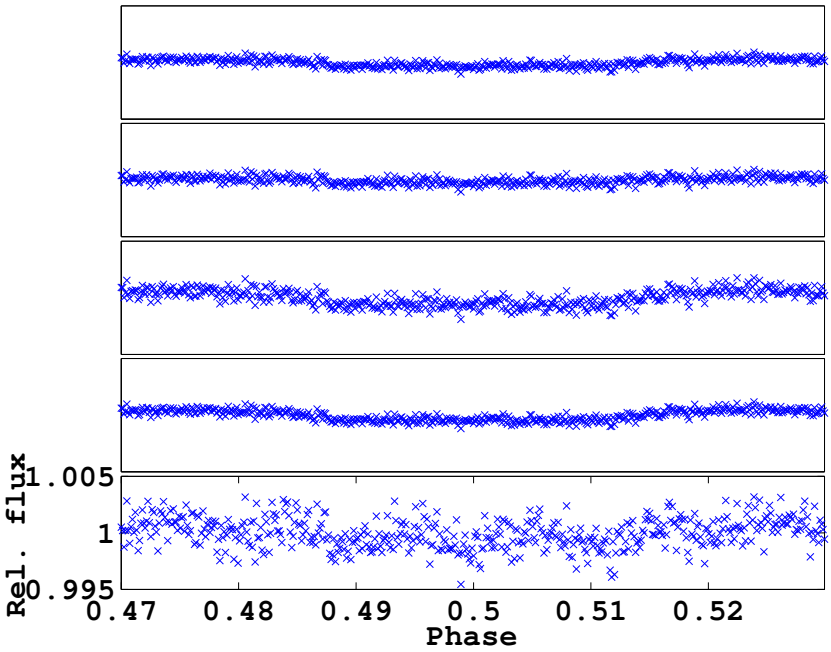


Figure 3.11: As in figure 3.3, the signals, **s**, in figure 3.10 were mixed using a random mixing matrix **A** to obtain the 'observed signals', **x**, shown in this diagram. The algorithm takes the lightcurves in this diagram as starting values. No further input is provided or assumptions on the underlying signals made.

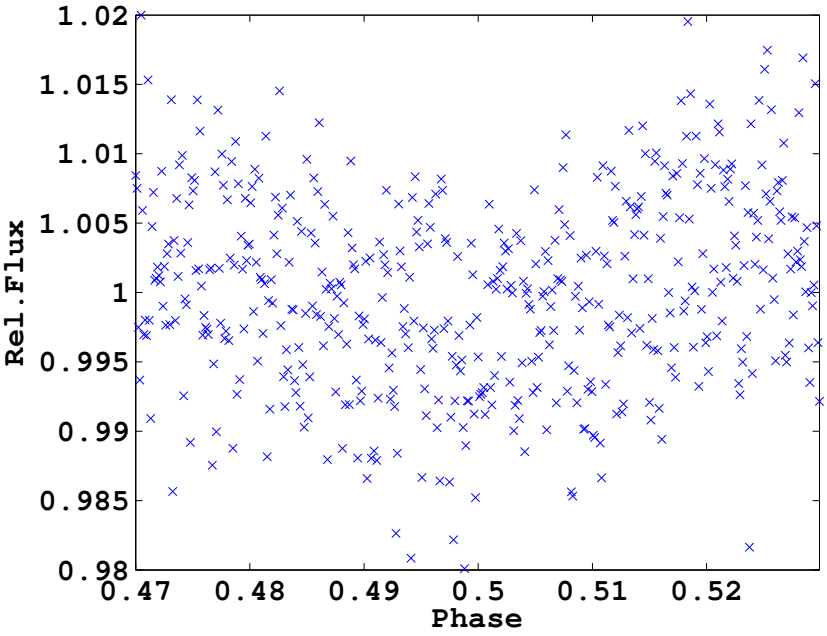


Figure 3.12: showing an enlarged version of the first row of figure 3.11 to illustrate the poor signal to noise conditions induced by the additional Gaussian noise added.

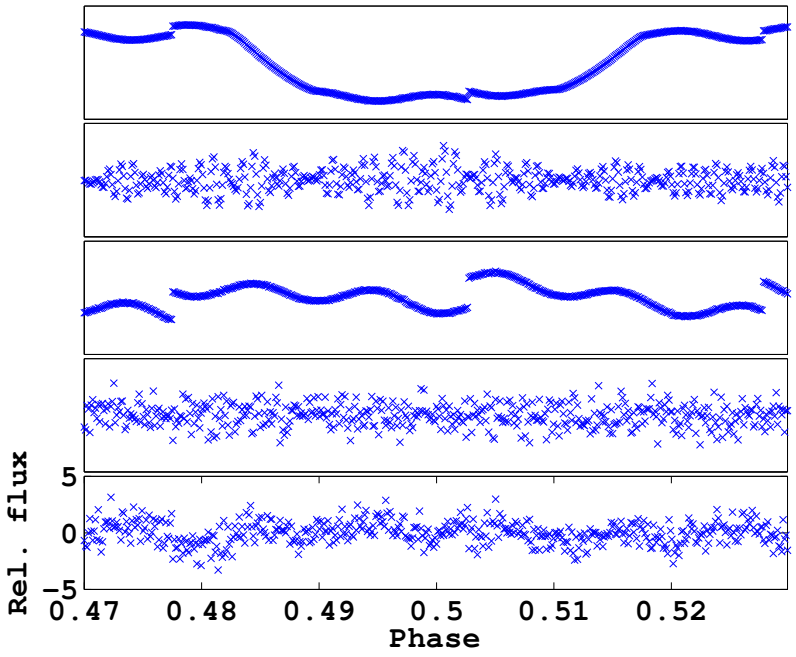


Figure 3.13: Results of the blind-source-separation of the signals in figure 3.11. It is clear that the separation was not optimal and none of the systematic noise components nor the lightcurve signal were fully separated from each other. The scaling of the ordinate is identical for all subplots.

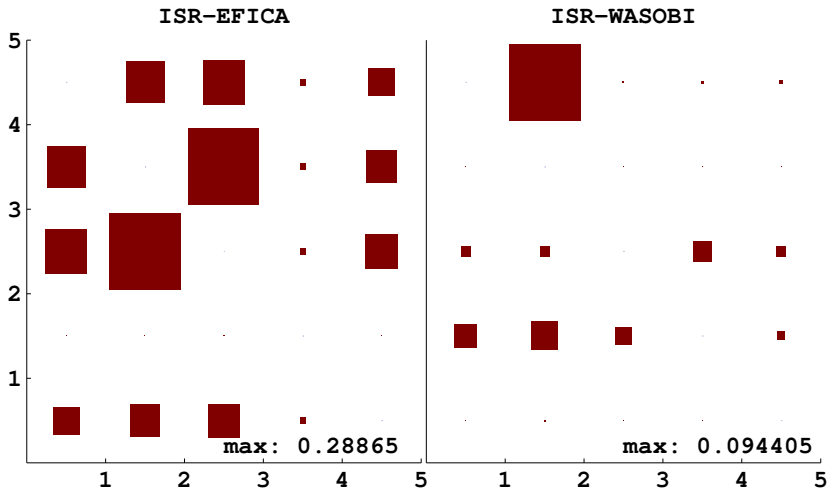


Figure 3.14: Hinton diagram of the EFICA and WASOBI interference-over-signal matrices for signals in figure 3.13. The polygon areas are normalised to the highest value in the matrix (given in the bottom corners). The smaller the off-diagonal elements of the matrix, the higher the signal separation efficiency of the algorithm. In this case the WASOBI algorithm outperform EFICA. Altogether the ISRs are higher here than for examples in sections 3.4.1 & 3.4.2 indicating an overall poorer signal separation .

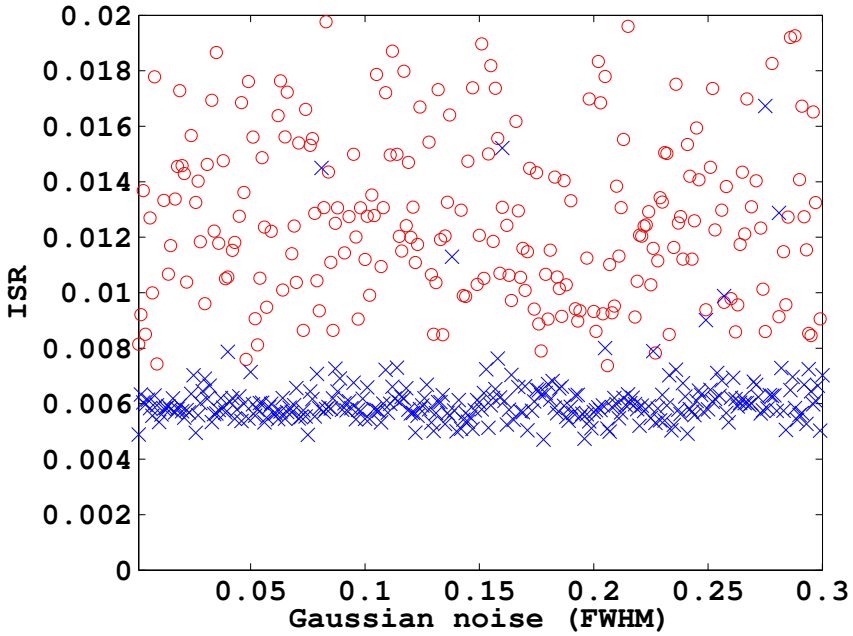


Figure 3.15: The same than in figure 3.9. As opposed to the non-Gaussian examples in sections 3.4.1 & 3.4.2, the WASOBI (blue crosses) algorithm performs better with Gaussian mixtures than the EFICA (red circles) algorithm in all cases.

any remaining non-Normal signal present. Furthermore, by comparing figures 3.15 and 3.18 we can see that the kernel pre-processing step significantly improved the efficiency of the EFICA algorithm, but WASOBI remains dominant.

3.5 Discussion

In the previous sections I have shown that for a set of simultaneously observed time series data (e.g. following an exoplanetary eclipse with a spectrograph) we can describe the data by an instantaneous mixing model (equation 3.3). This allows the separation of non-Gaussian, time and spatially-correlated signals from one another. The degeneracy caused by not being able to retrieve the component's signs or amplitudes can be circumvented in two ways: *Method 1*) The separated signals are used to construct a linear transformation to filter the astrophysical signal from the originally observed data and hence preserve all scaling information; *Method 2*) The separated astrophysical signal is not used directly but instead all systematic noise components are combined to form a 'systematic noise model' which can then be used to correct the original observed data.

I have investigated two possible scenarios: 1) the observations perfectly follow the instantaneous mixing model (equation 3.3) or 2) they do not follow this model and more than one component of the data contains Gaussian noise (equation 3.41, section 3.4.3). In the first case, a very good degree of de-mixing can be obtained without the use of PCA assisted dimensionality reduction or the need to pre-filter the data. In the latter case, the separation is severely limited by the additional Gaussian noise in every component. This scenario is in reality somewhat unlikely but can be solved with the use of pre-filtering techniques. In this case we applied a Gaussian kernel smoothing to our raw data before attempting the blind-source deconvolution. This pre-processing yielded a remarkable improvement in the signal separation. Such a study is important to verify the claims in section 3.3.1 that the instantaneous mixing model is not broken by data pre-processing' steps but the accuracy of the result can be dramatically improved in the case of corrupted data. Being able to combine the blind-source separation techniques presented here with an additional data cleaning method such as kernel regression, dimensionality reduction or Fourier and wavelet filtering allows for the broad applicability in current data-analysis.

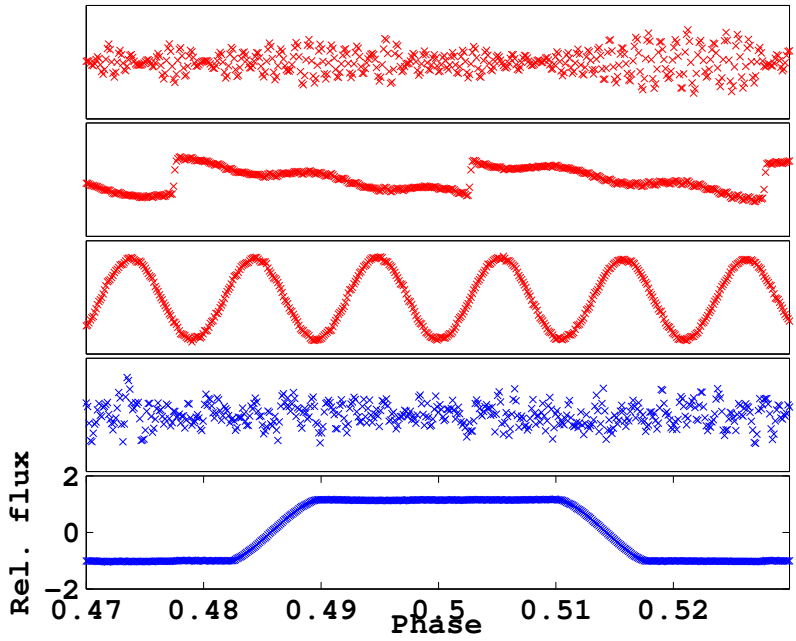


Figure 3.16: Results of the blind-source separation for the kernel-smoothed case. The top three signals in red were identified by the algorithm to comprise the systematic noise model, $\hat{\mathbf{s}}_{\mathbf{sn}}$. The 4th signal was correctly identified to be Gaussian noise and the bottom to be the lightcurve signal. Note that the blind-source-separation does not preserve signs nor scaling of the estimated signals.

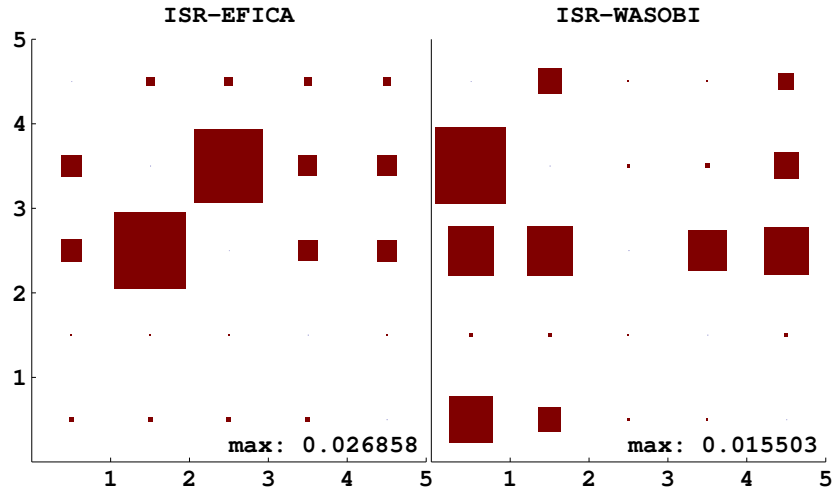


Figure 3.17: Hinton diagram as in figures 3.6 and 3.14 for the kernel-smoothing example.

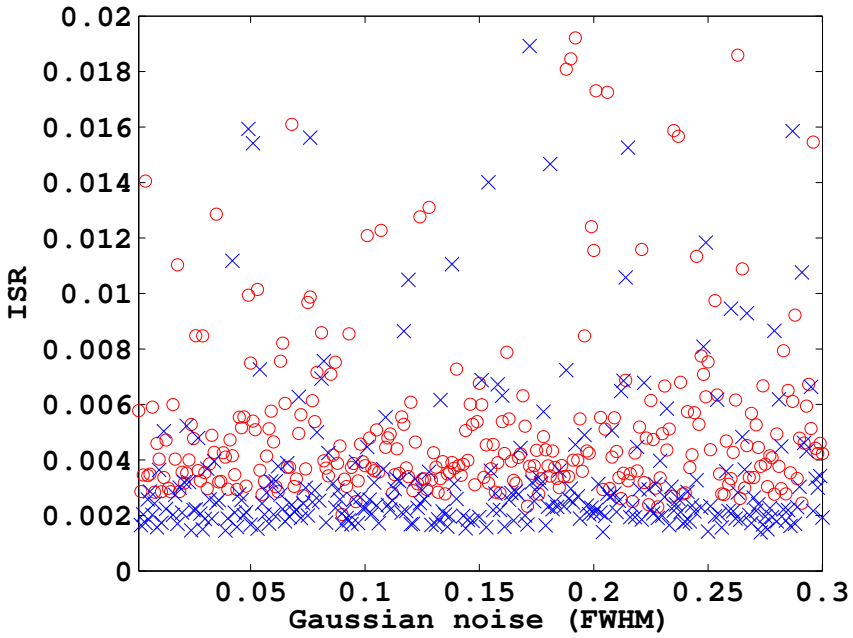


Figure 3.18: mean ISRs of the EFICA (red circles) and WASOBI (blue crosses) algorithms as described in figure 3.9. Compared to figure 3.15, the kernel regression pre-processing step has yielded a significant improvement in efficiency of the EFICA algorithm. Nonetheless, we find the WASOBI algorithm to be dominant.

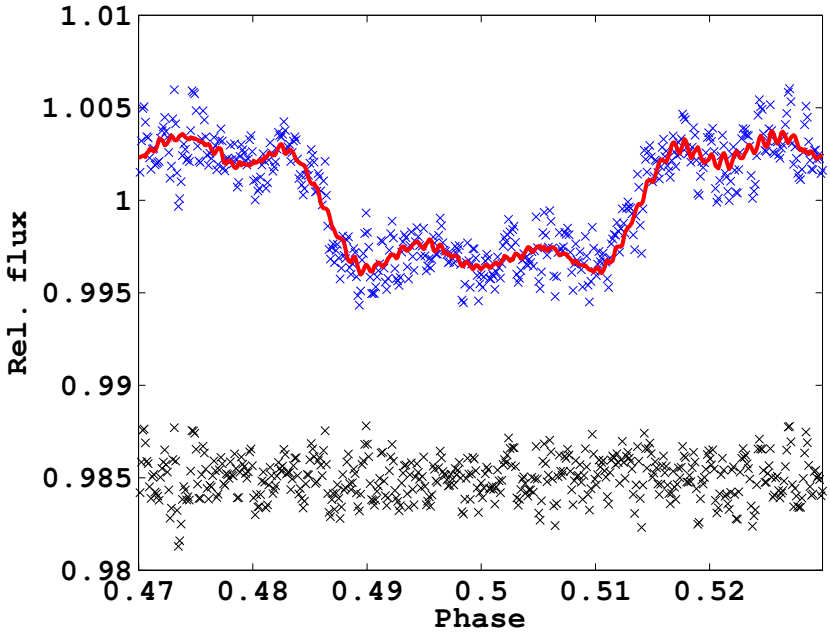


Figure 3.19: showing the normalised raw lightcurve after kernel regression (blue) with the model fit (red) overlaid and the fitting residuals plotted underneath (black). Note the improvement in signal to noise from figure 3.12 due to the kernel smoothing pre-processing.

3.6 Conclusion

In this chapter I have introduced the concepts of non-parametric, blind-source separation of a mixture of non-Gaussian signals. Starting from the concept of making several simultaneous observations of a mixture of voices, known as the 'Cocktail Party Problem', I show that it is possible to de-convolve these individual signals only using the underlying assumptions of the signals being statistically independent and non-Gaussian. These assumptions lead to the concept of independent component analysis which I introduce in section 3.2.2 alongside the definitions of projection pursuit and second-order statistical deconvolution techniques such as WASOBI. After the *instantaneous mixing model* had been defined for spectrophotometric observations of extrasolar planets, I proceed in section 3.3 to describe the proposed blind-source separation algorithm for the analysis of time-resolved exoplanetary spectra. The performance of said algorithm is then tested using simulated data for a variety of scenarios in section 3.4.

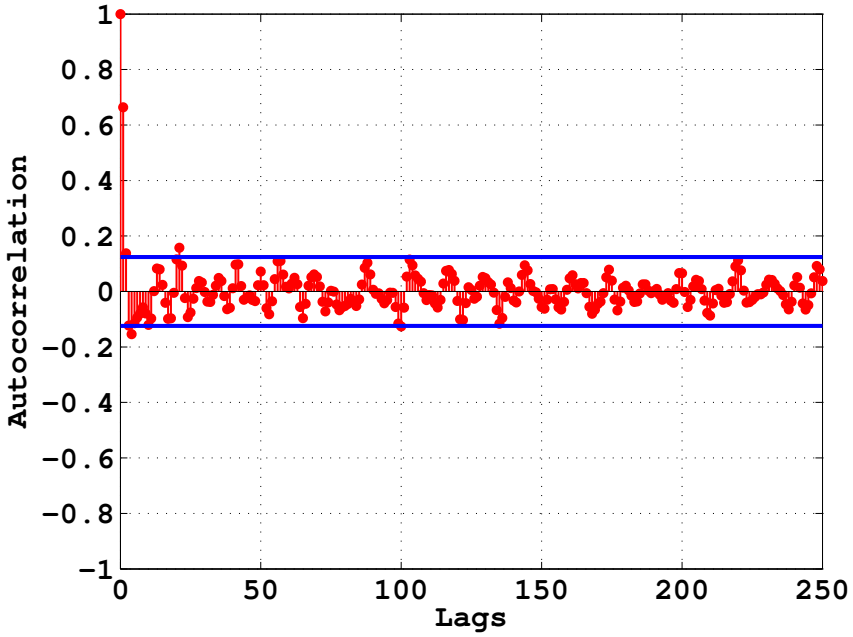


Figure 3.20: showing the auto-correlation function for 250 lags (red). The 3σ confidence limits that the observed residual is normally distributed are shown in blue. All but three lags are within the confidence limits, strongly suggesting that the residual follows a Gaussian distribution.

3.7 Appendix

3.7.1 Maximum entropy distribution

In statistical thermodynamics we know the distributions maximising entropy is such that it the distribution corresponding to the macrostate contains the highest possible number of microstates. In thermodynamics the maximum entropy distribution is given by the Maxwell-Botlzmann distribution.

In information theory, we can derive the function, g, maximising the entry $H(g)$ by requiring g to satisfy the following:

1. $g(x) \geq 0$, equal for all values outside the set $\mathcal{S}$
2. $\int_{\mathcal{S}} g(x)\mathrm{d}x = 1$
3. $\int_{\mathcal{S}} g(x) r_i(x)\mathrm{d}x = \alpha_i \quad \text{for } 1 \leq i \leq m.$

where $g(x)$ is a probability density over the set $\mathcal{S}$ for given moment constraints $\alpha_1, \alpha_2, \ldots, \alpha_m$. One can show that the functional form of $g(x)$ satisfying the above constraints is given by

$$g(x) = e^{\lambda_0 - 1 + \sum_{i=1}^{m} \lambda_i r_i(x)}, \qquad x \in \mathcal{S} \tag{3.42}$$

where λ_i are chosen constants so $g(x)$ satisfies the constraints. The prove for equation 3.42 is given in Cover and Thomas (2006). For a given mean μ and variance σ^2, we can state that the Gaussian distribution, $\mathcal{N}$,

$$\mathcal{N}(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{(x-\mu)^2}{2\sigma^2}} \tag{3.43}$$

is of the same functional form as equation 3.42. We can now prove that $\mathcal{N}(\mu, \sigma^2)$ indeed has a higher entropy than the arbitrary probability density function $f(x)$ by considering the Kullback-Leibler divergence, D_{KL}, between the Gaussian pdf $g(x)$ and $f(x)$. The Kullback-Leibler divergence can be understood in terms of coding length as the additional information needed to fully describe a given probability distribution with the base of another. It can also be understood as the informational 'distance' between two given probability distributions. The informational 'distance' between pdfs p_1 and p_2 is always non-negative, i.e. $D_{KL}(p_1||p_2) \geq 0$ and is only zero if and only if $p_1 = p_2$. Its units are bits for a $\log_2$ base and its equation is

$$D_{KL}(p_1||p_2) = \int p_1(\mathbf{u}) \log \frac{p_1(\mathbf{u})}{p_2(\mathbf{u})} \mathrm{d}\mathbf{u} \tag{3.44}$$

from the above equation we can see that the Shannon entropy, H (equation 3.13), is the distance of a given pdf to a uniform distribution $D_{KL}(U||p)$. Let us now consider the Gaussian distribution g and the arbitrary distribution f:

$$\begin{aligned} D_{KL}(f||g) &= \int f(x)\log\left(\frac{f(x)}{g(x)}\right)\mathrm{d}x \\ &= \int f(x)\log f(x)\mathrm{d}x - \int f(x)\log(g(x))\mathrm{d}x \\ &= -H(f) - \int f(x)\log(g(x))\mathrm{d}x \end{aligned} \tag{3.45}$$

substituting equation 3.43 for $g(x)$

$$\begin{aligned} 0 \leq D_{KL}(f||g) &= -H(f) - \int f(x)\log\left(\frac{1}{\sqrt{2\pi\sigma^2}}e^{\frac{(x-\mu)^2}{2\sigma^2}}\right)\mathrm{d}x \qquad (3.46) \\ &= -H(f) - \int f(x)\log\frac{1}{\sqrt{2\pi\sigma^2}}\mathrm{d}x + \log(e)\int f(x)\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)\mathrm{d}x \\ &= -H(f) - \frac{1}{2}\log(2\pi\sigma^2) - \frac{\sigma^2\log(e)}{2\sigma^2} \\ &= -H(f) - \frac{1}{2}(\log(2\pi\sigma^2) + \log(e)) \\ &= -H(f) - \frac{1}{2}\log(2\pi e\sigma^2) = -H(f) + H(g) \qquad (3.47) \\ 0 &\leq H(g) - H(f) \qquad (3.48) \end{aligned}$$

we can see that the entropy difference between a Gaussian distribution and an arbitrary pdf is always non-zero and only tends to zero if and only if both pdfs are Gaussian with the same variance. We can also see from equation 3.47 that the entropy of a Gaussian pdf is given by $H(g) = \frac{1}{2}\log(2\pi e\sigma^2)$ or for a multivariate Gaussian distribution we have

$$\begin{aligned} H(g) &= \frac{1}{2}(n + n\log(2\pi) + \log|\mathbf{\Sigma}|) \qquad (3.49) \\ &= \frac{1}{2}\log|\mathbf{\Sigma}| + \frac{n}{2}[1 + \log 2\pi] \qquad (3.50) \\ &= \frac{1}{2}\log(2\pi e)^n|\mathbf{\Sigma}| \qquad (3.51) \end{aligned}$$

where $|\mathbf{\Sigma}|$ is the determinant of the covariance matrix $\mathbf{\Sigma}$ and n the number of dimensions.

3.7.2 Differential Entropy Transformations

Here we demonstrate two properties of the differential entropy (also known as Shannon entropy), given in equation 3.13. Proofs can be found in the standard literature (e.g. Hyvärinen et al., 2001; Comon and Jutten, 2010; Cover and Thomas, 2006).

From the definition of differential entropy, we can see that it is translationally insensitive

$$H(X + c) = H(X) \tag{3.52}$$

where c is a constant. However, as mentioned in section 3.2.2, we find Shannon entropy to be sensitive to scaling by a factor of a. This introduces a scale and sign ambiguity as the data needs to be normalised. The proof of this property is as follows:

Let $\mathbf{y}$ be a scaled column vector of $\mathbf{x}$ such as $\mathbf{y} = \mathbf{A}\mathbf{x}$, this gives us the pdf

$$f_{\mathbf{y}}(y) = \frac{1}{|\mathbf{A}|} f_{\mathbf{x}}\left(\frac{\mathbf{y}}{\mathbf{A}}\right) \tag{3.53}$$

and the new entropy $H(\mathbf{A}\mathbf{x})$

$$\begin{aligned} H(\mathbf{A}\mathbf{x}) &= -\int f_{\mathbf{y}}(y)\log f_{\mathbf{y}}(y)\mathrm{d}y \qquad (3.54)\\ &= -\int \frac{1}{|\mathbf{A}|} f_{\mathbf{x}}\left(\frac{\mathbf{y}}{\mathbf{A}}\right) \log\left(\frac{1}{|\mathbf{A}|} f_{\mathbf{x}}\left(\frac{\mathbf{y}}{\mathbf{A}}\right)\right) \mathrm{d}y \\ &= -\int f_{\mathbf{x}}(x)\log f_{\mathbf{x}}(x)\mathrm{d}x + \log|\mathbf{A}| \\ &= H(\mathbf{x}) + log|\mathbf{A}| \qquad (3.55) \end{aligned}$$

from equation 3.55 we can appreciate that a scaling of $\mathbf{x}$ by $\mathbf{A}$ increases the entropy by a factor of $\log|A|$.

However it is now easy to prove that *negentropy*, $J(\mathbf{x})$, is scale invariant, unlike differential entropy. For the same transformation as above, $\mathbf{y} = \mathbf{A}\mathbf{x}$, we have the covariance $E[\mathbf{y}\mathbf{y}^{\mathbf{T}}] = \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^{\mathbf{T}}$. Following from equations 3.50 & 3.55, we can express the transformed *negentropy* as

$$J(\mathbf{Ax}) = \frac{1}{2}\log|\mathbf{A\Sigma A^T}| + \frac{n}{2}[1+\log(2\pi)] - (H(\mathbf{x}+\log|\mathbf{A}|)) \tag{3.56}$$

$$= \frac{1}{2}\log|\mathbf{\Sigma}| + 2\frac{1}{2}\log|\mathbf{A}| + \frac{n}{2}[1+\log(2\pi)] - H(\mathbf{x}) - \log|\mathbf{A}| \tag{3.57}$$

$$= \frac{1}{2}\log|\mathbf{\Sigma}| + \frac{n}{2}[1+\log(2\pi)] - H(\mathbf{x}) \tag{3.58}$$

$$= H(g) - H(\mathbf{x}) = J(\mathbf{x}) \tag{3.59}$$

3.7.3 Signal separation

In order to identify the non-white (i.e. systematic) signals in our estimated signal matrix $\hat{\mathbf{S}}$, we use the Ljung-Box portmanteau test (Brockwell and Davis, 2006). The test statistic, usually denoted by Q, is defined by summing the normalised autocorrelations of the individual time series, $\hat{\mathbf{s}}_\mathbf{l}$ over a range of lags:

$$Q = n(n+2)\sum_{\tau=1}^{m}\frac{\hat{\rho}_\tau^2}{m-\tau} \tag{3.60}$$

where $\hat{\rho}_\tau^2$ is the autocorrelation at lag τ and m is the number of observations in the time series. The hypothesis of the time series being solely white noise is rejected if Q is bigger than a pre-specified fraction of the chi-squared distribution

$$Q > \chi^2_{1-\alpha,h} \tag{3.61}$$

where $\chi^2_{1-\alpha,h}$ is the α-quantile of the chi-squared distribution with h degrees of freedom (Brockwell and Davis, 2006). Here we take $\alpha = 0.05$.

3.7.4 Code design

The programme developed for this chapter and chapter 4 is written in the commercially available *MATLAB*[1] programming language and packages. The use of high-level commercially maintained software such as *MATLAB* or *IDL* has the advantage of featuring extensive statistical libraries, optimised for programatic efficiency. In particular, the *Statistical Toolbox*[2] is heavily made use of in this programme and throughout this thesis.

Code philosophy and design

The de-trending algorithm described here is designed to require as little user input as possible to ensure an unbiased analysis of the data and easy usage. I therefore designed the algorithm

[1] MATLAB R2012a, The MathWorks Inc., Natick, MA, 2012, http://www.mathworks.co.uk/products/matlab/
[2] http://www.mathworks.co.uk/products/statistics/

to be fully automated with the user only providing the input data (contained in either an external file or a *MATLAB* array) and a minimal number of parameters contained in an external parameter file. The code is fully dynamic and sub-routines decide on a case-by-case basis on the appropriate methods to be used and pass messages to inform subsequent routines of update parameters without the need of manual user-input.

The outline of the de-trending algorithm is shown in figure 3.1 and explained in section 3.3. Given the high number of individual processing steps, sub-routines and parameters, a clean architectural design is required. Here I employ an object-orientated programming structure, which passes two main objects through the required set of subroutines. I elaborate on these below.

The `params` object:

At the start of the code, a parameter sub-routine is run which reads in pre-specified parameters from an external parameter file and initialises the parameter object, `params`, containing all parameters listed in the external file as well as any interactive user defined parameters. This object is passed to all functions and sub-routines but is considered a static object after initialisation. This object constitutes the main control 'unit' of the code. A dynamic copy of the parameters is also carried in the `data` object.

The `data` object:

After the `params` object has been initialised at the beginning of the code, a sub-routine initialises the `data` object. The `data` object is organised in 1^{st} and 2^{nd} order sub-structure levels. A structure is a *MATLAB* specific data type that allows one to group related information in individual data containers. This is programmatically similar to dictionary type approaches in *C#* or *Python*. However, unlike dictionaries, structures within structures are allowed, here I refer to these as sub-structures.

The fundamental 1^{st} order level of the `data` object includes two sub-structures: 1) DATA and 2) DPAR, where DATA contains the raw and de-trended data as well as all intermediate results in 2^{nd} order structures. The DPAR structure on the other hand contains a dynamic copy of the `params` object and capabilities of message passing between sub-routines. Given the fully automated nature of the de-trending algorithm, sub-routines dynamically decide on parameters changes found in the `params` object and record the updated parameters in the DPAR structure. The comparisons between the static `params` object and the dynamically changing DPAR structure allows for transparency and easy de-bugging capabilities of the code. Furthermore, the DPAR structure collects all warning and de-bugging messages generated by individual sub-routines.

Given input parameters derived from the `params` object, the data object loads the (to be de-trended) data into the DATA sub-structure of the `data` object. The DATA sub-structure contains six 2^{nd} level structures corresponding to the main blocks in figure 3.1. In these sub-structures, every intermediate data product is saved and collectively passed to the individual sub-routines and functions.

The sub-routines:

All routines described in section 3.3 have been coded from scratch and comprise ~ 30 individual *MATLAB* programmes that are called by the main program. Besides the functionality described in the main text, these subroutines include post-analysis scripts that take the final `data` object as input and issue a summary report on convergence properties and plot relevant result and intermediate data products. A rudimentary self-written implementation of the FastICA algorithm is included but due to a superior optimisation of the code, the *MATLAB* implementation of *MULTICOMBI*[3] by Tichavsky et al. (2006) was adopted and incorporated into this de-trending algorithm.

Benchmarking:

It is difficult to benchmark the above described algorithm as the processing time strongly depends on the quality of the data-set. For the cases presented in the following chapter, *Hubble*/NICMOS observations of HD189733b and XO1b, convergence is usually achieved on time-scales of 10 to 30 seconds, with the number of individual iterations not exceeding 500 for all cases. The efficiency of the code is clearly dependent on the size of the data set to be analysed. However, large spectroscopic data-sets can usually be said to be over-complete and the principal component analysis part of the algorithm will then try to reduce the data size of the input data by disregarding principal components with negligibly small eigenvalues.

[3] http://itakura.kes.tul.cz/zbynek/multicombi.htm

Chapter 4

Analysing Space-Based Data

4.1 Introduction

Following from the previous chapter, I will here apply the blind-source separation techniques to spectroscopic data obtained by the *Hubble* Space Telescope and individual time series obtained by the Kepler Space Mission. Exoplanetary spectroscopy had started with *Hubble* measurements of NaI lines in the visible part of the spectrum (Charbonneau et al., 2002). These early measurements detected individual, strong electronic transition lines. It was hence easy to take an out-of-transit (OOT) and an in-transit (INT) measurement of the extrasolar planet using UV and visible spectrographs such as *Hubble*/STIS and difference these spectra to obtain the exoplanetary absorption feature. This approach works well for individual electronic transitions in the UV and visible but is less effective in the infra-red where the spectrum is dominated by roto-vibrational transitions which feature less discrete and much broader absorption and/or emission features. As the search for molecular features (e.g. H_2O, CH_4, CO) commenced in 2007 (Tinetti et al., 2007), we needed to adopt new methods. The most common is that of spectrophotometry. Here, we follow the eclipse event with a sequence of short exposures using a spectrograph and hence obtain a lightcurve per resolution element of the spectrograph (see section 2.1). Time resolved measurements have the advantage of allowing a more precise measurement of the in-transit eclipse depth as it allows us to average over short period fluctuations of the instrument and host star and more accurately determine outliers and the systematic noise of the instrument.

Swain et al. (2008c) was amongst the first ones to use this spectrophotometric technique on the *Hubble*/NICMOS instrument to obtain a spectrum of the hot-Jupiter HD189733b. The now decommissioned *Hubble*/NICMOS was a multi-purpose imager in the blue/visible and near

infra-red (NIR) wavelengths. It also had the option of obtaining low-resolution spectra using various grisms attached to its filter wheel, most notable the G206 and G141 grisms covering the H-K band regions. The time-resolved spectroscopy approach quickly caught on with most observations being conducted in that way (e.g. Swain et al., 2009a,b; Tinetti et al., 2010; Burke et al., 2010; Sing et al., 2011; Pont et al., 2008).

One of the central challenges with *Hubble* as well as for most multi-purpose instruments, is the correction of their instrument systematics. Leading on from previous discussions in Chapter 2, we will briefly summarise the parametric approach to detrending of spectroscopic data. We will then proceed to showcase the non-parametric approach as postulated in Chapter 3 on two *Hubble*/NICMOS examples, HD189733b and XO1b. Finally, we extend the non-parametric approach to the deconvolution of stellar systematics from a single, consecutive time series obtained by Kepler.

4.2 Parametric Decorrelation

The parametric correction model (Swain et al., 2008c, 2009b,a; Tinetti et al., 2010; Gibson et al., 2011), is based on a linear regression of auxiliary information vectors, also known as 'optical state vectors' (OSV) to the observed data. Let us consider a single observation spanning an eclipse event using a spectrograph. Each spectral channel constitutes a time series, y_i, with N number of spectral resolution elements, i.e. time series. We collect these observations in the column vector $\mathbf{y} = (y_1, y_2, \ldots, y_N)$, where the time dependence t has been dropped for clarity. Additional to the observations contained in $\mathbf{y}$, we retrieve optical state vectors of the spectrum, namely 1) the X-axis position of the spectrum on the detector, 2) the Y-axis position, 3) the full-width-half-maximum (FWHM) width of the spectrum along the dispersion axis' normal, 4) the inclination of the spectrum on the detector. For a discussion on the retrieval of these vectors, please refer to section 2.3. We store these OSV in the $N \times K$ matrix $\mathbf{X}$, where K is the number of OSV. We can now use the Gauss-Markov theorem to describe our observations $\mathbf{y}$ in terms of the OSV and a residual noise vector, ϵ, which is assumed to be Gaussian:

$$\mathbf{y} = \mathbf{X} \cdot \beta + \epsilon \tag{4.1}$$

where β is a column vector of weighting coefficients for each OSV in $\mathbf{X}$ per time series. We now need to fit for β giving the ordinary least square estimator $\hat{\beta}$,

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \tag{4.2}$$

with the residual being given by:

$$\epsilon = \mathbf{y} - \mathbf{X} \cdot \hat{\beta} \tag{4.3}$$

this leads to the iteration scheme:

$$\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{N}(y_i - \sum_{j=1}^{K}\hat{\beta}_j X_{ij})^2 \tag{4.4}$$

which is usually minimised over each time series individually but can also be minimised simultaneously. Figure 4.1 shows an example of the the raw data, the noise model and the individual OSVs obtained for the primary transit of HD189733b.

The regression of equation 4.2, as used by Swain et al. (2008c) and subsequent publications of this group, is clearly linear in nature. Gibson et al. (2011) published a criticism of this approach stating that it is as reasonable to assume a quadratic model or any combination of the two to describe the noise model, leading to stark differences in results from which Gibson et al. (2011) conclude the science result to be inherently degenerate with the correction model used.

4.3 Non-Parametric Decorrelation

The parametric correction has two discernible disadvantages: 1) it assumes an underlying model, 2) its parameters are derived. The assumption of an underlying linear model is the simplest assumption of an instrument when no specific instrument response function is known and hence satisfies Occam's razor. The issue of derived parameters to model the noise is a far more serious one. Whilst it can be shown for instruments such as *Hubble*/NICMOS that the OSVs discussed above are correlated to the systematics found in the data, this cannot be said in general. Often it is not obvious which parameters cause a certain systematic noise morphology or a correlation exists but is weak. Another issue is the quality of the derived optical state vectors. When the variances of the individual OSVs is high, convergence of the regression model is seriously impaired and the variance from the individual OSVs is propagated into the science result. Swain et al. (2011) attributed this problem to be the main cause of discrepancy between Swain et al. (2008c) and Gibson et al. (2011). Gibson et al. (2012) proposed a semi-parametric algorithm that replaces equations 4.1 & 4.2 with a regression model based on Gaussian Processes (GP). This approach allows the user to minimise over the model and to optimise the model choice. The major criticism are twofold, without any additional knowledge of the instrument, a linear model is the simplest assumption and already satisfies Occam's razor. The GP approach continues to

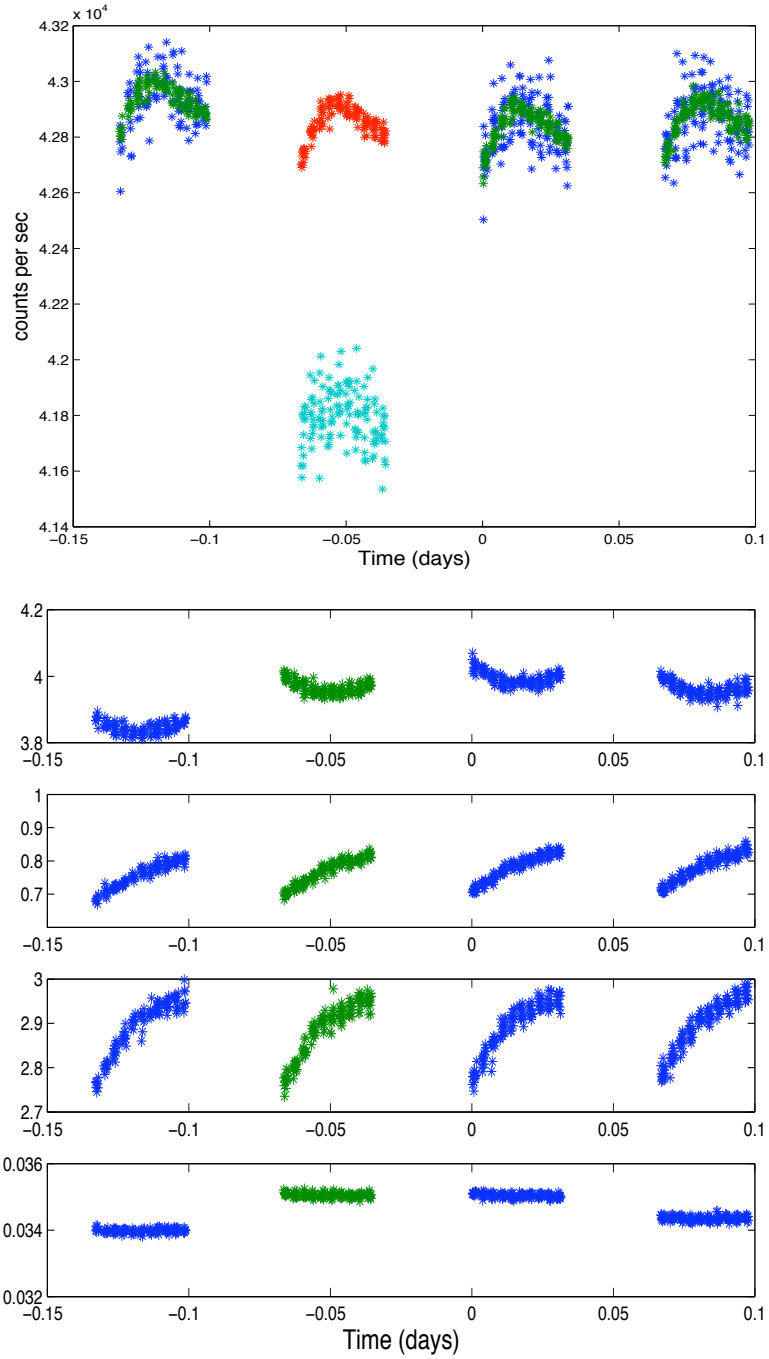


Figure 4.1: Top: example raw lightcurve, out-of-transit orbits in dark blue, in-transit orbit in turquoise. Superimposed is the parametric correction model in green and red. Bottom four panels show derived optical state vectors 1) the X-coordinate of the spectrum, 2) the Y-coordinate, 3) the width of the spectrum dispersion along the slit axis (y), 4) the angle of the spectrum (Swain et al., 2008c).

heavily rely on OSVs and all problems with regards to their quality remain.

This chapter follows on from chapter 3 which introduced 'blind-source deconvolution' based on the principle of independent component analysis. Chapter 3 introduced the main statistical concepts and applied the proposed algorithm to simulated data. In the following section I will apply the algorithm to two *Hubble*/NICMOS observations, discuss the differences in de-trending efficiency and proceed to compute the full *Hubble*/NICMOS spectrum of HD189733b.

