From space to ground: planetary atmospheres revealed through a machine learning approach

Mario Damiano

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

of

University College London.

Department of Physics and Astronomy

University College London

December 20, 2018

I, *Mario Damiano*, 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 work.

Abstract

In recent years, the study of exoplanetary atmospheres has flourished well beyond expectations. Current data are unveiling the key properties of hot massive planets orbiting very close to their stars, but sometimes results are not easy to interpret due to systematics affecting the data and degeneracies across the parameter space. The focus of my thesis is the study of exoplanetary atmospheres through spectroscopic observations using space and ground-based observatories.

The first part of the thesis describes the development of a new pipeline to analyse the low-resolution exoplanet data recorded with the WFC3 (Wide Field Camera 3) on-board the Hubble Space Telescope (HST). The focus is on a particular dataset: HAT-P-32b which is one of the most inflated Hot-Jupiters to date. Two different approaches are presented: a more standard parametric method and the use of a machine learning technique such as independent component analysis (ICA) applied for the first time on HST dataset. Water vapour and possibly more exotic metal-oxides such as VO and TiO are found in the atmosphere of HAT-P-32b. Further observations at longer wavelengths are needed to confirm these and other chemical compounds.

The second part describes the development of a new pipeline to analyse high resolution datasets recorded with ground based instruments (VLT/CRIRES, TNG/GIANO-B). High-resolution spectroscopy (HRS) allows to resolve molecular bands into individual lines. Using radial velocity measurements and techniques such as Cross-Correlation Function, it is possible to separate three physically different sources: telluric absorption, stellar absorption and planetary transmission spectrum, which are normally entangled. The standard method used in the literature to analyse HRS data consists on applying corrections for the airmass and for the stellar signal and the use of ad-hoc masks to eliminate residual, strong features. The analysis method that I have developed is based on a novel use of Principal Component Analysis (PCA) that aims to maximise the planetary signal without any manual corrections.

The two approaches are highly complementary and may be used to constrain

the thermal structure and the composition of the planetary atmosphere.

Impact statement

Part of the work presented in this thesis is obtained by using machine learning techniques, which currently find application in a number of scientific areas and sectors beyond astrophysics.

Acknowledgements

This document is the result of a journey started three years ago and many people need to be acknowledge for their moral and practical support.

In primis, I would like to say thanks to my parents for giving me the possibility to choose my path since my childhood, supporting and encouraging my passions. To my little brother for his help to dissipate my doubts and fears. To my fiancé for the huge support along the way and for believing in me.

I want to say thanks to my Ph.D. supervisors Prof. Giovanna Tinetti and Dr. Giuseppina Micela for the extraordinary possibility that they have given to me, teaching and suggesting the values of a scientist.

Another huge acknowledge is reserved to Prof. Giovanni Peres that with his passion supervised me trough my undergraduate studies becoming an example for me.

To Dr. Ingo Waldmann that inspired me with his open mind, on studying and applying the tools of machine learning and deep learning.

To Dr. Angelos Tsiaras that has treated me more as a brother than as colleague helping with useful discussions and suggestions.

I would like to thank also my Ph.D. colleagues Tiziano Zingales, Gordon Kai Yip, Billy Edwards and Dr. Giuseppe Morello for the good time spent together during these years working and exchanging useful ideas and point of views.

An acknowledgement is reserved to the two examiners Prof. Peter Doel and Prof. David Pinfield 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 closest friends.

Contents

1 Introduction **27**

1.1 Radial Velocity 29

1.2 Transit 32

1.3 Constraining the bulk composition through density 33

1.3.1 Gaseous and icy planets 34

1.3.2 Terrestrial planets 35

1.4 Thesis outline 36

2 Atmospheric characterization **39**

2.1 Transmission spectroscopy 40

2.2 Emission and reflection spectroscopy 42

2.3 Spectral modelling 44

2.4 Transit method 45

2.4.1 Primary transit observations 47

2.4.2 Eclipse and phase curve observations 48

2.5 Direct imaging 50

2.6 High-Resolution Spectroscopy 51

2.6.1 HRS observations 52

2.7 High and low resolution 54

3 Signal decomposition **57**

3.1 Observing a Cocktail Party 59

3.2 Principal Component Analysis 60

3.3 SYSREM algorithm 63

3.4 Independent Component Analysis 64

3.4.1 Interference-to-Signal Ratio 65

3.4.2 COMBI and MULTICOMBI 66

4 Space observations 69

4.1 Code overview 70

4.2 HST/WFC3 and the spatial scanning mode 71

4.3 Data reduction 74

4.3.1 Zero Read subtraction 75

4.3.2 Non-linearity correction 76

4.3.3 Dark current subtraction 77

4.3.4 Gain conversion 77

4.3.5 Sky background 77

4.3.6 Flat field correction 77

4.3.7 Bad pixels and cosmic rays correction 78

4.4 Calibration and extraction 78

4.4.1 Point source position 79

4.4.2 Position shift 80

4.4.3 Wavelength-dependent photon trajectories (w.d.p.t.) 81

4.4.4 Extraction 86

4.5 Obtaining the 1D spectrum 88

4.5.1 Fitting the white light-curve 89

4.5.2 Fitting the spectral light-curves 90

4.6 HAT-P-32b dataset 92

4.6.1 Using ICA to de-trend the light-curves 101

4.6.2 Atmospheric retrieval 104

4.6.3 Discussion 106

5 Ground observations 111

5.1 Code overview 112

5.2 Model and input data handling 113

5.3 Calibration and spikes correction 116

5.4 PCA and SYSREM 119

5.4.1 Time or wavelength domain? 121

5.4.2 Comparison of the algorithms 126

5.5 Cross-Correlation Function (CCF) 129

5.6 Output 131

5.6.1 SNR map 131

5.6.2 Whelch' T-test statistics 132

5.7 VLT/CRIRES datasets 136

5.7.1 Data reduction 137

5.7.2 HD209458b 138

5.7.3 HD189733b 149

5.7.4 Discussion 158

5.8 TNG/Giano-B datasets 160

5.8.1 Data reduction 160

5.8.2 Data analysis 162

6 Conclusion 165

Bibliography 168

List of Figures

1.1 The graph depicts the exoplanets population updated at October 2018. 28

1.2 Cumulative number of exoplanets detection per year and per different techniques. Figure source: `https://exoplanetarchive.ipac.caltech.edu/exoplanetplots/`. 29

1.3 The cartoon shows the principle behind the radial velocity method. Star and planet orbit around a common centre of gravity. The wobble of the star results in a Doppler shift according to the system inclination along the observer's line of sight. 30

1.4 Geometry of the transit method with relative light-curve. Figure source: `https://heasarc.gsfc.nasa.gov/docs/tess/primary-science.html`. 33

1.5 Mass–radius relation for giant planets, updated at Feb 2015. Figure adopted from: Fulton et al. (2015). 34

1.6 The graph depicts mass–radius relation for planets with masses <20 $M_\oplus$, updated at January 2018. Figure source: Damasso et al. (2018). 36

2.1 Geometry of transmission spectroscopy. Source of the figure: Hollis et al. (2013). 41

2.2 Phase-curve of the hot super-Earth 55 Cancri e (Demory et al., 2016). 46

2.3 Detection of potassium in the atmosphere of XO-2b (Sing et al., 2011). 47

2.4 First consistent population study using observations from SST/IRAC, HST/WFC3 and HST/STIS (Sing et al., 2016). 49

2.5 The infrared spectrum of 51Eri b. The picture is part of the work of Rajan et al. (2017). 51

2.6 **Top panel:** synthetic H_2O spectrum at different resolutions. **Bottom panel:** synthetic CO spectrum at different resolution (Birkby, 2018). 52

2.7 **Top panel:** motion of the planet around the host star. **Bottom panel:** Doppler shift of the planetary CO signal as a function of the orbital phase. The planetary signal is identified without degeneracies due to higher radial velocity amplitude (Birkby, 2018). . . . 53

2.8 CO signal in transmission in the atmosphere of HD209458b after the data have been cross-correlated with the synthesised CO model. Figure adopted from Snellen et al. (2010). 54

4.1 Flow chart of the pipeline for analysing HST/WFC3 datasets. The box colours indicate different classes of action: green boxes represent external inputs coming from other models or different sources, e.g. users. The blue boxes contain the modules and routines of the code. 71

4.2 Raw image from the HAT-P-32b dataset, the scanning mode technique disperses the signal along the vertical axes. The total scan length is approximately 40 pixels. 73

4.3 WFC3/IR non linearity effect in time. In red the expected electrons recorded and in black the actual value recorded. Figure adopted from: `http://www.stsci.edu/hst/wfc3/pipeline/wfc3_pipeline`. 76

4.4 Left-top: sum along the columns of the first (continuous) and the last (dashed) spatially scanned spectra of the HAT-P-32b dataset. Left-bottom: difference between the two profiles before and after shifting, indicated by dashed and continuous lines, respectively. Right: same plots for the sum along the rows of the first non-destructive read. 82

4.5 Horizontal (top) and vertical (bottom) shift relative to the first one for each image of the HAT-P-32b dataset. 83

4.6 Relative positions of the spectral trace (red line), of the star ($x^\star$, $y^\star$) in the direct image, and a random point on the trace (x_λ, y_λ). θ indicates the tilt of 0.5 degree of the grism respect to the detector' rows. Image adopted from: Tsiaras et al. (2016a). 84

4.7 Calculated wavelength solution of one frame of the HAT-P-32b dataset. **Top panel**: Trajectories of the dispersed photons of different wavelengths (coloured points) while the star moves along the scan direction (white arrow). **Bottom panel**: Left and right edges of the spectrum where it is possible to appreciate how the grid follows the signal on the detector. 87

4.8 Extracted raw white light-curve of the dataset HAT-P-32b. 89

4.9 **Top panel**: normalised white light-curve of HAT-P-32b before fitting for transit and systematics model. **Middle panel**: white light-curve divided by the best-fit model of the systematics. **Bottom panel**: fitting residuals, it can be seen that the model fails to fit the egress. The possible reasons for this behaviour are either non-optimal orbital parameters, limb-darkening coefficients or remaining systematics. 91

4.10 Single reduced frame of the HAT-P-32b dataset before the extraction. It is possible to see how the signal relative to each of the two stars is blended with each other. 94

4.11 Multiple extractions of the HAT-P-32b dataset. The 12 stripes correspond to the NDRs recorded by Hubble Space Telescope. In each of the stripes the two signals relative to the two stars are divided, so that possible contamination can be avoided. 97

4.12 1D spectra of the first 6 stripes of the HAT-P-32b dataset. 98

4.13 1D spectra of the last 6 stripes of the HAT-P-32b dataset. 99

4.14 Final 1D spectra resulted from calculating one spectral light-curve per bin (stacked) and averaging the spectral $(R_p/R_\star)_\lambda$ for each bin for each stripe (weighted). The two spectra are equivalent within 0.3σ. 100

4.15 The figure shows two spectral light-curve ($1.3657-1.3901\mu$m) referred to the two method adopted (stacked and weighted) before and after the ICA correction. Font: priv. comm. of Dr. Giuseppe Morello 103

4.16 Comparison between all the spectra obtained by using all the combinations of the method proposed. An horizontal shift has been introduced to make easier the visualisation. Font: priv. comm. of Dr. Giuseppe Morello. 103

4.17 Stacked spectra obtained with the parametric pipeline (magenta) and with the ICA + white residuals stacked approach (blue). 104

4.18 Transmission spectrum of HAT-P-32b obtained with the parametric pipeline (black) and best-fitting model (light-blue). 106

4.19 Posterior distributions to the fit for the WFC3 spectrum of the giant planet HAT-P-32b. Even though we tested the presence of many other molecules in this atmosphere, here we show only the posterior of H_2O because it is the only significant one. All of the other molecules do not show a statistically significant contribution to the fit. 108

5.1 Flow chart of the pipeline for high spectral resolution data analysis. The box colours indicate different classes of action: green boxes represent an external input coming from other models or different sources. The red box includes all the different reduction pipelines supported. Finally, the blue boxes contain most of the calculations that I developed. 113

5.2 1D extracted spectrum for the four CRIRES detectors (HD209458b dataset). 116

5.3 Telluric transmission spectrum in the nominal wavelengths of the CRIRES spectrograph generated by the SKYCALC tool for clear sky scenario. 117

5.4 Pixel-wavelength relationship for detector 1 of the HD209458b dataset. **Top panel** shows the relationship between pixel numbers and wavelength. **Mid panel** shows the residuals after a linear fit, a trend is still visible. **Bottom panel** shows residuals after fourth order fit. 118

5.5 The HD209458b (full discussion see Sec. 5.7) dataset after calibration. The y-axis maps the orbital phase. The x-axis the wavelength grid. 119

5.6 Detector 1 of the HD189733b dataset prior to the application of any pre-processing algorithm. 120

5.7 **Top Panel:** detector 1 of the dataset HD189733b with the mean of each column subtracted. **Bottom Panel:** same image but with the standardisation applied as pre-process step. 120

5.8 **Top Panel:** covariance matrix of the detector 1 in the HD189733b dataset, where the input matrix has been transposed (TDM case): the dimension of this matrix is the number of spectra taken during the observation (i.e. 45). **Bottom Panel:** same as top but the input matrix has not been transposed (WDM case) and the dimension is equal to the number of the wavelength bins (i.e. 1024). 122

5.9 **Left panels** show the first five eigenvectors of the TDM case (45×45) (**top panel** Fig. 5.8). These components depict the time correlation of the data along the time domain. **Right panels** show the five eigenvectors of the WDM covariance matrix (1024×1024) (**bottom panel** Fig. 5.8). 123

5.10 Linear relation between the first component in the time domain and the recorded airmass. 124

5.11 **Top panel** shows the explained variance ratio (eigenvalues) of the decomposition of the TDM covariance matrix. The graph is related to the first detector of the HD189733b dataset. **Bottom panel** is the same as top panel but in the WDM case. In both figures the first component has most of the total variance (<70%). 125

5.12 **Top figure:** the input matrix (PCA input) for reference. **Left column:** the PCA's results in the TDM case. **Central column:** the same as **left column** but in the WDM case. **Right column:** the difference between the wavelength and the time decomposition. Telluric residuals are present in the difference, meaning that the two methods are not equivalent. 126

5.13 Results obtained with the decomposition algorithms after subtracting the first five components (from top to bottom). **Left column:** PCA algorithm, **right column:** SYSREM. 127

5.14 Differences between SYSREM and PCA residuals obtained by applying these two algorithms to the input data. Structured signals are visible and the differences are 1/10 of the signal showed in Fig. 5.13. 128

5.15 **Top left panel:** the four CCF of the four CRIRES' detectors summed together. **Bottom left panel:** same as top but with the model injected. The injection is $1\times$ the synthesised model ($R_p/R_\star$ $\sim 10^{-3}$). **Top right panel:** cross-correlation after changing the reference frame from the Earth to the rest frame of the exoplanet. In this frame the planetary cross-correlation signal is aligned to zero $\mathrm{km\,s^{-1}}$. **Bottom right panel:** same as top right panel but with the injection. The injected signal is aligned to zero $\mathrm{km\,s^{-1}}$ in the exoplanet's rest frame. 130

5.16 In-transit summed cross-correlations for the input data and for different injections. The injection perturbs the surroundings of the CCF peak,the stronger the injection the higher is the perturbation around the peak. 132

5.17 **Top Panel:** SNR map in case of planetary signal detection (the maximum SNR is in agreement with the orbital K_p and the planetary v_{rest}) (Full discussion Sec. 5.7). **Bottom panel:** same map but the planetary signal is not detected. 133

5.18 **Top panel:** SNR variation with respect to K_p at fixed v_{rest} value of the SNR map in Fig. 5.17 top panel. The red vertical lines indicate the uncertainty interval. **Bottom panel:** same as **top panel** but referred to Fig. 5.17 bottom panel. 134

5.19 Co-added in-transit cross-correlations at different K_p. The red square determine which values are $in-trail$, i.e. those inside the box, and values that are $out-trail$, i.e. outside the box. 135

5.20 Distribution of the statistical Welch's T-test performed on HD209458b dataset. In-trail (orange) and out-trail (blue) distributions are drawn and in both cases the in-trail distribution is shifted towards positive CCF values. 136

5.21 CRIRES' reduction process. **Top left panel**: raw frame in the nodding position B. **Bottom left panel**: raw frame in the nodding position A. **Right panel**: combination of the two nodding position. 138

5.22 HD209458b dataset. In **(a)**, the data are shown after calibration, normalisation and spikes correction. In **(b)**, the data are shown after the MFS method has been applied (Sec. 5.4). In **(c)**, the results of PCA are shown. In **(d)**, the data are shown after the application of PCA and the injection of the CO model. 142

5.23 Detectors' variances of the PCA decomposition relative to the HD209458b dataset. The first component always carries more than 75% of the information. However, the variance is different for each of the detectors. 143

5.24 Summed in-transit cross-correlations with and without the injection of carbon monoxide. The injection perturbs the surroundings of the CCF peak. The co-added CCFs are relative to the planetary rest frame of HD209458b ($K_p = 145.041$ km s^{-1}). This graph has been generated considering components from 7 to 28 of the PCA decomposition (the same of Fig. 5.25). 144

5.25 SNR map for the carbon monoxide. The maximum point is compatible with the planetary orbital parameters. 144

5.26 Distributions (i.e. *in-trail* and *out-trail*) used to compute the Whelch's T-Test. The null hypothesis is rejected with a confidence >7σ. 145

5.27 SNR map for the carbon monoxide when SYSREM is used. The maximum point is compatible with the planetary parameters but the peak value is lower than the one obtained using PCA. 145

5.28 Cross-correlations of water vapour co-added in-transit. The injected signal and the planetary signal are still present after using PCA. The co-added CCFs are relative to HD209458b rest frame ($K_p =$ 145.041 $\mathrm{km\,s^{-1}}$). This graph has been generated considering PCA components from 33 to 43. 146

5.29 SNR map of the water vapour. The peak is compatible with the planetary parameters. 146

5.30 Distribution (i.e. *in-trail* and *out-trail*) used to compute the Whelch's T-Test. The null hypothesis is rejected with a confidence of 6.56σ. 147

5.31 **Top panel**: SNR map of the dataset HD209458b cross-correlated with the telluric model. **Bottom panel**: SNR map of the telluric model including the same range of components of the reported water vapour detection (see Fig. 5.29). No signal is visible at the orbital parameters of the planet. 148

5.32 HD189733b dataset. In **(a)**, the data are shown after calibration, normalisation and spikes correction. In **(b)**, the MFS method has been applied to the data (Sec. 5.4). In **(c)**, the data are shown after PCA is applied. In **(d)**, the same figure of *c* but with the injection of CO. 152

5.33 Detectors' variance of the PCA decomposition relative to HD189733b dataset. The first component always carries more than 75% of the information. 153

5.34 The picture shows the co-added in-transit cross-correlations for the input data and for different levels of injections of carbon monoxide. The injected signal and the planetary signal are present after using PCA. The co-added CCFs shown in picture are calculated aligning the singular CCFs to the theoretical orbital velocity of the planet HD189733b ($K_p = 152.564\,\mathrm{km\,s^{-1}}$). This graph has been generated using PCA components from $22th$ to $36th$, same of Fig. 5.35 154

5.35 The figure shows the SNR map for the carbon monoxide for the HD189733b dataset. The peak of the matrix is compatible with the results reported by Brogi et al. (2016) but not with the theoretical radial velocity of HD189733b ($K_p = 152.564$ km s^{-1}). 154

5.36 Cross-correlations of the planetary signal and of the injected water vapour co-added in-transit. The co-added CCFs are calculated at the theoretical orbital velocity of the planet HD189733b ($K_p = 152.564$ km s^{-1}). The CCFs are the result of the combination of the PCA components from $12th$ to $27th$. 155

5.37 SNR map of water vapour for the HD189733b dataset. The maximum point is compatible with Brogi et al. (2016) and with the planetary parameters. 155

5.38 The figure shows the relation between SNR and K_p at fixed v_{rest} of the peak value of SNR map shown in Fig. 5.37. The red vertical lines indicate the uncertainty interval. 156

5.39 The figure shows the two distribution (i.e. *in-trail* and *out-trail*) used to compute the Whelch's T-Test. The null hypothesis is rejected with a confidence of 5.21σ. 156

5.40 SNR map of the HD189733b dataset cross-correlated with the telluric model in the same range of components of the result reported in Fig. 5.37. 157

5.41 GIANO-B' reduction process. **Top panels**: raw images observed in nodding A (left) and B (right). **Bottom panel**: difference A-B: the B trace has negative (dark) values. Figure source: GOFIO manual `https://atreides.tng.iac.es/monica.rainer/gofio`. 161

5.42 GIANO-B' extracted spectra. All the 50 orders are shown and some of them are highly affected by the telluric absorption. 163

5.43 Same of Fig. 5.42 but here the saturated orders have been neglected and the remaining orders have been normalised. 163

5.44 **Top panel**: GIANO-B' order 33 before the pre-processing. The differences on the nodding baselines are visible. **Bottom panel**: same matrix as top panel but the MFS method has been applied. The difference between the left and right areas of the detector is visible. 164

5.45 GIANO-B' order 33 variance of the PCA decomposition relative to the WASP-33b dataset. The first component carries less than 50% of the information. 164

List of Tables

4.1 Relation between the length of the direct image and the SUBTYPE parameter. 80

4.2 Horizontal offset for the different WFC3/IR filters. 80

4.3 Calibration coefficients for the G141 grism. 84

4.4 HAT-P-32 system information (Hartman et al., 2011). 92

4.5 White light-curve fitting results. 95

4.6 White light-curve fitting results for every stripe. Comparison between staked and weighted results. 96

4.7 Limb darkening coefficients a_{1-4} and transit depth $(R_p/R_*)^2$ for the wavelength channels. 105

4.8 Fitting results for HAT-P-32 b atmosphere 107

5.1 HD209458 system information 140

5.2 The table shows the results obtained for the HD209458b dataset. . . 141

5.3 HD189733 system information. 150

5.4 The table shows the results obtained for the dataset HD189733b. . . 151

Chapter 1

Introduction

"A journey of a thousand miles begins with a single step, even the longest and most difficult ventures."

– Laozi - 550 BC

Young Ellie: *Dad, do you think there are people on other planets?*

Ted Arroway: *I don't know, Sparks. But I guess I'd say if it is just us... seems like an awful waste of space.*

This was the opening of a movie which drastically marked my childhood. *Contact* in 1997 was released worldwide and quickly became a cult movie. Obviously, it was not the first work outlining the space observation and exploration nor was it the last. Since the dawn of humankind we have been looking at the sky with passionate curiosity, questioning if we are alone in the universe. After millions of years we still have this conundrum in mind. However, we have started moving some steps towards the answer. Finally, in 1992, the first exoplanet was discovered. The millisecond pulsar PSR1257+12, has shown a non regular rotational period due to the presence of small celestial bodies gravitationally bounded (Wolszczan & Frail, 1992). This was a breakthrough result at that time. Three years later Mayor & Queloz (1995) discovered a planetary system around a solar-type star. 51Peg is indeed a G2 star (5793 K) like our Sun. The orbiting planet (51Peg b) is an *Hot-Jupiters*, a kind of planet that is not present in our solar system. 51Peg b is a Jupiter-like planet in terms of radius and mass, but it is '*Hot*' (>1000 K) due to its small distance from the host star. Since these discoveries everything has changed and the era of extra-solar planets (exoplanets) has begun.

Recent studies have shown that each star of our galaxy, the Milky Way, statistically hosts at least one planet (Cassan et al., 2012). Given that, our galaxy is populated by billions of stars, billions of planets are then waiting to be discovered. If we imagine that our galaxy is just one among millions or even billions of galaxies, the total number of the planets in our universe probably is uncountable.

A quarter of century after the first exoplanet discovery, we count approximately 4000 exoplanets in our catalogues. These planets are different in terms of physical and chemical properties. We can see planets with the same dimension of the Earth, planets a few times bigger than the Earth, Jupiter-like planets and planets much bigger than our gas giants. To this, a great variety of temperatures seems to be present (Fig. 1.1).

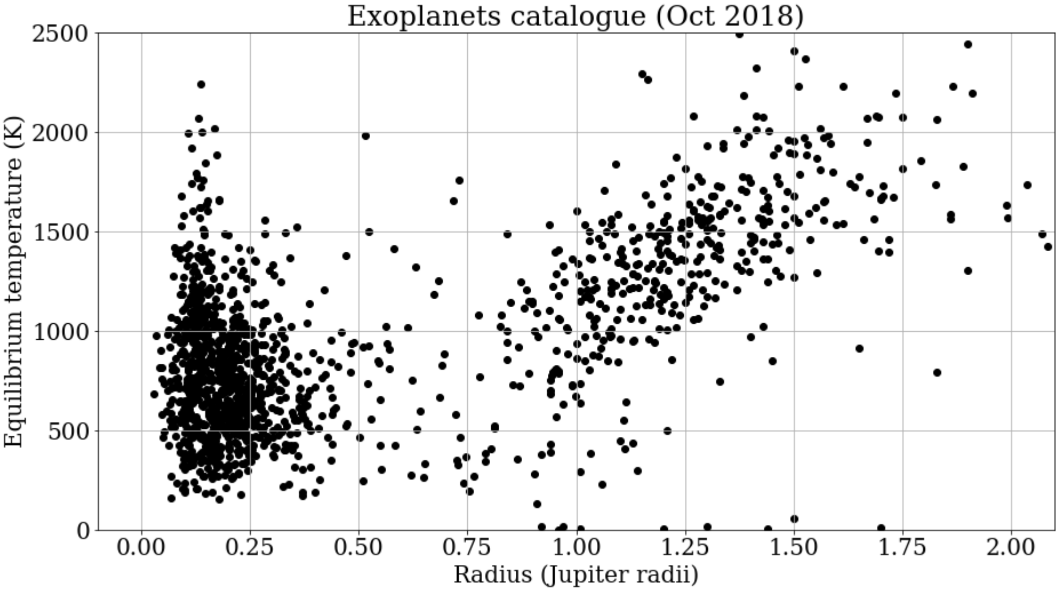

Figure 1.1: The graph depicts the exoplanets population updated at October 2018.

A long-term goal is to classify exoplanets in the same way we classify stars but we have no clues today whether a planetary analogue of the H-R diagram (Russell, 1910; Hertzsprung, 1912) makes sense or not.

Fig. 1.2 shows the number of planets detected per year and with different techniques. The most successful methods are radial velocity and transit.

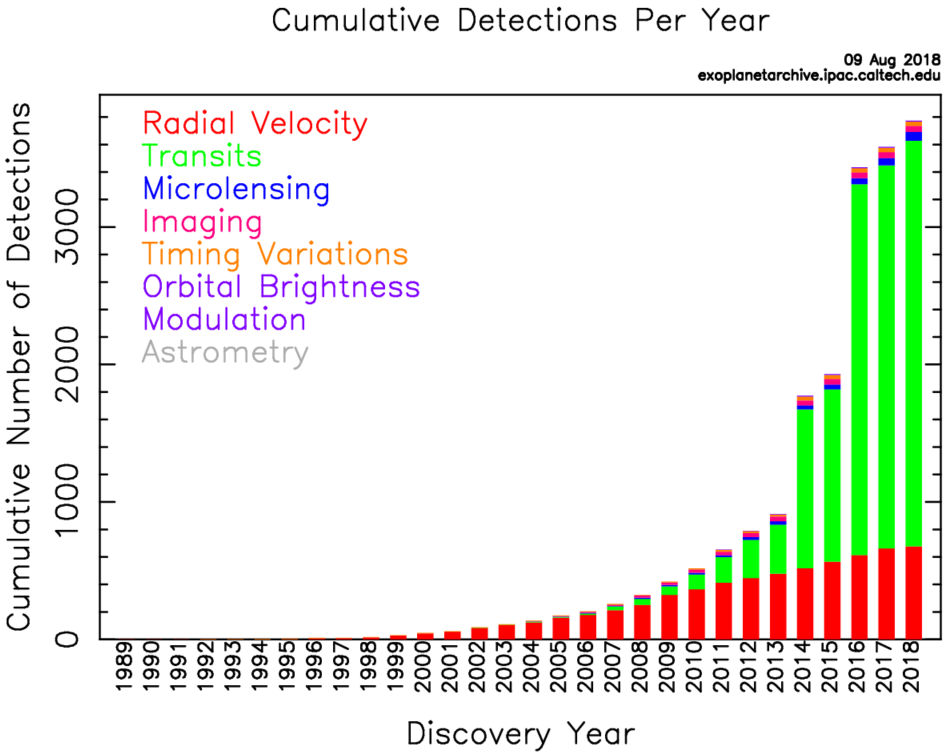

Figure 1.2: Cumulative number of exoplanets detection per year and per different techniques. Figure source: `https://exoplanetarchive.ipac.caltech.edu/exoplanetplots/`.

Both transit and radial velocity provide the orbital period of the planet. The measurement of the planetary radius through transit observations (Charbonneau et al., 2000; Henry et al., 2000) combined with the measurement of the mass with radial velocity (Mayor & Queloz, 1995; Mazeh et al., 2000), allows a preliminary estimate of the planetary bulk density.

Finally, most of the techniques we use today are indirect detections because they rely on the observations of effects on stars caused by the presence of planets. The only exception is the direct image method which is the only direct method.

1.1 Radial Velocity

The first exoplanet around a main sequence star was 51Peg b (Mayor & Queloz, 1995), and it was discovered using radial velocity technique. This method consists on observing periodic *Doppler shifts* in the spectrum of a star due to the presence of an orbiting planet. Since the planet and the star orbit around a common centre of mass (centre of gravity of the system). When we observe along our line of sight we might detect (if the inclination of the system is not face-on, i.e. $i = 0°$) the stellar motion through the Doppler shifts of the star spectrum (Fig. 1.3). Of course, due to the difference in mass between the star and the planet this effect is tiny on the star spectrum.

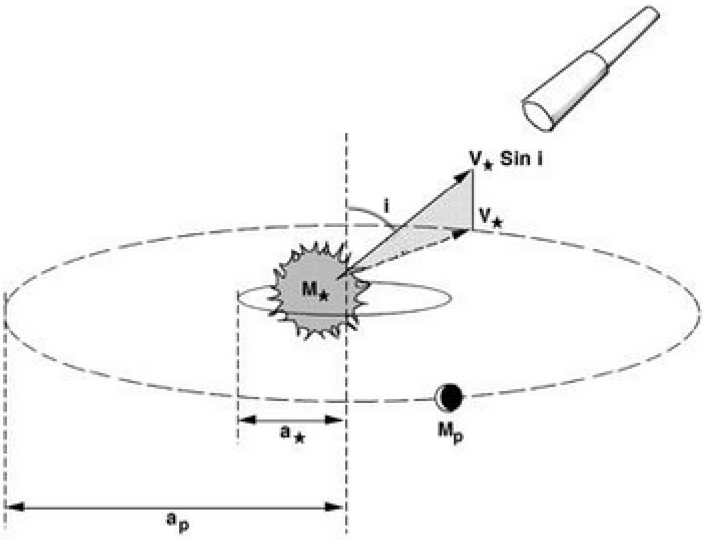

Figure 1.3: The cartoon shows the principle behind the radial velocity method. Star and planet orbit around a common centre of gravity. The wobble of the star results in a Doppler shift according to the system inclination along the observer's line of sight.

The explanation of the Doppler effect was firstly included within the special relativity (Einstein, 1905), regarding a photon moving in a flat space time. The definitive description of the effect was finally comprised in the general relativity taking into account massive celestial bodies (Eq. 1.1) (Lovis & Fischer, 2010)

$$\lambda = \lambda_0 \frac{1 + \frac{1}{c}\mathbf{k}\cdot\mathbf{v}}{1 - \frac{\Phi}{c^2} - \frac{v^2}{2c^2}} \tag{1.1}$$

where λ_0 is the wavelength of an emitted photon in the rest frame of the source, λ is the wavelength recorded by an observer moving with respect to the emitter, $\mathbf{v}$ is the velocity vector of the source relative to the observer, **k** is the unit vector pointing towards the source from the observer, Φ is the Newtonian gravitational potential at the source ($\Phi = GM/r$ at a distance r of a spherically source M) and c is the speed of light.

However, if the intensity of the velocity vector **v** is very low compared to the speed of light c (e.g. exoplanets' orbital velocities) we can well describe observations by using a classical approach, neglecting the relativistic terms. In this case Eq. 1.1 can be reduced to the more familiar classical expression of the Doppler shift (Lovis & Fischer, 2010): let $v_\star$ be the radial velocity of the star around the system barycentre, i the orbital inclination of the system plane with respect to the observer's line of sight, c the speed of light and λ the observational wavelength, then the shift is calculated as follows

$$\Delta\lambda = \lambda \frac{v_\star \cdot \sin(i)}{c} \tag{1.2}$$

For a Sun like star and a close-in Jupiter-like planet the radial velocity is typically ~100 m/s. This value drops to a few cm/s if we consider an Earth-like planet around a Sun-like star.

From Eq. 1.2 we derive the projection of the orbital velocity of the star along the line of sight $K_\star = v_\star \sin(i)$. This allows to calculate the lower limit of the mass of the planet, resolving Eq. 1.3 (Lovis & Fischer, 2010), with the inclination of the planetary orbit being unknown.

$$\frac{m_p^3 \sin^3(i)}{(M_\star + m_p)^2} = \frac{PK_\star^3}{2\pi G}\left(1-e^2\right)^{3/2} \tag{1.3}$$

In the last equation m_p is the mass of the planet, $M_\star$ is the mass of the star that can be determined from stellar models and observations, P is the orbital period and can be calculated from the periodicity of the Doppler shifts, e is the eccentricity of the planet's orbit and finally, i is the inclination of the orbital plane, such that $i = 90°$ is edge-on and $i = 0°$ is face-on with respect to the line of sight of the observer.

1.2 Transit

The first dedicated mission to discover exoplanets using the transit technique was CoRoT (Convection, Rotation and planetary Transits) (Auvergne et al., 2009). It was launched in 2006 and was operational until 2013 (Barge et al., 2008; Alonso et al., 2008; Csizmadia et al., 2015). A substantial number of exoplanet detections through transit were also made using ground-based facilities e.g. HATNeT (Bakos, 2018), SuperWASP (Street et al., 2003) and TRAPPIST (Jehin et al., 2011).

From Fig. 1.2 we note that the number of planets rocketed in 2014 and in 2016. This is due to the Kepler's double program data releases (Kepler main mission and K2) (Burke et al., 2014; Rowe et al., 2015; Coughlin et al., 2016). Kepler was launched in 2009 and it was designed to discover Earth-size planets orbiting other stars in our galaxy (Borucki et al., 2003). Kepler observed continuously main sequence stars in a fixed field of view. In 2013 due to a failure of two reaction wheels the telescope was unable to continue the primary mission and in 2014 the K2 mission (Kapler2) (Howell et al., 2014) has began using the remained capabilities of the spacecraft.

The idea behind the transit technique is to observe the luminosity of the star through time and study its variability. The luminosity of a star can vary for different reasons, first of all, intrinsic variability and activity. However, it could also be due to the presence of a planet if the variability follows a particular pattern. Similarly to what happens when Venus crosses in front of the Sun's disc, the luminosity of a star drops when a planet passes in front of it (Fig. 1.4).

In Fig. 1.4 with L_1 we define the luminosity of the star before the transit, i.e. when the planet is outside the projected disc of the star, and with L_2 we define the luminosity of the star during the transit, i.e. when the planet passes in front of the star. We can then write:

$$L_1 = f_\star * \pi R_\star^2$$

$$L_2 = f_\star * \pi \left(R_\star^2 - R_p^2\right)$$

where $f_\star$ is the flux of the star, $R_\star$ and R_p are respectively the radius of the star and

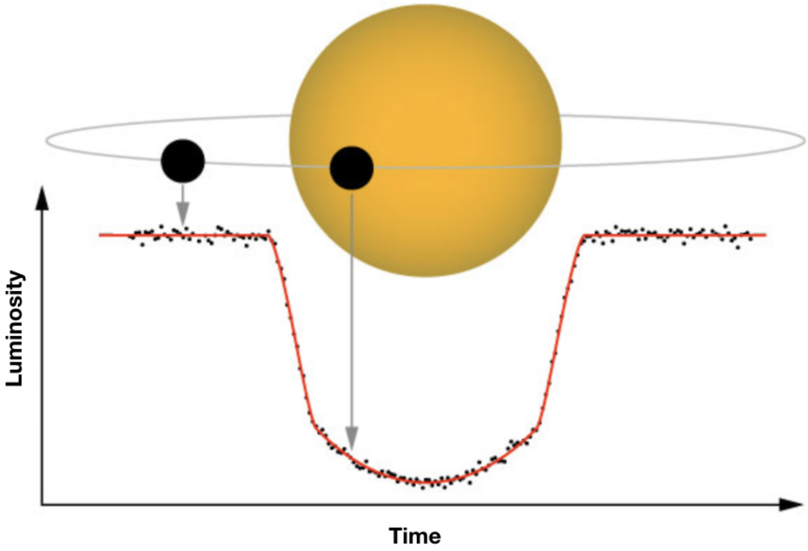

Figure 1.4: Geometry of the transit method with relative light-curve. Figure source: https://heasarc.gsfc.nasa.gov/docs/tess/primary-science.html.

the planet. The dip in the light-curve (Fig. 1.4) depends on the type of planet and star

$$\frac{L_1 - L_2}{L_1} = \left(\frac{R_p}{R_\star}\right)^2 \tag{1.4}$$

From Eq. 1.4, we can estimate the radius of the planet R_p in units of stellar radii.

1.3 Constraining the bulk composition through density

Using the results derived from radial velocity and transit methods, described in previous sections, it is possible to estimate the mean density of a planet. However, the mass-radius relationship has degenerate solutions in terms of bulk composition and planetary characteristics. The iconic example is offered by Earth and Venus which are twin planets in terms of radius and mass but they are very different worlds. Venus is inhospitable with a thick atmosphere made of CO_2 and sulphuric acid clouds due to a combined effect of volcanism and photochemical processes. Venus' surface temperature may reach 700K (Walker, 1975). On Earth, molecular nitrogen and oxygen are the main molecules in the atmosphere and the temperature is appropriate for life.

1.3.1 Gaseous and icy planets

Thanks to the explorations and studies conducted in our solar system we know that our giant planets are mainly composed of hydrogen and helium with other chemical compounds, some condensed due to low temperature (Atreya et al., 2003). Giant exoplanets are also expected to have similar composition, but due to the higher temperature (see Fig. 1.1), chemical species such as H_2O, CO, CO_2, CH_4, NH_3 are present in gas phase.

Fig. 1.5 shows the mass-radius relation for Jupiter-like planets. From this graph it is not possible to infer the precise interior structure and bulk composition.

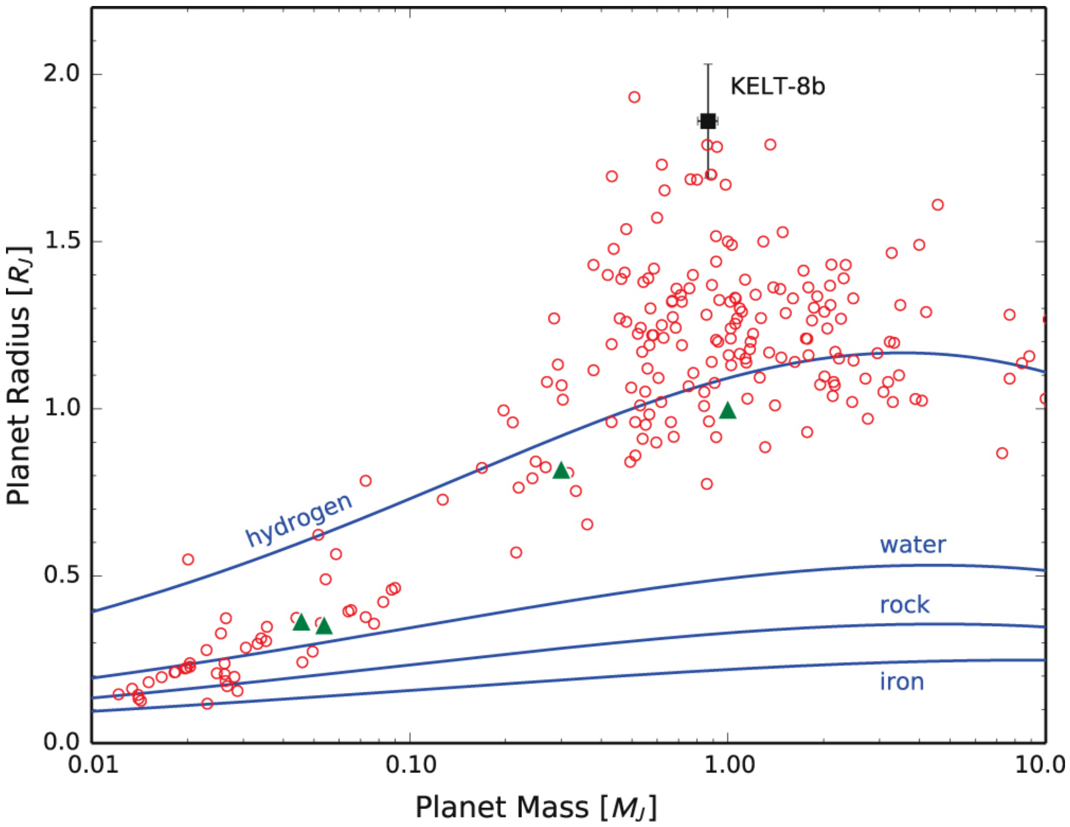

Figure 1.5: Mass–radius relation for giant planets, updated at Feb 2015. Figure adopted from: Fulton et al. (2015).

To start with, we do not know the age of these planets. We expect younger planets to be more inflated. More in general, giant exoplanets appear to be more puffed-up, but the reason for this is not well understood. A number of explanations have been proposed in the literature, e.g. Batygin et al. (2011), Wu & Lithwick (2013).

1.3.2 Terrestrial planets

Our direct knowledge about terrestrial planets is limited to the four examples in our solar system plus the giant planets moons. In spite of this, Super-Earths (not present in our solar system) are likely to be the most common type of planets in our galaxy (Fressin et al., 2013; Fulton et al., 2017). They are typically bigger than the Earth (a few Earth masses up to $\sim$10 $M_{\oplus}$).

Fig. 1.6 shows the mass-radius relation for planets with masses up to 20 $M_{\oplus}$. Explaining the diversity that we observe is not easy because it is the result of the combination of a variety of factors (e.g formation and evolution processes). For example, super-Earths may form in the same way as giant planets but then due to the interaction with parent star, they could lose part of the H/He envelope (Leitzinger et al., 2011; Owen & Jackson, 2012; Owen & Wu, 2013; Hansen & Zink, 2015). According to other models super-Earths could form later in the lifetime of the disc and they could accrete heavier elements (Lee & Chiang, 2016). The final result of what we observe depends also where the planets formed in the disc.

In the early history of the planet, volcanic activity can transport from the interior to the atmosphere heavy sulphur-, carbon- and silicates-based molecules. Impacts with meteorites or comets could contaminate the original atmospheric chemistry (Frei & Rosing, 2005; Moynier et al., 2009; Kleine et al., 2009; Willbold et al., 2011). Finally, in the case of Earth, life modified dramatically the atmospheric composition. We can speculate that a similar course of event might occur on other planets.

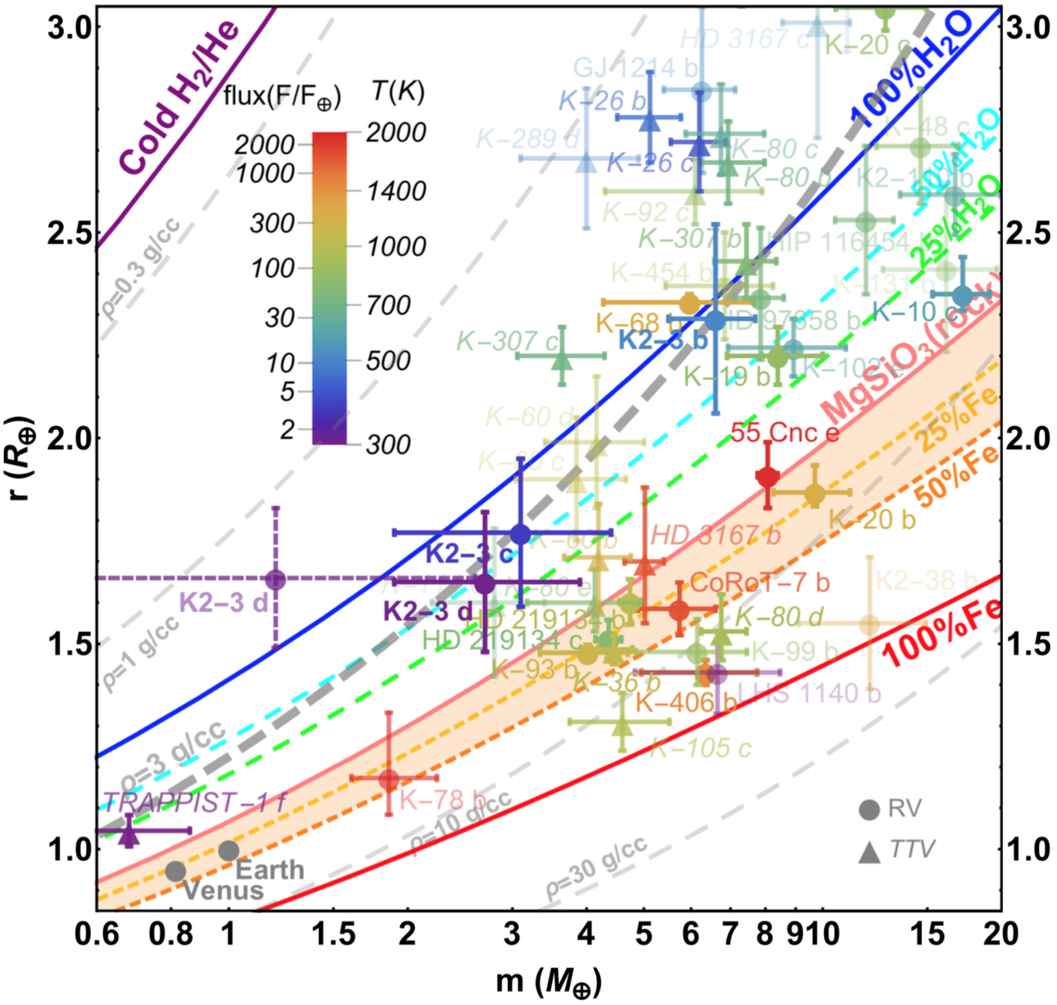

Figure 1.6: The graph depicts mass–radius relation for planets with masses <20 $M_\oplus$, updated at January 2018. Figure source: Damasso et al. (2018).

1.4 Thesis outline

The classification of planets is expected to be even less straightforward than our understanding of stars. Many effects have to be taken into account. All of them combined create a complex picture that is difficult to interpret if we focus on a small number of objects.

The focus of this Thesis is the development and the discussion of data analysis algorithms used to study space and ground observations to unveil exoplanetary atmospheres, their composition and their general characteristics.

In **chapter 2**, I will give a technical and detailed discussion on the characterisation of the planetary atmospheres through transit spectroscopy from space satellites and radial velocity from ground observatories.

In **chapter 3**, I will discuss the role played by machine and deep learning algorithms applied to data analysis.

In **chapter 4** the development of a new pipeline will be described. This is used to reduce the low-resolution exoplanet data recorded with the WFC3 (Wide Field Camera 3) on-board the Hubble Space Telescope (HST). The focus is on a particular dataset (i.e. HAT-P-32b) related to one of the most inflated Hot-Jupiters to date.

In **chapter 5**, I will present the data analysis of high-resolution ground observations (taken by VLT/CRIRES) and the technical details on a new developed pipeline will be discussed. Then, I will present the application of such pipeline to two datasets: HD189733b and HD209458b. Finally, I will show the work in progress on a different instrument (i.e. TNG/GIANO-B).

Chapter 6 is a brief summary of the work presented in this thesis and discusses future projects.

Chapter 2

Atmospheric characterization

"The important thing is not to stop questioning. Curiosity has its own reason for existence. One cannot help but be in awe when he contemplates the mysteries of eternity, of life, of the marvellous structure of reality."

– Albert Einstein - 1955

It is clear, from chapter 1, that the exoplanets population is more variegated than we have thought for years. To answer fundamental questions, like:

- how have they formed?
- what are exoplanets made of?
- how have they evolved since their formation?
- how can we explain what we observe today?

the chemistry of their atmospheres can provide a powerful diagnostics.

