A&A 421, 281–294 (2004)
DOI: 10.1051/0004-6361:20040127

Astronomy & Astrophysics

Estimating stellar parameters from spectra⋆

I. Goodness-of-fit parameters and lack-of-fit test

L. Decin[1,⋆⋆], Z. Shkedy[2], G. Molenberghs[2], M. Aerts[2], and C. Aerts[1]

[1] Department of Physics and Astronomy, Institute for Astronomy, K.U. Leuven, K.U. Leuven, Celestijnenlaan 200B, 3001 Leuven, Belgium
[2] Biostatistics, Center for Statistics, Limburgs Universitair Centrum, Universitaire Campus, Building D, 3590 Diepenbeek, Belgium

Received 22 April 2003/ Accepted 18 January 2004

Abstract. Estimating stellar parameters from spectrophotometric data is a key tool in the study of stellar structure and stellar evolution. Although many methods have been proposed to estimate stellar parameters from ultraviolet (UV), optical and infrared (IR) data using low, medium or high-resolution observational data of the target(s), only a few address the problem of the uncertainties in the stellar parameters. This information is critical for a meaningful comparison of the derived parameters with results obtained from other data and/or methods. Here we present a *frequentist* method to estimate these uncertainties. We demonstrate that the combined use of *both* a *local* and a *global* goodness-of-fit parameter alters the uncertainty intervals as determined from the use of only one of these deviation estimating parameters. This technique using both goodness-of-fit parameters is applied to the infrared 2.38−4.08 μm ISO-SWS data (Infrared Space Observatory – Short Wavelength Spectrometer) of α Boo, yielding an effective temperature range from 4160 K to 4300 K, a logarithm of the gravity range from 1.35 to 1.65 dex and a metallicity from −0.30 to 0.00 dex. However, using a lack-of-fit test, it is shown that even the "best" theoretical models are still not capable of capturing all the structure in the data, and this is due to our incomplete knowledge and modelling of the full physical stellar structure or due to problems in the data reduction process.

Key words. methods: data analysis – methods: statistical – techniques: spectroscopic – stars: fundamental parameters – stars: individual: Alpha Boo

1. Introduction

Everything we know about the structure of stellar objects being studied is the result of a comparison between theoretical predictions and stellar observations. To give realistic answers to many physical questions induced by stellar phenomena, not only are accurate theoretical models indispensable, but also realistic uncertainty estimates on the parameters being deduced are required. In astronomy, uncertainty estimates (if assessed at all) are still often based on too simple a study of the sample of observables, resulting in too "optimistic" error bars. A realistic knowledge of the uncertainties is however crucial and has to be taken into account if one, for example, wants to test the proposed physical mechanism explaining certain phenomena (see, e.g., De Bruyne et al. 2003). The present paper is the first of a series devoted to the development and description of a statistical method to assess the uncertainties of stellar atmospheric parameters deduced from astronomical spectra.

One way of estimating stellar atmospheric parameters and drawing inferences from them consists of comparing the observed spectrum of the target being studied with a collection of synthetic spectra (e.g., Decin et al. 2000, for an application to cool stars; Bailer-Jones 2000, for an application using neural networks). Depending on the quality, the resolution and the wavelength coverage of the data, different stellar parameters can be traced. In this paper we focus on the three most important stellar parameters for the model structure: the effective temperature $T_{\rm eff}$, the gravity g and the metallicity [Fe/H]. Other parameters such as the abundance pattern or the microturbulence are treated as known. Let $\Omega = (T_{\rm eff}, \log g, [{\rm Fe/H}])$ present the parameters of the stellar atmosphere. A synthetic spectrum, $\theta^{(m)}$, $m = 1, 2, \ldots, M$, is calculated for specific values of the parameters, $\Omega^{(m)} = ({T_{\rm eff}}^{(m)}, \log g^{(m)}, [{\rm Fe/H}]^{(m)})$ and compared to the observed spectrum. When the synthetic spectrum and the observed spectrum agree the parameters of the stellar atmosphere are assumed to be known. The first question which arises when applying this kind of method is how to

Send offprint requests to: L. Decin,
e-mail: Leen.Decin@ster.kuleuven.ac.be

⋆ Based on observations with ISO, an ESA project with instruments funded by ESA Member States (especially the PI countries France, Germany, The Netherlands and the UK) and with the participation of ISAS and NASA.