4.3.1 *Hubble*/NICMOS: HD189733b

First presented by Swain et al. (2008c), this data set of the primary eclipse of HD189733b was recorded using *Hubble*/NICMOS in the G206 grism setting spanning five consecutive orbits. The *Hubble* pipeline calibrated data were downloaded from the MAST[1] archive and the spectrum was extracted using both standard IRAF[2] routines as well as a custom built routine for optimal spectral extraction (see section 2.3 for more details). Both extractions are in good accord with each other but the custom built routine was found to yield a better signal to noise and was subsequently used for all further analysis. A binning of 10 spectral channels ($\sim 0.08\mu$m) was used resulting in 10 light curves across the G206 grism band. Figure 4.2 shows the obtained time series which serve as input to the algorithm. It can be seen that each time series is strongly affected by instrument systematics propagating from the blue side of the spectrum (bottom light curve) to the red with varying intensity and even sign. Swain et al. (2008c) showed that these systematics are correlated to instrument state vectors such as orbital phase, relative positions and angles of the spectrum on the detector, instrument temperature, etc. We can hence expect that these systematics are statistically independent from the recorded astrophysical signal (the light curve) and it should therefore be in principle possible to de-correlate the signal.

We here demonstrate the de-trending on an individual light curve at $\sim$1.694μm (8^{th} one down in figure 4.2). All time series in figure 4.2 were taken as input to the algorithm described above to estimate the de-mixing matrix $\mathbf{\tilde{W}}$, the astrophysical signal vectors, $\mathbf{\hat{s}_a}$ and the systematic noise vectors, $\mathbf{\hat{s}_{sn}}$. The interference over signal (ISR) matrix indicated the good separation of four main components figure 4.3 with the rest of the components being classified as predominantly Gaussian or weakly systematic. The existence of more than one Gaussian component ($l_{wn} > 1$) indicates that the set is overcomplete. However since the data-set is small enough, no PCA dimensionality reduction was performed. After the algorithm has identified the correct astrophysical signal, it proceeded to reconstruct the light curve using both methods described in section 3.3.3.

[1] http://archive.stsci.edu/
[2] http://iraf.noao.edu/

Method 1: The astrophysical signal was filtered using equations 3.36 & 3.37. Figure 4.4 shows the raw light curve (blue circles) with the de-trended time series, $\mathbf{x_a}$ underneath (green squares). Superimposed light curves were computed using Mandel and Agol (2002) with orbital parameters were taken from Winn et al. (2007) and limb-darkening parameters from Claret (2000). It is clear that the de-trended light curve is an improvement to the raw time series but that systematics still remain in the data. This is further illustrated by plotting the autocorrelation function of the model-fit residual in figure 4.8 (red squares). Here, residual correlation can be observed in particular at low lags. This is a consequence of the astrophysical signal, $\hat{\mathbf{s}}_\mathbf{a}$, being well separated but as shown in figure 4.3 (component 1), there remains some weak interference between the $\hat{\mathbf{s}}_\mathbf{a}$ and other vectors, which is a consequence of equation 3.20 and to be expected for real data-sets.

Method 2: The second method is a less direct approach. Instead of filtering for the astrophysical signal directly, we try to construct a 'systematic noise model' that is then subtracted off the raw data. Using equation 3.39 and a simplex downhill algorithm (Nelder and Mead, 1965) we estimated the scaling matrix, $\mathbf{O}$, by fitting the the systematic noise vectors, $\hat{\mathbf{s}}_\mathbf{sn}$ to the four out of transit orbits. The scaled systematic noise vectors are shown in figure 4.6 which combine to form the systematic noise model, m_k, in figure 4.5. It should be noted that $\mathbf{O}$ is only a scaling matrix of the individual vectors as the scaling information is not preserved by the independent component analysis. Hence, relative intra and inter-orbit variations are preserved. Figure 4.7 shows the corrected data by subtracting the systematic noise model off the raw data. Inspecting the fitting residual's autocorrelation function in figure 4.8 (black circles) indicates the residual to be statistically white and a maximal de-correlation of the data has been achieved.

4.3.2 *Hubble*/NICMOS : XO1-b

Originally presented by Tinetti et al. (2010), the primary eclipse of XO1b was observed using the *Hubble*/NICMOS instrument in the G141 grism setting. The *Hubble* pipeline calibrated data was downloaded and the spectra extracted using the same settings as for section 4.3.1. This yielded 10 light curves and which serve as input to the algorithm, see figure 4.9. Similar to HD189733b the algorithm retrieved four main components, the light curve signal and three main systematic noise components. The ISR matrix is shown in figure 4.10. We now proceeded to de-trending the light curve at the very red end of the spectrum (first from top in figure 4.9) as it, after visual inspection, exhibits the most prominent systematics of the 10 time series. Light curve fits assumed limb-darkening and orbital parameters by Burke et al. (2010).

Method 1: Figure 4.11 shows the raw time series and the de-trended light curve using equation 3.37. The light curve is significantly de-trended but systematics remain in the data as also shown

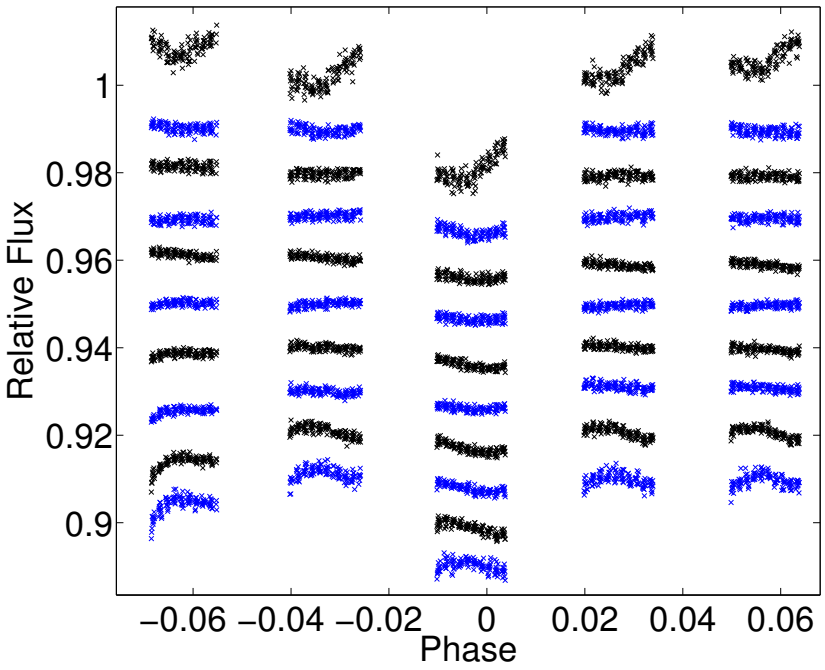


Figure 4.2: showing 'raw', extracted *Hubble*/NICMOS light-curves of HD189733b primary eclipse. Light curves are offset for clarity.

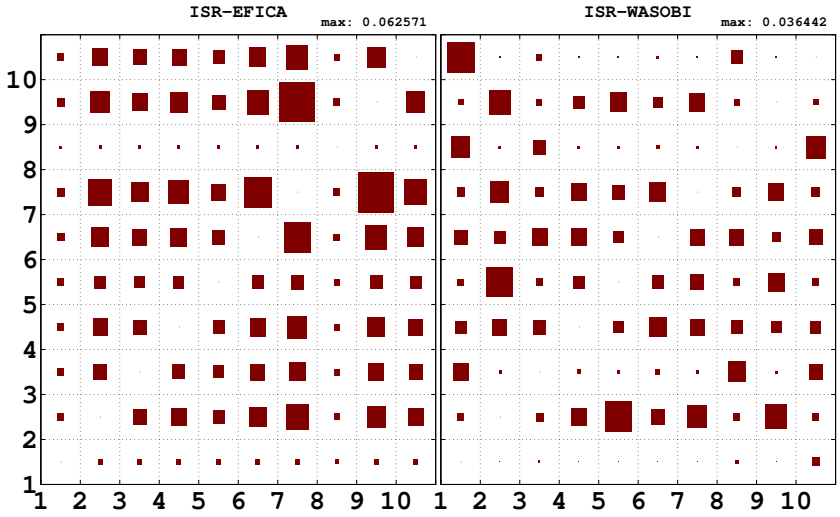


Figure 4.3: the Interference over Signal (ISR) matrix of the component separation for both the EFICA and the WASOBI algorithms. All values were normalised with the maximum ISR = 0.0626. Components 1, 3, 5 & 8 yielding the lowest ISR values and correspond to the astrophysical light curve signal (comp. 1) and the three most prominent systematic noise vectors in figure 4.6. Other components were identified as predominantly Gaussian or weakly systematic by the pipeline.

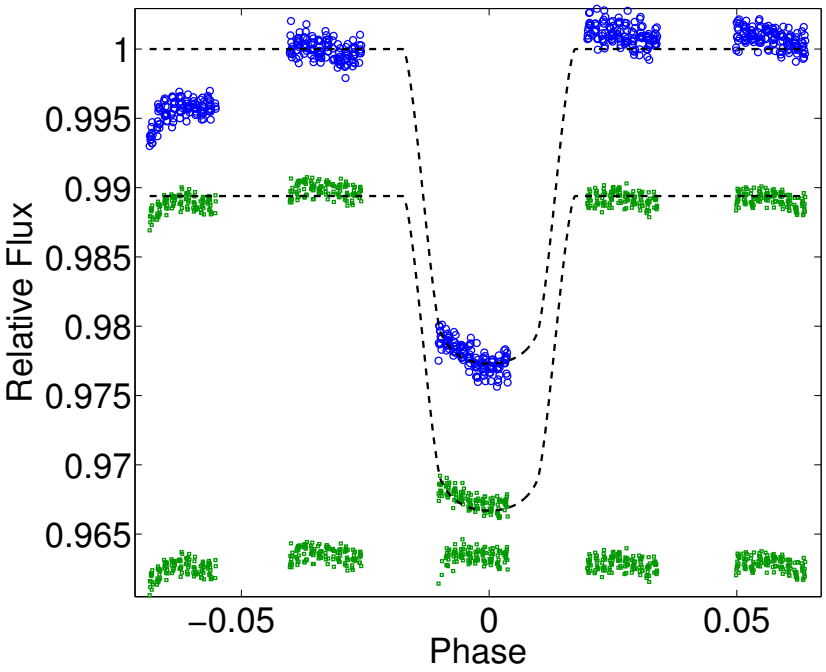


Figure 4.4: showing the raw-data light curve (blue crosses) and the corrected light curve (green squares) offset below. In this example, we used equations 3.36 & 3.37 as light curve filter. The systematic noise components were reduced but residual systematics remain in the final light curve.

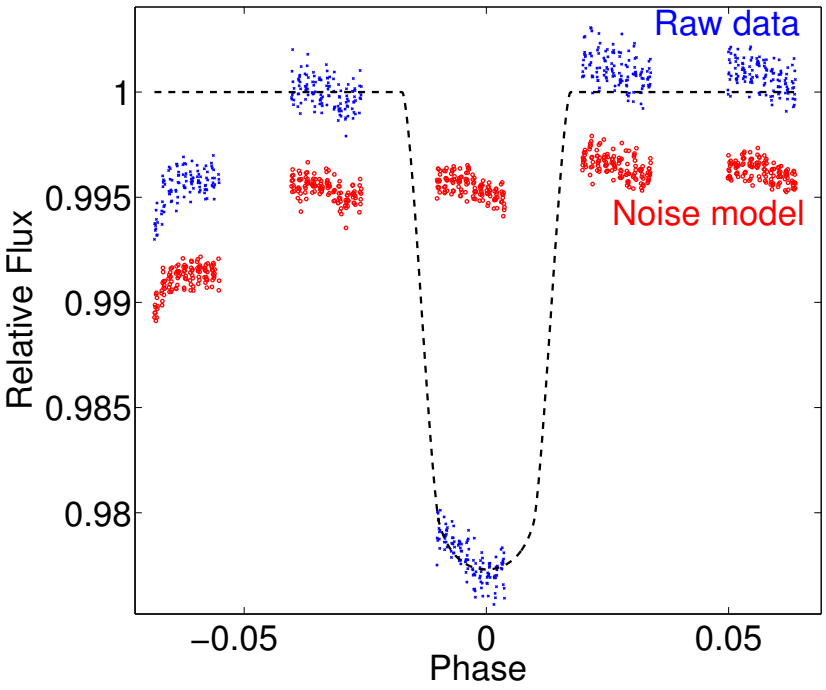


Figure 4.5: showing the same 'raw' light curve as in Figure 4.4 (blue crosses) and the calculated systematic noise model (red circles) offset below.

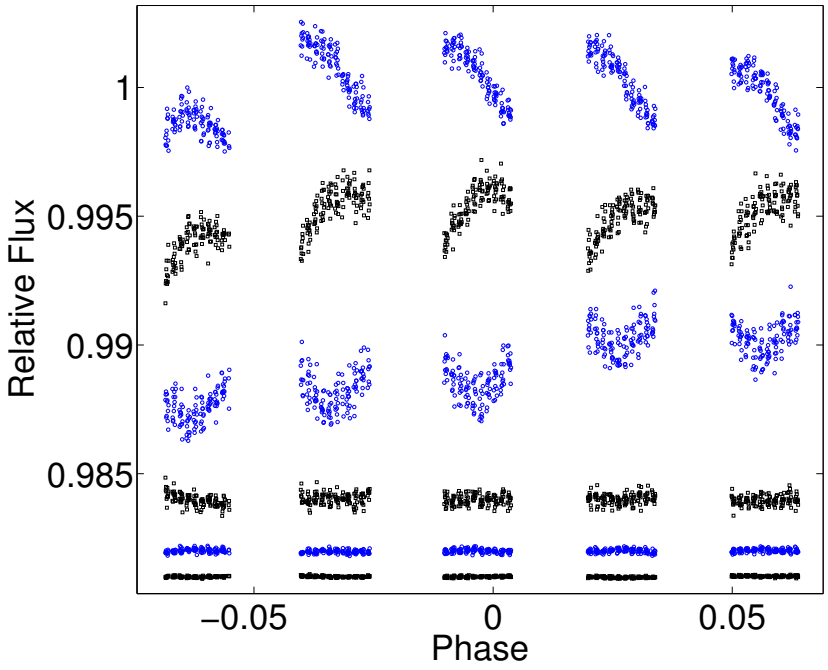


Figure 4.6: Individual systematic noise vectors, $\hat{\mathbf{s}}_{\mathbf{sn}}$, of HD189733b, with the appropriate scaling. Combined they form the systematic noise model in figure 4.5 (red circles).

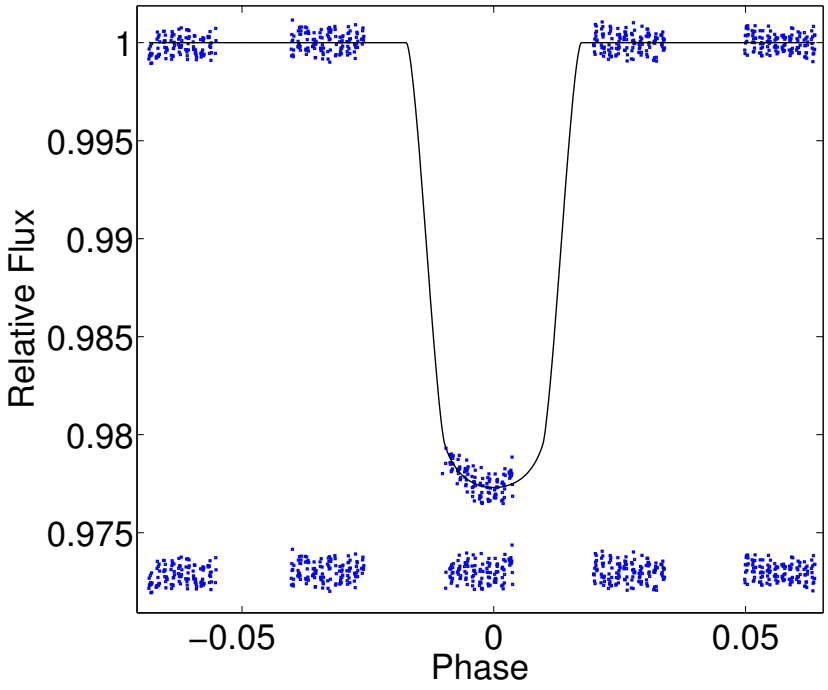


Figure 4.7: showing the de-trended data by subtracting the noise model of the raw data.

by the autocorrelation function (red circles) in figure 4.15.

Method 2: As described in previous sections, figure 4.13 shows the retrieved systematic noise vectors and figure 4.14 features the 'raw' data with the combined systematic noise model (red) underneath. The autocorrelation function of the model fit residual is shown in figure 4.15 (black crosses) and shows a factor 2 improvement on the de-correlation in the lower lags.

Figure 4.12 compares the de-trended light curves of *method 1* and *method 2* and shows the residual of *method 1* - *method 2* (black crosses). There is little difference between both methods indicating that the signal separation for this data-set is close to its maximum with the data being partially de-correlated. This is in contrast to the HD189733b example where a near perfect de-correlation was achieved and can be attributed to the systematics being mostly wavelength invariant in the case of XO1b. In other words, systematic noise components which have a constant weighting throughout the data set cannot be de-correlated using ICA or PCA methods, which is to be expected following equations 3.1 - 3.3.

4.3.3 Discussion and comparison of HD189733b and XO1b

Above I tested the algorithm proposed in chapter 3 on two *Hubble*/NICMOS spectra using equal settings for both cases. Both data sets feature primary eclipse events recorded with the same instrument but differ in the spectral grism used. We find the underlying systematics to be sufficiently different in both cases to allow us to study the behaviour of the ICA based algorithm in de-correlating the data sets. In the case of HD189733b, we find the algorithm to be very efficient in separating the non-Gaussian components from one another as seen in figure 4.6. When we compare these systematic components to the parametrically derived optical state vectors in figure 4.1 we find a high degree of comparability. Please note that the systematic components in figure 4.6 do not necessarily feature the same sign and may appear flipped compared to the OSVs in figure 4.1. It is also worth noting that the systematic-noise vectors calculated by the non-parametric method do not need to resemble the OSVs. The non-parametric systematic components can generally be assume to be the more accurate set of OSVs since these are forced to be statistically independent whilst the parametrically derived OSVs can be correlated with each other. The noise-model as computed by *Method 2* (equation 3.39) is very similar to the linear regression model of equation 4.1 and it is hence not surprising that the non-parametric correction is of very similar quality than the parametric approach of Swain et al. (2008c). We can draw two main conclusions from this results: 1) it is possible to derive the statistically independent systematic noise vectors out of the data itself without the use of auxiliary information and 2) the OSVs computed by Swain et al. (2008c) closely approach statistical independence in this

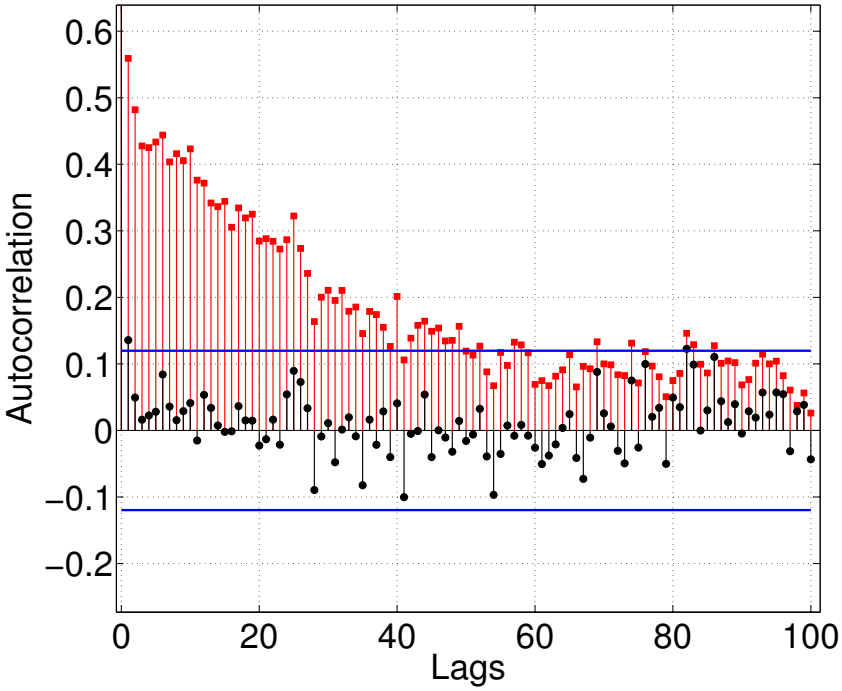


Figure 4.8: showing the autocorrelation function for 100 lags of the fitting residual in figure 4.4 (red squares) and figure 4.7 (black circles). The blue lines signify 3σ limits for a Gaussian distribution. The fitting residual of figure 4.4 shows high amounts of residual correlation, particularly at lower lags whilst the fitting residual of figure 4.7 follows a Gaussian distribution.

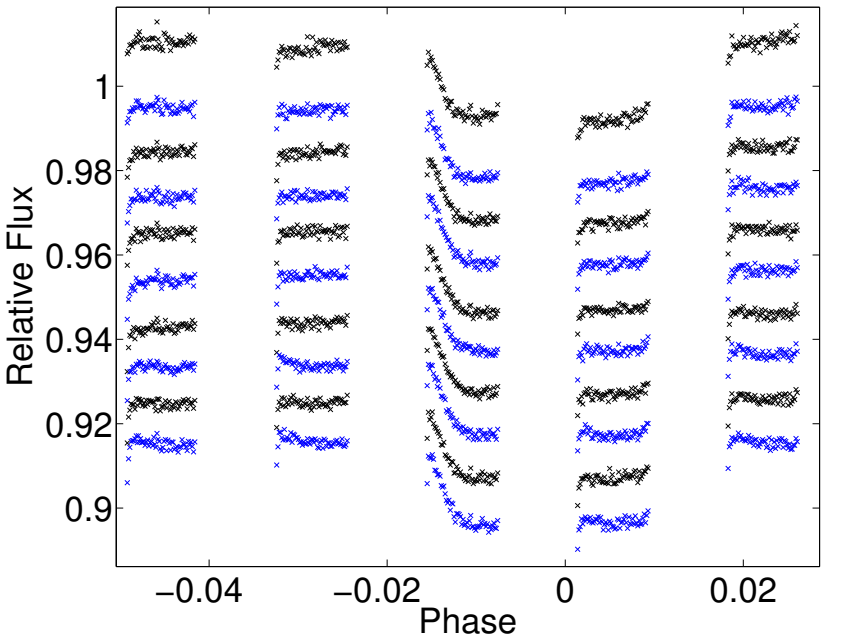


Figure 4.9: showing 'raw', extracted *Hubble*/NICMOS light-curves of HD189733b primary eclipse. Light curves are offset for clarity, bluest at the bottom to reddest at the top.

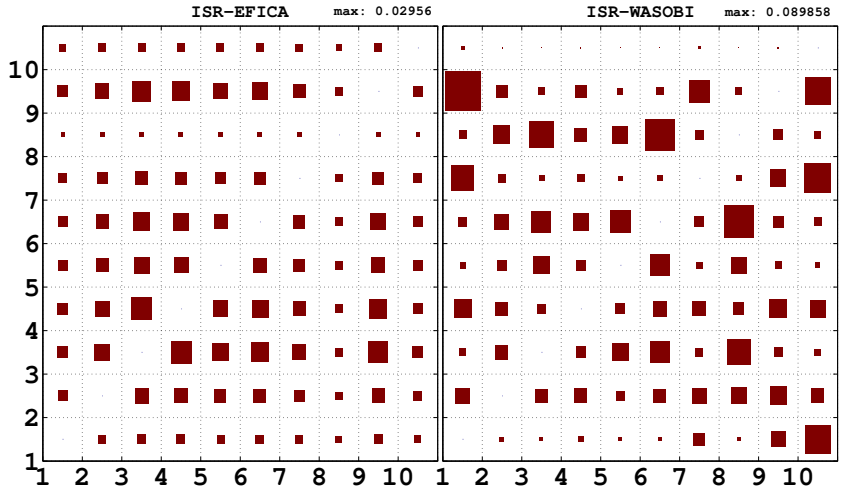


Figure 4.10: same than for figure 4.3. The light curve vector (component 1) shows residual interference with other vectors for both EFICA and WASOBI algorithms. Overall the EFICA algorithm outperforms WASOBI.

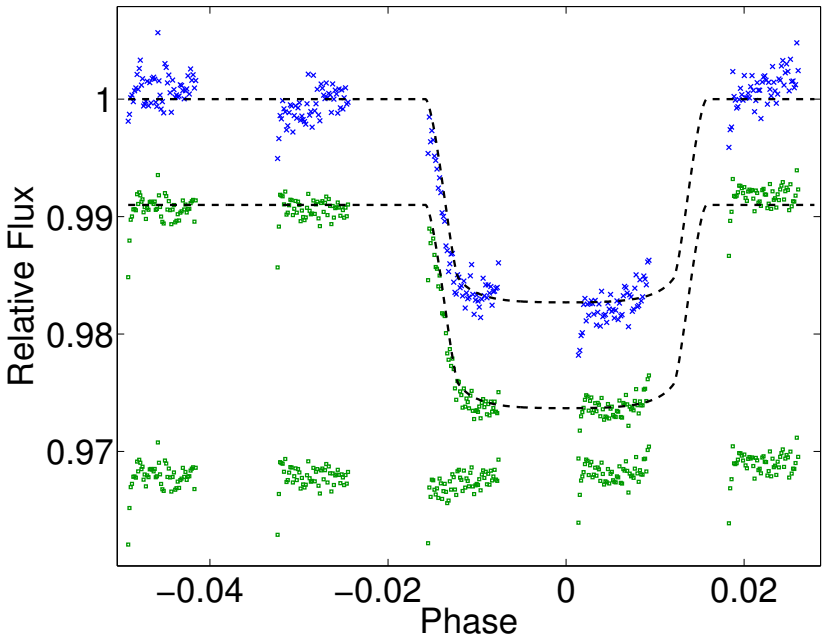


Figure 4.11: showing the raw-data light curve (blue circles) and the corrected light curve (green squares) offset below. In this example, we used equations 3.36 & 3.37 as light curve filter. The systematic noise components were reduced but residual systematics remain in the final light curve.

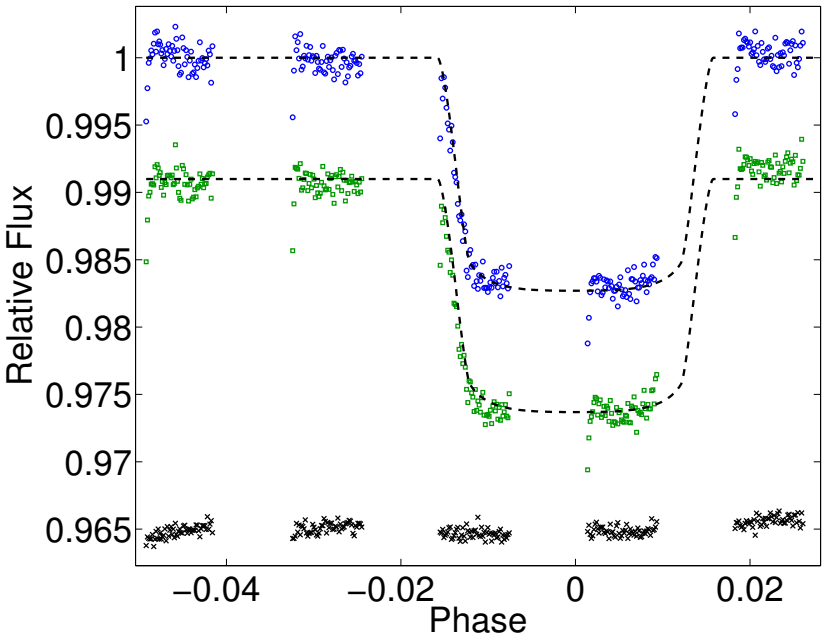


Figure 4.12: showing the de-trended data using method 1 (top blue circles) and method 2 (bottom green squares) offset from each other. Both results show little differences between them as seen by the residual of method 1 - method 2 (black crosses).

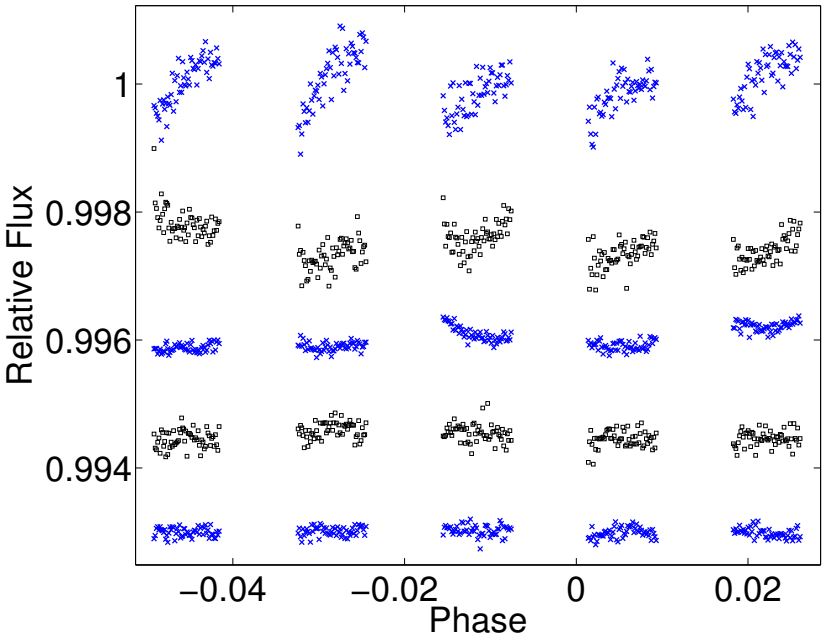


Figure 4.13: Individual systematic noise vectors, $\hat{\mathbf{s}}_{\mathbf{sn}}$, of XO1b, with the appropriate scaling. Combined they form the systematic noise model in figure 4.5 (red circles).

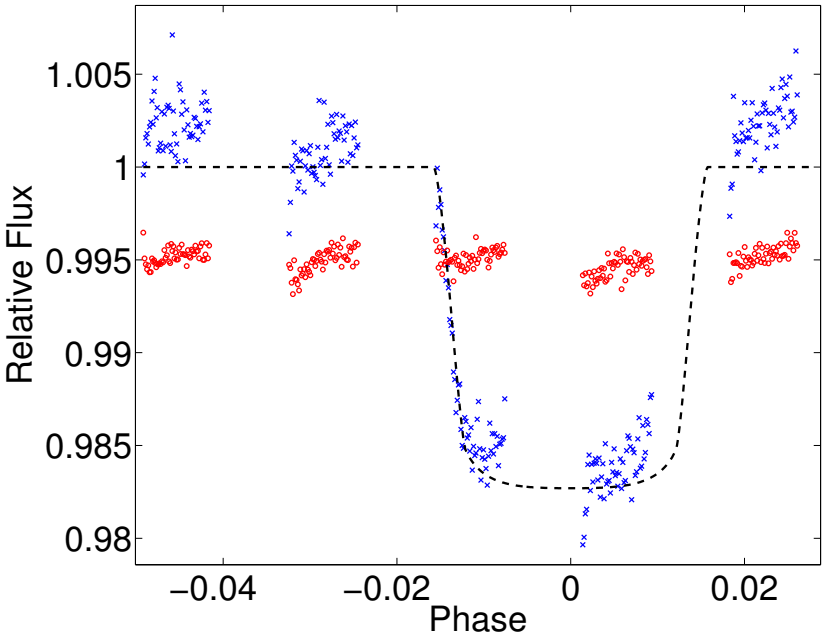


Figure 4.14: showing the same 'raw' light curve as in Figure 4.11 (blue squares) and the calculated systematic noise model using the systematic noise vectors in figure 4.13.

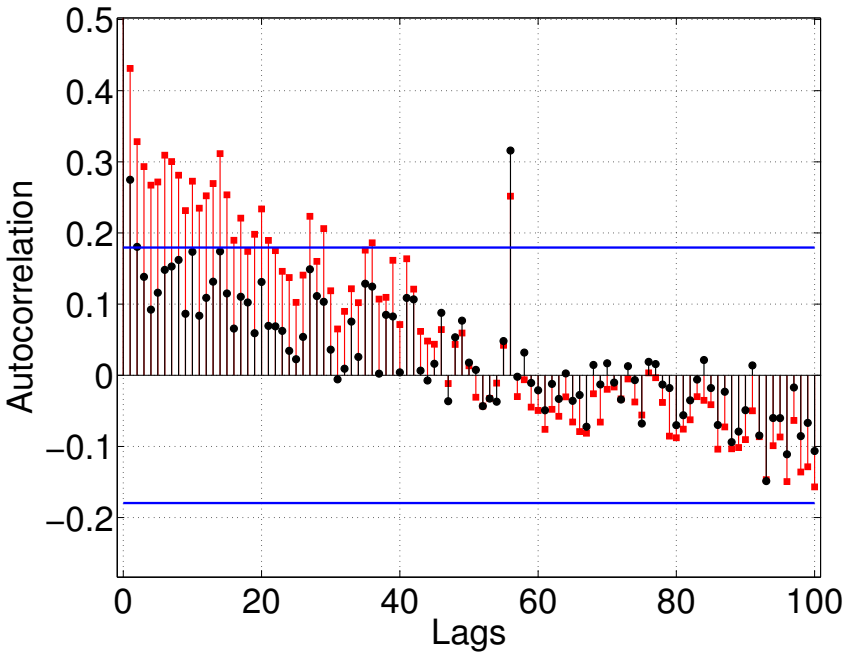


Figure 4.15: showing the autocorrelation function for 100 lags of the fitting residual for method 1 (red squares) and method 2 (black circles). The blue lines signify 3σ limits for a Gaussian distribution. The fitting residual of method 1 shows residual correlation, particularly at lower lags whilst the fitting residual of method 2 is by a factor of two better de-correlated in the lower lags.

particular case.

The de-trending results for XO1b differ in several respects to those of HD189733b. Here, the de-trending is not as complete as for the previous case with auto-correlative noise significantly reduced (figure 4.15) but continuing to exhibit time-correlated structure in the final result. This is starkly different to the case of HD189733b where de-correlation is very much complete. As discussed in section 4.3.2, we computed the de-correlated spectrum using both, *Method 1* and *Method 2*. In the case of HD189733b we find *Method 2* to perform significantly better in de-correlating the data. However in the case of XO1b we see little difference (figure 4.12, black crosses) between both approaches. This is indicative of the de-correlation, given the systematic noise vectors, to be at its optimum. We can hence argue that the systematic noise vectors are not optimally derived by the algorithm, which is also supported by the ISR matrix in figure 4.10. There can be several reasons for this behaviour: 1) the residual systematic noise is stationary in wavelength and/or time, 2) the higher scatter in the raw data impaired the de-convolution process. If the residual systematics are stationary in wavelength and/or time, we find the weighting coefficients a_{ij} comprising the mixing matrix $\mathbf{A}$ (equation 3.3) to be constants. In this case, the instantaneous mixing model cannot discern any differences between systematics and astrophysical signal. This is a fundamental limit of the method. Figure 4.13 shows the systematic noise components retrieved. Compared to those of HD189733b we can see a much increased scatter in the individual components, which can impair the minimisation process. In section 3.4.3 we described a possible pre-processing step using regressive kernel smoothing of the raw data. However, in the case of XO1b we found little improvement to the scatter of the systematic components. The de-correlation of Gaussian data will be further discussed in chapter 7.

4.4 Computing the full spectrum of HD189733b

Following on from the previous analysis of *Hubble*/NICMOS HD189733b and XO1b data sets, we proceeded to extend the data de-trending to compute the spectrum across the entire spectral range (Waldmann et al., 2012a). It should be noted that the previous examples were based on the spectral extraction used in Waldmann (2012) whilst the raw data of this section and in Waldmann et al. (2012a) is based on a better optimised spectral extraction of the observed data. It is hence not surprising that the raw lightcurves and computed systematic noise components are slightly different to the previous example.

The G206 grism covers the spectral range of ∼1.51 - 2.43 μm. In order to minimise inter- and intra-pixel variability of the NICMOS detector, the instrument was slightly de-focused to a full-width-half-maximum (FWHM) of ∼5 spectral channels per resolution element. This sets a limit on the maximum resolution, R, achievable. In the previous analysis, we used a binning of 10 spectral channels but reduce this to a higher resolution of 8 spectral channel bins ($\sim 0.09\mu$m). This lies slightly above the minimum of 5 in order to boost the signal-to-noise (SNR) without a significant reduction in R. This resulted in 11 light curves across the G206 grism band. We found the first of the 5 orbits to be very noisy and negatively impacting the efficiency of the algorithm, in particular at the edges of the spectrum, and excluded the first orbit from all further analysis. This is in accord to other analysis in the literature (Swain et al., 2008c; Gibson et al., 2011, 2012). An example of the 'raw' light-curves' quality, at ∼2.33 μm, can be found in figure 4.17.

As described in the previous sections and Waldmann (2012), we used the extracted light-curves as input to the ICA algorithm to calculate the mixing matrix, $\mathbf{A}$, and its (pseudo)inverse the de-mixing matrix $\mathbf{W} = \mathbf{A^{-1}}$. Once the de-mixing matrix had been determined, the algorithm tested the estimated components for their nongaussianity and returned four main systematic noise components which do not correlate with the expected light-curve morphology. These components, comprising $\mathbf{s_{sn}}$, were extracted over the entire spectral range of the grism (from hereon referred to as 'global') and showed a good degree of separation (figure 4.16, left side). In section 4.3.1 we could show that *Method 2* is more efficient in the de-trending of the data and we proceed to use *Method 2* only. A Nelder-Mead minimisation algorithm (Press et al., 2007) was used to fit for $\mathbf{O_k}$ in the noise model $m_k = \sum_{l_{sn}}^{N_{sn}} \mathbf{O_k s_{sn}}$. The scaling amplitude of each component for each light-curve is given in figure 4.16 (right side). Once m_k was determined, we subtracted it from the raw data to get the corrected timeseries $y_k = x_k - m_k$, see figure 4.17 for an example. The corrected light-curve, y_k, was then fitted using Mandel and Agol (2002) with orbital and limb-darkening parameters taken from Swain et al. (2008c) and Claret (2000) respectively, leaving the transit depth, δ, as only free parameter. The transit depths of these 11

light-curves constitute the exoplanetary spectrum.

4.4.1 Determining the Error-bar

The previous sections did not feature a definition of the spectral error-bar associated to each lightcurve. This was not necessary as the error on the depth is determined in the standard way by computing the sum of squares of the out-of-transit scatter σ_{OOT} , in-transit scatter σ_{INT} and a a correlated-noise error (CNE), σ_{CNE}. However, by computing the full spectrum rather than detrending individual lightcurves, we can introduce a fourth error term: a systematic-noise-model error (SNME), σ_{SNME}, which leads to the following definition of the spectral error-bar:

$$\sigma_\delta = \sqrt{\sigma_{OOT}^2 + \sigma_{INT}^2 + \sigma_{CNE}^2 + \sigma_{SNME}^2} \tag{4.5}$$

The first two terms are dominated by photon or white noise whilst the last two determine the amount of nongaussian noise and correction errors. The CNE is calculated using 100 iterations of a classic Monte-Carlo boot-strap (or Jackknife) analysis (Press et al., 2007). Finally, the SNME accounts for possible over-corrections of the global SNM to individual, poorer constraind light-curves. This term becomes non-zero when the scaling of a systematic noise component, $o_{kl_{sn}}$ (figure 4.16, right hand side), shows a 3σ significant deviation from the mean scaling of all other light-curves, $\bar{o}_{l_{sn}}$. In other words, we expect the scaling of an individual systematic noise component, $\mathbf{s_{sn}}$, to be a slowly varying function over wavelength for 'globally' estimated systematic-noise components. If individual light-curves show a significantly larger positive or negative scaling than expected, we can assume that the nongaussian noise of the affected light-curve cannot be explained by the global model. In larger data-sets it is easier to exclude the affected light-curve from any further analysis, whilst in small data-sets we take the amplitude of the scaling from its mean scaling as the error, i.e. $\sigma_{SE} = |(o_{kl_{sn}} - \bar{o}_{l_{sn}})|$.

4.4.2 Results

Figure 4.17 shows the raw light-curve at $\sim$2.48 μm in black crosses with its corrected counterpart (blue circles) offset below. Here, much of the autoregressive noise in the original data could be captured by the noise model (red squares) and removed from the final result. All corrected light-curves are shown in figure 4.18 and the resulting spectrum presented in figure 4.19 and table 4.1. We find the retrieved spectrum to be in good agreement with Swain et al. (2008c) and Gibson et al. (2012). The underlying noise of the spectral point at $\sim$2.06μm was flagged by the algorithm to be discrepant with the global systematic noise model. Here the first systematic

noise component is indicative of an overcorrection which is reflected in the error-bar as described in the previous section.

4.4.3 Discussion

In this analysis we have computed the global SNM and found an excellent agreement with previously published results. The error-bars reported here are marginally larger than those of Swain et al. (2008c) towards the red part of the spectrum. Both methods use a bootstrap error-analysis and therefore allow for a direct comparison of the parametric de-correlation approach by Swain et al. (2008c) and the non-parametric, blind approach described in this publication. We attribute the differences to the performance of the global SNM for different parts of the spectrum. We find the global SNM approach to be most sensitive to slowly varying systematic trends across one's data-set and local nongaussian deviations tend not to be captured. It is therefore possible to additionally compute a 'local' SNM for a sub-set of spectral bins or before binning on the individual 'raw' spectral channels, should the SNR permit it. It is important to remember that $k \geq N_{sn} + 1$ as the input to the algorithm. In other words, at least as many observed time series, x_k, are required as input to the algorithm than total number of nongaussian components in the data. In this case, the minimum 'local' SNM would include 5 spectral bins (4 systematic noise and 1 astrophysical component). With 11 spectral bins in total, the NICMOS data-set is too small for this approach. However, for larger sets (e.g. Waldmann et al., 2012b), this two stage 'global' + 'local' detrending becomes a viable solution.

Figure 4.19 compares the spectrum obtained with our ICA approach to Swain et al. (2008c) as discussed above as well as Gibson et al. (2012), who used *Gaussian Processes* to marginalise over auxiliary information parameters (optical state vectors) of the instrument. The very different nature of all three analysis techniques used and the excellent agreement between their respective results showcases the high degree of stability of this exoplanetary spectrum.

It is interesting to note that the spectral point at $\sim$2.06μm is difficult to constrain using non-parametric techniques, which is reflected in larger error-bars (compared to neighbouring points) in both the analysis presented here and that by Gibson et al. (2012). This spectral region is of particular importance as it is the main constraint on CO_2 abundance of HD189733b in the limited wavelength range of the grism. A higher R_p/R_s value than reported by Swain et al. (2008c) could be indicative of a stronger CO_2 contribution than initially thought, which would also be supported by the planet's observed day-side emissions (Swain et al., 2009b).

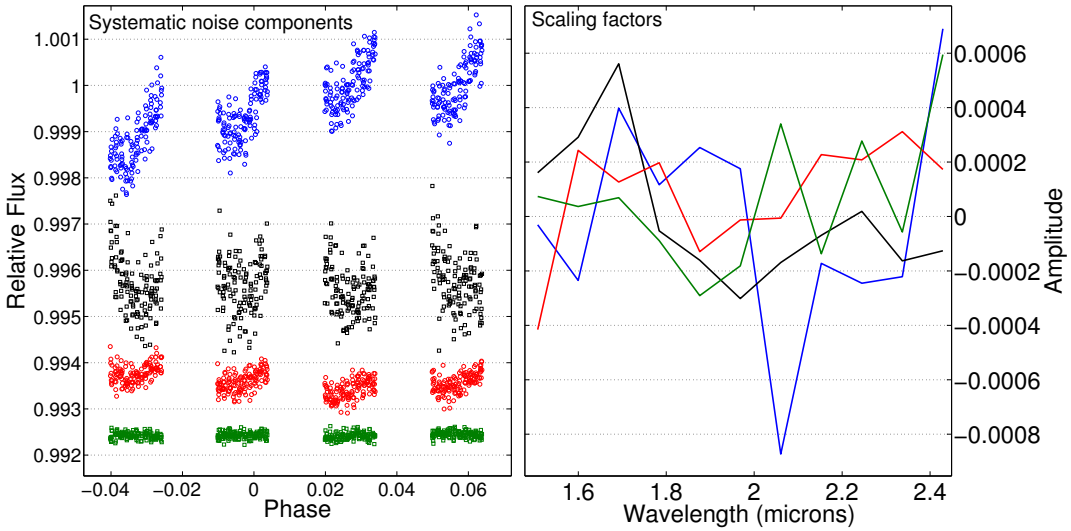


Figure 4.16: LEFT: Four retrieved nongaussian systematic noise components in the order of importance. They were computed over the whole spectral range of the G206 grism and describe the systematic noise (instrumental and/or stellar) common to all spectral channels. RIGHT: Scaling factors of the systematic noise components on the left. The colour coding is identical for both plots. We can see that the first component at 2.06μm is sharply deviating from its own pass-band mean and the mean of all components at 2.06μm. This can indicate that the 'global' systematic noise model does not well describe the systematics in this channel and may be prone to over-corrections.