There are three main techniques to study and characterise exoplanetary atmospheres. The first one is the well known and established *transit spectroscopy*, thanks to which today we have the largest number of results. Another method relies on taking an image of the target planet, i.e. *direct imaging*. Finally – inheriting Doppler observations from the radial velocity technique and combining it with transit spectroscopy – *high dispersion spectroscopy* is bringing some valuable insights into the atmospheric studies. These techniques rely on absorption/emission/reflection of the particles, atoms and molecules present in the atmosphere to unveil aspects of these faraway worlds.

2.1 Transmission spectroscopy

Absorption occurs when the light interacts with a gas and the temperature of the emitter is higher than the one of the absorber. The same phenomenon takes place when the planetary atmosphere absorbs part of the starlight.

At zero-order approximation planetary atmospheres can be modelled as an annulus surrounding the planetary body of radius R_p which is always opaque (i.e. optically thick). The annulus that absorbs the incoming starlight has a radial height of a few scale heights (generally four or five in the IR, (Tinetti et al., 2013)). The scale height H is equal to $kT/\mu g$, where k is the Boltzmann's constant, T is the equilibrium temperature of the planet, μ is the mean molecular weight of the atmosphere and g is the planet's gravity. A $\mu \sim 2$ a.m.u. is a typical value for Jupiters-like planets i.e. mainly composed of H/He. The amplitude of the absorption can be approximated as (Tinetti et al., 2013)

$$\delta \sim 5 \cdot \left(\frac{2R_pH}{R_\star^2} \right) \tag{2.1}$$

where R_p and $R_\star$ are respectively the radius of the planet and the radius of the star. The amplitude is especially large for hot planets characterised by light atmospheres and low gravity.

If we want, however, to determine more precise information about the composition of the atmosphere and the atomic and molecular abundances we cannot rely on Eq. 2.1. Given the geometry of the event (Fig. 2.1), the absorption/filtering of the light is described by the Beer-Bouguer-Lambert law:

$$I(\lambda, z) = I_0(\lambda) e^{-\tau(\lambda, z)} \tag{2.2}$$

where λ is the wavelength, τ is the optical depth and $I(\lambda, z)$ and $I_0(\lambda)$ are respectively the intensity of the light after the absorption and the incoming one. Note that the optical depth is function of the altitude, z, and to calculate the total absorption

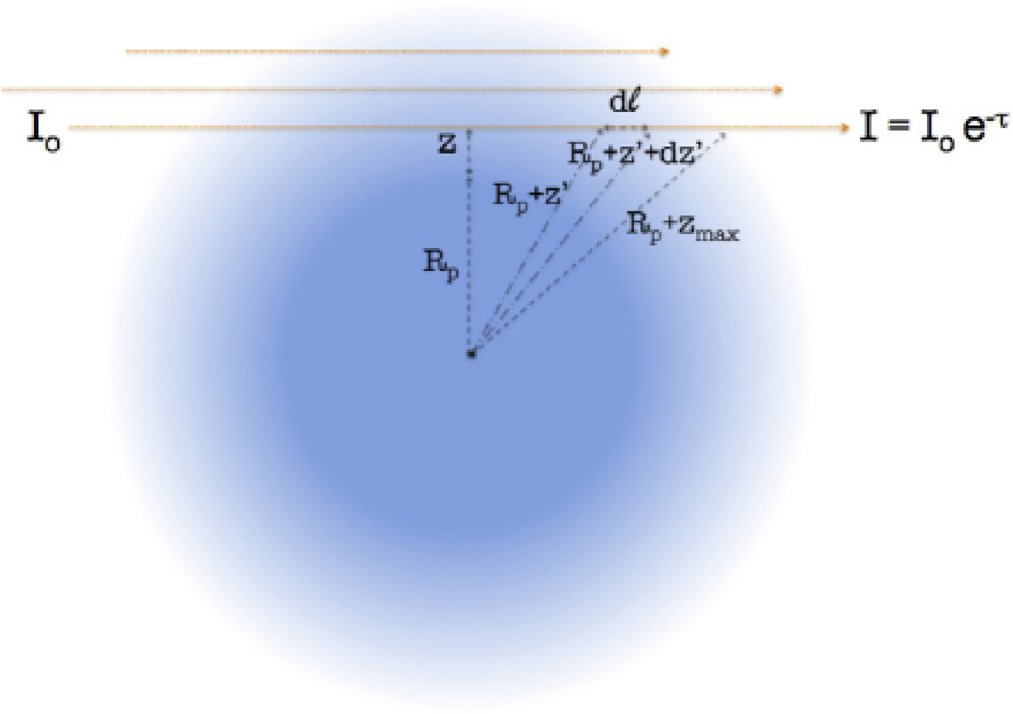

Figure 2.1: Geometry of transmission spectroscopy. Source of the figure: Hollis et al. (2013).

caused at one particular wavelength, $A(\lambda)$, we need to integrate the atmospheric opacity, $1-e^{-\tau(\lambda,z)}$, multiplied by the differential ring element of the projected atmosphere, $2\pi(R_p+z)dz$, for all the z values:

$$A(\lambda) = 2\pi \int_0^{z_{max}} (R_p+z)\left(1-e^{-\tau(\lambda,z)}\right) dz \tag{2.3}$$

where z_{max} is generally five scale heights of the planetary atmosphere. Finally, the total amplitude of the absorption per wavelength, $\delta(\lambda)$, is calculated by adding the absorption per wavelength, $A(\lambda)$, on top of the optically thick core of the projected planet, πR_p^2, divided by the projected disk of the star $\pi R_\star^2$:

$$\delta(\lambda) = \frac{\pi R_p^2 + A(\lambda)}{\pi R_\star^2} \tag{2.4}$$

where R_p and $R_\star$ are the planetary radius and the stellar radius, respectively.

2.2 Emission and reflection spectroscopy

Another way to study the composition of a gas is to observe its proper emission. In a planetary atmosphere, this is a direct measure of the vertical and thermal structure and its chemical composition. At zero-order approximation the amplitude of the emission of the planetary atmosphere with respect to the host star emission can be expressed by using the black body law (i.e. Plank function) (Tinetti et al., 2013)

$$\eta(\lambda) = \left(\frac{R_p}{R_\star}\right)^2 \frac{B_p(\lambda, T_p)}{B_\star(\lambda, T_\star)} \tag{2.5}$$

where λ is the specific wavelength of the observation, B_p and $B_\star$ are the black body laws for the planet and star respectively, T_p and $T_\star$ are the temperatures of the planet and the star.

For a more detailed discussion on the emission signal of a planetary atmosphere, we need to consider the absorption and emission processes ongoing in the atmosphere itself. For a non-scattering atmosphere in local thermodynamic equilibrium (good approximation in IR light), a beam of light of intensity $I(\lambda)$ emitted by the planet can be described by using the Schwarzschild's equation (Eq. 2.6 or 2.7) (Böhm-Vitense, 1992; Sportisse, 2009)

$$\mathrm{d}I(\lambda) = -\left(I(\lambda) - B_p(\lambda, T_p)\right) k(\lambda)\rho \mathrm{d}z \tag{2.6}$$

$$I(\lambda, z) = I_0(\lambda)e^{-\tau(\lambda,z)} + \int_0^{z_{max}} B_p(\lambda, T_p(z))e^{-\tau(\lambda,z)}k(\lambda)\rho \mathrm{d}z \tag{2.7}$$

where $k(\lambda)$ is the absorption coefficient, ρ is the density of the gas, z is the altitude of the atmosphere, $B_p(\lambda, T_p(z))$ is the Planck function of the planetary atmosphere and τ is the optical depth.

The first term on the right-hand side of the equal sign, on both Eq. 2.6 and 2.7, is the Beer-Bouguer-Lambert law and it describes the absorption process that the light emitted by the source undergoes in the atmosphere. The second term is the source function which describes the emission of thermal radiation along the optical path. Moreover, it also unveils insights on the vertical thermal structure of the atmosphere given by the presence on the temperature vertical profile, $T_p(z)$ inside the Planck function.

Let us consider two cases: (a) $\tau(\lambda,z) \ll 1$, i.e. the atmosphere has a small optical depth. This is the case of *optically thin* atmosphere. In this case, the source function in Eq. 2.7 is predominant with respect to the absorption term. (b) $\tau(\lambda,z) \gg 1$, i.e. *optically thick* atmosphere. In this case, the absorption term is larger than the source function and on the contrary, we expect to observe absorption lines (Böhm-Vitense, 1992).

For a more general discussion covering a broader interval of wavelength (not limited to solely IR) we need to take into account also scattering processes. The effect of such processes is to change only the direction of a photon. Therefore, we need to add two terms to Eq. 2.6 that takes into account the incoming flux of the star scattered and the scatter of the light coming from deeper layers of the atmosphere (Liou, 2002; Sportisse, 2009)

$$\begin{aligned}\frac{\mathrm{d}I(\lambda)}{\mathrm{d}z} = -I(\lambda) + \omega_a B_p(\lambda, T_p(z)) + \frac{\omega_d}{4\pi} F_\star e^{-\tau} P(\Omega,\Omega') + \\ + \frac{\omega_d}{4\pi}\int P(\Omega'',\Omega''') I(\lambda,\Omega''')\mathrm{d}\Omega'''\end{aligned} \tag{2.8}$$

where ω_a and ω_d are respectively the absorption and scattering albedos, Ω is the solid angle, $P(\Omega,\Omega')$ and $P(\Omega'',\Omega''')$ are the phase function which describe the probability of a photon to be scattered in a given direction (i.e. as a function of incident and scattering solid angle), $I(\lambda,\Omega''')$ is the intensity of the light scattered toward the solid angle Ω''' and $F_\star$ is the incoming flux from the star.

Three scattering regimes are usually distinguished: the Rayleigh scattering, the scattering represented by the optical geometry's laws and the so-called Mie scattering (Sportisse, 2009). To determine which of the three aforementioned regimes is stronger than others, we need to compare the *characteristic size*, d, of a molecule with the wavelengths, λ, of the observation. (a) if $d \ll \lambda$ the Rayleigh scattering is predominant; (b) if $d \gg \lambda$ we are in the optical geometry scattering regime, finally, (c) if $d \simeq \lambda$ the Mie scattering regime takes place (Sportisse, 2009).

Finally, another way to study the properties of planetary atmospheres is to consider the reflection spectroscopy. The light emitted by the host star not only is absorbed (Sec. 2.1) but in part it is also reflected by the planetary atmosphere. The amount of reflected light depends on the reflectivity of the planet, which is called *albedo*, a. At zero-order the reflected component is given by (Tinetti et al., 2013):

$$\eta(\lambda) = \left(\frac{R_p}{R_\star}\right)^2 a\zeta \left(\frac{R_\star^2}{D^2}\right) \frac{F_\star(\lambda)}{F_\star(\lambda)} = \left(\frac{R_p^2}{D^2}\right) a\zeta \tag{2.9}$$

where a is the albedo, D is the semi-major axis, ζ is the fraction of the projected planet disk illuminated by the star and R_p and $R_\star$ are respectively the planetary and stellar radii.

2.3 Spectral modelling

To interpret the spectra obtained analysing the data presented in the following chapters model spectra needed to be synthesised. The processes described in Sec. 2.1 and 2.2 are implemented in the $\mathscr{T}$-REx code (Waldmann et al., 2015b,a) which is a fully Bayesian inverse atmospheric retrieval framework. This can work in *forward* and in *inverse* mode. The former simulates the transmission or the emission spectrum of a planetary atmosphere given the chemical constituents of the atmosphere and physical parameters of the target system. The latter, on the other hand, is able to interpret an input spectrum, obtaining information on the chemical composition, in particular, the abundances of the constituents and orbital and physical parameters of the target system.

In transmission spectroscopy the key information is carried by the light after being filtered by the planetary atmosphere. The atmospheric spectrum model generated by $\mathscr{T}$-REx describes the transit depth, $(R_p/R_\star)^2$, as a function of wavelength. In emission spectroscopy the atmospheric model is expressed as ratio of the fluxes of the planet and the star $(F_p/F_\star)$.

On the processes reported in Sec. 2.1 and 2.2, the calculation of the optical depth, $\tau(\lambda,z)$, is the most complicated. This term reflects the geometrical structure of every single different molecular species considered in the atmosphere. The cross-sections, $\sigma(\lambda)$, used to calculate the optical depth, $\tau(\lambda,z)$, have been computed from molecular line lists obtained from ExoMol (Tennyson et al., 2016), HITRAN (Rothman et al., 2009) and HITEMP (Rothman et al., 2010). $\mathscr{T}$-REx is designed to work both with transmission or emission cross-sections. The optical depth, $\tau(\lambda, z)$, is given by

$$\tau(\lambda, z) = \sum_{m=1}^{N_m} 2 \int_0^{l(z)} \varsigma_m(\lambda) \chi_m(z) \rho_N(z) \mathrm{d}l \tag{2.10}$$

where m represents each absorbing molecule, N_m is the total number of molecules, $\varsigma_m(\lambda)$ is the absorption cross-section, $\chi_m(z)$ is the mixing ratio, $\rho_N(z)$ is the number density, and $l(z)$ is the optical path length in the atmosphere of the planet. The main molecular line lists considered in $\mathscr{T}$-REx are relative to: H_2O (Barber et al., 2006), CO (Rothman et al., 2010), CO_2 (Rothman et al., 2010), CH_4 (Yurchenko & Tennyson, 2014), NH_3 (Yurchenko et al., 2011), VO (McKemmish et al., 2016), HCN (Barber et al., 2014), and TiO (McKemmish et al., in prep). Inside $\mathscr{T}$-REx calculations, effects produced by Rayleigh scattering, the collision-induced absorption of H_2-H_2 and H_2-He and presence of clouds are also included (Waldmann et al., 2015b,a; Borysow et al., 2001; Borysow, 2002; Lee et al., 2013).

2.4 Transit method

As introduced in Sec. 1.2, a transit phenomenon occurs when a celestial body passes in front of another one with respect to our line of sight. In the solar system, transits of Venus and Mercury have been observed. These are special events since they are the closest that we can observe. These transits not only provide information on the planets but also insights into the photosphere and corona of our Sun (Mura et al., 2009; Chiavassa et al., 2015; Reale et al., 2015).

The same event can occur when the orbital plane of an exoplanet is aligned with our line of sight. As commonly reported in literature (Winn, 2010), when a planet crosses in front of the disk of its host star, we call this event *primary transit*. On the contrary, a passage behind the star with consequent occultation of the planet is called *secondary transit* or *eclipse*. In both cases, information on the planetary atmosphere are extracted from the differences in the flux recorded before one of the two events and in the middle of the event itself. During the primary transit the observed planet shows us its night side and terminator, on the contrary the day side is shown in proximity of the eclipse. In between these two events, the planet, while orbiting its host star, progressively changes the visible phase. The modulation observed as function of its orbital position is known as phase-variations or phase-curve (Fig. 2.2) (Knutson et al., 2007b, 2009b; Laughlin et al., 2009; Cowan et al., 2012; Demory et al., 2016; Krick et al., 2016).

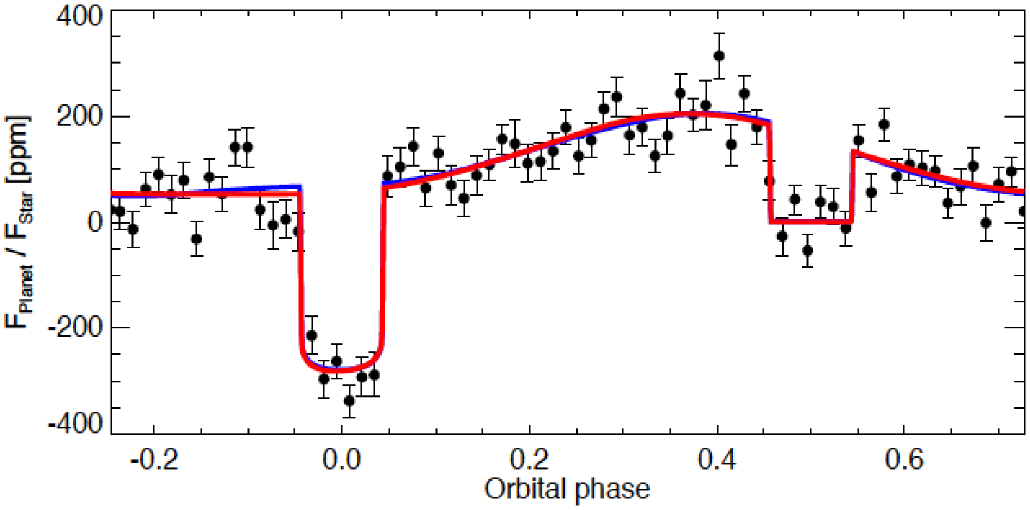

Figure 2.2: Phase-curve of the hot super-Earth 55 Cancri e (Demory et al., 2016).

During the primary transit we can use the calculations described in Sec. 2.1 to obtain information on the composition of the atmosphere and on the abundance of its chemical constituents. The discussion presented in Sec. 2.1 is only valid for those planets that transit their host stars.

The eclipse and the phase-curve variation can be studied with the equations presented in Sec. 2.2 and they can provide insights into the thermal structure and chemical compositions of planetary atmospheres. Phase-curve observations are not limited to transiting planets. Orbits inclined more than 85° are also expected to show phase-variation and thermal emission (Harrington et al., 2006; Crossfield et al., 2010).

2.4.1 Primary transit observations

This method has provided the largest number of results. Observing a transit at different wavelengths allows us to probe different parts of a planetary atmosphere.

Observations in the UV range show the behaviour of the upper atmosphere and the interaction with stellar radiation (Reale et al., 2015). We can observe if a planetary atmosphere is in equilibrium or in hydrodynamic escape regime (Vidal-Madjar et al., 2003; Ben-Jaffel, 2007; Linsky et al., 2010). Detecting ions suggests in fact that these species are moving out up to the Roche lobe and beyond (Vidal-Madjar et al., 2004, 2013).

Moving to longer wavelengths, i.e. optical range, the transit technique shows the presence of atomic species. In particular, the attention is driven on sodium and potassium which have strong absorption/emission lines (sodium resonance doublet at 5893 Å, K line at 7665 Å) (Charbonneau et al., 2002; Redfield et al., 2008; Huitson et al., 2012; Sing et al., 2011, 2015) (see Fig. 2.3 and Fig. 2.4). The peak values are related to the abundances of sodium and potassium, but the shape of the features (wing shape) is linked to the presence of clouds or hazes, which could mask part of these features.

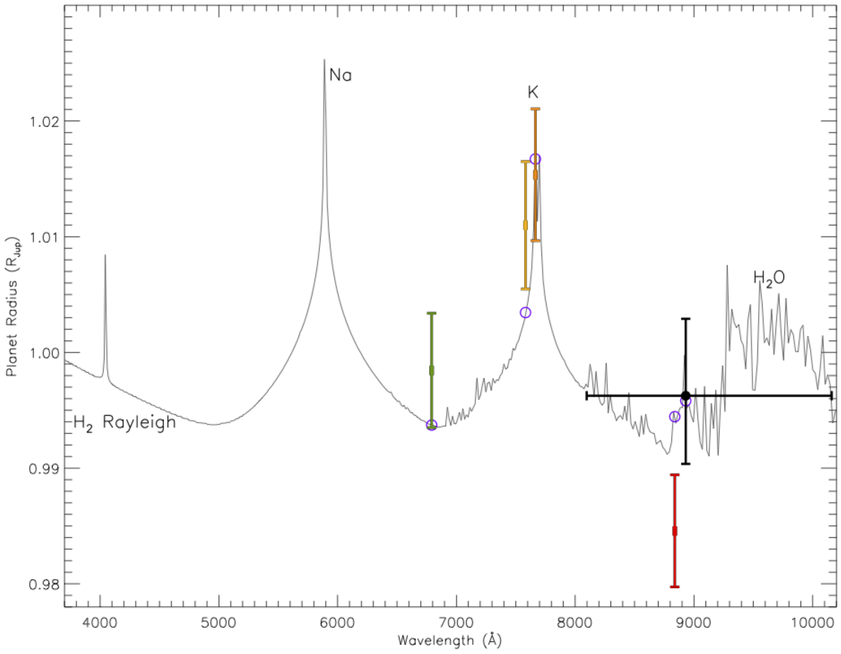

Figure 2.3: Detection of potassium in the atmosphere of XO-2b (Sing et al., 2011).

At even longer wavelengths in the near-IR and mid-IR molecules are active due to their roto-vibrational energetic bands. Space observations have been performed by using Spitzer Space Telescope (SST) and Hubble Space Telescope (HST). HST/WFC3 performs observations in the near infrared ($1.1-1.7\mu$m), while SST/IRAC provided observations at longer wavelengths in four photometric channels (3.6, 4.5, 5.8 and 8.0μm). The majority of the planets observed to date are hot and gaseous, as they are the easiest targets to probe. Transit observations in the IR have started to provide important insights into the chemical composition and structure of the atmospheres of gas giants orbiting very close to their star. Common atmospheric components detected include water vapour (e.g. Barman (2007); Tinetti et al. (2007); Grillmair et al. (2008); Deming et al. (2013); Fraine et al. (2014); Kreidberg et al. (2014); Sing et al. (2016); Damiano et al. (2017); Tsiaras et al. (2016a,b, 2018)) (Fig. 2.4). Condensates or hazes have also been identified (e.g. Knutson et al. (2014a); Sing et al. (2016)). Some of the data also suggest that carbon-bearing or more exotic species, such as TiO and VO (e.g. Swain et al. (2009a,b); Snellen et al. (2010); Evans et al. (2016); Line et al. (2016)), are present in some of these atmospheres. Finally, CH_4 has also been detected from space observations in the atmosphere of HD189733b (Swain et al., 2008; Waldmann et al., 2012).

Other molecules are difficult to be detected since the spectral resolution of space instruments do not allow to disentangle degeneracies when bands of different molecules overlap in the same wavelength range such as CO, CO_2 and CH_4 itself.

2.4.2 Eclipse and phase curve observations

Observations of the emission spectra provide insights into the vertical temperature-pressure profile of the atmosphere. If a stratosphere is present, for example, molecular features might be seen either in absorption or in emission (Encrenaz et al., 2004; Tinetti et al., 2013).

Combining the near-infrared with the mid-infrared, spectra observed during the eclipse show modulation due to the presence of H_2O, CH_4, CO and CO_2. These results have been achieved using both Spitzer IRAC/IRS/MIPS (Deming et al., 2005,

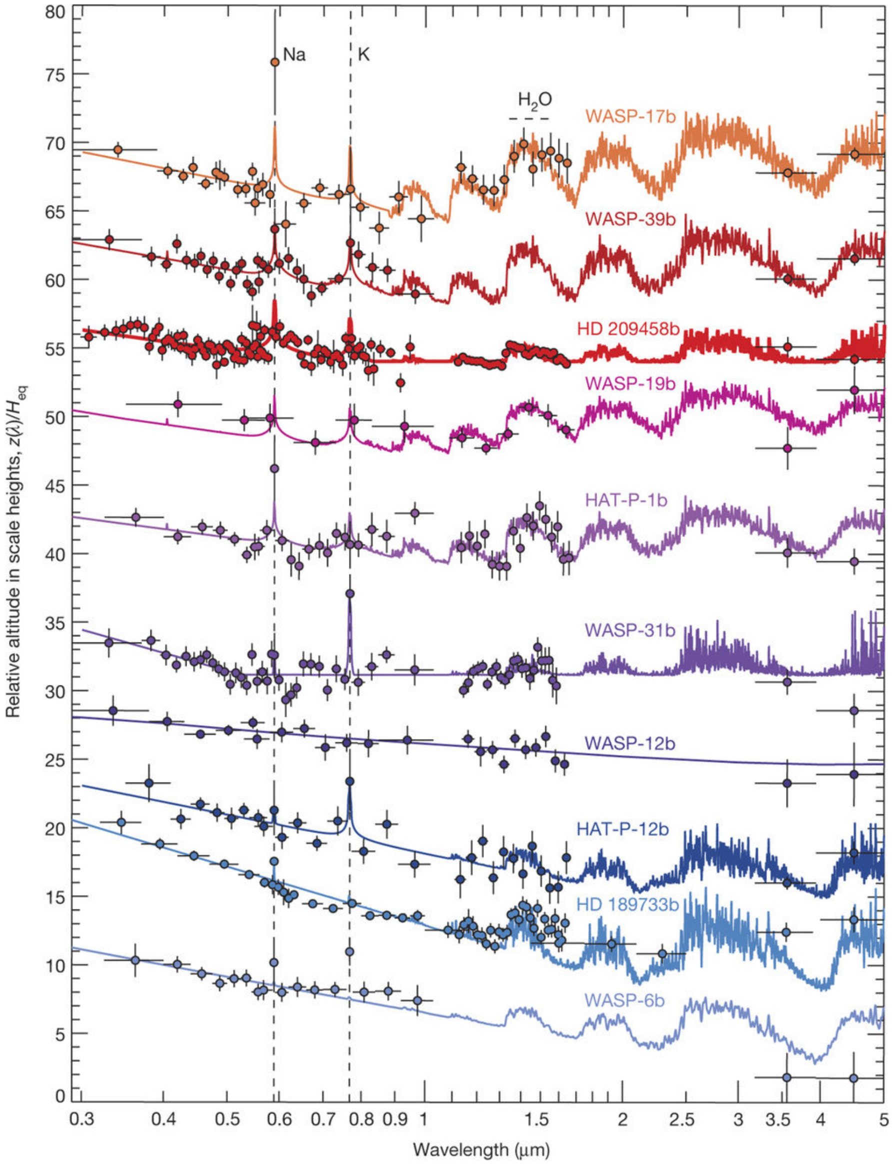

Figure 2.4: First consistent population study using observations from SST/IRAC, HST/WFC3 and HST/STIS (Sing et al., 2016).

2006; Knutson et al., 2007b; Charbonneau et al., 2008; Grillmair et al., 2008; Swain et al., 2008) and Hubble NICMOS (Swain et al., 2009a,b).

Other studies have suggested thermal inversion on HD209458b, TrES-4b and HAT-P-32b (Charbonneau et al., 2008; Knutson et al., 2008, 2009a; Zhao et al., 2014), for HD209458b, however, there are some studies claiming the contrary (Evans et al., 2015; Hoeijmakers et al., 2015). The possible correlation of thermal inversion with the presence of molecules such as TiO and VO has been reported by

Spiegel et al. (2009).

Finally, spectroscopic observations of phase-variations are very useful since they provide at the same time information on the chemical composition and on the vertical and horizontal thermal profile of the atmosphere. So far, only one planet has been studied with this technique. Stevenson et al. (2014) observed the hot-Jupiter WASP-43b which has a very short period (P=0.813 days). They have extracted the planetary emission spectrum and finally retrieved the temperature-pressure profile at different orbital phases of the planet.

2.5 Direct imaging

The planetary population observed through direct imaging is different from the objects observed through transit or eclipse The imaged planets are distant from their parent stars and they are also hot and young. To perform such observations dedicated instruments have been developed, equipped with an integrated field spectrograph. Also this technique relies on observation of the emission/reflection spectra (Sec. 2.2) to unveil the constituents of the atmosphere of the targetted planets.

The first spectrum obtained with direct imaging was recorded using the instrument NACO mounted on ESO Very Large Telescope (VLT) (Janson et al., 2010). Subsequent observations of the same target (HR 8799c) have highlighted the presence of carbon monoxide and water in its atmosphere (Konopacky et al., 2013). Using this technique, 51 Eri b (Fig. 2.5) (Macintosh et al., 2015; Rajan et al., 2017; Samland et al., 2017) and HD 131399Ab (Wagner et al., 2016) have also been studied.

One limitation of spectra obtained through direct imaging, is the unknown mass and radius of the planet. If these parameters are poorly constrained, the uncertainties on the atmospheric structure and chemical abundances are much larger.

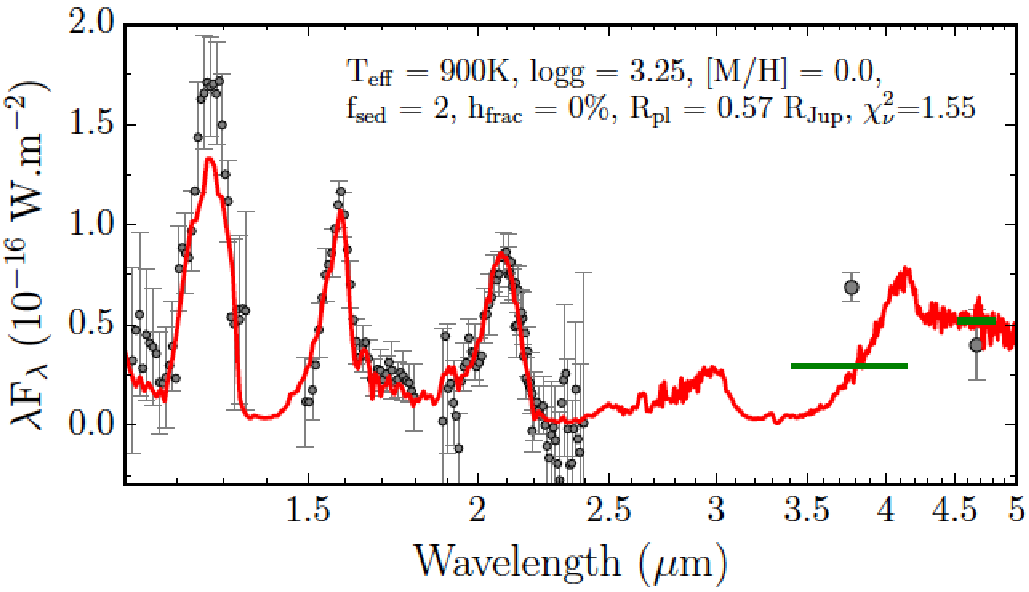

Figure 2.5: The infrared spectrum of 51Eri b. The picture is part of the work of Rajan et al. (2017).

2.6 High-Resolution Spectroscopy

In previous sections (in particular Sec. 2.4) I have discussed some details and achievements of the transit spectroscopy, however, the resolution was not explicitly discussed. Observations performed from space-based instruments are made at low spectral resolution ($\lambda/\Delta\lambda = R < 300$). One limitation is the degeneracy of the results when bands of different molecules overlap if the spectral range probed is not broad enough. A different approach has been pioneered by Snellen et al. (2010) using the ground-based VLT/CRIRES instrument (CRyogenic high-resolution InfraRed Echelle Spectrograph). In their work a first detection of CO in the atmosphere of HD209458b was presented.

This technique combines transit spectroscopy (see Sec. 2.1 and 2.2) with the radial velocity technique (see Sec. 1.1) to sound the atmosphere of the planet and it is based on the fact that high-resolution spectroscopy (HRS) ($\lambda/\Delta\lambda = R > 10'000$) resolves molecular bands into individual lines. Fig. 2.6 shows the change on the information shown by a spectrum when different resolutions are explored.

The main idea is to disentangle the planetary signal from its host star signal,

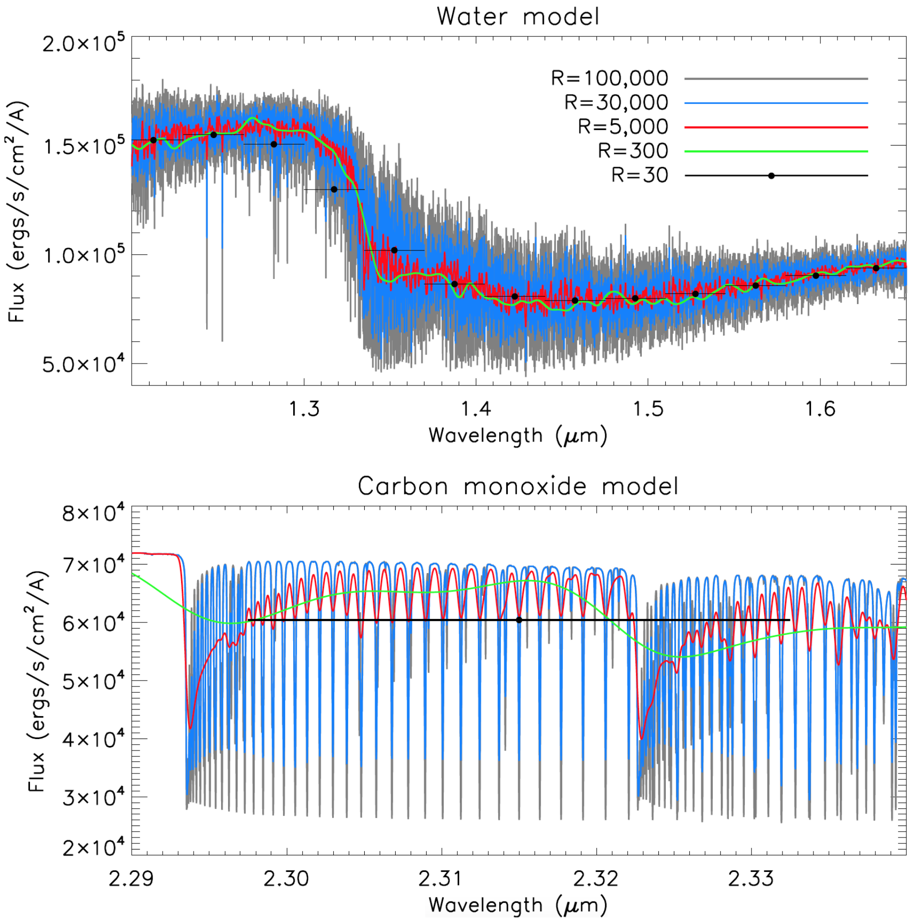

Figure 2.6: **Top panel:** synthetic H_2O spectrum at different resolutions. **Bottom panel:** synthetic CO spectrum at different resolution (Birkby, 2018).

the Earth absorption (telluric absorption), because observations are performed in ground-based observatories, and correlated instrumental noise. This is possible by using a radial velocity approach because the amplitude of the Doppler shift of the planetary signal (ΔRV$\sim$ km s^{-1}) is higher than the host star signal (ΔRV$\sim$ m s^{-1}) and the telluric absorption (rest frame of the observer). Finally, the extracted planetary signal is compared with different atmospheric models to determine which one correlates more to give us information on the chemical composition.

2.6.1 HRS observations

The idea of using high resolution spectroscopy to sound exoplanetary atmosphere was explored not long after the discovery of 51 Peg b (Mayor & Queloz, 1995). First attempts, using high-resolution spectroscopy in the optical wavelengths, have

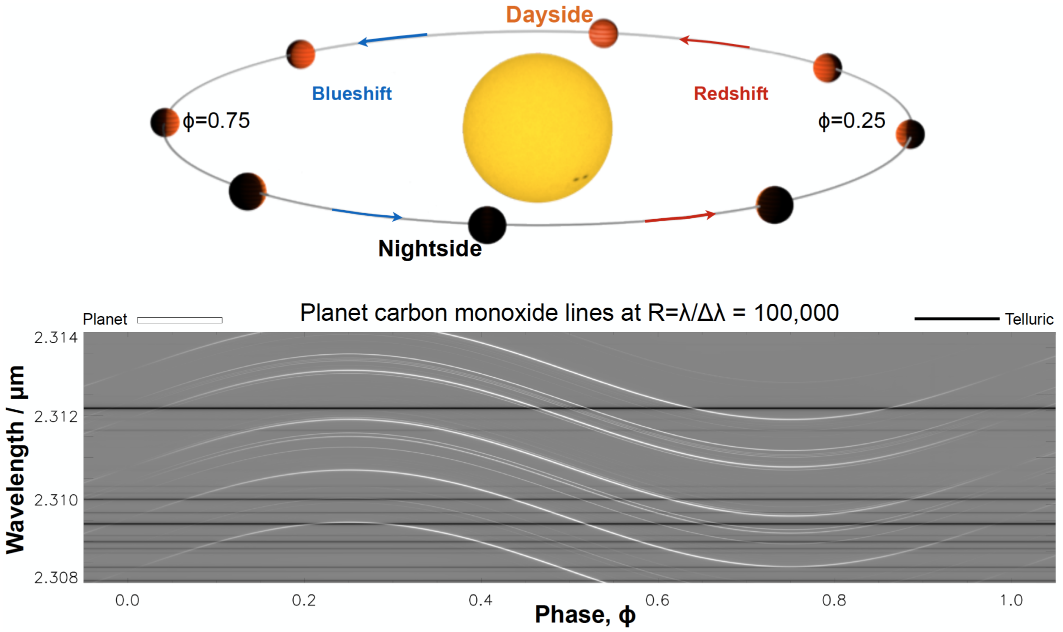

Figure 2.7: **Top panel:** motion of the planet around the host star. **Bottom panel:** Doppler shift of the planetary CO signal as a function of the orbital phase. The planetary signal is identified without degeneracies due to higher radial velocity amplitude (Birkby, 2018).

found upper limits (Charbonneau et al., 1998, 1999; Collier Cameron et al., 1999). Multiple attempts have been made to study exoplanetary atmosphere in the IR but again only upper limits have been obtained (Brown et al., 2002; Deming et al., 2005; Barnes et al., 2007a,b).

In 2010 using VLT/CRIRES combined with adaptive optics (MACAO at VLT) which improved the stability of the data, CO and H_2O have been detected on transiting planets (i.e. HD209458b (Fig. 2.8), HD189733b) (Snellen et al., 2010; de Kok et al., 2013; Birkby et al., 2013; Brogi et al., 2016, 2018) and non-transiting planets such as 51Peg b, τ-Bootis b, HD179949b (Brogi et al., 2013; Snellen et al., 2014; Brogi et al., 2014; Birkby et al., 2017). A more exotic molecule such as TiO has been recently detected on WASP-33b (Nugroho et al., 2017). An attempt to detect sodium and calcium on the atmosphere of 55Cnc e has been also done by using this technique in the optical wavelength range (Ridden-Harper et al., 2016).

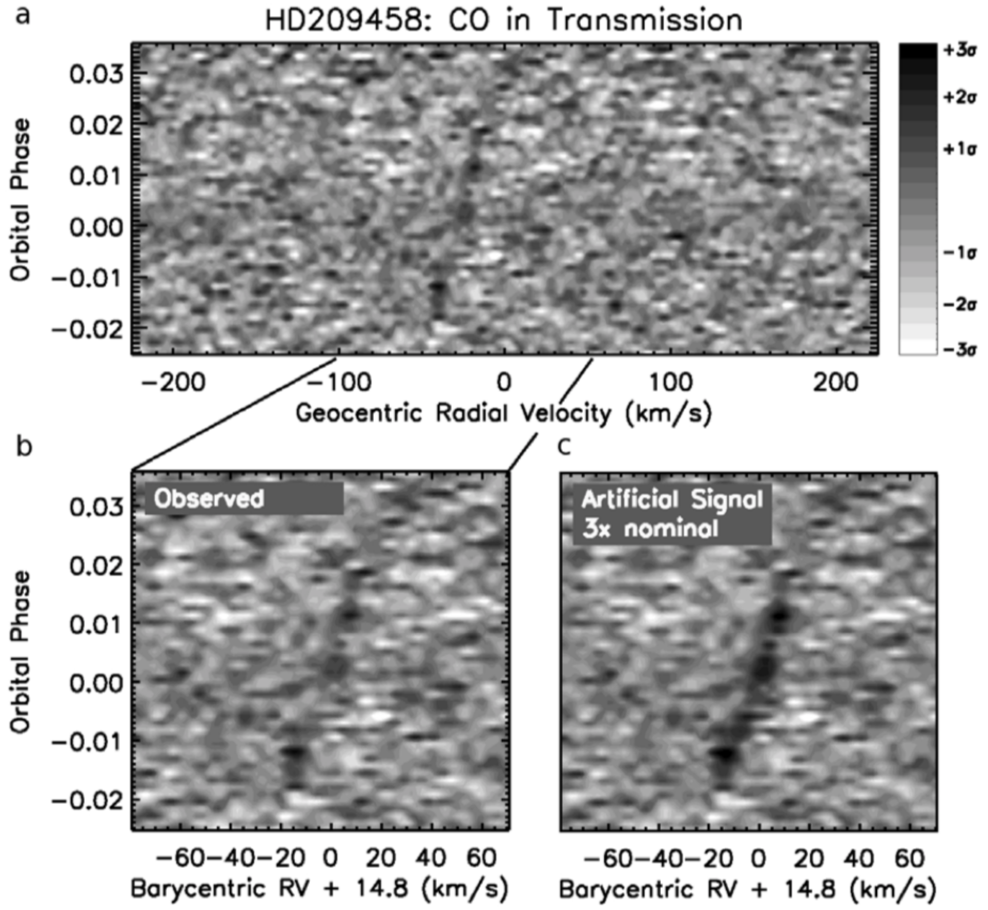

Figure 2.8: CO signal in transmission in the atmosphere of HD209458b after the data have been cross-correlated with the synthesised CO model. Figure adopted from Snellen et al. (2010).

2.7 High and low resolution

From the previous sections (Sec. 2.4 and Sec. 2.6) it is clear that the two spectroscopic techniques are used in two different regimes. Low-resolution spectroscopy ($\Delta\lambda/\lambda = R < 300$) is accessible from both space- and ground-based facilities. However, water and other chemical compounds, present in the Earth's atmosphere, contaminate and cover the signal coming from the target which may contain same molecules. Over a broad wavelength range observations can only be performed by satellites.

High-resolution ($\Delta\lambda/\lambda = R > 10'000$) can only be performed from ground-based observatories. Very Large Telescope(VLT) has four telescopes with 8 m primary mirrors and the next generation observatory, the European Extreme Large Telescope (E-ELT) is going to be equipped with a 39 m primary mirror. For these reasons, these facilities can only be built on the ground.

In the next chapters (chapter 4 and chapter 5) I will give details on low- and high-resolution methods describing two pipelines for data analysis that work in these two regimes.

Chapter 3

Signal decomposition

"How can a bunch of dumb particles moving around according to the laws of physics exhibit behavior that we'd call intelligent?"

– Max Tegmark - Life 3.0 - 2017

The technological progress has delivered valuable instruments not only to simplify our job but also to improve the performances and the accuracy we can reach. Data analysis in the scientific domain has to correct for systematics that are not always kept into account by reduction procedures. Data analysis is therefore sometimes purely experiment, because unknown systematics and errors can be present. These effects can sometimes be treated ad-hoc, with manual intervention. However, this is not the best scenario since this procedure could introduce biased, non reproducible and therefore non-scientific results.

In 1959 for the first time the word *machine learning* was used (Samuel, 1959) to describe an ensemble of techniques aimed learning trends from the data. Today we have moved forward by developing an entire new branch of algorithms capable of mimicking the human brain with astonishing results in a variety of fields (Krizhevsky et al., 2012; Goodfellow et al., 2014; Lecun et al., 2015; van den Oord et al., 2016).

Some of these algorithms are currently used in the astrophysical environment producing successful results. In the spectral retrieval context, for example, pattern recognition could help to pre-select molecules and abundances to be fed as prior to a retrieval code (Waldmann, 2016). More interestingly, the entire classical retrieval process can be substituted with a quantitative pattern recognition (Zingales & Waldmann, 2018).

The instrumentations used to perform observations are not perfect and they contribute to introduce errors and systematics that can heavily affect the data. It is common, for example, that infrared detectors produce a ‘charge-trap’ that prevents electrons to be read and properly counted creating ‘ramps-type’ systematics in the light-curve (see chapter 4 and Agol et al. (2010)). The sub-optimal solution to this is to fit an ad-hoc model to correct the data. A better approach would be let a machine ‘understand’ the trends and effectively remove them.

Time-series systematic effects may be associated, for example, with the varying atmospheric conditions, the variability of the detector efficiency or point spread function (PSF) changes. However, these effects might vary from star to star, depending on the stellar colour or on the position of the star on the CCD, a dependence which is not always known. Therefore, the removal of such effects might not be trivial.

For these reasons in this thesis I have worked on two automated pipelines. One is able to analyse data recorded by the Hubble Space Telescope (see chapter 4) resolving also the aforementioned ‘charge-trap’ issue. Alongside that pipeline I will describe an alternative way that has been used involving Independent Component Analysis (ICA) to demonstrate that unsupervised machine learning algorithms could be an alternative solution for interpreting time-series data.

In chapter 5 I will discuss a different pipeline which effectively uses Principal Component Analysis (PCA) (see Sec. 3.2) to de-trend data from time variation systematics. PCA is an unsupervised linear transformation technique that is widely used across different fields, most prominently for feature extraction, dimensionality reduction, exploratory data analyses and de-noising of signals (Jolliffe, 2002). This is an optimal algorithm to analyse high-resolution spectra (chapter 5) that are affected by time-series systematics (e.g. airmass variation).

An additional algorithm discussed is SYSREM (see Sec. 3.3) that can remove systematic effects, such as those associated with atmospheric extinction, detector efficiency, or point spread function changes over the detector. PCA and SYSREM algorithms have been compared on same datasets in chapter 5.

3.1 Observing a Cocktail Party

Blind-source separation is a technique that allows to disentangle independent or uncorrelated sources. This process is also known as 'Cocktail Party Problem' (Hyvärinen & Oja, 2000; Hyvärinen, 2012). Imagine that there are n people talking to each other in a room. The speech signals of these people are indicated by $s_1(t)$, $s_2(t)$ ... $s_n(t)$. In the same room there are m microphones that record the signals. The observed signals $x_1(t)$, $x_2(t)$ ... $x_m(t)$ can then be expressed as:

$$\begin{aligned}
x_1(t) &= a_{11}s_1(t) + a_{12}s_2(t) + \ldots + a_{1n}s_N(t) \\
x_2(t) &= a_{21}s_1(t) + a_{22}s_2(t) + \ldots + a_{2n}s_N(t) \\
\vdots \\
x_m(t) &= a_{m1}s_1(t) + a_{m2}s_2(t) + \ldots + a_{mn}s_N(t)
\end{aligned} \tag{3.1}$$

where a_{mn} is the weighting factor (in the example proposed is the distance of the speaker to the microphone, squared) and m, n are respectively the number of observations and the total number of signals. Eq. 3.1 can be expressed in a compact form using a matrix notation:

$$\boldsymbol{x} = \boldsymbol{A} \cdot \boldsymbol{s} \tag{3.2}$$

where the rows of $\boldsymbol{x}$ contain the individual time series, x_m, $\boldsymbol{s}$ is the signal matrix of the individual source signal s_n and finally, $\boldsymbol{A}$ is the mixing matrix. The Eq. 3.2 is often recognised as the classical 'Cocktail Party Problem' (Hyvärinen & Oja, 2000; Hyvärinen, 2012). The challenge is to estimate both the mixing matrix and the signal matrix which are unknown using no prior information but the input matrix $\boldsymbol{x}$.

Many algorithms have been proposed that address the described problem and among them we can find PCA (Pearson, 1901; Oja, 1992; Manly, 1994; Jolliffe, 2002; Press et al., 2007), ICA (Comon, 1994; Hyvärinen, 1999; Hyvärinen & Pajunen, 1999; Hyvärinen & Oja, 2000; Hyvärinen, 2001; Comon & Jutten, 2010).

3.2 Principal Component Analysis

PCA is a technique for data simplification used in multivariate analysis. The method was theorised at the beginning of '900. Today it is a well established algorithm being useful in a variety of fields, from finance to research and it has been classified into the family of machine learning algorithms. PCA can be used to find correlations in the data with no prior knowledge but the data itself. The aim is to decompose signals into their most statistically significant elements.

Let $\boldsymbol{X}$ be a non squared matrix of dimension n, we use feature extraction to transform the data into a new d-dimensional space. As a result of transforming the original n-dimensional data into this new d-dimensional space, the first principal component will have the largest possible variance (the most statistically significant direction), and all the other principal components will have decreasing value of variance given the constraint that these components are orthogonal to the other principal components. Even if the input features are correlated, the resulting principal components will be mutually orthogonal and therefore uncorrelated (Jolliffe, 2002).

A key point to keep in mind is that the principal component process decomposition is highly sensitive to data scaling. We need to rescale the data if the features were measured on different scales and we need to assign equal importance to all the features.

Among the rescaling methods the most popular are *standardisation*, *normalisation* and *MinMax* scaling. The first one consists on subtracting the mean of each feature and dividing by its standard deviation in order to have the mean equal to zero and the standard deviation equal to one. The *normalisation* consists, as the name suggests, on dividing every feature by its mean. Finally, *MinMax* consists on taking the max and the minimum of the data and map those into a new range typically [0,1] (Raschka & Mirjalili, 2017).

The scaling method to choose depends on the problem analysed. If the features of the matrix are not measured in the same unit (e.g. survey on commercial products that takes into account different characteristics) the standardisation is the best option. In some cases one could also choose a custom scaling depending on its set-up, for example dividing a spectrum in different bins and applying *MinMax* on each singular bin (Zingales & Waldmann, 2018).