⋆⋆ Postdoctoral Fellow of the Fund for Scientific Research, Flanders.

measure the goodness-of-fit between observational and theoretical data. A second – equally important – question is then how to assess the uncertainties on the derived stellar parameters. These questions (and answers) become even more complicated when we want to take measurement errors into account.

A search of the astronomical literature reveals that statistical tests are often restricted to *local* deviation estimating parameters, as e.g. the ordinary least square method (OLS) (or derivatives from it) (Bailer-Jones 2000; Valenti & Piskunov 1996; Erspamer & North 2002; Katz et al. 1998). The first goal of this paper is to demonstrate that parameter ranges as determined from the use of a *local* goodness-of-fit parameter can be optimized by combining a *local* with a *global* goodness-of-fit parameter (Sects. 3 and 4). In both methods, the estimate for Ω is the value of $\Omega^{(m)}$ which minimises the proposed goodness-of-fit parameter. Using the results of this first part of the study, our second goal is to discuss (Sect. 5) how the differences between observed and synthetic spectra can be used to check the appropriateness of a proposed set of stellar parameters – a step often neglected by astronomers. This qualification can be performed using *lack-of-fit* tests.

The methodology developed in this paper has broad applications in astronomy as it relies only on observed spectra and on theoretical predictions thereof. To test our method, we have chosen to apply it to the spectra of a stellar target of which the basic stellar parameters are already very well known from successful comparisons with models. Accurate estimations of stellar parameters for cool standard stars were done using data of the ISO-SWS (Decin et al. 2003a–c). We therefore have chosen to illustrate our *general* methodology on the ISO-SWS observations of one such star, the case study of the K2IIIp giant Alpha Bootis (Arcturus, HD 124897).

Before doing the analysis in Sects. 3–5, we give in Sect. 2 a description of the observational and theoretical data on which the method will be tested. The results of both Part I (Sects. 3 and 4) and Part II (Sect. 5) are discussed in Sects. 4.3 and 5.3 respectively. We end with a summary and some conclusions in the last section, Sect. 6. How to treat observational errors in this kind of study will be discussed in a forthcoming paper of this series.

2. Observational and synthetic data

This section describes the used observational ISO-SWS and theoretical data. The grid of synthetic spectra calculated for the test-case α Boo is specified in Sect. 2.3.

2.1. Observational data y

The observational data for this study consist of near-infrared (2.38−4.08 μm) spectra of α Boo observed with the SWS (Short Wavelength Spectrometer, de Graauw et al. 1996) on board ISO (Infrared Space Observatory, Kessler et al. 1996). The spectrometer was used in the SWS observing mode AOT01 (=a single up-down scan for each aperture with four possible scan speeds at degraded resolution) with scanner speed 4, resulting in a resolving power of ~1500. The observation lasted for 6538 s and was performed during revolution 452[1].

Table 1. Resolution and factors used to shift the sub-bands.

Sub-band	Wavelength range [μm]	Resolution	Factor
1A	2.38–2.60	1300	1.007
1B	2.60–3.02	1200	1.013
1D	3.02–3.52	1500	1.018
1E	3.52–4.08	1000	1.011

The reduction was made using the SWS Interactive Analysis Package IA[3] (de Graauw et al. 1996) using calibration files and procedures equivalent with pipeline version 10.0. Further data processing consisted of bad data removal (σ = 2.0), aligning of the 12 detectors to their average level. Since the grid of observational pixel values does not have a fixed resolution, we first want to "summarise" the observational pixel values, and then make a comparison between this summary (denoted as y) and a synthetic spectrum (θ) with the same resolution. The standard way to resample the input data is by "rebinning". To summarise the ISO-SWS data to a fixed resolution we have applied a flux conserving non-parametric rebinning method – i.e. for each bin the flux value is calculated using the trapezoidal rule – with an oversampling of 4. This means that the resolution bin used is 4 times the grid separation determined by the resolution for a specific wavelength range of the ISO-SWS data. To fully recover the intervening flux values it can be shown in the context of "rectangular filtering" that taking 4 points in an interval of length Δt is enough to optimise the signal-to-noise (S/N) ratio (Bracewell 1985). The rebinning used in the data reduction procedure introduces a correlation between the data point values. The appropriate resolving power was taken to be the most conservative resolving power as determined by Lorente in Leech et al. (2002) (see Table 1), with the exception being band 1A[2] for which this value has been changed from 1500 to 1300 (Decin et al. 2003b).