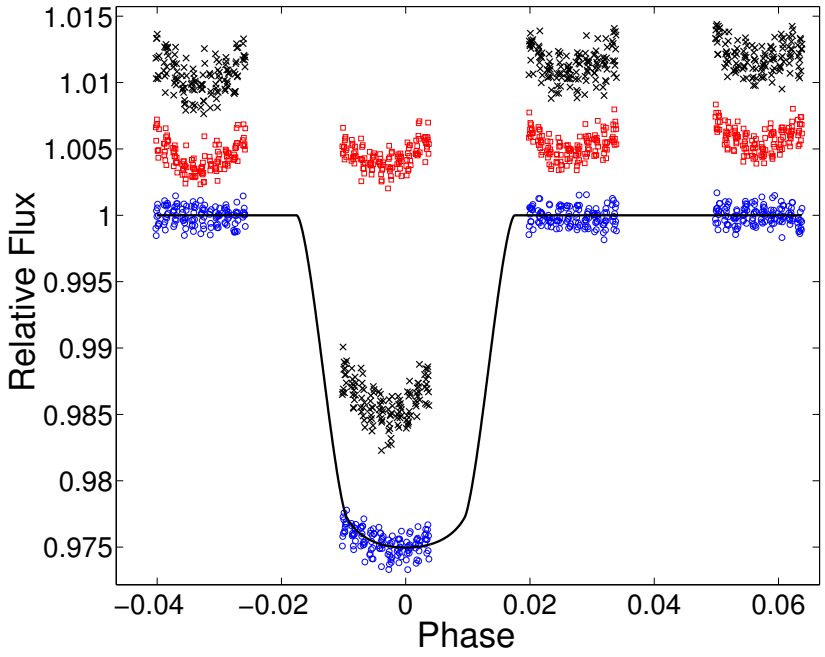


Figure 4.17: Raw light-curve at ~2.33 μm (black crosses), its respective systematic noise model (red squares), $m_k(t)$, composed out of the systematic components in figure 4.16. The de-trended final light-curve is shown underneath (blue circles) with a Mandel and Agol (2002) fit overlaid.

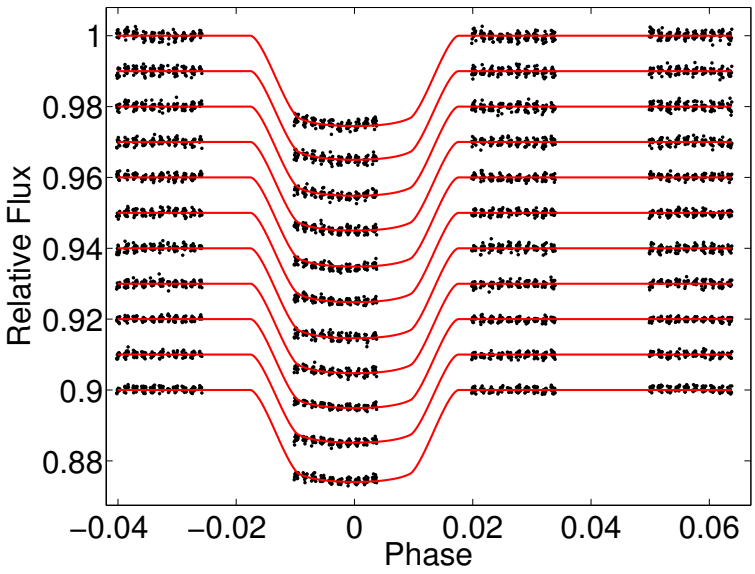


Figure 4.18: Final de-trended light-curves from 1.51μm (bottom) to 2.43μm (top) with fitted Mandel and Agol (2002) model overlaid.

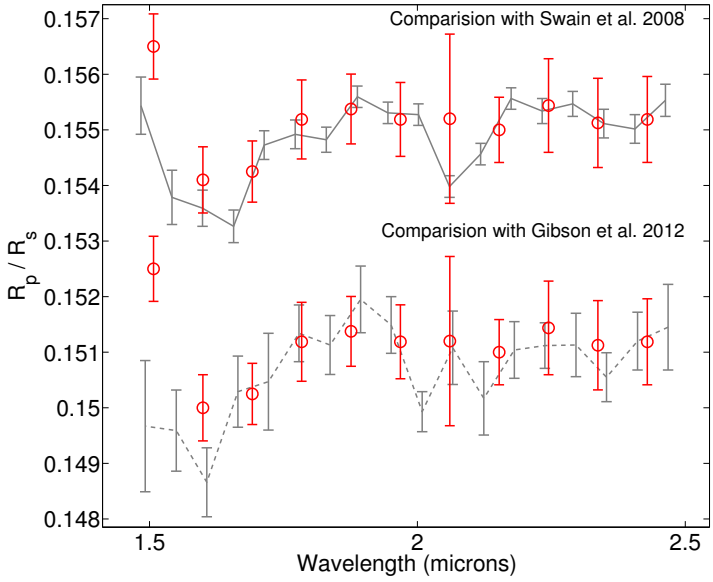


Figure 4.19: Final spectrum (red circles) obtained with the ICA algorithm described here and in Waldmann et al. (2012b), overlaid on the Swain et al. (2008c) result in grey (top,solid line) and Gibson et al. (2012) (bottom, discontinuous line). The bottom spectrum was shifted by $R_p/R_s = 0.004$ for clarity. The spectrum reported here is in good agreement with both, the 'parametric' analysis of Swain et al. (2008c) and the 'non-parametric' analysis of Gibson et al. (2012) and show-cases the robustness of this methodology and the stability of the result as a whole.

4.5 Photometric lightcurves, the case of Kepler

In the previous examples we have shown that spectroscopic datasets can be de-correlated effectively. It now becomes interesting to investigate how well the proposed algorithm can de-correlate consecutively observed data-sets. This is of particular interest in cases where no multi-channel data are available and the time series data are contaminated with time-correlated noise, be it stellar or instrumental.

As opposed to the previous examples, where several time series, x_k, were observed simultaneously, here we take a single time series covering several consecutive eclipse features and cut the time series into segments spanning equal lengths over each eclipse event. Using these segments as inputs to the algorithm clearly violates the underlying assumptions of the independent component analysis, as the mixing is not instantaneous. In this case, the ICA analysis can be understood as a Projection Pursuit (PP) analysis, see section 3.2.2 and (Hyvärinen and Oja, 2000; Stone, 2004; Huber, 1985). Here the ICA algorithm, in the absence of a working ICA data-model, will try to extract as many non-Gaussian components as possible and return the rest of the data in its original form. This is very similar to Projection Pursuit, where the data is not described by an underlying data model at all but only the most non-Gaussian component is retrieved. In other words, we can only expect to retrieve the eclipse signal component, $\hat{\mathbf{s}}_\mathbf{a}$, with any degree of accuracy. As a result we will not be able to retrieve systematic noise components, $\hat{\mathbf{s}}_\mathbf{sn}$, and we can only use *Method* 1 (in section 3.3.3) to de-trend the data.

We have downloaded data observed by the Kepler space telescope (Borucki et al., 1996, 2010; Jenkins et al., 2010; Caldwell et al., 2010; Koch et al., 2010) for a planet-hosting candidate star observed over the second and third data-release quarters (Q2 & Q3). The time series, with the Kepler ID: 10118816, exhibits highly variable features and significant time-correlated noise (see figure 4.20, blue crosses). Given Kepler's superb instrument calibration, we can assume this time-correlated noise to be due to stellar variability. Using the periodogram calculator on the NSteD database[3], we identified four main periodically recurring signals in the data-set. Choosing the second strongest feature with a period of 0.040915 days, we phase folded the data and cut the time series in 10 equally sized segments. Gaps in the data were filled with a constant value to ensure that all segments are correctly phase-folded. Since these gaps do not appear at the same position for the phase-folded curves, the algorithm does not get biased towards these interpolations and the interpolations are later on removed from the final result. As for the previous spectroscopic examples we now took these pseudo simultaneous time series segments as input to the algorithm.

[3] *http://nsted.ipac.caltech.edu/applications/ETSS/Kepler_index.html*

We performed our de-correlation as for the previous examples but using *Method 1* only. Figure 4.21 shows the ISR matrix of the separation indicating a relatively poor separation of the components but two (components 4 & 9). As discussed above, this behaviour is to be expected with the breaking of the instantaneous mixing model. Nonetheless, we obtained a clear feature (component 4) in our analysis which is over plotted (red circles) on the mean, phase-folded data (blue crosses) in figure 4.22. Here the de-correlated signal has a much reduced scatter compared to the mean of the phase-folded feature, which indicates that much of the unwanted stellar variability has been removed. It is also clear from this figure that we are not dealing with an exoplanetary light curve but a stellar pulsation feature. As expected, the remaining components returned from the algorithm (figure 4.23) are the residuals of the input data minus the component shown in figure 4.22. Hence we only used the component in figure 4.22 to re-construct the original time series. This was done by using equation 3.37 on each segment of the time series, followed by adding the segments back together in the order they were originally split up.

Figure 4.20 shows the original input data (blue crosses) with the filtered signal (red circles) over plotted on top. In the bottom plot (a zoomed in version of the time series), it is clear that the desired feature remains in the filtered time series whilst the contribution of other time-correlated stellar noise is substantially reduced.

4.6 Discussion

Here we tested the ICA algorithm on phase folded consecutive data. Such an approach is a strict violation of the instantaneous mixing model (IMM) as outlined in equations 3.2 & 3.3 and independent component analysis cannot theoretically accept such a data input. However, in section 3.2.2 we discuss the similarity of a deflationary-ICA code to the more general concept of Projection Pursuit (PP). The deflationary-ICA algorithm is not technically concerned with the validity of the IMM as it computes each independent component in series, always aligned to the axis of highest non-Gaussianity. In this case the deconvolution of time consecutive data becomes feasible as the feature of the highest 'common' non-Gaussianity is in fact the feature over which we phase-folded the time series data. It is needless to say that the act of phase-folding inherently biases the ICA algorithm to pick out the feature over which we have folded the data. In other words, for time consecutive data, we need to cherry pick one specific recurring feature in the data which will then be 'cleaned' by the ICA algorithm, removing or dampening all other features from the time series. Such a 'cherry picking' is usually fine since we are mostly interested in the periodically recurring transit feature rather than stochastic or systematic noise in the time

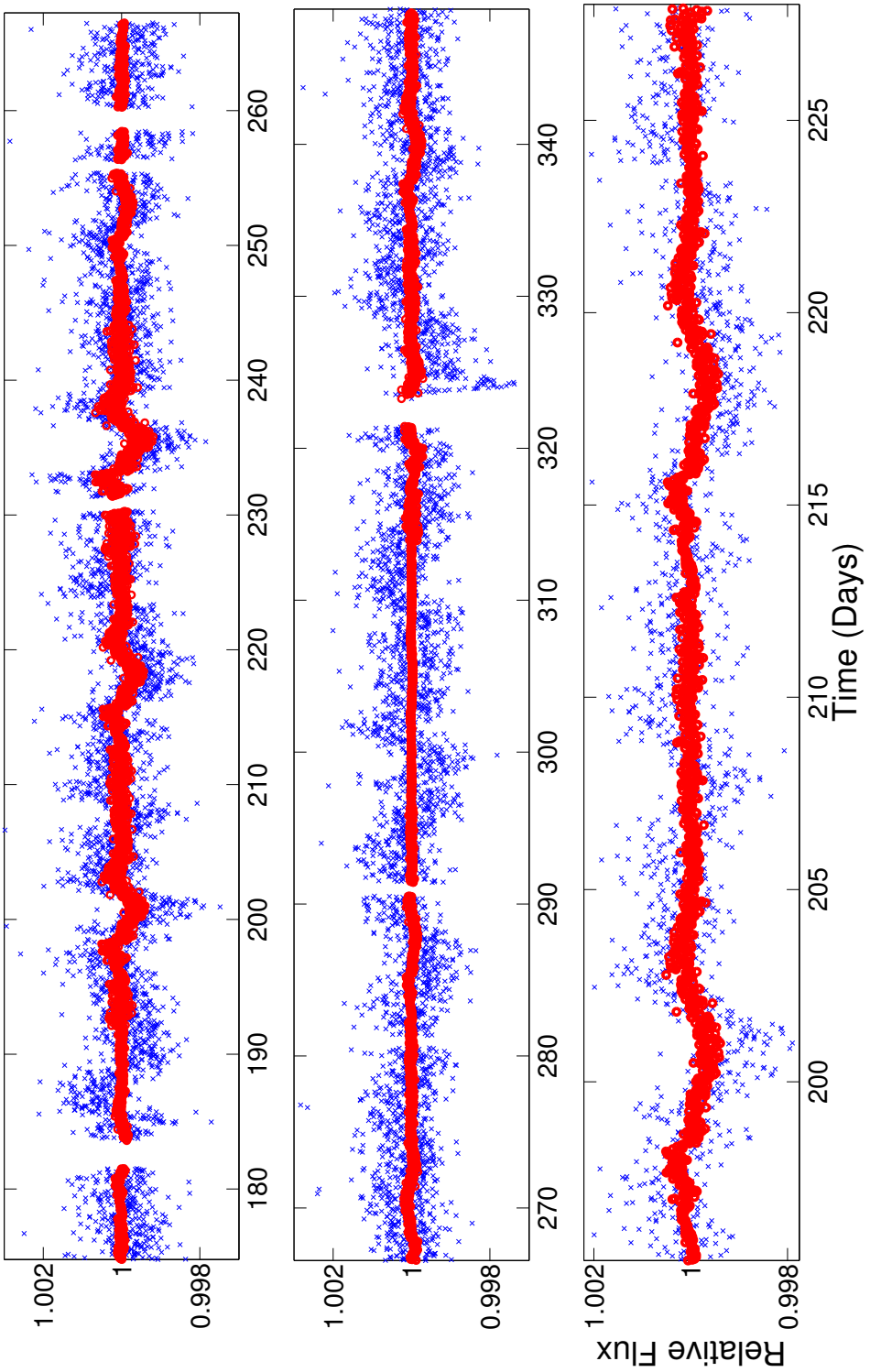


Figure 4.20: Input time series (blue crosses) with filtered signal using *Method* 1 over plotted (red circles). Bottom plot is a zoomed in part of the time series above.The algorithm effectively filtered for the desired feature and strongly decreased contributions from autocorrelated noise.

λ (μm)	R_p/R_s	$\Delta(R_p/R_s) \times 10^{-4}$
2.429	1.55187	7.73230
2.336	1.55125	8.02697
2.244	1.55437	8.42058
2.152	1.55000	5.86704
2.060	1.55200	1.52237
1.968	1.55187	6.64653
1.876	1.55375	6.28246
1.784	1.55187	7.10625
1.691	1.54250	5.50374
1.599	1.54000	5.95214
1.507	1.56500	5.85648

Table 4.1: NICMOS transmission spectrum of HD189733b for a 'global' systematic noise model correction and plotted in figure 4.19. The columns are wavelength (λ), planet-star-ratio (R_p/R_s) and the respective error-bar ($\Delta(R_p/R_s)$).

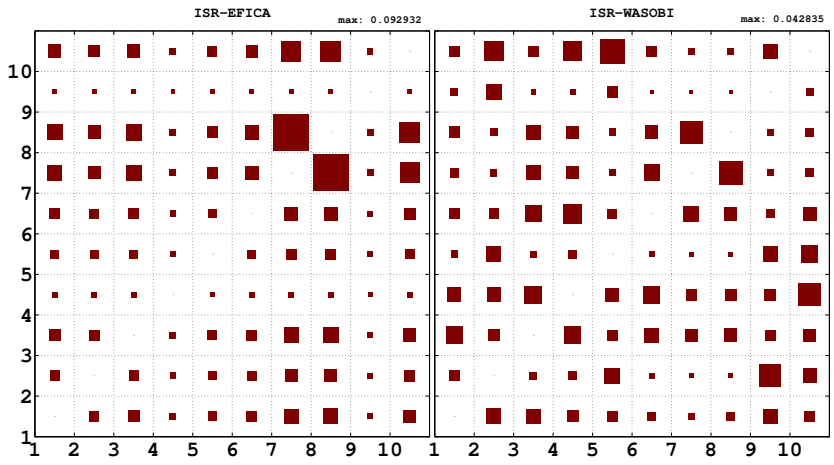


Figure 4.21: the Interference over Signal (ISR) matrix of the component separation for both the EFICA and the WASOBI algorithms. All values were normalised with the maximum ISR = 0.09293. Components 4 and 9 are the best separated, with component 4 being the desired signal component.

series data.

In this analysis we used a combination of EFICA and WASOBI (see chapter 3) algorithms to de-trend the data. It is interesting to note that both these algorithms are strictly speaking parallel and not deflationary. This means they are technically bound by the IMM and cannot allow violations of the same. Nonetheless we find the single signal to be non-Gaussian enough to be retrieved by the parallel code, forcing the other components to be estimated in a pseudo deflationary way. Such a behaviour is particularly interesting given the advantages with regards to convergence of a statistically efficient parallel code compared to a simpler deflationary model.

The application to time-consecutive transits is particularly important for treating variability of the host-star, in particular for photometry cases with only one wavelength channel to work with, i.e. Kepler. In particular for such higher accuracy measurements as provided by Kepler but not excluding the most general case, the stellar activity can significantly impair the quality of the final science result (eg. Czesla et al., 2009; Boisse et al., 2011; Aigrain et al., 2011; Ballerini et al., 2012). Using ICA as selective and optimal filter can provide a viable non-parametric solution to this problem.

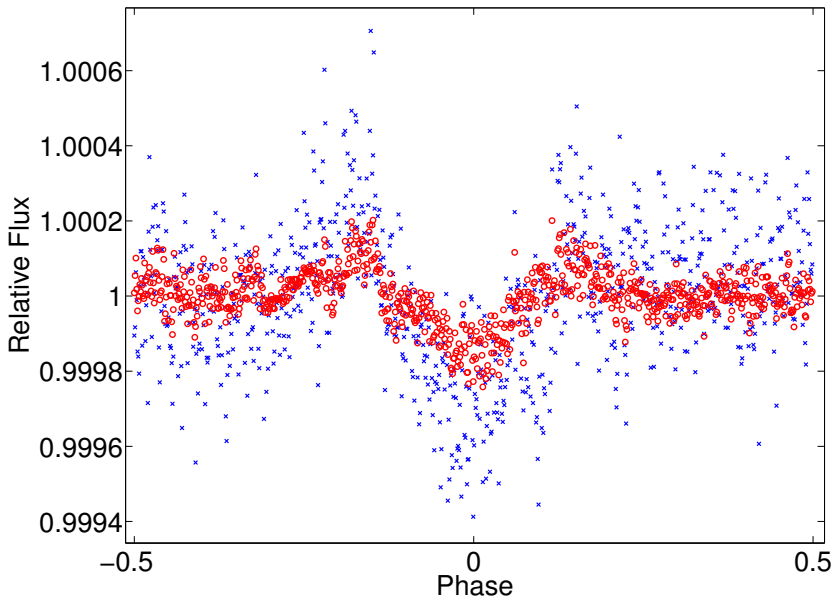


Figure 4.22: showing the mean, phase-folded feature (blue crosses) with the ICA filtered signal component (red circles) over plotted. The ICA filtered signal shows a significant reduction in scatter and auto-correlative noise compared to the simply phase folded data.

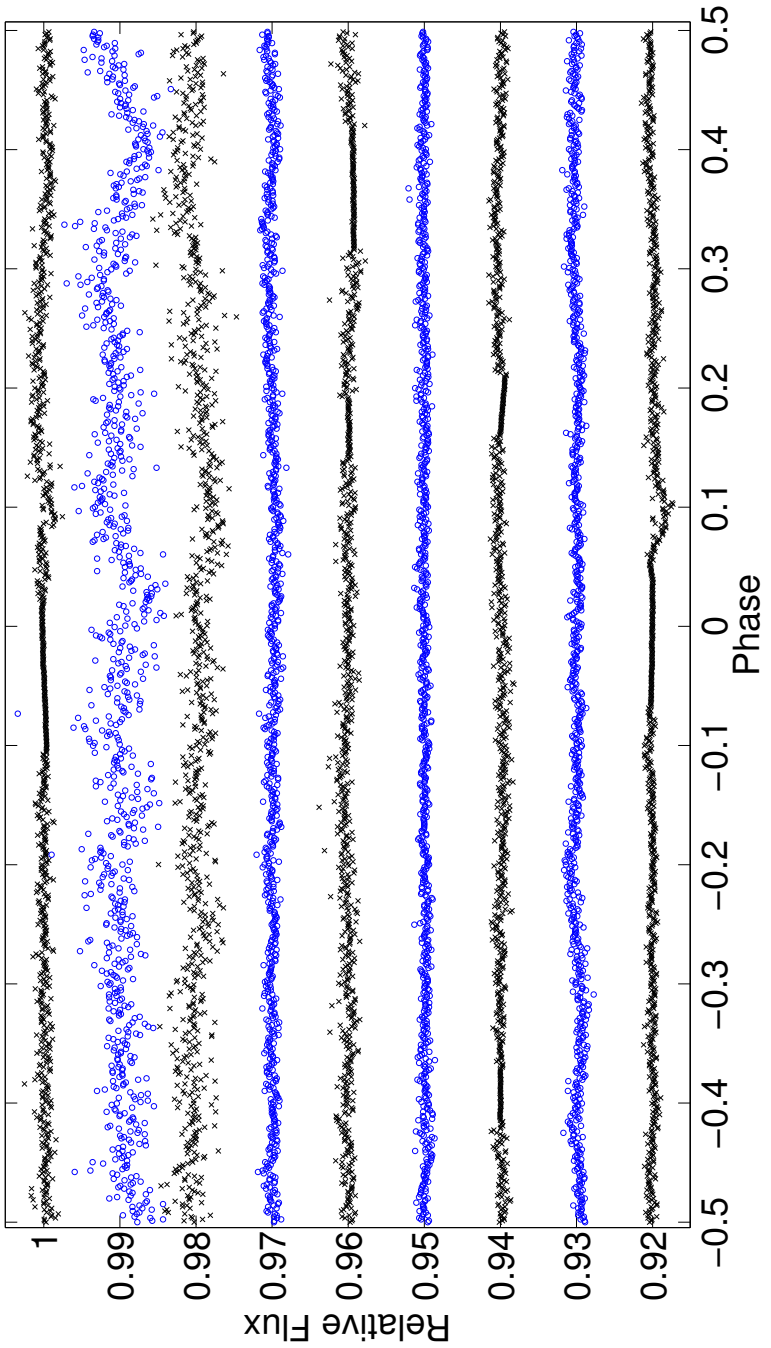


Figure 4.23: Addition components to the signal in figure 4.22 as calculated by the algorithm.

4.7 Conclusion

This chapter was concerned with data applications of the blind-source deconvolution algorithm introduced in chapter 3. We tested the proposed algorithm on three data sets, two *Hubble*/NICMOS spectroscopic observations of HD189733b and XO1b and one time-consecutive observation of a pulsating star observed with the Kepler spacecraft. We began by de-trending individual lightcurves for HD189733b and XO1b and compared the de-trending efficiency of the algorithm for different grisms of the *Hubble*/NICMOS instrument, namely grisms G206 and G141 respectively. We found a near perfect de-correlation of the HD189733b data whilst that of XO1b was only partially de-trended. Following on from this initial study, we proceeded to extend this de-trending effort to the entire spectral range of the G206 grism and found the results to be in excellent agreement with previously reported results, demonstrating not only the validity of this approach but also the stability of the scientific result. From this study we can learn the following insight on blind-source deconvolution of exoplanetary spectroscopic data:

- In data sets with low inherent Gaussian noise and variable non-Gaussian systematic noise, we can achieve a near perfect de-trending of the data as for HD189733b.
- *Method 2* which computes the noise-model by iteratively fitting the systematic noise components to out-of-transit data can be shown to yield better de-trended results than *Method 1* which filters the astrophysical signal directly.
- The algorithm cannot de-convolve non-Gaussian noise that is either static in wavelength and/or time, as the weighting coefficients comprising the mixing matrix $\mathbf{A}$ become constants and the non-Gaussian signal becomes indistinguishable from the astrophysical signal.
- We can speak of 'global' and 'local' systematic noise models, depending whether the whole of the spectral range was taken as initial input or localised parts. The 'global' model can effectively de-trend slowly and consistently varying systematic noise sources well but is prone to over-correction if localised parts of the spectral range exhibit starkly varying noise properties. The 'global' model can be complemented with a localised noise model where additionally to the slowly varying systematics, the localised morphology of the non-Gaussianities is captured by rerunning the algorithm on individual areas.
- From the distribution of the systematic noise component scaling factors (figure 4.16 right hand side), we can identify possible over-corrections by the 'global' noise model and either choose to exclude this spectral range from further analysis or propagate this over-correction to the associated error-bar, in a self-consistent way.

Following from the analysis of spectroscopic data sets that feature a set of simultaneously observed time series, we tested the algorithm a time-consecutive time series of an active star obtained by the Kepler spacecraft. The detection of faint exoplanetary eclipses are often made difficult by time-correlated activity of the host star. We demonstrated, using a single Kepler time series, that much of the stellar variability can be removed in time series that span several exoplanetary eclipse events. Based on the broader concept of Projection Pursuit we discuss the feasibility of selective filtering of features by phase-folding the time series about the period of the desired morphology. This pseudo simultaneous data set clearly violates the assumptions of the instantaneous mixing model but we can show that the predominant, non-Gaussianity over which we phase-folded the data can be retrieved with good accuracy and the data be re-constructed with all other non-Gaussian noise sources dampened or suppressed. Furthermore, we showed that interpolations of data-gaps do not impair the efficiency of the algorithm.

Chapter 5

Exoplanetary spectroscopy from the ground

"The struggle itself is enough to fill a man's heart.
One must imagine Sisyphus happy."

— **The myth of Sisyphus (Albert Camus)**

5.1 Introduction

In the previous chapters I discussed the de-trending of space-based data. Whilst I showcased the removal of systematic noise for *Hubble* and *Kepler* data, we could apply these techniques to any other time-series measurement, such as mid-infrared measurements with the *Spitzer* telescope. Despite the residual instrument systematics, which we can remove in most cases, space-based data are often characterised by a high degree of temporal stability and a very high signal-to-noise ratio of the final science products. It is therefore unsurprising that they have been the preferred choice for the characterisation of exoplanetary atmospheres in the past decade.

Unfortunately, this fairy tale of space-borne remote-sensing of foreign worlds is half-way through its final chapter. With the depletion of the helium cryogen tank on *Spitzer* in the summer of 2009, we have effectively lost our eye and all spectroscopic facilities in the mid-IR, leaving us with two photometric *Spitzer*/IRAC channels at 3.6 and 4.5 μm in the extended mission. However without active cooling, the remaining IRAC channels feature a much higher background noise level, deteriorating the measuring accuracy. In the UV to near-IR, our eye in the sky is and has been the *Hubble Space Telescope*. *Hubble* has played a pivotal role in the first discoveries of exoplanetary atmospheres (Charbonneau et al., 2002) and continued its success with the near-IR *Hubble*/NICMOS instrument and its successor WFC3. In the spring 2009, the

HST received its last servicing mission and with the end of the Space-Shuttle era in 2011, no more servicing missions are planned nor possible.

With the progressive loss of the current space-based facilities and their successors' (JWST, EChO, FINESSE) launch dates being 6-12 years from now, it is unsurprising that ground-based facilities enjoy increasing attention. Whilst it is true that the photon-collection areas of ground-based telescopes are a multifold of that of *Hubble*, *Spitzer* or *Kepler*, observations of exoplanetary atmospheres through our own telluric atmosphere are amongst the most challenging measurements in contemporary astronomy. Telluric absorption is highly variable in time and wavelength given the numerous molecular constituents which severely limit the spectral ranges through which we can observe our targets. The key molecular species contributing to the telluric opacity (e.g. H_2O, CO_2, CH_4) are also the main molecules we search for in exoplanetary spectra. It is therefore needless to elaborate on the difficulties of observing these molecular species in an exoplanetary atmosphere through a highly variable telluric atmosphere (see chapter 6 for a more in-depth discussion).

Different detection techniques

As inarguably challenging as this is, various groups have succeeded in the detection of metal lines and complex molecules (Redfield et al., 2008; Snellen et al., 2008, 2010b; Swain et al., 2010; Bean et al., 2010). In order to obtain the desired observations, different groups have developed different techniques. These can be divided into three main categories:

1. Time-unresolved techniques: usually one or more high signal-to-noise (SNR) spectra are taken in and out of transit. Both in and out-of-transit spectra are then differenced with the additional use of a telluric model. Care needs to be taken not to over-correct and remove the exoplanetary signal (Mandell et al., 2011).

2. Time-resolved high-resolution: this is sensitive to very thin and strong emission lines where the exoplanet eclipse is followed with many consecutive exposures and the emission line is identified by the varying doppler shift of the planet as it transits (Snellen et al., 2010b).

3. Time-resolved mid-resolution: as above, the exoplanetary eclipse is followed by many consecutive exposures with a mid-resolution spectrograph making this method sensitive to broad roto-vibrational transitions. The use of telluric corrections with a synthetic model is not necessary since we obtain a normalised lightcurve per spectral channel of which the transit depths constitute the spectral signatures (Swain et al., 2010; Thatte et al., 2010; Waldmann et al., 2012b; Danielski et al., 2012; Zellem et al., 2012).

Out of these three methodologies, the first still needs to prove its worth whereas the latter two techniques have been highly successful in the past two years. For the purpose of this work we concentrate on the time-resolved, mid-resolution method to detect broad roto-vibrational transitions of molecules in the infra-red.

Very low-SNR measurements

As in the case of space-based observations, time-resolved spectroscopy allows us to obtain the whole lightcurve morphology and thus to determine stellar, planetary and orbital parameters. This is clearly preferable to obtaining a few out-of-transit spectra and a few in-transit spectra and blindly taking the difference. Therefore, the main difference between space and ground is the SNRs we can achieve. For the space-based case, typical SNRs per resolution element of the instrument are in the 10s and 100s, whereas for an individual resolution element using a stable ground-based instrument, we achieve SNRs of the order of unity. For example, the secondary eclipse measurement of HD189733b using the *IRTF*/SpeX instrument, extensively discussed in chapter 6, features SNRs of $\sim$0.7 per spectral resolution element.

Here the stochastic, white noise contribution is of the same order of magnitude as the science signal in the data. Component separation algorithms, such as the ones described in chapter 3, are severely limited by the high Gaussian noise in the data and the deconvolution reaches its limits for two reasons: 1) the non-Gaussian sources are difficult to be detected by the algorithm and signal separation is only partial, 2) fitting a systematic noise model, as was done very effective for *Hubble*/NICMOS , becomes almost impossible. The scatter in the out-of-transit time series data prevents for any meaningful convergence of the noise model. For very weak signals, this circumstance quickly becomes a fundamental limit.

Thatte et al. (2010) used principal-component-analyis (PCA) to filter the secondary eclipse feature of HD189733b from ground-based data. As explained in chapter 3, PCA is a much more simplistic approach to component separation than independent-component-analysis (ICA) based techniques and may prove more robust in very low SNR conditions. Nonetheless, as described before, PCA based approaches are not optimised for signal de-composition of such data and the lightcurve obtained can be biased by other non-Gaussianities or stochastic noise.

Rather than trying to separate systematics from astrophysical signals as in the case of space-based observations, here the problem is the amplification of the desired signal and suppression of the unwanted systematics. A common method to boost the SNR is to compute the mean of a stack of lightcurves with similar wavelengths. In accord with the Central-Limit Theorem (CLT), this binning improves the SNR by a factor of $\sqrt{N}$, where N is the number of spectral

resolution elements. However, despite the data being dominated by Gaussian noise, they also contain a significant amount of wavelength dependent non-Gaussianity that renders a simple stacking approach ineffective. In short, the systematic noise in the data disallows a simple binning procedure, whilst the high amount of Gaussian noise prevents the effective removal of non-Gaussian noise from the data using component decomposition techniques.

In the following sections we will outline how signal amplification and filtering can be achieved in Fourier and Wavelet space. The robustness of these Fourier based amplification mechanisms are tested using simulations. The application to ground-based data will then be discussed in chapter 6.

5.2 Lightcurve analysis in Fourier space

From the discussion above, it is evident that at SNRs ~ 1, the spectral emission features of a secondary eclipse event are too small to be statistically significant for an individual spectral channel. High signal to noise detections require a low spectral resolution, i.e. binning the data in λ. This can be done more efficiently in the frequency domain as explained below.

5.2.1 The sparsity of lightcurves

Let us, as first order approximation, assume the transiting lightcurve to be a trapezoidal function in the time-domain. Taking the Fourier transform of a trapezoid with N data-points, we obtain a well known sync function (Riley et al., 2002)

$$\begin{aligned} f_{trap}(t) &= 8\sqrt(2)\delta\left(\sin(1/\tau) + \frac{\sin(3/\tau)}{9} - \frac{\sin(5/\tau)}{25} - ...\right) \\ &= 8\sqrt(2)\delta\sum_{k=1,3,5...}^{N}\left(\frac{\sin(k/4\tau) + \sin(3k/4\tau)}{k^2}\right) \end{aligned} \tag{5.1}$$

where δ and τ are the signal amplitude and duration respectively and will be discussed later. We can see that the Fourier series, in equation 5.1, is rapidly converging with a rate of $1/k^2$, where k is the index of the Fourier coefficient in the series. Such a simple study showcases an important advantage of the lightcurve signal in the Fourier space compared to the more common time-domain representation. In the time domain, all N values in our time series, $x(t)$, are required to describe the signal, whilst in the frequency domain only the very first few out of the N total coefficients are needed to describe the lightcurve signal. The amplitudes of the other coefficients in the series converge asymptotically to zero. In signal processing, these are

commonly referred to as 'dense' and 'sparse' arrays or data-sets. In dense arrays, every datum contains significant information regarding the system (the time-domain case), whist in the sparse array, most values are or tend to zero, and only a few coefficients are needed to describe the full signal (the frequency-domain case). In figure 5.1 I show a Fourier transform of a secondary-eclipse lightcurve computed using the non-limb-darkened Mandel and Agol (2002) model. This is a very rapidly converging series with most of its energy contained in the first 3-4 coefficients.

We can furthermore extend the argument to limb-darkened lightcurves that exhibit a markedly rounder morphology. These are a natural extension to the trapezoidal case and it is generally true that the 'rounder' the eclipse shape, the less power is contained in Fourier coefficients above $k = 1$. Hence the series are converging even faster and are consequently sparser in Fourier space.

5.2.2 Signal amplification through self-weighted convolution

Following from the concept of sparsity, we can appreciate that the desired signal is contained only in a few coefficients. If we now consider our time-resolved spectroscopic data set, we have one time series per spectral channel i, $x_i(t)$, for M number of total resolution elements. It is important to note that the lightcurve feature is assumed to be common and near identical to all adjacent time series for $\delta\lambda \to 0$. If we now multiply the discrete-Fourier transforms (DFT) of m wavelength adjacent time series, $\mathcal{F}[x_i(t)]$, we effectively amplify the signal, whilst self-filtering the noise components. Fourier coefficients common to all channels, i.e. the lightcurve signal, are amplified whilst other varying signals effectively cancel out. We can then normalise this product by taking the mth root of the product

$$\mathcal{F}[\bar{x}(t)] = \left(\prod_{i=1}^{m} \mathcal{F}[x_i(t)] \right)^{1/m} \tag{5.2}$$

where $\mathcal{F}$ is the discrete Fourier-transform and $x_i(t)$ is the time series for the spectral channel i for m number of spectral channels in the Fourier product ($m \in \mathbb{Z}^+$). Since the input time series are always real and the Fourier transforms are Hermitian, we can take the mth root of the real-part of the final product without losing information. Note that $x_i(t)$ has been re-normalised to a zero mean to avoid windowing effects. Finally, we can take the inverse Fourier transform to obtain our filtered lightcurve, $\bar{x}(t)$

$$\bar{x}(t) = \mathcal{F}^{-1} \left[\left(\prod_{i=1}^{m} \mathcal{F}[x_i(t)] \right)^{1/m} \right] \tag{5.3}$$

In the time-domain, this operation is equivalent to a consecutive convolution of x_i with x_{i+1}, equation 5.4.

$$(x_i * x_{i+1})[n] \stackrel{def}{=} \sum_{t=1}^{n} x_i[t] x_{i+1}[n-t] \tag{5.4}$$

We can appreciate from equation 5.4 that one eclipse time series is the weighting function of the other. The consecutive repetition of this process for all remaining (i-1) time series, effectively filters the convolved time series with the weighting function that is another time series. This process has the effect of smoothing out noise components, whilst preserving the signal common to all the time series sets (Pagiatakis et al., 2007; Swain et al., 2010). The final result is the geometric mean of all time series. For a single time series, the eclipse signal may not be statistically significant, but the simultaneous presence of the eclipse in all the time series allows us to amplify the eclipse signal to a statistical significance by suppressing the noise. The *convolution theorem* states that the Fourier transform of a convolution is equivalent to the dot product of the Fourier transforms

$$\mathcal{F}(x_i * x_{i+1}) \equiv k \otimes \mathcal{F}(x_i) \otimes \mathcal{F}(x_{i+1}) \tag{5.5}$$

where $\otimes$ signifies multiplication in the Fourier space and k is a normalisation constant. It is important to note that this process is self-filtering and no prior knowledge of the signal is injected in the data.

5.3 Limitations in low SNR conditions

Given the low SNRs of the raw data, it is critical to understand how robust is the Fourier convolution described above and what SNR constitutes a lower bound. Whilst it is true that systematic noise present in all the time series, $x_i(t)$, cannot be differentiated from the astrophysical signal and it will be amplified as well, the Fourier coefficients describing the systematic noise can either be discarded or the systematics can be removed in the time-domain. The effects of Gaussian, white noise are much harder to account for and to remove and, as we will demonstrate in this section, define the shape of the retrieved signal $y(t)$.

5.3.1 Noise in the frequency domain

As explained in the previous paragraphs, the information of the lightcurve signal is contained in discrete harmonic peaks in the frequency domain. The power of these Fourier coefficients will fall off rapidly with higher orders. In a noise free case (equations 5.6 - 5.3), this does not pose a problem. However, in the presence of white or other coloured noise, the Fourier series will be

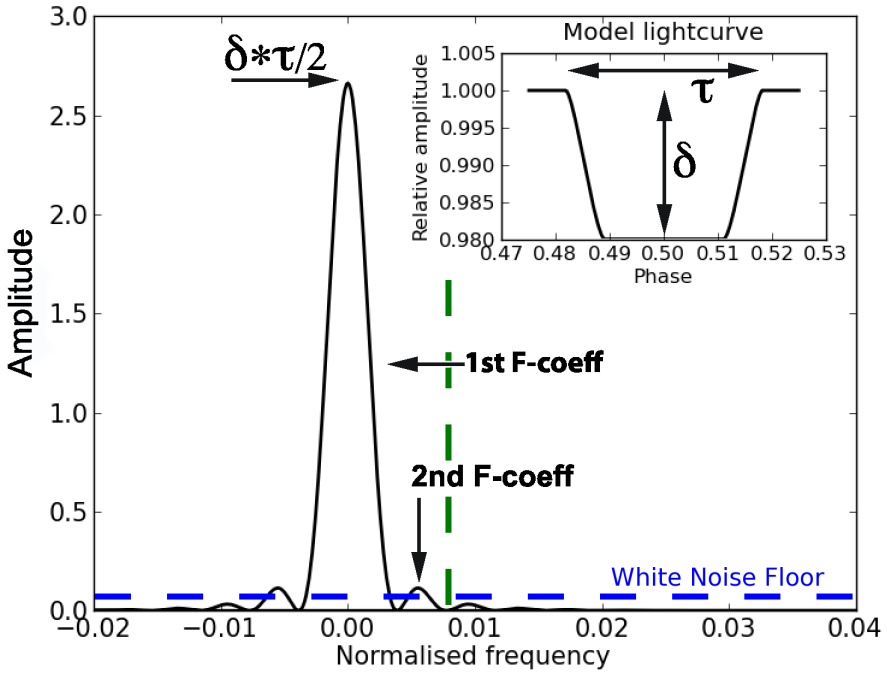


Figure 5.1: Power spectrum of a Mandel and Agol (2002) model lightcurve of HD189733b (inset). It can clearly be seen that most power of the lightcurve signal is contained in the first Fourier coefficient. Discontinuous line illustrates a constant white noise floor below which the lightcurve signal cannot be retrieved.

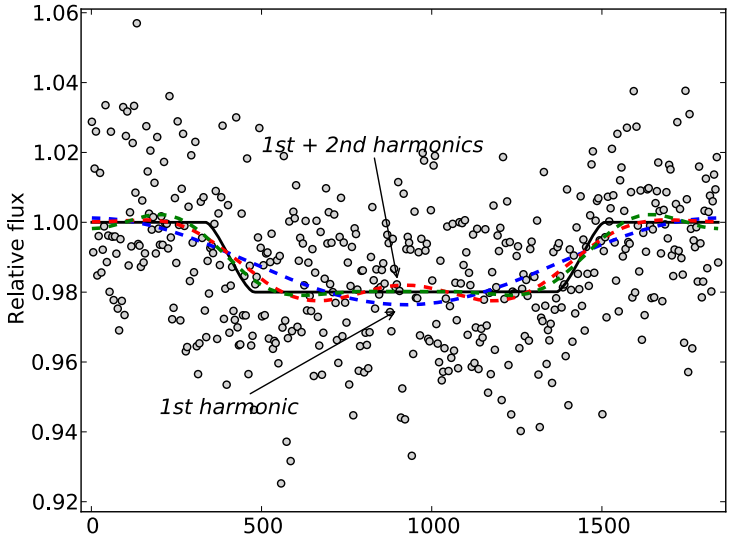


Figure 5.2: Simulated secondary eclipse lightcurve for HD189733b with SNR of 1.25. Grey dots present simulated data with white noise. Over-plotted, Mandel and Agol (2002) curve (black), first Fourier coefficient only (blue), first two coefficients (red), first three coefficients (green).

truncated and the information contained in higher-order coefficients will be lost. Adding white noise, $N_i(t)$ to our original time series, $X_i(t)$ produces equation 5.6.

$$\frac{a_0}{2} + \sum_{k=1}^{N}[a_k \cos(kt) + b_k \sin(kt) + G(t)] = \left(\prod_{i=1}^{m} \mathcal{F}[X_i(t) + G_i(t)] \right)^{1/m} \tag{5.6}$$

Since the Gaussian noise, $G_i(t)$, is constant at all frequencies, it will have a floor effect and k cannot tend to infinity. Therefore, noise truncates the series and acts as a low-pass filter in the frequency domain, leading to a loss of lightcurve information. This is a well known fact of Fourier series and can be schematically represented by figure 5.1, showing a power spectrum of the model-lightcurve with a constant white noise floor. When the lightcurve coefficients fall below this threshold, they become indistinguishable from the Gaussian noise. Schematically this is shown in figure 5.1. The vertical, dotted, green line marks the low-pass cut-off due to the noise floor. In general, a_k or $b_k \rightarrow 0$ if a_k or $b_k << G$. If G is small, coefficients to a high order of k can be retrieved and the lightcurve shape is well defined. However, should the SNR of the observation be low, i.e. < 1, we may observe a truncation effect after $k = 2$ or 3. Having a lightcurve shape defined by only very few coefficients, inherently results in distortions of the lightcurve morphology that needs to be accounted for. Figure 5.2 shows a model-lightcurve (black continuous line) over-plotted by the same lightcurve with only $k = 1$ (blue) and $k = 2$ (red) coefficients. In the $k = 1$ case, one can see that the lightcurve is only defined by a single sine or cosine curve, whilst with $k = 2$, more details are retained but distortions such as a mid-transit bump are visible.

Note that lightcurve information in higher harmonics is not lost but simply irretrievable with the SC method when the harmonics fall below the noise floor. Following this argument, the effect applies to all lightcurve observations but for high SNR data, the noise floor in the frequency domain is too low to effectively suppress the first three to four harmonics of the lightcurve signal.

5.3.2 Simulations

For practical uses, such as atmospheric spectroscopy of exoplanets, the only parameter of relevance is the transit-depth. Examining the lightcurve distortions due to noise is crucial to understand these distortions and to test whether the transit-depths can be recovered from a set of very low SNR lightcurves.

For these tests, I have generated white noise lightcurves, taking a secondary transit of HD 189733b as example. I have used equation 8 from Mandel and Agol (2002), assuming no

limb-darkening and a transit depth of $F_p/F_* = 0.02$. In this paper the SNR is defined to be the ratio of the signal strength (ie. transit depth) and a full-width-half-maximum (FWHM) of the random noise for a single measurement in time. Hence, the SNR for the full lightcurve feature, SNR_{full}, is given by $\mathrm{SNR}_{full} = \mathrm{SNR} \times \sqrt{N/2}$, since the in-transit and out-of-transit points can be averaged which reduces the noise by $\sqrt{N}$ in accord with the Central Limit Theorem. We used a Gaussian distributed (white) noise and 'synthesised' it by using a Mersenne Twister pseudo-random number generator. Each time series has 500 data points along time and the noise was added to the simulated lightcurve in the frequency domain. Figure 5.2 is an example of the lightcurves generated at SNR = 1.25. For every given SNR and model morphology, 10^5 lightcurves were generated to guarantee convergence.