The PCA algorithm used in this work has been developed from scratch to understand the process on its every single step (see Sec. 5.4). I summarise here the necessary steps (Jolliffe, 2002; Raschka & Mirjalili, 2017):

1. apply a pre-process method for feature scaling;
2. calculate the covariance matrix Σ;
3. decompose the covariance matrix Σ into eigenvectors $\boldsymbol{v}_k$ and eigenvalues λ_k
4. calculate the projection matrix $\boldsymbol{W}$ from the eigenvectors
5. derive the new projected space $\boldsymbol{X}$'
6. reconstruct the original space with less components.

After applying one of the pre-processing methods aforementioned, the covariance matrix can be calculated. The covariance σ of two features $\mathbf{x}_j$ and $\mathbf{x}_z$, measures their joint variability. In the case of a matrix the correspondents covariance matrix measures how the elements are correlated with each other. Each element of such matrix is calculated

$$\sigma_{jz} = \frac{1}{n}\sum_{i=1}^{n}\left(x_j^{(i)} - \mu_j\right)\left(x_z^{(i)} - \mu_z\right) \tag{3.3}$$

where μ_j and μ_z are the mean of each respective feature j and z, n is the total number of elements (i) contained into each feature array. If a covariance σ_{jz} is positive, then the two elements vary in accordance, i.e. increasing or decreasing together. On the other hand, in case of negative covariance, they vary in the opposite direction. The covariance matrix Σ is a $k \times k$ matrix and has the property of being symmetric and normal. For this, we can calculate a basis of eigenvectors that satisfy the following relation:

$$\left(\Sigma_{k\times k} - \lambda_k \boldsymbol{I}_{k\times k}\right)\boldsymbol{V}_{k\times k} = 0 \tag{3.4}$$

where Σ is the covariance matrix, λ_k are the eigenvalues that will define the eigenvectors' variance, $\boldsymbol{I}$ is the diagonal identity matrix and $\boldsymbol{V}$ is a matrix that contains eigenvectors as columns. These eigenvectors are also known as *principal components*.

Since the eigenvalues define the importance of the eigenvectors, one can use them as indicator of which eigenvector has most information content. One can define the *explained variance ratio* (EVR)

$$EVR_j = \frac{\lambda_j}{\sum_{j=1}^{k} \lambda_j} \tag{3.5}$$

In a nutshell EVR indicates in percentage how much information of the data is stored in an eigenvector. After eigenvectors are sorted in decreasing EVR, the projection matrix $\boldsymbol{W}$ is constructed as follows: d eigenvectors are chosen storing them as columns in a new matrix. The projection matrix is a $k \times d$ matrix where k are the features of the original matrix (length of eigenvector array), and d are the features of the new one (number of eigenvectors chosen). If we choose the whole set of eigenvectors so that $k = d$; the projected space $\boldsymbol{X'}$ is equal to (Raschka & Mirjalili, 2017):

$$\boldsymbol{X'}_{n\times k} = \boldsymbol{X}_{n\times k}\boldsymbol{W}_{k\times k} \tag{3.6}$$

$\boldsymbol{X'}$ is similar to the original dataset; it has the same dimension, but the data is rotated into a different coordinates frame where the axes are along the eigenvectors directions that maximise the variance.

Since the projection matrix $\boldsymbol{W}$ is composed of the eigenvectors $\boldsymbol{v}_k$ that are an orthogonal basis for the covariance matrix (symmetric and normal), $\boldsymbol{W}$ is an orthogonal matrix, i.e. $\boldsymbol{W}^{-1} = \boldsymbol{W}^T$, so the original space can be expressed by re-projecting $\boldsymbol{X'}$ back into $\boldsymbol{X}$ using the following equation:

$$\boldsymbol{X}_{n\times k} = \boldsymbol{X'}_{n\times k}\left(\boldsymbol{W}_{k\times k}\right)^T \tag{3.7}$$

At this point, if $k \neq d$, then the original space described with less component is obtained by taking d columns from ***X'*** and ***W***

$$\boldsymbol{X}_{n\times k} = \boldsymbol{X'}_{n\times d}\left(\boldsymbol{W}_{k\times d}\right)^{T} \tag{3.8}$$

3.3 SYSREM algorithm

The original formulation was published by Gabriel & Zamir (1979); Tamuz et al. (2005). More recently, it has been used to correct systematics among the data and in particular trying to correct the telluric absorption and its variability during the night mainly due to airmass variations (Birkby et al., 2013, 2017; Nugroho et al., 2017).

Let us define the input matrix $\boldsymbol{R}$ with i wavelength bins or features (rows) and j observations (columns). σ_{ij} is defined as the uncertainty of the element ij of the matrix $\boldsymbol{R}$ and it is calculated by computing the root sum square of the standard deviations of the corresponding row i and column j. Finally, we define a column array $\boldsymbol{c}$ and a row array $\boldsymbol{a}$ whose product $\boldsymbol{c}\cdot\boldsymbol{a}$ computes the approximation of $\boldsymbol{R}$. Giving the initial value $\boldsymbol{a}^{(0)}$ we then search for the optimal $\boldsymbol{c}^{(0)}$ that minimises

$$S_i^2 = \sum_j \frac{\left(r_{ij} - c_i^{(0)} a_j^{(0)}\right)^2}{\sigma_{ij}^2} \tag{3.9}$$

It is possible to evaluate $\boldsymbol{c}^{(0)}$ by differentiating the above equation

$$\boldsymbol{c}^{(0)} = \frac{\sum_j \left(r_{ij} a_j^{(0)} / \sigma_{ij}^2\right)}{\sum_j \left({a_j^{(0)}}^2 / \sigma_{ij}^2\right)} \tag{3.10}$$

With the same logic, $\boldsymbol{a}^{(1)}$ needs to be evaluated by minimising

$$S_j^2 = \sum_i \frac{\left(r_{ij} - c_i^{(0)} a_j^{(1)}\right)^2}{\sigma_{ij}^2} \tag{3.11}$$

and again by calculating the derivative of the above equation, we can resolve for $\mathbf{a}^{(1)}$:

$$\boldsymbol{a}^{(1)} = \frac{\sum_i \left(r_{ij} c_i^{(0)} / \sigma_{ij}^2\right)}{\sum_i \left({c_i^{(0)}}^2 / \sigma_{ij}^2\right)} \tag{3.12}$$

At this point we have to calculate again $\boldsymbol{c}^{(1)}$ by inserting the new $\boldsymbol{a}^{(1)}$ into Eq. 3.10. After n iterations the product $\boldsymbol{c}^{(n)} \cdot \boldsymbol{a}^{(n)}$ approximates well the original matrix **R** and thus can be subtracted: $\boldsymbol{R}^{(1)} = \boldsymbol{R} - \boldsymbol{c}^{(n)} \cdot \boldsymbol{a}^{(n)}$. The process starts again to calculate the second product that best represents $\boldsymbol{R}^{(1)}$.
The quantity that has to be minimised is

$$S^2 = \sum_{ij} \frac{\left(r_{ij} - c_i a_j\right)^2}{\sigma_{ij}^2} \tag{3.13}$$

Iteratively, then, the algorithm finds the systematics as long as the global minimum of S^2 is achieved. Eq. 3.13 is chi-square like where $\boldsymbol{R}$ is the observation, $\boldsymbol{c} \cdot \boldsymbol{a}$ is the model that should describe the observation and σ_{ij} is the uncertainty. If the uncertainties σ_{ij} are unitary, the algorithm is reduced to the conventional PCA. However, an important point is that this method does not guarantee the orthogonality of the calculated $\boldsymbol{a}$ (Tamuz et al., 2005).

Finally, unlike PCA, SYSREM does not provide simultaneously all the components, but they need to be calculated one by one iteratively. So, the more components are searched for the more calculation time the algorithm will require. Because of this issue the algorithm is kept to run only for the first 10-20 components (Birkby et al., 2013; Nugroho et al., 2017).

In chapter 5, I will compare PCA and SYSREM highlighting performances and limitations.

3.4 Independent Component Analysis

The general idea of ICA is to change the space from an m-dimensional to an n-dimensional space such that the new space with the transformed variables (components) describes the essential structure of the data having the constraint that the new components (new space) are mutually independent in complete statistical sense. Among its virtues, ICA has a good performance in pattern recognition, noise reduction and data reduction. Unlike PCA, the basis vectors in ICA are neither orthogonal nor ranked in order, and the algorithm is higher-order statistic since it retrieves components directly from input data.

While orthogonality is an intuitive concept (see Sec. 3.2), the concept of *statistical independence* is related to the *probability density function* (PDF) of a random variable. Variables are defined statistically independent if and only if the joint PDF is equals to the product of the single variables' PDF:

$$p_{x,y}(x,y) = p_x(x)p_y(y) \tag{3.14}$$

where x and y are random variables, p_x and p_y are the respective PDF and $p_{x,y}$ is their joint PDF.

In literature many algorithms can be found that implement the ICA technique. The one that we adopted for analysing HST observations (see chapter 4) is the MULTICOMBI algorithm (Tichavsky et al., 2008). MULTICOMBI is an hybrid model that performs the separation of non-Gaussian and/or time-correlated sources. This task is achieved by combining the strength of two popular algorithms, i.e. EFICA (Hyvärinen & Oja, 2000) and WASOBI (Yeredor, 2000).

3.4.1 Interference-to-Signal Ratio

Before describing the ICA algorithm, it is propaedeutic to introduce the *interference-to-signal ratio* (ISR) matrix. Knowing the mixing matrix $\boldsymbol{A}$ (see Eq. 3.2) and the de-mixing matrix $\boldsymbol{W}$, we evaluate the quality of the decomposition by defining the gain matrix $\boldsymbol{G}$:

$$\boldsymbol{G} = \boldsymbol{W} \cdot \boldsymbol{A} \approx \boldsymbol{I} \tag{3.15}$$

Ideally, the de-mixing matrix $\boldsymbol{W}$ is the inverse of the matrix $\boldsymbol{A}$ and the gain matrix is the unity matrix. In a real case problem the gain matrix $\boldsymbol{G}$ contains non-diagonal terms that account for the residuals among components. The $\boldsymbol{ISR}$ matrix is then defined as:

$$isr_{ij} = \frac{g_{ij}^2}{g_{ii}^2} \approx g_{ij}^2 \tag{3.16}$$

The specific interference-to-signal ratio associated to the i-th component is defined as

$$isr_i = \frac{\sum_{j=1, j \neq i}^{n} g_{ij}^2}{g_{ii}^2} \tag{3.17}$$

The ***ISR*** matrix, for a real case problem, can not be calculated from Eq. 3.16 since the mixing matrix ***A*** is unknown and the gain matrix ***G*** can not be obtained. However, for specific algorithms the ***ISR*** matrix can be estimated (Yeredor, 2000; Koldovsky et al., 2006; Tichavsky et al., 2006; Tichavsky et al., 2008).

3.4.2 COMBI and MULTICOMBI

EFICA (Hyvärinen & Oja, 2000) and WASOBI (Yeredor, 2000) are two algorithms able to estimate the de-mixing matrix ***W*** and allowing to calculate the independent components, **s**:

$$\boldsymbol{s} = \boldsymbol{Wx} \tag{3.18}$$

where $\boldsymbol{x}$ is the input data. They have been designed to treat observations with certain characteristics. EFICA is able to decompose a mixture of non-Gaussian signals, while, WASOBI is designed to separate Gaussian autoregressive signals. Real case problems may show both non-Gaussian and autoregressive behaviour, for this reason an hybrid model that takes into account the strengths of the two aforementioned models has been developed. The COMBI algorithm (Tichavsky et al., 2006) can be used as follows:

1. apply both EFICA and WASOBI to the input data $\boldsymbol{x}$: calculate the source signals $\boldsymbol{s}^{EF}$ and $\boldsymbol{s}^{WA}$, the respective interference matrices $\boldsymbol{ISR}^{EF}$ and $\boldsymbol{ISR}^{WA}$ and the correspondent vectors $\boldsymbol{isr}^{EF}$ and $\boldsymbol{isr}^{WA}$;

2. define $E = min\ isr_k^{EF}$ and $W = min\ isr_k^{WA}$;

3. retain all the source signals from $\boldsymbol{s}^{EF}$ such that $isr_k^{EF} < W$, and similarly retain all the source signals from $\boldsymbol{s}^{WA}$ such that $isr_k^{WA} < E$; reject the rest of the remaining signals;

4. if the rejected signals are more than one, start again from the first step.

The MULTICOMBI algorithm is a hierarchical multistep COMBI in which the input matrix is divided in clusters that are analysed separately (Tichavsky et al., 2008; Morello, 2015).

Chapter 4

Space observations

"The time will come when man will know even what is going on in the other planets and perhaps be able to visit them."

– Henry Ford - 1930

Hubble Space telescope was launched in 1990 and after almost 30 years it is still in operation thanks to instrument substitutions and upgrades. HST orbits in a low Earth orbit making possible for astronauts to reach it and replace its equipments. In 2009 the *Wide Field Camera 3* (WFC3) has been installed. This is a versatile instrument with two channels collecting light in 200 – 1000 and 800 – 1700 nm spectral windows and each channel is equipped with a variety of prisms and grisms enabling wide-field low-resolution spectroscopy. Initially, the only way to record exoplanetary spectra was using the *staring mode* but this was not efficient in observing bright targets because the detectors saturated fast. The spatial scanning technique has then been introduced to avoid this issue. During a spatial scanning exposure, the instrument moves slowly along the perpendicular direction of the dispersed spectrum instead of staring at the target. As a result, the total number of photons collected is much larger, increasing the signal-to-noise ratio (S/N), without the risk of saturation. However, because of geometrical distortions, the shifted staring-mode spectra, which construct each spatially scanned spectrum, are not identical to each other. This was either partially or not taken into account in previous analyses.

In this chapter, I present a stand-alone, dedicated pipeline, which is able to produce 1D spectra from the raw scanning-mode spectroscopic images. This pipeline uses a new method to calibrate and extract the 1D spectra, eliminating possible issues caused by the scanning process (Tsiaras et al., 2016a). Adopting such an approach allows the efficient analysis of even longer scans, thus extending the capabilities of the spatial scanning technique.

Said pipeline has been published in Tsiaras et al. (2016a) and my contribution consisted of: a multiple extraction aperture method and the introduction of spectral bins division with the goal of achieve similar flux per bin. During my Ph.D. I have updated said pipeline to analyse the near-infrared transit spectrum of the hot Jupiter HAT-P-32b (Teq = 1786 K; Hartman et al. (2011)) obtained with the WFC3 camera on board the HST (Damiano et al., 2017). In particular, the aforementioned dedicated WFC3 pipeline (Tsiaras et al., 2016a) was used to extract the transit light-curves per wavelength bin and obtain the planetary spectrum. Additionally, we used in parallel Independent Component Analysis to correct for the instrumental systematics, and we investigated the effect of different analysis techniques on the same dataset (Damiano et al., 2017). Finally, the obtained planetary spectrum was interpreted using the fully Bayesian spectral retrieval code, $\mathscr{T}$-REx (Sec. 2.3) (Waldmann et al., 2015b,a).

4.1 Code overview

The code was specifically designed and developed to properly analyse HST/WFC3 datasets recorded using spacial scanning-mode, from raw images to the final 1D spectrum. The code is entirely written in *python* and it is freely available on GitHub[1]. In Fig. 4.1 it is shown the flowchart of the HST/WFC3 pipeline. The analysis is mainly divided in three steps: reduction of the data, analysis of light-curves and systematics and finally, generation of the outputs.

[1] https://github.com/ucl-exoplanets/Iraclis

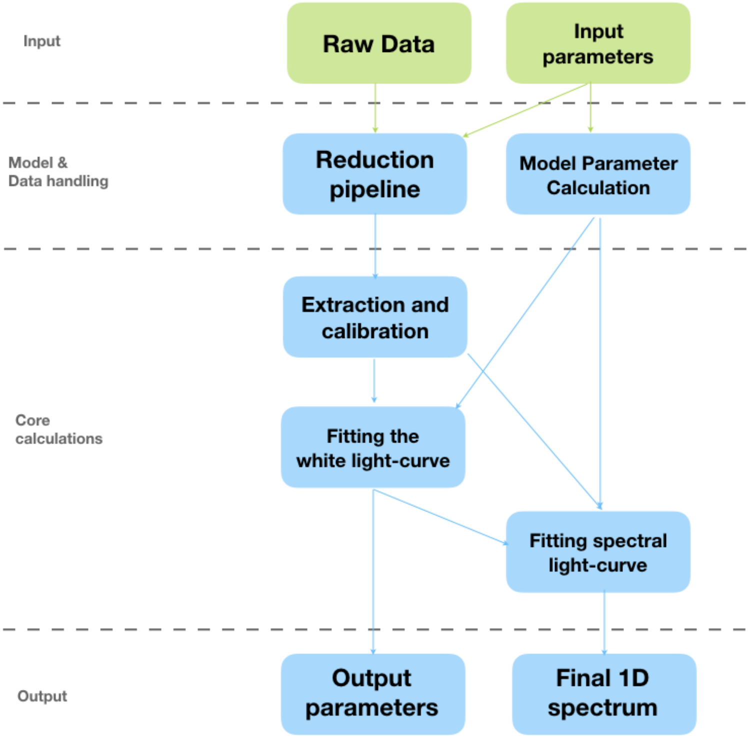

Figure 4.1: Flow chart of the pipeline for analysing HST/WFC3 datasets. The box colours indicate different classes of action: green boxes represent external inputs coming from other models or different sources, e.g. users. The blue boxes contain the modules and routines of the code.

4.2 HST/WFC3 and the spatial scanning mode

The WFC3 was not initially designed and installed to perform spatially scanned observations. This approach introduces distortions on the signal on the frames due to the geometry of the camera itself. In order to obtain the best result from these observations an ad-hoc script that can correct for these distortions is needed. To understand the type of corrections required, it is worth to describe the instrument.

The camera consists of two channels: one ultraviolet/optical channel (UVIS) and the near infrared camera (IR). The last one, which is the one the code has been developed for, has got a HgCdTe detector, that is kept constantly at 145 K, 15 filters for the direct image for calibration reasons and two grisms, the G102 (0.8 – 1.15µm, R=210 at 1.0µm) and the G141 (1.075 – 1.7µm, R=130 at 1.4µm).

The camera has been in use since 2009 and since 2012 the spatial scanning-mode has been available. With respect to the staring-mode the new technique allows for a larger number of photons to be collected in a single exposure without the risk of saturation. As a result, overheads are reduced and the achieved signal-to-noise ratio (S/N) is increased.

As described on *WFC3 Instrument Handbook*[2], the detector is composed by a mosaic of 4 quadrants of 512×512 pixels for a total of 1024×1024 pixels. However, not all the pixels are sensitive to the incoming light, indeed, the outer five rows and columns are called *reference pixels* and they are used to measure the bias drift from the zero read onwards (see Sec. 4.3.1). Indeed, an important characteristic of the detector is that it allows measurements of the collected electrons without resetting the charge in the pixels. These *non destructive reads* (NDRs) permit each WFC3/IR exposure to be the result of many intermediate exposures (the first one is called *zero read*). The final image is the result of the sum of all the intermediate images starting from the zero read onwards. The number of intermediate images (NSAMP) can be chosen based on the exposure time suitable for the particular target observation, NSAMP can vary from a minimum of two to a maximum of 16 NDRs in a single image. This characteristic is useful since it allows to split the entire observation in sub-observations that can be analysed separately and finally summed. This is the strategy used for the first time in the context of this pipeline to analyse the HAT-P-32b dataset (full discussion Sec. 4.6, (Damiano et al., 2017)).

The NDRs images are recorded in the entire detector, by default, however, it is possible to choose a sub-array of the detector to increase the number of exposures within the available memory and reduce in this way the overheads (time lost in non-observation processes). The possible sub-array (SUBTYPE keyword in a WFC3/IR exposure header, see Tab. 4.1) that can be chosen are: FULLIMAGE (full detector array 1024×1024), SQ512SUB (522×522 pixels sub-array), SQ256SUB (266×266 pixels sub-array), SQ128SUB (138×138 pixels sub-array), SQ64SUB (74×74 pixels sub-array). All the sub-array types have the same centre of the full detector

[2] `http://www.stsci.edu/hst/wfc3/documents/handbooks/currentIHB/wfc3_ihb.pdf`

array. In the exoplanet environment, however, the most common are SQ512SUB and SQ256SUB, since the scan length in a frame varies from short (∼30 pixels) to long (∼350 pixels, e.g. 55Cnc e dataset (Tsiaras et al., 2016b)).

Finally, a spatially scanned spectrum can be described as the superposition of many staring-mode spectra, with each one slightly shifted along the vertical axis of the detector (see Fig. 4.2). The most common approach to produce 1D spectra from 2D spatially scanned frames, is to sum along the detector columns (Deming et al., 2013). However, the "building blocks" of a spatially scanned spectrum are neither identical to each other nor parallel to the detector rows, because:

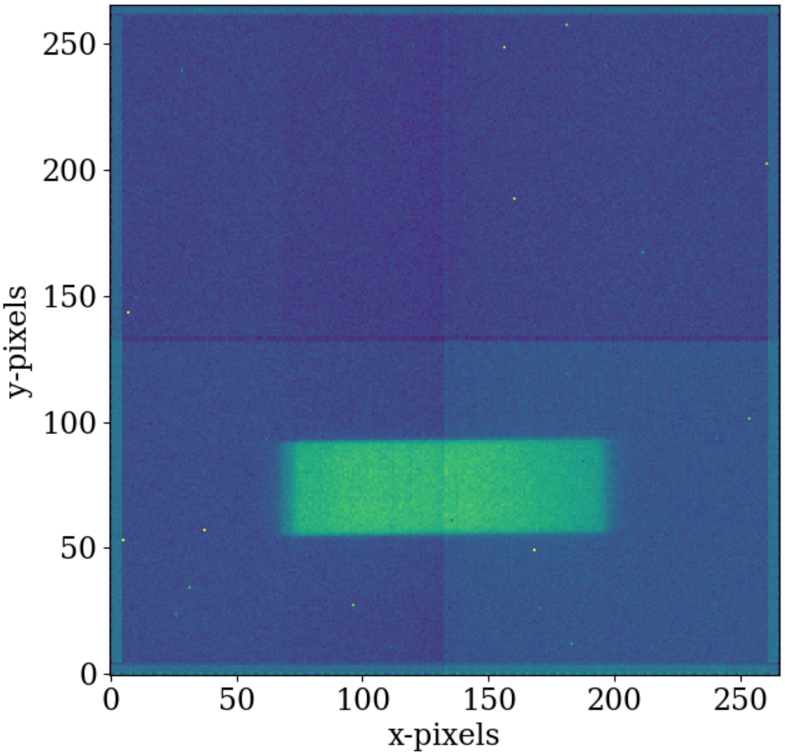

Figure 4.2: Raw image from the HAT-P-32b dataset, the scanning mode technique disperses the signal along the vertical axes. The total scan length is approximately 40 pixels.

- the detector is tilted by 24 degree with respect to its horizontal axis, resulting in significant dispersion variations along the vertical axis of the WFC3/IR detector (from about 4.47–4.78 nm/pix);

- the first-order spectrum of the G141 grism used, is inclined by 0.5 degrees with respect to the WFC3/IR detector rows.

The combined effect is the introduction of distortions in the signal on the detector. Because of these dispersion variations, the wavelength associated to a column is not constant along it, and in particular, the wavelenght associated to the bottom of a specific column in the detector is longer than that associated to the top. To have an idea about the quantity of distortions, let us consider two datasets that have different scan length (total length of the signal along y-axis). In the case of HD 209458b (Tsiaras et al., 2016a) (scan length of 170 pixels), for a column at 1.2 µm, the wavelength difference between the lower and the upper edge of the spatially scanned spectrum is 30 Å, while at 1.6 µm the difference is 70 Å. These values correspond to 0.6 and 1.5 pixels, respectively. As a result, 1D spectra resulting from summing along the columns of the detector vary by up to 1% between an intermediate scan of 60 pixels and the final scan of 170 pixels. For longer scans, such as 55 Cancri e (Tsiaras et al. (2016b), 340 pixels), the effect is stronger and the discrepancy can be more than 2%.

An effort to correct for dispersion variations has been made by Kreidberg et al. (2014) with a row-by-row interpolation, which rearranges the flux in each row to create a uniformly repeated spectrum along the scanning direction. Although this is a possible approach, it may restrict the achievable precision level, because the dispersion direction is inclined by 0.5 degree and, therefore, the "building blocks" of the spatially scanning spectrum are not parallel to the detector rows. In Sec. 4.4 I will explain our approach to this problem and describe the advantages derived from it.

4.3 Data reduction

The standard HST pipeline, *CalWF3*[3], and the spectroscopic package *aXe*[4] can reduce the HST staring-mode spectroscopic images and extract their respective 1D

[3] http://www.stsci.edu/hst/wfc3/pipeline/wfc3_pipeline

[4] http://axe-info.stsci.edu/

spectra. By contrast, scanning-mode spectroscopic images have a much more complicated structure (see Sec. 4.2). Due to this, only an intermediate product of the CalWF3 package (IMA images) is valid when applied to scanning-mode datasets. In addition, the calibration/extraction routines included in the aXe package cannot be applied to spatially scanned spectra. In the literature, analyses of datasets obtained in scanning-mode include custom routines to further reduce the IMA images and extract their calibrated 1D spectra (Deming et al., 2013).

For the reasons mentioned a new reduction pipeline has been developed to treat properly spatially scanned observations (Tsiaras et al., 2016a). The steps that account for the image reduction are:

1. zero Read subtraction;
2. non-linearity correction;
3. dark current subtraction;
4. gain conversion;
5. sky background subtraction;
6. flat field correction;
7. bad pixels and cosmic rays correction;

4.3.1 Zero Read subtraction

The nature of the images that are sequences of non destructive images (NDRs) helps to subtract the signal stored in the detector before the beginning of the observations. This step is necessary since Hubble Space Telescope lacks a shutter. The first read of the sample sequence (referred as *zero read*, ZR) is corrected for the called *super-zero read*, ZR* that contains the bias level of the WFC3/IR detector which is a calibration file (Hilbert (2014), *u1k1727mi_lin.fits*).

All the NDRs samples are then subtracted by ZR and the reference flux level becomes the ZR itself and this is important since it will be used in the non-linearity correction.

Finally, from the ZR the mean of the reference pixels is calculated (NDR$_{ref}$) and it is subtracted to the subsequent NDRs to correct for the bias drift (see Sec. 4.2).

4.3.2 Non-linearity correction

On the IR detector on board HST, electrons are recorded linearly up to a certain point. In particular, the more the recorded flux approaches to the saturation limit, the more electrons are not recorded. Fig. 4.3 shows the described behaviour: here the non linearity effect reaches 5% difference at $\sim 70'000$ e$^-$.

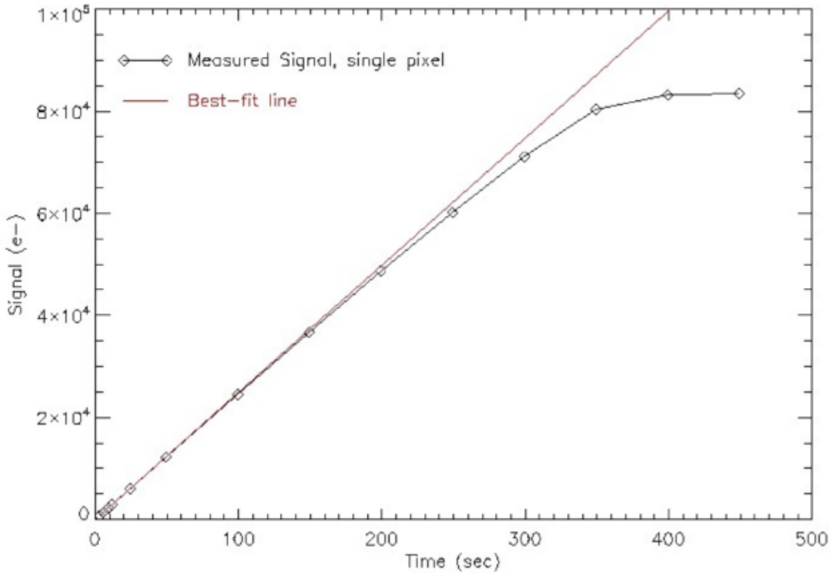

Figure 4.3: WFC3/IR non linearity effect in time. In red the expected electrons recorded and in black the actual value recorded. Figure adopted from: `http://www.stsci.edu/hst/wfc3/pipeline/wfc3_pipeline`.

To correct this effect it is necessary to apply the following correction (Hilbert, 2008)

$$F_c = \left(1 + c_1 + c_2 F + c_3 F^2 + c_4 F^3\right) F \tag{4.1}$$

where F is the non-linear recorded signal, F_c is the linear corrected flux and c_{1-4} are the non-linearity coefficients (Hilbert (2008) *u1k1727mi_lin.fits* calibration file).

Since the flux reference for the NDRs is the ZR, before applying this correction it is necessary to sum the ZR again and in particular:

$$NDRs = F_c\,(NDRs + ZR) - F_c\,(ZR) \tag{4.2}$$

4.3.3 Dark current subtraction

In absence of external illumination, the detector collects electrons. The correction consists on subtracting a master dark that matches the same observational option of the dataset (e.g. NSAMP and integration time). From the provided master darks (Dulude et al., 2014) the suitable one is than picked and it is subsequently subtracted from all the NDRs.

4.3.4 Gain conversion

This step is performed in the same way as the official pipeline *CalWF3*. This correction accounts for the conversion from electrons (e^-) to digital number (DN). The gain values for each of the four quadrants of the detector are included in the calibration file *t2c16200i_ccd.fits* ($\sim$ 2.35e^-/DN). The NDRs are then multiplied by these values.

4.3.5 Sky background

This step is not included in the official pipeline, however, a master sky file calibration is provided (i.e. *WFC3.IR.G141.sky.V1.0.fits*). In Kümmel et al. (2011) it is said that this step is necessary before the wavelength dependent flat-field correction is applied, and also that the master-sky needs to be scaled with the sky ratio of the dataset under investigation. The scaling factor (or sky ratio) is calculated by dividing an area of the detector, that is not affected by the astrophysical signal (sky area), by the master sky frame. The NDRs are therefore subtracted by the scaled master sky.

4.3.6 Flat field correction

The reduction process needs two more steps to be completed (i.e. flat-field correction and spikes correction). However, since the flat-field is wavelength dependent, the calibration step needs to be performed to calculate the wavelength solution. The calibration step is explained in Sec. 4.4.

Assuming that the signal has been calibrated, the wavelength dependent flat-field, F(λ), is described by a 3^{rd} order polynomial (Kuntschner et al. (2011), *WFC3.IR.G141.flat.2.fits* file):

$$F(\lambda) = \sum_{i=0}^{i=3} F_i \left(\frac{\lambda - \lambda_{min}}{\lambda_{max} - \lambda_{min}} \right)^i \tag{4.3}$$

where F_i are the flat-field coefficients, λ is the wavelength, and λ_{max} and λ_{min} are the maximum and minimum wavelength detectable by the grism (G141 in this case).

4.3.7 Bad pixels and cosmic rays correction

The final step in our reduction process is the correction of bad pixels and cosmic rays. Bad pixels have been studied during the calibration cycles and stored in the calibration file *y711520di_bpx.fits* (Hilbert, 2012).

In contrast, cosmic rays are randomly positioned on the detector and have to be identified in each image, independently. To do so, two values are taken into account for each pixel: an x-flag (the median difference from the six horizontally neighbourhood pixels) and y-flag (the median difference from the six vertically neighbourhood pixels). If both x-flag and y-flag are more than 3σ from the respective medians then the point is identified as cosmic ray. In this way the 2D structure of the scanned spectrum is taken into account in both directions. Both bad pixels and cosmic rays are corrected by performing a 2D polynomial interpolation on the image and substitute the flagged pixels with the value of the interpolation.

4.4 Calibration and extraction

Before all the frames are reduced completely, the calibration step and the calculation of the wavelength solution take place. The calibration process is divided into three steps (Tsiaras et al., 2016a):

- checking the position of the signal in each frame and take into account any shifts with respect each other;

- determining the position of the star in the direct image;
- drawing the wavelength-dependent photon trajectory (w.d.p.t.) with the geometry of the signal on the detector as constraint.

4.4.1 Point source position

As reported in Kuntschner et al. (2009), to calibrate properly the signal dispersed by the grism G141 (IR channel), the absolute position (position in the full array detector) of the star ($x^{\star}, y^{\star}$) needs to be determined on the direct image. In the case of spatially scanned spectra, the vertical position ($y^{\star}$) is not constant and so it cannot be determined from the direct image. On the other hand, the horizontal position ($x^{\star}$) is given by the equation:

$$x^{\star} = x_0 + (507 - (L/2)) + \Delta_{off} + \Delta_{ref} \tag{4.4}$$

where x_0 is horizontal position of the star in the direct image calculated via 2D Gaussian fitting, L is the size of the direct image and is related with the SUBTYPE chosen for the observation (see Tab. 4.1), 507 - $L/2$ is the correction needed to transform from the sub-array to the full detector coordinates, because, the sub-array and the full detector have the same centre, the reference pixels are not included in the calculation, so, the centre is 507 instead of 512. Δ_{off} is the difference in centroid offset in the $x - axis$ between the filter used for the direct image and the filter used for calibration (F140W in the case of G141 grism); the values of the centroid offset per filter are reported in Tab. 4.2 (Sabbi et al., 2010). Finally, Δ_{ref} is the difference in the x-position of the chip reference pixel, x_{ref} (the pixel that corresponds to the coordinates of the target), between the WFC3 aperture used for the direct image and the WFC3 aperture used for the dispersed image. This difference originates from the different centring of the detector: the direct image is centred on the star, while during the scanning observations the center is on the first order of the dispersed spectrum. This correction also includes any shifts indicated by the observer through the POSTARG1 keyword in the fits file header (converted to pixels. i.e. POSTARG1/x_{scale}). The POSTARG1 movement is indicated by the observers and it is the shift of the telescope during, or just before or after, the exposure. The values for x_{ref} and x_{scale} for different WFC3 apertures can be found in the *WFC3 aperture file*[5].

SUBTYPE	$L[pix]$	SUBTYPE	$L[pix]$	SUBTYPE	$L[pix]$
FULLIMAG	1014	SQ512SUB	512	SQ128SUB	128
		SQ256SUB	256	SQ64SUB	64

Table 4.1: Relation between the length of the direct image and the SUBTYPE parameter.

filter	$x_{off}[pix]$	filter	$x_{off}[pix]$	filter	$x_{off}[pix]$	filter	$x_{off}[pix]$
F098W	0.150	F127M	0.131	F126N	0.264	F125W	0.046
F140W	0.083	F128N	0.026	F167N	0.196	F110W	-0.037
F153M	0.146	F130N	0.033	F164N	0.169	F105W	0.015
F139M	0.110	F132N	0.039	F160W	0.136		

Table 4.2: Horizontal offset for the different WFC3/IR filters.

4.4.2 Position shift

Since HST has not been designed to perform spatially scanned observation, after every scanning it fails to reach the exact starting position. This results in shifts across both horizontal and vertical position from a frame to another. This would not be a problem if every frame had its own direct image paired, because in that case it would be possible to calibrate images singularly. But since the direct image is referred only to the first of the scanned sequence images, every subsequent frame needs to be related to the first one. If not corrected, this could introduce systematics that result in variations of up to 250 ppm in the final spectrum.

To calculate the horizontal shifts, we compare the structure of the first spatially scanned spectrum with all subsequent spectra, using the normalised sum along their

[5] `http://www.stsci.edu/hst/observatory/apertures/wfc3.html`

columns (Fig. 4.4, left panels), similar to Kreidberg et al. (2014). The sum calculated from the first spatially scanned spectrum is interpolated and fitted to the sum from each subsequent spatially scanned spectrum. The results of the fittings are the horizontal shift Δx_i and the normalisation factor which account for flux variation. Note that this step is performed before the wavelength dependent flat-field correction. The sum above is, corrected for the non-wavelength dependent flat-field to avoid the introduction of undesired bias.

Shifts of the vertical position (Δy_i) are calculated from the first non-destructive read of each exposure. We apply the same method as for the horizontal shifts described above, with the difference that the sum is calculated along the rows instead of the columns (Fig. 4.4, right panels). The horizontal position of the star and the vertical position at the beginning of the scan for each frame are then:

$$x_i^\star = x_1^\star + \Delta x_i \tag{4.5}$$

$$y_{si}^\star = y_{s1}^\star + \Delta y_i \tag{4.6}$$

where i varies from 1 to N which is the total number of frames in the visit.

Finally, we calculate the scan length (l_i) by fitting an extended Gaussian function on the sum along the columns of the last non-destructive read of each image.

4.4.3 Wavelength-dependent photon trajectories (w.d.p.t.)

The 24 degree tilt of the WFC3/IR with respect to the horizontal axis and the 0.5 degree with respect to the detector rows make the dispersed spectrum trace not only non-parallel to the detector rows but also to be field-dependent (where on the detector is the trace). In the *aXe User Manual version 2.3* (calibration routine for staring mode observations) is reported that once the detector is illuminated by a source dispersed by the G141 grism, the trace of the dispersed spectrum is described on the detector by:

$$y - y^\star = a_t(x - x^\star) + b_t \tag{4.7}$$

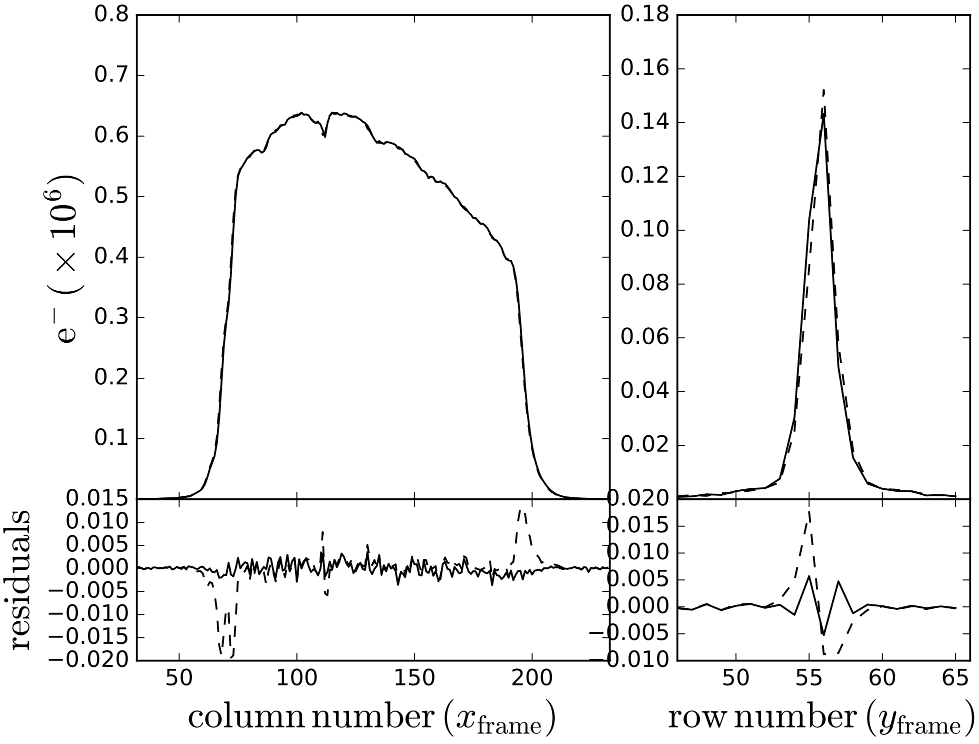

Figure 4.4: Left-top: sum along the columns of the first (continuous) and the last (dashed) spatially scanned spectra of the HAT-P-32b dataset. Left-bottom: difference between the two profiles before and after shifting, indicated by dashed and continuous lines, respectively. Right: same plots for the sum along the rows of the first non-destructive read.

where:

$$\begin{aligned}
\underbrace{a_t}_{\text{or DYDX\_A \_1}} &= \begin{cases} a_{t0} + a_{t1}x^\star + a_{t2}y^\star + \\ a_{t3}x^{\star 2} + a_{t4}x^\star y^\star + a_{t5}y^{\star 2} \end{cases} \\
\underbrace{b_t}_{\text{or DYDX\_A \_0}} &= b_{t0} + b_{t1}x^\star + b_{t2}y^\star
\end{aligned} \tag{4.8}$$

Moreover, the associated wavelength of a point on the dispersed trace is a linear function of the distance, d, between the point and the star along the trace:

$$\lambda = a_W d + b_W \tag{4.9}$$

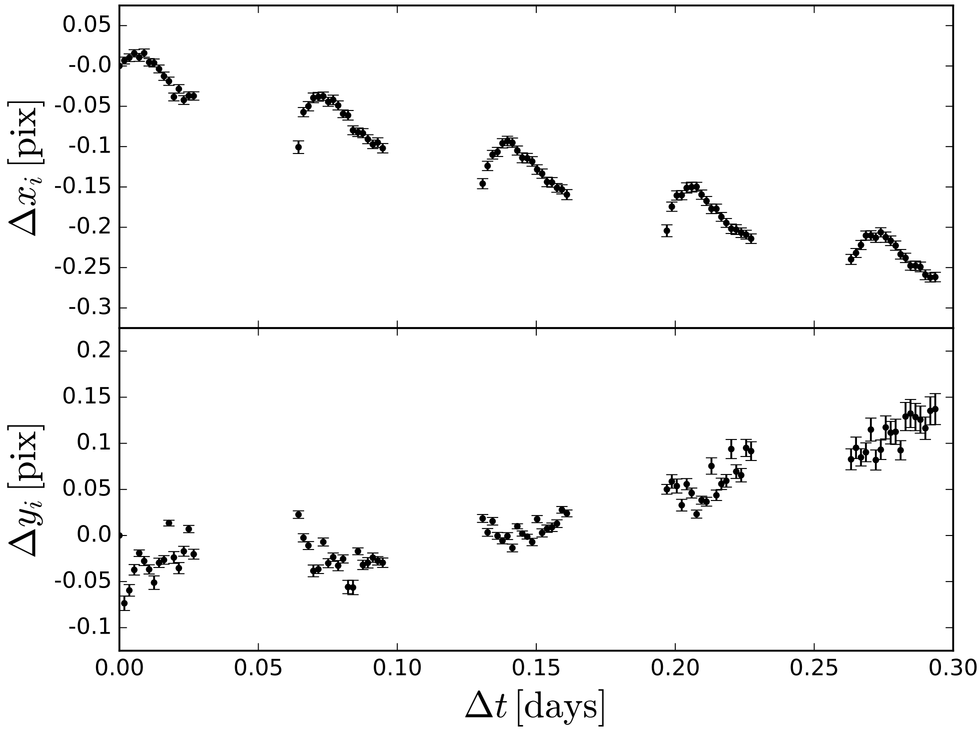

Figure 4.5: Horizontal (top) and vertical (bottom) shift relative to the first one for each image of the HAT-P-32b dataset.

where:

$$\begin{aligned} \underbrace{a_W}_{\text{or DLDP\_A \_1}} &= \begin{cases} a_{W0} + a_{W1}x^\star + a_{W2}y^\star + \\ a_{W3}x^{\star 2} + a_{W4}x^\star y^\star + a_{W5}y^{\star 2} \end{cases} \\ \underbrace{b_W}_{\text{or DLDP\_A \_0}} &= b_{W0} + b_{W1}x^\star + b_{W2}y^\star \end{aligned} \tag{4.10}$$

the coefficients a_{t0-5}, b_{t0-2}, a_{W0-5}, b_{W0-2} are contained in the configuration file *WFC3.IR.G141.V2.5.conf* (Kuntschner et al. (2009), Tab. 4.3), and $x^\star$ and $y^\star$ are the coordinates of the star in the direct image (see Sec. 4.4.1).

To help visualising the geometry that has just been discussed, Fig. 4.6 shows the relative position of the star on the direct image and the dispersed spectral trace in the frames of the visits. After the position of the star on the full detector has been determined, $P^\star(x^\star, y^\star)$, the photon of a particular wavelength, λ, is dispersed on the position $P_\lambda(x_\lambda, y_\lambda)$. Let us draw the line perpendicular to the spectral trace passing

	$n=0$	$n=1[x^\star]$	$n=2[y^\star]$	$n=3[x^{\star 2}]$	$n=4[x^\star y^\star]$	$n=5[y^{\star 2}]$
a_{tn}	1.04275E-02	-7.96978E-06	-2.49607E-06	1.45963E-09	1.39757E-08	4.8494E-10
b_{tn}	1.96882E+00	9.09159E-05	-1.93260E-03			
a_{Wn}	4.51423E+01	3.17239E-04	2.17055E-03	-7.42504E-07	3.48639E-07	3.09213E-07
b_{Wn}	8.95431E+03	9.35925E-02	0.0			

Table 4.3: Calibration coefficients for the G141 grism.

for the position of the star. The crossing point is identified as $P_1(x_1, y_1)$.

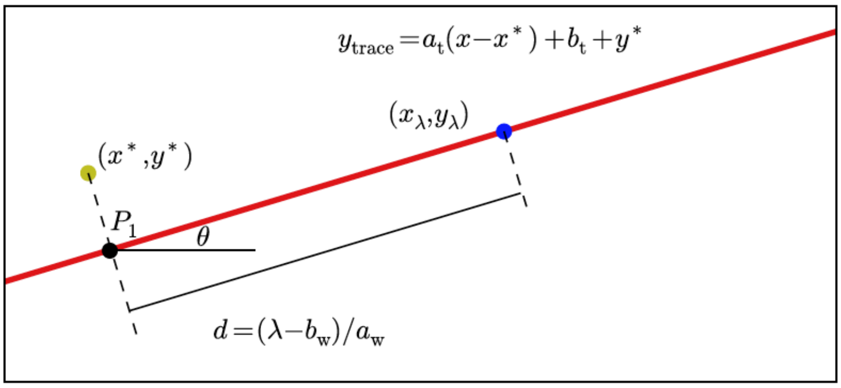
$y_{\rm trace} = a_{\rm t}(x - x^*) + b_{\rm t} + y^*$
(x_λ, y_λ)
(x^*, y^*)
P_1
θ
$d = (\lambda - b_{\rm w})/a_{\rm w}$

Figure 4.6: Relative positions of the spectral trace (red line), of the star $(x^\star, y^\star)$ in the direct image, and a random point on the trace (x_λ, y_λ). θ indicates the tilt of 0.5 degree of the grism respect to the detector' rows. Image adopted from: Tsiaras et al. (2016a).

The coordinates of the point P_1 can be calculated using Eq. 4.7 (which becomes Eq. 4.11) and the property of orthogonality between the spectrum trace and the projection of the star position on the trace.

$$y_1 = a_t(x_1 - x^\star) + b_t + y^\star. \tag{4.11}$$

Since the vector $\mathbf{P}^\star\mathbf{P}_1$ =$(x_1 - x^\star, y_1 - y^\star)$ is orthogonal to $\mathbf{V}$=$(1, a_t)$ (vector parallel to the trace) we have that their scalar product is zero by definition:

$$\mathbf{P}^\star\mathbf{P}_1 \cdot \mathbf{V} := 0 = x_1 - x^\star + a_t(y_1 - y^\star). \tag{4.12}$$

Substituting Eq. 4.11 into Eq. 4.12 it is possible to calculate the x-coordinate of P_1:

$$x_1 = x^{\star} - \frac{a_t b_t}{1+a_t^2} \tag{4.13}$$

To generalise this expressions to any points on the trace we just need to introduce the inclination of the trace with respect to the detector rows. This is indeed indicated as θ in Fig. 4.6; it is defined as $\theta = \tan^{-1}(a_t)$ and from the geometry of the triangle indicated by the trace, the rows and the columns of the detector, we can calculate:

$$\cos(\theta) = \frac{x_\lambda - x_1}{d} \tag{4.14}$$

where d has been calculated from Eq. 4.9. By replacing x_1 from Eq. 4.13 and re-arranging the last equation for x_λ we have:

$$x_\lambda = x^{\star} - \frac{a_t b_t}{1+a_t^2} + \frac{\lambda - b_W}{a_W} \cos\left(\tan^1(a_t)\right) \tag{4.15}$$

and from Eq. 4.7, y_λ:

$$y_\lambda = a_t(x_\lambda - x^{\star}) + b_t + y^{\star}. \tag{4.16}$$

Using these last two equations we are able to determine the wavelength solution for every point on a single trace. However, we are interested in calculating the wavelength solution of the entire scanning spectra, for this reason the following recipe is proposed (Tsiaras et al., 2016a):

1. we assume that $x_\star$ is constant on a single frame, but it is different from frame to frame ($x_i^\star$ from Eq. 4.5);

2. let $y^\star$ vary from the start of the scan, $y_{si}^\star$ (Eq. 4.6) to the end, $y_{si}^\star + l_i$, where l_i is the scan length on the i^{th} frame;

3. let λ vary on the spectral wavelength range covered by the grism ($1.0 - 1.7\mu$m);

4. we use $y^{\star}$ and λ from the last two steps and use Eq. 4.15 and Eq. 4.16 to calculate all the relative x_{λ} and y_{λ} in order to create a large grid ($y^{\star}$, λ, x_{λ}, y_{λ});

5. we fit on the grid points the function of the w.d.p.t.:

$$y = \left(\frac{c_1}{c_2+\lambda}+c_3\right)+\left(\frac{s_1}{s_2+\lambda}+s_3\right)x \tag{4.17}$$

For a particular wavelength, the trajectory is a straight line across the detector, this indicates where the photons of a particular wavelength intersects the detector during the scan. Fig. 4.7 shows the calculated wavelength solution (w.d.p.t) for a single frame of the HAT-P-32b dataset (Damiano et al., 2017). It is possible to appreciate how the photon trajectories follow accurately the shape of the signal on the detector.