The individual sub-band spectra can show jumps in flux level at the band-edges when combining them into a single spectrum. These band-to-band discontinuities can have several causes: uncertainties in flux calibration, the low responsivity at the band edges, pointing errors, and a problematic dark current subtraction in combination with the RSRF (Relative Spectral Response Function) correction, from which the pointing errors are believed to have the largest impact for this high-flux observation. Hence, the individual sub-bands were

[1] Each observation is determined uniquely by its observation number (8 digits), in which the first three digits represent the revolution number. The observing data can be calculated from the revolution number which is the number of days after 17 November 1995.

[2] The bands are combinations of detector array, aperture and grating orders such that for each band its detector array sees a unique order of light, and hence a unique wavelength t. Band 1 (2.38−4.08 μm) is subdivided in 4 sub-bands: band 1A: 2.38−2.60 μm, band 1B: 2.60−3.02 μm, band 1D: 3.02−3.52 μm, and band 1E: 3.52−4.08 μm.

multiplied by a factor to construct a smooth spectrum (see Table 1). These factors were determined using the SED (Spectral Energy Distribution) of α Boo as constructed in Decin et al. (2003b) as a reference. The estimated 1σ uncertainty on these factors is 10% (Leech et al. 2002).

2.2. Synthetic data θ

The synthetic spectra used in this study have been generated using model photospheres calculated with the MARCS code, version May 1998. This version is a major update of the MARCS model-photosphere programs first developed by Gustafsson et al. (1975), and further improved by, e.g. Plez et al. (1992), Jørgensen et al. (1992), Edvardsson et al. (1993).

The common assumption of spherical stratification in homogeneous stationary layers, hydrostatic equilibrium and Local Thermodynamic Equilibrium (LTE) were made. Energy conservation was required for radiative and convective flux, where the energy transport due to convection was treated through a local mixing-length theory. The mixing-length l was chosen as $1.5\,H_{\rm p}$, with $H_{\rm p}$ the pressure scale height. Turbulent pressure was neglected. The reliability of these assumptions is discussed in Plez et al. (1992). The continuous absorption as well as the new models will be fully described in a series of forth-coming papers (Gustafsson et al.; Jørg ensen et al.; Plez et al., all in preparation).

Using the computed model atmospheres, the synthetic spectra were generated by solving the radiative transfer at a high wavelength resolution ($\Delta t \sim 1\,{\rm km\,s^{-1}}$, corresponding to $t/\Delta t \sim 330\,000$). With a microturbulent velocity $\xi_t \sim 2\,{\rm km\,s^{-1}}$, this means we are sure to sample all lines in the atomic and molecular database in the generation of the synthetic spectrum. This is necessary so as not to overestimate the absorption in regions with a high line density, or to underestimate it in regions with a low line density (Ryde & Eriksson 2002). For the line opacity in the ISO-SWS range a database of infrared lines including atoms and molecules has been prepared. For the molecular lines, the same data have been used as in Decin et al. (2000). The accuracy and completeness of these line lists are discussed in Decin (2000). For the atomic transitions, the newly generated atomic linelist of J. Sauval (priv. comm.) based on the FTS-ATMOS (Atmospheric Trace Molecule Spectroscopy) spectrum of the Sun (Farmer & Norton 1989; Geller 1992) has been included. The emergent synthetic spectra are then convolved with a Gaussian instrumental profile with the same resolution as the ISO-SWS sub-bands (see Table 1).

2.3. Synthetic data for α Boo

As described in the introduction, we will calculate a grid of synthetic spectra over discrete values in the parameter vector Ω. In the following sections, we will use this grid over the vector parameter Ω to estimate Ω with $\Omega^{(*)}$ for which the synthetic spectrum $\theta^{(*)}$ is the "closest" to the observed spectrum (see Sects. 3 and 4).