For the purpose of this simulation we will restrict ourselves to the use of white noise only. This is done for the following reasons: 1) It is very difficult to produce a meaningful simulation of systematic noise, since instrumental, astrophysical and telluric noise sources vary greatly between observations, targets and instruments. 2) Red or systematic noise will not necessarily be constant like white noise, but a function of k, $G = f(k)$. Nonetheless we can assume that it affects more the high-order coefficients containing very little power rather than by the low-order ones where most of the information is contained, making the impact of systematic noise very similar (but not identical) to that of white noise.

5.3.3 Altered lightcurve morphology

Depending on the number of Fourier coefficients that can be extracted from the constant noise floor, the original lightcurve can be recovered with more or less fidelity.

Monte Carlo simulations show that the transit-depth recovery of lightcurves with SNR < 0.15 becomes problematic. Figures 5.5 & 5.8, show a well defined discontinuity and a sudden increase in transit depth recovered for simulations with SNRs < 0.15. This behaviour can be attributed to the noise suppressing too many coefficients of the Fourier series so that the lightcurve information becomes irretrievable. We refer to this as the ultra-low-signal-to-noise regime (ULSNR).

For data with a SNR > 0.15, we see a very different picture. Figure 5.4 shows the self-coherence outcome with the input signal having a SNR of 1.1. Here, the transit depth can be recovered. The upward bump at the centre of the distribution is due to only the first two coefficients, $k = 2$, of the lightcurve signal being recovered and not the third. Here, the third harmonic lies below the noise-floor and has been suppressed. Figure 5.2 (red-curve) is an example of the expected eclipse morphology containing the first two harmonics solely. From hereon this case will be referred to as medium-SNR (MSNR).

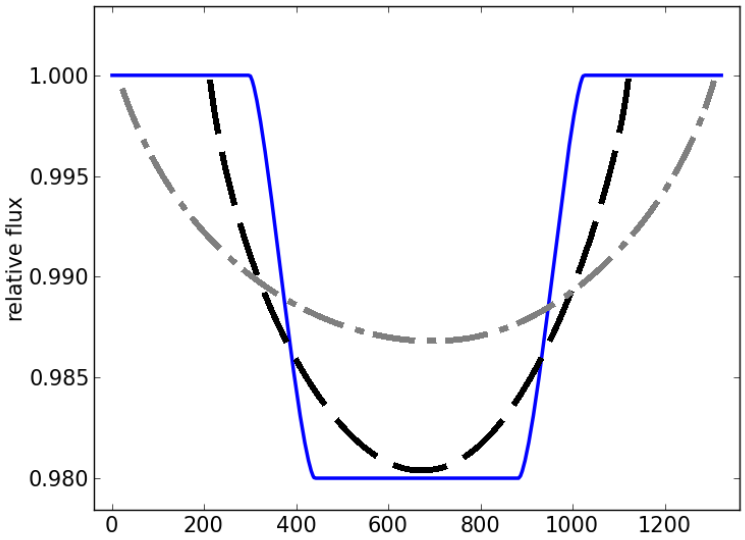


Figure 5.3: showing a symmetric model-lightcurve (blue solid line). The black (– –) and grey (– - –) dotted lines schematically sketch different shapes of the first harmonic for different out-of-transit durations. If only few harmonics are recovered, the shape of the first harmonic becomes important. With excessive out-of-transit (OOT) data, the first sine/cosine curve of the Fourier series can be understood to be stretched out (black – – line) compared to the case of less OOT data (grey – - – line). This effect can result in a loss of transit depth retrieved.

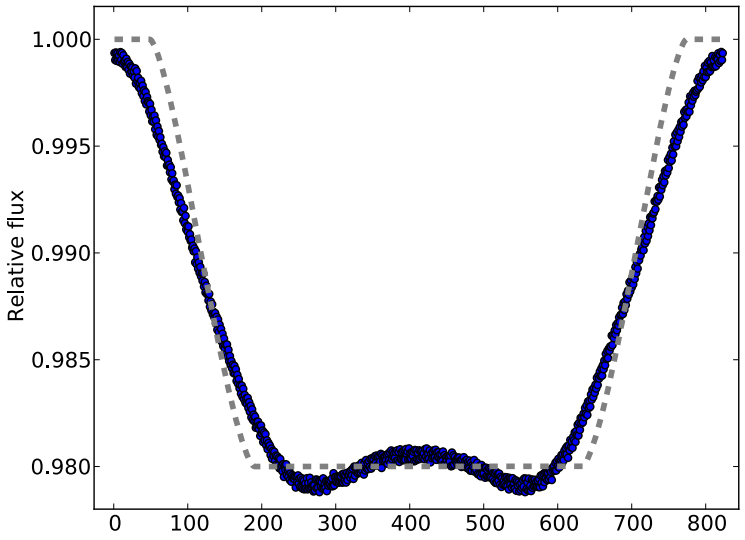


Figure 5.4: The self-coherence result of 10^5 symmetric lightcurves (grey dotted line) with white noise added at SNR = 1.1. Enough power is contained in the first two Fourier coefficients to lift them above the noise floor. This central bump at mid-transit is indicative of two Fourier coefficients being present as shown by the red curve in figure 5.2.

Since we can often only recover the first two harmonics of the lightcurve, we find that the retrieved eclipse shape does not only depend on signal-to-noise, as discussed above, but also on the data-set's own morphology. The result of the SC-method is also dependent on the length of the out-of-transit data gathered and the duration of mid-transit. Furthermore, the morphology of the eclipse signal recovered depends on whether the time series data-set is symmetric or asymmetric. In a symmetric case, the pre and post-transit data is of the same duration, whilst for the asymmetric case it is not. The differences can be attributed to the different Fourier transform sampling for mathematically even and not-even functions. These differences are rather theoretical since we can always set up our observations as mid-point symmetric or asymmetric. Using the symmetric case, we will briefly discuss the effects of out-of-transit sampling lengths to the Fourier product in equation 5.2.

The symmetric case

The differences in behaviour of the Fourier convolution for symmetric and asymmetric data sets can be significant. For mid-point symmetric signals, such as the trapezoidal shaped lightcurve, we observe a redundancy of 50%. Such systems are Hermitian and no information is lost if the signal is under-sampled by a factor of two. These signals are known as *even* functions, which can be described fully by either the sine or cosine terms of the Fourier transform. In other words, if a signal is Hermitian, most fast-Fourier transform (FFT) algorithms will discard half of the available information, effectively reducing the number of retrieved coefficients by a factor of two. As a consequence, the effect of coefficients being suppressed by white noise is much more pronounced than for the asymmetric case. This allows us to study the most extreme case of morphological distortions of the signal due to white noise.

If only the two first, even harmonics ($a_1 + a_2$) are recovered in the symmetric case, the flat out-of-transit parts become poorly defined since the first harmonic of the distribution will always extend from the beginning of the time series till the end. This is schematically illustrated in figure 5.3. The black, dotted line represents the schematic shape of the first harmonic with a large out-of-transit (OOT) part whilst the grey, dotted line shows how the same harmonic would look like if the data set had with less OOT points. The stretching out of the first harmonic by excessive out-of-transit data will result in a loss of transit-depth recovered. This loss is a function of both OOT data present and SNR and can be shown for a given number of OOT points to be a well behaved logarithmic function of SNR for the MSNR regime. In other words, as OOT $\rightarrow \infty$, the eclipse-signal approaches a delta-Dirac function, $\delta(t)$, and will consequently become under-sampled. Figure 5.5 shows the parameter space of transit-depth recovered (in percent)

against SNR from 0.001 to 1.25 and total OOT points of 100 to 2600 symmetrically distributed before and after transit. Each node on the grid is a Monte Carlo simulation of 10^5 lightcurves as described in section 5.3.2. As discussed above, at SNRs < 0.15, the Fourier coefficients cannot confidently be recovered. On the other hand the surface for the MSNR regime seems relatively well behaved.

Continuing this exploration of the parameter space, figure 5.6 shows the transit-depth (held constant at 0.02) recovered as a function of in-transit length and out-of-transit length. This case is a little more theoretical since for a known system the transit-duration is well known. The case of in-transit-data tending to zero is equivalent to OOT $\rightarrow \infty$, resulting in an under-sampling of the eclipse in the Fourier transform and a potential loss of eclipse-depth recovered which needs to be accounted for.

The asymmetric case

In the case of an asymmetric time series, ie. more out-of-transit data at either one side of the eclipse, the time series is not a reflection of itself about the mid-point any longer. Therefore the limitations of the *even* functions are no longer valid and for the same noise induced low-pass cut-off, we retrieve all, a_k & b_k, Fourier coefficients. This results in a twice better retrieval of the true eclipse signal for the MSNR case. In order to simulate the eclipse retrieval for the asymmetric case, we have taken the same model as used for the symmetric case but increased the post-egress data by 200 data points in length. Figure 5.7 shows two lightcurves generated with the same input model SNRs of 0.1 (blue) and 1.0 (green). It can clearly be seen that, despite the out-of-transit part of the lightcurve being better defined now, the distortions to the eclipse curve recovered are significant. It is generally true that the higher the SNR of the data, the better the SC result traces the true feature. Taking the means of the out-of-transit and in-transit data in order to measure the transit-depth recovered (see figure 5.8) results in a very similar picture as it is the case for the symmetric data set (figure 5.5) with transit-depths fully recovered down to a SNR ~ 0.4. The differences between the symmetric and asymmetric surface plots can be attributed to the difficulty in accurately measuring the out-of-transit level for the symmetric case and the improved information content of the asymmetric Fourier-product.

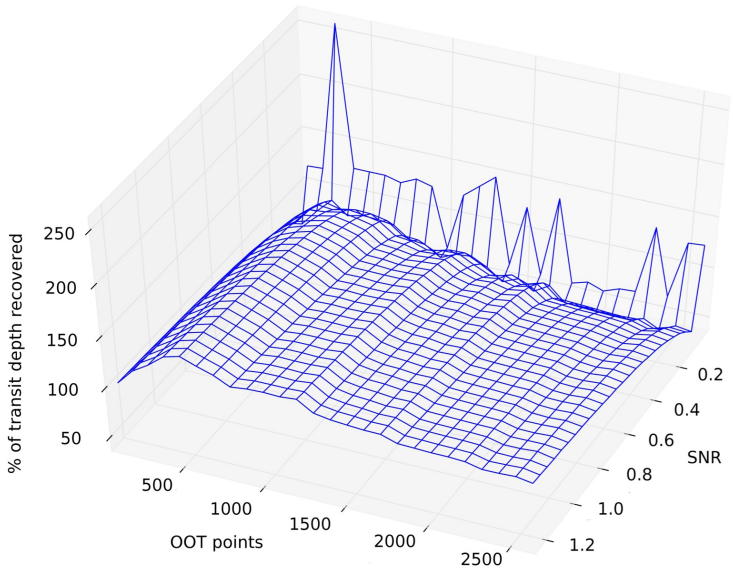


Figure 5.5: Transit depth recovered (in %) as a function of out-of-transit (OOT) data and SNR, where the OOT points are symmetrically distributed for a transit model of HD 189733b with a transit depth of 0.02. It can be seen that the transit-depth retrieved is a well-behaved function of OOT and SNR for SNRs > 0.15. Below SNR ~ 0.15, transit-depth retrieval becomes problematic due to the suppression of all coefficients, but a_0, by noise.

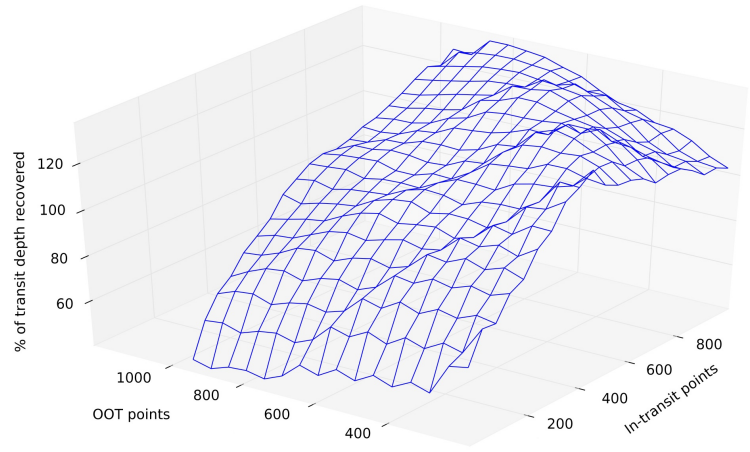


Figure 5.6: Transit depth recovered (in %) as a function of out-of-transit (OOT) data and in-transit points, where the OOT points are symmetrically distributed and 450 in-transit points representing the real in-transit duration for HD 189733b. The SNR was fixed to 0.5.

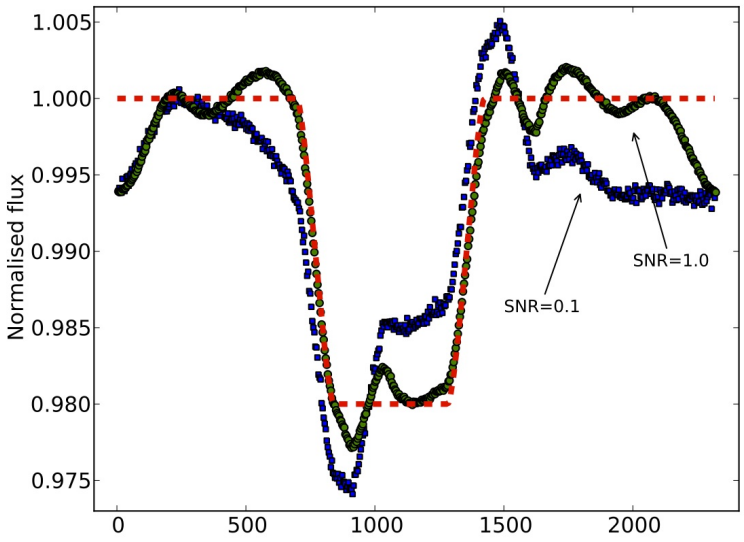


Figure 5.7: The self-coherence results for the input model shown by the red-dashed line, with SNRs of 0.1 (blue squares) and 1.0 (green circles). The out-of-transit data is asymmetrically distributed with more data post-egress. In the asymmetric case, the function is not mathematically *even*, resulting in a 2x improvement in the retrieval of the eclipse shape for the same SNR conditions when compared to the symmetric case in fig. 5.4. The progressive convergence to the true eclipse-shape can be seen with the SNR = 1.0 case being better behaved than the SNR = 0.1 case.

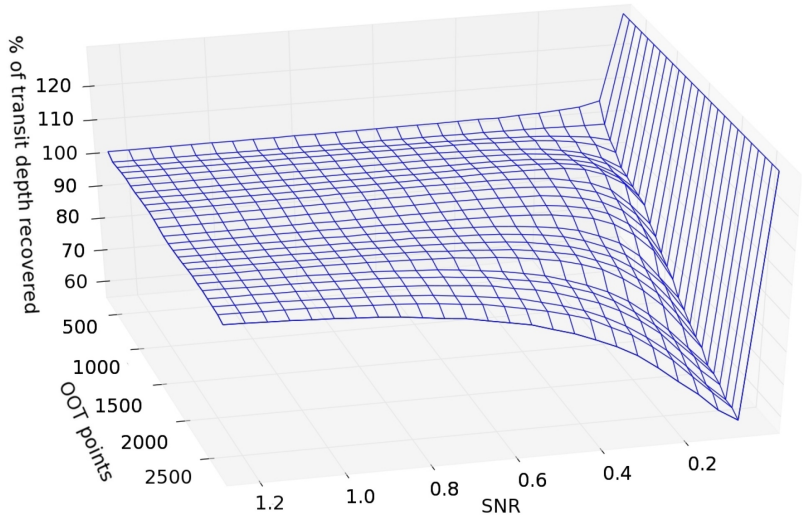


Figure 5.8: Transit depth recovered (in %) as a function of out-of-transit (OOT) data and SNR, where the OOT points are asymmetrically distributed for a transit model of HD 189733b with a transit depth of 0.02. The SNR > 0.15 surface is much better behaved than in the symmetric case but here too, transit-depth retrieval at SNRs below $\sim$0.15 become problematic.

5.4 Measuring the resulting lightcurve

Given this specific behaviour of Fourier convolutions in the presence of significant amount of white noise, we need to understand how to retrieve the transit-depth parameter from our measurements.

5.4.1 Measuring the lightcurve in the time-domain

Following from the previous sections, the optimal strategy is: 1) The data-set should be asymmetric, 2) the SNR per time series should not fall below ~ 0.2, 3) the self-coherence result should not be directly model-fitted, due to the well-understood, yet significant distortions present.

Fitting the standard lightcurve models (eg. Mandel and Agol, 2002; Seager and Mallén-Ornelas, 2003) to the outcome of the Fourier convolution will only be a crude approximation to the real value, due to the distortions present. This issue can easily be addressed by applying to the model the same procedure that has been applied to the data. Figure 5.9 shows the resulting SC distribution of a simulated, asymmetric, white noise data set for 100 lightcurves (blue squares) and 10^5 lightcurves (red circles) as input to the SC-method. In this figure is clear that the 100 lightcurve case has not fully converged to the real expectation value but is a good tracer of it. In a spectroscopic data-set one can assume to have tens to hundreds of lightcurves

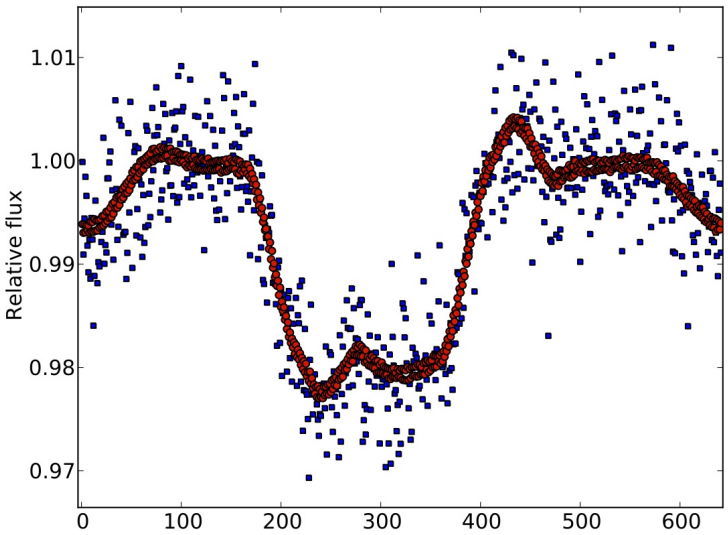


Figure 5.9: The self-coherence result for 100 time series (blue squares) and 10^5 time series (red dots). All have transit-depths of 0.02, asymmetric out-of-transit data and white noise added at SNR = 0.5. The 100 spectral channel curve (blue) serves as a typical example of what to expect for real data. For real data, the eclipse-depth can be retrieved more easily in the asymmetric case.

as input.

Using simulations as described in section 5.3.2, we can iteratively fit the self-coherence result of the real data. We have tested this approach by simulating white-noise data-sets ranging from SNRs of 0.2 to 1 and transit depths of $F_p/F_* = 0.001$ to 0.1. We used a downhill, simplex algorithm (Nelder and Mead, 1965) to iteratively converge the model described in section 5.3.2 to the 'real' data set. Here the SNR and transit-depth were left as free parameters. We find that the mean error on the transit-depth retrieved is $\sim$3% of the original eclipse-depth. For the sake of computational speed, I used 5000 lightcurves in my Monte Carlo simulations. By increasing the number of randomly generated lightcurves for the model, the error on the transit-depth recovered will decrease further. The final uncertainty of the fit, for a white noise case, will inevitably depend on the level of convergence of one's real-data to the expectation value.

The quick and dirty way: taking means

For reasons of computational speed and analytic purpose, the transit depths for figures 5.5,5.6 & 5.8, have been computed by simply taking the mean of the out-of-transit data and the in-transit data and differencing those. For SNRs $>\sim 0.4$, this method is very stable, especially for asymmetric data sets. However, when using means, the loss in transit-depth recovered at SNRs < 0.4 needs to be corrected for.

5.4.2 Measuring the lightcurve in the frequency domain

As previously mentioned, the noise residuals are in part generated during the conversion of the data from the frequency domain to the time domain, in part are due to systematics. We can remove some of these residuals, by measuring the eclipse depth directly in the frequency domain, assuming that most systematic noise is found at different frequencies to the eclipse signal.

As shown in figure 5.1, where we took the Fourier transform of the secondary eclipse model shown in the insert, the frequency spectrum is centred on the first Fourier coefficient. It is clear that most of the power is contained in the first Fourier coefficient and the series rapidly converges asymptotically to zero after the third coefficient. Following Fourier series properties, the modulus of the coefficient amplitude, $|A|$, of the coefficients in equation 5.1 is directly proportional to the transit depth δ and the transit duration τ, where $\tau = t_{1-4}/t_s$ and t_{1-4} is the transit duration from the first to fourth contact point and t_s is the sampling rate (i.e. exposure time + overheads).

$$|A|_{trapez} = \frac{\tau\delta}{2} \sum_{k=1,3,5...}^{\infty} \frac{1}{k^2} \tag{5.7}$$

The amplitude of the Fourier coefficients above $k = 1$ decreases by $1/k^2$ for a trapezoidal-shape function and it is an even faster converging series for limb-darkened lightcurve shapes. From equation 5.7 we see that for the first Fourier coefficient, $k = 1$, the relationship between the transit depth, δ, and the Fourier coeffcient amplitude, $|A|$, is simply given by $|A_{k=1}| = (\tau/2)\delta$. From the analytical arguments presented above, we know that τ is the transit duration (in units of number of observed spectra). We checked the consistency of the theory with the data, by calculating the value of τ numerically. To calculate τ we produced secondary eclipse curves with the transit duration and sampling rate of 10 seconds per exposure (Mandel and Agol, 2002, equation 8). We generated 300 curves with transit depths (τ) ranging from 0.0001 to 0.1 and measured the corresponding amplitude ($|A_{k=1}|$). Here, the derivative, $d(|2A|)/d\delta$ gives us the value of τ. We find $\tau = 116$ in-eclipse measurements, which agrees with the number of in-transit points used as input.

N spectra are usually obtained at a constant sampling interval of t_s, giving us a sampling rate of $R = 1/t_s$ in the frequency domain. For a complete representation of the data, the sampling rate is equal to the Nyquist rate, $R = 2B$, where B is the spectral bandwidth of the Fourier transform. The total number of Fourier coefficients, K, is then given by $K = 2BN$. It follows that the resolution in the frequency domain is determined by $\Delta f = 1/N$. In other words, the more measurements are available the more Fourier coefficients can be extracted to describe the data and consequently the frequency range covered by each coefficient is smaller for a fixed sampling rate.

The fact that Δf is finite ($\Delta f \to 0$ for infinitely sampled data-sets), means that the first Fourier coefficient can be contaminated by remaining noise signals very similar in frequency. To estimate the error bar on this contamination, we varied the out of transit (oot) length N_{oot} by 50% and calculated the resulting spectrum for each Δf. The error is then estimated as the standard deviation to the mean of all computed spectra.

5.5 Dampening Gaussian noise with Wavelets

In the previous sections we extensively discussed the impact of noise on the Fourier product in equation 5.2. Given the importance of minimising white noise contributions it is useful to pre-filter the data before attempting the convolution. Based on the de-noising approach proposed by Thatte et al. (2010), I have opted for a wavelet filtering of the individual time series using the 'Wavelet Toolbox' in *MATLAB*. There are clear advantages to wavelet de-noising compared to simple smoothing algorithms: 1) wavelet de-composition is a non-parametric algorithm and hence does not assume prior information on the signal or noise properties, making it an easy to use and

objective de-noising routine; 2) contrary to smoothing algorithms (e.g. kernel regression) high and low signal frequencies are retained; 3) temporal phase information of the signal is preserved during the de- and re-construction of the signal. This allows for an optimal white and $1/f$ noise reduction at varying frequency pass-bands (Carter and Winn, 2009; Donoho, 1995; Percival and Walden, 2000).

In the next paragraphs, I will briefly introduce the concepts of the wavelet transform and multi-resolution analysis and outline how it is used to selectively minimise white noise in time series data. An application to real ground-based data follows in chapter 6.

5.5.1 Definition of Wavelets

Similar to a Fourier Transform (FT), the Wavelet Transform (WT) decomposes a given time series signal into its frequency components. Where the FT uses sine and cosine functions that extend over the full range of the data, the WT uses highly localised impulses. These impulses or 'wavelets' scan through the time series and much like a tuning fork to an instrument, 'resonate' with localised features in the time series that are akin to the wavelet's shape and scaling. The individual wavelet basis functions are derived from a single mother wavelet $\psi(t)$ through translation and dilation of the mother wavelet (Percival and Walden, 2000). Different wavelets exist with different analytical properties, here we use the orthogonal basis wavelets of the Daubechies (db) family (Daubechies, 1988). Examples of the Daubechies wavelet family for different numbers of vanishing moments[1] are given in figure 5.11 (red lines). The wavelet analogue to the Fourier transform of the times series $x(t)$ is given by:

$$c_{\tau,s} = \int_{\mathbb{R}} x(t)\psi_{\tau,s}(t)\mathrm{d}t \tag{5.8}$$

The scaling and translation of the mother wavelet for the continuous wavelet transform (CWT) is given by

$$\psi_{\tau,s} = \frac{1}{\sqrt{2}}\psi\left(\frac{t-\tau}{s}\right) \tag{5.9}$$

where τ and s are the translation and scaling of the mother wavelet respectively. The wavelet base is orthogonal and we can hence reconstruct the data by taking the sum of the product of all coefficients for a given scale and translation, $c_{\tau,s}$, with the respectively scaled and translated mother wavelet

[1] Vanishing moments (VM) are statistical moments of the wavelet series that tend to zero, e.g. a wavelet with one VM has zero mean. The number of VMs determines the order of a polynomial a wavelet can describe exactly, i.e. a db4 wavelet contains two VMs and can perfectly describe all polynomials up to second order, or a db8 can describe a fourth-oder polynomial.

$$x(t) = \sum_{s \in \mathbb{Z}} \sum_{\tau \in \mathbb{Z}} c_{\tau,s} \psi_{\tau,s}(t) \tag{5.10}$$

For a more in-depth definition of wavelets and their respective properties I refer the interested reader to Daubechies (1992) and Percival and Walden (2000).

5.5.2 Multi-resolution analysis

The above equations apply to the CWT case. The wavelet coefficients describe the correlation between the wavelet at varying scales (or frequencies). These can be calculated by changing the scale of the wavelet, i.e. the analysis window. We can hence speak of a multi-resolution decomposition, where each scaling of the mother wavelet denotes a given resolution. Here, the analogy to the Fourier Transform would be band-pass filters of varying size. Most times it is more sensible to exploit the discrete nature of the data and to define the discrete wavelet transform (DWT). The DWT is significantly easier to implement and faster to compute. Similarly to the continuous case, in the DWT we have a 'mother' wavelet and a scaling function, also known as 'father' wavelet. Here, the 'mother' wavelet is denoted by $h[n]$ and the 'father' by $g[n]$ (Daubechies, 1992; Percival and Walden, 2000; Press et al., 2007). Examples of these wavelet forms for Daubechies family of wavelets are shown in figure 5.11. In this figure, the 'mother' wavelet, $h[n]$, is shown in red and the scaling function, $g[n]$ is shown in blue. *It is important to note that unlike in the CWT case, where the 'mother' wavelet itself is scaled to represent different frequencies in the data, this is not the case in the DWT. In the DWT, in analogy with the Fourier Transform, $h[n]$ and $g[n]$ can be thought of as high-pass and low-pass frequency filters respectively. Different scalings are then achieved by progressively 'down-sampling' the data as explained later.*

The DWT is best understood by following the individual steps of the algorithm that computes the transform:

1. The observed, discrete time series, $x[n]$, is convolved with the 'mother' wavelet $h[n]$:

$$cD_s[n] = (x * h)[n] = \sum_{k=-\infty}^{\infty} x[k] \cdot h[n-k]. \tag{5.11}$$

   Here cD_s represent the 'mother' wavelet coefficients for a given scale s. As mentioned earlier, the 'mother' wavelet, $h[n]$ acts as a high-pass filter, sensitive to the high frequencies or details of the time series. We hence refer to the coefficients of $h[n]$ as *detail coefficients*.

2. The next step is to convolve the same time series, $x[n]$, with the scaling function or 'father'

wavelet:

$$cA_s[n] = (x * g)[n] = \sum_{k=-\infty}^{\infty} x[k] \cdot g[n-k]. \tag{5.12}$$

As opposed to the 'mother' wavelet, the 'father' wavelet acts as a low-pass filter of the time series and its coefficients can be viewed as a moving average of the underlying trend of $x[n]$. We hence refer to its coefficients as *average coefficients* and denote them with cA_s. Furthermore, the low-pass filter $g[n]$ is related to the high-pass filter by

$$g[L-1-n] \;=\; (-1)^n \;\cdot\; h[n] \tag{5.13}$$

where L is the filter length and corresponds to the number of points in the time series $x[n]$.

3. We now have two sets of time series, a low-pass filtered, moving-average time series, cA_s, and a high-pass filtered time series, cD_s. We record the cD_s coefficients and proceed with our analysis of the average coefficients, cA_s. Since half of the frequencies in cA_s (namely the high-pass ones) have been removed by equation 5.12, the Nyquist theorem tells us that we are oversampled by a factor of two. We can hence remove every second coefficient in cA_s without losing information. This operation is termed 'down-sampling' and abbreviated by $\downarrow 2$, leaving us with $N/2$ coefficients to describe cA_s. Similar applies to the detailed coefficients cD_s. The detail and average coefficients are hence given by:

$$cA_s[k] = \sum_n x[n] \cdot g[2k-n] \tag{5.14}$$

$$cD_s[k] = \sum_n x[n] \cdot h[2k-n] \tag{5.15}$$

4. The Nyquist down-sampling introduces the concept of scaling or multiple resolutions. If we now repeat steps 1-3 on the down-sampled average coefficients, cA_s, we obtain a new set of coefficients (cA_{s+1} and cD_{s+1}) on a scale that is double the size of the previous decomposition. This is illustrated in figure 5.10 for a three level decomposition.

For a given scale, s, the data can now be reconstructed by reversing the above process:

$$x_s[n] = (cA_{s=\mathcal{S}}[k] \cdot g[-n+2k]) + \sum_s \sum_{k=-\infty}^{\infty} (cD_s \cdot h[-n+2k]) \tag{5.16}$$

where $\mathcal{S}$ are the total number of decompositions.

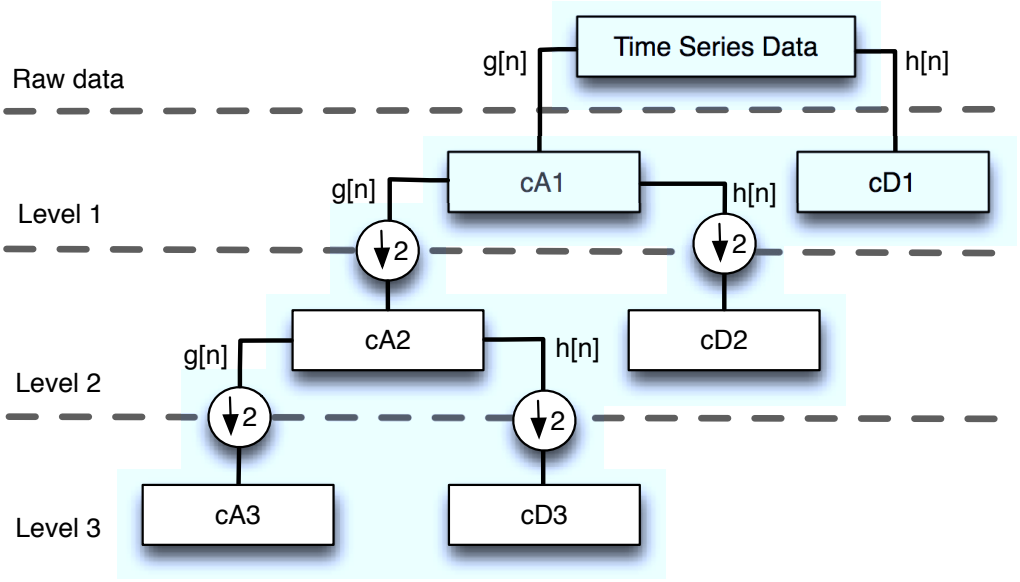


Figure 5.10: *Outline of multi-resolution wavelet decomposition down to the 3rd decomposition level.*

The quadrature mirror filter

A very fast and simple implementation of the DWT for multi-resolution decomposition is by constructing the quadrature mirror transformation matrix. For a Daubechies-4 wavelet we obtain four coefficients comprising $h[n]$ and similarly $g[n]$. Rather than convolving the time series, $x[n]$ with both filters separately and then down-sample, we can also construct a matrix where each odd row contains $h[n]$ and each even row contains $g[n]$ coefficients. This automatically down-samples the data to the new resolution $s+1$. Such a matrix is called a 'quadrature mirror filter' (QMF) and equation 5.17 is an example of such (Press et al., 2007).

$$\begin{bmatrix}
c_0 & c_1 & c_2 & c_3 & & & & & & & & \\
c_3 & -c_2 & c_1 & -c_0 & & & & & & & & \\
 & & c_0 & c_1 & c_2 & c_3 & & & & & & \\
 & & c_3 & -c_2 & c_1 & -c_0 & & & & & & \\
\vdots & \vdots & & & & & \ddots & & & & & \\
 & & & & & & & c_0 & c_1 & c_2 & c_3 & \\
 & & & & & & & c_3 & -c_2 & c_1 & -c_0 & \\
c_2 & c_3 & & & & & & & & c_0 & c_1 & \\
c_1 & -c_0 & & & & & & & & c_3 & -c_2 &
\end{bmatrix} \tag{5.17}$$

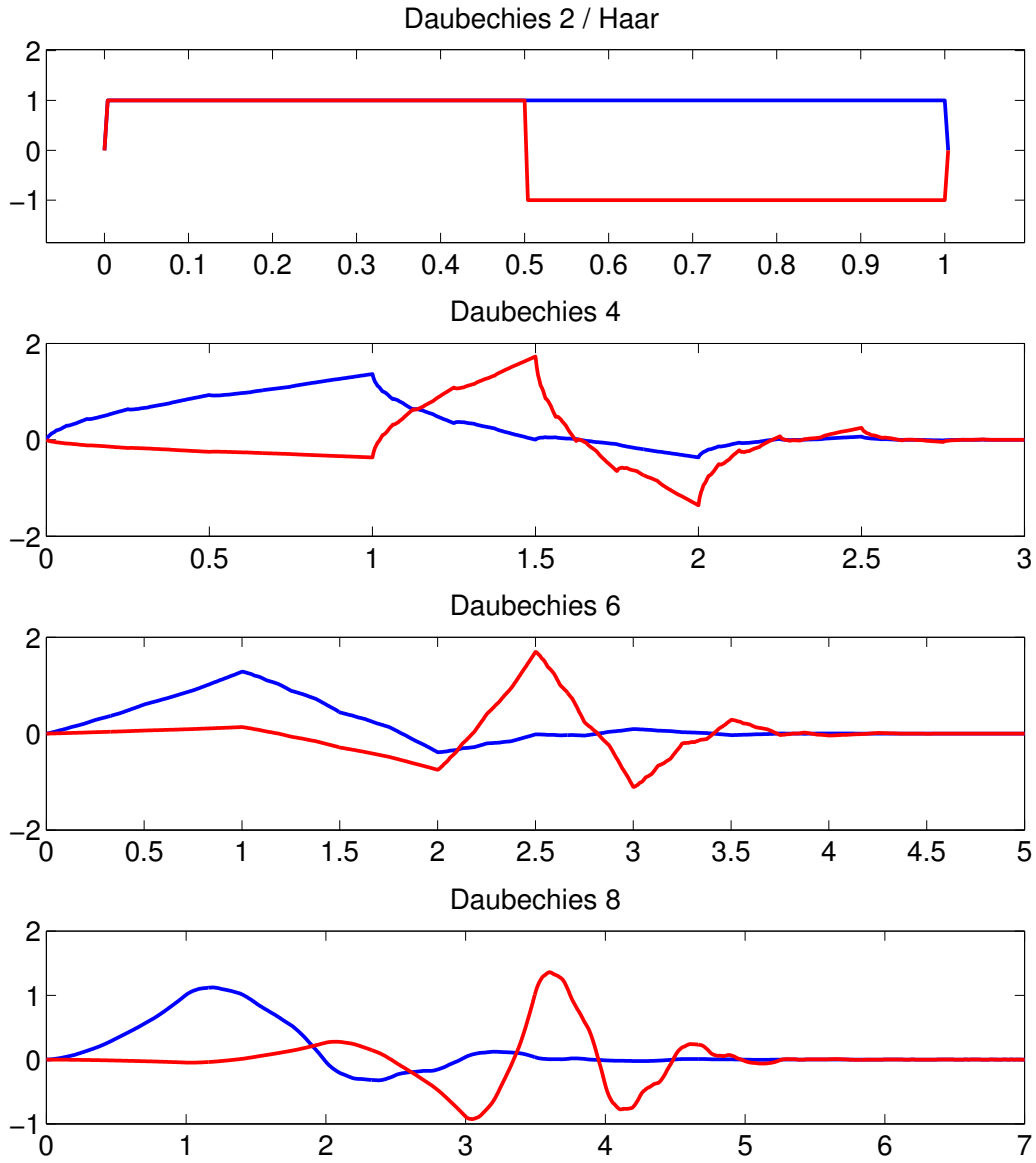


Figure 5.11: showing four wavelets of the Daubechies wavelet family. Top: Daubechies 2 wavelet, also known as Haar wavelet, features two vanishing moments and is the simplest wavelet form as it represents a top-hat impulse function. The red curve (in all plots) shows the mother-wavelet, whilst the blue curve shows the scaling function, also known as the father-wavelet. Below are examples of mother and father wavelets featuring 4, 6 and 8 vanishing moments, able to describe higher order polynomial shapes more precisely.

where the empty spaces denote zeros. To obtain a DWT using this QMF, we multiply the QMF with the column vector containing the time-series on the right. Note the circular behaviour of the matrix at the bottom, where the wavelet coefficients wrap around to the beginning. This has important consequences as it indicates that the DWT wraps around the data and the transform at the end of the time-series is sensitive to data at the beginning of the time series. This effect can be avoided by adding sufficient zero-valued points to the time series at its beginning and end. This process is also known as 'zero-padding'.

5.5.3 Denoising the data by thresholding

We can now use the concept of multi-resolution analysis to selectively suppress the white noise in the time series data. Let us assume the following data model

$$x[n] = s_a[n] + \sigma_{e[n]} \tag{5.18}$$

where $x[n]$ is our discrete input time series, $s_a[n]$ our desired astrophysical signal and $\sigma_{e[n]}$ the noise contribution from a Gaussian process $e[n]$. The key objective is to suppress $\sigma_{e[n]}$. This can be done by 'thresholding' the detailed coefficients, cD_s. The de-noising sequence using soft coefficient thresholding (Donoho, 1995) is:

1. Perform a multi-resolution decomposition to a given level, $\mathcal{S}$.

2. The detailed coefficients will contain most of the Gaussian noise and cD_1 will almost entirely consist of high-frequency Gaussian noise. Following the thresholding algorithm by Donoho (1995), the coefficients for each resolution are now thresholded in two steps:

   1) $|cD_s[k]| = 0$ if $cD_s[k] < \mathcal{T}$, where

   $$\mathcal{T} = \gamma\sigma\sqrt{2\log(N)/N} \tag{5.19}$$

   and γ is a constant;

   2) if $|cD_s[k]| > \mathcal{T}$ then

   $$cD_s[k] = \mathrm{sgn}(cD_s[k])(|cD_s[k] - \mathcal{T}). \tag{5.20}$$

3. The time series $x[n]$ is reconstructed using the thresholded coefficients.

This process significantly reduces the white noise contribution of the time series data and effectively improves the Fourier convolution process described in section 5.2.2.

5.6 Conclusion

In this chapter I introduced and discussed the framework for lightcurve signal amplification in the frequency domain and suppression of Gaussian noise in the wavelet domain. Such approaches are necessary for the analysis of ground-based, time-resolved spectroscopic data of extrasolar planets, given that signal-to-noise levels are too low for more standard de-trending methods. This chapter can be summarised by the following points:

- Due to the sparsity of the lightcurve signal in the frequency domain, we can effectively self-filter the time-resolved spectroscopic data by convolving wavelength adjacent channels together.
- I show that the constant white noise floor in the frequency domain effectively suppresses high-order Fourier coefficients of the lightcurve signal. This leads to a truncation of the Fourier series and results in distortions of the lightcurve morphology.
- Using simulations, I find that the truncation of the Fourier series directly impacts the retrievability of the correct transit-depth of the lightcurve.
- I find that below the threshold of SNR ~ 0.15 per lightcurve, the suppression of eclipse-information by noise becomes too dominant to recover the original signal, which constitutes a fundamental lower limit.
- The lightcurve depth can be retrieved by measuring the amplitude of the first coefficient in the lightcurve's Fourier series. Using this method, we can also estimate the contribution of systematic noise to the transit-depth parameter.
- Due to the strong impact of Gaussian noise in the Fourier analysis, it is important to reduce its contribution in the time series data. This can be achieved using thresholding of detailed wavelet coefficients.

Chapter 6

Non-LTE methane emissions from hot-Jupiter HD189733b

6.1 Introduction

In the previous chapter I introduced and discussed a possible method for data-analysis of very low signal-to-noise (SNR) exoplanetary spectroscopy data obtained using ground-based instruments. These techniques are rooted in the frequency domain due to the preferential properties of the astrophysical lightcurve signal in Fourier space. In this chapter I will use these Fourier and wavelet based methods to analyse four secondary eclipse events of the hot-Jupiter HD189733b.

HD189733b is one of the most studied extrasolar planets known. First discovered by Bouchy et al. (2005), the hot-Jupiter orbits its bright (V=7.7), late K-type host star every ∼2.219 days at a distance of ∼0.031 AU (astronomical units). This planet features a highly inflated atmosphere at an average effective temperature of ∼1200 K. Since 2007, several groups have successfully analysed its atmosphere in various wavelength bands ranging from the UV to the mid-IR. In the visible part of the spectrum, metallic lines such as sodium have been detected using a ground-based instrument (Redfield et al., 2008) whilst in the IR, the roto-vibrational transitions of H_2O observed by *Spitzer*/IRAC, suggested the presence of water in the atmosphere (Tinetti et al., 2007). Using the *Hubble*/NICMOS and *Spitzer*/IRS instruments, the discovery of CH_4, CO_2 and CO quickly followed (Swain et al., 2008c, 2009b), with other contributions to the field by (Grillmair et al., 2008; Pont et al., 2008; Charbonneau et al., 2008; Knutson et al., 2007a; Thatte et al., 2010; Sing et al., 2011; Gibson et al., 2012). Due to the planet's strong day-side emission, it has even been possible to obtain the first spatially resolved temperature maps of an extrasolar planet by 'slicing' the planet as it enters and exists the secondary eclipse (Knutson et al., 2007a;

Majeau et al., 2012).

The secondary eclipse is not only good for 'slicing' but grants us valuable information on the emission signatures of atmospheric molecules of highly irradiated exoplanets. These spectral signatures are best observed using space-based observatories since the measurement of the secondary eclipse depth requires a very stringent photometric precision of $<10^{-4}$ in flux. At the end of the *Spitzer* cold-phase, increased efforts need to be undertaken to ensure spectroscopic capabilities using ground-based observatories (see chapter 5). As inarguably difficult as this is, various groups have succeeded in the detection of metal lines and complex molecules (Bean et al., 2010; Redfield et al., 2008; Snellen et al., 2008, 2010b; Swain et al., 2010) from the ground. Using ground-based facilities has two unique advantages over space-based observatories: 1) observations are more readily repeatable and 2) ground-based spectrographs may cover spectral regions that were inaccessible to space-based instrumentation.

Having exploited the spectral ranges of *Hubble* and *Spitzer* , Swain et al. (2010), hereafter S10, used the SpeX instrument on the NASA-*IRTF* telescope on Mauna Kea, to observe the secondary eclipse of HD189733b simultaneously in the K- and L-bands. Figure 6.1, illustrates the data coverage of HD189733b in terms of wavelengths observed. The blue and green points represent *Spitzer* IRAC and IRS data respectively whilst the red-points show *Hubble*/NICMOS results. Whilst the K-band (2.1-2.5 μm), is well observed with NICMOS and the wavelengths above 5μm are well covered by *Spitzer* , the L-band region (3.0-4.5μm) was uncharted territory. The results of the study by S10 showed a good agreement of the *IRTF*/SpeX data (in black) with *Hubble*/NICMOS in the K-band whilst a very strong and highly unexpected peak was observed at $\sim$3.25μm. This peak corresponds to the ν_3 band emission of methane and was interpreted as non-LTE fluorescence.