Finally, since the wavelength solution has been calculated, it is possible to complete the reduction process by performing the wavelength dependent flat-field and the bad pixels and cosmic rays correction (Sec. 4.3.6 and 4.3.7).

4.4.4 Extraction

After the reduction and calibration process, the signal is extracted and summed, resulting on a single value that corresponds to the flux value. With this process for every frame, it is possible to obtain the light-curve of the star during the observation. From WFC3/IR images two different kind of light-curves are extracted: white and spectral ones. We will refer to the first one when talking about the entire signal extracted from frames and we will refer to spectral one when talking about the signal extracted in a particular wavelength interval. Choosing the right wavelength bins is an important step since this will define the resolution of the final spectrum but also may introduce/avoid scatter on the 1D spectrum. For instance as reported in Sec. 4.6 (Damiano et al., 2017), I decided to chose the wavelength bins with the constraint that every bin contains the same amount of signal (flux) and iterate the total number of wavelength bins on a typical value (between 20 and 25) to find the optimal set.

In the literature the extraction process is commonly performed by summing over the column, here, the UCL pipeline extraction follows the structure of the

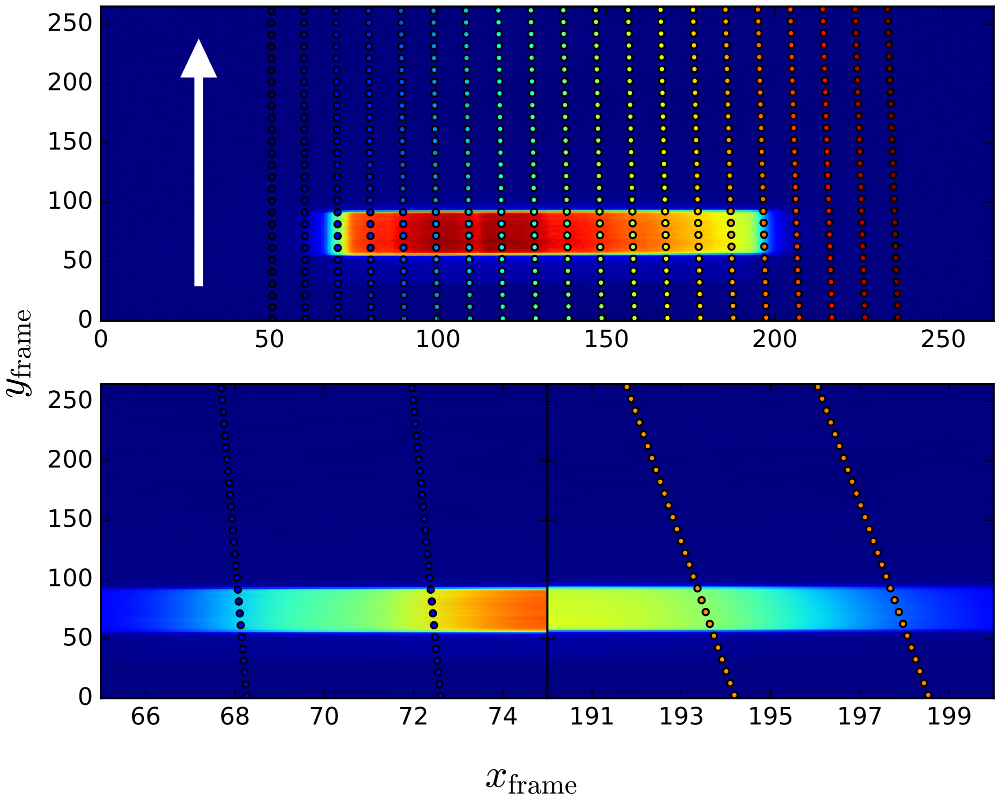

Figure 4.7: Calculated wavelength solution of one frame of the HAT-P-32b dataset. **Top panel**: Trajectories of the dispersed photons of different wavelengths (coloured points) while the star moves along the scan direction (white arrow). **Bottom panel**: Left and right edges of the spectrum where it is possible to appreciate how the grid follows the signal on the detector.

signal (wavelength/scan coordinates) rather than the detector coordinates (column/row).

The signal is extracted from apertures of quadrangular shape. In the case of spectral light-curves these apertures are calculated for each wavelength bin (λ_1 - λ_2) per frame. The left and right edges of each quadrilateral are given by the w.d.p.t. (Eq. 4.17) and the upper and lower edges are given by the spectrum trace (Eq. 4.7). A little margin ($\sim$ 15 pixels) is given on top and bottom sides to include the tails of the signal. However, this extraction method brings some issues. Since we use quadrangles that are not parallel with the column/row coordinates, fractional pixels need to be taken into account and this can introduce scattering on the final 1D spectrum. To avoid these issues one can fit a second-order 2D polynomial function on the perimeter pixels and their next surroundings, such that the integral of this function is equal to the flux level of the pixels under the curve, and then take a fraction of this flux.

Finally, the signal can be extracted all at once by using a large extraction box on the last NDR (e.g. Tsiaras et al. (2016a,b)) or split the signal across the scan direction using the differential NDRs $(NDR(t) - NDR(t-1))$ referring them as 'stripes' (Fig. 4.11) (Damiano et al., 2017). I introduced, to the pipeline, this last method to analyse the HAT-P-32b dataset since it is a binary system and the two stellar signals are blended on the detector. Moreover, we also found that the strategy of using multiple aperture extraction (i.e. stripes) reduced the scatter in the final light-curves when we analysed other HST datasets (Tsiaras et al., 2018).

4.5 Obtaining the 1D spectrum

As described in the previous section the output consists of the white light-curve and the spectral ones. The raw white light-curve of the HAT-P-32b dataset is shown in Fig. 4.8. The light-curves obtained with HST are not continuous; between the signal points there are gaps because of HST orbits. A pair of single signal set and single gap coincides with one complete orbit of HST around the Earth. Moreover, it is also possible to observe systematics at the beginning of every observational orbit. It is, indeed, known from previous studies that using the WFC3 camera both in staring-mode (Berta et al., 2012; Swain et al., 2013; Wilkins et al., 2014) and in scanning-mode (Deming et al., 2013; Kreidberg et al., 2014; Knutson et al., 2014a,b) introduces two time-dependent systematics to the light-curves: one long-term throughout the visit, with an approximately linear behaviour, and one short-term throughout each HST orbit, with an approximately exponential behaviour. These systematics are commonly referred as the "ramps". Moreover, these systematics are also correlated with the brightness of the observed target. The brighter the target the stronger is the ramp (Deming et al., 2013; Tsiaras et al., 2016b,a; Damiano et al., 2017; Tsiaras et al., 2018).

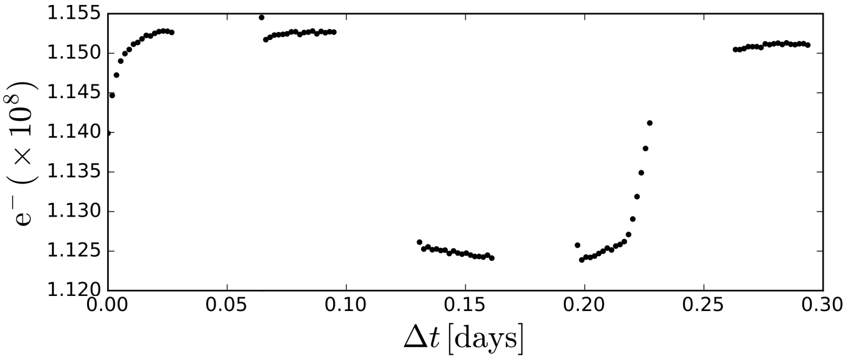

Figure 4.8: Extracted raw white light-curve of the dataset HAT-P-32b.

4.5.1 Fitting the white light-curve

At first order the model that can fit the white light-curve is a combination of a transit model, $F(t)$, a function of the systematics, $R(t)$, and a normalisation factor, n_W:

$$M(t) = n_W R(t) F(t) \tag{4.18}$$

The transit model describes the ratio between the stellar flux during the out-of-transit and during the in-transit at every time-step of a given time series based on a set of input parameters. The input parameters are used to calculate the star-planet projected separation and the relative flux occulted by the planet. The transit model calculation is implemented in the Python package PyLightcurve[6] (Tsiaras et al. (2016a) and Dr. Angelos Tsiaras, Ph.D. thesis).

In synthesis, in the model, the star-planet projected distance is calculated for every time-step of the time-series. This parameter is useful to calculate the flux blocked by the planet in agreement with the limb darkening law used. The flux of the stellar disc relatively to the centre of the disc – i.e. limb darkening law – in our pipeline is modelled using a 4-coefficients law (Claret, 2000):

$$I(a_n, r) = 1 - \sum_{n=1}^{n=4} a_n \left(1 - \left(1 - r^2\right)^{n/4}\right). \tag{4.19}$$

The a_n coefficient are calculated by fitting the ATLAS model (Kurucz, 1970; Howarth, 2011; Espinoza & Jordán, 2015) created from the input parameters (Tab.

[6] https://github.com/ucl-exoplanets/pylightcurve

5.1) and the sensitivity curve of the grism used (WFC3/IR-G141 in this case) in the same wavelength limits of the extracted light-curve.

The calculated transit model, F($T_0, P, i, a/R_\star, R_p/R_\star, e, \omega, t$) is then multiplied by an instrumental systematics function, $R(t)$, similarly to Kreidberg et al. (2014); Stevenson et al. (2014); Kreidberg et al. (2015):

$$R(t) = (1 - r_a(t - T_0))\left(1 - r_{b1}e^{-r_{b2}(t-t_0)}\right) \tag{4.20}$$

where T_0 is the mid-transit time, t_0 is the time when each orbit starts, t is the time sequence, r_a is the slope of the linear long term ramp and (r_b1, r_b2) are the coefficients of the exponential short-term ramp.

On the real case, for the analysis of the light-curves, we excluded the first orbit since it has different short and long term ramps. Removing the first orbit is a common procedure (e.g. Deming et al. (2013); Huitson et al. (2013); Haynes et al. (2015); Tsiaras et al. (2016a); Damiano et al. (2017)).

Fig. 4.9 (mid panel) shows the white light-curve after systematics have been corrected. The fitting residuals (Fig. 4.9 bottom panel) are not Gaussian distributed especially in the egress. This can be due to non-optimal parameters (e.g. i and $a/R_\star$) that have been fixed due to lack of ingress points, or remaining systematics not included in the $R(t)$ function.

4.5.2 Fitting the spectral light-curves

To extract the planetary spectrum from the spectral light-curves, we follow the approach described by Kreidberg et al. (2014). According to this method every spectral light-curve is divided by the white light-curve and then it is fitted with a model that takes into account a wavelength-time dependent linear slope, a normalisation factor, the best transit model that fit the white light-curve and a wavelength-time dependent transit model $F(\lambda, t)$:

$$M(\lambda, t) = n_\lambda \frac{1 + r_{\lambda a1}(t - T_0)}{F_W(t)} F(\lambda, t) \tag{4.21}$$

where n_λ is the normalisation factor, $r_{\lambda a1}$ is the coefficient for the wavelength-time linear slope, $F_W(t)$ is the best model that fit the white light-curve, t is the

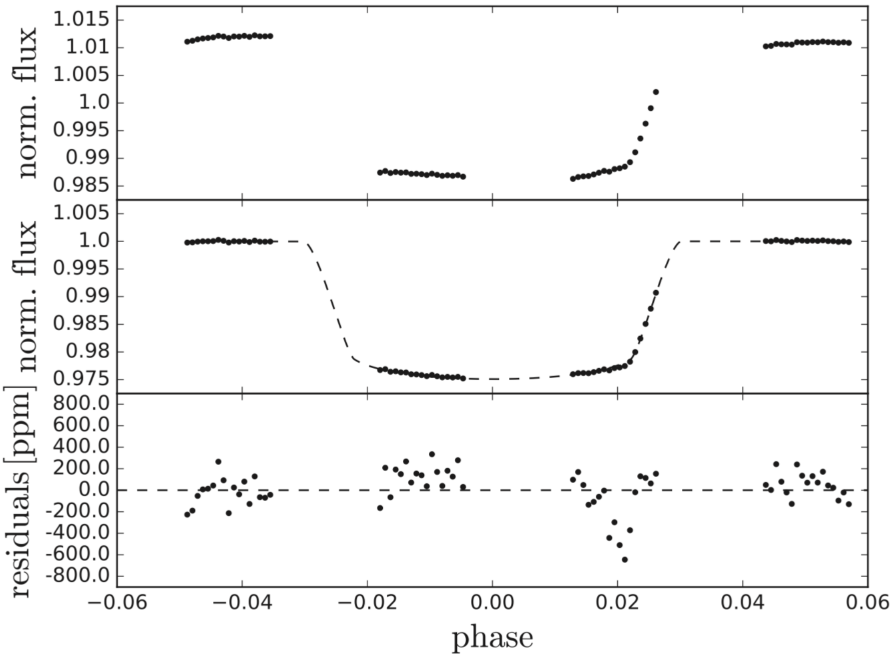

Figure 4.9: **Top panel**: normalised white light-curve of HAT-P-32b before fitting for transit and systematics model. **Middle panel**: white light-curve divided by the best-fit model of the systematics. **Bottom panel**: fitting residuals, it can be seen that the model fails to fit the egress. The possible reasons for this behaviour are either non-optimal orbital parameters, limb-darkening coefficients or remaining systematics.

observational time series and $F(\lambda,t)$ is a wavelength-time dependent model specific for each spectral bin. In all $F(\lambda,t)$ the only free parameters is the ratio between the planetary and the star radius, since all the other parameters are fixed to those calculated by the white light-curve fitting.

The advantage of using this method is that the residuals of the spectral fittings do not show the same behaviour as the white ones since all the possible systematics are corrected by dividing by the white light-curve.

From the fitting process described above, the ratio $R_p/R_\star$ is obtained for each spectral bin, and the final 1D spectrum can be determined (e.g. Fig. 4.14).

4.6 HAT-P-32b dataset

In this section I report the analysis of the near-infrared transit spectrum of the hot Jupiter HAT-P-32 b ($T_{eq} = 1786$K, Hartman et al. (2011), see Tab. 4.4) obtained with the WFC3 camera on board the HST (Damiano et al., 2017). HAT-P-32b is one of the most inflated exoplanets discovered, being less massive than Jupiter ($M_p = 0.86\ M_{Jup}$) but having almost twice its radius ($R_p = 1.789\ R_{Jup}$). The atmosphere of HAT-P-32b has been observed with ground-based instruments in the optical wavelengths, revealing a featureless transmission spectrum (Gibson et al., 2013; Zhao et al., 2014; Mallonn & Strassmeier, 2016; Nortmann et al., 2016). In addition, Zhao et al. (2014) suggested the presence of a thermal inversion in the thermal profile of the atmosphere of HAT-P-32b to interpret eclipse observations suggesting the presence of exotic molecules that can cause the inversion.

Stellar parameters	
T_{eff} (K)	6207 ± 88
$M_\star$ ($M_\odot$)	1.160 ± 0.041
$R_\star$ ($R_\odot$)	1.219 ± 0.016
$log(g_\star)$ (csg)	4.33 ± 0.01
Fe/H (dex)	-0.04 ± 0.08
Planetary parameters	
T_{eq} (K)	1786 ± 26
a (AU)	0.0343 ± 0.0004
R_p (R_{Jup})	1.789 ± 0.025
M_p (M_{Jup})	0.860 ± 0.164
P (days)	2.150008 ± 0.000001
T_0 (BJD)	$2454420.44637 \pm 0.00009$
i (deg)	88.9 ± 0.4

Table 4.4: HAT-P-32 system information (Hartman et al., 2011).

The spatially scanned spectroscopic images of HAT-P-32b were obtained with the G141 grism, they are available on the MAST archive[7]: ID program: 14260 and PI: Deming Drake. The dataset contains five consecutive HST orbits, each exposure is the result of 14 NDRs, with a size of 256×256 pixels with a total exposure time of 88.435623s and the total scan length is approximately 40 pixels per frame. During the light-curve analysis, the first of the five orbits was discarded (see Sec. 4.5.1). Of the remaining four HST orbits, the first and the fourth provide the out-of-transit baseline, while the second and the third capture the transit. Moreover, the first points of the first orbit and the third one have been recognised as outliers and discarded from the analysis. The dataset contains, for calibration purposes, a non-dispersed (direct) image of the target, obtained using the F139N filter. Before extracting the light-curves (white and spectral), all frames were reduced using the routines described in the sections above.

HAT-P-32A has an M1.5 stellar companion, HAT-P-32B ($T = 3565 \pm 82$K, Zhao et al. (2014)). The dispersed signals from HAT-P-32A and HAT-P-32B are blended when using the scanning-mode (Fig. 4.10). However, these two stars are separated enough ($2.''923 \pm 0.''004$, Zhao et al. (2014)) to avoid blending when the multiple aperture extractions (i.e. stripes, see Sec. 4.4.4) are considered (Fig. 4.11). For each stripe, we determined the photometric aperture, taking into account the wavelength-dependent photon trajectories (see Sec. 4.4.3 and Tsiaras et al. (2016a)) and obtained a set of 12 white light-curves. The same criterion was used to extract the spectral light-curves, obtaining a set of 12 time-series for each one of the 20 spectral bins. The wavelength range of each bin was chosen in order to have the same flux level across all bins.

The light-curves analysis was performed following the recipe presented in Sec. 4.5.1 and Sec. 4.5.2. However, I decided to proceed using two different paths:

- (stacked) all the extracted 12 white light-curves are summed together to obtain a unique reference light-curve, and the same is performed for every spectral bin;

- (weighted) fitting each white and spectral light-curve alone, then taking the

[7] `https://archive.stsci.edu/`

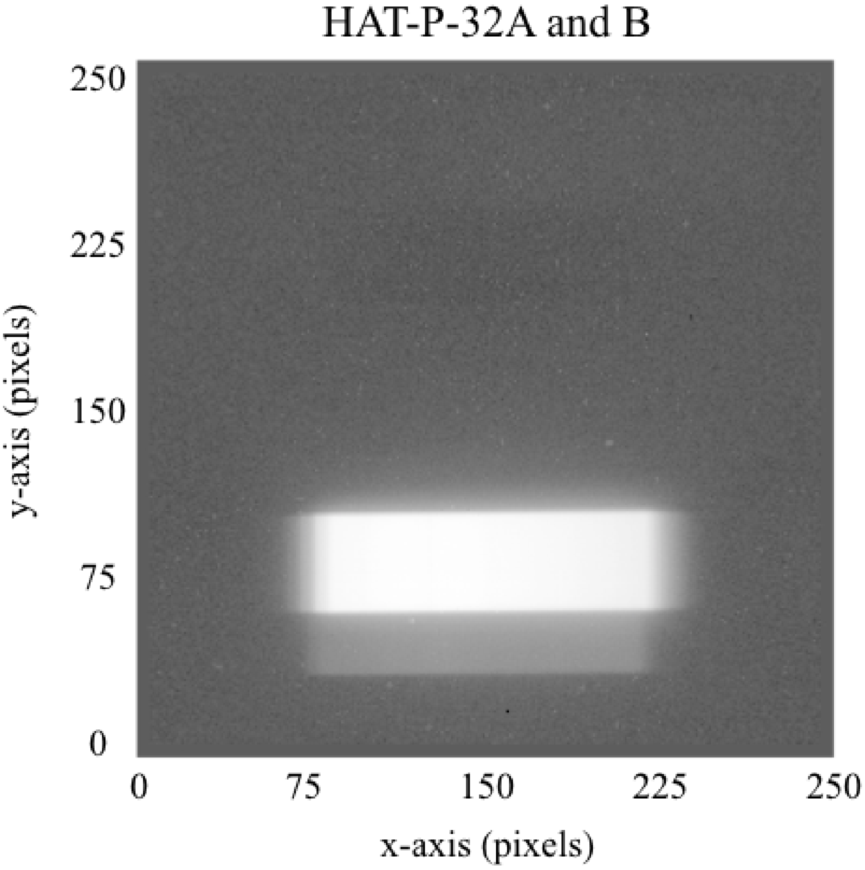

Figure 4.10: Single reduced frame of the HAT-P-32b dataset before the extraction. It is possible to see how the signal relative to each of the two stars is blended with each other.

weighted mean for each spectral bin.

The stacked method led to a similar analysis reported in previous sections, having a unique white light-curve to fit and the 20 spectral light-curves. On the other hand, using the weighted method we needed to perform 12 white light-curve fit (one for each stripe) and 12×20 spectral light-curves fit (one for each stripe for each bin). Note that using the weighted method we worked with more noisier signals because it was split.

Following the stacked method, we obtained the white light-curve shown in Fig. 4.9 (top panel). We followed the process described in Sec. 4.5.1 to find the best model that fits the raw white light-curve and correct for the systematics. In this case the ‘ramps’ are not strong because the star is relatively faint ($K_{mag} = 9.99$). The result of the fitting is shown in Fig. 4.9 (mid and bottom panel) and in Tab. 4.5.

Limb-darkening Coefficients (1.125 - 1.650 µm)	
a_1	0.603336
a_2	-0.223032
a_3	0.281379
a_4	-0.13988
Fitting Results	
T_0 (HJD)	$2457408.95783 \pm 0.00004$
$R_p/R_\star$	0.1521 ± 0.0003

Table 4.5: White light-curve fitting results.

Using the weighted method instead, 12 white light-curves were obtained and analysed singularly. These light-curves are the result of the extraction of lower signal and, therefore, their fit is noisier than the previous case. In Tab. 4.6 the result of the fit performed in every stripe is reported. In the bottom section of the table we note that the results of the two methods agree with each other. However, the stacked method resulted one order of magnitude better than the weighted one for the estimation of the mid transit point.

Every spectral light-curve of every stripe has been fitted and 12 spectra of the planet have been generated (Fig. 4.12 and 4.13). The final 1D spectrum obtained following the weighted method is calculated by taking the weighted mean spectrum of the 12 spectra (Fig. 4.14). Both methods give the same 1D spectrum, with the exception of a few bins where the differences are within 0.3σ. The result is shown in Fig. 4.14: from this we can not choose one method over the other, however, from the white light-curve fittings, we choose the stacked method for a better constraint on the mid-transit point.

Stripes fitting result		
# of stripe	T_0 (HJD)	$R_p/R_\star$
1	2457408.9573971 ± 0.00013061	0.15290952 ± 0.00032594
2	2457408.95755273 ± 0.0001529	0.15156066 ± 0.00036521
3	2457408.957628 ± 0.00022603	0.15214494 ± 0.00053603
4	2457408.95778216 ± 0.00018649	0.15227708 ± 0.0004362
5	2457408.95777772 ± 0.00022194	0.15232754 ± 0.00051591
6	2457408.95752518 ± 0.00017632	0.15254122 ± 0.00042601
7	2457408.95787749 ± 0.00021696	0.15196633 ± 0.00048721
8	2457408.95802174 ± 0.00015749	0.15210755 ± 0.00034092
9	2457408.95794628 ± 0.0002358	0.15168565 ± 0.00053574
10	2457408.9582092 ± 0.00015189	0.15276672 ± 0.00035866
11	2457408.95807645 ± 0.00019223	0.15239069 ± 0.00042786
12	2457408.95864508 ± 0.00018721	0.1525417 ± 0.00041915
Fitting Results		
Stacked method	2457408.95783 ± 0.00004	0.1521 ±0.0003
Weighted method	2457408.9578 ± 0.0004	0.1523 ± 0.0004

Table 4.6: White light-curve fitting results for every stripe. Comparison between staked and weighted results.

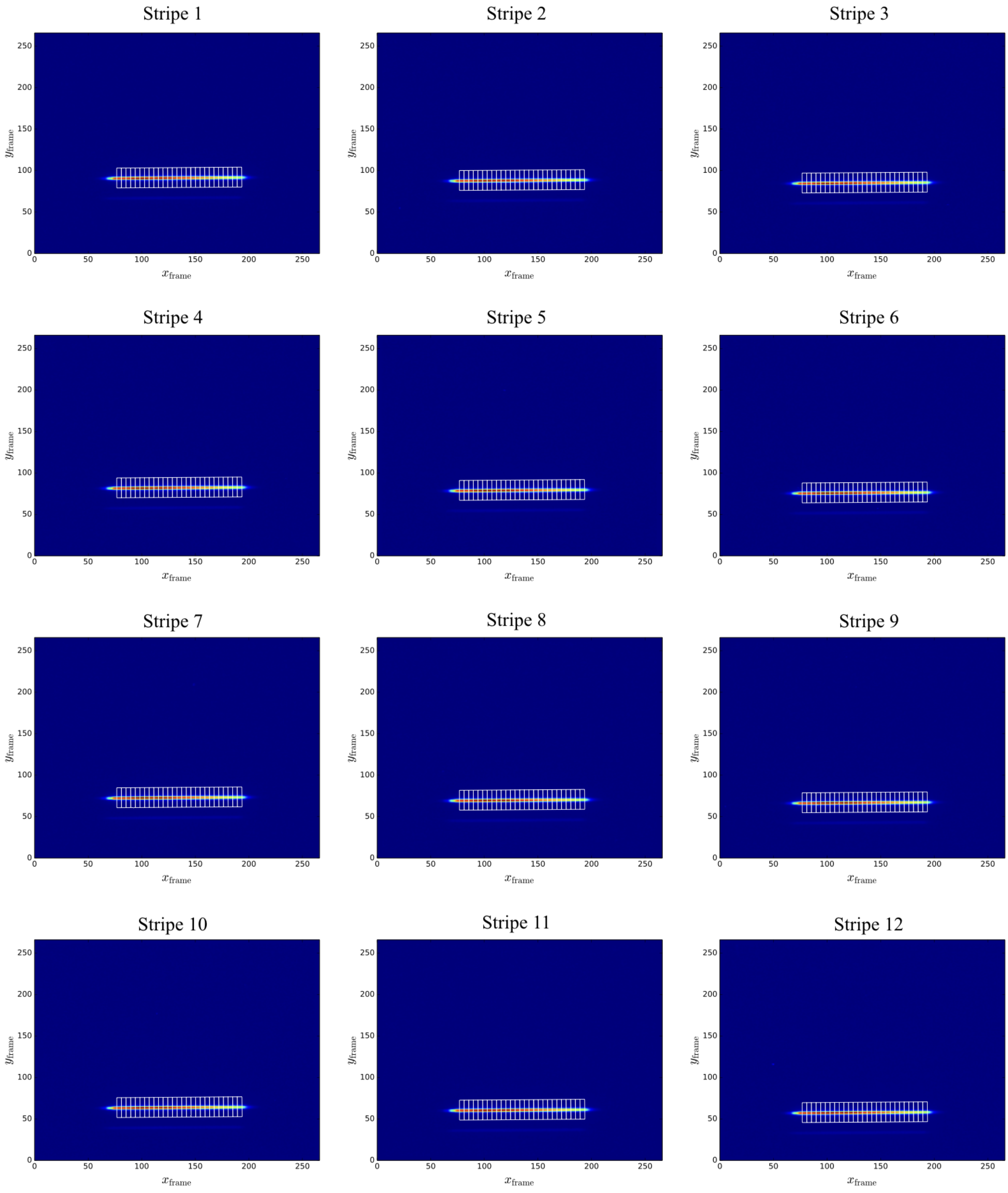

Figure 4.11: Multiple extractions of the HAT-P-32b dataset. The 12 stripes correspond to the NDRs recorded by Hubble Space Telescope. In each of the stripes the two signals relative to the two stars are divided, so that possible contamination can be avoided.

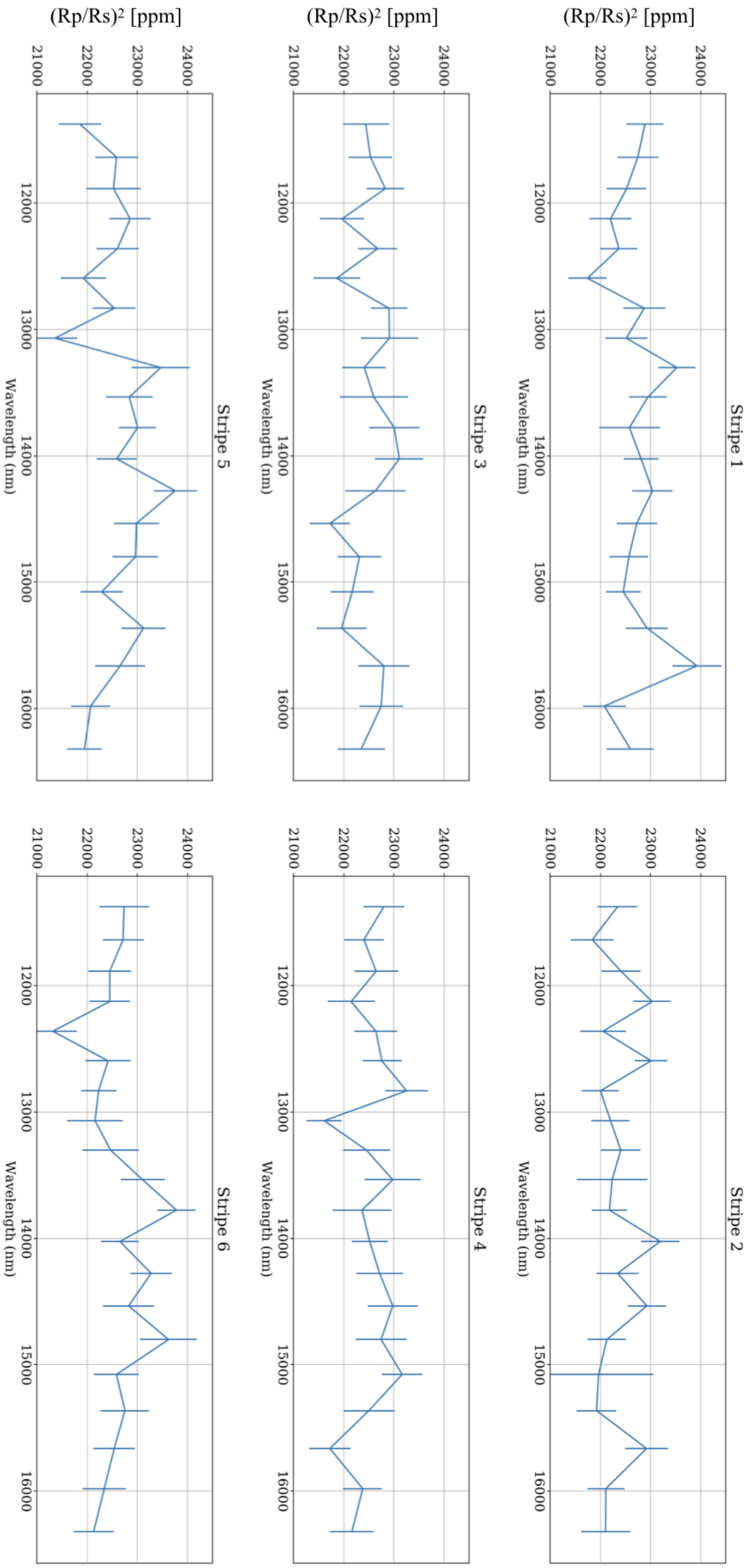

Figure 4.12: 1D spectra of the first 6 stripes of the HAT-P-32b dataset.

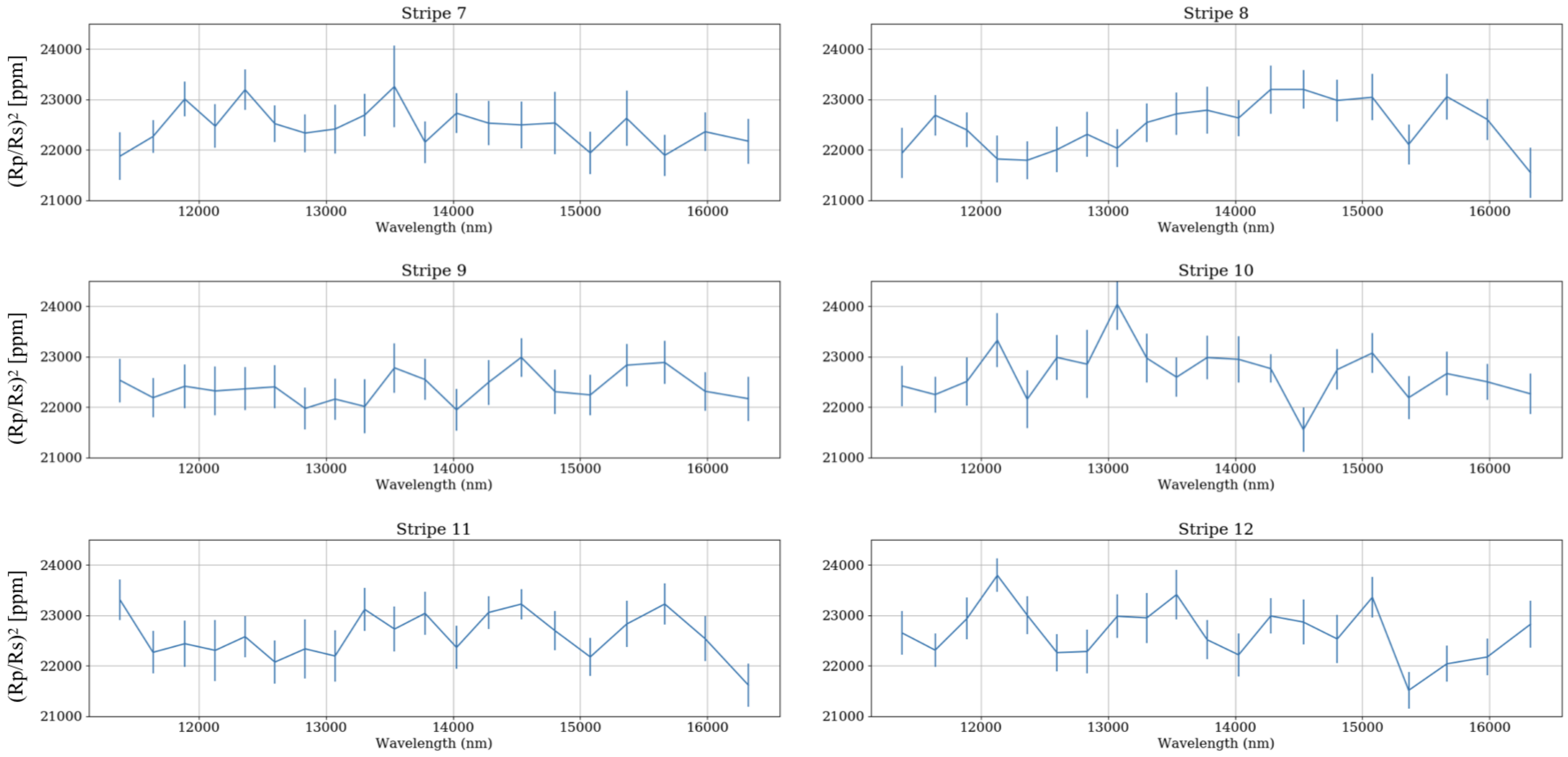

Figure 4.13: 1D spectra of the last 6 stripes of the HAT-P-32b dataset.

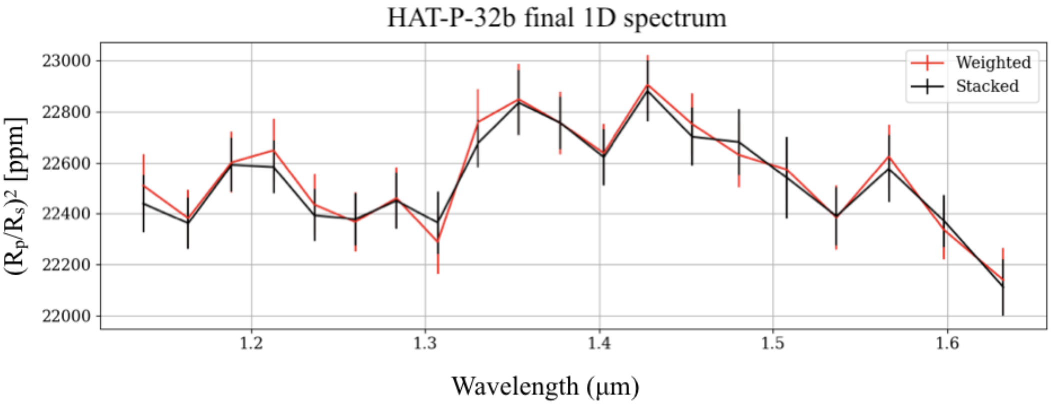

Figure 4.14: Final 1D spectra resulted from calculating one spectral light-curve per bin (stacked) and averaging the spectral $(R_p/R_\star)_\lambda$ for each bin for each stripe (weighted). The two spectra are equivalent within 0.3σ.

4.6.1 Using ICA to de-trend the light-curves

As described in Sec. 4.4.4 I introduced the multi aperture extraction method and therefore we had the possibility to map the signal contained in a frame into a time-series. In particular, we obtained 12 temporal sequenced light-curves for each of the 20 spectral bins. This gives us the opportunity to perform a different method to de-trend the instrumental systematics from the data. Independent Component Analysis (ICA, see chapter 3) has been used effectively to remove instrument systematics and other astrophysical signals in exoplanetary light-curves obtained with Kepler, HST/NICMOS (Waldmann, 2012; Waldmann et al., 2013), Spitzer/IRS (Waldmann, 2014) and Spitzer/IRAC (Morello et al., 2014, 2015; Morello, 2015; Morello et al., 2016).

Thanks to the collaboration with Dr. Giuseppe Morello, we have been able to apply ICA to de-trend HST/WFC3 light-curves for the first time. The process follows these steps:

1. ICA transformation of the time-series to get the independent components;
2. identification of the transit components by inspection;
3. fitting a linear combination of non-transit components plus the calculated transit model $F(t)$ (see Sec. 4.5.1), to the white light-curve, using the coefficients of the terms of the linear combination as free parameters;
4. subtraction of the non-transit components, with coefficients determined by the fitting, from the light-curves.

To obtain the ICA components we have adopted the MULTICOMBI algorithm (Tichavsky et al., 2008). This is an hybrid technique that is able to separate non-Gaussian and time-correlated sources, mixing respectively two popular algorithms: EFICA (Hyvärinen & Oja, 2000) and WASOBI (Yeredor, 2000). Among all the components, one contains most of the transit signal and it is excluded, the others (non-transiting components) are then simultaneously fitted together with a transit model to a reference light-curve. Also in this case we followed the two steps as presented in the previous section, in particular the light-curves were:

- stacked, i.e. we sum all the stripe light-curves to obtain a unique light-curve per spectral bin;

- weighted, i.e. we fit every single spectral light-curve per bin and we take the weighted result.

The transit model, $F(\lambda,t)$, used is the same described in Sec. 4.5.1 using the parameters shown in Tab. 4.4. In both cases the model used to perform the fitting on the white light-curve is:

$$M(\lambda,t) = \left(\sum_j o_j C_j\right) + n_\lambda F(\lambda,t) \tag{4.22}$$

where o_j are the coefficients of the ICA components, C_j are the retrieved components, n_λ is the coefficient for the transit model and $F(\lambda,t)$ is the transit model per spectral bin. In this case the free parameters are all the coefficients for the linear combination and the parameter $R_p/R_\star$ of the transit model.

A possible evolution to this approach is also to include the residuals obtained for the fitting on the white light-curve as an additional component (Eq. 4.23). This process can account for possible systematics common to all wavelength bins.

$$M(\lambda,t) = \left(\sum_j o_j C_j\right) + n_\lambda F(\lambda,t) + n_W white_{res} \tag{4.23}$$

Following the ‘weighted’ approach all the single stripe spectral light-curves are noisier than the stacked one mainly due to correlated noise and lower signal. ICA, however, is effective in correcting these sources of noise. The systematics are hard to visualise also due to their low amplitude in this dataset. Fig. 4.15 shows an example of two raw spectral light-curves, single stripe and summed (stacked), and the de-trended ones, with and without the white residuals component.

Finally, after processing all the fittings for the two methods proposed (stacked and weighted) for every spectral bin using both Eq. 4.22 and 4.23, we were able to obtain the final 1D spectrum for all the combinations (see Fig. 4.16). The final spectra are all consistent witch each other.

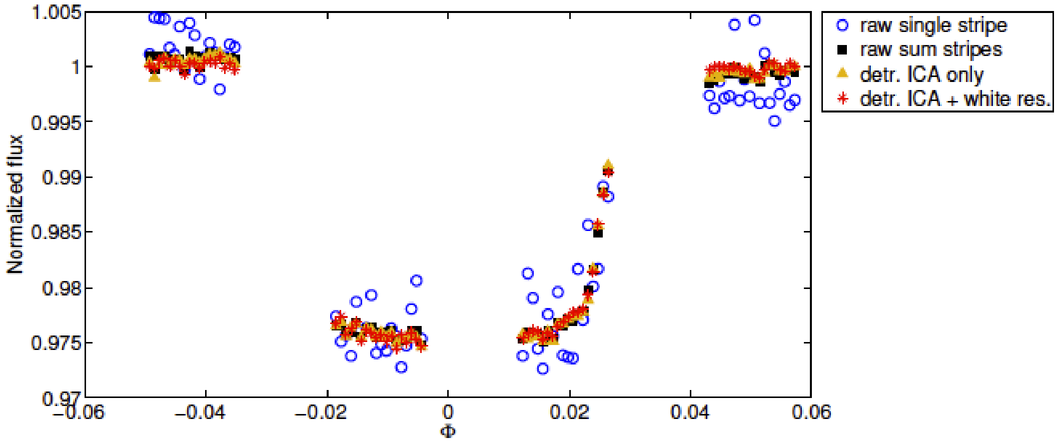

Figure 4.15: The figure shows two spectral light-curve (1.3657 – 1.3901μm) referred to the two method adopted (stacked and weighted) before and after the ICA correction. Font: priv. comm. of Dr. Giuseppe Morello

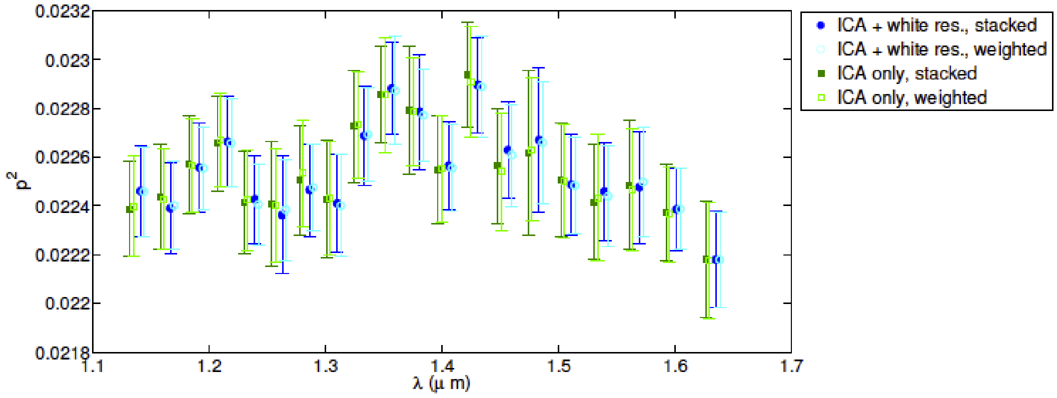

Figure 4.16: Comparison between all the spectra obtained by using all the combinations of the method proposed. An horizontal shift has been introduced to make easier the visualisation. Font: priv. comm. of Dr. Giuseppe Morello.

If we compare, the 1D final spectrum obtained using the parametric method, to the one obtained using ICA, the only noticeable differences are the larger error bars of the ICA method (Fig. 4.17). This is due to an additional error term that needs to be added to take into account the uncertainty of the decomposition process (Morello, 2015; Morello et al., 2016).

$$\sigma_{ICA}^2 = \sum_j o_j^2 \mathbf{isr}_j \tag{4.24}$$

where $\mathbf{isr}_j$ is the so-called interference-to-signal-ratio (Eq. 3.17, Sec. 3.4.1) (Morello, 2015; Morello et al., 2016), and o_j are the coefficients of the non-transit components determined by the fitting. The σ_{ICA} term is the weighted sum of the errors attributed to the independent components extracted by ICA. Finally, the comparison between the two different analysis methods is plotted in Fig. 4.17 and the numerical values are shown in Tab. 4.7.

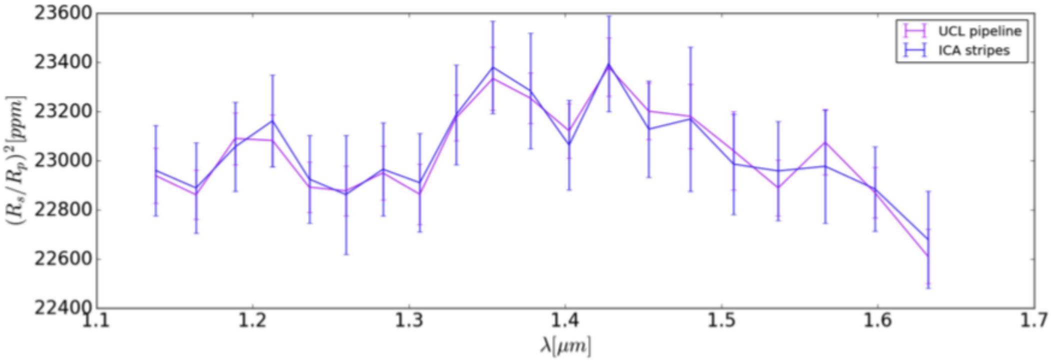

Figure 4.17: Stacked spectra obtained with the parametric pipeline (magenta) and with the ICA + white residuals stacked approach (blue).

4.6.2 Atmospheric retrieval

The last step of our analysis is to interpret the 1D spectrum to understand which molecules may cause the modulation in wavelengths. To accomplish this last step we used the spectral retrieval $\mathscr{T}$-REx code (Sec. 2.3) (Waldmann et al., 2015b,a).

As input, to the $\mathscr{T}$-REx code, we assumed an atmosphere dominated by molecular hydrogen and helium, with a mean molecular weight of 2.3 amu. We considered as candidate trace gases a broad range of molecules, including H_2O, C_2H_2, CH_4, CO_2, CO, HCN, NH_3, VO and TiO. However, the RobERt (Robotic Exoplanet Recognition, Waldmann (2016)) module restricted the list of detectable molecules, based on the observed spectral pattern, to H_2O, VO and TiO.

Given the relatively narrow spectral range probed, we assumed an isothermal profile and molecular abundances constant with pressure. In addition, we set uniform priors to the fitted parameters, which were: the mixing ratios of the molecules (10^{-12} – 10^{-1}), the effective temperature of the planet (1400–2100 K), the radius

$\lambda_1 - \lambda_2\,(\mu\mathrm{m})$		a_1	a_2	a_3	a_4	$(R_\mathrm{p}/R_*)^2$ (ppm)	$(R_\mathrm{p}/R_*)^2$ (ppm)
						UCL pipeline	ICA
1.1250	1.1511	0.632741	-0.481904	0.701108	-0.306091	22940 ± 112	22961 ± 184
1.1511	1.1767	0.619205	-0.434713	0.64011	-0.282483	22862 ± 100	22890 ± 184
1.1767	1.2011	0.614294	-0.41589	0.610565	-0.272242	23091 ± 105	23057 ± 181
1.2011	1.2247	0.599151	-0.360648	0.544934	-0.247917	23083 ± 105	23163 ± 186
1.2247	1.2480	0.584001	-0.29953	0.465487	-0.216442	22893 ± 102	22926 ± 179
1.2480	1.2716	0.581928	-0.282551	0.441745	-0.210655	22878 ± 102	22863 ± 242
1.2716	1.2955	0.58946	-0.229732	0.322997	-0.169253	22951 ± 110	22966 ± 189
1.2955	1.3188	0.57237	-0.227002	0.362724	-0.181489	22864 ± 123	22911 ± 200
1.3188	1.3421	0.569522	-0.202303	0.325228	-0.166816	23176 ± 94	23188 ± 203
1.3421	1.3657	0.564634	-0.163366	0.265035	-0.14235	23335 ± 129	23381 ± 189
1.3657	1.3901	0.561817	-0.127278	0.200548	-0.113503	23255 ± 103	23285 ± 236
1.3901	1.4152	0.561832	-0.0979712	0.148201	-0.0914278	23122 ± 111	23064 ± 182
1.4152	1.4406	0.572262	-0.100901	0.133369	-0.0848254	23382 ± 119	23396 ± 195
1.4406	1.4667	0.58462	-0.111943	0.124656	-0.0799948	23202 ± 115	23129 ± 196
1.4667	1.4939	0.600205	-0.136878	0.140204	-0.0874595	23181 ± 130	23170 ± 294
1.4939	1.5219	0.609784	-0.134319	0.11158	-0.0721681	23041 ± 160	22987 ± 204
1.5219	1.5510	0.626375	-0.139701	0.0839621	-0.0555132	22890 ± 114	22959 ± 201
1.5510	1.5819	0.647904	-0.193435	0.120068	-0.0635888	23076 ± 131	22978 ± 230
1.5819	1.6145	0.663831	-0.223633	0.124246	-0.0583813	22871 ± 102	22886 ± 171
1.6145	1.6500	0.686226	-0.267069	0.137329	-0.0557593	22611 ± 111	22680 ± 198

Table 4.7: Limb darkening coefficients a_{1-4} and transit depth $(R_p/R_*)^2$ for the wavelength channels.

of the planet (1.56-2.10 R_{Jup}), and the cloud top pressure (10^{-3} – 10^{6}Pa).

The transmission spectrum of HAT-P-32b and the best fit calculated by $\mathscr{T}$-REx are shown in Fig. 4.18. The best-fitting values and the posterior distributions are shown in Tab. 4.8 and Fig. 4.19. With the exception of water vapour, the fitted values for all of the other molecular mixing ratios are smaller than 10^{-7}. This result means that they are not detectable from this dataset. The water vapour mixing ratio oscillates, instead, between $\log \mathrm{H_2O} = -3.45^{+1.83}_{-1.65}$ depending on the clouds' top pressure, which could occur between 5.16 and 1.73 bar. A strong correlation between the water vapour mixing ratio, the clouds' top pressure, the planetary radius at 10 bar, and temperature is noticeable in Fig. 4.19, indicating there is a degeneracy in the solutions space.

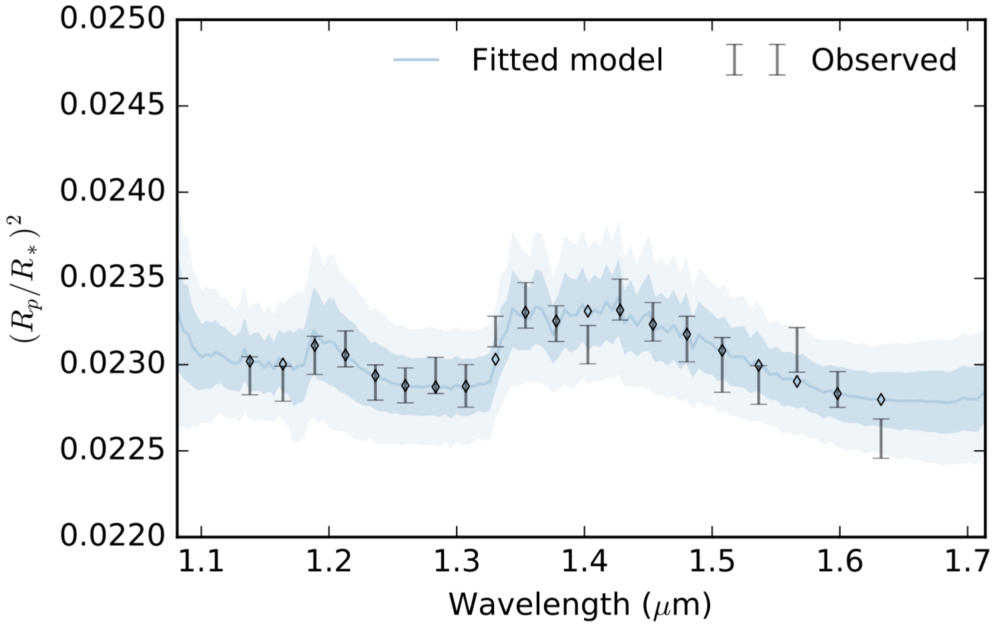

Figure 4.18: Transmission spectrum of HAT-P-32b obtained with the parametric pipeline (black) and best-fitting model (light-blue).

4.6.3 Discussion

As mentioned in the previous section, the error bars obtained with ICA are larger by a factor of ~1.6–1.8 compared to the ones obtained with the parametric fitting. The

Table 4.8: Fitting results for HAT-P-32 b atmosphere

Atmospheric Retrieval results	
$\log \mathrm{H_2O}$	$-4.66^{+1.66}_{-1.93}$
$T_{\mathrm{eff}}\,[\mathrm{K}]$	1553^{+174}_{-91}
$R_p\,[R_{\mathrm{jup}}]$	$-1.76^{+0.05}_{-0.04}$
$P_{\mathrm{cld,top}}\,[\mathrm{bar}]$	$3.39^{+1.77}_{-1.66}$

larger error bars obtained with ICA are the trade-off for higher objectivity, due to the lack of any assumption about the instrument systematics compared to the parametric approach. The ICA error bars are worst-case estimates. It is worth noting that the discrepancies between the spectra obtained with the different methods are smaller than the parametric error bars, suggesting that, in this case, the ICA error bars might be overly conservative.

Previous ground-based observations of the transit of HAT-P-32b in the optical wavelengths (Gibson et al., 2013; Zhao et al., 2014; Mallonn & Strassmeier, 2016; Nortmann et al., 2016) did not find evidence of spectral modulations due to atoms, ions and molecules suggesting a cloudy atmosphere. Our cloud top pressure is consistent with their measurements within 1σ, hence the water detection in the infrared is not controversial.

Water vapour has been detected, to date, in the atmospheres of about 25 hot Jupiters (Iyer et al., 2016; Tsiaras et al., 2018). Stevenson (2016) identify two classes of hot-Jupiters, essentially mostly cloudy or with a strong water signature. The observed trend suggests that hotter ($T_{eq} > 700\,\mathrm{K}$) and more inflated ($\log g > 2.8$) planets are more likely to have a strong water signature than cooler and smaller ones, but the current sample is not statistically significant. In agreement with this scenario, we find that HAT-P-32b ($T_{eq} = 1786\,\mathrm{K}$; $\log g > 2.8$) has one of the strongest water features so far detected (~ 500 ppm, 5.3 σ).

The detection of the water vapour is important to confirm previous studies and theories. This molecule is indeed predicted to be among the most abundant (if not the most abundant) molecular species after hydrogen in the atmospheres of

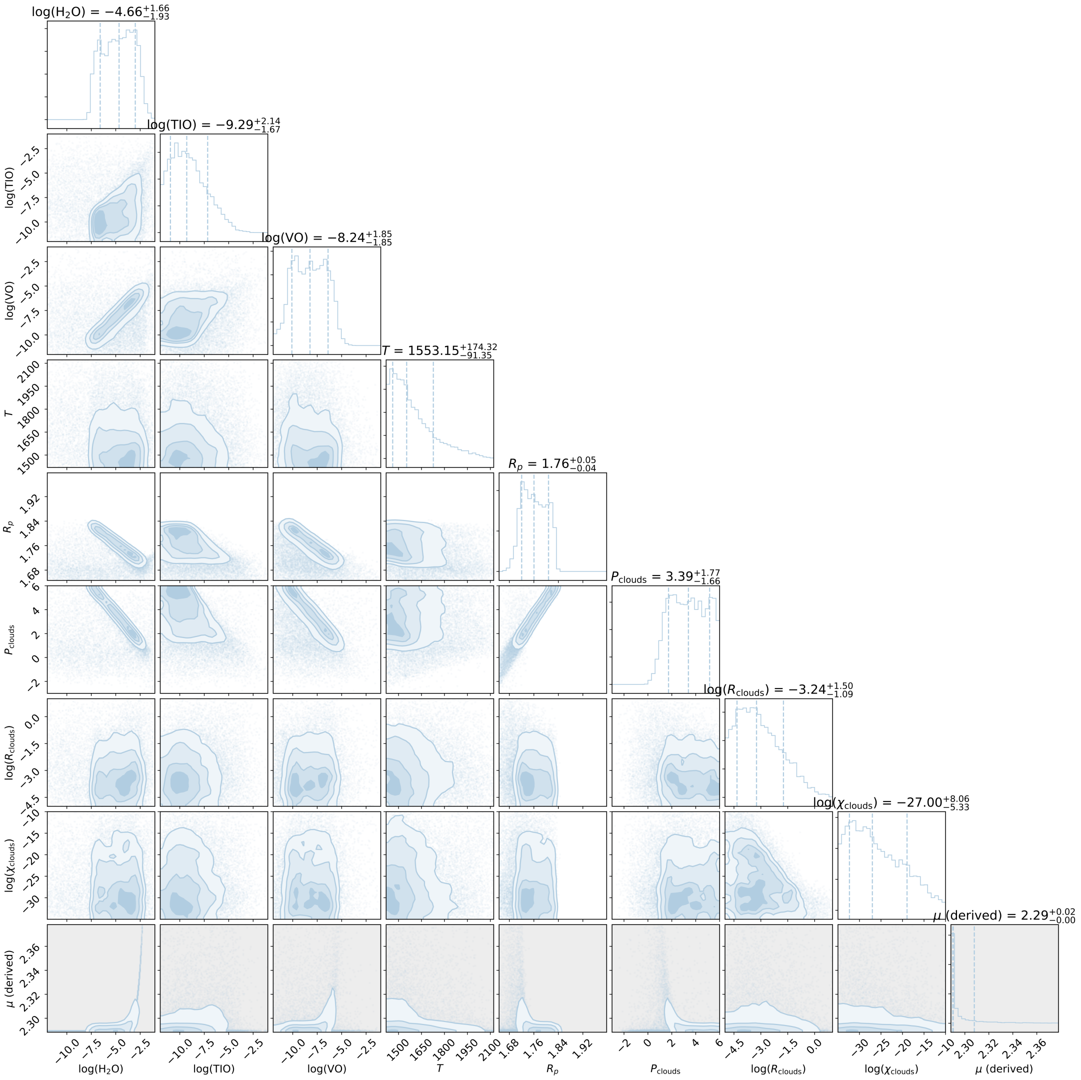

Figure 4.19: Posterior distributions to the fit for the WFC3 spectrum of the giant planet HAT-P-32b. Even though we tested the presence of many other molecules in this atmosphere, here we show only the posterior of H_2O because it is the only significant one. All of the other molecules do not show a statistically significant contribution to the fit.

close-in extrasolar giant planets, i.e. hot-Jupiters (Seager & Sasselov, 2000; Brown, 2001). Moreover, the abundance of water vapour could help to characterize the environment where the planet formed and evolved. The presence of water combined to a detection of a carbon-based molecule (e.g. CO, CO_2, CH_4) can led to the calculation of the C/O ratio (Madhusudhan, 2012). In an environment where the

C/O value ranges from 0.5 (solar value) to 2, the H_2O and CH_4 abundances can vary by several orders of magnitude. Carbon-based molecules such as HCN and C_2H_2 become prominent for C/O $\geq$ 1, while the CO abundance remains almost unchanged. Moreover, a C/O $\geq$ 1 can prevent a strong thermal inversion due to TiO and VO in a hot-Jupiter atmospheres, since TiO and VO are naturally under-abundant in a carbon-rich environment (Madhusudhan, 2012).

From Fig. 4.19 it is possible to note that the mixing ratio of the water vapour is correlated with the clouds top pressure. Higher are the clouds in the atmosphere lower is the retrieved concentration of the water. Generally the clouds' altitude is regulated by the atmospheric temperature (Barstow et al., 2017), and in the case of primary transit (e.g. the dataset presented in this chapter) the temperature can not effectively constrained.

Spectroscopic data over a broader wavelength range, especially in the IR, will be needed to de-correlate the water vapour's mixing ratio from clouds and identify other possible molecular species in HAT-P-32b atmosphere. Even if in the optical range, observations did not show significant modulation (Gibson et al., 2013; Zhao et al., 2014; Mallonn & Strassmeier, 2016; Nortmann et al., 2016), it will be worth to try at longer wavelength. Additionally, a powerful technique, that can help to break degeneracies, is using high-resolution observations and this is described in the next chapter.

Chapter 5

Ground observations

"We must trust to nothing but facts: these are presented to us by nature and cannot deceive. We ought, in every instance, to submit our reasoning to the test of experiment, and never to search for truth but by the natural road of experiment and observation."

– Antoine Lavoisier - 1790

High-resolution spectroscopy (HRS) allows us to resolve molecular bands into individual lines. Using radial velocity measurements and techniques such as Cross-Correlation Function (CCF) it is possible to separate three physically different sources: telluric absorption, stellar signal and the planetary spectrum, which are entangled in the recorded spectrum. The aim – but also the biggest challenge – is to recognise the planetary signal among the telluric and the stellar signals, which can be orders of magnitude stronger. The standard method used in the literature to analyse HRS data is to apply a number of manual corrections which involve the correction of airmass variations, the subtraction of a modelled stellar spectrum from the data and the use of ad-hoc masks to eliminate residual strong features (Snellen et al., 2010; Birkby et al., 2013, 2017; Birkby, 2018; Brogi et al., 2014, 2016, 2018).

In this chapter I present and assess an alternative automatic procedure to reduce HRS data which requires no manual intervention that could interfere with the objectivity and repeatability of the analysis. My analysis method is based on a use of Principal Component Analysis (PCA) and Cross-Correlation Function (CCF). The exoplanetary atmosphere has been simulated using $\mathscr{T}$-REx (Waldmann et al., 2015b) and line lists from the ExoMol project (Tennyson et al., 2016).

The technique described here has been developed to initially analyse VLT/CRIRES data, but, it may also be used on other current or future instruments. These include VLT/CRIRES+ (Follert et al., 2014), TNG/GIANO-B, a high dispersion spectrograph (Oliva et al., 2012) which covers 0.9 to 2.5 μm with a resolution of R=50,000, SUBARU/IRCS (Kobayashi et al., 2000), which uses a lower resolution (R=20,000) but covers a broader range (from 1 to 5 μm), SUBARU/HDS an high dispersion spectrograph with a resolution of R=165,000 in a narrow wavelength range (0.62 – 0.88µm) (Noguchi et al., 2002) and CARMENES at Calar Alto Observatory (Quirrenbach et al., 2014) with a spectral resolution up to 80,000 in the near-IR ($0.9 - 1.7$ μm).

5.1 Code overview

I have initially developed the code to analyse VLT/CRISES datasets. But it may easily handle also different instruments with few modifications (e.g. see Sec. 5.8). Fig. 5.1 shows the flow chart of the code. This is divided in four main phases and it is assembled in routines and modules for easier upgrades and debugging.

The process starts with the download of raw datasets from the instrument archive, the VLT one in this case. A parameter object is then created where all the known parameters of the planetary system are stored. The data are then aligned to the telluric spectrum so that all the spectra are in the telluric frame. The data are corrected using unsupervised algorithms and correlated subsequently with a synthetic exoplanet atmospheric model to obtain the final result (SNR map and Welch's T-test). An additional module, which can be disabled, allows to inject a synthetic atmospheric model to the data to test the effect of subsequent corrections and in which conditions the injected signal is retrieved.

The code is written entirely in *Python2.7* except for the "reduction pipelines" module (see. Fig. 5.1, the red box) that may be written in different languages (e.g. JavaScript or C).

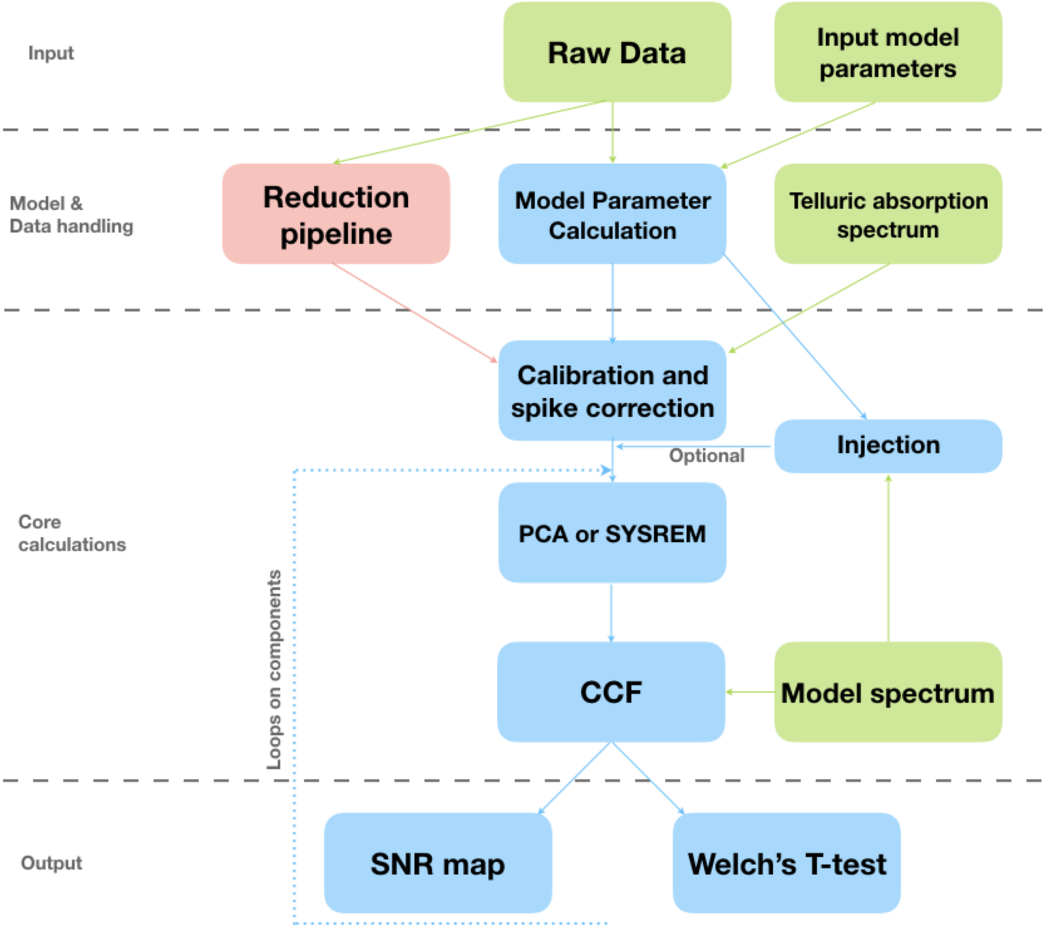

Figure 5.1: Flow chart of the pipeline for high spectral resolution data analysis. The box colours indicate different classes of action: green boxes represent an external input coming from other models or different sources. The red box includes all the different reduction pipelines supported. Finally, the blue boxes contain most of the calculations that I developed.

5.2 Model and input data handling

The process begins by obtaining the raw images from the respective archives[1][2]. For the technical discussion I will use the VLT/CRIRES datasets for the rest of this chapter, in Sec. 5.8 I will show the work in progress on TNG/GIANO-B frames.

Orbital parameters and values used by the pipeline are estimated. In particular,

[1] VLT/CRIRES: `http://archive.eso.org/eso/eso_archive_main.html`

[2] TNG/GIANO-B: `http://archives.ia2.inaf.it/tng/faces/search.xhtml?dswid=3218`

in the “input parameters” in Fig. 5.1 the user may provide stellar parameters (i.e. radius and temperature) and planetary parameters such as radius, semi-major axis, orbital period, mid-transit point, inclination and, most importantly, the systemic velocity (i.e. the relative velocity between the Sun and the target star). If these parameters are not provided, the Open Exoplanet Catalogue python package (Rein, 2012; Varley, 2016) is then called to complete the list. All the input parameters are converted in the MKS system. Other parameters that should be included are: type of observation (e.g. transit, eclipse or phase variation, for transiting planet and not) and which molecules the pipeline should consider for cross-correlation with the data.

The orbital phase is calculated as follows:

$$\phi(t) = \frac{t - T_0}{P_{orb}} \tag{5.1}$$

where ϕ is the orbital phase, t is the time of the observation, T_0 is the mid transit point and P is the orbital period. Knowing the orbital phase of each observation is useful for computing the planetary radial velocity at a specific time

$$V_p(t) = K_p \cdot \sin(\phi(t)) \tag{5.2}$$

$$K_p \;=\; v_{orb}\, sin(i) \tag{5.3}$$

$$v_{orb} \;=\; \frac{2\pi a}{P_{orb}} \tag{5.4}$$

where $V_p(t)$ is the radial velocity of the planet at time t, K_p is the radial velocity amplitude, v_{orb} is the orbital velocity, i is the inclination and a is the semi-major axis.

An important correction is then calculated for each observation. Since the telescope is on Earth which is moving around the barycentric point of the solar system, we choose a reference frame where the observer is at rest. The barycentric velocity correction, v_{bary}, is calculated using the *baryCorr* function inside the PyAstronomy python package[3].

At this stage the transit time is also evaluated to determine which image of the dataset contains the beginning of the transit (ingress) and which one contains the end of it (egress). We start by defining a parameter that effectively affects the transit time: the impact parameter b is defined as the sky-projected distance between the centre of the stellar disc and the centre of the planetary disc at conjunction (middle of transit)

$$b = \frac{a\cos(i)}{R_\star} \tag{5.5}$$

where a is the semi-major axis, i is the orbital inclination and $R_\star$ is the radius of the star. If the orbital inclination is 90° ($b = 0$), then the planet will cross the star at the equator and the transit time will be maximum, otherwise the higher is b the less will be the duration of the transit. An expression for the transit duration, is given by (Seager & Mallén-Ornelas, 2003; Kipping, 2010)

$$T_{transit} = \frac{P}{\pi}\arcsin\left(\frac{\sqrt{1-b^2}}{a_R\sin(i)}\right) \tag{5.6}$$

where P is the orbital period, and a_R is the semi-major axis in units of stellar radii ($a_R = a/R_\star$).

To detect the weak planetary signal, the basic calibration is not accurate enough. Generally the reduction pipeline use a molecular lamp with known wavelength line position (e.g. ThAr and OH). However normally the lamp lines are not enough to ensure a calibration sufficiently precise for our goal. For this reason, we followed the procedure described in the literature (de Kok et al., 2013; Snellen et al., 2010; Birkby et al., 2013, 2017; Brogi et al., 2013, 2014, 2016), which involves a further calibration with telluric absorption spectrum(see Sec. 5.3) using the ESO Sky Model Calculator tool[4]. SKYCALC simulates the telluric absorption spectrum

[3] https://github.com/sczesla/PyAstronomy

[4] https://www.eso.org/observing/etc/bin/gen/form?INS.MODE=swspectr+INS.NAME=SKYCALC

for a specific night.

5.3 Calibration and spikes correction

During the previous steps raw images have been analysed and the 1D spectrum for each image has been extracted (see Sec. 5.7 and 5.8). The first step is to normalise each spectrum of each detector by dividing it by its median. This step is necessary to avoid differences of baseline across spectra due to e.g. slit loss. In Fig. 5.2 the spectrum extracted from one of the images for the four CRIRES detectors is shown. The signal is still dominated by telluric absorption lines In Fig. 5.3 we show the telluric transmission spectrum generated using the SKYCALC tool, setting the option for clear sky.

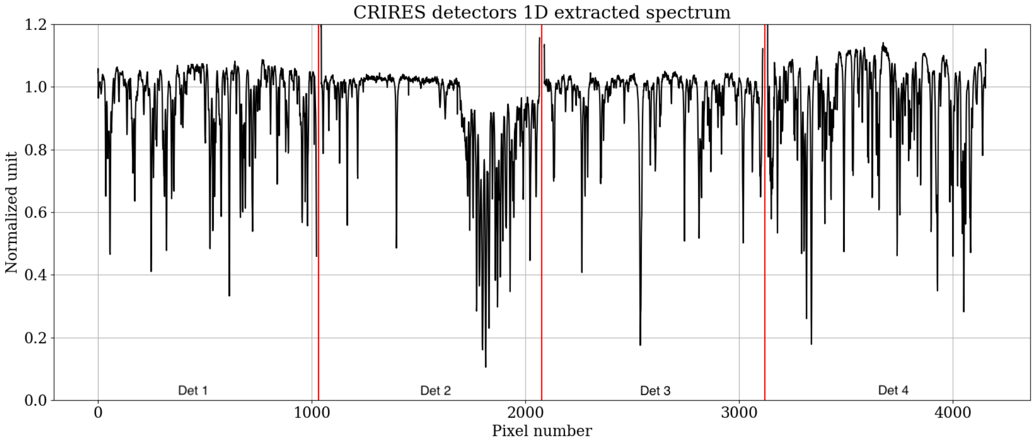

Figure 5.2: 1D extracted spectrum for the four CRIRES detectors (HD209458b dataset).

The goal of this process is to determine the relationship pixel-wavelength, by comparing the position of strongest lines in both the telluric transmission and the 1D extracted spectra.

Working on each detector at a time, we consider the mean spectrum. Here, the strongest lines (all the lines reaching a minimum < 0.8) have been identified as homogeneously distributed as possible to cover the whole x-axis range. These same lines are also been identified within the telluric template. A Gaussian fit is then performed for each of these lines and the centroid is taken. The extracted spec-

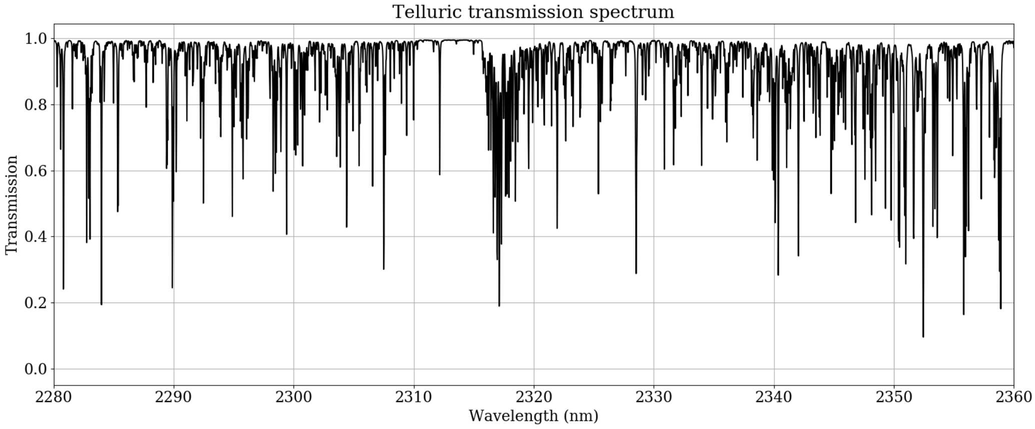

Figure 5.3: Telluric transmission spectrum in the nominal wavelengths of the CRIRES spectrograph generated by the SKYCALC tool for clear sky scenario.

trum centroids indicate the pixel number position. In the telluric template, instead, they indicate wavelength positions of the lines. The pixel positions are then plotted against the wavelengths (Fig. 5.4, top panel). At a first look the relationship may appear as a linear trend, but, the residuals of a linear fit still show some correlations (Fig. 5.4, mid panel) and therefore an higher order fit is required.

A second and third order fits reduce the standard deviation of the residuals but they are not good enough as they well reproduce the central part of the spectrum but fail to reproduce the wings. A fourth order fit is required for all of the four detectors. In Fig. 5.4 (bottom panel), the residuals of a fourth order fit are shown and no correlations appear. We can estimate the precision of a fit, in terms of velocity by using

$$\Delta V = \frac{std(residuals) \cdot c}{\lambda_c} \tag{5.7}$$

where $std(residuals)$ is the standard deviation of the residuals, c is the speed of light and λ_c is the central wavelength of the spectrum. The error in terms of velocity for all the detectors is about 1 $\mathrm{km\,s^{-1}}$. This is not a problem when analysing a GIANO-B dataset, since the pixel resolution is $\sim$ 3 $\mathrm{km\,s^{-1}}$. The CRIRES spectrograph, on the other hand, has a resolution of $\sim$ 1.5 $\mathrm{km\,s^{-1}}$ per pixel and the current calibration method could affect the result. This method needs to be improved when

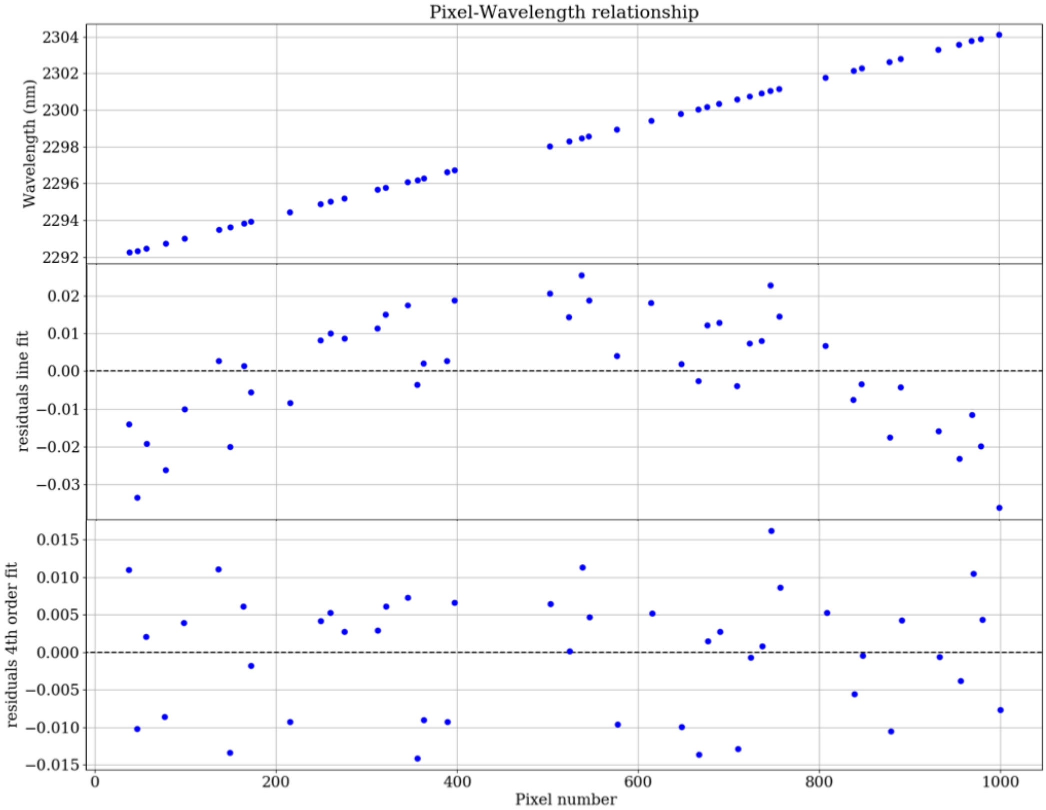

Figure 5.4: Pixel-wavelength relationship for detector 1 of the HD209458b dataset. **Top panel** shows the relationship between pixel numbers and wavelength. **Mid panel** shows the residuals after a linear fit, a trend is still visible. **Bottom panel** shows residuals after fourth order fit.

working with spectrographs of even higher spectral resolution such as the SUBARU/HDS ($R = 165,000$).

All the single spectra are then interpolated via a third order spline to the derived wavelength grid to have the same grid for all the spectra. This procedure has been used previously (Snellen et al., 2010; Brogi et al., 2013, 2018), but, for future instruments it needs to be improved to obtain a precision less than 0.5 $\mathrm{km\,s^{-1}}$.

I analysed each detector separately as a two-dimensional matrix, where the x-axis contains wavelengths and the y-axis time: every row of this matrix is a spectrum, every column is a temporal-series at a given wavelength (see Fig. 5.5).

Finally, the pipeline removes all the cosmic rays or spikes that could occur at the edges of the spectra due to the spline interpolation to the wavelength grid. The

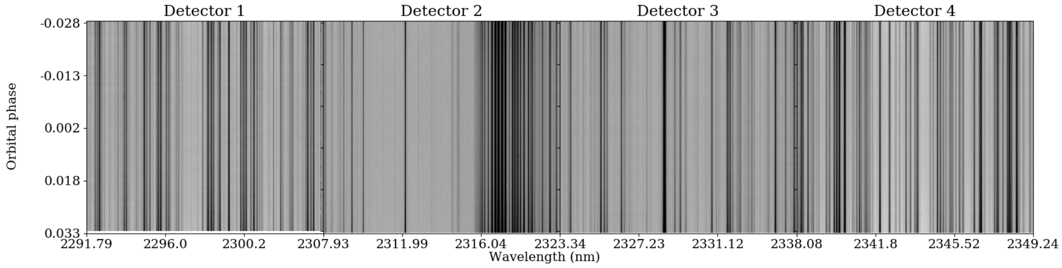

Figure 5.5: The HD209458b (full discussion see Sec. 5.7) dataset after calibration. The y-axis maps the orbital phase. The x-axis the wavelength grid.

pipeline takes one column at a time of each 2D matrix, it calculates the median of the column and all the values outside 3σ from the median are set to the median value.

5.4 PCA and SYSREM

At this point, the data are treated with the algorithms developed and described in Sec. 3.2 and 3.3. The user selects which algorithm to use from the parameter file.

Pre-processing is needed at this point (Sec. 3.2), one can select: standardisation (ST) and mean features subtraction (MFS). In Fig. 5.6 is depicted the first detector of the HD189733b dataset (see Sec. 5.7 for full discussion) after the calibration process, the two pre-processing methods have been applied and the result is shown in Fig. 5.7. The main differences are:

- The MFS matrix shows the variation of telluric lines throughout the visit mainly due to airmass variations, while, in the ST matrix these features are flatter and more extended.

- In the colour-map meter in Fig. 5.6, the range of values in the MFS in the first case has been reduced, while in the ST case, the range is even bigger than the data prior to pre-processing (range= $[0-1.1]$).

After trying both methods I noticed that MFS is more effective in terms of final SNR output (Sec. 5.6.1).

The datasets are now analysed by PCA or SYSREM (see chapter 3). However, while the decomposition using SYSREM is not affected by the orientation of the

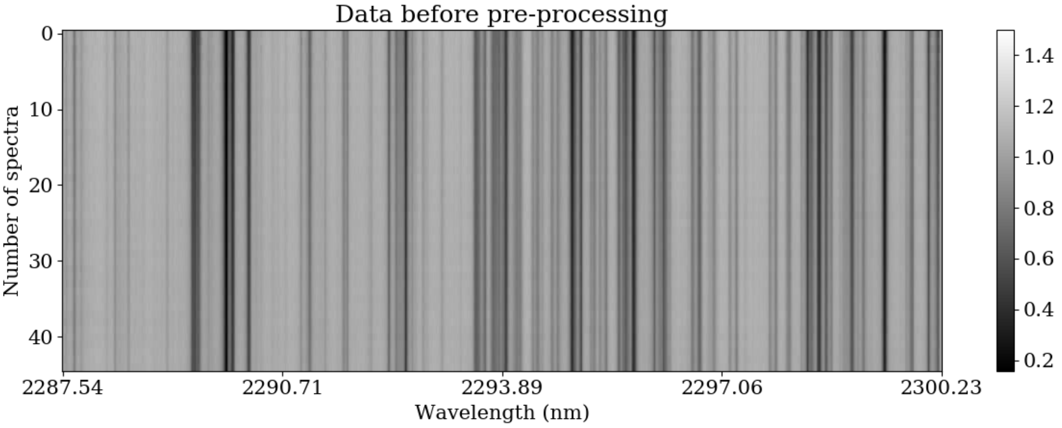

Figure 5.6: Detector 1 of the HD189733b dataset prior to the application of any pre-processing algorithm.

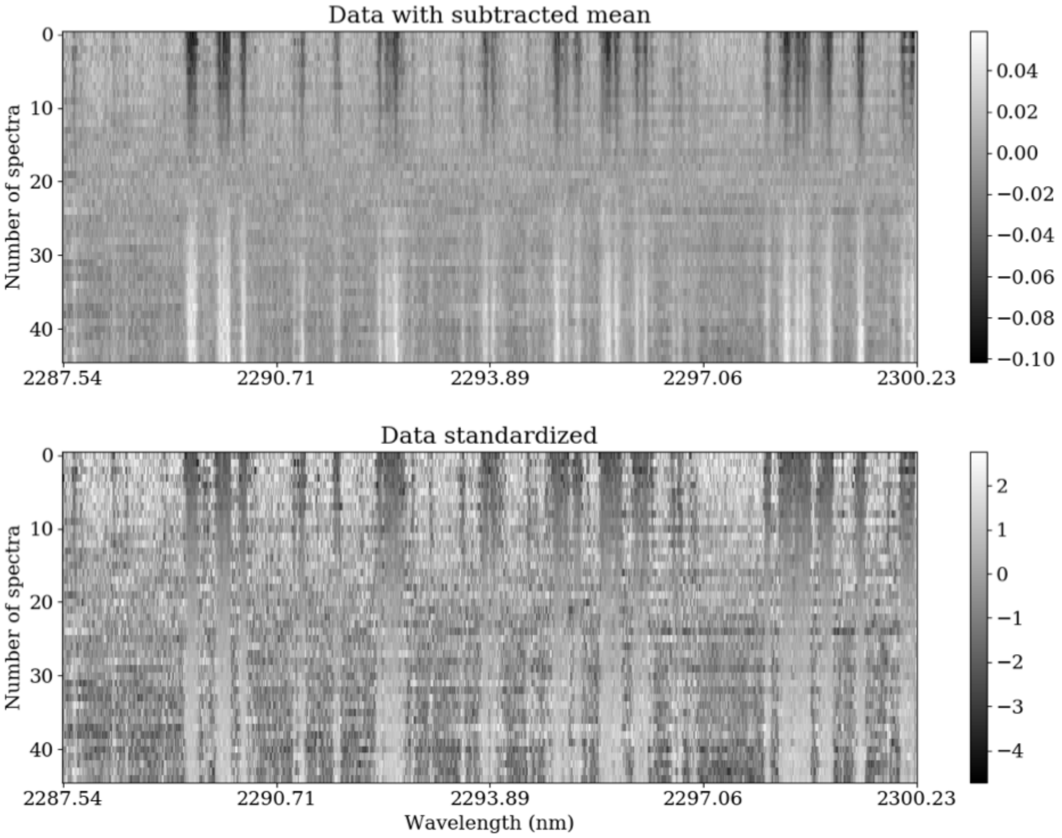

Figure 5.7: **Top Panel:** detector 1 of the dataset HD189733b with the mean of each column subtracted. **Bottom Panel:** same image but with the standardisation applied as pre-process step.

input matrix, PCA components change if the algorithm is applied to a matrix or to its transpose. It is worth to explore what happens when the data matrix is analysed.

5.4.1 Time or wavelength domain?

On a typical high resolution spectroscopy dataset the number of spectra are less than the wavelength bins, resulting in matrices that have way more columns than rows. In the HD189733b dataset, for example, the data matrix per detector has the dimensionality 45×1024. Using the algorithm described in Sec. 3.2 the dimension of the covariance matrix and the number of principal components (eigenvectors) are equal to the number of rows of the input matrix. Two cases are then considered:

- **time domain matrix (TDM)**; we use the spectra as row arrays and wavelength bins as columns. In this case the covariance matrix has the dimensionality 45×45 and 45 components will be obtained.
- **wavelength domain matrix (WDM)**; we transposed the input matrix to use rows to indicate the wavelength bins and columns for the number of spectra. In this case the covariance matrix will have the dimensionality 1024×1024 and 1024 components will be calculated;

On top of different computational efficiencies, these two processes return two different spaces. In the WDM case the components contain the information of wavelength correlations among the data. In the TDM case (using the transposed input matrix) the components matrix contains information of the correlations along the time direction. In the literature both procedures are used (de Kok et al., 2013; Piskorz et al., 2016, 2017). Here I will compare both methods to show differences and/or advantages.

We apply PCA to the matrix shown in Fig. 5.7 top panel. Both WDM and TDM covariance matrices are shown in Fig. 5.8. By resolving the characteristic polynomial of these matrices it is possible to calculate the eigenvectors and eigenvalues.

In Fig. 5.9 we show the first five components in order of variance of the 45×45 covariance matrix (left) and the fist five of the 1024×1024 covariance

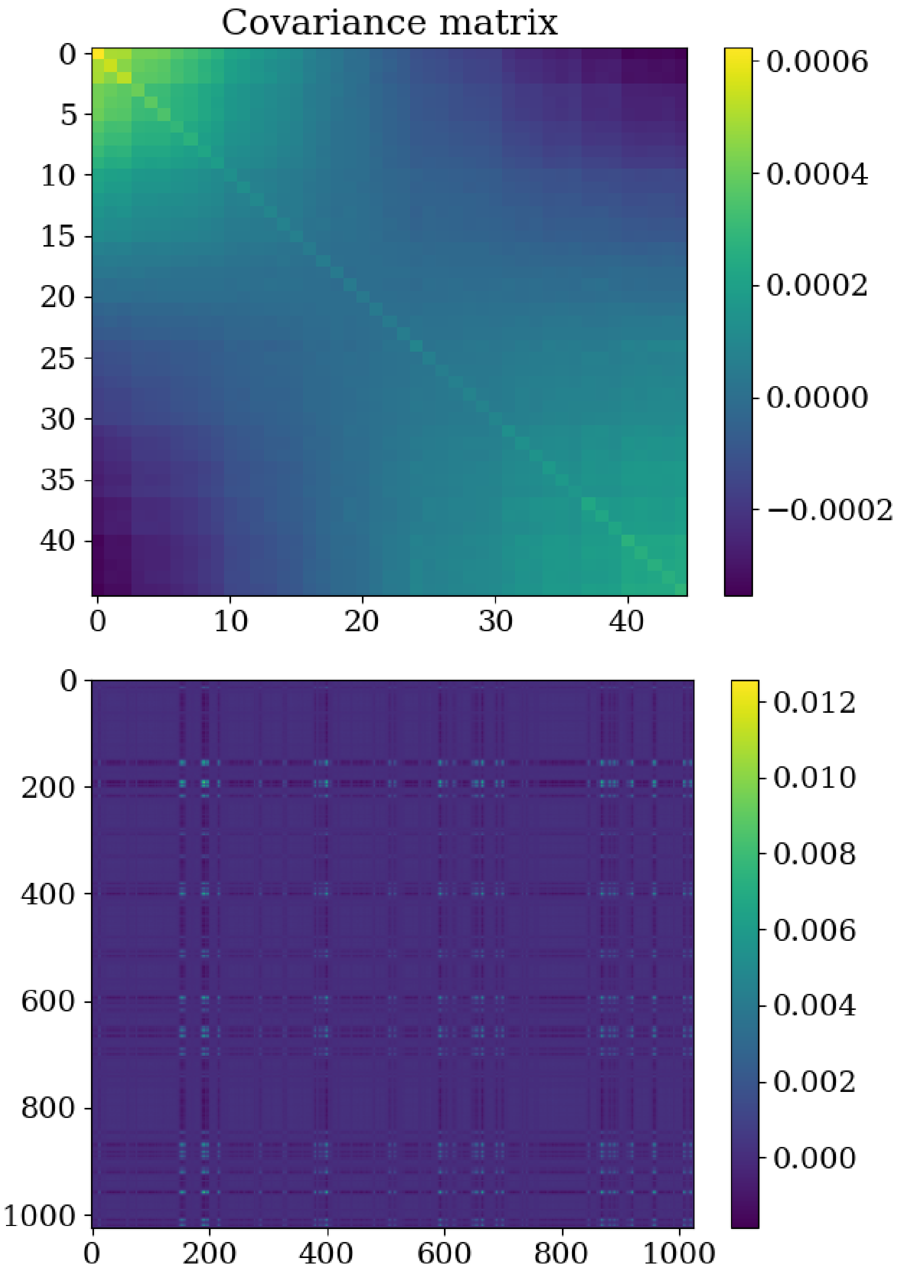

Figure 5.8: **Top Panel:** covariance matrix of the detector 1 in the HD189733b dataset, where the input matrix has been transposed (TDM case): the dimension of this matrix is the number of spectra taken during the observation (i.e. 45). **Bottom Panel:** same as top but the input matrix has not been transposed (WDM case) and the dimension is equal to the number of the wavelength bins (i.e. 1024).

matrix (right). As the data are highly affected by the telluric absorption, the first components are indeed closely linked to this effect. The first components on the

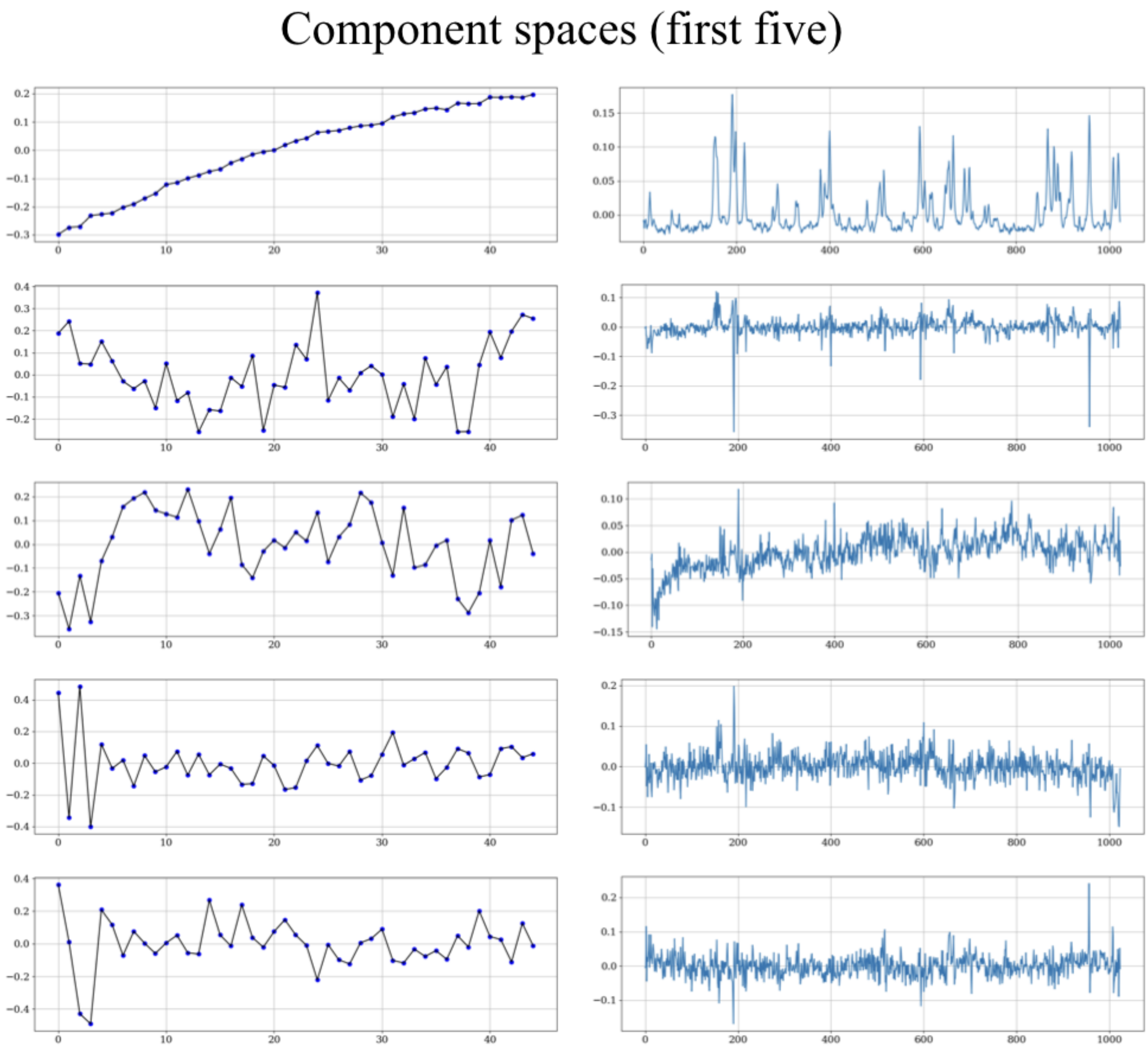

Figure 5.9: **Left panels** show the first five eigenvectors of the TDM case (45×45) (**top panel** Fig. 5.8). These components depict the time correlation of the data along the time domain. **Right panels** show the five eigenvectors of the WDM covariance matrix (1024×1024) (**bottom panel** Fig. 5.8).

left-hand side of Fig. 5.9 contain information on the time-domain and the first one, in particular, is linked to the variations of the airmass: these are linearly correlated as we can appreciate from Fig. 5.10. Components on the right-hand side show correlation in the wavelength domain and in particular all the components are correlated with the telluric transmission spectrum. A good example is the strong feature around 200 (Fig. 5.9 x-axis unit) (∼2290nm) that persists in all the components.