Whilst not uncommon in our own solar system (see section 6.6), this was the first detection of a non-LTE atmospheric process on an extrasolar planet and triggered significant debate in the community. The results by Swain et al. (2010), from here S10, were based on a single night of observation and featured a very low spectral resolution covering the L-band feature, leading to a justifiably critical view in the eyes of the community.

In Waldmann et al. (2012b), we re-analyse the original S10 data as well as three additional planetary eclipses observed with the *IRTF*/SpeX instrument. One eclipse, in particular, was obtained with a reference star in the slit. We used the time-resolved mid-resolution method outlined in chapter 5 and an improved data-preprocessing routine over the one presented by S10. The additional data in conjunction with the more advanced techniques adopted, secured results at higher spectral resolution and smaller error bars. Furthermore, we are able to thoroughly test

our data to eliminate/quantify the residual telluric contamination.

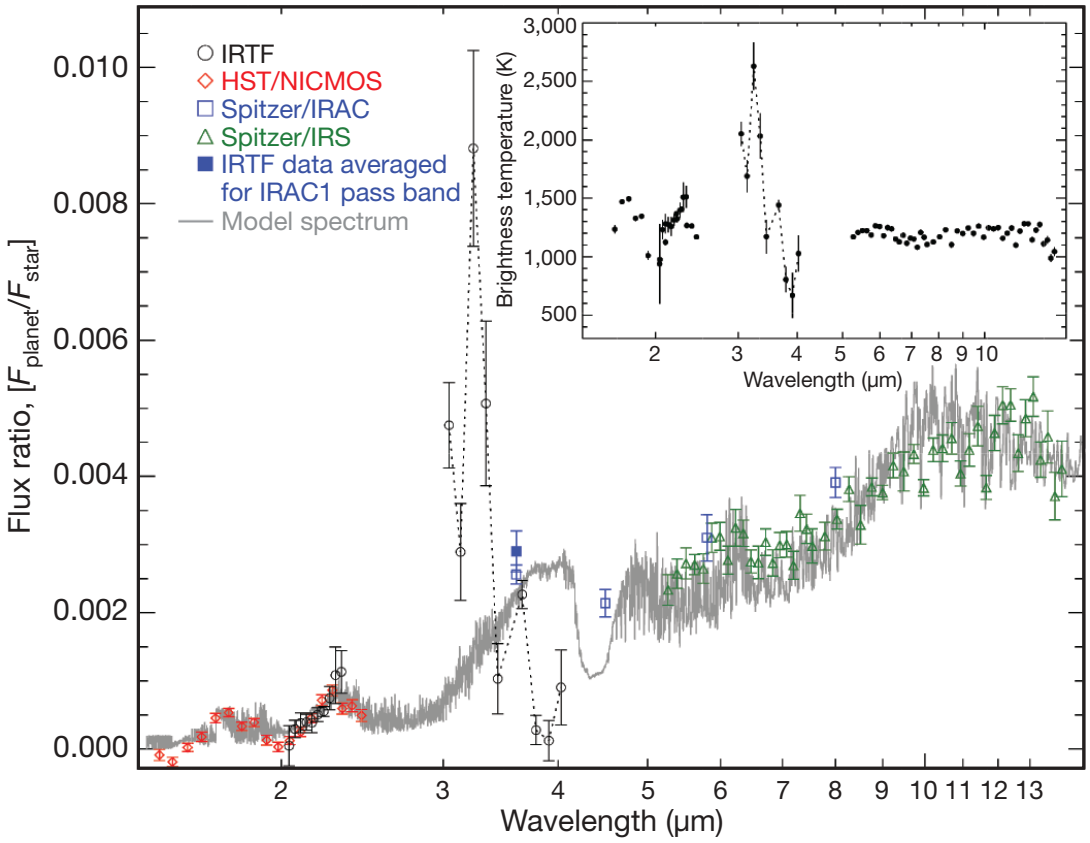


Figure 6.1: Unexpectedly strong 3.25-μm emission present in the dayside spectrum. The brightness temperature of the 3.25-μm emission feature indicates the likely presence of a non-LTE emission mechanism. The dayside emission spectrum is based on the new measurements reported in S10 (black), together with previous results from *Hubble* spectroscopy (red), *Spitzer* spectroscopy (green), and *Spitzer* photometry (blue); all data are shown 1σ errors. A radiative transfer model (grey) assuming LTE conditions and consistent with the measurements made with the *Spitzer* and *Hubble* space telescopes fails to describe the emission structure at 3.14.1 μm. (Swain et al., 2010).

6.2 Observations and data reduction

Secondary eclipse data of the hot-Jupiter HD189733b were obtained on the nights of August 11th 2007 (previously analysed by S10), June 22nd 2009 and the 12th of July 2009 using the SpeX instrument (Rayner et al., 2003) on the NASA Infrared Telescope Facility (*IRTF*). The observations were timed to start approximately one to two hours before the secondary eclipse event, until one to two hours post-egress. The instrumental setup was not changed for these three nights. The raw detector frames were reduced using the standard SpeX data reduction package, SpexTool, available for IDL (Cushing et al., 2004), resulting in sets of 439, 489 and 557 individual stellar spectra for each secondary eclipse event respectively. The extraction was done using the aperture photometry setting with a two arc-second aperture.

In addition we have analysed a fourth secondary eclipse of HD189733b observed on July 3rd 2010 using the same instrument. As opposed to the other three nights, we observed HD189733b in the L-band only, with a single order, long-slit setting. The one arc minute slit allowed us to simultaneously observe our target and a reference star with a K-band magnitude of 8.05 (2MASS 20003818+2242065). For not saturating the target star, we kept the exposure time at 8 seconds and employed the standard ABBA nodding sequence throughout the night. Each AB set was differenced to remove the background and the final spectra were extracted using both a custom built routine and standard IRAF routines. We found both extractions to yield the same results but the custom built routine performs better in terms of the final scatter observed. The flux received from the reference star is on average 27 times less than that of the target.

The secondary eclipses in the obtained raw spectra (from here onwards, 'raw' refers to the flat fielded, background corrected, wavelength calibrated and extracted spectra) are dominated with systematic (telluric and instrumental) noise. Consequently, the spectral reduction step is followed by data de-noising and signal amplification steps as described in the following sections.

6.3 Extraction of the exoplanetary spectrum

We describe in the following subsections how the planetary signal was extracted from the raw spectra. With the nature of the observations being a combined light (planet and stellar flux) measurement, we employ time-differential spectrophotometry during the time of the secondary eclipse. Standard photometric calibration routines typically achieve a $\sim$1% level of photometric accuracy, hence further de-noising is necessary to reach the required precision. We first removed the instrument systematics in the data (data cleaning) and then we extracted the planetary signal in the cleaned data (spectral analysis).

6.3.1 Data-cleaning

To achieve the accuracy we need, a robust cleaning of the data is required. The cleaning process comprises three main steps: 1) Normalising the spectra, getting rid of flux offsets in the time series and correcting for airmass variations. 2) Correcting wavelength shifts between spectra by re-aligning all spectra with respect to one reference spectrum. This step removes $\sim 80\%$ of outliers. 3) Filtering the time series of each spectral channel with adaptive wavelets. This step removes white and pink noise contributions at multiple passbands without damaging the underlying data structure - see Percival and Walden (2000).

Normalisation

Firstly, we discarded the spectral information outside the intervals of 2.1 - 2.45μm and 2.9 - 4.0μm to avoid the edges of the K and L photometric bands respectively. Then, we corrected for airmass and instrumental effects. This was achieved in a two step process. We first calculated a theoretical airmass function, $AF = exp(-b \times airmass(t))$, for each night and divided the data by this function. However, we found this procedure insufficient since the baseline curvature is caused not only by the airmass but by other instrumental effects (e.g. changing gravity vectors of the instrument). We hence additionally fitted a second order polynomial to the pre- and post-eclipse baseline of each time series and divided each single time series by the polynomial. Furthermore, we normalised each observed spectrum by its mean calculated in a given wavelength band (equation 6.1).

$$\hat{F}_n(\lambda) = \frac{F_n(\lambda)}{\bar{F}_n} \begin{cases} \lambda = 2.1 - 2.45\mu m & K - band \\ \lambda = 2.9 - 4.0\mu m & L - band \end{cases} \tag{6.1}$$

$$\bar{F}_n = \frac{\int_{\lambda_0}^{\lambda_1} F_n(\lambda) \mathrm{d}\lambda}{\lambda_1 - \lambda_0}$$

where $F(\lambda)$ is the flux expressed as a function wavelength, λ, for each spectrum obtained, n. $\bar{F}_n$ and $\hat{F}_n(\lambda)$ is the normalised spectrum. In the case of an idealised instrument and constant airmass, the normalisation would be superfluous. However, due to pixel sensitivity variations and bias off-sets on the detector, the individual spectra need to be normalised to avoid frequent 'jumps' in the individual time series. In the domain of the high-interference limit (Pagiatakis et al., 2007; Swain et al., 2010), the astrophysical signal is preserved. We investigated the effects of normalising the spectrum over a whole wavelength band or smaller sub-sections of the spectrum and various combinations of both, but found the differences to be negligible.

Spectra re-alignment and filtering

After the normalisation, we constructed 2D images with rows representing spectra of the planet-star system at a specific time, and columns representing time series for specific wavelengths (see figure 6.2A). In figure 6.2A, the main sources of outliers in individual time series, are mis-alignments by up to 4 pixels along the wavelength axis. We corrected this effect by fitting Gaussians to narrow (FWHM $\sim$ 5px) emission and absorption lines to estimate the line centres to the closest pixel. When the shift occurred for all the lines, the spectrum was corrected with respect to a reference spectrum, i.e. the first spectrum in the series. Then cosmic rays were removed by a 2D median filter replacing 5σ outliers with the median of its surrounding 8 pixels.

Wavelet de-noising

Due to variations in detector efficiency, the cumulative flux of each spectrum depends on the exact position of the spectrum on the detector (horizontal bands in figure 6.2A), resulting in high frequency scatter in each individual time series. This effect was already attenuated by the normalisation step but further removal of systematic and white noise is required. Based on the de-noising approach proposed by Thatte et al. (2010), we have opted for a wavelet filtering of the individual time series. There are clear advantages to wavelet de-noising compared to simple smoothing algorithms. With wavelets we can specifically filter the data for high frequency 'spikes' and low frequency trends without affecting the astrophysical signal or losing temporal phase information. This allows for an efficient reduction of white and pink noise in the individual time series. By contrast, smoothing algorithms, such as kernel regression, will impact the desired signal since these algorithms smooth over the entire frequency spectrum.. For a more detailed discussion see section 5.5 and Thatte et al. (2010); Donoho (1995); Percival and Walden (2000); Stein (1981); Sardy (2000). The use of the wavelet filtering to each individual time series yielded a factor of two improvement on the final error bars. The final results were generated with and without wavelet de-noising and found to be consistent within the respective error bars. An example of the final de-noised data can be seen in figure 6.2B.

6.3.2 Measuring the exoplanetary spectrum

After the data were de-noised as described in the previous subsection, we focused on the extraction of the planetary signal. We based our analysis on the approach described in chapter 5. The spectral emission features of a secondary eclipse event are too small to be statistically significant for an individual spectral channel. High signal to noise detections require a low spectral resolution, i.e. binning the data in λ. This can be done more efficiently in the frequency domain

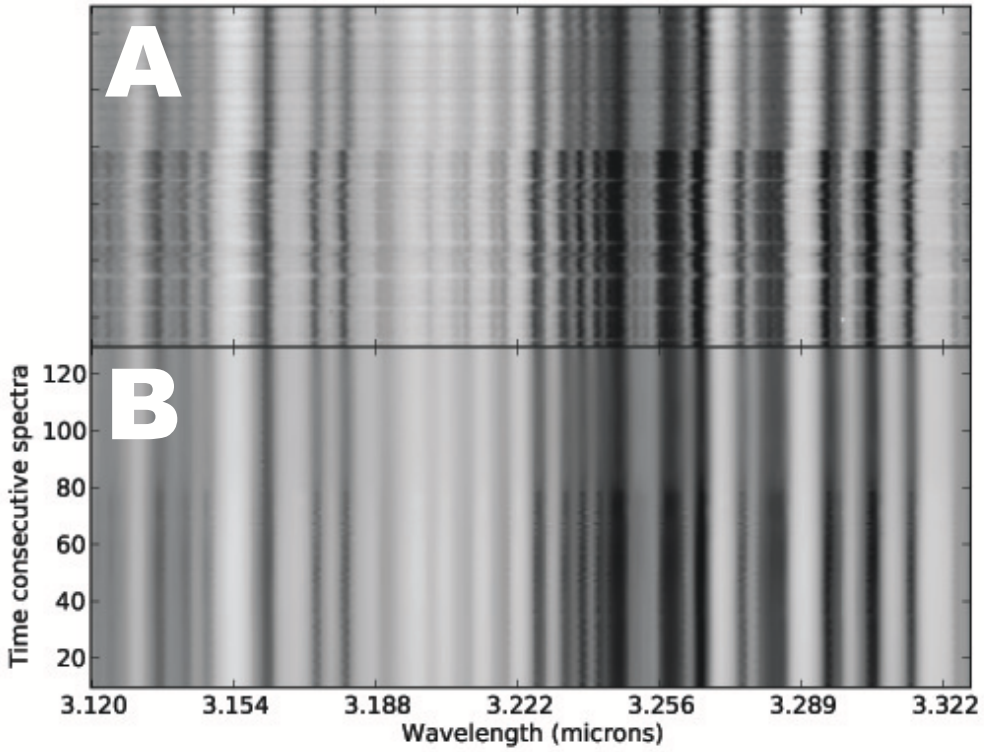


Figure 6.2: Zoomed in fraction of the data prior to the cleaning process (A) and post cleaning (B). Each column is a time series at a specific wavelength and each is an individual spectrum (n) taken at a specific time.

for reasons discussed in chapter 5 and below. Each time series $X_i(t)$ (here i denotes the spectral channel) was re-normalised to a zero mean to minimise windowing effects in the frequency domain. The discrete fast-Fourier transform (DFT) was computed for each time series and, depending on the final binning, m number of Fourier-transformed time series were multiplied with each other and finally normalised by taking the geometric mean (equation 6.2).

$$\mathcal{F}[\bar{X}(t)] = \left(\prod_{i=1}^{m} \mathcal{F}[X_i(t)] \right)^{1/m} \tag{6.2}$$

where $\mathcal{F}$ is the discrete Fourier-transform and $X_i(t)$ is the time series for the spectral channel i for m number of spectral channels in the Fourier product ($m \in \mathbb{Z}^+$). Since the input time series are always real and the Fourier transforms are Hermitian, we can take the n'th root of the real-part of the final product without losing information. For a more in-depth discussion of the Fourier convolution, please refer to chapter 5.

Time-domain analysis

Having calculated the Fourier product, $\mathcal{F}[\bar{X}(t)]$, for m spectral channels, we can take the inverse of the Fourier transform to obtain the filtered lightcurve signal.

$$\bar{X}(t) = \mathcal{F}^{-1}(\mathcal{F}[\bar{X}(t)]) \tag{6.3}$$

The lightcurves were then re-normalised by fitting a second-order polynomial to the out-of-transit baseline. We modeled the final lightcurves with equation 8 of Mandel and Agol (2002), using the system parameters reported in Bakos et al. (2006), with the transit depth as the only free parameter left. Here we ignore the Fourier induced distortions of the lightcurve morphology (see section 5.4.1) as the remaining systematic noise contribution masks intrinsic Fourier effects.

As clear from the lightcurves presented in section 6.5, the systematic noise in the data is higher in areas of low transmissivity. Systematic noise increases the scatter of the obtained lightcurves as well as the error-bars of the final spectra and places a lower limit of $m = 50$ channels (~ 2.88nm) on the currently achievable spectral bin size. This is a noticeable improvement compared to the original S10 analysis which reported a lower limit of $m = 100$ and 150 spectral channels for the K and L-bands respectively.

Frequency-domain analysis

The generated lightcurves are of high quality and ready for accurate spectroscopic measurements. However, as previously mentioned, a certain amount of periodic and systematic noise is still

present in the time series. The noise residuals are in part generated during the conversion of the data from the frequency domain to the time domain, in part are due to systematics. We can remove some of these residuals, by measuring the eclipse depth directly in the frequency domain, assuming that most systematic noise is found at different frequencies to the eclipse signal.

As we have seen in section 5.2 we can define the secondary eclipse lightcurve as a trapezoidal function of which the Fourier transform is a sinc like function. For reasons of continuity we repeat the equations 5.1 and 5.7 below, outlining the Fourier series for a trapezoidal function

$$\begin{aligned} f_{trap}(t) &= 8\sqrt(2)\delta \left(\sin(1/\tau) + \frac{\sin(3/\tau)}{9} - \frac{\sin(5/\tau)}{25} - ... \right) \\ &= 8\sqrt(2)\delta \sum_{k=1,3,5...}^{N} \left(\frac{\sin(k/4\tau) + \sin(3k/4\tau)}{k^2} \right) \end{aligned} \tag{6.4}$$

where δ corresponds to the transit depth of the lightcurve and τ to the duration between first and fourth contact points of the lightcurve and k is the Fourier coefficient index. We can hence relate the amplitude of a given Fourier coefficient, $|A|$, to the transit depth of the eclipse feature by

$$|A|_{trapez} = \frac{\tau\delta}{2} \sum_{k=1,3,5...}^{\infty} \frac{1}{k^2} \tag{6.5}$$

Section 5.4.2 describes how to determine the error-bars on the here derived transit depth. The advantages of such a direct measurement in the frequency domain is the reduction of systematic noise contribution to the transit depth measurement, since all frequencies but one are ignored.

6.3.3 Application to data

We have applied the same procedure described in sections 6.3.1 and chapter 5 to the four data sets. In addition to the individual analysis, we also combined in the frequency domain the three data sets recorded with the same observational technique. Given that the low-frequency systematics –such as residual airmass function, telluric water vapour content, seeing, etc– are significantly different for each individual night, by combining multiple data sets, we can amplify the lightcurve signal and reduce the systematic noise.

To generate the final K and L-band spectra, we chose in equation 6.2 $m = 100$ spectral channels. From $R_{spectra} = \lambda_{centre}/\Delta\lambda$, we get a final spectral resolution of R $\sim$ 50. Combining all three data-sets together ($\sim$33 spectral channels taken from each observed planetary eclipse) we obtain a spectral resolution of $\sim$170 and $\sim$ 185 for the K and L-bands respectively. We note

that the spectral resolving power for the SpeX instrument, considering the seeing, is R $\sim$ 800.

6.4 Model

We have simulated planetary emission spectra using line-by-line radiative transfer models as described in Tinetti et al. (2005, 2006) with updated line lists at the hot temperatures from UCL ExoMol and new HITEMP (Barber et al., 2006; Yurchenko et al., 2011; Rothman et al., 2009). Unfortunately accurate line lists of methane at high temperatures covering the needed spectral range are not yet available. We combined HITRAN 2008 (Rothman et al., 2009), and the high temperature measurements from (Thiévin et al., 2008). These LTE-models were fitted to the spectra presented in section 6.5.

Additional to the standard LTE model, we considered possible non-LTE models to fit the presented data. Upper atmospheres of planetary atmospheres are subject to non-LTE emissions; although negligible in most part of the near infrared spectrum, these emissions become dominant in the strongly absorbing vibration bands of molecular constituents, like CO_2 in telluric planets and CH_4 in giant planets (and Titan). A synthetic model of the spectrum in the L band has been adapted from a model of Giant Planets fluorescence of CH_4 developed for ISO/SWS (Drossart et al., 1999). The main steps involved in the radiative transfer with redistribution of frequency in non-LTE regime can be summarised as follows:

- We first calculate the solar (stellar) flux absorbed from all bands of CH_4. Although classical, this part of the model can be cumbersome as all the main absorption bands corresponding to the stellar flux have to be (in principle) taken into account. Limitations come from the knowledge of the spectroscopy of the hot bands. In this model, the following bands are taken into account: Pentad (3.3 micron) Octad (2.3 micron); Tetradecad (1.8 micron). An estimate of the accuracy of the approximation in neglecting hotter bands will be given below. Following an approach given by Doyennette et al. (1998), the spectroscopy of CH_4 is simplified by dividing the vibrational levels in stretching and bending modes: therefore we have a reduced number of superlevels (overtones of the methane $\nu 4$ emission band), instead of the full 29 potential sub-levels of the molecule. It is also assumed that for each super-level belonging to a polyad, thermal equilibrium is achieved within the population. This assumption comes from the observation that intra-vibrational transitions within polyads have a higher transition rate than inter-vibrational transitions.

- The population of the vibrational levels is then calculated within each "super-level" of CH_4. The vibrational de-excitation is assumed to follow the bending mode de-excitation

scheme (Appleby, 1990).

- From the population of each super-level, the radiative rate of each level can be calculated to determine the emission within each of the bands (fundamental, octad-dyad and tetradecad-pentad) contribute to the 3.3 micron domain. A schematic diagramme of the de-excitation scheme is given in figure 6.3.

- If hot band emission can be proven to remain optically thin down to deep levels of the atmosphere, the resonant fluorescence is not the same, as self-absorption is an essential ingredient of the fluorescence. Evidently, photons absorbed, on average, at a tau=1 level have the same probability to be re-absorbed as re-emitted upwards. The optically thick fluorescence, including absorption and re-emission, is therefore applied to the resonant band.

The increase in CH_4 vibrational temperature of 5% is presently an ad-hoc hypothesis: it simply describes the amount of non-LTE population required to fit the observations, pure LTE populations being insufficient. The source of this population increase can come for a variety of sources: XUV illumination from the star, electron precipitations, etc.which are presently not constrained at all. Such effects are nonetheless known in planetary physics, such as on Jupiter, where H_2 vibrational temperatures in the upper atmosphere have been demonstrated to be out of equilibrium through Ly-alpha observations (Barthélemy et al., 2005), with a 1.4-1.5 fold increase in vibrational temperature.

6.5 Results

6.5.1 Validation of the method used

As described in previous sections, we analysed four nights of observations: three in multi-order mode, with only HD 189733b in the slit (referred to as 'short-slit nights') and one night in L-band with single order, long-slit set up, observing HD 189733b and a fainter reference star simultaneously. While the long-slit observation covers a narrower spectral interval compared to the other eclipse observations, it is a critical test of the methodology with its simultaneous observations of the target and the reference star. In figure 6.13 we present two lightcurves: HD 189733 and the reference star. Both are centred at 3.31μm with a binning width of 50 channels ($\sim$ 2.88nm). As expected the HD 189733 time series (top) shows the distinctive lightcurve shape whilst the reference star (bottom) time series shows a null result. We have fitted a Mandel and Agol (2002) secondary eclipse lightcurve to both and found the HD 189733b transit depth

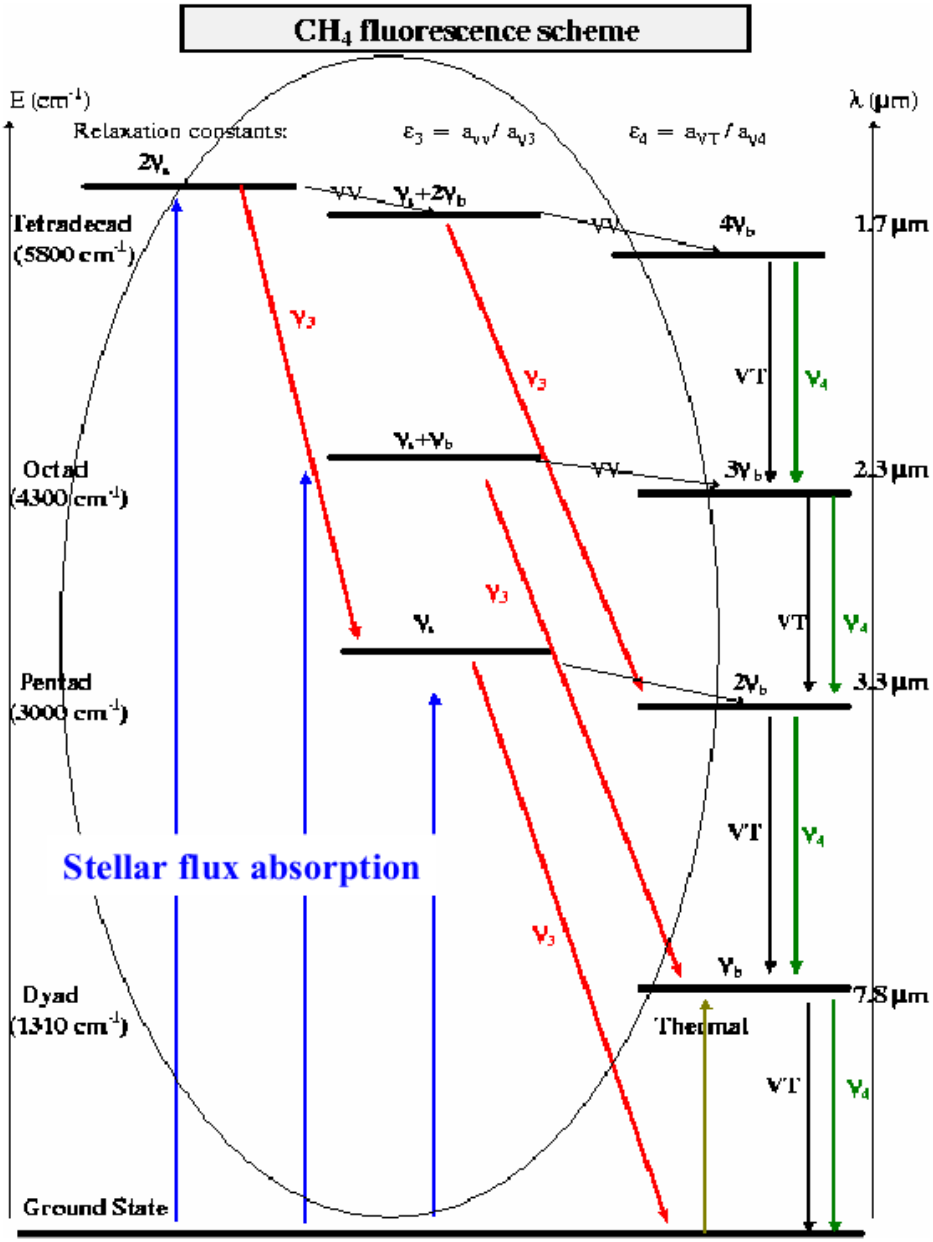


Figure 6.3: Scheme of the fluorescence calculation in methane for standard stellar flux illumination. Polyads of methane are condensed in "superlevels" corresponding to identical stretching/bending quanta. Thermal equilibrium is assumed within each super-level. Transitions between different levels are due to stellar absorption (blue), fluorescent emission in L band (red) and collisional or radiative de-excitation (black/green respectively). Other transitions can be neglected.

to be $\delta_{HD189} = 0.0078 \pm 0.0003$ and $\delta_{REF} = 0.0 \pm 0.0007$ respectively. These results are in good agreement with the spectra presented below.

6.5.2 K and L-band spectra

The same analysis was undertaken for the three short-slit nights: illustrative lightcurves are presented in figures 6.4 & 6.5. In figure 6.4 are plotted the lightcurves of the 'three-nights-combined' analysis for the K and L-band bands centred at 2.32, 3.20, 3.31 3.4 and 3.6 microns, with 50 channel ($\sim$ 2.88 nm) bins. The residual systematic noise is most pronounced in the areas of low atmospheric transmissivity, which is reflected in the error bars of the lightcurves and of the retrieved spectra. We also show the lightcurves centred on the methane ν_3 branch at $\sim$3.31μm for all individual nights, figure 6.5.

Having verified the detection of HD 189733b eclipse in all data sets, we have generated K- and L-band spectra for each individual night as well as for all the three nights combined. The three individual nights are plotted in figures 6.6 and 6.8 for K and L bands respectively. All spectra are consistent with each other and are within the error bars of the initial S10 results. This said, we find the nights of the 11th of August 2007 and July 12th 2009 of higher quality and in better agreement. The single night analysis supports the assumption that intra-night variations are negligible which allowed us to averaged the data sets and hence increase the signal to noise of the final spectra. We could hence push the resolution to R $\sim$ 170-180 for the final combined spectra. Figures 6.9 and 6.10 are the three-nights-combined K and L-band spectra respectively. We include in these figures the comparison with black body emission curves and LTE models. It is clear from the figures that the strong features observed in the L-band cannot be explained by standard LTE processes.

6.5.3 Comparison of the observations with atmospheric LTE and non-LTE models

Even if many uncertainties subsist on the thermal vertical profile of HD189733b, the thermal methane emission needed to reproduce the observed spectrum would lead to brightness temperatures of $\sim$3000 K, which not only are unlikely given the star-planet configuration, but would also appear in other bands –e.g. in the ν4 band at 7.8 μm– hypothesis ruled out from *Spitzer* observations. While LTE models cannot explain such temperatures, non-LTE models with only stellar photons as pumping mechanism do not supply enough excess flux. This result is not unexpected since the contribution of stellar reflection from the planet is smaller in L band than the thermal emission, and fluorescence is only a redistribution of the stellar flux (even if a small

enhancement comes from the redistribution of frequency in the fluorescence cascade). However, a good fit can be obtained by assuming a vibrational temperature excess for methane by 5% due to an enhancement of the octad level population in methane which is higher than expected by stellar flux pumping (figure 6.10). This increase is currently an ad-hoc hypothesis and simply describes the amount of vibrational temperature increase required to explain the observed feature.

In the case of the K-band spectrum, it is less obvious whether LTE or non-LTE processes are prevalent. We show in figure 6.9 a comparison with two LTE simulations, one including CH_4 plus CO_2 in absorption as suggested by other data sets. Another model was obtained with LTE emission of methane. However, neither of the two simulations perfectly capture the spectrum observed. Given the stronger non-LTE emission features detected at $\sim$3.3 μm, one can expect to find non-LTE effects in the K-band as well. Further observations are required in order to build up the required spectral resolution to decisively constrain the excitation mechanisms at work.

6.6 Discussion

In figure 6.6 we present the K-band spectra of the three separate nights. This plot shows a slight discrepancy between the night of the 22nd of June 2009 compared to the other two nights analysed. We can observe a systematic off-set in both the K and L-bands (figure 6.8) with this night giving consistently lower emission results. We associate this effect to the poorer observing conditions and a degraded quality of the data compared to the data obtained in the other two nights: a very high intrinsic scatter of the data may in fact reduce the eclipse depth retrieved. We estimated the average spectra excluding the night of June 22nd 2009 (figure 6.7) and found the results to be in good agreement with the 3 nights-combined spectrum. This test demonstrates the robustness of the final retrieved spectrum. It should be noted that this issue is less severe in the L-band, since the overall signal strength is higher, than in the K-band.

Whilst the K-band spectra could be explained with LTE models, we encounter a quite different picture in the L-band. The observed emission around $\sim$3.3μm exhibits a very poor match with the predicted LTE scenario. By contrast, non-LTE emission of methane can capture the behaviour of the ν_3 branch. Similar fluorescence effects have been observed in our own solar system, mainly CO_2 in telluric and CH_4 in giant solar system planets (Barthélemy et al., 2005). Figure 6.12 is an example of *ISO*/SWS data obtained in the L-band for Jupiter and Saturn. Both planets exhibit a strong fluorescence of methane in the same spectral range as it has been observed in HD189733b. Similarly, CH_4 fluorescence has been observed on Titan in three *Cassini* fly-bys, figure 6.11. Here the spectral shape of the CH_4 emission at an altitude of 1700km is close

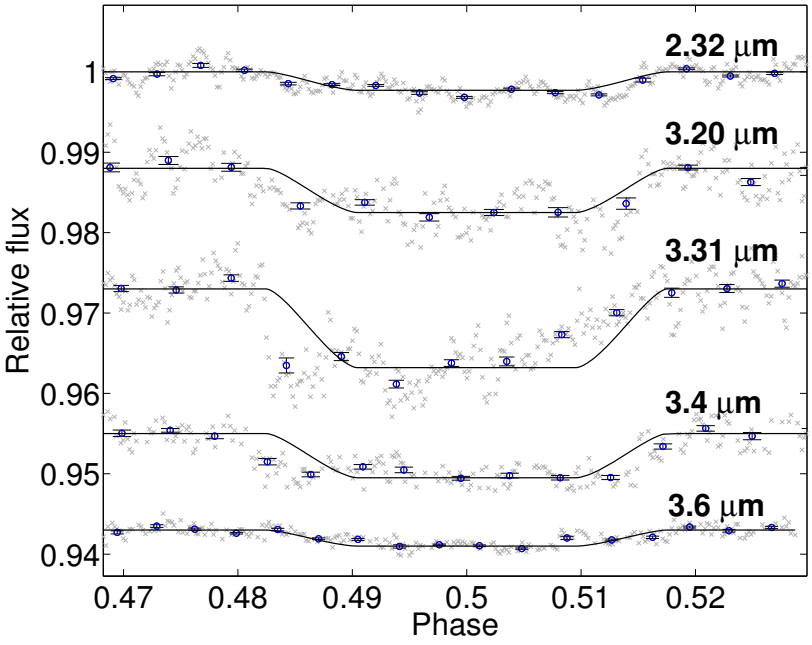


Figure 6.4: Lightcurves of the 'three-night-combined' analysis for the K and L bands. Lightcurves are offset vertically for clarity.

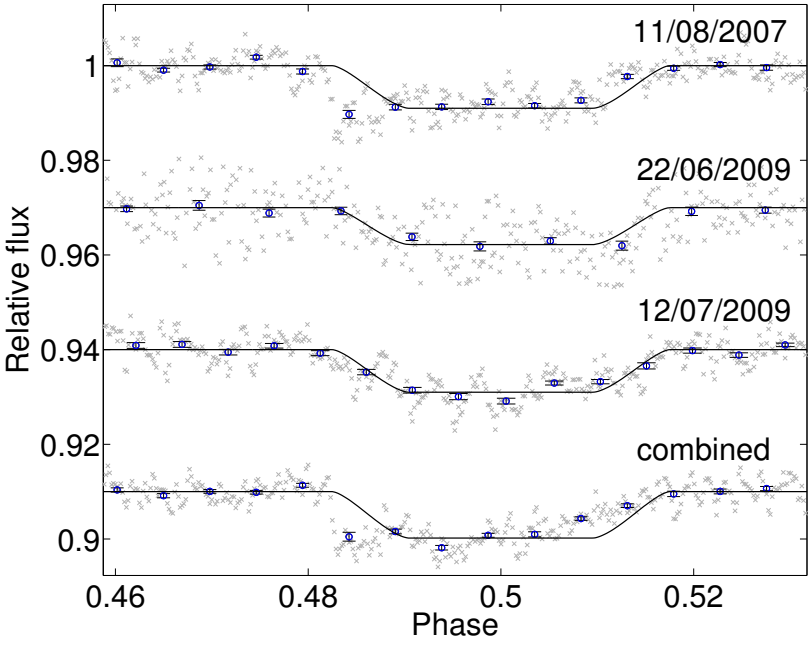


Figure 6.5: Lightcuves centred at 3.31μm with a bin size of 50 channels ($\sim$ 2.88 nm) for the three individual nights and 'three-nights-combined'.

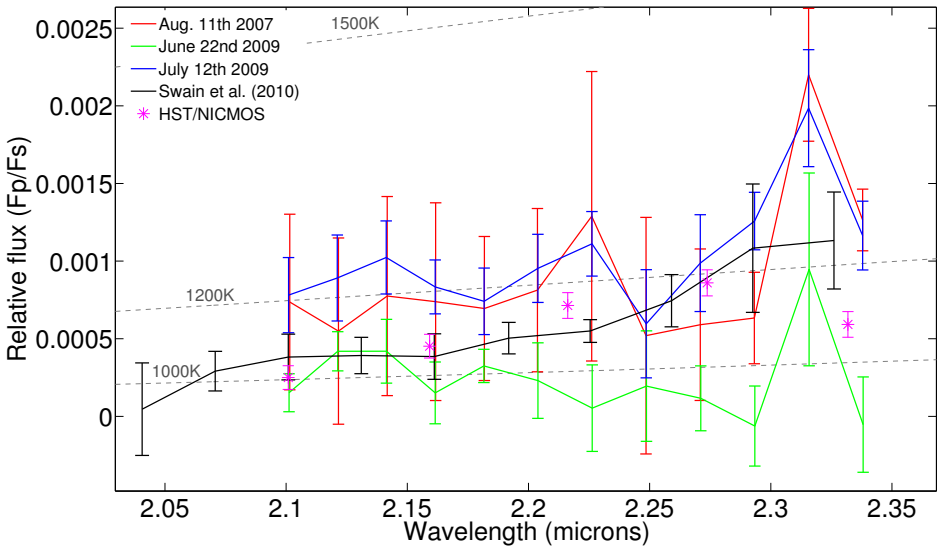


Figure 6.6: showing K-band planetary signal for the three nights separate: August 11th 2007, June 22nd 2009 and the 12th of July 2009 in red, green and blue respectively. The night of June 22nd 2009 had poor observing conditions and the data was significantly noisier and planetary emissions retrieved are systematically lower for this night in both K and L-band. Results from S10 are shown in black.

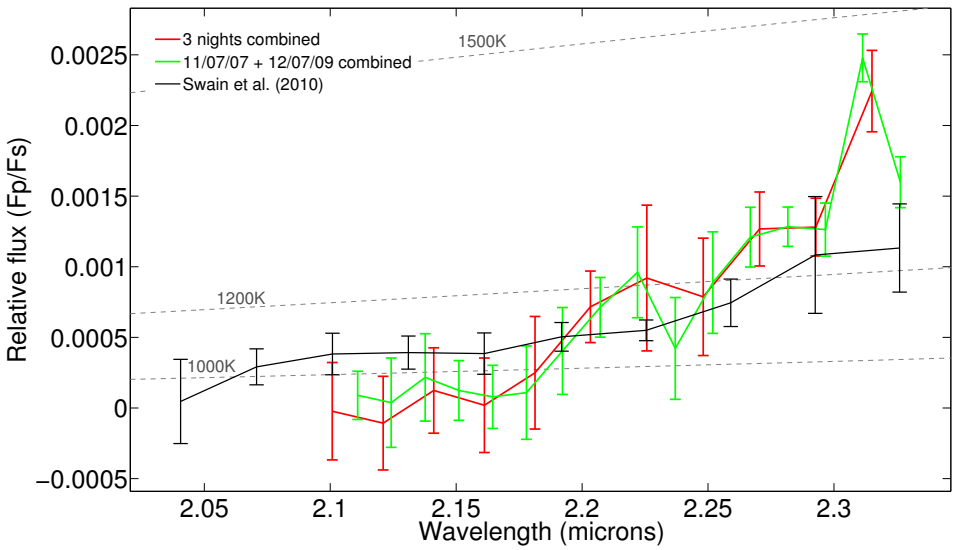


Figure 6.7: showing the combined K-band planetary signal for the nights of August 11th 2007 and July 12th 2009 only (green), excluding the poor data quality of the June 22nd 2009 night. For comparison the spectrum of all three nights combined (red) is overplotted. The difference between both spectra is small and indicates the night of June 22nd 2009 having a small effect on the overall result. Ground-based results from S10 and *Hubble*/NICMOS data (Swain et al., 2008c), are shown in black and purple respectively.

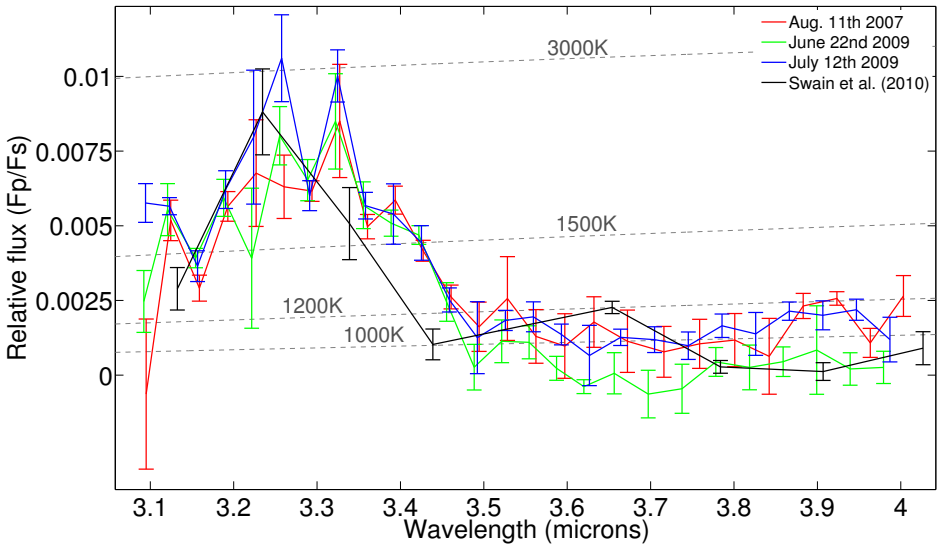


Figure 6.8: showing L-band planetary signal for the three nights separate: August 11th 2007, June 22nd 2009 and the 12th of July 2009 in red, green and blue respectively. Similar to figure 6.6, the night of June 22nd 2009 shows a systematic lower emission. As described previously, this may be a result of the poor data quality of this night. Results from S10 are shown in black.

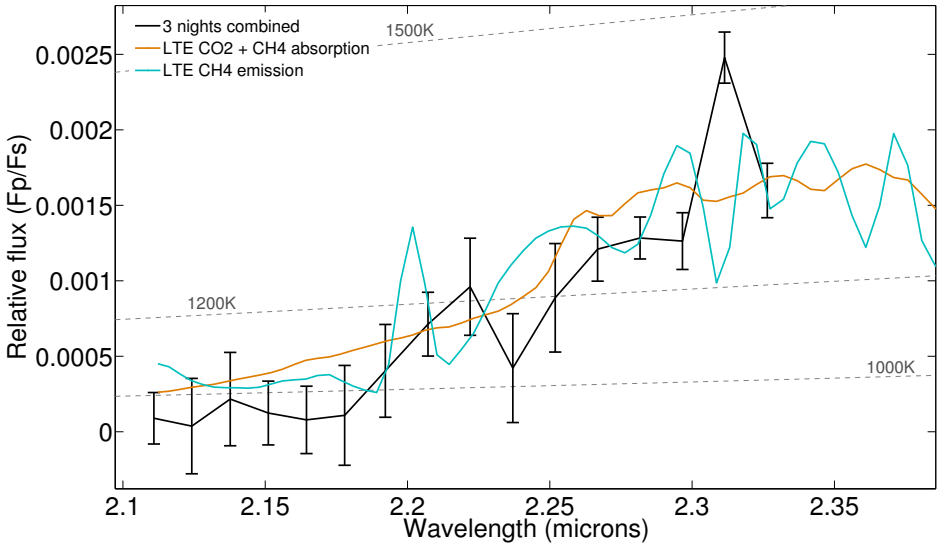


Figure 6.9: Three night combined K band spectrum compared with three black body curves at 1000, 1500, 2000 K. Furthermore two LTE models of CH_4 in emission (turquoise) and CH_4 plus CO_2 in absorption (orange).

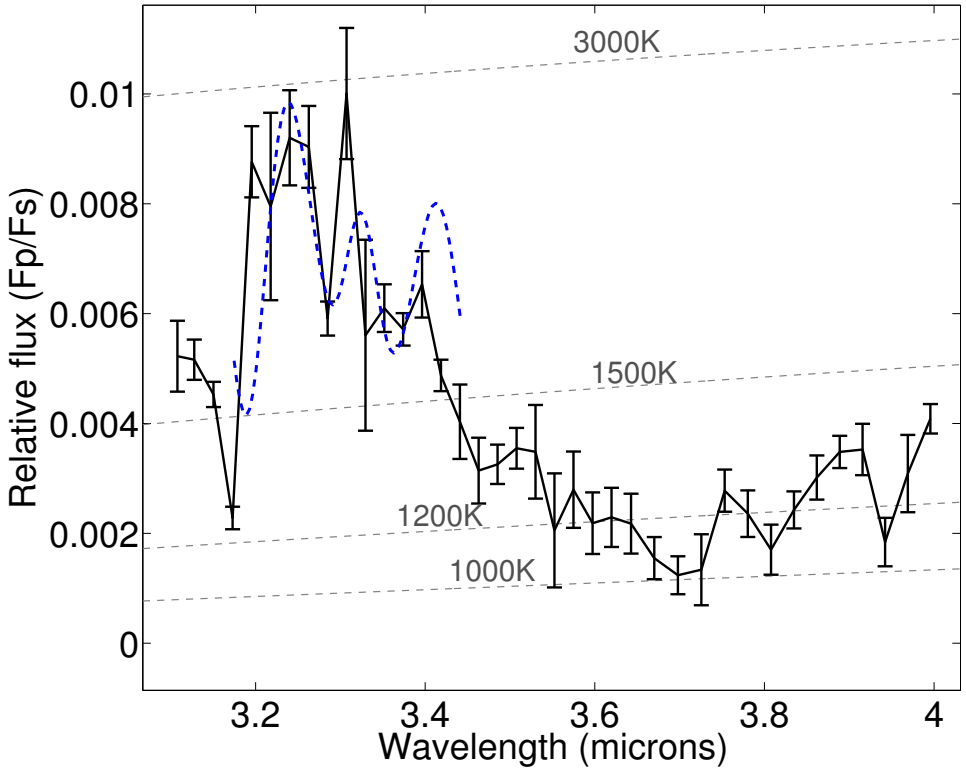


Figure 6.10: Three nights combined L-band spectrum. The blue discontinuous line shows a comparison of the observations with the "enhanced fluorescent" model; non-thermal population enhancement in the octad level with a 5% increase of vibrational temperature of CH_4. Overlaid are black body curves at 1000, 1500, 2000, 3000 K.

to what is observed in HD189733b, giving further support to the conclusion that these observed effects originate in the upper atmosphere of this inflated hot-Jupiter.