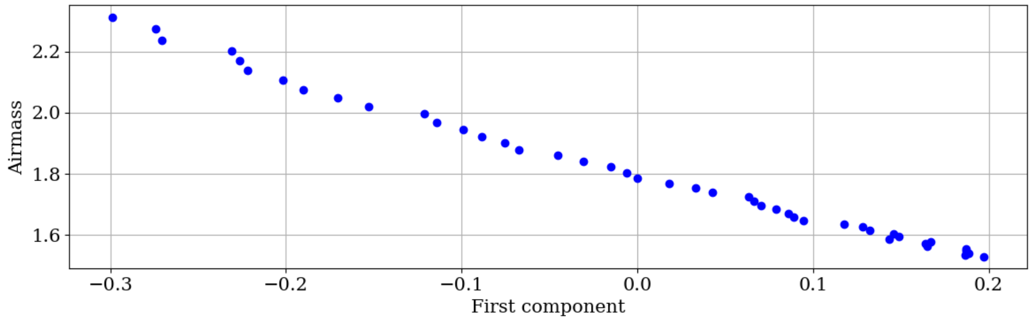

Figure 5.10: Linear relation between the first component in the time domain and the recorded airmass.

An useful information comes from the *explained variance ratio* (EVR) (see Sec. 3.2) that highlights the information in terms of percentage carried by every single component. In the case of the TDM (45 components) (Fig. 5.11, top panel) the variance of the first component includes most of the information and only the last one has zero variance. Also in the WDM case (1024 components) (Fig. 5.11, bottom panel) the first component has the majority of the variance, but then that drops to zero after a certain value which corresponds to the number of spectra in the dataset. Since the dimension (rank) of the input matrix is equal to the number of spectra, the number of uncorrelated vectors (i.e. the basis of the space) can not be greater than the dimension of the matrix itself. This implies that in both cases, the same number of eigenvectors is calculated, but these have different information content.

The first five components of the WDM case show some persistent features, reflecting the fact that the algorithm is not able to condense the information in a few components. On the contrary, when we use the TDM decomposition, a better correction is obtained. In Fig. 5.12 the results of the PCA algorithm after the subtraction of the first five components are shown for both TDM (left panels) and WDM (central panels) case (Eq. 3.8). The third column shows the difference of the respective WDM and TDM results, highlighting the persistence of telluric signal (the input matrix is also shown for reference).

A way to address this issue is to use a telluric mask after the calibration step to

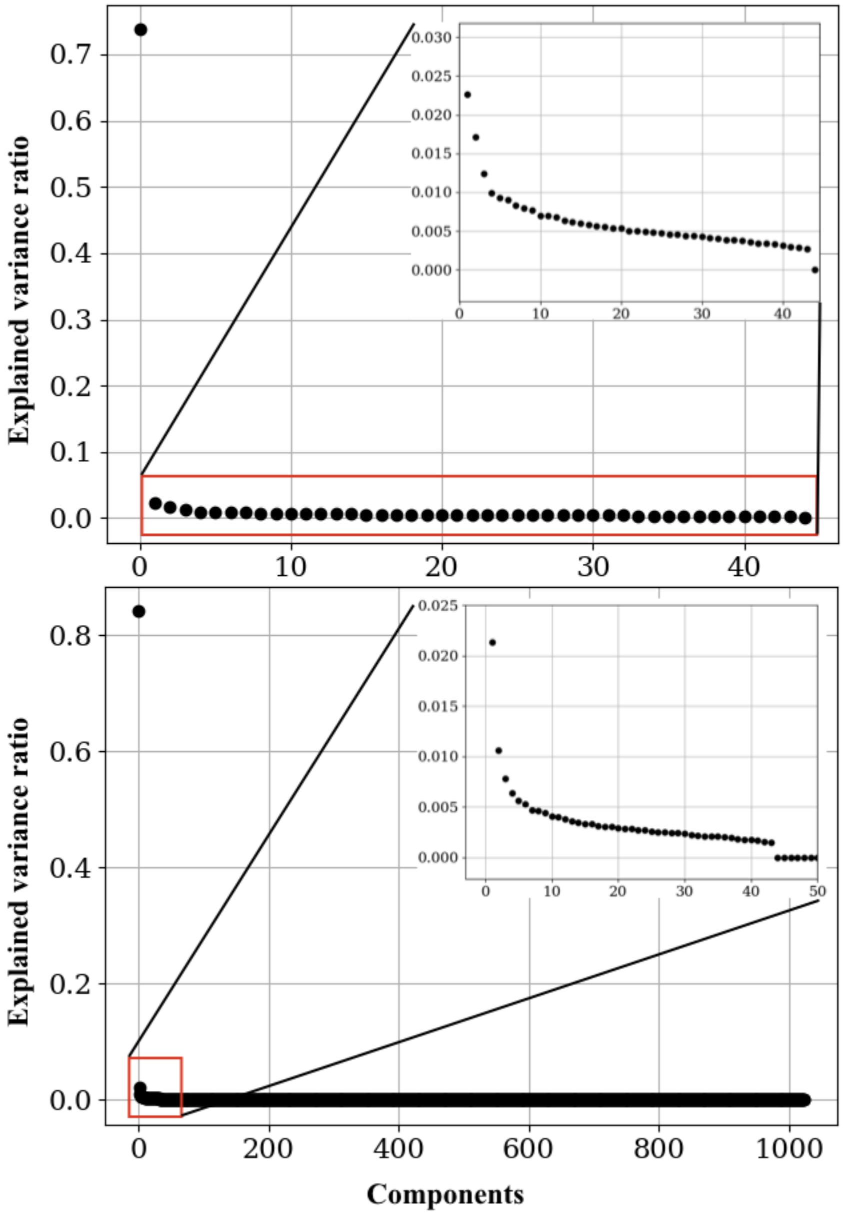

Figure 5.11: **Top panel** shows the explained variance ratio (eigenvalues) of the decomposition of the TDM covariance matrix. The graph is related to the first detector of the HD189733b dataset. **Bottom panel** is the same as top panel but in the WDM case. In both figures the first component has most of the total variance (<70%).

neglect most prominent telluric features.

The TDM case has been chosen as best method for the following reasons:

- the WDM component space is degenerate since there are more variables than

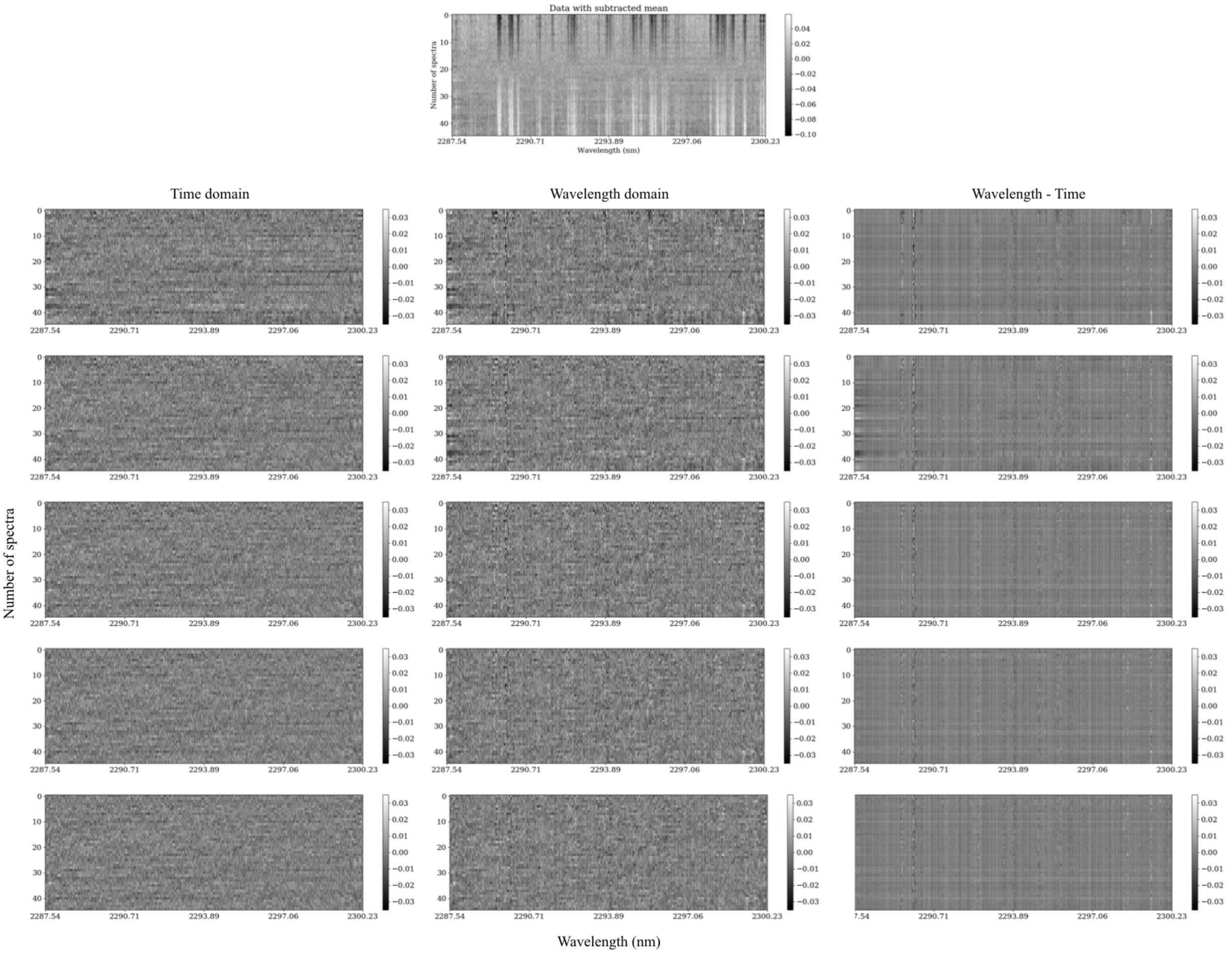

Figure 5.12: **Top figure:** the input matrix (PCA input) for reference. **Left column:** the PCA's results in the TDM case. **Central column:** the same as **left column** but in the WDM case. **Right column:** the difference between the wavelength and the time decomposition. Telluric residuals are present in the difference, meaning that the two methods are not equivalent.

observations;

- the application of a telluric mask is required if the WDM case is chosen.

5.4.2 Comparison of the algorithms

In Tamuz et al. (2005) it is stated that the orthogonality of the components retrieved by SYSREM is not assured and this may introduce systematics. The planetary signal is small compared to the overwhelming signal from star and the telluric absorption, for this reason choice of the right correction algorithm is crucial.

Fig. 5.13 shows the results of PCA (left panels) and SYSREM (right panels) after the subtraction of the first five components. At first look no significant differences can be spotted, however, if we calculate the difference between the matrices, structured signals become detectable (Fig. 5.14). The colour-map scale of the differences indicates that these effects are 1/10 of the signal. At this stage both algorithms are equally good.

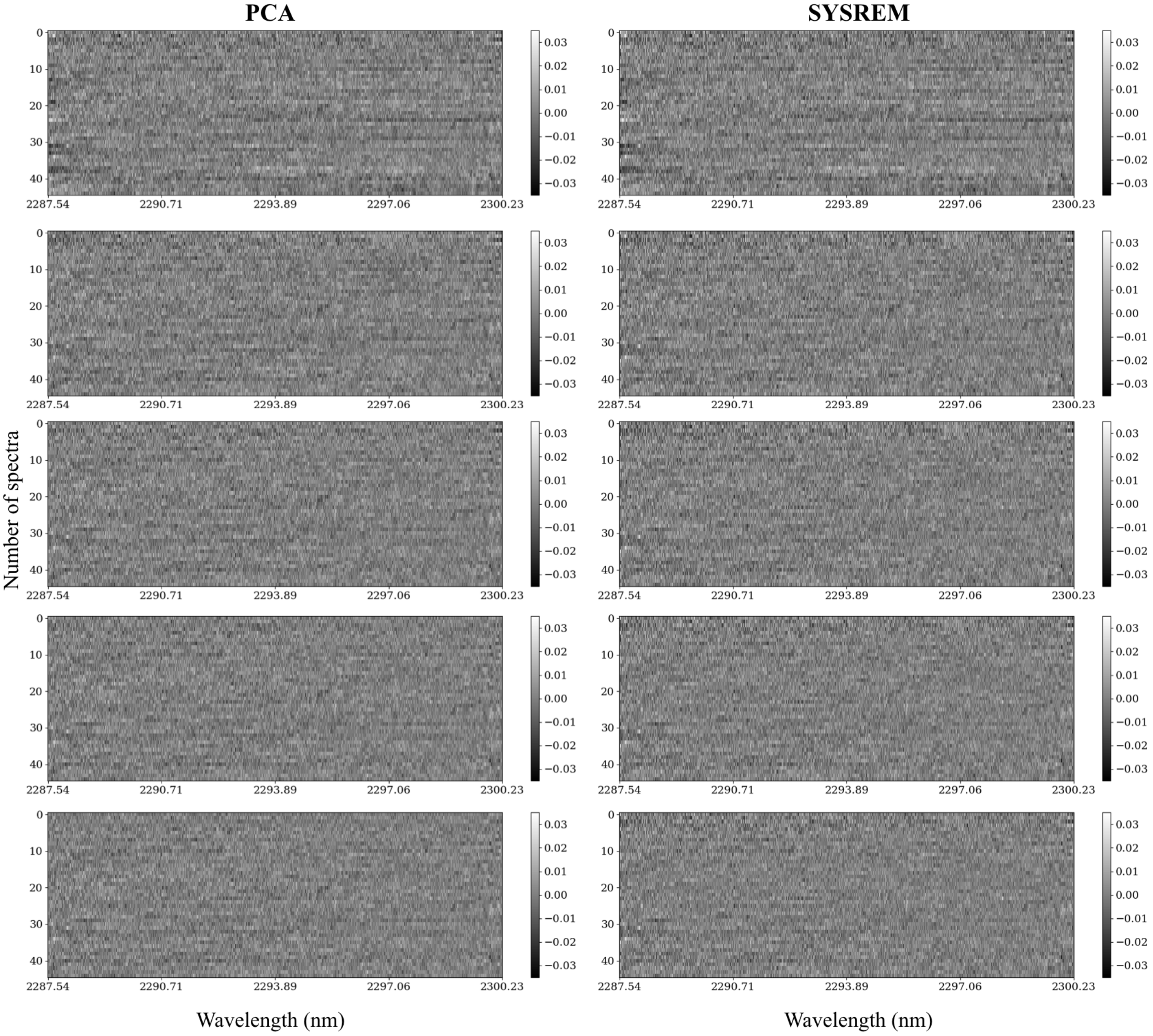

Figure 5.13: Results obtained with the decomposition algorithms after subtracting the first five components (from top to bottom). **Left column:** PCA algorithm, **right column:** SYSREM.

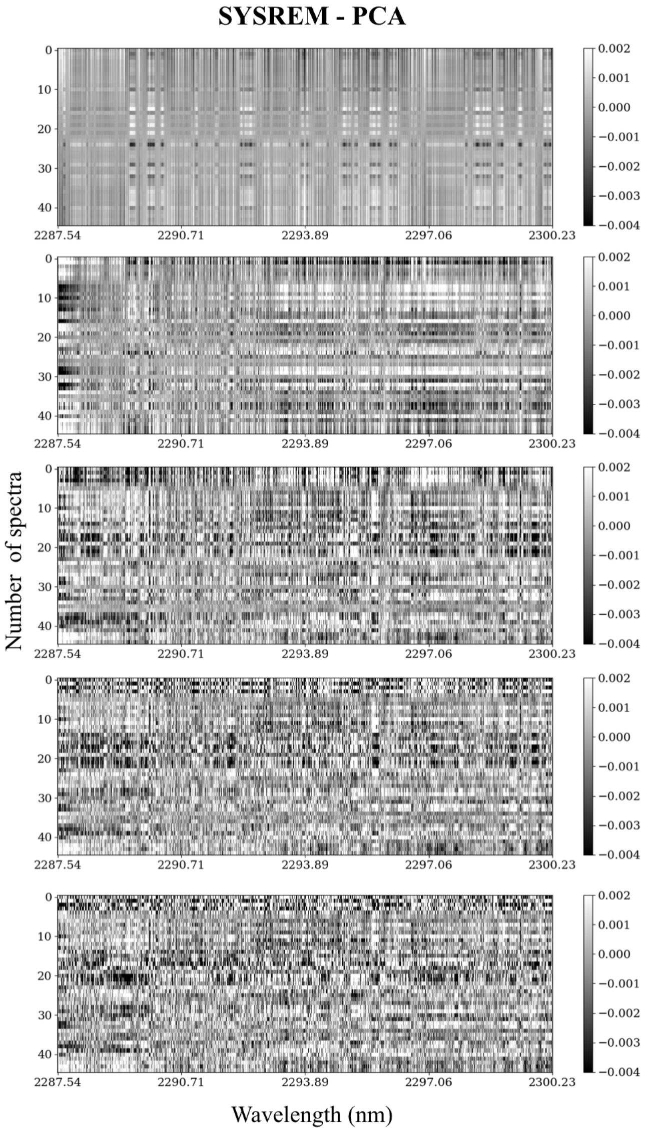

Figure 5.14: Differences between SYSREM and PCA residuals obtained by applying these two algorithms to the input data. Structured signals are visible and the differences are 1/10 of the signal showed in Fig. 5.13.

After the application of PCA or SYSREM each column of the output matrix has been divided by its standard deviation to restore the SNR of the processed data (de Kok et al., 2013; Birkby et al., 2013; Ridden-Harper et al., 2016; Nugroho et al., 2017).

5.5 Cross-Correlation Function (CCF)

The cross-correlation function measures the similarity of two signals. It is also often called *sliding dot product* since it returns a single value from the product of two signals in which one slides over the other. Considering two series $\boldsymbol{x}$ and $\boldsymbol{y}$, the normalised cross-correlation CCF at the delay d, for discrete series, is defined as follows (Bracewell, 1965)

$$CCF(d) = \frac{\sum_i \left((x(i) - \bar{x}) \cdot (y(i-d) - \bar{y})\right)}{\sqrt{\sum_i (x(i) - \bar{x})^2} \cdot \sqrt{\sum_i (y(i-d) - \bar{y})^2}} \tag{5.8}$$

where $\bar{x}$ is the mean of the array $\boldsymbol{x}$, $\bar{y}$ is the mean of the array $\boldsymbol{y}$ and $i = 0, 1, 2...N-1$. The idea of using such function is to find possible similarities between the data and an atmospheric model. The cross-correlation aims at matching similarities between the two signals.

The chemical models have been simulated using $\mathscr{T}$-REx (Sec. 2.3) (Waldmann et al., 2015a,b).

Every row of the data matrix (every single spectrum), after the application of PCA or SYSREM, is cross-correlated with the simulated atmospheric spectrum. This is interpolated to the same wavelength grid of the data, and it is then shifted from -100 to 100 $\mathrm{km\,s^{-1}}$ with 1.0 km/s as step. The step is chosen based on the precision obtained during the calibration step and on the velocity resolution of the instrument (see Sec. 5.3).

The CCF transforms the matrices (one for each of the four CRIRES' detectors) from the wavelength domain to the velocity domain. The CCF matrices are then added together to obtain one single matrix (we will refer to it as CCF matrix).

At this stage the planetary signal is not visible (see Fig. 5.15 top left panel). The injection of a synthesised model is performed after the calibration and the effects of this process are not visible at that stage because the signal intensity is at least three order of magnitude less than the telluric and star signals. When the CCF is performed, the injected model cross-correlates with itself resulting, for example, in the signal showed in Fig. 5.15 bottom panels. The model is injected using the orbital parameters of the planet. The stronger is the injected signal the higher will be the variance of those components that describe the signal. This means that injection should not be more than 1x times the signal not to interfere with the decomposition.

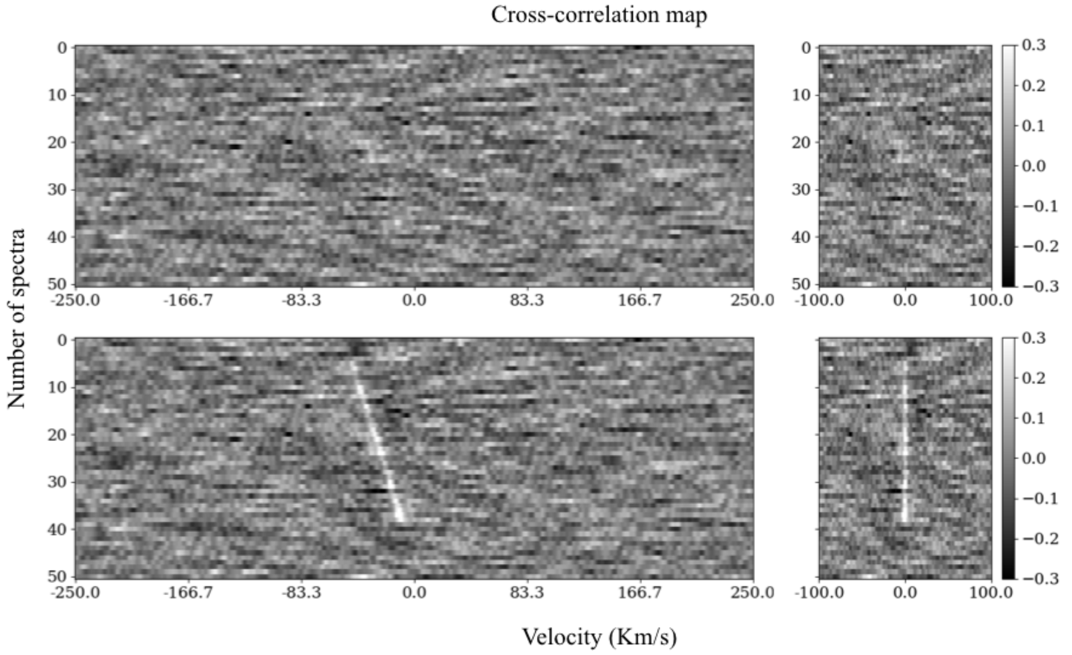

Figure 5.15: **Top left panel:** the four CCF of the four CRIRES' detectors summed together. **Bottom left panel:** same as top but with the model injected. The injection is 1× the synthesised model ($R_p/R_\star \sim 10^{-3}$). **Top right panel:** cross-correlation after changing the reference frame from the Earth to the rest frame of the exoplanet. In this frame the planetary cross-correlation signal is aligned to zero $\mathrm{km\,s^{-1}}$. **Bottom right panel:** same as top right panel but with the injection. The injected signal is aligned to zero $\mathrm{km\,s^{-1}}$ in the exoplanet's rest frame.

Since the signal of the planet is not visible, we summed the in-transit cross-correlations to increase the correlated signal. Before doing this, a change of reference frame is needed. All the previous steps have been performed working in the Earth's reference frame where the data have been aligned to the telluric spectrum. The reference frame is changed to the exoplanet's rest frame where the planetary signal is at rest. To perform this step the following transformation in the velocity space is applied:

$$V_p \;=\; K_p\, sin\left[2\pi\phi(t)\right] \;+ v_{sys} \;+ v_{bary}(t) \tag{5.9}$$

where V_p is the velocity correction, K_p is the radial velocity of the planet (defined in Eq. 5.3), $\phi(t)$ is the orbital phase (see Eq. 5.1), v_{sys} is the relative velocity between the Sun system and the target system and $v_{bary}(t)$ is the correction for changing from the Earth reference frame to the barycentric frame of the solar system. Fig. 5.15 top and bottom right panels show the cross-correlation map after the reference frame is changed. Here, the injected signal is at rest. Since the planetary signal is now aligned to its natural reference frame we summed only the in-transit cross-correlations to increase the planetary SNR.

5.6 Output

When all the in-transit cross-correlations are summed together, the 2D cross-correlation matrix is reduced to 1D signal (see Fig. 5.16) relative to the theoretical orbital velocity of the planet. To explore different orbital velocities we proceeded as follows:

1. we let K_p varying from 0 to 250 km/s with 1 km/s step;
2. for each K_p we apply the correction (Eq. 5.9) to every single CCF in the CCF matrix;
3. we sum only the in-transit cross-correlations.

In this way we were able to explore all possible orbital velocities including those corresponding to the host star. Following the previous steps, we obtained a matrix with K_p on y-axis and velocity rest frame along x-axis (v_{rest}). From this matrix two different outputs are extracted: the SNR map and the T-test statistic.

5.6.1 SNR map

We refer to the last matrix obtained, i.e. K_p on y-axis and v_{rest} on x-axis. We calculated the standard deviation of this matrix excluding those points potentially connected to the planetary signal ($|v_{rest}| < 15 \text{ km s}^{-1}$) and we divided the entire matrix for this value. In this way we derived the SNR matrix (see Fig. 5.17).

To assign an uncertainty to the K_p value we followed the same procedure as reported in Brogi et al. (2016): i.e. we took the maximum of the matrix and, fix-

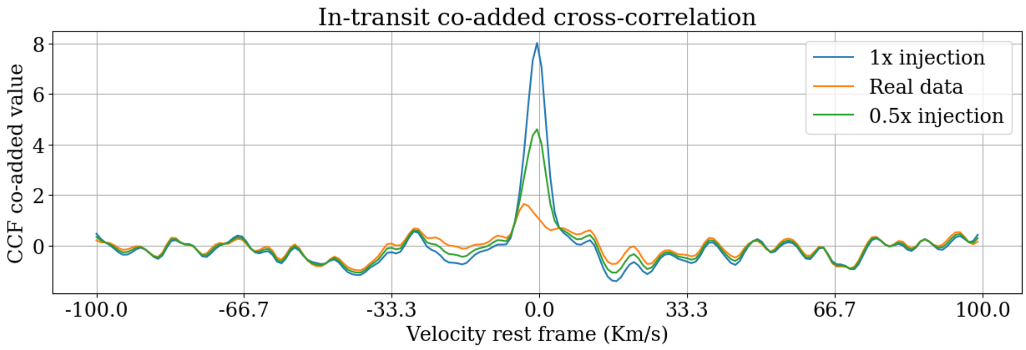

Figure 5.16: In-transit summed cross-correlations for the input data and for different injections. The injection perturbs the surroundings of the CCF peak,the stronger the injection the higher is the perturbation around the peak.

ing the relative v_{rest}, we calculated the K_p interval where the SNR drops by a unit around the K_p peak (Fig. 5.18). The same approach has been used to determine the uncertainty for the v_{rest}.

The SNR map is not only useful to visually represent the result but also to see if spurious signals or telluric residuals are present. These signals may have high SNR value but located at K_p and/or v_{rest} different from those expected for the planetary signal.

Finally, Fig. 5.17 and Fig. 5.18 contain an extra information: when PCA and SYSREM have been introduced we have not mentioned how to decide the number of components to neglect. We performed a computational loop which comprises PCA, CCF and SNR map calculation to explore the component space; this loop is sketched in Fig. 5.1. If SYSREM is chosen as de-trending method, there is only one loop subtracting one component at a time iteratively. If PCA is chosen, two loops are calculated: one neglects higher variance components onwards and the other subtracts lower variance components backwards.

5.6.2 Whelch' T-test statistics

The Welch's T-test is a method used in statistics to test the hypothesis that two populations have equal means (Welch, 1947). Welch's T-test, differently from Student's T-test, is more reliable when the two samples have unequal variances and unequal

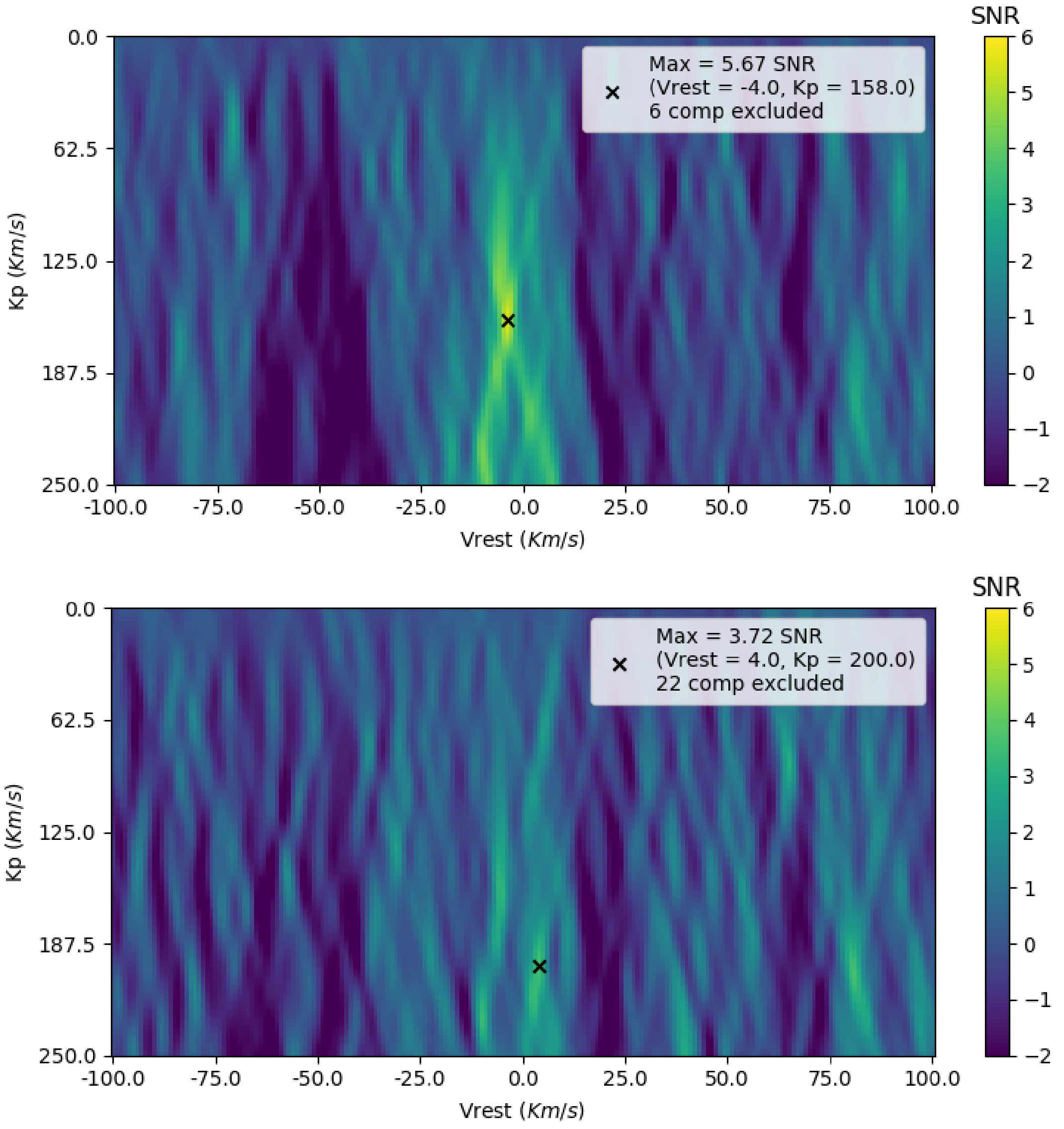

Figure 5.17: **Top Panel:** SNR map in case of planetary signal detection (the maximum SNR is in agreement with the orbital K_p and the planetary v_{rest}) (Full discussion Sec. 5.7). **Bottom panel:** same map but the planetary signal is not detected.

sample sizes. Moreover, the Welch's T-test assumes that the two populations have normal distributions. The idea to use this method is to compare the population of points on the CCF map interested by the planetary signal with those that are not.

From the output CCF matrix (Fig. 5.19 and Sec. 5.6), we defined, as done in the literature (Brogi et al., 2016; Nugroho et al., 2017):

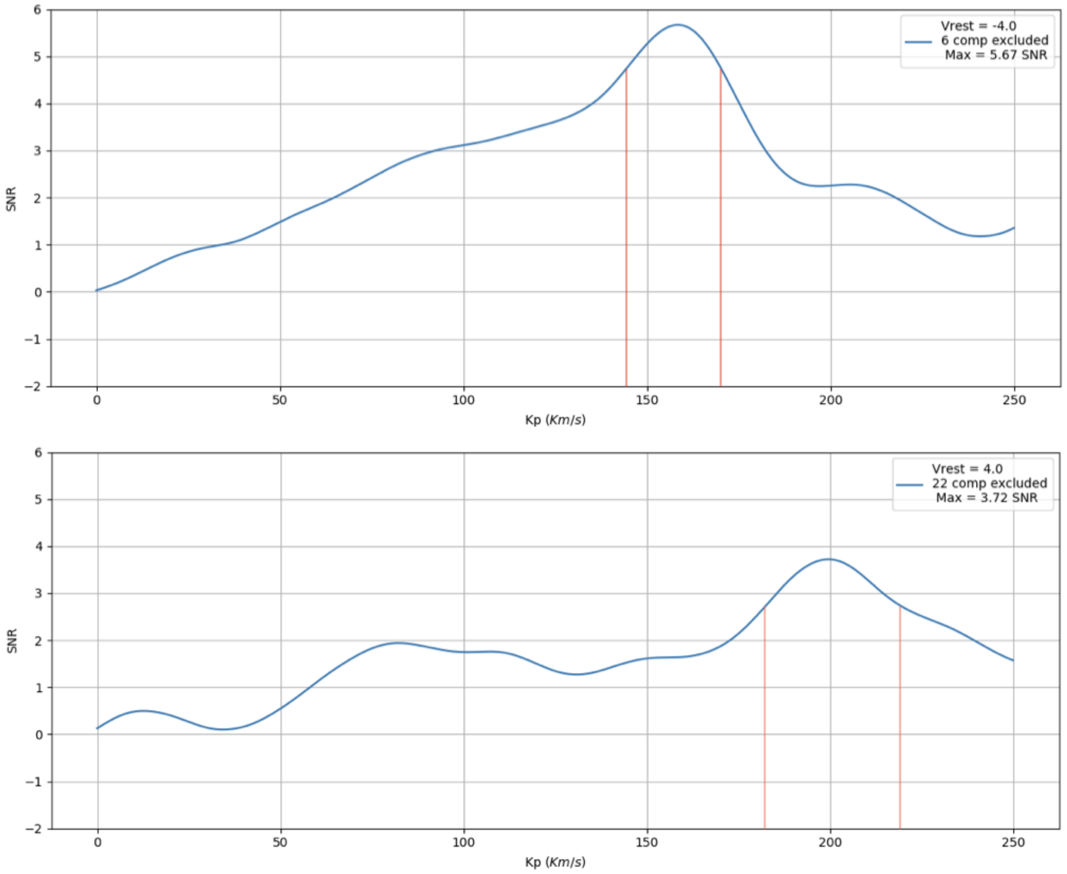

Figure 5.18: **Top panel:** SNR variation with respect to K_p at fixed v_{rest} value of the SNR map in Fig. 5.17 top panel. The red vertical lines indicate the uncertainty interval. **Bottom panel:** same as **top panel** but referred to Fig. 5.17 bottom panel.

- *in-trail*, those values inside a squared box centred on the CCF' peak with a radius of $\pm 15 \mathrm{km\,s^{-1}}$;

- *out-trail*, those values outside the *in-trail* box

We extracted two families of values from the output CCF matrix and these are compared through Welch's T-test (Fig. 5.20). From the figure we can appreciate how the in-trail distribution is shifted with respect to the out-trail distribution, centred on zero-mean. The Welch's T-test (calculated using *scipy.stats.ttest_ind* in *python*) provides a *p-value* (two-tailed) which is converted into *σ-value* (significance interval) through the inversion of the *survival function* (SF)

$$\sigma_{value} = SF^{-1}(p\text{-}value \ / \ 2) \tag{5.10}$$

where the SF^{-1} is the inverse of the survival function that is calculated from the *cumulative density function* (CDF) as follows:

$$SF = 1 - CDF. \tag{5.11}$$

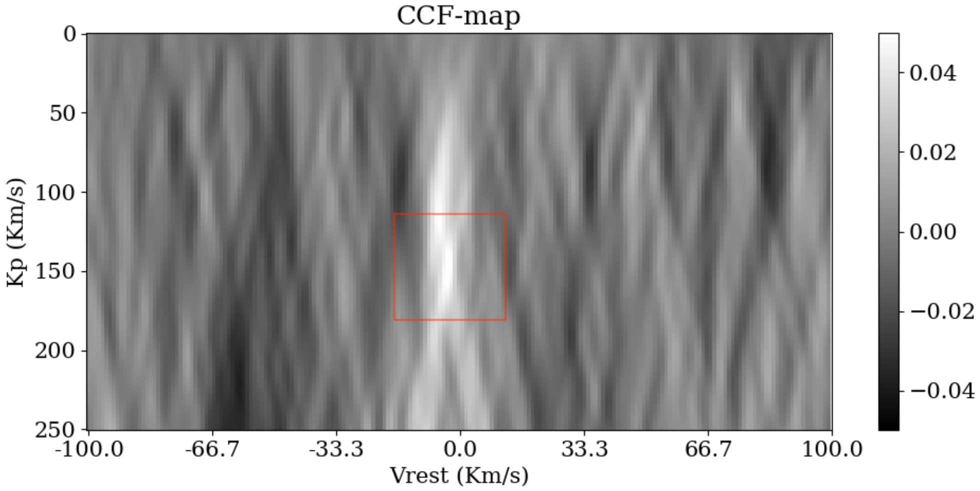

Figure 5.19: Co-added in-transit cross-correlations at different K_p. The red square determine which values are $in-trail$, i.e. those inside the box, and values that are $out-trail$, i.e. outside the box.

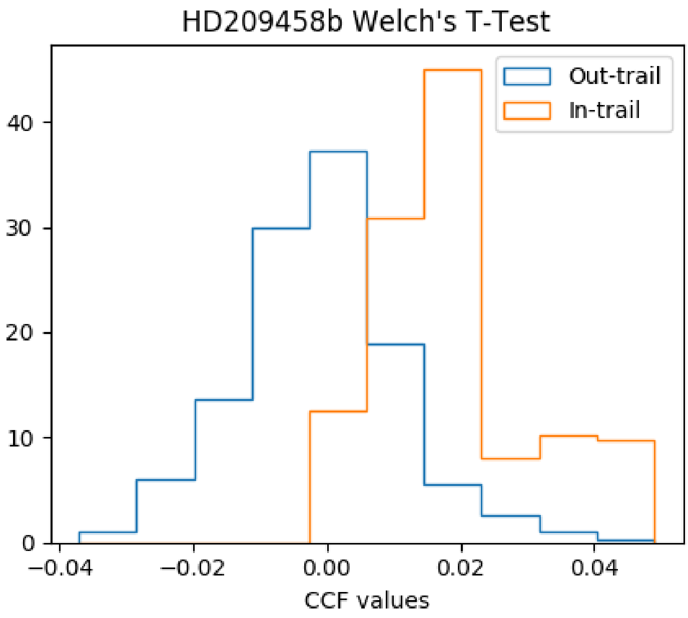

Figure 5.20: Distribution of the statistical Welch's T-test performed on HD209458b dataset. In-trail (orange) and out-trail (blue) distributions are drawn and in both cases the in-trail distribution is shifted towards positive CCF values.

5.7 VLT/CRIRES datasets

The CRyogenic high-resolution InfraRed Echelle Spectrograph (CRIRES) at ESO's Very Large Telescope (VLT) was proposed a few years after the first radial velocity detection (Wiedemann, 1996; Wiedemann et al., 2000; Kaeufl et al., 2004) and for the first time a robust detection of CO in the infrared wavelength range has been obtained by Snellen et al. (2010). They observed HD209458b, an hot-Jupiter orbiting around a G0-type star ($\sim$ 6000K), and they were able to estimate the orbital velocity of the planet, to detect carbon monoxide in the atmosphere, and finally to obtain some insights into the dynamic of the atmosphere. The success of this work is, in part, due to the high stability not only of the instrument but also of the telescope that used Multi-Applications Curvature Adaptive Optics (MACAO) to optimise the signal-to-noise ratio and the spatial resolution.

The VLT high-resolution spectrograph, CRIRES, provided a resolving power up to 100′000 (two pixels) in the wavelength range from 1 to 5.3 μm. It was provided with a mosaic of four detectors (4096×512 pixels) in the focal plane.

5.7.1 Data reduction

The CRIRES reduction pipeline has been embedded into my code thanks to ESO's *EsoRex* which is a command-line driven utility that can launch pipeline reduction routines (they are referred as recipes). These are individual scripts that perform specific actions to the input data. The recipe used to reduce the CRIRES raw data is the *crires_spec_jitter*[5]. This recipe performs the following operations:

- dark subtraction;
- correction for detector non-linearity;
- flat-fielding;
- combination of nodding exposures;
- spectrum extraction;
- wavelength calibration.

The master reduction files (e.g. dark and flat) are provided with the raw data, while the specific non-linearity correction files need to be downloaded from the archive[6].

The nodding is an observational strategy that allows to record subsequent images alternating two different positions on the detector, these are generally referred as *nod A* and *nod B* (ABBA or ABAB) (see Fig. 5.21). The strategy consists on subtracting images relative to two subsequent positions, creating AB or BA couples. This allows to correct the sky background, possible bad pixels and glowing effects. The two signals (relative to *nod A* and *nod B*) are separately extracted or combined together depending on the data analysis strategy (combining images reduces the time resolution). In the CRIRES' datasets the two signals are merged together (Fig. 5.21) and then optimally extracted (Horne, 1986).

[5] https://www.eso.org/observing/dfo/quality/CRIRES/pipeline/recipe_science.html

[6] https://www.eso.org/sci/facilities/paranal/instruments/crires/doc/VLT-MAN-ESO-14200-4032_v91.pdf

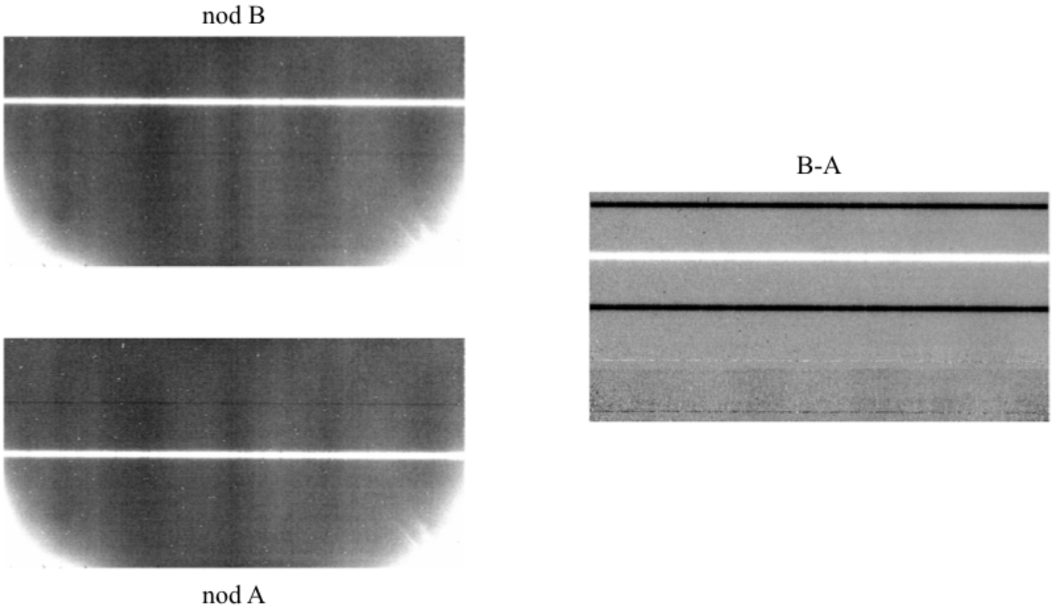

Figure 5.21: CRIRES' reduction process. **Top left panel**: raw frame in the nodding position B. **Bottom left panel**: raw frame in the nodding position A. **Right panel**: combination of the two nodding position.

The extracted spectra are calibrated with the arc frame provided alongside the master reduction files aforementioned. As explained in Sec. 5.3 this calibration process is not enough. An additional calibration with the telluric spectrum is required for a better correction of the telluric absorption effects.

In the following two sections I will describe the application of the pipeline previously presented on the HD209458b and HD189733b datasets.

5.7.2 HD209458b

We downloaded the HD209458b dataset which is publicly available on the ESO archive. It is part of the 383.C-0045(A) program (PI: Snellen, I.). HD209458b has been observed with CRIRES at the highest resolution available (R$=100,000$) through the $0''.2$ slit. The dataset covers a narrow wavelength range, i.e. $2291.79-2349.25$ nm with three gaps in between due to the physical separation of CRIRES detectors (283 pixels). The dataset has been recorded with the nodding method ABBA (Snellen et al., 2010), and *EsoRex* has been used to reduce the data. The number of spectra extracted by the reduction process (Sec. 5.7.1) is 51.

The data are then corrected using the Earth's absorption spectrum (telluric spectrum) and cosmic rays are removed (Fig. 5.22). The data are decomposed by PCA or SYSREM. The variances of the eigendecomposition for each of the four detectors are shown in Fig. 5.23. The first component contains most of the information (>75%) but its value varies for different detectors suggesting that choosing the same number of components for all the detectors may not be the optimal solution.

For the cross-correlation process, the planetary transmission spectrum was modelled using $\mathscr{T}$-REx (Waldmann et al., 2015a,b) (Sec. 2.3). The CO and H_2O line-lists were taken from ExoMol (Tennyson & Yurchenko, 2012; Tennyson et al., 2016). We assumed an isothermal T/p profiles at $T = 1400$K, with pressure varying from 10^{-5} to 10^{4} Pa. We did not include clouds or lines broadening due to the rotation of the planet. We used 10^{-3} as Volume Mixing Ratio (VMR), this value is compatible with chemical models' predictions for Hot-Jupiters atmospheres (Venot et al., 2012). The same value was also used by Snellen et al. (2010).

The cross-correlations are aligned to the planetary rest frame (Eq. 5.9) and thanks to the duration of the transit calculated with Eq. 5.6, only those CCFs between 9 and 40 (51 in total) are summed, i.e. in-transit CCFs. We explored the components' space for CO and H_2O and for both of them a signal has been found.

The signal obtained for CO, using PCA, peaks at SNR=5.7 (Fig. 5.25). The signal is compatible with the planetary orbital parameters ($K_p = 148^{+16}_{-15}$ km s^{-1}, $v_{rest} = -3.0^{+1.3}_{-1.1}$ km s^{-1}). This result has been obtained by considering components from the $7th$ to the $28th$. The in-transit co-added cross-correlation is shown in Fig. 5.24 in case of injection or not. Since the injection effect, in this graph, is not null, the PCA did not erase the planetary signal. The result is also confirmed by the Whelch's T-Test (Fig. 5.26). Using a box of radius 15 km s^{-1} (Sec. 5.6.2) the null hypothesis is rejected with a confidence greater than 7σ, the shift of the *in-trail* population is noticeable with respect to the *out-trail* values that are, instead, distributed as a Gaussian centred to zero.

If we use SYSREM, the SNR peak is not strong as in the PCA case (see Fig. 5.27). The SNR peak is compatible with the orbital parameters but its value is

Stellar parameters	
T_{eff} (K)	6065 ± 50[1]
$M_{\star}$ ($M_{\odot}$)	1.119 ± 0.033[1]
$R_{\star}$ ($R_{\odot}$)	$(1.155^{+0.014}_{-0.016})$[1]
$log(g_{\star})$ (csg)	4.361 ± 0.008[1]
v_{sys} (km s^{-1})	-14.7652 ± 0.0016[2]
Planetary parameters	
T_{eq} (K)	1449 ± 12[1]
a (AU)	$(0.04707^{+0.00046}_{-0.00047})$[1]
R_p (R_{Jup})	$(1.359^{+0.016}_{-0.019})$[1]
M_p (M_{Jup})	0.685 ± 0.015[1]
P (days)	3.52474859(38)[3]
T_0 (BJD_{UTC})	2452826.629283(87)[3]
i (deg)	86.71 ± 0.05[1]

[1]Torres et al. (2008), [2]Mazeh et al. (2000), [3]Knutson et al. (2007a)

Table 5.1: HD209458 system information

SNR$= 4.03$. As described in Sec. 5.6.1 the components in SYSREM are subtracted one at a time and the non orthogonality of the SYSREM's components may have erased part of the planetary signal during the process.