In section 6.4 we outline a plausible model for the creation of such a prominent feature. As previously mentioned, the increase in CH_4 vibrational temperature of 5% is presently an ad-hoc hypothesis: it simply describes the amount of non-LTE population required to fit the observations, pure LTE populations being insufficient. The source of this population increase can come for a variety of sources: XUV illumination from the star, electron precipitations, etc.which are presently not constrained at all. Such effects are nonetheless known in planetary physics, such as on Jupiter, where H_2 vibrational temperatures in the upper atmosphere have been demonstrated to be out of equilibrium through Ly-alpha observations (Barthélemy et al., 2005), with a 1.4-1.5 fold increase in vibrational temperature.

6.6.1 Validation of observations

The results presented here are found to be consistent with the results initially presented by S10, HST/NICMOS data in the K-band (Swain et al., 2008c) and verified in the L-band by the *Spitzer*/IRAC 3.6μm broadband photometry (Charbonneau et al., 2008), see figures 6.7 & 6.8. However, Mandell et al. (2011), from here M11, recently published a critique of the original S10 result reporting a non-detection of any exoplanetary features in their analysis. Since the results of this publication are in good accord with S10, the fundamental discrepancy between the findings presented here and those by M11 needs to be addressed.

M11 argue that the L-band features reported by S10 were likely due to un-accounted for telluric water emissions rather than exoplanetary methane. This hypothesis poses four main questions which will be addressed below: (1) Do the L-band features look like water emissions? (2) Are the results repeatable? (3) Do or do we not see similar lightcurve features in the reference star? (4) Can we quantify the amount of residual telluric contamination in the data?

Do the L-band features look like water?

Here the simple answer is no. As discussed in section 6.5 and shown in figures 6.8 & 6.10, the improved spectral resolution of these results shows that we are clearly dealing with methane signatures. As M11 pointed out, a temporary change in telluric opacity due to atmospheric water (or methane) could mimic a secondary eclipse event. However, for temporal atmospheric variations to mimic an eclipse signal in the combined result of all three nights, the opacity variations, as well as the airmass function, would need to be identical or at least very similar in all data sets. The likelihood of such hypothesis is very small.

In addition, we have retrieved weather recordings from near-by weather stations. These include periodic temperature, relative humidity and pressure readings from the the CFHT[1] as well as atmospheric opacity (tau) readings at 225 μm obtained by the CSO[2] (see figure 6.18). Spread over all three eclipsing events, we found no significant correlations between these parameters and the secondary transit shape expected.

Are the results repeatable?

A main focus throughout this publication is to demonstrate the repeatability of the observations. In section 6.5 we present spectra retrieved for each individual observing run of the three 'short-slit' nights and found them consistent with each other within the error-bars. For the methane ν_3 band which is the most difficult to achieve measurement we present lightcurves for all three observing runs considered, figure 6.5. These do vary in quality from night to night but are found to be consistent with one another over a measured timescale ranging from August 11th 2007 to July 12th 2009. This test of repeatability is of paramount importance in asserting the validity of the analysis as a whole.

Do or do we not see similar lightcurve features in the reference star?

We do not see any lightcurve features in the reference star's time series. As described in previous sections, we have obtained a fourth night in addition to the three main nights analysed here. This fourth night was taken in the single-order, L-band only mode with a one arc-minutes long slit. This allowed us to simultaneously observe the target HD189733b and a fainter reference star, 2MASS 20003818+2242065, over the course of a secondary eclipse on July 3rd 2010. We have equally applied the same routines outlined in section 6.3 to both, the target and the reference. In figure 6.13 we plot the resulting lightcurves of both stars centred at 3.31μm using the standard 50 channel bin. We find the transit depth for HD189733b to be within the error bars of the other nights analysed, see figure 6.15, whilst the reference star time series is flat. Hence, the routines used produce a null result where a null detection is expected.

Furthermore it is important to note that a faulty background subtraction would have much stronger effects on the fainter reference star than on the target, as any residual background is a proportionally larger fraction of the stellar signal. We find the mean observed flux for a single exposure to be F_{HD189} $\sim$24300e$^-$ and F_{REF} $\sim$900e$^-$ for the target and the reference stars respectively. We can now state that the observed flux is a sum of the stellar flux and a background contribution: $F_{observed} = F_{star} + F_{back}$. We also assume that the background flux, F_{back}, is the

[1]http://mkwc.ifa.hawaii.edu/
[2]http://www.cso.caltech.edu/

same for both stars as they were observed simultaneously on the same detector. Whatever the value of F_{back} may be, its relative contribution on the overall flux would be ∼27 times higher for F_{REF} than for F_{HD189}. Following this argument, if we now assume the lightcurve feature to be due to an inadequate background correction (as postulated by M11), we would expect a ∼27 times deeper lightcurve signal in the reference star time series than in HD189733b. To illustrate the severity of this effect, we re-plotted the time series presented in figure 6.13 with an additional 27 times deeper transit than that of HD189733b underneath. Given the flat nature of the reference star's time series though, we can confidently confirm an adequate treatment of telluric and other backgrounds.

Can we quantify the residual telluric contamination in the data?

Using the Fourier based techniques described in chapter 5, we can quantify the remaining contribution of systematic noise and the residual telluric components in the spectra shown in sec. 6.5. As described in section 5.4.2, we are mapping individual Fourier coefficients of the lightcurve signal in the frequency domain. Any systematic noise or telluric contamination can therefore only contribute to this one frequency bin. The degree of residual contamination by systematics on that frequency bin can hence be estimated by running the routine described in section 5.4.2 on only out-of-transit and only in-transit data, i.e. removing the eclipse signal. Figure 6.16 and 6.17 show the planet signal (black) and out-of-transit and in-transit measurements of the contamination in red and green respectively. We conclude that the amplitude of the systematic noise and the residual telluric component is within the error bars of the planetary signal.

6.7 Conclusion

In this chapter and Waldmann et al. (2012b) I presented new data on the secondary eclipse of HD189733b recorded with the SpeX instrument on the *IRTF* . Our data analysis algorithm for time-resolved, ground-based spectroscopic data, is based on a thorough pre-cleaning of the raw data and subsequent spectral analysis using Fourier based techniques outlined in chapter 5. The results of this analysis can be summarised as follows;

- By combining three nights of observations, with identical settings, and a further development of the data analysis methodology presented in S10, we could to increase the spectral resolution to R $\sim$ 175.
- We confirm the existence of a strong feature at $\sim 3.3\mu$m, corresponding to the methane ν_3 branch, which cannot be explained by LTE models. Non-LTE processes are most likely the origin of such emission and we propose a plausible scheme to explain it.
- The possibility of telluric contamination of the data is thoroughly tested but we demonstrate that the residual due to atmospheric leakage is well within the error-bars, both by using Fourier based techniques and additional observations with a reference star in the slit. This critical test demonstrates the robustness of our calibration method and its broad applicability in the future to other space and ground exoplanet data.

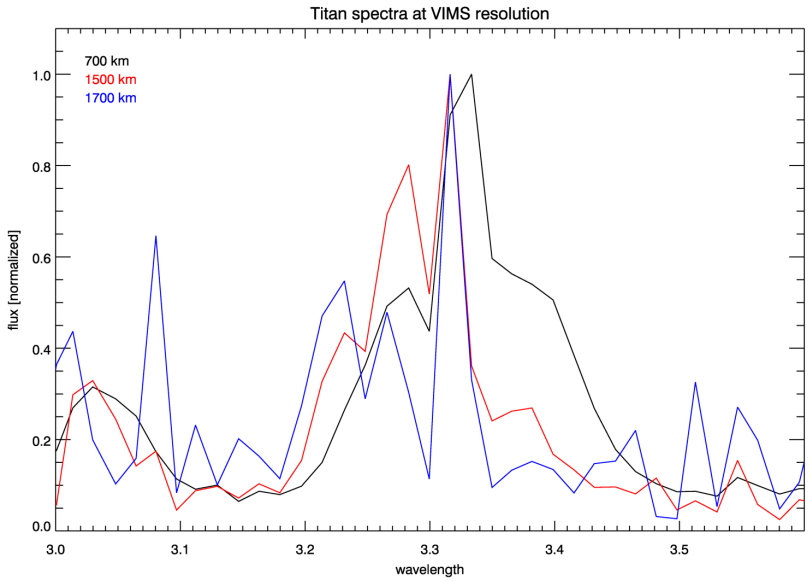


Figure 6.11: A sequence of three *Cassini*/VIMS spectra taken off Titan's limb showing CH_4 ν_3 band fluorescence and progressive weakening of the P branch with increasing altitude. At 1700km, the Titan spectral shape is a good match to the HD189733b spectrum in figure 6.1 (Swain et al., 2010).

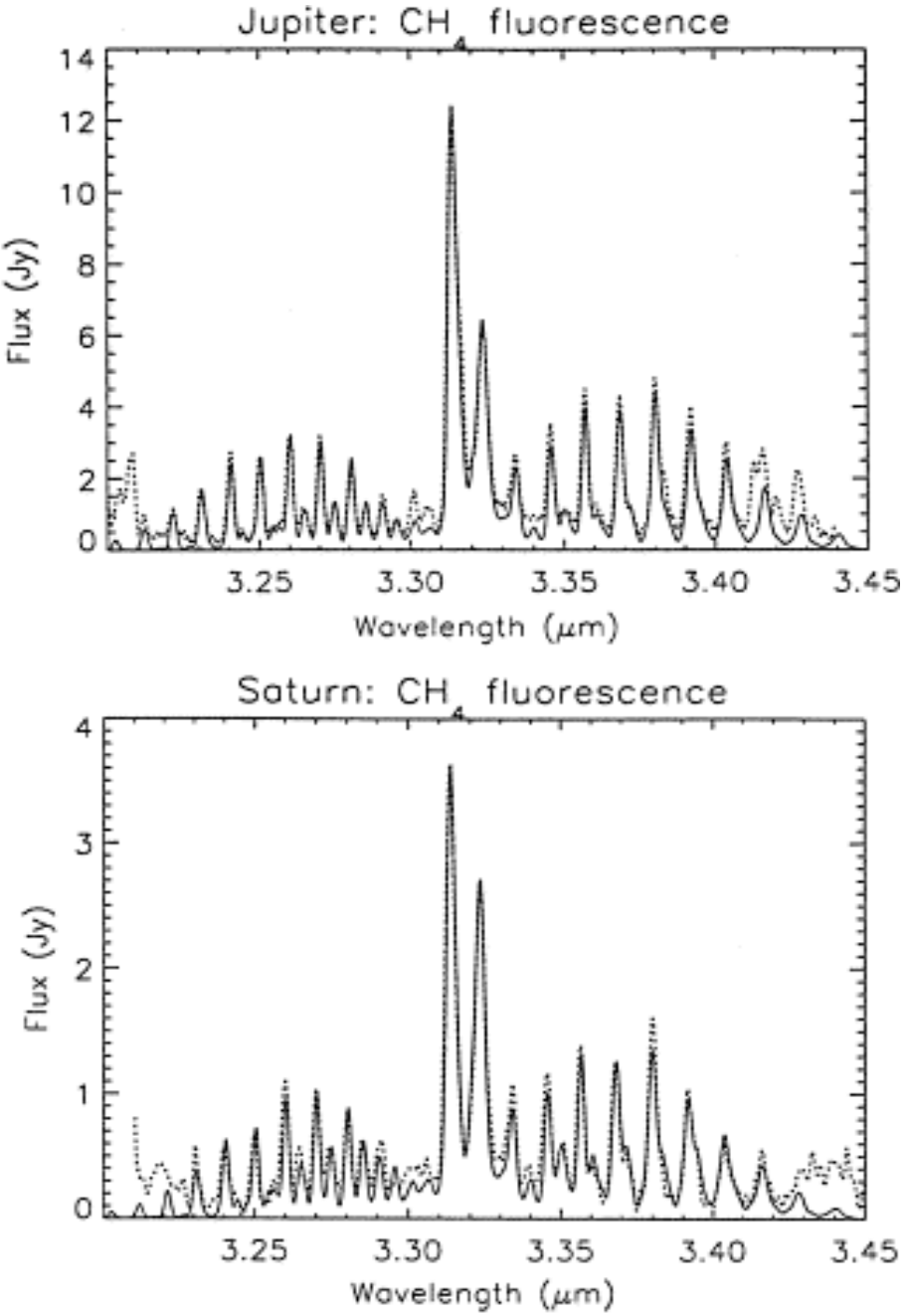


Figure 6.12: Synthetic spectral model of the CH_4 ν_3 branch fluorescence (solid) over plotted on ISO spectral data of Jupiter (top plot, dotted line) and Saturn (bottom plot, dotted line), taken from Drossart et al. (1999).

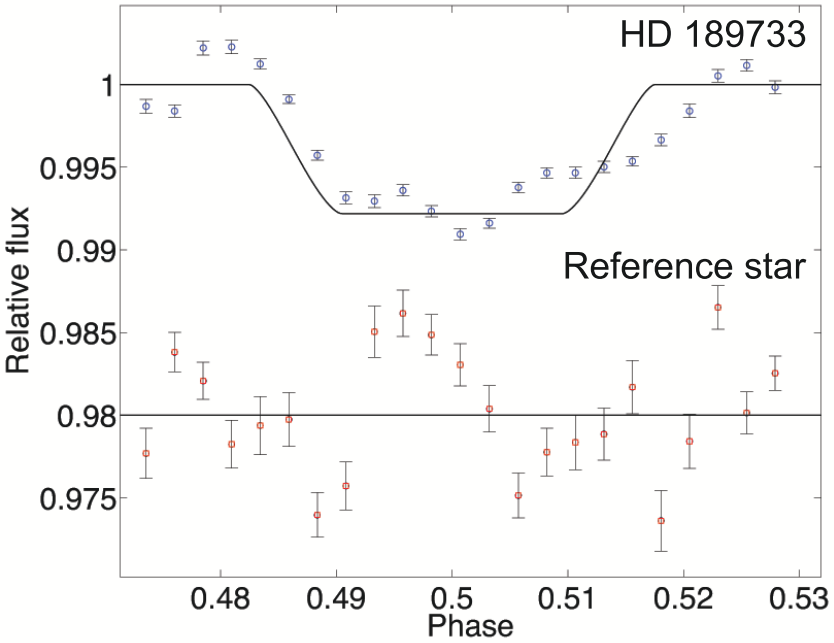


Figure 6.13: showing the lightcurves of the long-slit analysis of HD189733b and the simultaneously observed fainter reference star beneath, centred at 3.31μm with the standard 50 channel binning. Over-plotted are two fitted Mandel and Agol (2002) curves for the secondary eclipse. The HD189733b lightcurve is in good agreement with the other results of this analysis whilst the reference star's time series is noticeably flat.

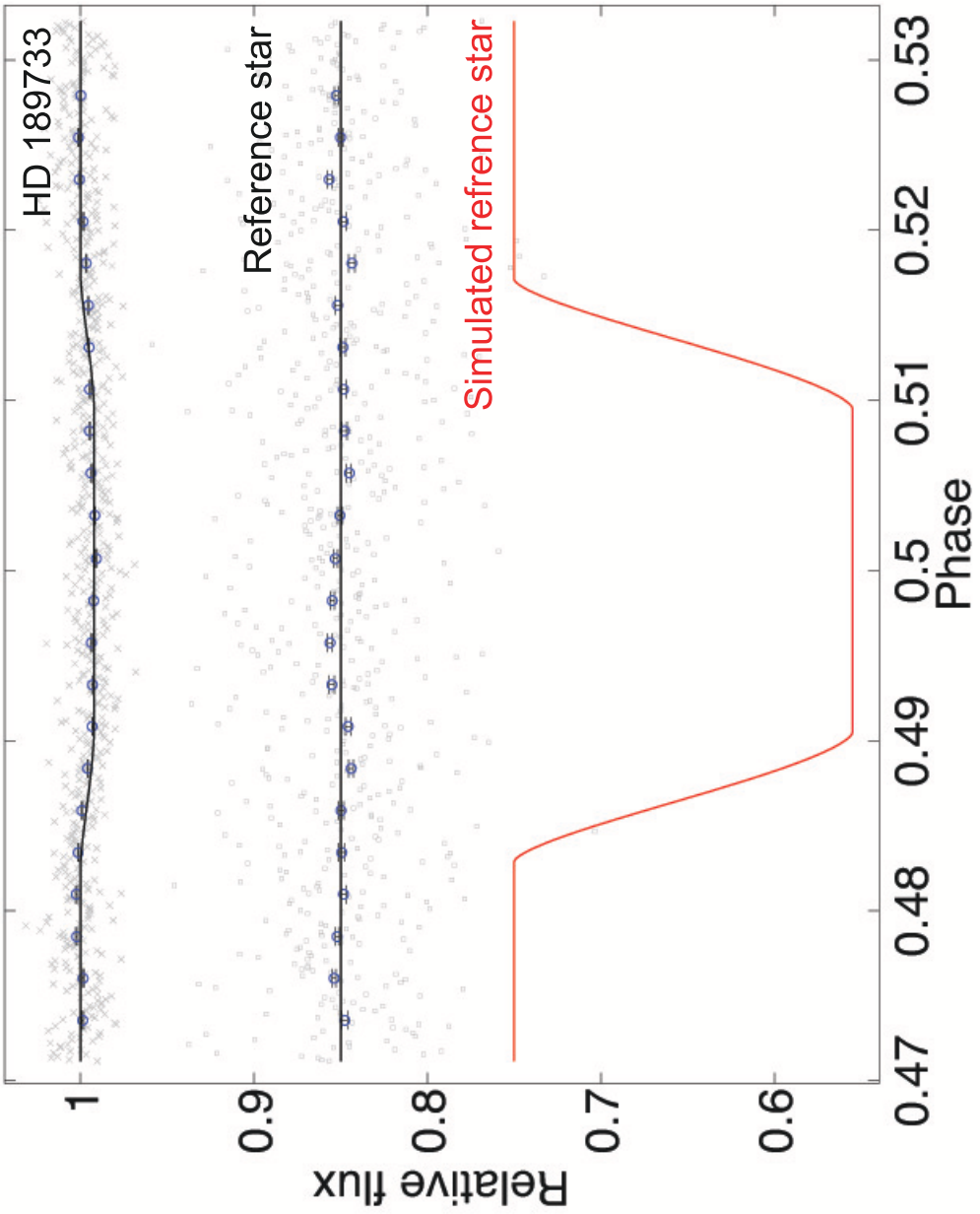


Figure 6.14: showing on the top the observed lightcurve of HD189733b, beneath the simultaneously observed flat time series of the fainter reference star. At the bottom in red is the simulated reference star lightcurve expected to be observed under the assumption that the observed signal in HD189733b is due to an imperfect background subtraction. The flat nature of the observed reference star lightcurve is a strong indication that the background subtraction was treated adequately.

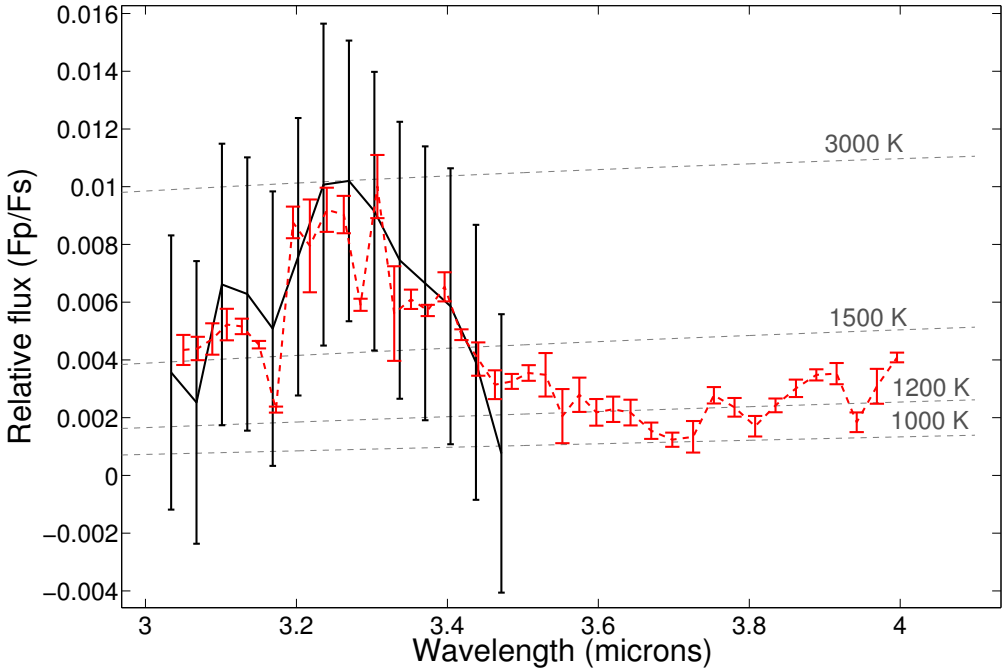


Figure 6.15: Black, continuous line: Spectrum retrieved for HD189733b observed in the long-slit setting. Red, discontinuous line: Three nights combined spectrum from figure 6.10. Grey, discontinuous lines: Black body curves for various temperatures. Despite the much poorer data quality of the long-slit setting night, the retrieved spectrum is in agreement with the three-nights-combined result.

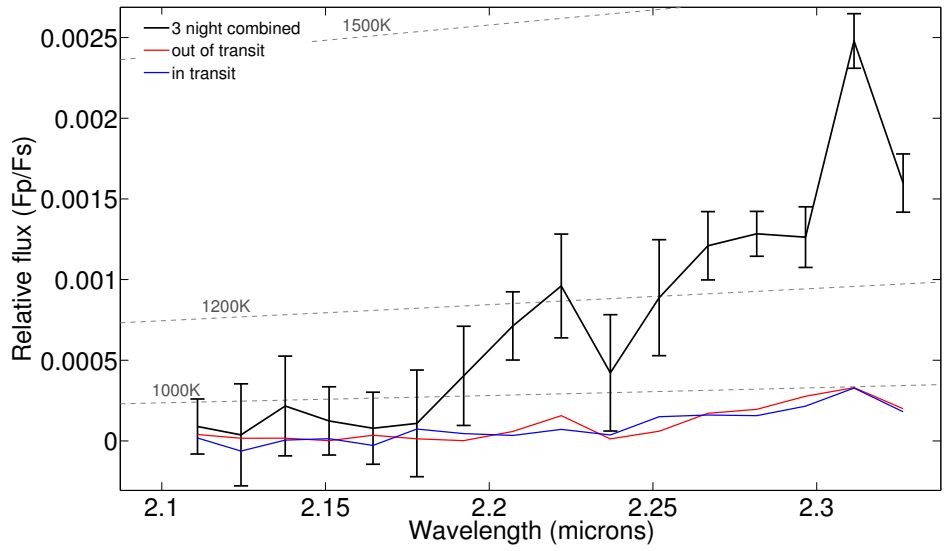


Figure 6.16: showing the three night combined K-band result (black), in-transit and out-of-transit contamination measures are plotted in blue (dash-dotted) and red (dashed) respectively. It can clearly be seen that the contamination by telluric components is much smaller than the planetary signal and that its amplitude lies within the signal's error bar.

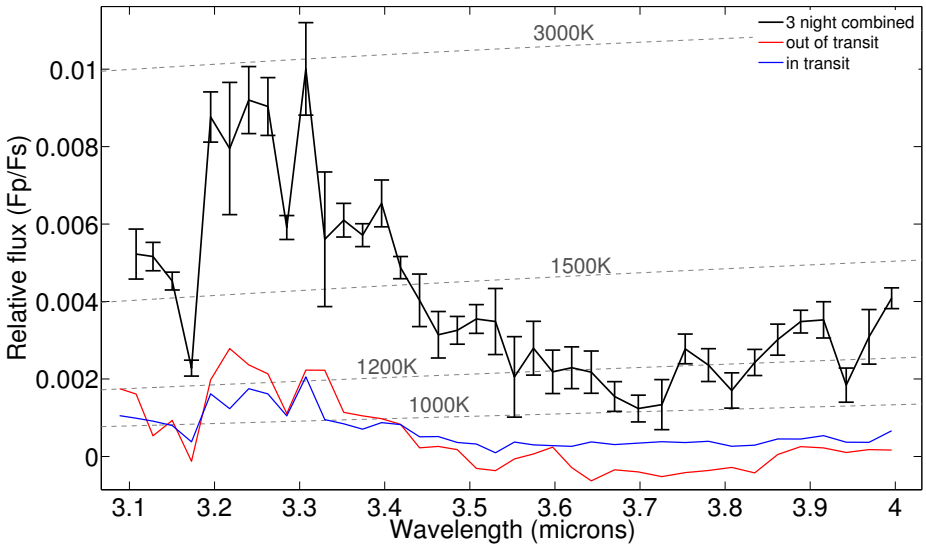


Figure 6.17: showing the three night combined L-band result (black), in-transit and out-of-transit contamination measures are plotted in blue (dash-dotted) and red (dashed) respectively. It can clearly be seen that the contamination by telluric components is much smaller than the planetary signal and that its amplitude lies within the signal's error bar.

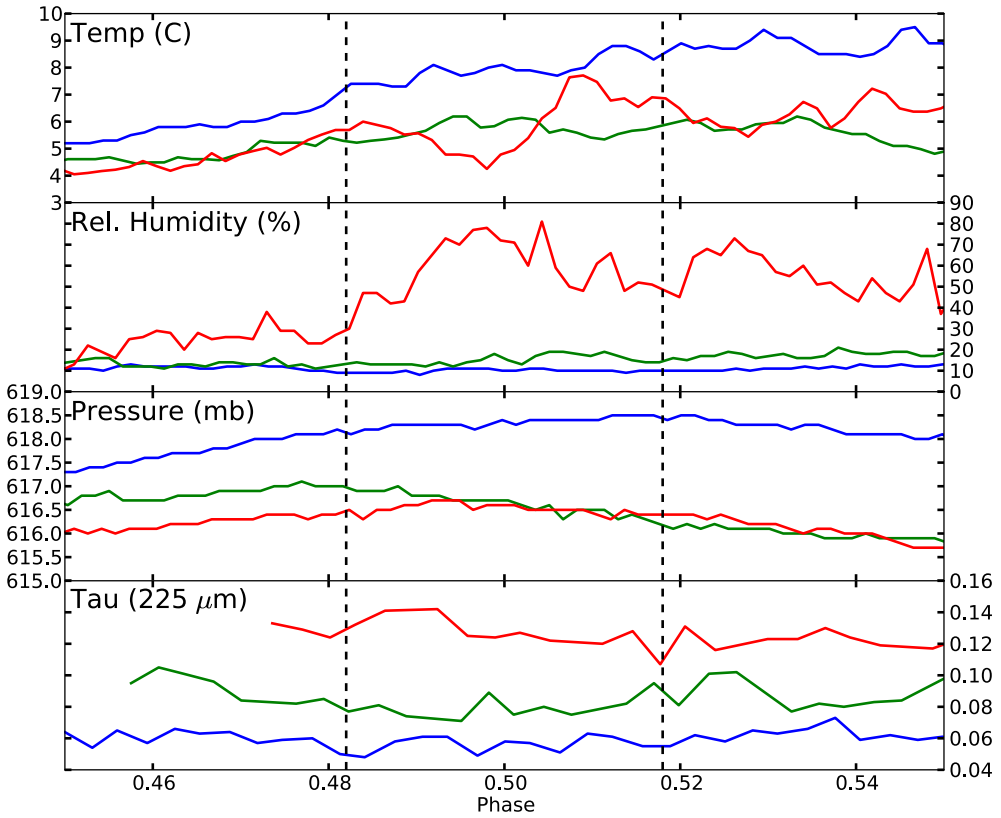


Figure 6.18: showing from top to bottom: Temperature (deg. C, CFHT Weather station), Rel. Humidity (%,CFHT), Pressure (mb,CFHT) and optical depth, tau (225μm, CSO) for the 12nd Aug. 2007 (blue), 22nd June (green) and 12th July 2009 (red). The discontinuous vertical lines mark the secondary transit duration.

Chapter 7

Future Work

"Ideas are like rabbits. You get a couple and learn
how to handle them, and pretty soon you have a dozen."

— John Steinbeck

7.1 Introduction

In the previous chapters I presented two main methods of data de-trending: 1) by component separation and 2) by frequency domain analysis. Both techniques have distinct advantages and disadvantages. Whilst the component separation, described in chapter 3, is a powerful tool to disentangle systematic noise from the astrophysical signal, it is limited in its performance to high signal-to-noise (SNR) data with little to no Gaussian noise. On the other hand, spectral analysis techniques such as the Fourier convolution filtering described in chapter 5 robustly retrieve the desired transit-depth parameter in very low-SNR conditions but come at the expense of distorting the overall astrophysical lightcurve signal. Furthermore, in the Fourier convolution we have no information of the systematic noise components' morphologies but can measure their amplitudes, whereas in the component-separation case, we know the systematics' morphologies well but their respective amplitudes are unknown.

It is easy to see that the properties of both methods are highly complementary with each other. An ideal de-trending algorithm should hence have the non-Gaussian component separation abilities of the ICA algorithm with the Gaussian noise damping abilities and low-SNR robustness of the frequency domain convolution. One such possible algorithm is based on the implementation of the component deconvolution algorithm, described in chapter 3, via wavelet masks in the frequency domain.

I will here briefly outline and describe such a possible algorithm as pathway for future exo-

planetary data analysis.

7.2 Sparsity re-visited

In section 5.2 I discussed the sparsity of the lightcurve signal in the Fourier domain. It could be shown that the Fourier series of a lightcurve shape is a rapidly converging series with <5 Fourier coefficients describing the lightcurve shape. This lead to the concept of the Fourier series being overcomplete (i.e. containing more coefficients than necessary) and the astrophysical lightcurve signal being sparse (i.e. very few of the coefficients present are needed to describe the desired signal). This is true for the Fourier series as much as for a multi-resolution-analysis (MRA) using a discrete or continuous wavelet transform. We described the MRA in section 5.5 for the purpose of selective Gaussian noise damping using the discrete wavelet transform.

We can now extend the concept of lightcurve sparsity to the wavelet MRA space. Figure 7.1 shows the multi-resolution decomposition of a secondary eclipse lightcurve and a sinusoidal curve. We can see that similar to the Fourier analysis, few coefficients are required in the MRA case to completely describe the signal. Both the lightcurve and sine-curve have very characteristic signatures in the MRA space.

7.3 From PCA/ICA to MRA masks

We can now extend the study in figure 7.1 to any non-Gaussian signals retrieved from the independent-component-analysis (ICA) or a principal-component-analysis (PCA) of the multi-variate time series data as discussed in chapters 3 & 4. For convenience we repeat equations 3.3, 5.8 and 5.10 as 7.1, 7.3 and 7.4 respectively. For N simultaneously observed time series, we obtain the observed signal vector $\mathbf{x}$ where the instantaneous mixing model is given by

$$\mathbf{x} = \mathbf{A}\mathbf{s} \tag{7.1}$$

where $\mathbf{A}$ is the mixing matrix and $\mathbf{s}$ the signal component vector made out of the individual signals, s_l, where l is the signal component index. For an individual time series in $\mathbf{x}$, we can also express this as

$$x_k(t) = \sum_l a_{k,l} s_l(t) \tag{7.2}$$

where k is the time series index and $a_{k,l}$ a weighting factors comprising $\mathbf{A}$. We now turn to the definition of the continuous wavelet transform (CWT), although we use the discrete form for

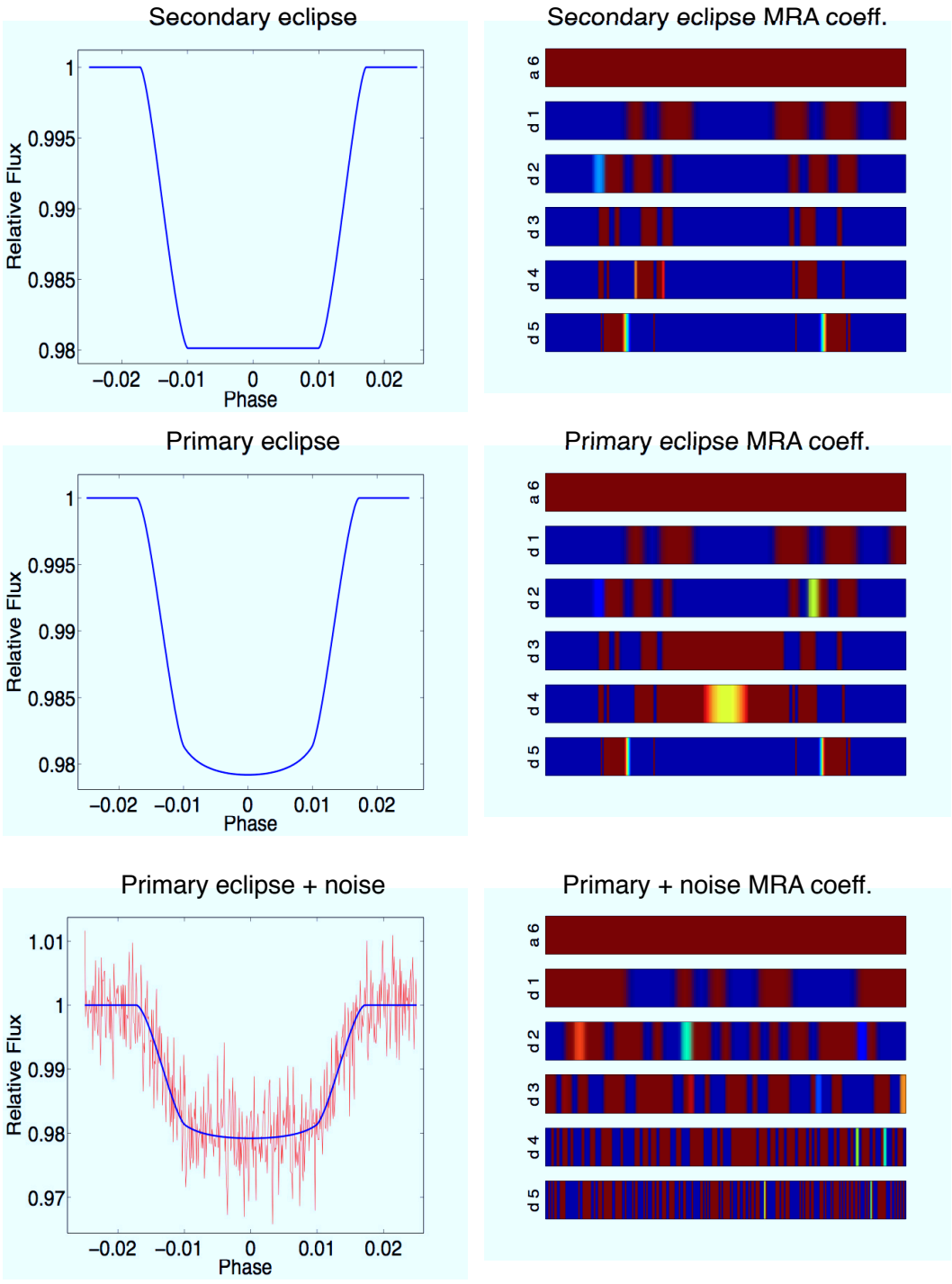


Figure 7.1: **Left column:** Time-domain plots of a secondary eclipse lightcurve, a primary eclipse lightcurve and a primary eclipse lightcurve with a SNR of ∼10 (red). **Right column:** wavelet coefficients of 5 level MRA, scale down sampled by a factor of two (↓ 2) from top to bottom with d5 containing the highest frequency coefficients. All dark coefficients have negligible amplitude with red and green coefficients $|cD_s|$ or $|cA_S| >> 0$. We can see that in the first two cases most coefficients are close to zero with the lightcurves having distinctive signatures in wavelet space. The noisy case (bottom) requires far more coefficients to describe than the noiseless cases. Also note the increased number of coefficients needed in d3 & d4 to describe limb-darkening in the primary eclipse case compared to the secondary eclipse.

the actual computation, we here use the CWT for its simpler notation. For a given lag, τ, and scale, s, we obtain the wavelet coefficients by

$$c_{\tau,s} = \int_{\mathbb{R}} x(t)\psi_{\tau,s}(t)\mathrm{d}t \tag{7.3}$$

where $\psi_{\tau,s}(t)$ is the mother wavelet at given lag and scale. The inverse is given by

$$x(t) = \sum_{s}\sum_{\tau} c_{\tau,s}\psi_{\tau,s} \tag{7.4}$$

We can now replace the time series $x(t)$ in equation 7.3 with the non-Gaussian signal $s_l(t)$ from equation 7.2 to obtain a MRA decorposition of the s_l signal

$$\mathcal{L}_{l,\tau,s} = \int_{\mathbb{R}} s_l(t)\psi_{\tau,s}(t)\mathrm{d}t \tag{7.5}$$

where $\mathcal{L}_{l,\tau,s}$ denotes the MRA coefficient for a given signal s_l, lag and scale. From figure 7.1 we can appreciate that most signals are sparse in MRA space and we can disregard all coefficients close to zero amplitude in the wavelet MRA without loss of information. Similar to section 5.5.3 where we used thresholding of coefficients to dampen white noise contributions, we can now define a threshold for $\mathcal{L}_{l,\tau,s}$ to create a boolean mask of the relevant wavelet coefficients to the signal s_l. This mask consists of 1's and 0's denoting the relevant coefficients and thresholding all others to zero. The mask is given by

$$\mathcal{M}_{l,\tau,s} = \begin{cases} 0, |\mathcal{L}_{\tau,s}| < \mathcal{T} \\ 1, |\mathcal{L}_{\tau,s}| > \mathcal{T} \end{cases} \tag{7.6}$$

where $\mathcal{T}$ is some threshold and $\in \mathbb{R}$. It is now possible to use the mask $\mathcal{M}_{l,\tau,s}$ and to apply it to the MRA of an individual time series, $x_k(t)$. This has the effect of only selecting the wavelet coefficients relevant to the non-Gaussian signal s_l contained in $x_k(t)$. In other words we apply an optimal filter designed for the spectral shape of s_l to the data. Since $\mathcal{M}_{l,\tau,s}$ is a boolean filter, we do not alter the amplitude information contained in the wavelet coefficients. Hence, we can retrieve the non-Gaussian signal s_l with the correct scaling from the data. The 'masking' process is given by

$$\hat{x}_{k,l}(t) = \sum_{s\in\mathbb{Z}}\sum_{\tau\in\mathbb{Z}} \psi_{\tau,s}(t)\left[\mathcal{M}_{l,\tau,s}\int_{\mathbb{R}} x_k(t)\psi_{\tau,s}(t)\mathrm{d}t\right] \tag{7.7}$$

where $\hat{x}_{k,l}(t)$ is the filtered or 'masked' time series for a given non-Gaussian signal s_l.

Figure 7.2 shows the MRA decompositions of a secondary eclipse lightcurve and sinusoid.

On the right, we have used a thresholding of $\mathcal{T} = 0.5$ to create wavelet coefficient masks of the decompositions shown to their respective lefts. This mask is then applied to the individual time series data. A flowchart overview of the proposed algorithm is given in figure 7.4.

We can furthermore define the selective SNR for each non-Gaussian component in each time series $SNR_{l,k}$ as

$$SNR_{l,k} = \frac{\mathcal{M}_{l,\tau,s} \int_{\mathbb{R}} x_k(t)\psi_{\tau,s}(t)\mathrm{d}t}{\mathcal{N}_{l,\tau,s} \int_{\mathbb{R}} x_k(t)\psi_{\tau,s}(t)\mathrm{d}t} \tag{7.8}$$

where $\mathcal{N}_{l,\tau,s}$ is the inverse mask of $\mathcal{M}_{l,\tau,s}$ and defined as

$$\mathcal{M}_{l,\tau,s} = \begin{cases} 0, |\mathcal{L}_{\tau,s}| > \mathcal{T} \\ 1, |\mathcal{L}_{\tau,s}| < \mathcal{T} \end{cases} \tag{7.9}$$

7.4 Discussion of the algorithm

The above proposed algorithm has several advantages over pure component separation or Fourier filtering algorithms.

- Non-Gaussian signals obtained by either ICA or PCA or ICA+PCA, can be used to create optimal masks in wavelet space. Filtering for the non-Gaussianity in the original signal allows us to retrieve the correct scaling of s_l which is not retrieved by the ICA algorithm and cannot be retrieved otherwise in low-SNR conditions. Whereas in chapter 4 we fitted for the scaling of s_l using out-of-transit data in high-SNR conditions, this minimisation technique does not converge in low-SNR conditions.

- The masking process is a natural extension to the soft-thresholding scheme proposed by Donoho (1995) and outlined in section 5.5.3. Hence, additionally to the masking procedure described above, we can furthermore selectively dampen Gaussian noise in the data.

- By using the output of the ICA algorithm as input to the wavelet masking, we ensure the entire process to be non-parametric without prior or additional knowledge of the instrument, giving us un-biased estimates of the astrophysical and systematic noise signals.

- In addition to the fully non-parametric approach, any semi- to fully-parametric definitions are allowed.

- Using wavelet masks we can selectively obtain the SNR of each non-Gaussian component per time series. This allows us to estimate the relative significance of a given signal

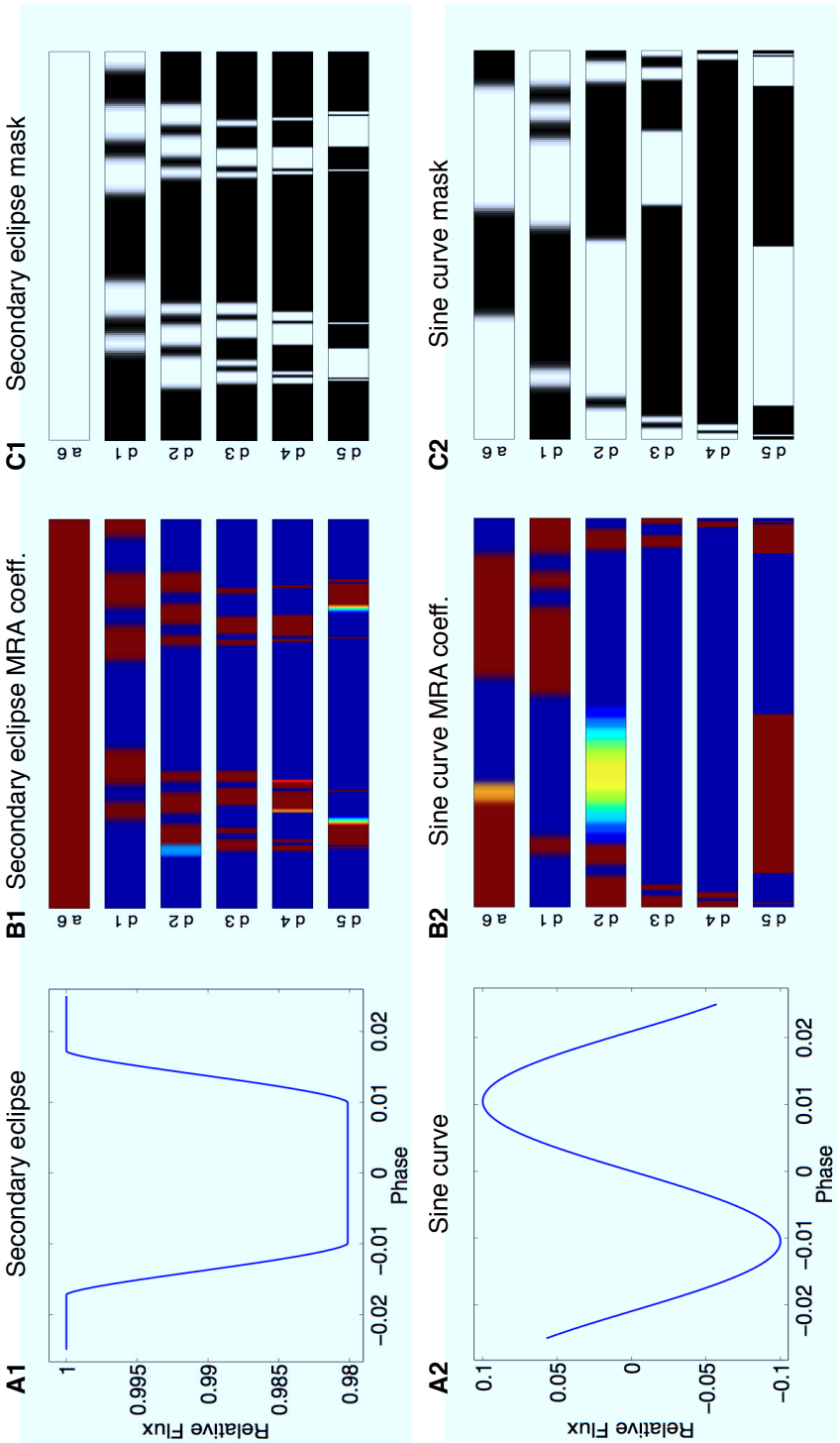


Figure 7.2: Discrete wallet transform multi-resolution analysis (MRA) of two signals: 1) secondary eclipse lightcurve; 2) sinusoidal curve. A-diagrammes: time domain plot of signal; B-diagrammes: wavelet coefficients of 5 level MRA, scale down sampled by a factor of two ($\downarrow 2$) from top to bottom with d5 containing the highest frequency coefficients. All dark coefficients have negligible amplitude with red and green coefficients $|cD_s|$ or $|cA_S| >> 0$. We can see that both signals are highly sparse with characteristic signatures in MRA space. C-diagrammes: Coeffcient mask $\mathcal{M}_{l,\tau,s}$ with $\mathcal{T} = 0.5$. Black and white areas represent 0's and 1's respectively.