The signal of the water vapour is more difficult to detect since the Earth's atmosphere also contains water. To determine the planetary signal a good telluric correction is required, to do so, several components have been subtracted using PCA. A signal at the compatible planetary parameters is observable in the SNR map in Fig. 5.29. The maximum peaks at SNR=3.95, $K_p = 140^{+25}_{-16}$ km s^{-1} and $v_{rest} = -4.0^{+1.4}_{-1.6}$ km s^{-1} and it is obtained considering components from the $33th$ to the $43th$. To demonstrate that the H_2O planetary signal survives after 33 components have been subtracted, Fig. 5.28 shows the in-transit co-added cross-correlation relative to the range of components aforementioned. Both the injected and non-injected signal survive to the PCA correction (note that the injected signal does not include any atmospheric dynamics, so it is not blue-shifted as the planetary signal). Moreover, the co-added cross-correlation value is lower with respect to the CO case (Fig. 5.28 and 5.24) meaning that the concentration of water is lower than CO or that PCA has erased part of the signal. Finally, the Whelch's T-Test is performed on the *in-trail* and *out-trail* populations (Fig. 5.30). In this case the shift of the *in-trail* population is not as strong as in the CO case but the null hypothesis is rejected with a confidence greater than 6σ.

When we used SYSREM, no signal was found for the water vapour.

Another test we performed was to cross-correlate the telluric model used in the calibration process (Sec. 5.3) with the data, to check if any telluric signal still persists. In Fig. 5.31 we show two SNR map of the dataset cross-correlated with the telluric model: the top panel considers the entire component space. This result not only highlights the correction of the telluric signal but also that the water vapour has a different spectral signature at different temperature (300 K on the Earth and 1400 K on HD209458b).

Tab. 5.2 summarises the results obtained for the analysis of the HD209458b dataset. The results of both molecules are in agreement. The negative v_{rest} (i.e. blueshift of the signal in the SNR map) is compatible with high altitude winds. Snellen et al. (2010) reported a blueshift of $\sim$2 km s^{-1} explaining that it is compatible with the presence of high altitude winds.

Results (this work)	CO	H_2O
SNR	5.7	3.95
K_p (km s^{-1})	148^{+16}_{-15}	140^{+25}_{-16}
v_{rest} (km s^{-1})	$-3.0^{+1.3}_{-1.1}$	$-4.0^{+1.4}_{-1.6}$
W T-Test	21.62σ	6.56σ

Table 5.2: The table shows the results obtained for the HD209458b dataset.

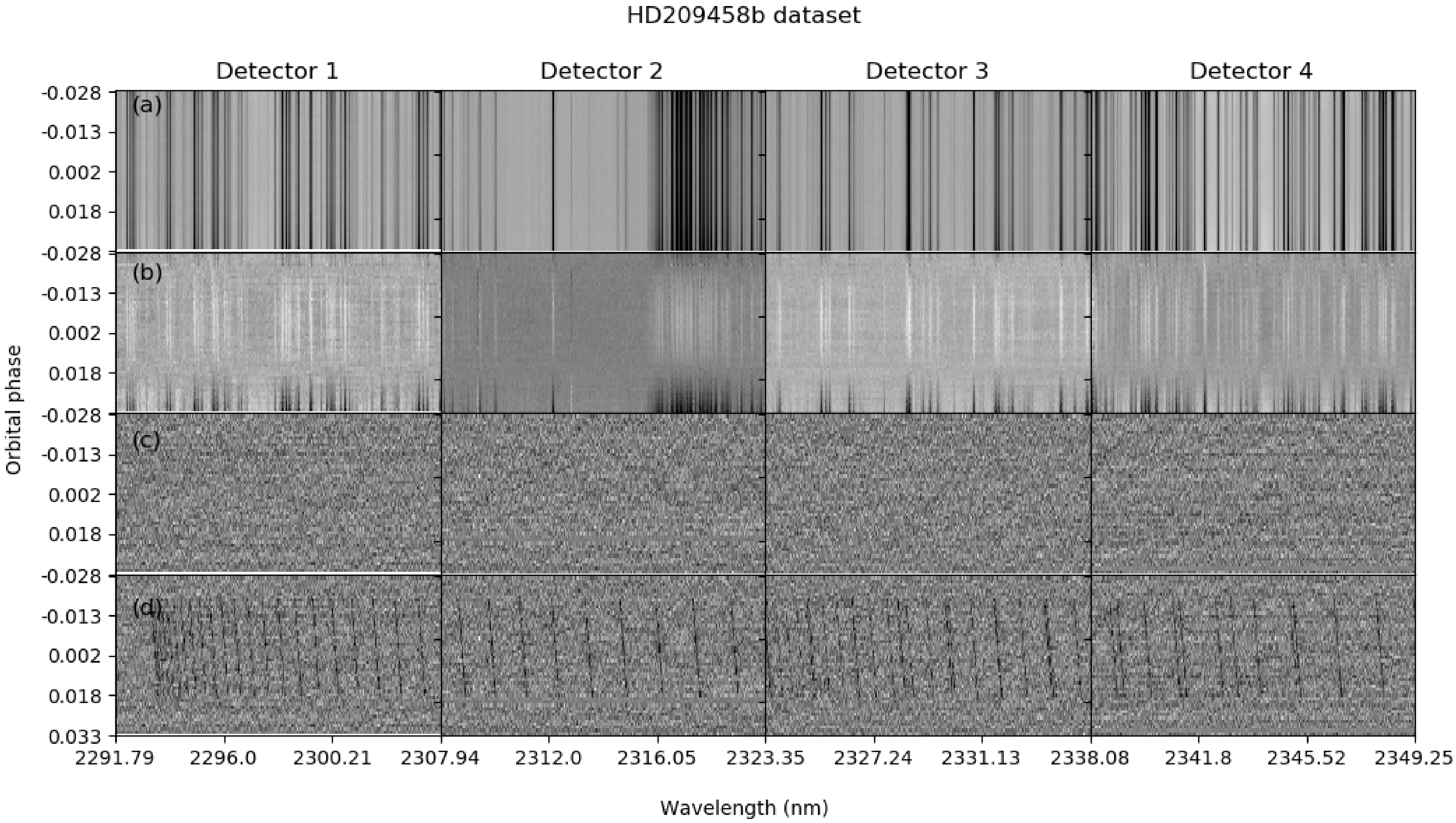

Figure 5.22: HD209458b dataset. In **(a)**, the data are shown after calibration, normalisation and spikes correction. In **(b)**, the data are shown after the MFS method has been applied (Sec. 5.4). In **(c)**, the results of PCA are shown. In **(d)**, the data are shown after the application of PCA and the injection of the CO model.

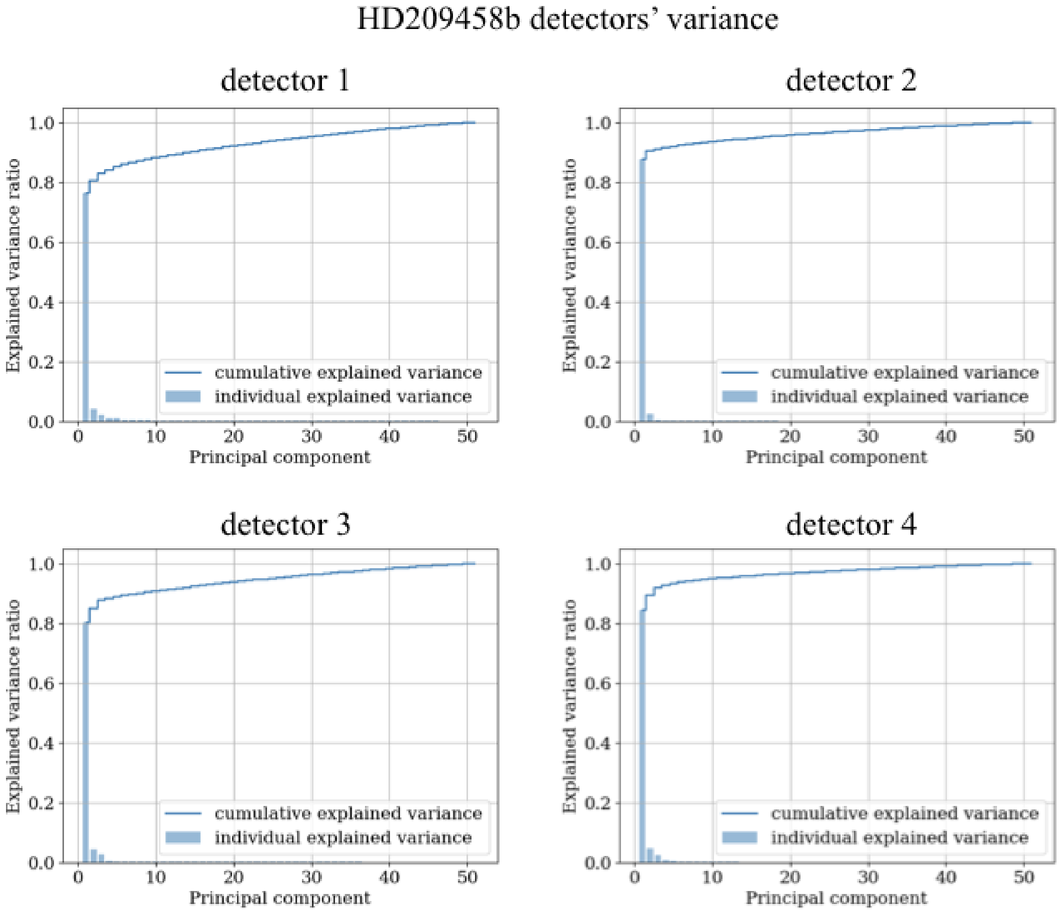

Figure 5.23: Detectors' variances of the PCA decomposition relative to the HD209458b dataset. The first component always carries more than 75% of the information. However, the variance is different for each of the detectors.

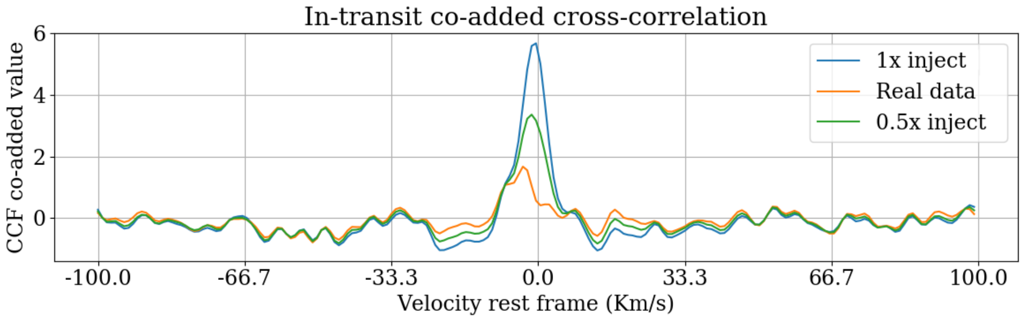

Figure 5.24: Summed in-transit cross-correlations with and without the injection of carbon monoxide. The injection perturbs the surroundings of the CCF peak. The co-added CCFs are relative to the planetary rest frame of HD209458b ($K_p = 145.041\ \mathrm{km\,s^{-1}}$). This graph has been generated considering components from 7 to 28 of the PCA decomposition (the same of Fig. 5.25).

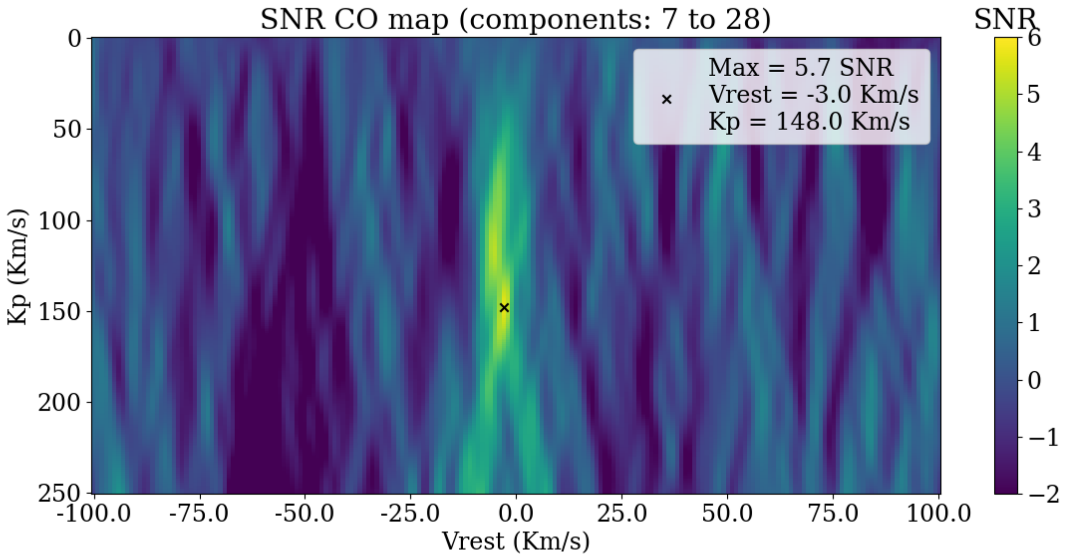

Figure 5.25: SNR map for the carbon monoxide. The maximum point is compatible with the planetary orbital parameters.

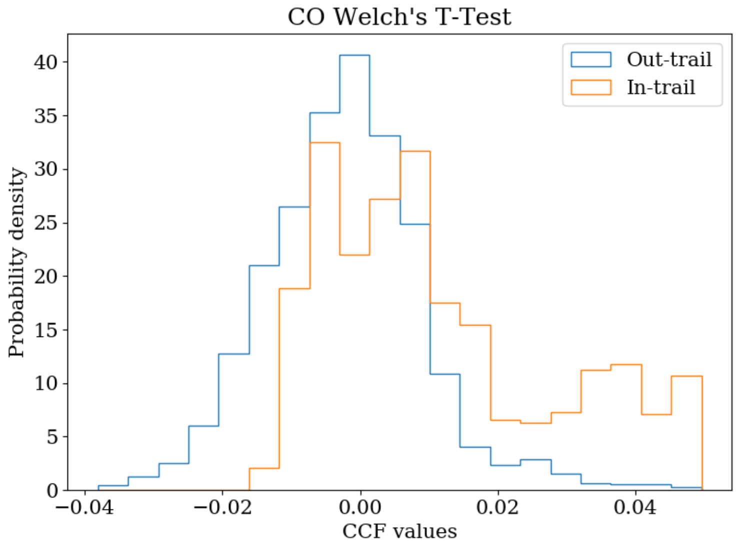

Figure 5.26: Distributions (i.e. *in-trail* and *out-trail*) used to compute the Whelch's T-Test. The null hypothesis is rejected with a confidence >7σ.

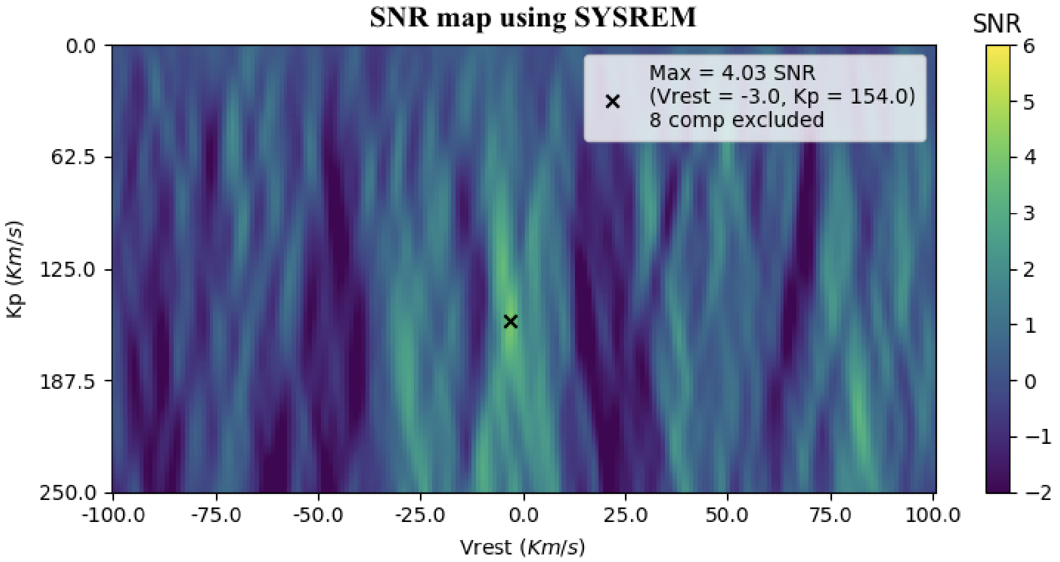

Figure 5.27: SNR map for the carbon monoxide when SYSREM is used. The maximum point is compatible with the planetary parameters but the peak value is lower than the one obtained using PCA.

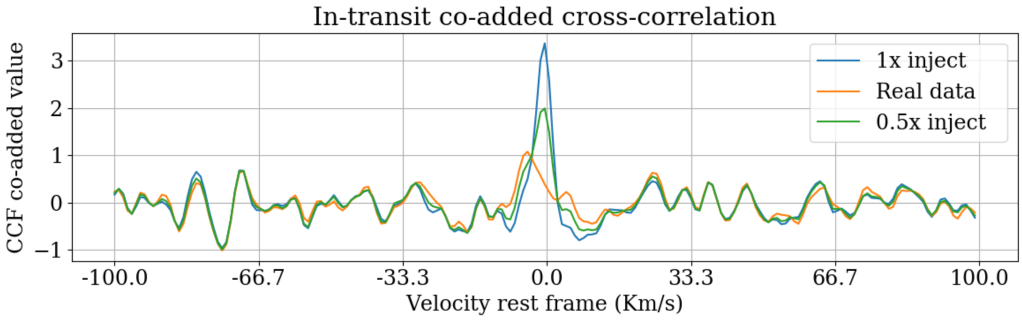

Figure 5.28: Cross-correlations of water vapour co-added in-transit. The injected signal and the planetary signal are still present after using PCA. The co-added CCFs are relative to HD209458b rest frame ($K_p = 145.041\ \mathrm{km\,s^{-1}}$). This graph has been generated considering PCA components from 33 to 43.

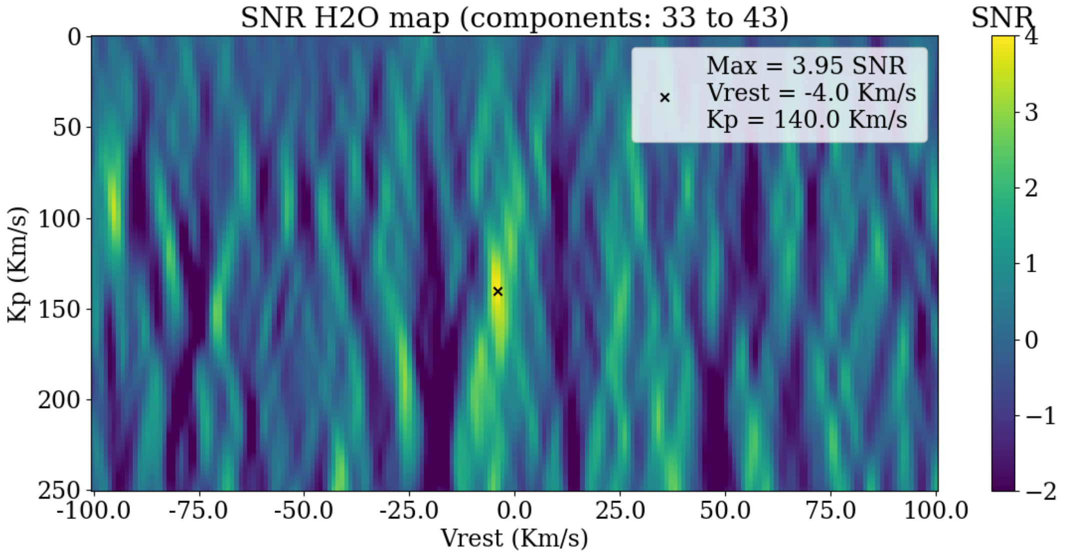

Figure 5.29: SNR map of the water vapour. The peak is compatible with the planetary parameters.

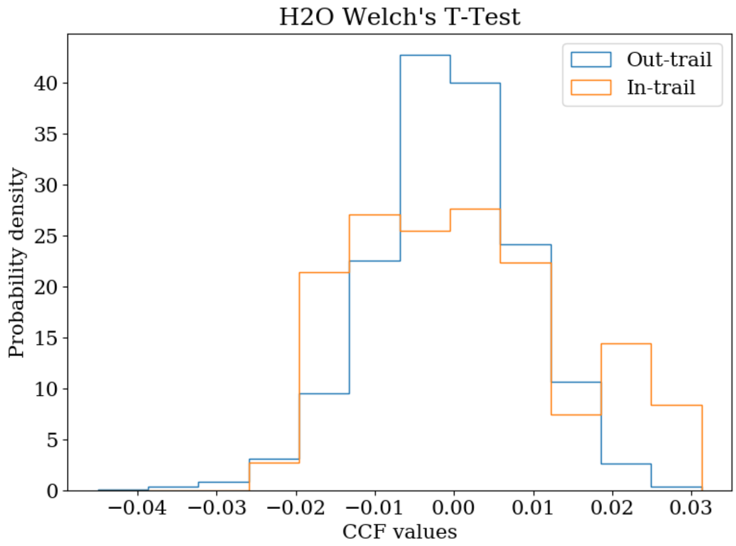

Figure 5.30: Distribution (i.e. *in-trail* and *out-trail*) used to compute the Whelch's T-Test. The null hypothesis is rejected with a confidence of 6.56σ.

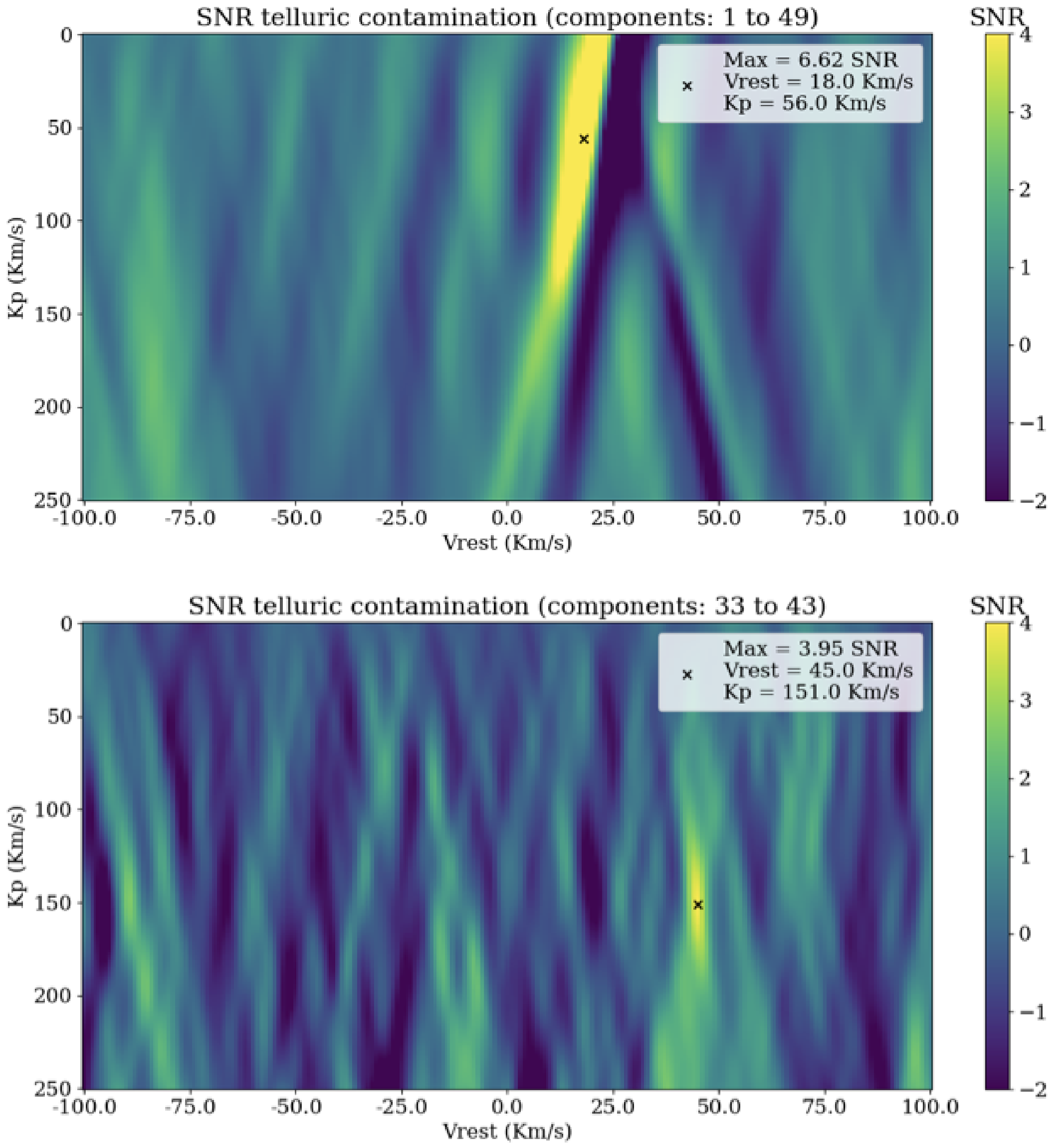

Figure 5.31: **Top panel**: SNR map of the dataset HD209458b cross-correlated with the telluric model. **Bottom panel**: SNR map of the telluric model including the same range of components of the reported water vapour detection (see Fig. 5.29). No signal is visible at the orbital parameters of the planet.

5.7.3 HD189733b

The HD189733b dataset is publicly available on the ESO archive. It is part of the 289.C-5030(A) program (PI: Snellen, I.). HD189733b has been observed with CRIRES at the highest resolution available (R= $100,000$) through the $0''.2$ slit. The datasets, cover a narrow wavelength range, i.e. 2287.54 – 2345.34nm. The dataset has been recorded with the nodding method ABBA (Brogi et al., 2016), and *EsoRex* has been used to reduce the data. The final number of spectra extracted after using the reduction process (Sec. 5.7.1) are 45. The same analysis done for the HD209458b dataset has been performed here.

For the cross-correlation process the planetary transmission spectrum was modelled using $\mathscr{T}$-REx (Waldmann et al., 2015a,b) (Sec. 2.3). The CO and H_2O line lists were adopted from ExoMol (Tennyson & Yurchenko, 2012; Tennyson et al., 2016). We assumed isothermal T/p profiles at $T = 1000$K, pressure varying from 10^{-5} to 10^4 Pa and we did not include clouds or any lines broadening due to the rotation of the planet. We used 10^{-3} as Volume Mixing Ratio (VMR), this value is compatible with chemical models' predictions for Hot-Jupiters (Venot et al., 2012). The same value was also used by Brogi et al. (2016) in their analysis.

All the cross-correlations are aligned to the planetary rest frame (Eq. 5.9) and thanks to the duration of the transit calculated through Eq. 5.6, only CCFs from 7 to 45 (45 in total) are summed up.

The CO detection is highly difficult since the star, being a K-type star (T$\sim$ 4900 K), contains CO in the outer regions. In Brogi et al. (2016) a master stellar spectrum has been simulated and subtracted to the data, but the stellar contamination continued to be persistent also in the result. In this work, PCA was not as effective as in the HD209458b case because the star spectrum moves 1-2 pixels on the detector preventing an optimal correction. The result (see Fig. 5.35 and Tab. 5.4) is compatible with the one claimed by Brogi et al. (2016) (SNR=5.1, $K_p = 194^{+19}_{-41}$ km s^{-1}, $v_{rest} = -1.7^{1.1}_{1.2}$ km s^{-1}), however the error on the K_p, being smaller than the one reported in literature, does not include the theoretical value of the orbital velocity of the planet calculated with Eq. 5.4. The signal determined at

Stellar parameters	
T_{eff} (K)	5040 ± 50[1]
$M_\star$ ($M_\odot$)	0.806 ± 0.048[1]
$R_\star$ ($R_\odot$)	0.756 ± 0.018[1]
$log(g_\star)$ (cgs)	4.587 ± 0.015[1]
v_{sys} (km s^{-1})	-2.361 ± 0.003[2]
Planetary parameters	
T_{eq} (K)	(1201^{+13}_{-12})[1]
a (AU)	(0.03120(27))[3]
R_p (R_{Jup})	$(1.178^{+0.016}_{-0.023})$[3]
M_p (M_{Jup})	$(1.144^{+0.057}_{-0.056})$[1]
P (days)	2.21857567(15)[4]
T_0 (BJD$_{UTC}$)	2454279.436714(15)[4]
I (deg)	85.710 ± 0.024[4]

[1]Torres et al. (2008), [2]Bouchy et al. (2005), [3]Triaud et al. (2009)
[4]Agol et al. (2010)

Table 5.3: HD189733 system information.

lower K_p ($\sim$ 85 km s^{-1}) is a contamination signal of the star, that results from the Rossiter-McLaughlin effect combined with the change of reference frame from the Earth to the barycentric one (Brogi et al., 2016).

The data in-transit co-added and cross-correlated at the theoretical velocity of the planet (Fig. 5.34) show a peak at the position of the injected signal, but that is not strong enough to emerge from the noise.

Concerning water vapour, the same discussion done for HD209458b can be applied here. The planetary water signal needs to be disentangled from the telluric absorption. The result obtained is compatible with both literature and theoretical parameters, e.g. see Fig. 5.36 where the planetary signal is compared with the injected one. The injected signals do not account for $v_{rest} \neq 0$ km s^{-1}, we can appreciate the data being blue-shifted. Finally the Whelch's T-Test confirms that the null hypothesis can be rejected with a confidence greater than 5σ (Fig. 5.39).

As in the case of HD209458b, we have cross-correlated the data with the telluric model. Fig. 5.40 shows the SNR map of the telluric model in the component space determined for the water vapour detection (Fig. 5.37). Even in this case the telluric signal is not present at the theoretical position of the planet suggesting the effectiveness of the correction.

Results (this work)	CO	H_2O
SNR	5.24	3.69
K_p ($\mathrm{km\,s^{-1}}$)	190^{+16}_{-16}	167^{+32}_{-21}
v_{rest} ($\mathrm{km\,s^{-1}}$)	$-3.0^{+1.0}_{-1.3}$	$-4.0^{+2.0}_{-1.8}$
W T-Test	N/A	5.21σ

Table 5.4: The table shows the results obtained for the dataset HD189733b.

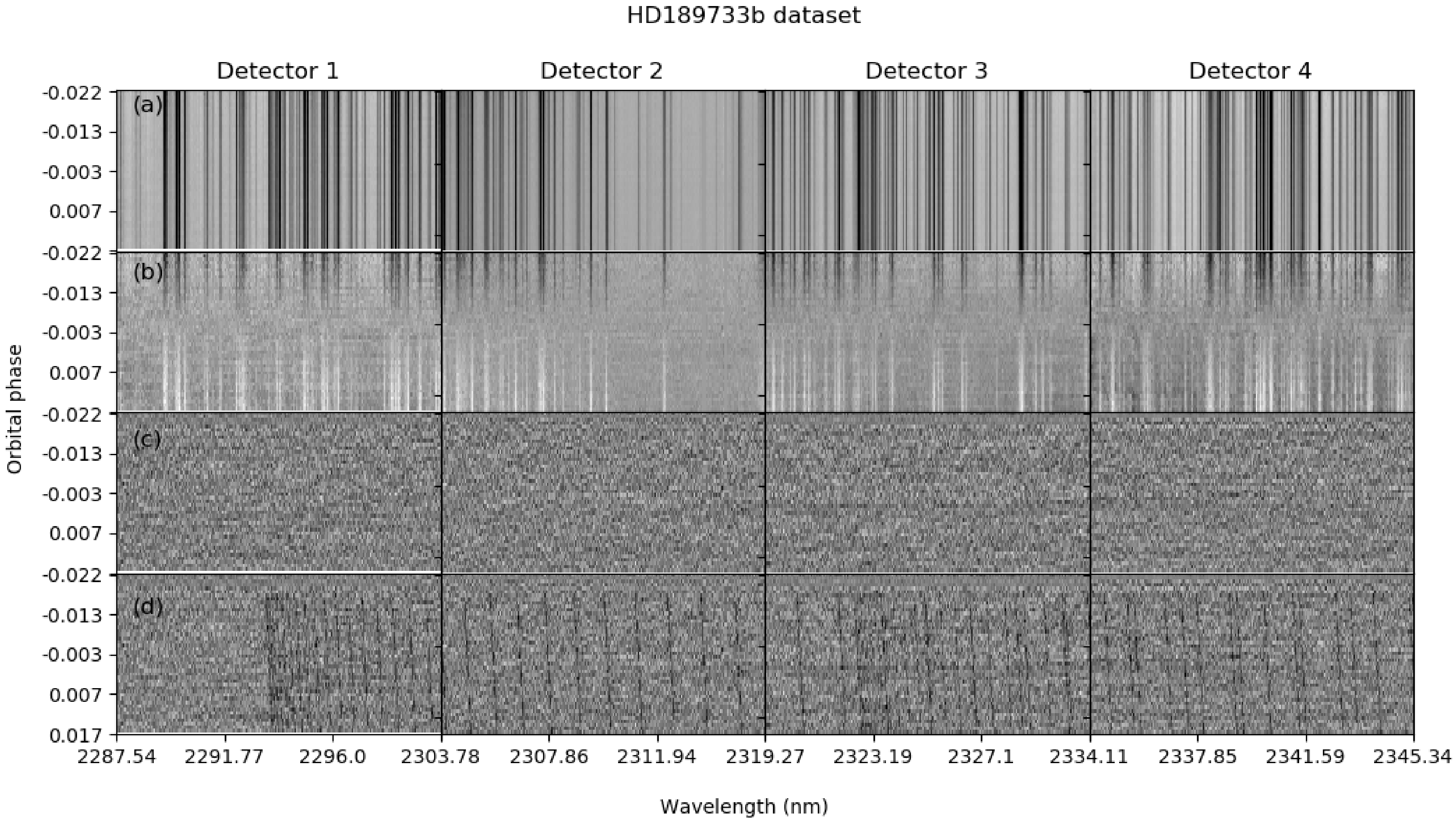

Figure 5.32: HD189733b dataset. In **(a)**, the data are shown after calibration, normalisation and spikes correction. In **(b)**, the MFS method has been applied to the data (Sec. 5.4). In **(c)**, the data are shown after PCA is applied. In **(d)**, the same figure of *c* but with the injection of CO.

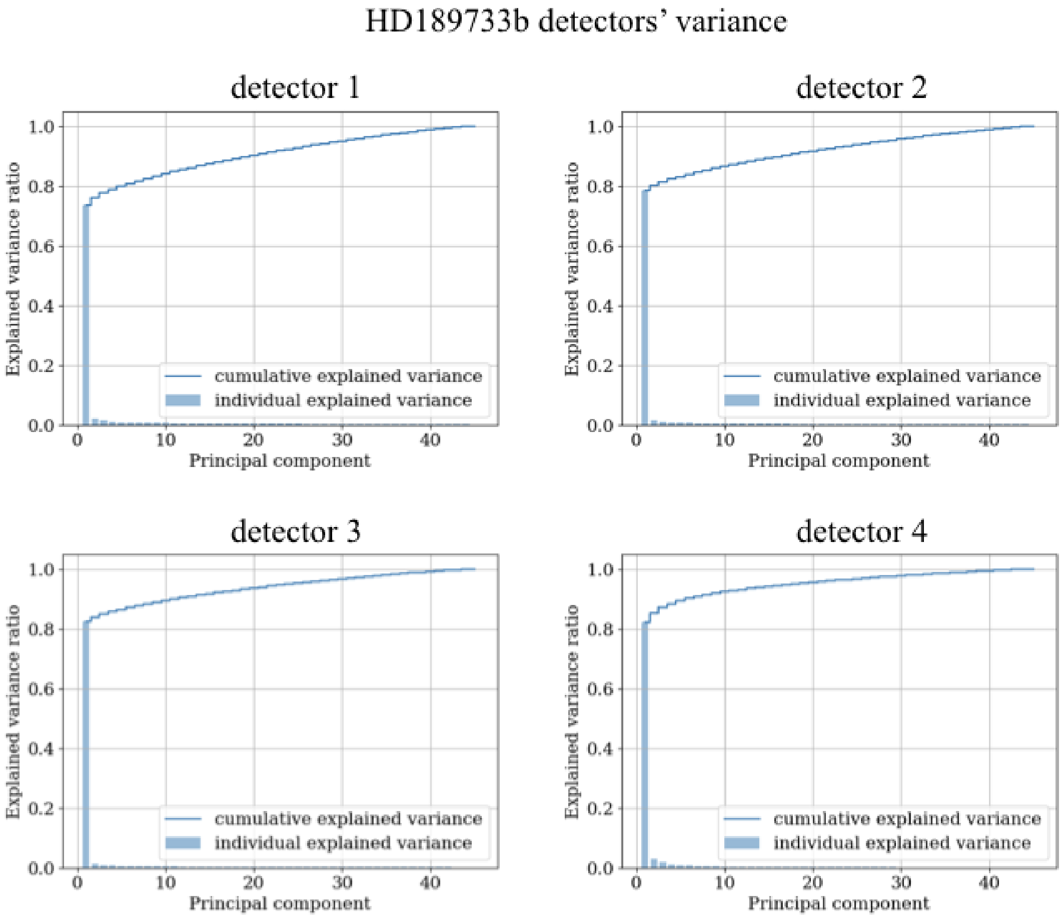

Figure 5.33: Detectors' variance of the PCA decomposition relative to HD189733b dataset. The first component always carries more than 75% of the information.

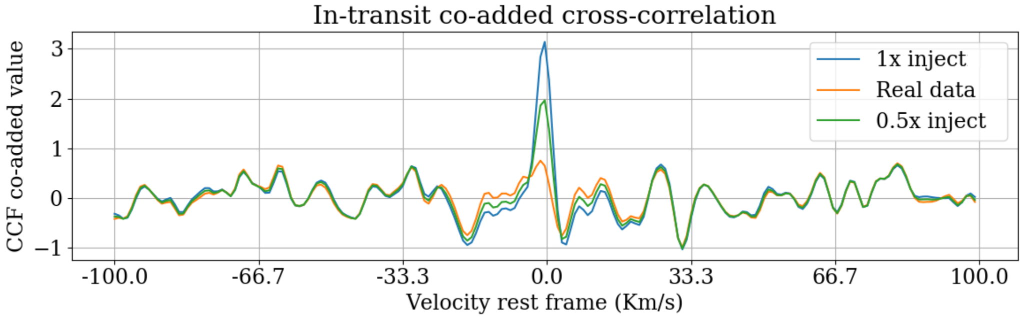

Figure 5.34: The picture shows the co-added in-transit cross-correlations for the input data and for different levels of injections of carbon monoxide. The injected signal and the planetary signal are present after using PCA. The co-added CCFs shown in picture are calculated aligning the singular CCFs to the theoretical orbital velocity of the planet HD189733b ($K_p = 152.564 \text{ km s}^{-1}$). This graph has been generated using PCA components from $22th$ to $36th$, same of Fig. 5.35

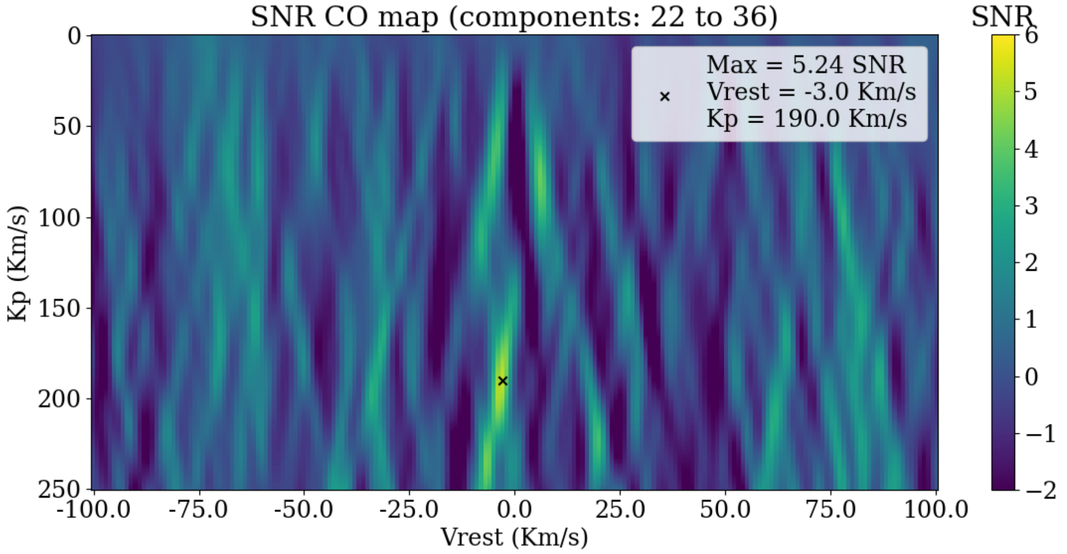

Figure 5.35: The figure shows the SNR map for the carbon monoxide for the HD189733b dataset. The peak of the matrix is compatible with the results reported by Brogi et al. (2016) but not with the theoretical radial velocity of HD189733b ($K_p = 152.564 \text{ km s}^{-1}$).

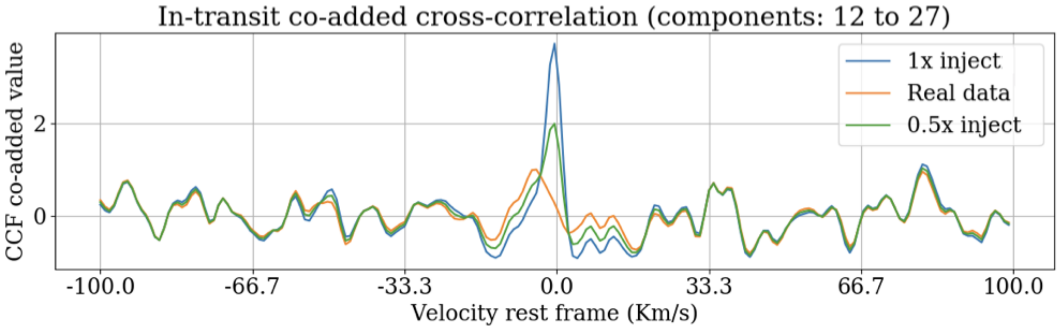

Figure 5.36: Cross-correlations of the planetary signal and of the injected water vapour co-added in-transit. The co-added CCFs are calculated at the theoretical orbital velocity of the planet HD189733b ($K_p = 152.564 \text{ km s}^{-1}$). The CCFs are the result of the combination of the PCA components from $12th$ to $27th$.

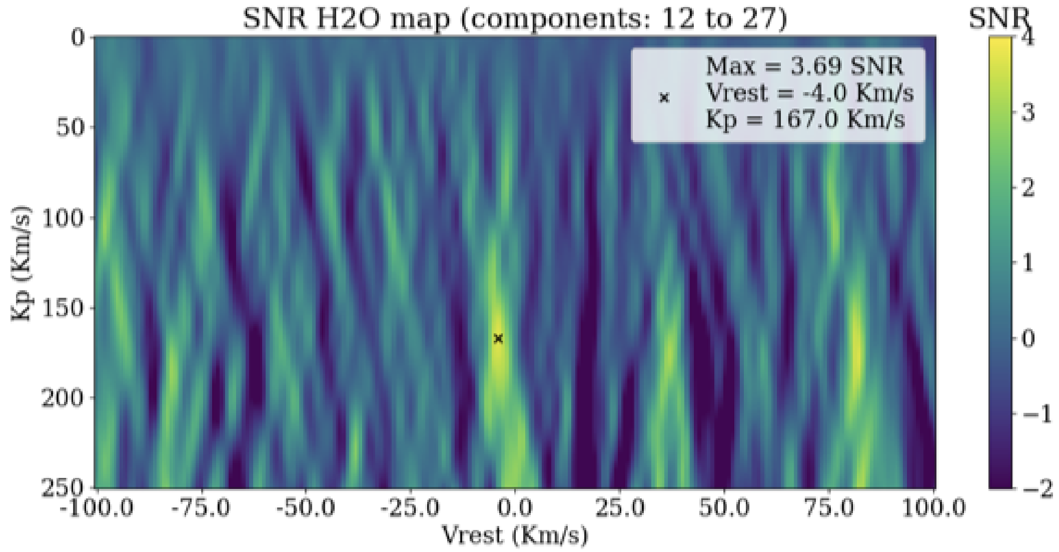

Figure 5.37: SNR map of water vapour for the HD189733b dataset. The maximum point is compatible with Brogi et al. (2016) and with the planetary parameters.

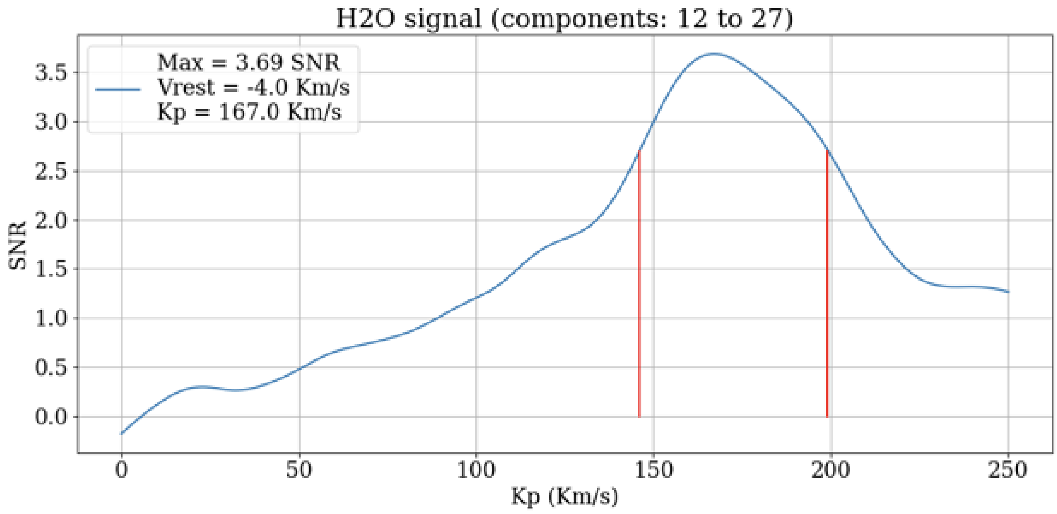

Figure 5.38: The figure shows the relation between SNR and K_p at fixed v_{rest} of the peak value of SNR map shown in Fig. 5.37. The red vertical lines indicate the uncertainty interval.

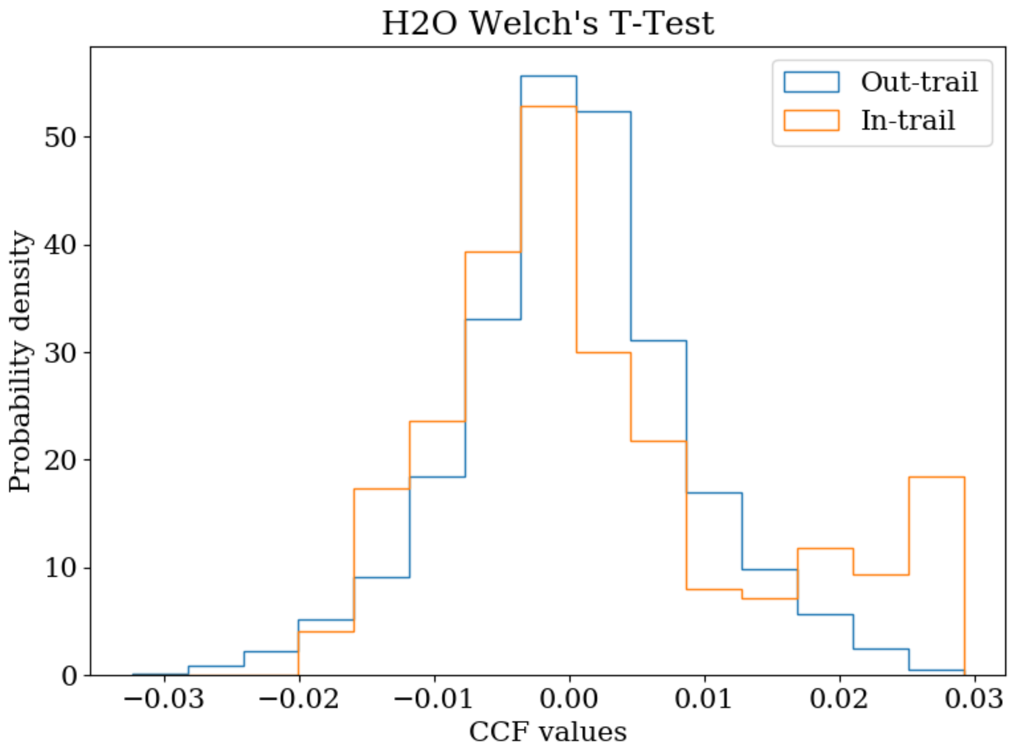

Figure 5.39: The figure shows the two distribution (i.e. *in-trail* and *out-trail*) used to compute the Whelch's T-Test. The null hypothesis is rejected with a confidence of 5.21σ.

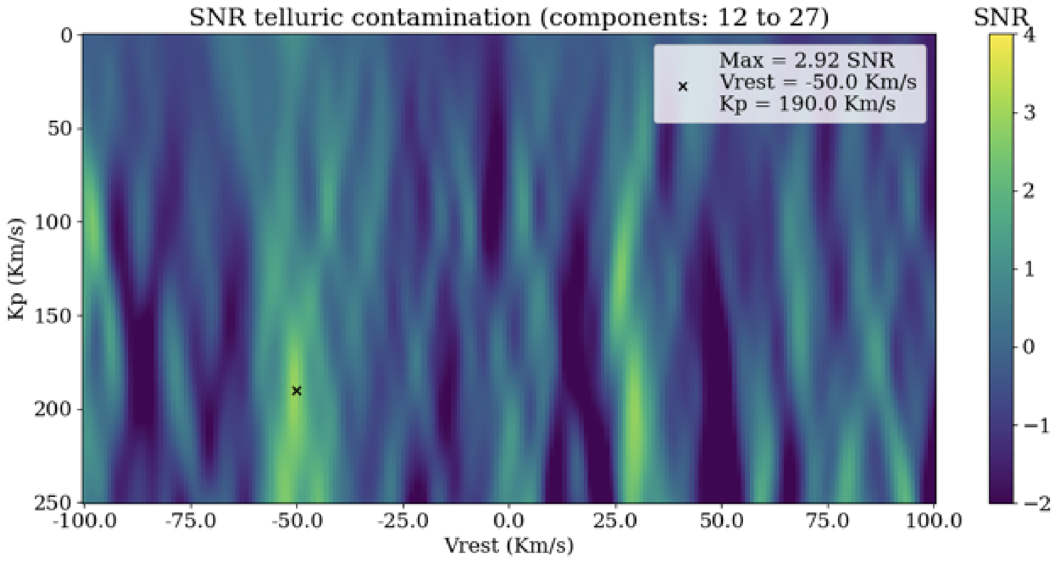

Figure 5.40: SNR map of the HD189733b dataset cross-correlated with the telluric model in the same range of components of the result reported in Fig. 5.37.

5.7.4 Discussion

As explained in the previous sections, to test the pipeline that I have developed and described here, I have re-analysed two CRIRES' datasets (HD209458b and HD189733b). CO and H_2O have been detected in the HD209458b dataset, and H_2O in the HD189733b dataset. The detection of CO in the HD209458b atmosphere is supported by an SNR peak of 5.7 at K_p and v_{rest} compatible with the planetary orbital parameters. Contrary to CO, H_2O is present in the Earth's atmosphere and therefore an accurate telluric correction is required. The lower SNR peak may be due to a lower concentration of H_2O with respect to CO in the atmosphere of HD209458b, or part of the signal might have been removed by PCA. In both detections a blueshift has been observed and this could be explained with high altitude winds. The results presented here are in agreement with the results published by Snellen et al. (2010).

Concerning HD189733b, using an unsupervised approach, we have been able to detect H_2O. The CO signal we obtained, is compatible with the literature (Brogi et al., 2016) but it is not in agreement with the theoretical radial velocity of the planet, and could be due to stellar signal contamination (K-type star contains CO). The in-transit co-added cross-correlations at the planetary velocity show an hint of the presence of CO, but the signal is buried into the noise (Fig. 5.34). Even in this dataset a blueshift of the signal has been observed and also in this case it can be associated with high altitude winds.

Pipeline performance

In the literature, previous works concerning high-resolution data analysis have been completed adopting ad-hoc corrections. In particular these corrections involved masks to neglect strong telluric lines and linear or second order polynomial fit to correct the airmass variation (Snellen et al., 2010; Brogi et al., 2012, 2013, 2014, 2016, 2018). In general this approach may led to results that are not easily reproducible. The automatic pipeline presented here does not have prior knowledge on the shape or depth of the telluric lines or on the variation of the airmass. This makes the algorithm more general, and it could be applied to high-resolution observations taken by other instruments.

The SYSREM algorithm (Tamuz et al., 2005) has been used in the context of high-resolution data analysis allowing for the detection of H_2O (Birkby et al., 2013) and TiO (Nugroho et al., 2017). Since in Tamuz et al. (2005) it is stated that the orthogonality of the calculated components in not guaranteed, we compared the performance of SYSREM and PCA. We noted that the results obtained by PCA are better in terms of the calculated SNR (see Fig. 5.25 and 5.27). This may be due to two factors:

1. Using PCA we can access all the components at the same time, allowing to neglect those components with higher and lower associated variance. This allowed us to subtract signal we were not interested in and to reduce the correlated noise enhancing the SNR. SYSREM allows only to neglect calculated components iteratively from the most important onwards.

2. SYSREM does not guarantee the orthogonality of the calculated components meaning that more than one signal can be contained in a particular component rising the risk of small signal to be erased.

The pipeline, however, has some limitations. It can cross-correlate synthetic models that contain one molecule at time. For the HD209458b datasets, for example, the CO and the H_2O signals have been found in different components. If a combined model (e.g. CO + H_2O) is cross-correlated with the data a different components range is obtained. Since different molecules may be present at different concentration in the atmosphere, during the the decomposition analysis, the signals of different molecules are comprised in different components. Finally, an important point to mention is the different signal decompositions across the four CRIRES' detectors. The EVR is different on each detector (Fig. 5.23 and 5.33) and this means that the signal relative to a particular molecule is found in different components across the detectors. An optimal approach should choose different numbers of components per detector based on their variance.

In 2014 CRIRES has been dismissed for an upgrade. A new set of gratings are going to be installed converting the instrument into a high-dispersion spectrograph with 10x the range that it had. Even if this pipeline has been designed for analysing CRIRES datasets, its general approach can be extended to different instruments. In the next section I will describe the work in progress with the instrument TNG/GIANO-B.

5.8 TNG/Giano-B datasets

Thanks to the collaboration with the Observatory of Palermo, during my Ph.D. I have joined the GAPS team[7]. The team has access to high-resolution data recorded with the instruments GIANO-B (Oliva et al., 2012; Origlia et al., 2014) (50'000 resolution power, IR band) and HARPS-N (Cosentino et al., 2012) (115'000 resolution power, optical band) installed on the TNG (Telescopio Nazionale Galileo) in La Palma. The two instruments work simultaneously to offer a coverage from optical to IR band (0.3 – 2.5 μm) and they, together, are referred to as GIARPS (Claudi et al., 2016). I will focus only on the IR part of the instrument. GIANO-B is a near-infrared high-resolution spectrograph and provides cross-dispersed spectroscopy at a resolution of $\sim 50'000$ (velocity resolution per pixel ~ 3 $\mathrm{km\,s^{-1}}$) in the near-infrared 0.9 – 2.45 μm spectral range in a single exposure. Unlike the VLT/CRIRES, GIANO-B has a single detector of 2048x2048 pixels containing all the diffraction orders dispersed by the echelle slit. The instrument has been in commissioning phase until last year and it is now available for regular use.

I recently started to test my pipeline on datasets recorded by GIANO-B.

5.8.1 Data reduction

The data reduction pipeline, called GOFIO, is a python package with some subroutines written in fortran77[8]. It has been embedded into my pipeline, even if reduced frames are also available on the archive[9]. Similarly to VLT/CRIRES, the

[7] http://www.oact.inaf.it/exoit/EXO-IT/Projects/Entries/2011/12/27_GAPS.html

[8] https://atreides.tng.iac.es/monica.rainer/gofio

[9] http://archives.ia2.inaf.it/tng/faces/search.xhtml;jsessionid=d415ff68349076c3b060b0f802eb?dswid=-3961

main TNG/GIANO-B observing mode is the nodding mode. The object is observed first in the nodding position A, then B (or B and then A) to create the sequences ABBA or ABAB. The process starts by calculating the differences A-B or B-A to do sky and dark subtraction (see Fig. 5.41). On these images the flat correction is performed and subsequently the images are straightened and an optimal extraction (Horne, 1986) is performed order by order (for order we intend a row in the trace of the signal on the detector, see Fig. 5.41).

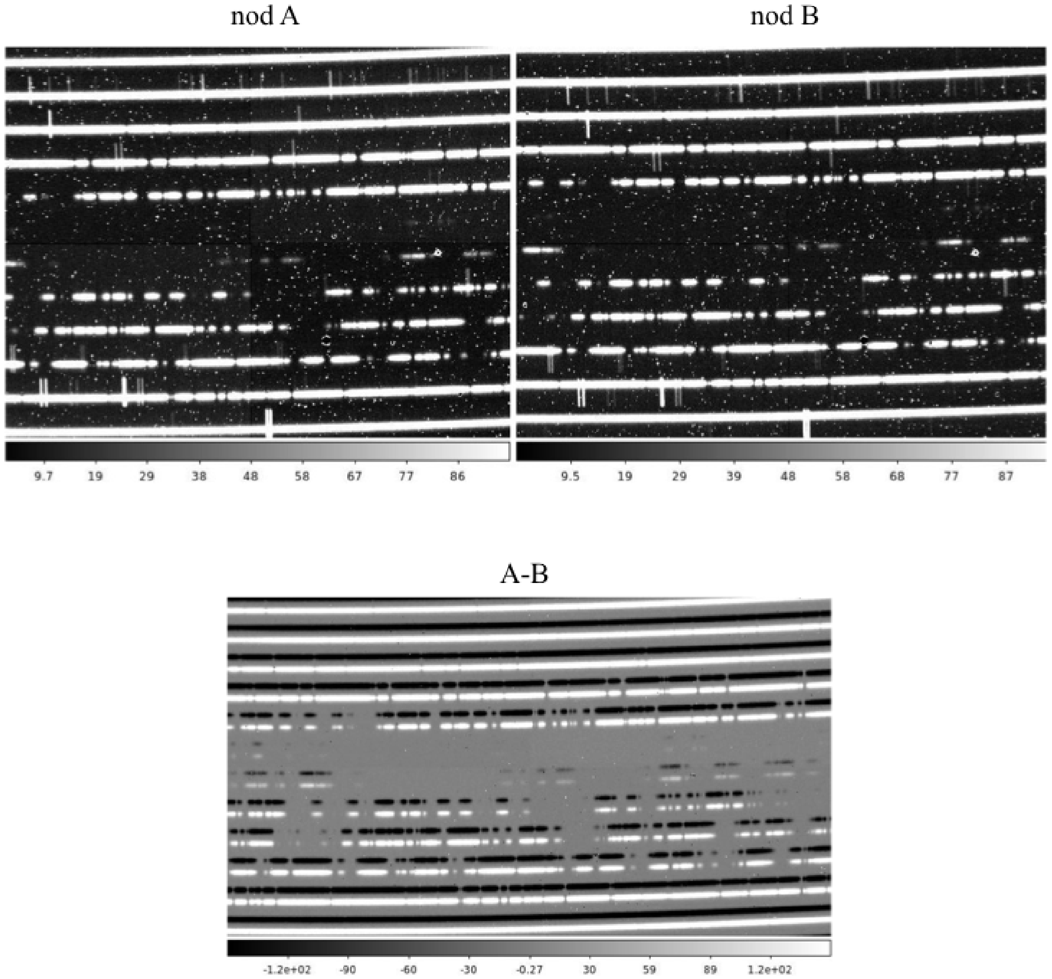

Figure 5.41: GIANO-B' reduction process. **Top panels**: raw images observed in nodding A (left) and B (right). **Bottom panel**: difference A-B: the B trace has negative (dark) values. Figure source: GOFIO manual `https://atreides.tng.iac.es/monica.rainer/gofio`.

Contrary to CRIRES' datasets, the nodding positions, here, are extracted separately to keep the original time resolution. Finally, the extracted 1D spectra are calibrated order by order using the arc lamp file recorded at the end of the observational night.

One of the main differences between CRIRES, which is a pre-dispersed high-resolution spectrograph, and GIANO-B (cross-dispersed spectrograph) is the simultaneous wavelength range coverage. In the latter all the orders are recorded at the same time in the detector, while in CRIRES the slit is positioned on the desired wavelength range.

5.8.2 Data analysis

To test my pipeline on a GIANO-B dataset, I started working on the WASP-33b dataset (Smith et al., 2011) (GAPS team proprietary data), which is an hot-Jupiter ($T \sim 3600$K) orbiting around a very hot star (~ 7500K).

After the reduction process (Fig. 5.42), all the spectra are stored in matrices (one per each order) which are in total 50 (from order 81 to 32). As shown in Fig. 5.42 some orders are saturated by telluric absorption and for those no results can be calculated and they are, then, excluded from the analysis. This will reduce the number of orders to about 25 (Fig. 5.43). After the calibration process (Sec. 5.3) we have noticed that the nodding A and the nodding B (extracted separately) show differences in the baseline level after the normalisation. In Fig. 5.44 top panel we can appreciate this effect. Two subsequent rows of the matrix are a AB or BA nodding couple. Moreover, after subtracting the mean of each column (MFS method, Sec. 5.4) (Fig. 5.44 bottom panel) it appears also a difference between the left half of the detector and the right one. Every order is then not only split in two different matrices containing only the nodding A or the nodding B sequences, but also divided in two sub matrices. In this way every matrix relative to a specific order is divided into four distinct matrices rising the total number of matrices to analyse to more than one hundred. This has a considerable impact on the execution time of the pipeline. Remember that CRIRES' datasets consisted of only four matrices.

Finally, some of the matrices show a different PCA decomposition in terms of EVR, see for example Fig. 5.45 where the first component does not carry more than

75% like in the case of CRIRES' datasets.

Since H_2O is not expected to be present at these temperature ($T \sim 3600$K) due to thermal dissociation, we have tried to cross-correlate CO and TiO using an isothermal T/p profile at $T_p = 3600$K and VMR equal to respectively 10^{-3} and 10^{-8} with pressure varying from $10^{-5} - 10^{2}$ (parameters adopted from Smith et al. (2011) and Nugroho et al. (2017)).

No detection has been determined, suggesting that a more detailed study on the pre-process and normalisation steps is required additionally to an optimal choice of PCA components per matrix.

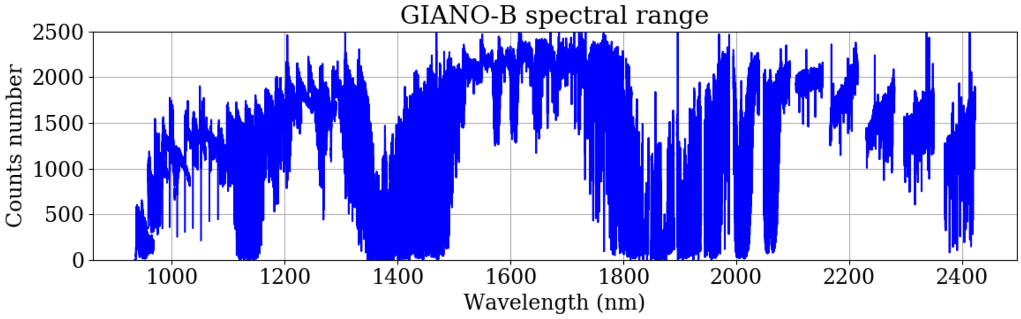

Figure 5.42: GIANO-B' extracted spectra. All the 50 orders are shown and some of them are highly affected by the telluric absorption.

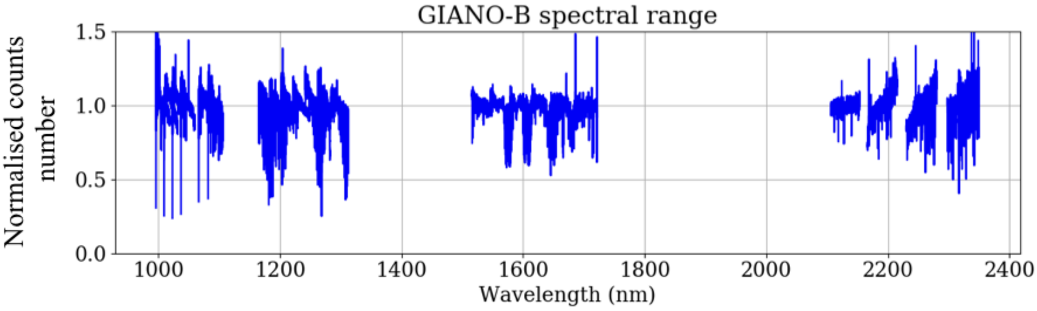

Figure 5.43: Same of Fig. 5.42 but here the saturated orders have been neglected and the remaining orders have been normalised.

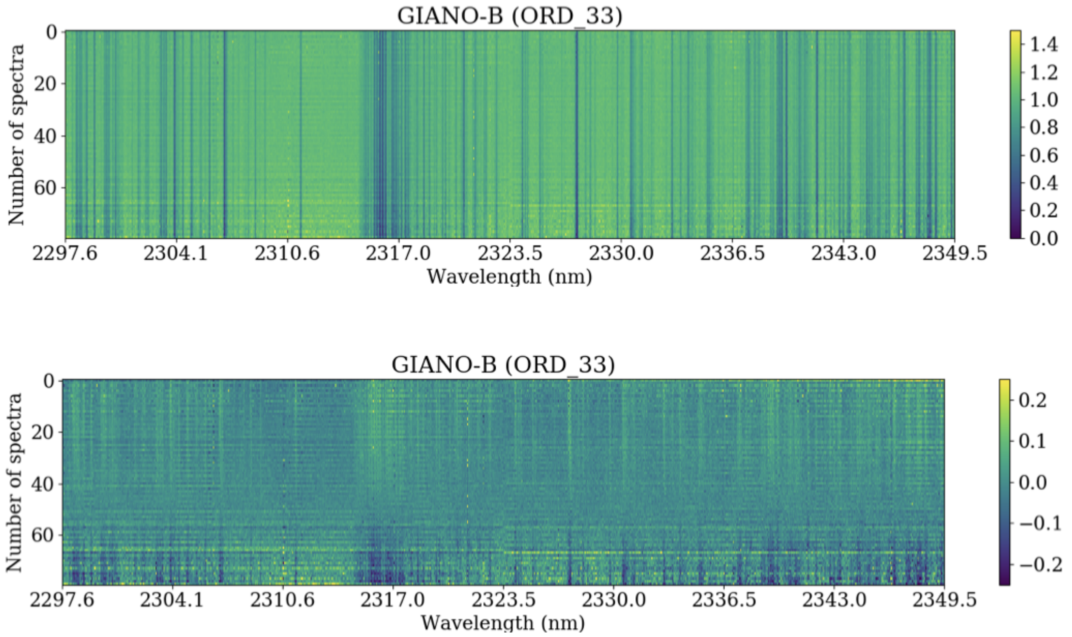

Figure 5.44: **Top panel**: GIANO-B' order 33 before the pre-processing. The differences on the nodding baselines are visible. **Bottom panel**: same matrix as top panel but the MFS method has been applied. The difference between the left and right areas of the detector is visible.

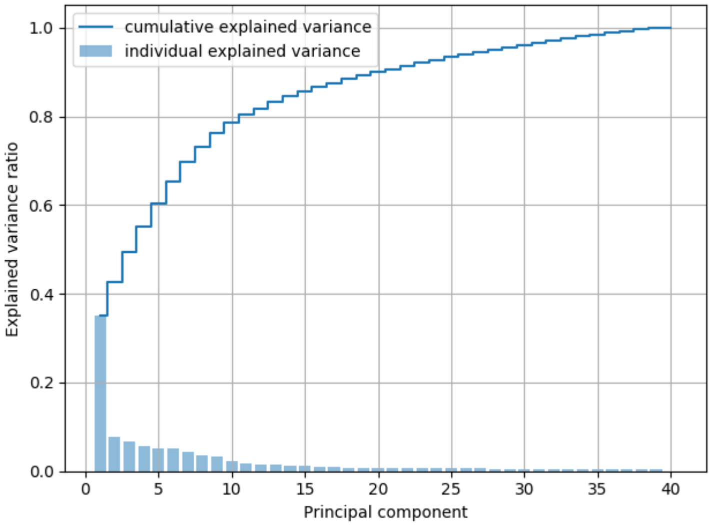

Figure 5.45: GIANO-B' order 33 variance of the PCA decomposition relative to the WASP-33b dataset. The first component carries less than 50% of the information.

Chapter 6

Conclusion

"Just when you think you know something, you have to look at in another way. Even though it may seem silly or wrong, you must try."

– Robbie Williams - 1989

The characterisation of an exoplanet's atmosphere can be achieved by using a good variety of observational and analysis strategies. In this thesis I focused on data analysis techniques that improve the performance of current methods, while limiting the introduction of the human bias through ad-hoc corrections.

In chapter 3 I have discussed well known techniques (i.e. ICA, PCA and SYSREM) for signal decomposition, highlighting differences and similarities not reported in the literature.

Part of my Ph.D. work has been focused on data analysis of space-based observations recorded by the HST/WFC3 camera (chapter 4). The UCL pipeline I have contributed to (Tsiaras et al., 2016a) can perform better with respect to other approaches that can be found in the literature (Deming et al., 2013; Wilkins et al., 2014). The combined effect of the multiple aperture extraction module (able to divide the signal in temporal sequences. i.e. 'stripes') and a binning module assuring similar SNR across all the bins has decreased the scattering on the final 1D spectrum (Tsiaras et al., 2018). Additionally, it has allowed the analysis of blended signals and had permitted the use of a different approach to treat instrument systematics trough a machine learning approach (Damiano et al., 2017). The method adopted for the HST/WFC3 could be adapted and extended to future space observations such as the James Webb Space Telescope (JWST)[1] and the Atmospheric Remote-sensing Exoplanet Large-survey (ARIEL)[2]. These two facilities will provide thousands of observations at low/medium-resolution over a broad spectral range.

In the second half of my Ph.D. I have developed from scratch a pipeline to analyse high spectral resolution data (chapter 5) recorded from the ground. Here, I demonstrated that the systematics introduced by telluric absorption can be removed without manual intervention or introduction of ad-hoc masks as done in previous papers(Snellen et al., 2010; Birkby et al., 2013; Brogi et al., 2016). This has been achieved by taking full advantage of PCA technique (Damiano et. al submitted). In particular, analysing the HD189733b dataset I have suggested that results presented in Brogi et al. (2016) may be affected by contamination of the stellar signal. Ground based high-resolution observations are rapidly growing and new facilities have recently started to be operative (e.g. ESPRESSO, CARMENES, HARPS-N, GIANO-B) and in the next decade a breakthrough is expected with the E-ELT[3] that will be equipped with a 39-meter class telescope.

The spectra obtained with WFC3 and ground-based observatories were compared with a chemical/atmospheric model to understand the composition and the structure of the exoplanet's atmosphere. In both projects we used $\mathscr{T}$-REx (Waldmann et al., 2015b,a) (Sec. 2.3) to synthesise planetary spectra using ExoMol line lists (Tennyson & Yurchenko, 2012). In particular the use of the ExoMol line lists can improve the data analysis since it has been demonstrated that at higher temperature (T>1000 K) they are more accurate than HITRAN which is measured at room temperature (e.g. for methane HITRAN vs ExoMol Lavie et al. (2017)).

In chapter 2 I have introduced the concept of high- and low-resolution observations. The differences, between them, have been highlighted when the two dedicated pipelines for data analysis have been discussed. However, ground-based high-resolution spectroscopy over a narrow spectral interval and space-based low/medium resolution over a broad wavelength range, are very complementary ap-

[1] https://www.jwst.nasa.gov

[2] https://ariel-spacemission.eu

[3] https://www.eso.org/sci/facilities/eelt/

proaches (de Kok et al., 2014; Brogi et al., 2017; Pino et al., 2018)). A long list of space and ground-based instruments will come online in the next decade. Having developed skills in both space and ground observations will be crucial to advance the study of exoplanetary atmospheres.

Bibliography

Agol, E., Cowan, N. B., Knutson, H. A., et al. 2010, ApJ, 721, 1861

Alonso, R., Auvergne, M., Baglin, A., et al. 2008, A&A, 482, L21

Atreya, S. K., Mahaffy, P. R., Niemann, H. B., Wong, M. H., & Owen, T. C. 2003, Planetary and Space Science, 51, 105

Auvergne, M., Bodin, P., Boisnard, L., et al. 2009, A&A, 506, 411

Bakos, G. Á. 2018, The HATNet and HATSouth Exoplanet Surveys, 111

Barber, R. J., Strange, J. K., Hill, C., et al. 2014, MNRAS, 437, 1828

Barber, R. J., Tennyson, J., Harris, G. J., & Tolchenov, R. N. 2006, MNRAS, 368, 1087

Barge, P., Baglin, A., Auvergne, M., et al. 2008, A&A, 482, L17

Barman, T. 2007, ApJ, 661, L191

Barnes, J. R., Barman, T. S., Prato, L., et al. 2007a, MNRAS, 382, 473

Barnes, J. R., Leigh, C. J., Jones, H. R. A., et al. 2007b, MNRAS, 379, 1097

Barstow, J. K., Aigrain, S., Irwin, P. G. J., & Sing, D. K. 2017, ApJ, 834, 50

Batygin, K., Stevenson, D. J., & Bodenheimer, P. H. 2011, ApJ, 738, 1

Ben-Jaffel, L. 2007, ApJ, 671, L61

Berta, Z. K., Charbonneau, D., Désert, J.-M., et al. 2012, ApJ, 747, 35

Birkby, J. L. 2018, ArXiv e-prints, arXiv:1806.04617

Birkby, J. L., de Kok, R. J., Brogi, M., et al. 2013, MNRAS, 436, L35

Birkby, J. L., de Kok, R. J., Brogi, M., Schwarz, H., & Snellen, I. A. G. 2017, AJ, 153, 138

Böhm-Vitense, E. 1992, Introduction to Stellar Astrophysics:, Introduction to Stellar Astrophysics (Cambridge University Press)

Borucki, W. J., Koch, D., Basri, G., et al. 2003, in ESA Special Publication, Vol. 539, Earths: DARWIN/TPF and the Search for Extrasolar Terrestrial Planets, ed. M. Fridlund, T. Henning, & H. Lacoste, 69–81

Borysow, A. 2002, A&A, 390, 779

Borysow, A., Jorgensen, U. G., & Fu, Y. 2001, Journal of Quantitative Spectroscopy and Radiative Transfer, 68, 235

Bouchy, F., Udry, S., Mayor, M., et al. 2005, A&A, 444, L15

Bracewell, R. 1965, The Fourier Transform and Its Applications. (New York: McGraw-Hill), 46 and 243

Brogi, M., de Kok, R. J., Albrecht, S., et al. 2016, ApJ, 817, 106

Brogi, M., de Kok, R. J., Birkby, J. L., Schwarz, H., & Snellen, I. A. G. 2014, A&A, 565, A124

Brogi, M., Giacobbe, P., Guilluy, G., et al. 2018, A&A, 615, A16

Brogi, M., Line, M., Bean, J., Désert, J. M., & Schwarz, H. 2017, ApJ, 839, L2

Brogi, M., Snellen, I. A. G., de Kok, R. J., et al. 2012, Nature, 486, 502

Brogi, M., Snellen, I. A. G., de Kok, R. J., et al. 2013, ApJ, 767, 27

Brown, T. M. 2001, The Astrophysical Journal, 553, 1006

Brown, T. M., Libbrecht, K. G., & Charbonneau, D. 2002, Publications of the Astronomical Society of the Pacific, 114, 826

Burke, C. J., Bryson, S. T., Mullally, F., et al. 2014, The Astrophysical Journal Supplement Series, 210, 19

Cassan, A., Kubas, D., Beaulieu, J. P., et al. 2012, Nature, 481, 167

Charbonneau, D., Brown, T. M., Latham, D. W., & Mayor, M. 2000, ApJ, 529, L45

Charbonneau, D., Brown, T. M., Noyes, R. W., & Gilliland, R. L. 2002, ApJ, 568, 377

Charbonneau, D., Jha, S., & Noyes, R. W. 1998, ApJ, 507, L153

Charbonneau, D., Knutson, H. A., Barman, T., et al. 2008, ApJ, 686, 1341

Charbonneau, D., Noyes, R. W., Korzennik, S. G., et al. 1999, ApJ, 522, L145

Chiavassa, A., Pere, C., Faurobert, M., et al. 2015, A&A, 576, A13

Claret, A. 2000, A&A, 363, 1081

Claudi, R., Benatti, S., Carleo, I., et al. 2016, in Ground-based and Airborne Instrumentation for Astronomy VI, Vol. 9908, 99081A

Collier Cameron, A., Horne, K., Penny, A., & James, D. 1999, Nature, 402, 751

Comon, P. 1994, Signal Process., 36, 287

Comon, P., & Jutten, C. 2010, Handbook of Blind Source Separation: Independent Component Analysis and Applications, 1st edn. (Orlando, FL, USA: Academic Press, Inc.)

Cosentino, R., Lovis, C., Pepe, F., et al. 2012, in Ground-based and Airborne Instrumentation for Astronomy IV, Vol. 8446, 84461V

Coughlin, J. L., Mullally, F., Thompson, S. E., et al. 2016, The Astrophysical Journal Supplement Series, 224, 12

Cowan, N. B., Machalek, P., Croll, B., et al. 2012, ApJ, 747, 82

Crossfield, I. J. M., Hansen, B. M. S., Harrington, J., et al. 2010, ApJ, 723, 1436

Csizmadia, S., Hatzes, A., Gandolfi, D., et al. 2015, A&A, 584, A13

Damasso, M., Bonomo, A. S., Astudillo-Defru, N., et al. 2018, A&A, 615, A69

Damiano, M., Morello, G., Tsiaras, A., Zingales, T., & Tinetti, G. 2017, AJ, 154, doi:10.3847/1538-3881/aa738b

de Kok, R. J., Birkby, J., Brogi, M., et al. 2014, A&A, 561, A150

de Kok, R. J., Brogi, M., Snellen, I. A. G., et al. 2013, A&A, 554, A82

Deming, D., Harrington, J., Seager, S., & Richardson, L. J. 2006, ApJ, 644, 560

Deming, D., Seager, S., Richardson, L. J., & Harrington, J. 2005, Nature, 434, 740

Deming, D., Wilkins, A., McCullough, P., et al. 2013, ApJ, 774, 95

Demory, B.-O., Gillon, M., de Wit, J., et al. 2016, Nature, 532, 207

Dulude, M. J., Baggett, S., & Hilbert, B. 2014, New WFC3/IR Dark Calibration Files, Tech. rep.

Einstein, A. 1905, Annalen der Physik, 322, 891

Encrenaz, T., Bibring, J. P., & Blanc, M. 2004, The solar system

Espinoza, N., & Jordán, A. 2015, MNRAS, 450, 1879

Evans, T. M., Aigrain, S., Gibson, N., et al. 2015, MNRAS, 451, 680

Evans, T. M., Sing, D. K., Wakeford, H. R., et al. 2016, ApJ, 822, L4

Follert, R., Dorn, R. J., Oliva, E., et al. 2014, in Ground-based and Airborne Instrumentation for Astronomy V, Vol. 9147, 914719

Fraine, J., Deming, D., Benneke, B., et al. 2014, Nature, 513, 526

Frei, R., & Rosing, M. T. 2005, Earth and Planetary Science Letters, 236, 28

Fressin, F., Torres, G., Charbonneau, D., et al. 2013, ApJ, 766, 81

Fulton, B. J., Collins, K. A., Gaudi, B. S., et al. 2015, ApJ, 810, 30

Fulton, B. J., Petigura, E. A., Howard, A. W., et al. 2017, AJ, 154, 109

Gabriel, R. K., & Zamir, S. 1979, Technometrics, 21, 489

Gibson, N. P., Aigrain, S., Barstow, J. K., et al. 2013, MNRAS, 436, 2974

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., et al. 2014, ArXiv e-prints, arXiv:1406.2661

Grillmair, C. J., Burrows, A., Charbonneau, D., et al. 2008, Nature, 456, 767

Hansen, B. M. S., & Zink, J. 2015, MNRAS, 450, 4505

Harrington, J., Hansen, B. M., Luszcz, S. H., et al. 2006, Science, 314, 623

Hartman, J. D., Bakos, G. Á., Torres, G., et al. 2011, ApJ, 742, 59

Haynes, K., Mandell, A. M., Madhusudhan, N., Deming, D., & Knutson, H. 2015, ApJ, 806, 146

Henry, G. W., Marcy, G. W., Butler, R. P., & Vogt, S. S. 2000, ApJ, 529, L41

Hertzsprung, E. 1912, Astronomische Nachrichten, 192, 309

Hilbert, B. 2008, WFC3 TV3 Testing: IR Channel Nonlinearity Correction, Tech. rep.

—. 2012, WFC3/IR Cycle 19 Bad Pixel Table Update, Tech. rep.

—. 2014, Updated non-linearity calibration method for WFC3/IR, Tech. rep.

Hoeijmakers, H. J., de Kok, R. J., Snellen, I. A. G., et al. 2015, A&A, 575, A20

Hollis, M. D. J., Tessenyi, M., & Tinetti, G. 2013, Computer Physics Communications, 184, 2351

Horne, K. 1986, Publications of the Astronomical Society of the Pacific, 98, 609

Howarth, I. D. 2011, MNRAS, 413, 1515

Howell, S. B., Sobeck, C., Haas, M., et al. 2014, PASP, 126, 398

Huitson, C. M., Sing, D. K., Vidal-Madjar, A., et al. 2012, MNRAS, 422, 2477

Huitson, C. M., Sing, D. K., Pont, F., et al. 2013, MNRAS, 434, 3252

Hyvärinen, A. 1999, Neural Comput., 11, 1739

—. 2001, Neural Comput., 13, 883

—. 2012, Philosophical Transactions of the Royal Society of London Series A, 371, 20110534

Hyvärinen, A., & Oja, E. 2000, Neural Netw., 13, 411

Hyvärinen, A., & Pajunen, P. 1999, Neural Netw., 12, 429

Iyer, A. R., Swain, M. R., Zellem, R. T., et al. 2016, ApJ, 823, 109

Janson, M., Bergfors, C., Goto, M., Brandner, W., & Lafrenière, D. 2010, ApJ, 710, L35

Jehin, E., Gillon, M., Queloz, D., et al. 2011, The Messenger, 145, 2

Jolliffe, I. T. 2002, Principal component analysis

Kaeufl, H.-U., Ballester, P., Biereichel, P., et al. 2004, in Ground-based Instrumentation for Astronomy, Vol. 5492, 1218–1227

Kipping, D. M. 2010, MNRAS, 407, 301

Kleine, T., Touboul, M., Bourdon, B., et al. 2009, Geochimica et Cosmochimica Acta, 73, 5150

Knutson, H. A., Benneke, B., Deming, D., & Homeier, D. 2014a, Nature, 505, 66

Knutson, H. A., Charbonneau, D., Allen, L. E., Burrows, A., & Megeath, S. T. 2008, ApJ, 673, 526

Knutson, H. A., Charbonneau, D., Burrows, A., O'Donovan, F. T., & Mandushev, G. 2009a, ApJ, 691, 866

Knutson, H. A., Charbonneau, D., Cowan, N. B., et al. 2009b, ApJ, 703, 769

Knutson, H. A., Charbonneau, D., Noyes, R. W., Brown, T. M., & Gilliland, R. L. 2007a, ApJ, 655, 564

Knutson, H. A., Charbonneau, D., Allen, L. E., et al. 2007b, Nature, 447, 183

Knutson, H. A., Dragomir, D., Kreidberg, L., et al. 2014b, ApJ, 794, 155

Kobayashi, N., Tokunaga, A. T., Terada, H., et al. 2000, in Optical and IR Telescope Instrumentation and Detectors, Vol. 4008, 1056–1066

Koldovsky, Z., Tichavsky, P., & Oja, E. 2006, IEEE Transactions on Neural Networks, 17, 1265

Konopacky, Q. M., Barman, T. S., Macintosh, B. A., & Marois, C. 2013, Science, 339, 1398

Kreidberg, L., Bean, J. L., Désert, J.-M., et al. 2014, Nature, 505, 69

Kreidberg, L., Line, M. R., Bean, J. L., et al. 2015, ApJ, 814, 66

Krick, J. E., Ingalls, J., Carey, S., et al. 2016, ApJ, 824, 27

Krizhevsky, A., Sutskever, I., & Hinton, G. E. 2012, in Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1, NIPS'12 (USA: Curran Associates Inc.), 1097–1105

Kümmel, M., Kuntschner, H., Walsh, J. R., & Bushouse, H. 2011, Master sky images for the WFC3 G102 and G141 grisms, Tech. rep.

Kuntschner, H., Bushouse, H., Kümmel, M., & Walsh, J. R. 2009, WFC3 SMOV proposal 11552: Calibration of the G141 grism, Tech. rep.

Kuntschner, H., Kümmel, M., Walsh, J. R., & Bushouse, H. 2011, Revised Flux Calibration of the WFC3 G102 and G141 grisms, Tech. rep.

Kurucz, R. L. 1970, SAO Special Report, 309

Laughlin, G., Deming, D., Langton, J., et al. 2009, Nature, 457, 562

Lavie, B., Mendonça, J. M., Mordasini, C., et al. 2017, AJ, 154, 91

Lecun, Y., Bengio, Y., & Hinton, G. 2015, Nature, 521, 436

Lee, E. J., & Chiang, E. 2016, ApJ, 817, 90

Lee, J.-M., Heng, K., & Irwin, P. G. J. 2013, ApJ, 778, 97

Leitzinger, M., Odert, P., Kulikov, Y. N., et al. 2011, Planetary and Space Science, 59, 1472

Line, M. R., Stevenson, K. B., Bean, J., et al. 2016, AJ, 152, 203

Linsky, J. L., Yang, H., France, K., et al. 2010, ApJ, 717, 1291

Liou, K. N. 2002, An Introduction to Atmospheric Radiation, 2nd edn., Vol. 84 (Academic Press)

Lovis, C., & Fischer, D. 2010, Radial Velocity Techniques for Exoplanets, 27–53

Macintosh, B., Graham, J. R., Barman, T., et al. 2015, Science, 350, 64

Madhusudhan, N. 2012, ApJ, 758, 36

Mallonn, M., & Strassmeier, K. G. 2016, A&A, 590, A100

Manly, B. J. F. 1994 (Chapman & Hall)

Mayor, M., & Queloz, D. 1995, Nature, 378, 355

Mazeh, T., Naef, D., Torres, G., et al. 2000, ApJ, 532, L55

McKemmish, L. K., Yurchenko, S. N., & Tennyson, J. 2016, MNRAS, 463, 771

Morello, G. 2015, ApJ, 808, 56

Morello, G., Waldmann, I. P., & Tinetti, G. 2016, ApJ, 820, 86

Morello, G., Waldmann, I. P., Tinetti, G., et al. 2015, ApJ, 802, 117

—. 2014, ApJ, 786, 22

Moynier, F., Koeberl, C., Quitté, G., & Telouk, P. 2009, Earth and Planetary Science Letters, 286, 35

Mura, A., Wurz, P., Lichtenegger, H. I. M., et al. 2009, Icarus, 200, 1

Noguchi, K., Aoki, W., Kawanomoto, S., et al. 2002, Publications of the Astronomical Society of Japan, 54, 855

Nortmann, L., Pallé, E., Murgas, F., et al. 2016, A&A, 594, A65

Nugroho, S. K., Kawahara, H., Masuda, K., et al. 2017, AJ, 154, 221

Oja, E. 1992, Neural Netw., 5, 927

Oliva, E., Origlia, L., Maiolino, R., et al. 2012, in Ground-based and Airborne Instrumentation for Astronomy IV, Vol. 8446, 84463T

Origlia, L., Oliva, E., Baffa, C., et al. 2014, in Ground-based and Airborne Instrumentation for Astronomy V, Vol. 9147, 91471E

Owen, J. E., & Jackson, A. P. 2012, MNRAS, 425, 2931

Owen, J. E., & Wu, Y. 2013, ApJ, 775, 105

Pearson, K. 1901, The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 2, 559

Pino, L., Ehrenreich, D., Wyttenbach, A., et al. 2018, A&A, 612, A53

Piskorz, D., Benneke, B., Crockett, N. R., et al. 2016, ApJ, 832, 131

—. 2017, AJ, 154, 78

Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. 2007, Numerical Recipes 3rd Edition: The Art of Scientific Computing, 3rd edn. (New York, NY, USA: Cambridge University Press)

Quirrenbach, A., Amado, P. J., Caballero, J. A., et al. 2014, in Ground-based and Airborne Instrumentation for Astronomy V, Vol. 9147, 91471F

Rajan, A., Rameau, J., De Rosa, R. J., et al. 2017, AJ, 154, 10

Raschka, S., & Mirjalili, V. 2017, Python Machine Learning, 2nd Ed., 2nd edn. (Birmingham, UK: Packt Publishing)

Reale, F., Gambino, A. F., Micela, G., et al. 2015, Nature Communications, 6, 7563

Redfield, S., Endl, M., Cochran, W. D., & Koesterke, L. 2008, ApJ, 673, L87

Rein, H. 2012, ArXiv e-prints, arXiv:1211.7121

Ridden-Harper, A. R., Snellen, I. A. G., Keller, C. U., et al. 2016, A&A, 593, A129

Rothman, L. S., Gordon, I. E., Barbe, A., et al. 2009, Journal of Quantitative Spectroscopy and Radiative Transfer, 110, 533

Rothman, L. S., Gordon, I. E., Barber, R. J., et al. 2010, Journal of Quantitative Spectroscopy and Radiative Transfer, 111, 2139

Rowe, J. F., Coughlin, J. L., Antoci, V., et al. 2015, The Astrophysical Journal Supplement Series, 217, 16

Russell, H. N. 1910, AJ, 26, 147

Sabbi, E., MacKenty, J., & Borders, T. 2010, Proposal 11913-IR Filter Wedge Check, Tech. rep.

Samland, M., Mollière, P., Bonnefoy, M., et al. 2017, A&A, 603, A57

Samuel, A. L. 1959, IBM J. Res. Dev., 3, 210

Seager, S., & Mallén-Ornelas, G. 2003, ApJ, 585, 1038

Seager, S., & Sasselov, D. D. 2000, ApJ, 537, 916

Sing, D. K., Désert, J. M., Fortney, J. J., et al. 2011, A&A, 527, A73

Sing, D. K., Wakeford, H. R., Showman, A. P., et al. 2015, MNRAS, 446, 2428

Sing, D. K., Fortney, J. J., Nikolov, N., et al. 2016, Nature, 529, 59

Smith, A. M. S., Anderson, D. R., Skillen, I., Collier Cameron, A., & Smalley, B. 2011, MNRAS, 416, 2096

Snellen, I. A. G., Brandl, B. R., de Kok, R. J., et al. 2014, Nature, 509, 63

Snellen, I. A. G., de Kok, R. J., de Mooij, E. J. W., & Albrecht, S. 2010, Nature, 465, 1049

Spiegel, D. S., Silverio, K., & Burrows, A. 2009, ApJ, 699, 1487

Sportisse, B. 2009, Fundamentals in Air Pollution: From Processes to Modelling (Springer Netherlands)

Stevenson, K. B. 2016, ApJ, 817, L16

Stevenson, K. B., Désert, J.-M., Line, M. R., et al. 2014, Science, 346, 838

Street, R. A., Pollaco, D. L., Fitzsimmons, A., et al. 2003, in Scientific Frontiers in Research on Extrasolar Planets, Vol. 294, 405–408

Swain, M., Deroo, P., Tinetti, G., et al. 2013, Icarus, 225, 432

Swain, M. R., Vasisht, G., & Tinetti, G. 2008, Nature, 452, 329

Swain, M. R., Vasisht, G., Tinetti, G., et al. 2009a, ApJ, 690, L114

Swain, M. R., Tinetti, G., Vasisht, G., et al. 2009b, ApJ, 704, 1616

Tamuz, O., Mazeh, T., & Zucker, S. 2005, MNRAS, 356, 1466

Tennyson, J., & Yurchenko, S. N. 2012, MNRAS, 425, 21

Tennyson, J., Yurchenko, S. N., Al-Refaie, A. F., et al. 2016, Journal of Molecular Spectroscopy, 327, 73

Tichavsky, P., Koldovsky, Z., Doron, E., Yeredor, A., & Gomez-Herrero, G. 2006, in 2006 14th European Signal Processing Conference, 1–5

Tichavsky, P., Koldovsky, Z., & Oja, E. 2006, IEEE Transactions on Signal Processing, 54, 1189

Tichavsky, P., Koldovsky, Z., Yeredor, A., Gomez-Herrero, G., & Doron, E. 2008, IEEE Transactions on Neural Networks, 19, 421

Tinetti, G., Encrenaz, T., & Coustenis, A. 2013, Astronomy and Astrophysics Review, 21, 63

Tinetti, G., Vidal-Madjar, A., Liang, M.-C., et al. 2007, Nature, 448, 169

Torres, G., Winn, J. N., & Holman, M. J. 2008, ApJ, 677, 1324

Triaud, A. H. M. J., Queloz, D., Bouchy, F., et al. 2009, A&A, 506, 377

Tsiaras, A., Waldmann, I. P., Rocchetto, M., et al. 2016a, ApJ, 832, doi:10.3847/0004-637X/832/2/202

Tsiaras, A., Rocchetto, M., Waldmann, I. P., et al. 2016b, ApJ, 820, doi:10.3847/0004-637X/820/2/99

Tsiaras, A., Waldmann, I. P., Zingales, T., et al. 2018, AJ, 155, 156

van den Oord, A., Dieleman, S., Zen, H., et al. 2016, ArXiv e-prints, arXiv:1609.03499

Varley, R. 2016, Computer Physics Communications, 207, 298

Venot, O., Hébrard, E., Agúndez, M., et al. 2012, A&A, 546, A43

Vidal-Madjar, A., Lecavelier des Etangs, A., Désert, J. M., et al. 2003, Nature, 422, 143

Vidal-Madjar, A., Désert, J. M., Lecavelier des Etangs, A., et al. 2004, ApJ, 604, L69

Vidal-Madjar, A., Huitson, C. M., Bourrier, V., et al. 2013, A&A, 560, A54

Wagner, K., Apai, D., Kasper, M., et al. 2016, Science, 353, 673

Waldmann, I. P. 2012, ApJ, 747, 12

—. 2014, ApJ, 780, 23

—. 2016, ApJ, 820, 107

Waldmann, I. P., Rocchetto, M., Tinetti, G., et al. 2015a, ApJ, 813, 13

Waldmann, I. P., Tinetti, G., Deroo, P., et al. 2013, ApJ, 766, 7

Waldmann, I. P., Tinetti, G., Drossart, P., et al. 2012, ApJ, 744, 35

Waldmann, I. P., Tinetti, G., Rocchetto, M., et al. 2015b, ApJ, 802, 107

Walker, J. C. G. 1975, Journal of Atmospheric Sciences, 32, 1248

Welch, B. L. 1947, Biometrika, 34, 28

Wiedemann, G. 1996, The Messenger, 86, 24

Wiedemann, G., Delabre, B., Huster, G., Moorwood, A. F., & Sokar, B. 2000, in Optical and IR Telescope Instrumentation and Detectors, Vol. 4008, 1076–1083

Wilkins, A. N., Deming, D., Madhusudhan, N., et al. 2014, ApJ, 783, 113

Willbold, M., Elliott, T., & Moorbath, S. 2011, Nature, 477, 195

Winn, J. N. 2010, Exoplanet Transits and Occultations, 55–77

Wolszczan, A., & Frail, D. A. 1992, Nature, 355, 145

Wu, Y., & Lithwick, Y. 2013, ApJ, 763, 13

Yeredor, A. 2000, IEEE Signal Processing Letters, 7, 197

Yurchenko, S. N., Barber, R. J., & Tennyson, J. 2011, MNRAS, 413, 1828

Yurchenko, S. N., & Tennyson, J. 2014, MNRAS, 440, 1649

Zhao, M., O'Rourke, J. G., Wright, J. T., et al. 2014, ApJ, 796, 115

Zingales, T., & Waldmann, I. P. 2018, AJ, 156, 268