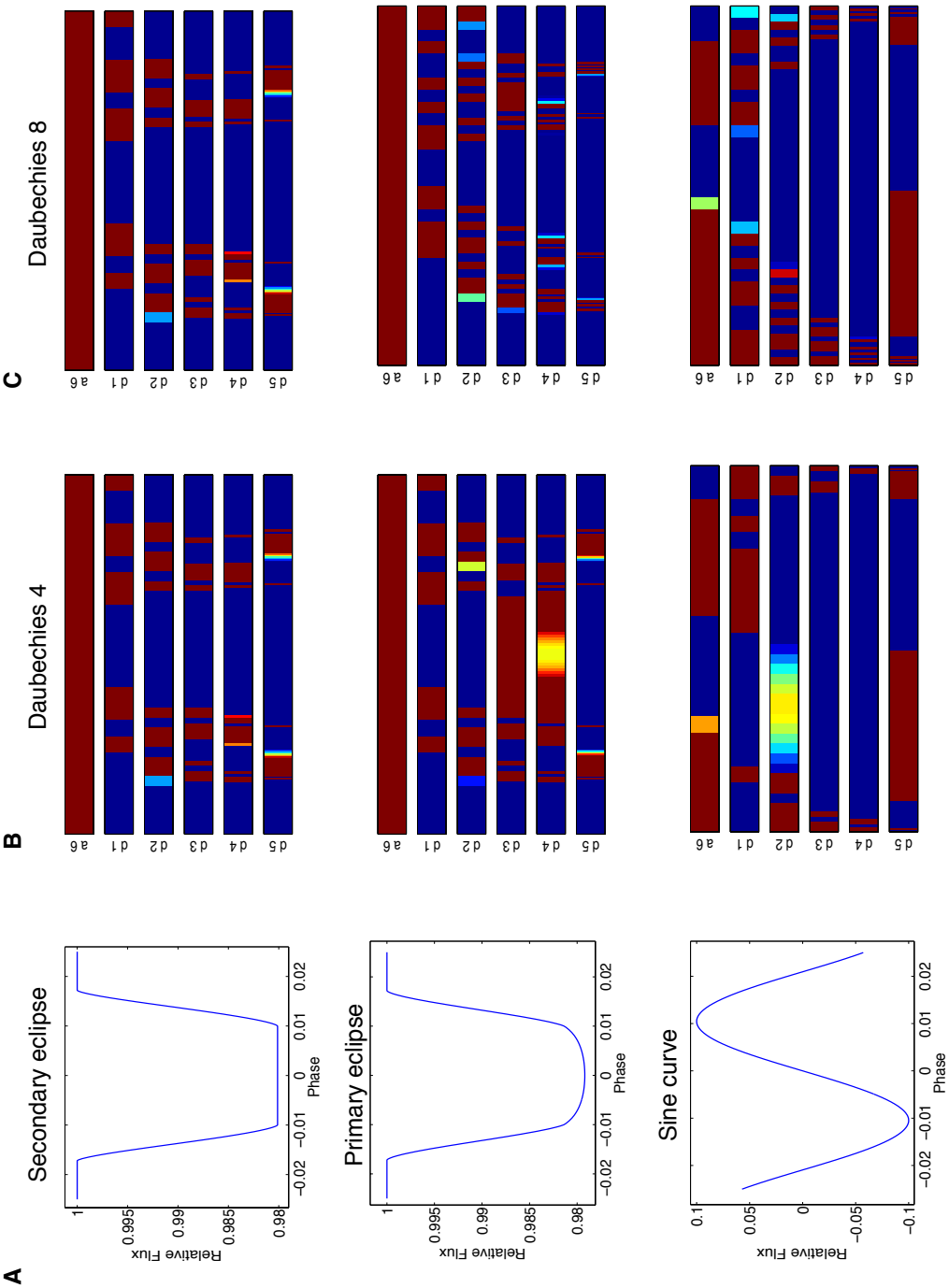


Figure 7.3: A) From top: lightcurves of the secondary eclipse and primary transit identical to figure 7.1; sinusoidal wave from figure 7.2. B) MRA coefficients of signals in A) using a Daubechies 4 (db4) 'mother' wavelet. C) Same as in B) but using a Daubechies 8 (db8) 'mother' wavelet. The db8 wavelet is able to better describe high-order polynomials, i.e. rounder shapes, due to the higher number of coefficients available. This allows db8 to describe the primary transit and sinusoid with fewer coefficients than are necessary for a db4 decomposition.

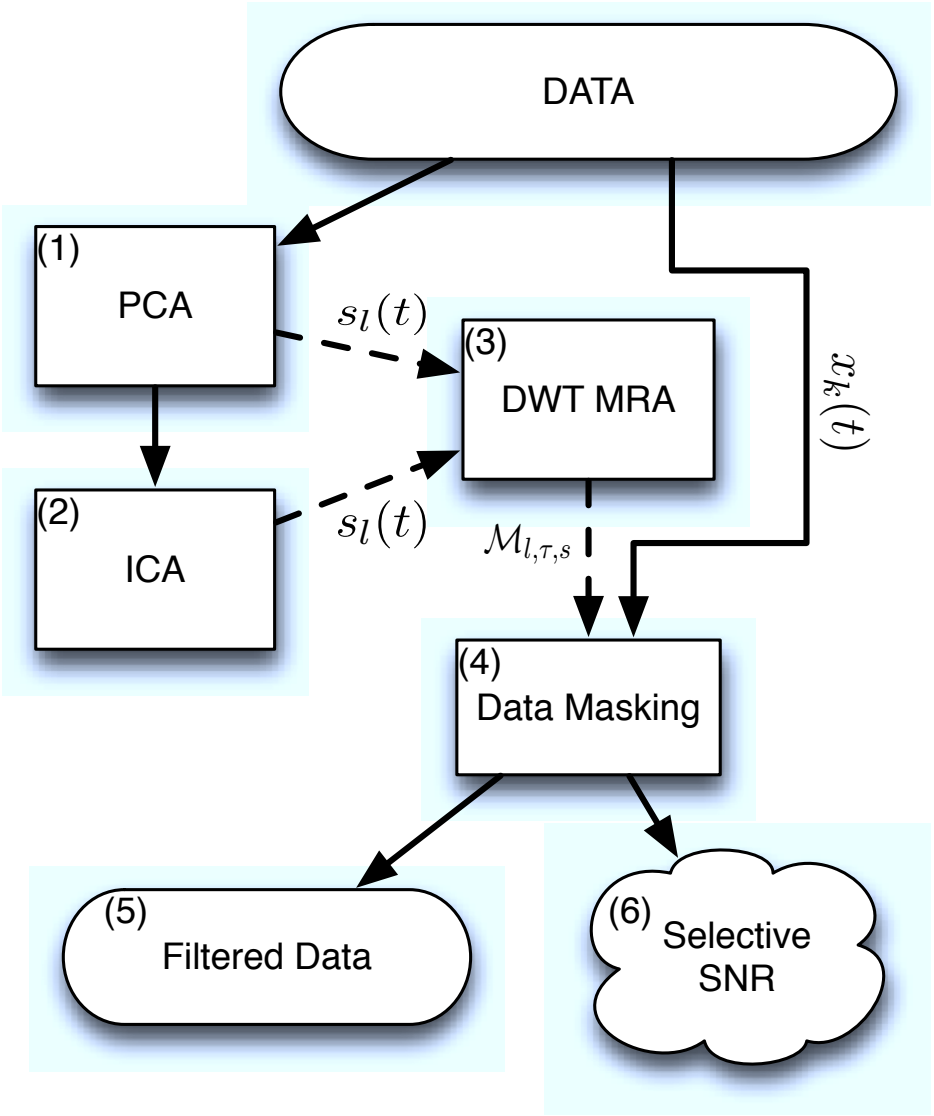


Figure 7.4: Flowchart of the proposed masking algorithm. The multivariate time series data is fed into the PCA/ICA deconvolution as described in chapter 3. Either the principal or independent components, both denoted by s_l are fed into the discrete wavelet transform multi-resolution analysis (DWT-MRA) to decompose the signals into their respective wavelet components, $\mathcal{L}_{l,\tau,s}$, from which the boolean mask, $\mathcal{M}_{l,\tau,s}$ is created given a certain threshold criterium $\mathcal{T}$. The mask is then combined with the individual time series data in (4) according to equation 7.7 to yield the filtered data and signal-to-noise (SNR) statistics.

compared to all other signals in the data set. Such a measure is invaluable in exploratory data analysis of highly convolved data sets.

7.4.1 To be done

What remains to be done in the development of this hybrid ICA/Wavelet algorithm is a detailed study of the thresholding criteria for different non-Gaussian sources and SNRs. This study will be done using simulations before applying the algorithm to real observed data to showcase its applicability.

7.5 The choice of wavelet

There are a great many different kinds of wavelets available to us ranging from 'mexican hat' shaped wavelets to 'symmetric', 'biorthogonal' and others. In these examples I will continue using wavelets from the Daubechies family as these contain the highest number of vanishing moments of any wavelet family, as mentioned earlier. It is nonetheless important to decide on the right 'mother' wavelet out of this family. In figures 7.1 & 7.2, I have used the Daubechies-4 (db4) wavelet as basis function to illustrate the different morphologies of various signals (e.g. primary transit to secondary eclipse). This wavelet (db4) has four coefficients and can hence exactly describe a second-order polynomial (as two coefficients are vanshing). This may not be an appropriate choice of coefficients however and for more 'rounded' shapes we may want to use higher order wavelets. In figure 7.3 I show the MRAs of three signals: secondary eclipse, primary transit and a sinusoid, and the respective MRAs for a db4 and a db8 basis. We can see the db8 decompositions being significantly sparser for rounder morphologies whilst being identical for the secondary eclipse case. In the secondary eclipse case, the additional coefficients are redundant but do not impair the decomposition whilst in the other cases they significantly increase the sparsity of the signal in wavelet space.

7.6 Comparing Fourier and Wavelet Analysis

Finally, I would like to return to the beginning and briefly discuss the major similarities and differences of Fourier Transforms (FT) and Wavelet Transform (WT).

As we have seen, wavelets are very closely related to FT and can in many cases be seen as an extension of the underlying concepts of FT. Both transforms decompose a given time-domain signal into its respective frequencies. The FT uses sine and cosine functions as basis functions, whereas the WT uses short impulses with a given morphology. Given these short

impulses, the WT has an additional spatial 'awareness' of the morphology of the signal in the time series, as opposed to the FT. In chapters 5 & 6 we predominantly use FT rather than WT for the determination of the eclipse depth. Given the more sophisticated nature of the WT this seems contradictory to intuition. Indeed, it is possible to use WT in the same way we used FT to self-filter the signal, however the higher complexity of WTs makes this a difficult task. In the FT case, the simplicity of the model allows us to analytically relate the transit depth parameter to the amplitude of the Fourier coefficients, equation 5.7. As these coefficients span the entirety of the time-domain, i.e. the sinusoid of the coefficient extends over the whole data-set, they capture the overall amplitude of the signal but not the signal's shape. On the other hand, wavelet coefficients contain spatial information and one would need to identify and average spatial out-of-transit and in-transit coefficients in the MRA decomposition. This is more difficult and more prone to error. In short, to determine a single amplitude parameter (the transit depth) from the data, a more complex model (such as WTs) does not provide additional information nor advantages.

For applications discussed in this chapter, which is based on the morphological identification of non-Gaussian signals in the frequency domain, the use of wavelets is more than ideal. Here FT would not be appropriate as it is much harder to obtain an accurate spatial awareness with FT than it is using wavelets.

7.7 Conclusion

In this chapter I have discussed a natural extension of the work presented in the previous chapters of this thesis. In chapters 3 and 5 we presented two algorithms for exoplanetary data analysis. The first algorithm is based on signal component separation using independent component analysis at its core, whilst the other algorithm is based on frequency domain techniques. As discussed in this chapter and showcased in previous chapters, both algorithms have clear advantages but are both limited by specific properties of the data, namely signal-to-noise ratios. Here, we presented a possible pathway to a hybrid algorithm combining all the advantages of component separation with the robustness of frequency domain analysis, to form a universally applicable algorithm with improved and novel properties. The continuation of such a study is paramount in the light of ever fainter and more complex ground-based data in the future.

Chapter 8

Summary and Conclusion

At the heart of any exoplanetary spectroscopy lies the data and any new insight into an exoplanetary atmosphere stands or falls with the quality of the result at hand. With 1 count in 10000 recorded on the detector coming from the exoplanetary atmosphere, the difficulty of filtering for these faint signals does not need to be elaborated upon. Furthermore, until the implementation of dedicated missions such as *EChO* or *FINESS* in six to twelve years time, we are limited to the use of general observatories and instruments that often do not feature an instrument calibration plan at the required level of accuracy.

Calibrating an instrument without knowing its instrument response function at the required level, has in fact become the central challenge of exoplanetary spectroscopy. Until now, most groups have tried to characterise the instrument response function using seemingly arbitrary instrument models based on derived instrumental state vectors. With different groups employing different instrument models and deriving their state vectors differently, it has become apparent that in most cases the scientific results are completely dependent on the instrument correction used. Using parametric methods, it is hard to perform an unbiased analysis of the data at the 10^{-4} level of accuracy and it is these parametric approaches that are the underlying cause for a plethora of controversy in the field.

In this thesis I worked towards unbiased, non-parametric methods to de-trend the observed data. As a result of these efforts, I have developed two de-trending algorithms for high signal-to-noise ratio (SNR) data and very low SNR data:

High SNR data

In the case of high to medium SNR data as often encountered in space-based measurement, we find ourselves often limited by the non-Gaussian instrumental systematics rather than the

photon-noise contribution. Based on the concepts of un-supervised machine learning and independent component analysis, I postulated a data de-trending algorithm in **Chapter 3** capable of 'learning' individual instrumental systematics and to construct an instrument noise-model without auxiliary nor prior knowledge of the instrument itself. Such an approach ensures the highest degree of objectivity in one's data analysis. In **Chapter 4**, I tested the proposed algorithm on various space-based data sets and find excellent agreement with previous results. This not only demonstrates the validity of such an approach but also ends controversies regarding the data sets analysed. I proceeded to extend our applications to single photometric time-series and demonstrated the ability of the algorithm to also effectively dampen or remove time-correlated stellar variability in photometric time-series. Considering the difficulty of modelling and removing stellar noise, we can assert this to be a critical result of the method.

Low SNR data

Ground-based observations often feature SNRs at the order of unity per resolution element of the spectrograph. These observations feature a too high Gaussian noise contribution for component separation to be an efficient means of de-trending but are also too contaminated by non-Gaussian noise to allow for more conventional binning approaches. Here we require an unbiased, non-parametric approach with a high degree of robustness in very poor SNR conditions. In **Chapter 5** I discussed the sparsity of lightcurves in the frequency domain and used the Fourier properties of lightcurves to build a data self-filtering mechanism. The robustness of this technique was tested using simulations of very low SNR data and determined the fundamental limit of this method to be a SNR ~ 0.2. Furthermore, I discussed an enhancement of the Fourier techniques by dampening the Gaussian noise present in the data by the use of wavelet based multi-resolution analysis and coefficient thresholding.

I used this robust Fourier and Wavelet based algorithm to analyse four observations of the secondary eclipse of HD189733b in the K- and L-bands obtained from the ground-based IRTF telescope in **Chapter 6**. Using the combination of these four nights and the algorithms developed, I was able to unambiguously detect a strong emission feature in the L-band which we associated with a non-LTE emission of the methane ν_3 band. Residual telluric contaminations were extensively tested for and I presented a tentative non-LTE methane model for the observations at hand.

Combining both algorithms

I conclude this work, in **Chapter 7**, by discussing the complementary natures of both algorithms and proposed an implementation of ICA and more general component separation algorithms to low SNR data via optimal multi-resolution masks in Wavelet space. Whilst the component separation, based on ICA and/or PCA, is good for feature extraction, it does not provide information on the features' amplitudes. On the other hand, the Fourier based approach provides a high degree of robustness in determining signal amplitudes but does not recover the features' morphologies well. Combining both algorithms via wavelet masks allows the determination of both, morphology and amplitude, and forms a natural extension of the wavelet de-noising formalism described by Donoho (1995).

Finally, what remains to be done, is to conclude this work by remaining faithful to the thesis' title (please see the next page for a cocktail recipe) and to bear in mind that ...

"What we call the beginning is often the end.
And to make an end is to make a beginning.
The end is where we start from."
— T. S. Eliot

Lexington Lemonade

2 measures	Maker's Mark Bourbon
1 measure	Grand Marnier or Cointreau
1	Orange
1/2	Lemon
2 - 3 drops	Angostura Bitters
	Ice

Serve in tall glass, shake Bourbon and Cointreau with ice and add the freshly squeezed orange and lemon juice. Top off with sparkling water or lemonade, add bitters, garnish with sliced orange and cocktail umbrella.

Bibliography

E. Agol, N. B. Cowan, H. A. Knutson, D. Deming, J. H. Steffen, G. W. Henry, and D. Charbonneau. The Climate of HD 189733b from Fourteen Transits and Eclipses Measured by Spitzer. *ApJ*, 721:1861–1877, October 2010. doi: 10.1088/0004-637X/721/2/1861.

S. Aigrain, F. Pont, and S. Zucker. A simple method to estimate radial velocity variations due to stellar activity using photometry. *MNRAS*, page 1898, November 2011. doi: 10.1111/j.1365-2966.2011.19960.x.

J. R. P. Angel, A. Y. S. Cheng, and N. J. Woolf. A space telescope for infrared spectroscopy of earth-like planets. *Nature*, 322:341–343, July 1986. doi: 10.1038/322341a0.

J. F. Appleby. CH4 nonlocal thermodynamic equilibrium in the atmospheres of the giant planets. *ICARUS*, 85:355–379, June 1990. doi: 10.1016/0019-1035(90)90123-Q.

J. Aumont and J. F. Macías-Pérez. Blind component separation for polarized observations of the cosmic microwave background. *MNRAS*, 376:739–758, April 2007. doi: 10.1111/j.1365-2966.2007.11470.x.

A. Baglin, M. Auvergne, and L. Boisnard. 36th cospar scientific assembly. In *COSPAR, Plenary COSPAR Plenary meeting*, volume 36, page 3749, 2006.

G. Bakos. Planet Hunting with HATNet and HATSouth. In *American Astronomical Society Meeting Abstracts 218*, page 103.04, May 2011.

G. Bakos, R. W. Noyes, G. Kovács, K. Z. Stanek, D. D. Sasselov, and I. Domsa. Wide-Field Millimagnitude Photometry with the HAT: A Tool for Extrasolar Planet Detection. *PASP*, 116:266–277, March 2004. doi: 10.1086/382735.

G. Á. Bakos, J. Lázár, I. Papp, P. Sári, and E. M. Green. System Description and First Light Curves of the Hungarian Automated Telescope, an Autonomous Observatory for Variability Search. *PASP*, 114:974–987, September 2002. doi: 10.1086/342382.

G. Á. Bakos, H. Knutson, F. Pont, C. Moutou, D. Charbonneau, A. Shporer, F. Bouchy, M. Everett, C. Hergenrother, D. W. Latham, M. Mayor, T. Mazeh, R. W. Noyes, D. Queloz, A. Pál, and S. Udry. Refined Parameters of the Planet Orbiting HD 189733. *ApJ*, 650:1160–1171, October 2006. doi: 10.1086/506316.

G. Á. Bakos, R. W. Noyes, G. Kovács, D. W. Latham, D. D. Sasselov, G. Torres, D. A. Fischer, R. P. Stefanik, B. Sato, J. A. Johnson, A. Pál, G. W. Marcy, R. P. Butler, G. A. Esquerdo, K. Z. Stanek, J. Lázár, I. Papp, P. Sári, and B. Sipőcz. HAT-P-1b: A Large-Radius, Low-Density Exoplanet Transiting One Member of a Stellar Binary. *ApJ*, 656:552–559, February 2007. doi: 10.1086/509874.

P. Ballerini, G. Micela, A. F. Lanza, and I. Pagano. Multiwavelength flux variations induced by stellar magnetic activity: effects on planetary transits. 01 2012. URL `http://arxiv.org/abs/1201.3514v1`.

G. E. Ballester, D. K. Sing, and F. Herbert. The signature of hot hydrogen in the atmosphere of the extrasolar planet HD 209458b. *Nature*, 445:511–514, February 2007. doi: 10.1038/nature05525.

R. J. Barber, J. Tennyson, G. J. Harris, and R. N. Tolchenov. A high-accuracy computed water line list. *MNRAS*, 368:1087–1094, May 2006. doi: 10.1111/j.1365-2966.2006.10184.x.

P. Barge, A. Baglin, M. Auvergne, H. Rauer, A. Léger, J. Schneider, F. Pont, S. Aigrain, J.-M. Almenara, R. Alonso, M. Barbieri, P. Bordé, F. Bouchy, H. J. Deeg, D. La Reza, M. Deleuil, R. Dvorak, A. Erikson, M. Fridlund, M. Gillon, P. Gondoin, T. Guillot, A. Hatzes, G. Hebrard, L. Jorda, P. Kabath, H. Lammer, A. Llebaria, B. Loeillet, P. Magain, T. Mazeh, C. Moutou, M. Ollivier, M. Pätzold, D. Queloz, D. Rouan, A. Shporer, and G. Wuchterl. Transiting exoplanets from the CoRoT space mission. I. CoRoT-Exo-1b: a low-density short-period planet around a G0V star. *A&A*, 482:L17–L20, May 2008. doi: 10.1051/0004-6361:200809353.

T. Barman. Identification of Absorption Features in an Extrasolar Planet Atmosphere. *ApJL*, 661:L191–L194, June 2007. doi: 10.1086/518736.

M. Barthélemy, J. Lilensten, and C. Parkinson. H_2 vibrational temperatures in the upper atmosphere of Jupiter. *A&A*, 437:329–331, July 2005. doi: 10.1051/0004-6361:20040257.

J. Bean. Extrasolar planets: Homing in on another Earth. *Nature*, 478:41–42, October 2011. doi: 10.1038/nature10578.

J. L. Bean, E. Miller-Ricci Kempton, and D. Homeier. A ground-based transmission spectrum of the super-Earth exoplanet GJ 1214b. *Nature*, 468:669–672, December 2010. doi: 10.1038/nature09596.

J. L. Bean, J.-M. Désert, P. Kabath, B. Stalder, S. Seager, E. Miller-Ricci Kempton, Z. K. Berta, D. Homeier, S. Walsh, and A. Seifahrt. The Optical and Near-infrared Transmission Spectrum of the Super-Earth GJ1214b: Further Evidence for a Metal-rich Atmosphere. *ApJ*, 743:92, December 2011. doi: 10.1088/0004-637X/743/1/92.

J.-P. Beaulieu, D. P. Bennett, P. Fouqué, A. Williams, M. Dominik, U. G. Jørgensen, D. Kubas, A. Cassan, C. Coutures, J. Greenhill, K. Hill, J. Menzies, P. D. Sackett, M. Albrow, S. Brillant, J. A. R. Caldwell, J. J. Calitz, K. H. Cook, E. Corrales, M. Desort, S. Dieters, D. Dominis, J. Donatowicz, M. Hoffman, S. Kane, J.-B. Marquette, R. Martin, P. Meintjes, K. Pollard, K. Sahu, C. Vinter, J. Wambsganss, K. Woller, K. Horne, I. Steele, D. M. Bramich, M. Burgdorf, C. Snodgrass, M. Bode, A. Udalski, M. K. Szymański, M. Kubiak, T. Więckowski, G. Pietrzyński, I. Soszyński, O. Szewczyk, Ł. Wyrzykowski, B. Paczyński, F. Abe, I. A. Bond, T. R. Britton, A. C. Gilmore, J. B. Hearnshaw, Y. Itow, K. Kamiya, P. M. Kilmartin, A. V. Korpela, K. Masuda, Y. Matsubara, M. Motomura, Y. Muraki, S. Nakamura, C. Okada, K. Ohnishi, N. J. Rattenbury, T. Sako, S. Sato, M. Sasaki, T. Sekiguchi, D. J. Sullivan, P. J. Tristram, P. C. M. Yock, and T. Yoshioka. Discovery of a cool planet of 5.5 Earth masses through gravitational microlensing. *Nature*, 439:437–440, January 2006. doi: 10.1038/nature04441.

J. P. Beaulieu, S. Carey, I. Ribas, and G. Tinetti. Primary Transit of the Planet HD 189733b at 3.6 and 5.8 μm. *ApJ*, 677:1343–1347, April 2008. doi: 10.1086/527045.

J. P. Beaulieu, D. M. Kipping, V. Batista, G. Tinetti, I. Ribas, S. Carey, J. A. Noriega-Crespo, C. A. Griffith, G. Campanella, S. Dong, J. Tennyson, R. J. Barber, P. Deroo, S. J. Fossey, D. Liang, M. R. Swain, Y. Yung, and N. Allard. Water in the atmosphere of HD 209458b from 3.6-8 μm IRAC photometric observations in primary transit. *MNRAS*, 409:963–974, December 2010. doi: 10.1111/j.1365-2966.2010.16516.x.

J.-P. Beaulieu, G. Tinetti, D. M. Kipping, I. Ribas, R. J. Barber, J. Y.-K. Cho, I. Polichtchouk, J. Tennyson, S. N. Yurchenko, C. A. Griffith, V. Batista, I. Waldmann, S. Miller, S. Carey, O. Mousis, S. J. Fossey, and A. Aylward. Methane in the Atmosphere of the Transiting Hot Neptune GJ436B? *ApJ*, 731:16, April 2011. doi: 10.1088/0004-637X/731/1/16.

A. Belouchrani, K. Abed-Meraim, J. F. Cardoso, and E. Moulines. *IEEE Trans. Signal Processing*, 45(434), 1997.

L. Ben-Jaffel. Spectral, Spatial, and Time Properties of the Hydrogen Nebula around Exoplanet HD 209458b. *ApJ*, 688:1352–1360, December 2008. doi: 10.1086/592101.

Z. K. Berta, D. Charbonneau, J.-M. Désert, E. Miller-Ricci Kempton, P. R. McCullough, C. J. Burke, J. J. Fortney, J. Irwin, P. Nutzman, and D. Homeier. The Flat Transmission Spectrum of the Super-Earth GJ1214b from Wide Field Camera 3 on the Hubble Space Telescope. *ApJ*, 747:35, March 2012. doi: 10.1088/0004-637X/747/1/35.

I. Boisse, F. Bouchy, G. Hébrard, X. Bonfils, N. Santos, and S. Vauclair. Disentangling between stellar activity and planetary signals. *A&A*, 528:A4, April 2011. doi: 10.1051/0004-6361/201014354.

I. A. Bond, A. Udalski, M. Jaroszyński, N. J. Rattenbury, B. Paczyński, I. Soszyński, L. Wyrzykowski, M. K. Szymański, M. Kubiak, O. Szewczyk, K. Żebruń, G. Pietrzyński, F. Abe, D. P. Bennett, S. Eguchi, Y. Furuta, J. B. Hearnshaw, K. Kamiya, P. M. Kilmartin, Y. Kurata, K. Masuda, Y. Matsubara, Y. Muraki, S. Noda, K. Okajima, T. Sako, T. Sekiguchi, D. J. Sullivan, T. Sumi, P. J. Tristram, T. Yanagisawa, P. C. M. Yock, and OGLE Collaboration. OGLE 2003-BLG-235/MOA 2003-BLG-53: A Planetary Microlensing Event. *ApJL*, 606:L155–L158, May 2004. doi: 10.1086/420928.

P. Bordé, D. Rouan, and A. Léger. Exoplanet detection capability of the COROT space mission. *A&A*, 405:1137–1144, July 2003. doi: 10.1051/0004-6361:20030675.

W. J. Borucki, E. W. Dunham, D. G. Koch, W. D. Cochran, J. D. Rose, D. K. Cullers, A. Granados, and J. M. Jenkins. FRESIP: A Mission to Determine the Character and Frequency of Extra-Solar Planets Around Solar-Like Stars. *APSS*, 241:111–134, March 1996. doi: 10.1007/BF00644220.

W. J. Borucki, D. Koch, J. Jenkins, D. Sasselov, R. Gilliland, N. Batalha, D. W. Latham, D. Caldwell, G. Basri, T. Brown, J. Christensen-Dalsgaard, W. D. Cochran, E. DeVore, E. Dunham, A. K. Dupree, T. Gautier, J. Geary, A. Gould, S. Howell, H. Kjeldsen, J. Lissauer, G. Marcy, S. Meibom, D. Morrison, and J. Tarter. Keplers Optical Phase Curve of the Exoplanet HAT-P-7b. *Science*, 325:709–, August 2009. doi: 10.1126/science.1178312.

W. J. Borucki, D. Koch, G. Basri, N. Batalha, T. Brown, D. Caldwell, J. Caldwell, J. Christensen-Dalsgaard, W. D. Cochran, E. DeVore, E. W. Dunham, A. K. Dupree, T. N. Gautier, J. C.

Geary, R. Gilliland, A. Gould, S. B. Howell, J. M. Jenkins, Y. Kondo, D. W. Latham, G. W. Marcy, S. Meibom, H. Kjeldsen, J. J. Lissauer, D. G. Monet, D. Morrison, D. Sasselov, J. Tarter, A. Boss, D. Brownlee, T. Owen, D. Buzasi, D. Charbonneau, L. Doyle, J. Fortney, E. B. Ford, M. J. Holman, S. Seager, J. H. Steffen, W. F. Welsh, J. Rowe, H. Anderson, L. Buchhave, D. Ciardi, L. Walkowicz, W. Sherry, E. Horch, H. Isaacson, M. E. Everett, D. Fischer, G. Torres, J. A. Johnson, M. Endl, P. MacQueen, S. T. Bryson, J. Dotson, M. Haas, J. Kolodziejczak, J. Van Cleve, H. Chandrasekaran, J. D. Twicken, E. V. Quintana, B. D. Clarke, C. Allen, J. Li, H. Wu, P. Tenenbaum, E. Verner, F. Bruhweiler, J. Barnes, and A. Prsa. Kepler Planet-Detection Mission: Introduction and First Results. *Science*, 327:977–, February 2010. doi: 10.1126/science.1185402.

W. J. Borucki, D. G. Koch, G. Basri, N. Batalha, T. M. Brown, S. T. Bryson, D. Caldwell, J. Christensen-Dalsgaard, W. D. Cochran, E. DeVore, E. W. Dunham, T. N. Gautier, III, J. C. Geary, R. Gilliland, A. Gould, S. B. Howell, J. M. Jenkins, D. W. Latham, J. J. Lissauer, G. W. Marcy, J. Rowe, D. Sasselov, A. Boss, D. Charbonneau, D. Ciardi, L. Doyle, A. K. Dupree, E. B. Ford, J. Fortney, M. J. Holman, S. Seager, J. H. Steffen, J. Tarter, W. F. Welsh, C. Allen, L. A. Buchhave, J. L. Christiansen, B. D. Clarke, S. Das, J.-M. Désert, M. Endl, D. Fabrycky, F. Fressin, M. Haas, E. Horch, A. Howard, H. Isaacson, H. Kjeldsen, J. Kolodziejczak, C. Kulesa, J. Li, P. W. Lucas, P. Machalek, D. McCarthy, P. MacQueen, S. Meibom, T. Miquel, A. Prsa, S. N. Quinn, E. V. Quintana, D. Ragozzine, W. Sherry, A. Shporer, P. Tenenbaum, G. Torres, J. D. Twicken, J. Van Cleve, L. Walkowicz, F. C. Witteborn, and M. Still. Characteristics of Planetary Candidates Observed by Kepler. II. Analysis of the First Four Months of Data. *ApJ*, 736:19, July 2011. doi: 10.1088/0004-637X/736/1/19.

F. Bouchy, S. Udry, M. Mayor, C. Moutou, F. Pont, N. Iribarne, R. da Silva, S. Ilovaisky, D. Queloz, N. C. Santos, D. Ségransan, and S. Zucker. ELODIE metallicity-biased search for transiting Hot Jupiters. II. A very hot Jupiter transiting the bright K star HD 189733. *A&A*, 444:L15–L19, December 2005. doi: 10.1051/0004-6361:200500201.

L. Brillouin. *Journal of Applied Physics*, 24(1152), 1953.

P. J. Brockwell and R. A. Davis. *Time Series: Theory and Methods*. Springer, 2006.

H. Bruntt, M. Deleuil, M. Fridlund, R. Alonso, F. Bouchy, A. Hatzes, M. Mayor, C. Moutou, and D. Queloz. Improved stellar parameters of CoRoT-7. A star hosting two super Earths. *A&A*, 519:A51, September 2010. doi: 10.1051/0004-6361/201014143.

C. J. Burke, P. R. McCullough, J. A. Valenti, C. M. Johns-Krull, K. A. Janes, J. N. Heasley, F. J. Summers, J. E. Stys, R. Bissinger, M. L. Fleenor, C. N. Foote, E. García-Melendo, B. L. Gary, P. J. Howell, F. Mallia, G. Masi, B. Taylor, and T. Vanmunster. XO-2b: Transiting Hot Jupiter in a Metal-rich Common Proper Motion Binary. *ApJ*, 671:2115–2128, December 2007. doi: 10.1086/523087.

C. J. Burke, P. R. McCullough, L. E. Bergeron, D. Long, R. L. Gilliland, E. P. Nelan, C. M. Johns-Krull, J. A. Valenti, and K. A. Janes. NICMOS Observations of the Transiting Hot Jupiter XO-1b. *ApJ*, 719:1796–1806, August 2010. doi: 10.1088/0004-637X/719/2/1796.

A. Burrows, E. Rauscher, D. S. Spiegel, and K. Menou. Photometric and Spectral Signatures of Three-dimensional Models of Transiting Giant Exoplanets. *ApJ*, 719:341–350, August 2010. doi: 10.1088/0004-637X/719/1/341.

D. A. Caldwell, J. J. Kolodziejczak, J. E. Van Cleve, J. M. Jenkins, P. R. Gazis, V. S. Argabright, E. E. Bachtell, E. W. Dunham, J. C. Geary, R. L. Gilliland, H. Chandrasekaran, J. Li, P. Tenenbaum, H. Wu, W. J. Borucki, S. T. Bryson, J. L. Dotson, M. R. Haas, and D. G. Koch. Instrument Performance in Kepler's First Months. *ApJL*, 713:L92–L96, April 2010. doi: 10.1088/2041-8205/713/2/L92.

J. A. Carter and J. N. Winn. Parameter Estimation from Time-series Data with Correlated Errors: A Wavelet-based Method and its Application to Transit Light Curves. *ApJ*, 704: 51–67, October 2009. doi: 10.1088/0004-637X/704/1/51.

J. A. Carter and J. N. Winn. The Detectability of Transit Depth Variations Due to Exoplanetary Oblateness and Spin Precession. *ApJ*, 716:850–856, June 2010. doi: 10.1088/0004-637X/716/1/850.

J. A. Carter, J. C. Yee, J. Eastman, B. S. Gaudi, and J. N. Winn. Analytic Approximations for Transit Light-Curve Observables, Uncertainties, and Covariances. *ApJ*, 689:499–512, December 2008. doi: 10.1086/592321.

A. Cassan, D. Kubas, J.-P. Beaulieu, M. Dominik, K. Horne, J. Greenhill, J. Wambsganss, J. Menzies, A. Williams, U. G. Jørgensen, A. Udalski, D. P. Bennett, M. D. Albrow, V. Batista, S. Brillant, J. A. R. Caldwell, A. Cole, C. Coutures, K. H. Cook, S. Dieters, D. D. Prester, J. Donatowicz, P. Fouqué, K. Hill, N. Kains, S. Kane, J.-B. Marquette, R. Martin, K. R. Pollard, K. C. Sahu, C. Vinter, D. Warren, B. Watson, M. Zub, T. Sumi, M. K. Szymański, M. Kubiak, R. Poleski, I. Soszynski, K. Ulaczyk, G. Pietrzyński, and Ł. Wyrzykowski. One or

more bound planets per Milky Way star from microlensing observations. *Nature*, 481:167–169, January 2012. doi: 10.1038/nature10684.

D. Charbonneau, T. M. Brown, D. W. Latham, and M. Mayor. Detection of Planetary Transits Across a Sun-like Star. *ApJl*, 529:L45–L48, January 2000. doi: 10.1086/312457.

D. Charbonneau, T. M. Brown, R. W. Noyes, and R. L. Gilliland. Detection of an Extrasolar Planet Atmosphere. *ApJ*, 568:377–384, March 2002. doi: 10.1086/338770.

D. Charbonneau, L. E. Allen, S. T. Megeath, G. Torres, R. Alonso, T. M. Brown, R. L. Gilliland, D. W. Latham, G. Mandushev, F. T. O'Donovan, and A. Sozzetti. Detection of Thermal Emission from an Extrasolar Planet. *ApJ*, 626:523–529, June 2005. doi: 10.1086/429991.

D. Charbonneau, H. A. Knutson, T. Barman, L. E. Allen, M. Mayor, S. T. Megeath, D. Queloz, and S. Udry. The Broadband Infrared Emission Spectrum of the Exoplanet HD 189733b. *ApJ*, 686:1341–1348, October 2008. doi: 10.1086/591635.

D. Charbonneau, Z. K. Berta, J. Irwin, C. J. Burke, P. Nutzman, L. A. Buchhave, C. Lovis, X. Bonfils, D. W. Latham, S. Udry, R. A. Murray-Clay, M. J. Holman, E. E. Falco, J. N. Winn, D. Queloz, F. Pepe, M. Mayor, X. Delfosse, and T. Forveille. A super-Earth transiting a nearby low-mass star. *Nature*, 462:891–894, December 2009. doi: 10.1038/nature08679.

G. Chauvin, A.-M. Lagrange, C. Dumas, B. Zuckerman, D. Mouillet, I. Song, J.-L. Beuzit, and P. Lowrance. A giant planet candidate near a young brown dwarf. Direct VLT/NACO observations using IR wavefront sensing. *A&A*, 425:L29–L32, October 2004. doi: 10.1051/0004-6361:200400056.

A. Cichocki and S. Amari. *Adaptive Blind Signal and Image Processing*. John Wiley & Sons Inc., 2002.

A. Claret. A new non-linear limb-darkening law for LTE stellar atmosphere models. Calculations for -5.0 log[M/H] +1, 2000 K = T_{eff} = 50000 K at several surface gravities. *A&A*, 363:1081–1190, November 2000.

A. Collier Cameron, D. M. Wilson, R. G. West, L. Hebb, X.-B. Wang, S. Aigrain, F. Bouchy, D. J. Christian, W. I. Clarkson, B. Enoch, M. Esposito, E. Guenther, C. A. Haswell, G. Hébrard, C. Hellier, K. Horne, J. Irwin, S. R. Kane, B. Loeillet, T. A. Lister, P. Maxted, M. Mayor, C. Moutou, N. Parley, D. Pollacco, F. Pont, D. Queloz, R. Ryans, I. Skillen, R. A. Street, S. Udry, and P. J. Wheatley. Efficient identification of exoplanetary transit candidates from SuperWASP light curves. *MNRAS*, 380:1230–1244, September 2007. doi: 10.1111/j.1365-2966.2007.12195.x.

P. Comon. *Signal Processing*, 36(287), 1994.

P. Comon and C. Jutten. *Handbook of Blind Source Separation*. Academic Press, 2010.

D. R. Coulter. NASA's Terrestrial Planet Finder mission: the search for habitable planets. In M. Fridlund, T. Henning, & H. Lacoste, editor, *Earths: DARWIN/TPF and the Search for Extrasolar Terrestrial Planets*, volume 539 of *ESA Special Publication*, pages 47–54, October 2003.

M. Cover, T. and J. A. Thomas. *Elements of Information Theory*. Wiley & Sons Inc., 2 edition, 2006.

B. Croll, L. Albert, R. Jayawardhana, E. Miller-Ricci Kempton, J. J. Fortney, N. Murray, and H. Neilson. Broadband Transmission Spectroscopy of the Super-Earth GJ 1214b Suggests a Low Mean Molecular Weight Atmosphere. *ApJ*, 736:78, August 2011. doi: 10.1088/0004-637X/736/2/78.

M. C. Cushing, W. D. Vacca, and J. T. Rayner. Spextool: A Spectral Extraction Package for SpeX, a 0.8-5.5 Micron Cross-Dispersed Spectrograph. *PASP*, 116:362–376, April 2004. doi: 10.1086/382907.

S. Czesla, K. F. Huber, U. Wolter, S. Schröter, and J. H. M. M. Schmitt. How stellar activity affects the size estimates of extrasolar planets. *A&A*, 505:1277–1282, October 2009. doi: 10.1051/0004-6361/200912454.

C. Danielski, G. Tinetti, M. R. Swain, P. Deroo, and I. P. Waldmann. *in prep.*, 2012.

I. Daubechies. *Commun. Pure Appl. Math.*, 41(909), 1988.

I. Daubechies. *Ten Lectures on Wavelets*. Soc. for Industrial Mathematics, 1992.

A. C. Davison. *Statistical Models*. Cambridge University Press, 2009.

J. Delabrouille, J.-F. Cardoso, and G. Patanchon. Multidetector multicomponent spectral matching and applications for cosmic microwave background data analysis. *MNRAS*, 346:1089–1102, December 2003. doi: 10.1111/j.1365-2966.2003.07069.x.

D. Deming, L. J. Richardson, and J. Harrington. 3.8-μm photometry during the secondary eclipse of the extrasolar planet HD209458b. *MNRAS*, 378:148–152, June 2007. doi: 10.1111/j.1365-2966.2007.11754.x.

J.-M. Désert, A. Lecavelier des Etangs, G. Hébrard, D. K. Sing, D. Ehrenreich, R. Ferlet, and A. Vidal-Madjar. Search for Carbon Monoxide in the Atmosphere of the Transiting Exoplanet HD 189733b. *ApJ*, 699:478–485, July 2009. doi: 10.1088/0004-637X/699/1/478.

J.-M. Désert, J. Bean, E. Miller-Ricci Kempton, Z. K. Berta, D. Charbonneau, J. Irwin, J. Fortney, C. J. Burke, and P. Nutzman. Observational Evidence for a Metal-rich Atmosphere on the Super-Earth GJ1214b. *ApJL*, 731:L40, April 2011. doi: 10.1088/2041-8205/731/2/L40.

D.L. Donoho. *IEE Trans. on Inf. Theory*, 41(3):613, 1995.

L. Doyennette, F. Menard-Bourcin, J. Menard, C. Boursier, and C. Camy-Peyret. *Phys.chem.A.*, 102(3849), 1998.

P. Drossart, T. Fouchet, J. Crovisier, E. Lellouch, T. Encrenaz, H. Feuchtgruber, and J. P. Champion. Fluorescence in the 3μm bands of methane on Jupiter and Saturn from ISO/SWS observations. In P. Cox & M. Kessler, editor, *The Universe as Seen by ISO*, volume 427 of *ESA Special Publication*, page 169, March 1999.

E. W. Dunham, G. I. Mandushev, B. W. Taylor, and B. Oetiker. PSST: The Planet Search Survey Telescope. *PASP*, 116:1072–1080, November 2004. doi: 10.1086/426303.

E. B. Ford. Improving the Efficiency of Markov Chain Monte Carlo for Analyzing the Orbits of Extrasolar Planets. *ApJ*, 642:505–522, May 2006. doi: 10.1086/500802.

T. Forveille, X. Bonfils, X. Delfosse, R. Alonso, S. Udry, F. Bouchy, M. Gillon, C. Lovis, V. Neves, M. Mayor, F. Pepe, D. Queloz, N. C. Santos, D. Segransan, J. M. Almenara, H. Deeg, and M. Rabus. The HARPS search for southern extra-solar planets XXXII. Only 4 planets in the Gl˜581 system. *ArXiv e-prints*, September 2011.

S. J. Fossey, I. P. Waldmann, and D. M. Kipping. Detection of a transit by the planetary companion of HD 80606. *MNRAS*, 396:L16–L20, June 2009. doi: 10.1111/j.1745-3933.2009.00653.x.

F. Fressin, G. Torres, J. F. Rowe, D. Charbonneau, L. A. Rogers, S. Ballard, N. M. Batalha, W. J. Borucki, S. T. Bryson, L. A. Buchhave, D. R. Ciardi, J.-M. Désert, C. D. Dressing, D. C. Fabrycky, E. B. Ford, T. N. Gautier, III, C. E. Henze, M. J. Holman, A. Howard, S. B. Howell, J. M. Jenkins, D. G. Koch, D. W. Latham, J. J. Lissauer, G. W. Marcy, S. N. Quinn, D. Ragozzine, D. D. Sasselov, S. Seager, T. Barclay, F. Mullally, S. E. Seader, M. Still, J. D. Twicken, S. E. Thompson, and K. Uddin. Two Earth-sized planets orbiting Kepler-20. *Nature*, 482:195–198, February 2012. doi: 10.1038/nature10780.

J. H. Friedman. *Journal of the American Statistical Association*, 82(249), 1987.

T. N. Gautier, III, D. Charbonneau, J. F. Rowe, G. W. Marcy, H. Isaacson, G. Torres, F. Fressin, L. A. Rogers, J.-M. Désert, L. A. Buchhave, D. W. Latham, S. N. Quinn, D. R. Ciardi, D. C. Fabrycky, E. B. Ford, R. L. Gilliland, L. M. Walkowicz, S. T. Bryson, W. D. Cochran, M. Endl, D. A. Fischer, S. B. Howel, E. P. Horch, T. Barclay, N. Batalha, W. J. Borucki, J. L. Christiansen, J. C. Geary, C. E. Henze, M. J. Holman, K. Ibrahim, J. M. Jenkins, K. Kinemuchi, D. G. Koch, J. J. Lissauer, D. T. Sanderfer, D. D. Sasselov, S. Seager, K. Silverio, J. C. Smith, M. Still, M. C. Stumpe, P. Tenenbaum, and J. Van Cleve. Kepler-20: A Sun-like Star with Three Sub-Neptune Exoplanets and Two Earth-size Candidates. *ArXiv e-prints*, December 2011.

N. P. Gibson, F. Pont, and S. Aigrain. A new look at NICMOS transmission spectroscopy of HD 189733, GJ-436 and XO-1: no conclusive evidence for molecular features. *MNRAS*, 411: 2199–2213, March 2011. doi: 10.1111/j.1365-2966.2010.17837.x.

N. P. Gibson, S. Aigrain, S. Roberts, T. M. Evans, M. Osborne, and F. Pont. A Gaussian process framework for modelling instrumental systematics: application to transmission spectroscopy. *MNRAS*, 419:2683–2694, January 2012. doi: 10.1111/j.1365-2966.2011.19915.x.

M. Gillon, A. A. Lanotte, T. Barman, N. Miller, B.-O. Demory, M. Deleuil, J. Montalbán, F. Bouchy, A. Collier Cameron, H. J. Deeg, J. J. Fortney, M. Fridlund, J. Harrington, P. Magain, C. Moutou, D. Queloz, H. Rauer, D. Rouan, and J. Schneider. The thermal emission of the young and massive planet CoRoT-2b at 4.5 and 8 μm. *A&A*, 511:A3, February 2010. doi: 10.1051/0004-6361/200913507.

A. Giménez. Equations for the analysis of the light curves of extra-solar planetary transits. *A&A*, 450:1231–1237, May 2006. doi: 10.1051/0004-6361:20054445.

P. C. Gregory. Bayesian exoplanet tests of a new method for MCMC sampling in highly correlated model parameter spaces. *MNRAS*, 410:94–110, January 2011a. doi: 10.1111/j.1365-2966.2010.17428.x.

P. C. Gregory. Bayesian re-analysis of the Gliese 581 exoplanet system. *MNRAS*, 415:2523–2545, August 2011b. doi: 10.1111/j.1365-2966.2011.18877.x.

C. J. Grillmair, D. Charbonneau, A. Burrows, L. Armus, J. Stauffer, V. Meadows, J. Van Cleve, and D. Levine. A Spitzer Spectrum of the Exoplanet HD 189733b. *ApJL*, 658:L115–L118, April 2007. doi: 10.1086/513741.

C. J. Grillmair, A. Burrows, D. Charbonneau, L. Armus, J. Stauffer, V. Meadows, J. van Cleve, K. von Braun, and D. Levine. Strong water absorption in the dayside emission spectrum of the planet HD189733b. *Nature*, 456:767–769, December 2008. doi: 10.1038/nature07574.

H. H. Harman. *Modern Factor Analysis*. University of Chicago Press, 2 edition, 1967.

A. P. Hatzes, M. Fridlund, G. Nachmani, T. Mazeh, D. Valencia, G. Hébrard, L. Carone, M. Pätzold, S. Udry, F. Bouchy, M. Deleuil, C. Moutou, P. Barge, P. Bordé, H. Deeg, B. Tingley, R. Dvorak, D. Gandolfi, S. Ferraz-Mello, G. Wuchterl, E. Guenther, T. Guillot, H. Rauer, A. Erikson, J. Cabrera, S. Csizmadia, A. Léger, H. Lammer, J. Weingrill, D. Queloz, R. Alonso, D. Rouan, and J. Schneider. The Mass of CoRoT-7b. *ApJ*, 743:75, December 2011. doi: 10.1088/0004-637X/743/1/75.

G. W. Henry, G. W. Marcy, R. P. Butler, and S. S. Vogt. A Transiting "51 Peg-like" Planet. *ApJ*, 529:L41–L44, January 2000. doi: 10.1086/312458.

Gregory W. Henry, Andrew W. Howard, Geoffrey W. Marcy, Debra A. Fischer, and John Asher Johnson. Detection of a transiting low-density super-earth. *ArXiv e-prints*, 09 2011. URL `http://arxiv.org/abs/1109.2549v1`.

P. J. Huber. *The Annals of Statistics*, 13(435), 1985.

A. Hyv̈arinen. Fast and robust fixed-point algorithms for independent component analysis. *Neural Networks, IEEE Transactions on*, 10(3):626 –634, may 1999. ISSN 1045-9227. doi: 10.1109/72.761722.

A. Hyvärinen and E. Oja. *Neural Networks*, 13(411), 2000.

A. Hyvärinen and P. Pajunen. *Neural Networks*, 12(429), 1999.

A. Hyvärinen, J. Karhunen, and Oja E. *Independent Component Analysis*. Wiley & Sons Inc., 2001.

J. M. Jenkins, D. A. Caldwell, H. Chandrasekaran, J. D. Twicken, S. T. Bryson, E. V. Quintana, B. D. Clarke, J. Li, C. Allen, P. Tenenbaum, H. Wu, T. C. Klaus, C. K. Middour, M. T. Cote, S. McCauliff, F. R. Girouard, J. P. Gunter, B. Wohler, J. Sommers, J. R. Hall, A. K. Uddin, M. S. Wu, P. A. Bhavsar, J. Van Cleve, D. L. Pletcher, J. A. Dotson, M. R. Haas, R. L. Gilliland, D. G. Koch, and W. J. Borucki. Overview of the Kepler Science Processing Pipeline. *ApJL*, 713:L87–L91, April 2010. doi: 10.1088/2041-8205/713/2/L87.

A. G. Jensen, S. Redfield, M. Endl, W. D. Cochran, L. Koesterke, and T. S. Barman. A Detection Of H-alpha In An Exoplanetary Exosphere. *ArXiv e-prints*, March 2012.

I. T. Jolliffe. *Principal Component Analysis*. Springer, 2 edition, 2002.

P. Kalas, J. R. Graham, E. Chiang, M. P. Fitzgerald, M. Clampin, E. S. Kite, K. Stapelfeldt, C. Marois, and J. Krist. Optical Images of an Exosolar Planet 25 Light-Years from Earth. *Science*, 322:1345–, November 2008. doi: 10.1126/science.1166609.

D. M. Kipping. Transiting planets - light-curve analysis for eccentric orbits. *MNRAS*, 389: 1383–1390, September 2008. doi: 10.1111/j.1365-2966.2008.13658.x.

D. M. Kipping. Transit timing effects due to an exomoon - II. *MNRAS*, 396:1797–1804, July 2009. doi: 10.1111/j.1365-2966.2009.14869.x.

D. M. Kipping. Investigations of approximate expressions for the transit duration. *MNRAS*, 407:301–313, September 2010. doi: 10.1111/j.1365-2966.2010.16894.x.

D. M. Kipping. LUNA: an algorithm for generating dynamic planet-moon transits. *MNRAS*, 416:689–709, September 2011. doi: 10.1111/j.1365-2966.2011.19086.x.

D. M. Kipping and G. Tinetti. Nightside pollution of exoplanet transit depths. *MNRAS*, 407: 2589–2598, October 2010. doi: 10.1111/j.1365-2966.2010.17094.x.

H. A. Knutson, D. Charbonneau, L. E. Allen, J. J. Fortney, E. Agol, N. B. Cowan, A. P. Showman, C. S. Cooper, and S. T. Megeath. A map of the day-night contrast of the extrasolar planet HD 189733b. *Nature*, 447:183–186, May 2007a. doi: 10.1038/nature05782.

H. A. Knutson, D. Charbonneau, R. W. Noyes, T. M. Brown, and R. L. Gilliland. Using Stellar Limb-Darkening to Refine the Properties of HD 209458b. *ApJ*, 655:564–575, January 2007b. doi: 10.1086/510111.

H. A. Knutson, D. Charbonneau, L. E. Allen, A. Burrows, and S. T. Megeath. The 3.6-8.0 μm Broadband Emission Spectrum of HD 209458b: Evidence for an Atmospheric Temperature Inversion. *ApJ*, 673:526–531, January 2008. doi: 10.1086/523894.

H. A. Knutson, N. Madhusudhan, N. B. Cowan, J. L. Christiansen, E. Agol, D. Deming, J.-M. Désert, D. Charbonneau, G. W. Henry, D. Homeier, J. Langton, G. Laughlin, and S. Seager. A Spitzer Transmission Spectrum for the Exoplanet GJ 436b, Evidence for Stellar Variability, and Constraints on Dayside Flux Variations. *ApJ*, 735:27, July 2011. doi: 10.1088/0004-637X/735/1/27.

D. G. Koch, W. J. Borucki, G. Basri, N. M. Batalha, T. M. Brown, D. Caldwell, J. Christensen-Dalsgaard, W. D. Cochran, E. DeVore, E. W. Dunham, T. N. Gautier, III, J. C. Geary,

R. L. Gilliland, A. Gould, J. Jenkins, Y. Kondo, D. W. Latham, J. J. Lissauer, G. Marcy, D. Monet, D. Sasselov, A. Boss, D. Brownlee, J. Caldwell, A. K. Dupree, S. B. Howell, H. Kjeldsen, S. Meibom, D. Morrison, T. Owen, H. Reitsema, J. Tarter, S. T. Bryson, J. L. Dotson, P. Gazis, M. R. Haas, J. Kolodziejczak, J. F. Rowe, J. E. Van Cleve, C. Allen, H. Chandrasekaran, B. D. Clarke, J. Li, E. V. Quintana, P. Tenenbaum, J. D. Twicken, and H. Wu. Kepler Mission Design, Realized Photometric Performance, and Early Science. *ApJL*, 713:L79, April 2010. doi: 10.1088/2041-8205/713/2/L79.

Z. Koldovsky, P. Tichavsky, and E. Oja. Efficient variant of algorithm fastica for independent component analysis attaining the cramer-rao lower bound. *Neural Networks, IEEE Transactions on*, 17(5):1265 –1277, sept. 2006. ISSN 1045-9227. doi: 10.1109/TNN.2006.875991.

S. Kullback and R. A. Leibler. *Annals of Mathematical Statistics*, 22(79), 1951.

A. Léger, J. M. Mariotti, B. Mennesson, M. Ollivier, J. L. Puget, D. Rouan, and J. Schneider. The DARWIN project. *Ap&SS*, 241:135–146, March 1996. doi: 10.1007/BF00644221.

A. Léger, D. Rouan, J. Schneider, P. Barge, M. Fridlund, B. Samuel, M. Ollivier, E. Guenther, M. Deleuil, H. J. Deeg, M. Auvergne, R. Alonso, S. Aigrain, A. Alapini, J. M. Almenara, A. Baglin, M. Barbieri, H. Bruntt, P. Bordé, F. Bouchy, J. Cabrera, C. Catala, L. Carone, S. Carpano, S. Csizmadia, R. Dvorak, A. Erikson, S. Ferraz-Mello, B. Foing, F. Fressin, D. Gandolfi, M. Gillon, P. Gondoin, O. Grasset, T. Guillot, A. Hatzes, G. Hébrard, L. Jorda, H. Lammer, A. Llebaria, B. Loeillet, M. Mayor, T. Mazeh, C. Moutou, M. Pätzold, F. Pont, D. Queloz, H. Rauer, S. Renner, R. Samadi, A. Shporer, C. Sotin, B. Tingley, G. Wuchterl, M. Adda, P. Agogu, T. Appourchaux, H. Ballans, P. Baron, T. Beaufort, R. Bellenger, R. Berlin, P. Bernardi, D. Blouin, F. Baudin, P. Bodin, L. Boisnard, L. Boit, F. Bonneau, S. Borzeix, R. Briet, J.-T. Buey, B. Butler, D. Cailleau, R. Cautain, P.-Y. Chabaud, S. Chaintreuil, F. Chiavassa, V. Costes, V. Cuna Parrho, F. de Oliveira Fialho, M. Decaudin, J.-M. Defise, S. Djalal, G. Epstein, G.-E. Exil, C. Fauré, T. Fenouillet, A. Gaboriaud, A. Gallic, P. Gamet, P. Gavalda, E. Grolleau, R. Gruneisen, L. Gueguen, V. Guis, V. Guivarc'h, P. Guterman, D. Hallouard, J. Hasiba, F. Heuripeau, G. Huntzinger, H. Hustaix, C. Imad, C. Imbert, B. Johlander, M. Jouret, P. Journoud, F. Karioty, L. Kerjean, V. Lafaille, L. Lafond, T. Lam-Trong, P. Landiech, V. Lapeyrere, T. Larqué, P. Laudet, N. Lautier, H. Lecann, L. Lefevre, B. Leruyet, P. Levacher, A. Magnan, E. Mazy, F. Mertens, J.-M. Mesnager, J.-C. Meunier, J.-P. Michel, W. Monjoin, D. Naudet, K. Nguyen-Kim, J.-L. Orcesi, H. Ottacher, R. Perez, G. Peter, P. Plasson, J.-Y. Plesseria, B. Pontet, A. Pradines, C. Quentin, J.-L. Reynaud, G. Rolland, F. Rollenhagen, R. Romagnan, N. Russ, R. Schmidt, N. Schwartz, I. Sebbag,

G. Sedes, H. Smit, M. B. Steller, W. Sunter, C. Surace, M. Tello, D. Tiphène, P. Toulouse, B. Ulmer, O. Vandermarcq, E. Vergnault, A. Vuillemin, and P. Zanatta. Transiting exoplanets from the CoRoT space mission. VIII. CoRoT-7b: the first super-Earth with measured radius. *A&A*, 506:287–302, October 2009. doi: 10.1051/0004-6361/200911933.

J. L. Linsky, H. Yang, K. France, C. S. Froning, J. C. Green, J. T. Stocke, and S. N. Osterman. Observations of Mass Loss from the Transiting Exoplanet HD 209458b. *ApJ*, 717:1291–1299, July 2010. doi: 10.1088/0004-637X/717/2/1291.

T. A. Lister, D. R. Anderson, and R. G. West. The Status of SuperWASP-South. In C. Afonso, D. Weldrake, & T. Henning, editor, *Transiting Extrapolar Planets Workshop*, volume 366 of *Astronomical Society of the Pacific Conference Series*, page 108, July 2007.

P. Lowell. In *Proc. Amer. Phil. Soc*, volume 42, page 364, 1903.

H. Lu, H. Zhou, J. Wang, T. Wang, X. Dong, Z. Zhuang, and C. Li. Ensemble Learning for Independent Component Analysis of Normal Galaxy Spectra. *AJ*, 131:790–805, February 2006. doi: 10.1086/498711.

N. Madhusudhan and S. Seager. A Temperature and Abundance Retrieval Method for Exoplanet Atmospheres. *ApJ*, 707:24–39, December 2009. doi: 10.1088/0004-637X/707/1/24.

D. Maino, A. Farusi, C. Baccigalupi, F. Perrotta, A. J. Banday, L. Bedini, C. Burigana, G. De Zotti, K. M. Górski, and E. Salerno. All-sky astrophysical component separation with Fast Independent Component Analysis (FASTICA). *MNRAS*, 334:53–68, July 2002. doi: 10.1046/j.1365-8711.2002.05425.x.

D. Maino, S. Donzelli, A. J. Banday, F. Stivoli, and C. Baccigalupi. Cosmic microwave background signal in Wilkinson Microwave Anisotropy Probe three-year data with FASTICA. *MNRAS*, 374:1207–1215, February 2007. doi: 10.1111/j.1365-2966.2006.11255.x.

C. Majeau, E. Agol, and N. B. Cowan. A two-dimensional infrared map of the extrasolar planet hd 189733b. *ApJ*, 02 2012. URL `http://arxiv.org/abs/1202.1883v1`.

K. Mandel and E. Agol. Analytic Light Curves for Planetary Transit Searches. *ApJl*, 580: L171–L175, December 2002. doi: 10.1086/345520.

A. M. Mandell, L. Drake Deming, G. A. Blake, H. A. Knutson, M. J. Mumma, G. L. Villanueva, and C. Salyk. Non-detection of L-band Line Emission from the Exoplanet HD189733b. *ApJ*, 728:18, February 2011. doi: 10.1088/0004-637X/728/1/18.

B. J. F. Manly. *Multivariate Statistical Methods - A Primer*. Chapman & Hall, 2 edition, 1994.

G. W. Marcy and R. P. Butler. A Planetary Companion to 70 Virginis. *ApJL*, 464:L147, June 1996. doi: 10.1086/310096.

M. Mayor and D. Queloz. A Jupiter-mass companion to a solar-type star. *Nature*, 378:355–359, November 1995. doi: 10.1038/378355a0.

M. Mayor, F. Pepe, D. Queloz, F. Bouchy, G. Rupprecht, G. Lo Curto, G. Avila, W. Benz, J.-L. Bertaux, X. Bonfils, T. Dall, H. Dekker, B. Delabre, W. Eckert, M. Fleury, A. Gilliotte, D. Gojak, J. C. Guzman, D. Kohler, J.-L. Lizon, A. Longinotti, C. Lovis, D. Megevand, L. Pasquini, J. Reyes, J.-P. Sivan, D. Sosnowska, R. Soto, S. Udry, A. van Kesteren, L. Weber, and U. Weilenmann. Setting New Standards with HARPS. *The Messenger*, 114:20–24, December 2003.

M. Mayor, X. Bonfils, T. Forveille, X. Delfosse, S. Udry, J.-L. Bertaux, H. Beust, F. Bouchy, C. Lovis, F. Pepe, C. Perrier, D. Queloz, and N. C. Santos. The HARPS search for southern extra-solar planets. XVIII. An Earth-mass planet in the GJ 581 planetary system. *A&A*, 507: 487–494, November 2009. doi: 10.1051/0004-6361/200912172.

P. R. McCullough, J. E. Stys, J. A. Valenti, S. W. Fleming, K. A. Janes, and J. N. Heasley. The XO Project: Searching for Transiting Extrasolar Planet Candidates. *PASP*, 117:783–795, August 2005. doi: 10.1086/432024.

P. S. Muirhead, J. A. Johnson, K. Apps, J. A. Carter, T. D. Morton, D. C. Fabrycky, J. S. Pineda, M. Bottom, B. Rojas-Ayala, E. Schlawin, K. Hamren, K. R. Covey, J. R. Crepp, K. G. Stassun, J. Pepper, L. Hebb, E. N. Kirby, A. W. Howard, H. T. Isaacson, G. W. Marcy, D. Levitan, T. Diaz-Santos, L. Armus, and J. P. Lloyd. Characterizing the Cool KOIs. III. KOI 961: A Small Star with Large Proper Motion and Three Small Planets. *ApJ*, 747:144, March 2012. doi: 10.1088/0004-637X/747/2/144.

E. A. Nadaraya. *Probability & its Applications*, 10(186), 1964.

D. Naef, D. W. Latham, M. Mayor, T. Mazeh, J. L. Beuzit, G. A. Drukier, C. Perrier-Bellet, D. Queloz, J. P. Sivan, G. Torres, S. Udry, and S. Zucker. Hd 80606 b, a planet on an extremely elongated orbit ". *Astronomy and Astrophysics*, 375(2):4, 2001. doi: 10.1051/0004-6361: 20010853.

J.A. Nelder and R. Mead. *Computer Journ.*, 7:308, 1965.

P. Nutzman and D. Charbonneau. Design Considerations for a Ground-Based Transit Search for Habitable Planets Orbiting M Dwarfs. *PASP*, 120:317–327, March 2008. doi: 10.1086/533420.

E. Oja. Principal components, minor components, and linear neural networks. *Neural Networks*, 5(6):927 – 935, 1992. ISSN 0893-6080. doi: 10.1016/S0893-6080(05)80089-9. URL `http://www.sciencedirect.com/science/article/pii/S0893608005800899`.

S. D. Pagiatakis, H. Yin, and M. A. El-Gelil. *Phys. Earth and Planet. Interiors*, 160:108, 2007.

A. Pál. Properties of analytic transit light-curve models. *MNRAS*, 390:281–288, October 2008. doi: 10.1111/j.1365-2966.2008.13723.x.

K. Pearson. *Phil. Mag.*, 2(559), 1901.

F. Pepe, M. Mayor, G. Rupprecht, G. Avila, P. Ballester, J.-L. Beckers, W. Benz, J.-L. Bertaux, F. Bouchy, B. Buzzoni, C. Cavadore, S. Deiries, H. Dekker, B. Delabre, S. D'Odorico, W. Eckert, J. Fischer, M. Fleury, M. George, A. Gilliotte, D. Gojak, J.-C. Guzman, F. Koch, D. Kohler, H. Kotzlowski, D. Lacroix, J. Le Merrer, J.-L. Lizon, G. Lo Curto, A. Longinotti, D. Megevand, L. Pasquini, P. Petitpas, M. Pichard, D. Queloz, J. Reyes, P. Richaud, J.-P. Sivan, D. Sosnowska, R. Soto, S. Udry, E. Ureta, A. van Kesteren, L. Weber, U. Weilenmann, A. Wicenec, G. Wieland, J. Christensen-Dalsgaard, D. Dravins, A. Hatzes, M. Kürster, F. Paresce, and A. Penny. HARPS: ESO's coming planet searcher. Chasing exoplanets with the La Silla 3.6-m telescope. *The Messenger*, 110:9–14, December 2002.

D. B. Percival and A. T. Walden. *Wavelet Methods for Time Series Analysis*. Cambridge University Press, 2000.

W. H. Pickering. An Explanation of the Martian and Lunar Canals. *Popular Astronomy*, 12: 439–442, August 1904.

D. L. Pollacco, I. Skillen, A. C. Cameron, D. J. Christian, C. Hellier, J. Irwin, T. A. Lister, R. A. Street, R. G. West, D. Anderson, W. I. Clarkson, H. Deeg, B. Enoch, A. Evans, A. Fitzsimmons, C. A. Haswell, S. Hodgkin, K. Horne, S. R. Kane, F. P. Keenan, P. F. L. Maxted, A. J. Norton, J. Osborne, N. R. Parley, R. S. I. Ryans, B. Smalley, P. J. Wheatley, and D. M. Wilson. The WASP Project and the SuperWASP Cameras. *PASP*, 118:1407–1418, October 2006. doi: 10.1086/508556.

F. Pont, H. Knutson, R. L. Gilliland, C. Moutou, and D. Charbonneau. Detection of atmospheric haze on an extrasolar planet: the 0.55-1.05 μm transmission spectrum of HD 189733b with the HubbleSpaceTelescope. *MNRAS*, 385:109–118, March 2008. doi: 10.1111/j.1365-2966.2008.12852.x.

F. Pont, S. Aigrain, and S. Zucker. Reassessing the radial-velocity evidence for planets around CoRoT-7. *MNRAS*, 411:1953–1962, March 2011. doi: 10.1111/j.1365-2966.2010.17823.x.

W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipes*. Cambridge University Press, 3 edition, 2007.

J. T. Rayner, D. W. Toomey, P. M. Onaka, A. J. Denault, W. E. Stahlberger, W. D. Vacca, M. C. Cushing, and S. Wang. SpeX: A Medium-Resolution 0.8-5.5 Micron Spectrograph and Imager for the NASA Infrared Telescope Facility. *PASP*, 115:362–382, March 2003. doi: 10.1086/367745.

S. Redfield, M. Endl, W. D. Cochran, and L. Koesterke. Sodium Absorption from the Exoplanetary Atmosphere of HD 189733b Detected in the Optical Transmission Spectrum. *ApJl*, 673: L87–L90, January 2008. doi: 10.1086/527475.

L. J. Richardson, D. Deming, K. Horning, S. Seager, and J. Harrington. A spectrum of an extrasolar planet. *Nature*, 445:892–895, February 2007. doi: 10.1038/nature05636.

K. F. Riley, M. P. Hobson, and S. J. Bence. *Mathematical Methods For Physics and Engineering*. Cambridge University Press, 2 edition, 2002.

L. S. Rothman, I. E. Gordon, A. Barbe, D. C. Benner, P. F. Bernath, M. Birk, V. Boudon, L. R. Brown, A. Campargue, J.-P. Champion, K. Chance, L. H. Coudert, V. Dana, V. M. Devi, S. Fally, J.-M. Flaud, R. R. Gamache, A. Goldman, D. Jacquemart, I. Kleiner, N. Lacome, W. J. Lafferty, J.-Y. Mandin, S. T. Massie, S. N. Mikhailenko, C. E. Miller, N. Moazzen-Ahmadi, O. V. Naumenko, A. V. Nikitin, J. Orphal, V. I. Perevalov, A. Perrin, A. Predoi-Cross, C. P. Rinsland, M. Rotger, M. Šimečková, M. A. H. Smith, K. Sung, S. A. Tashkun, J. Tennyson, R. A. Toth, A. C. Vandaele, and J. Vander Auwera. The hitran 2008 molecular spectroscopic database. *JQRST*, 110(533), 2009.

S. Sardy. *IEE Trans. on Sig. Proces*, 48:1023, 2000.

J. Schneider, C. Dedieu, P. Le Sidaner, R. Savalle, and I. Zolotukhin. Defining and cataloging exoplanets: the exoplanet.eu database. *A&A*, 532:A79, August 2011. doi: 10.1051/0004-6361/201116713.

S. Seager and G. Mallén-Ornelas. A Unique Solution of Planet and Star Parameters from an Extrasolar Planet Transit Light Curve. *ApJ*, 585:1038–1055, March 2003. doi: 10.1086/346105.

C. Shannon. *Bell System Tech. Journal*, 27(379), 1948.

D. K. Sing, F. Pont, S. Aigrain, D. Charbonneau, J.-M. Désert, N. Gibson, R. Gilliland, W. Hayek, G. Henry, H. Knutson, A. Lecavelier Des Etangs, T. Mazeh, and A. Shporer. Hubble Space Telescope transmission spectroscopy of the exoplanet HD 189733b: high-altitude atmospheric haze in the optical and near-ultraviolet with STIS. *MNRAS*, 416:1443–1455, September 2011. doi: 10.1111/j.1365-2966.2011.19142.x.

I. A. G. Snellen, S. Albrecht, E. J. W. de Mooij, and R. S. Le Poole. Ground-based detection of sodium in the transmission spectrum of exoplanet HD 209458b. *A&A*, 487:357–362, August 2008. doi: 10.1051/0004-6361:200809762.

I. A. G. Snellen, R. J. de Kok, E. J. W. de Mooij, and S. Albrecht. The orbital motion, absolute mass and high-altitude winds of exoplanet HD209458b. *Nature*, 465:1049–1051, June 2010a. doi: 10.1038/nature09111.

I. A. G. Snellen, E. J. W. de Mooij, and A. Burrows. Bright optical day-side emission from extrasolar planet CoRoT-2b. *A&A*, 513:A76, April 2010b. doi: 10.1051/0004-6361/200913338.

J. Southworth. Homogeneous studies of transiting extrasolar planets - I. Light-curve analyses. *MNRAS*, 386:1644–1666, May 2008. doi: 10.1111/j.1365-2966.2008.13145.x.

J. Southworth, P. J. Wheatley, and G. Sams. A method for the direct determination of the surface gravities of transiting extrasolar planets. *MNRAS*, 379:L11–L15, July 2007. doi: 10.1111/j.1745-3933.2007.00324.x.

A. Sozzetti. The Gaia astrometric survey. *Highlights of Astronomy*, 15:716–717, November 2010. doi: 10.1017/S1743921310011142.

K. R. Stapelfeldt, E. K. Holmes, C. Chen, G. H. Rieke, K. Y. L. Su, D. C. Hines, M. W. Werner, C. A. Beichman, M. Jura, D. L. Padgett, J. A. Stansberry, G. Bendo, J. Cadien, M. Marengo, T. Thompson, T. Velusamy, C. Backus, M. Blaylock, E. Egami, C. W. Engelbracht, D. T. Frayer, K. D. Gordon, J. Keene, W. B. Latter, T. Megeath, K. Misselt, J. E. Morrison, J. Muzerolle, A. Noriega-Crespo, J. Van Cleve, and E. T. Young. First Look at the Fomalhaut Debris Disk with the Spitzer Space Telescope. *ApJS*, 154:458–462, September 2004. doi: 10.1086/423135.

C.M. Stein. *Ann.Statist.*, 9(6):1135, 1981.

K. B. Stevenson, J. Harrington, S. Nymeyer, N. Madhusudhan, S. Seager, W. C. Bowman, R. A. Hardy, D. Deming, E. Rauscher, and N. B. Lust. Possible thermochemical disequilibrium in the atmosphere of the exoplanet GJ 436b. *Nature*, 464:1161–1164, April 2010. doi: 10.1038/nature09013.

F. Stivoli, C. Baccigalupi, D. Maino, and R. Stompor. Separating polarized cosmological and galactic emissions for cosmic microwave background B-mode polarization experiments. *MNRAS*, 372:615–629, October 2006. doi: 10.1111/j.1365-2966.2006.10769.x.

J. V. Stone. *Independent Component Analysis: A Tutorial Introduction*. A Bradford Book, 2004.

M. R. Swain. Finesse - A New Mission Concept For Exoplanet Spectroscopy. In *AAS Division for Planetary Sciences Meeting Abstracts 42*, volume 42 of *Bulletin of the American Astronomical Society*, page 1064, October 2010.

M. R. Swain, J. Bouwman, R. L. Akeson, S. Lawler, and C. A. Beichman. The Mid-Infrared Spectrum of the Transiting Exoplanet HD 209458b. *ApJ*, 674:482–497, February 2008a. doi: 10.1086/523832.

M. R. Swain, J. Bouwman, R. L. Akeson, S. Lawler, and C. A. Beichman. The Mid-Infrared Spectrum of the Transiting Exoplanet HD 209458b. *ApJ*, 674:482–497, February 2008b. doi: 10.1086/523832.

M. R. Swain, G. Vasisht, and G. Tinetti. The presence of methane in the atmosphere of an extrasolar planet. *Nature*, 452:329–331, March 2008c. doi: 10.1038/nature06823.

M. R. Swain, G. Tinetti, G. Vasisht, P. Deroo, C. Griffith, J. Bouwman, P. Chen, Y. Yung, A. Burrows, L. R. Brown, J. Matthews, J. F. Rowe, R. Kuschnig, and D. Angerhausen. Water, Methane, and Carbon Dioxide Present in the Dayside Spectrum of the Exoplanet HD 209458b. *ApJ*, 704:1616–1621, October 2009a. doi: 10.1088/0004-637X/704/2/1616.

M. R. Swain, G. Vasisht, G. Tinetti, J. Bouwman, P. Chen, Y. Yung, D. Deming, and P. Deroo. Molecular Signatures in the Near-Infrared Dayside Spectrum of HD 189733b. *ApJl*, 690: L114–L117, January 2009b. doi: 10.1088/0004-637X/690/2/L114.

M. R. Swain, P. Deroo, C. A. Griffith, G. Tinetti, A. Thatte, G. Vasisht, P. Chen, J. Bouwman, I. J. Crossfield, D. Angerhausen, C. Afonso, and T. Henning. A ground-based near-infrared emission spectrum of the exoplanet HD189733b. *Nature*, 463:637–639, February 2010. doi: 10.1038/nature08775.

M. R. Swain, P. Deroo, and G. Vasisht. Nicmos spectroscopy of hd 189733b. In A. Sozzetti, editor, *IAU Symposium*, volume 276, page 148, 2011.

P. Tenenbaum, J. L. Christiansen, J. M. Jenkins, J. F. Rowe, S. Seader, D. A. Caldwell, B. D. Clarke, J. Li, E. V. Quintana, J. C. Smith, M. C. Stumpe, S. E. Thompson, J. D. Twicken, J. Van Cleve, W. J. Borucki, M. T. Cote, M. R. Haas, D. T. Sanderfer, F. R. Girouard, T. C.

Klaus, C. K. Middour, B. Wohler, N. M. Batalha, T. Barclay, and J. E. Nickerson. Detection of Potential Transit Signals in the First Three Quarters of Kepler Mission Data. *ApJS*, 199: 24, March 2012. doi: 10.1088/0067-0049/199/1/24.

M. Tessenyi, M. Ollivier, G. Tinetti, J. P. Beaulieu, V. Coudé du Foresto, T. Encrenaz, G. Micela, B. Swinyard, I. Ribas, A. Aylward, J. Tennyson, M. R. Swain, A. Sozzetti, G. Vasisht, and P. Deroo. Characterizing the Atmospheres of Transiting Planets with a Dedicated Space Telescope. *ApJ*, 746:45, February 2012. doi: 10.1088/0004-637X/746/1/45.

A. Thatte, P. Deroo, and M. R. Swain. Selective principal component extraction and reconstruction: a novel method for ground based exoplanet spectroscopy. *A&A*, 523:A35, November 2010. doi: 10.1051/0004-6361/201015148.

J. Thiévin, R. Georges, S. Carles, A. Benidar, B. Rowe, and J.-P. Champion. High-temperature emission spectroscopy of methane. *JQSRT*, 109:2027–2036, July 2008. doi: 10.1016/j.jqsrt.2008.01.023.

P. Tichavský, E. Doron, A. Yeredor, and J. Nielsen. In *Proc. EUSIPCO-2006*, 2006.

P. Tichavsky, Z. Koldovsky, and E. Oja. Performance analysis of the fastica algorithm and crame rao bounds for linear independent component analysis. *Signal Processing, IEEE Transactions on*, 54(4):1189 – 1203, april 2006. ISSN 1053-587X. doi: 10.1109/TSP.2006.870561.

G. Tinetti, V. S. Meadows, D. Crisp, W. Fong , T. Velusamy, and H. Snively. Disk-Averaged Synthetic Spectra of Mars. *Astrobiology*, 5:461–482, August 2005. doi: 10.1089/ast.2005.5.461.

G. Tinetti, V. S. Meadows, D. Crisp, W. Fong, E. Fishbein, M. Turnbull, and J.-P. Bibring. Detectability of Planetary Characteristics in Disk-Averaged Spectra. I: The Earth Model. *Astrobiology*, 6:34–47, March 2006. doi: 10.1089/ast.2006.6.34.

G. Tinetti, A. Vidal-Madjar, M.-C. Liang, J.-P. Beaulieu, Y. Yung, S. Carey, R. J. Barber, J. Tennyson, I. Ribas, N. Allard, G. E. Ballester, D. K. Sing, and F. Selsis. Water vapour in the atmosphere of a transiting extrasolar planet. *Nature*, 448:169–171, July 2007. doi: 10.1038/nature06002.

G. Tinetti, P. Deroo, M. R. Swain, C. A. Griffith, G. Vasisht, L. R. Brown, C. Burke, and P. McCullough. Probing the Terminator Region Atmosphere of the Hot-Jupiter XO-1b with Transmission Spectroscopy. *ApJl*, 712:L139–L142, April 2010. doi: 10.1088/2041-8205/712/2/L139.

G. Tinetti, J. P. Beaulieu, T. Henning, M. Meyer, G. Micela, I. Ribas, D. Stam, M. Swain, O. Krause, M. Ollivier, E. Pace, B. Swinyard, A. Aylward, R. van Boekel, A. Coradini, T. Encrenaz, I. Snellen, M. R. Zapatero-Osorio, J. Bouwman, J. Y-K. Cho, V. Coudé du Foresto, T. Guillot, M. Lopez-Morales, I. Mueller-Wodarg, E. Palle, F. Selsis, A. Sozzetti, P. A. R. Ade, N. Achilleos, A. Adriani, C. B. Agnor, C. Afonso, C. Allende Prieto, G. Bakos, R. J. Barber, M. Barlow, P. Bernath, B. Bezard, P. Bordé, L. R. Brown, A. Cassan, C. Cavarroc, A. Ciaravella, C. O. U. Cockell, A. Coustenis, C. Danielski, L. Decin, R. De Kok, O. Demangeon, P. Deroo, P. Doel, P. Drossart, L. N. Fletcher, M. Focardi, F. Forget, S. Fossey, P. Fouqué, J. Frith, M. Galand, P. Gaulme, J. I. González Hernández, O. Grasset, D. Grassi, J. L. Grenfell, M. J. Griffin, C. A. Griffith, U. Grözinger, M. Guedel, P. Guio, O. Hainaut, R. Hargreaves, P. H. Hauschildt, K. Heng, D. Heyrovsky, R. Hueso, P. Irwin, L. Kaltenegger, P. Kervella, D. Kipping, T. T. Koskinen, G. Kovács, A. La Barbera, H. Lammer, E. Lellouch, G. Leto, M. Lopez Morales, M. A. Lopez Valverde, M. Lopez-Puertas, C. Lovis, A. Maggio, J. P. Maillard, J. Maldonado Prado, J. B. Marquette, F. J. Martin-Torres, P. Maxted, S. Miller, S. Molinari, D. Montes, A. Moro-Martin, J. I. Moses, O. Mousis, N. Nguyen Tuong, R. Nelson, G. S. Orton, E. Pantin, E. Pascale, S. Pezzuto, D. Pinfield, E. Poretti, R. Prinja, L. Prisinzano, J. M. Rees, A. Reiners, B. Samuel, A. Sanchez-Lavega, J. Sanz Forcada, D. Sasselov, G. Savini, B. Sicardy, A. Smith, L. Stixrude, G. Strazzulla, J. Tennyson, M. Tessenyi, G. Vasisht, S. Vinatier, S. Viti, I. Waldmann, G. J. White, T. Widemann, R. Wordsworth, R. Yelle, Y. Yung, and S. N. Yurchenko. EChO - Exoplanet Characterisation Observatory. *ArXiv e-prints*, December 2011a.

G. Tinetti, J. Y.-K. Cho, C. A. Griffith, O. Grasset, L. Grenfell, T. Guillot, T. T. Koskinen, J. I. Moses, D. Pinfield, J. Tennyson, M. Tessenyi, R. Wordsworth, A. Aylward, R. van Boekel, A. Coradini, T. Encrenaz, I. Snellen, M. R. Zapatero-Osorio, J. Bouwman, V. C. du Foresto, M. Lopez-Morales, I. Mueller-Wodarg, E. Pallé, F. Selsis, A. Sozzetti, J.-P. Beaulieu, T. Henning, M. Meyer, G. Micela, I. Ribas, D. Stam, M. Swain, O. Krause, M. Ollivier, E. Pace, B. Swinyard, P. A. R. Ade, N. Achilleos, A. Adriani, C. B. Agnor, C. Afonso, C. A. Prieto, G. Bakos, R. J. Barber, M. Barlow, P. Bernath, B. Bézard, P. Bordé, L. R. Brown, A. Cassan, C. Cavarroc, A. Ciaravella, C. Cockell, A. Coustenis, C. Danielski, L. Decin, R. De Kok, O. Demangeon, P. Deroo, P. Doel, P. Drossart, L. N. Fletcher, M. Focardi, F. Forget, S. Fossey, P. Fouqué, J. Frith, M. Galand, P. Gaulme, J. I. G. Hernández, D. Grassi, M. J. Griffin, U. Grözinger, M. Guedel, P. Guio, O. Hainaut, R. Hargreaves, P. H. Hauschildt, K. Heng, D. Heyrovsky, R. Hueso, P. Irwin, L. Kaltenegger, P. Kervella, D. Kipping, G. Kovacs, A. L. Barbera, H. Lammer, E. Lellouch, G. Leto, M. L. Morales, M. A. L. Valverde, M. Lopez-

Puertas, C. Lovi, A. Maggio, J.-P. Maillard, J. M. Prado, J.-B. Marquette, F. J. Martin-Torres, P. Maxted, S. Miller, S. Molinari, D. Montes, A. Moro-Martin, O. Mousis, N. N. Tuong, R. Nelson, G. S. Orton, E. Pantin, E. Pascale, S. Pezzuto, E. Poretti, R. Prinja, L. Prisinzano, J.-M. Réess, A. Reiners, B. Samuel, J. S. Forcada, D. Sasselov, G. Savini, B. Sicardy, A. Smith, L. Stixrude, G. Strazzulla, G. Vasisht, S. Vinatier, S. Viti, I. Waldmann, G. J. White, T. Widemann, R. Yelle, Y. Yung, and S. Yurchenko. The science of EChO. In A. Sozzetti, M. G. Lattanzi, & A. P. Boss, editor, *IAU Symposium*, volume 276 of *IAU Symposium*, pages 359–370, November 2011b. doi: 10.1017/S1743921311020448.

G. Tinetti, J. Tennyson, C. A. Griffith, and I. P. Waldmann. Water in exoplanets. *Philosophical Transactions A*, 2012.

G. Torres, J. N. Winn, and M. J. Holman. Improved Parameters for Extrasolar Transiting Planets. *ApJ*, 677:1324–1342, April 2008. doi: 10.1086/529429.

S. Udry and N.C. Santos. Statistical Properties of Exoplanets. *ARA and A*, 45:397–439, September 2007. doi: 10.1146/annurev.astro.45.051806.110529.

A. Vidal-Madjar, A. Lecavelier des Etangs, J.-M. Désert, G. E. Ballester, R. Ferlet, G. Hébrard, and M. Mayor. An extended upper atmosphere around the extrasolar planet HD209458b. *Nature*, 422:143–146, March 2003. doi: 10.1038/nature01448.

I. P. Waldmann. Of "Cocktail Parties" and Exoplanets. *ApJ*, 747:12, March 2012. doi: 10.1088/0004-637X/747/1/12.

I. P. Waldmann, G. Tinetti, and P. Deroo. Blind extraction of an exoplanetary spectrum: Independent component analysis applied to hd189733b. *ApJ*, submitted, 2012a.

I. P. Waldmann, G. Tinetti, P. Drossart, M. R. Swain, P. Deroo, and C. A. Griffith. Ground-based Near-infrared Emission Spectroscopy of HD 189733b. *ApJ*, 744:35, January 2012b. doi: 10.1088/0004-637X/744/1/35.

J. Wang, H. Xu, J. Gu, T. An, H. Cui, J. Li, Z. Zhang, Q. Zheng, and X.-P. Wu. How to Identify and Separate Bright Galaxy Clusters from the Low-frequency Radio Sky. *ApJ*, 723:620–633, November 2010. doi: 10.1088/0004-637X/723/1/620.

G.S Watson. *Sanky Series A*, 26(359), 1964.

J. N. Winn. Measuring accurate transit parameters. *IAU Symp. Proceedings*, July 2008.

J. N. Winn, M. J. Holman, G. W. Henry, A. Roussanova, K. Enya, Y. Yoshii, A. Shporer, T. Mazeh, J. A. Johnson, N. Narita, and Y. Suto. The Transit Light Curve Project. V.

System Parameters and Stellar Rotation Period of HD 189733. *AJ*, 133:1828–1835, April 2007. doi: 10.1086/512159.

J. N. Winn, J. A. Johnson, S. Albrecht, A. W. Howard, G. W. Marcy, I. J. Crossfield, and M. J. Holman. HAT-P-7: A Retrograde or Polar Orbit, and a Third Body. *ApJL*, 703:L99–L103, October 2009. doi: 10.1088/0004-637X/703/2/L99.

A. Wolszczan and D. A. Frail. A planetary system around the millisecond pulsar PSR1257 + 12. *Nature*, 355:145–147, January 1992. doi: 10.1038/355145a0.

A. Yeredor. Blind separation of gaussian sources via second-order statistics with asymptotically optimal weighting. *Signal Processing Letters, IEEE*, 7(7):197 –200, jul 2000. ISSN 1070-9908. doi: 10.1109/97.847367.

S. N. Yurchenko, R. J. Barber, and J. Tennyson. A variationally computed line list for hot NH_3. *MNRAS*, 413:1828–1834, May 2011. doi: 10.1111/j.1365-2966.2011.18261.x.

R. Zellem, C. A. Griffith, P. Deroo, M. R. Swain, I. P. Waldmann, and M. Zhao. *Demonstrating Palomar/TripleSpec's capability for exoplanet spectroscopy: an emission spectrum of HD 209458b.* in prep., 2012.