Based on the results in Decin et al. (2003a), 125 spectra have been calculated for α Boo, with parameter ranges:

$T_{\rm eff}$: 4160 K, 4230 K, 4300 K, 4370 K, 4440 K
log g : 1.20, 1.35, 1.50, 1.65, 1.80
[Fe/H] : 0.00, −0.15, −0.30, −0.50, −0.70.

The other parameters needed to compute a proper spherical symmetric spectrum – the mass, the abundances of C, N, O, Mg, and Si, the microturbulent velocity, and the $^{12}C/^{13}C$-ratio – were kept fixed, with values as determined in Decin et al. (2003a). For a detailed comparison between the stellar parameters as deduced in Decin et al. (2003a) and other literature values, we refer to Decin et al. (2003a). Each synthetic spectrum was thus calculated for a different combination of the atmospheric parameters. The fact that we are dealing with a 3-dimensional stellar parameter space made us decide to assign each theoretical model an (artificial) model number to plot all the results in one figure. The model numbers are specified in Table 2.

Since the ISO-SWS data are absolutely calibrated, one also has to compute the angular diameter to compare the rebinned observed and synthetic data properly. Therefore, the angular diameter $\theta_{\rm d}$ is deduced from the energy distribution of the synthetic spectrum between 2.38 and 4.08 μm and the absolute flux-values in this wavelength range of the ISO-SWS spectrum. We therefore have minimised the residual sum of squares

$$\sum_{t=1}^{n}\left(y(t)-\left(\frac{\pi}{4}\,{\theta_{\rm d}}^2\right)*\theta(t)\right)^2, \qquad (1)$$

with $y(t)$ and $\theta(t)$ representing, respectively, a (rebinned) observational and synthetic data point at the tth wavelength point. The derived angular diameters are listed in Table 2. The dependence of the angular diameter on the effective temperature is discussed in Decin (2000). A typical example of the ISO-SWS data of α Boo and a synthetic spectrum (model 1: $T_{\rm eff} = 4160$ K, log $g = 1.20$, and [Fe/H] $= -0.70$) is shown in Fig. 1.

3. Model selection based on the OLS criterion

We try for the first time in this paper to determine the parameter ranges of the effective temperature, the gravity and the metallicity of α Boo. We first describe in general the model selection based on the ordinary least square criterion (Sect. 3.1). This is followed by the application to our test-case, the ISO-SWS data of α Boo (Sect. 3.2).

3.1. Definition

For the high-quality (absolutely calibrated) spectroscopic data that we use, it is natural to estimate Ω with $\Omega^{(*)}$ for which the synthetic spectrum $\theta^{(*)}$ gives the best resemblance to the observed spectrum, y. By analogy to linear regression we can estimate Ω by minimising the residual sum of squares (also called the "ordinary least square" (OLS) method):

$$T^2(y,\theta^{(m)}) = \frac{1}{n}\sum_{t=1}^{n}\left(y(t)-\left(\frac{\pi}{4}\,{\theta_{\rm d}}^2\right)*\theta^{(m)}(t)\right)^2. \qquad (2)$$

Table 2. Angular diameters in mas and model numbers (in between brackets) associated with the different model parameters of the grid of synthetic spectra.

log g	$T_{\rm eff}$ [K]										
	4160		4230		4300		4370		4440		
log g = 1.20	21.16	(1)	20.95	(26)	20.72	(51)	20.51	(76)	20.28	(101)	
log g = 1.35	21.20	(6)	20.99	(31)	20.76	(56)	20.54	(81)	20.31	(106)	
log g = 1.50	21.24	(11)	21.02	(36)	20.79	(61)	20.57	(86)	20.34	(111)	[Fe/H] = −0.70
log g = 1.65	21.27	(16)	21.04	(41)	20.86	(66)	20.59	(91)	20.36	(116)	
log g = 1.80	21.29	(21)	21.06	(46)	20.99	(71)	20.61	(96)	20.38	(121)	
log g = 1.20	21.16	(2)	20.95	(27)	20.73	(52)	20.51	(77)	20.29	(102)	
log g = 1.35	21.21	(7)	21.00	(32)	20.76	(57)	20.54	(82)	20.31	(107)	
log g = 1.50	21.25	(12)	21.02	(37)	20.79	(62)	20.57	(87)	20.36	(112)	[Fe/H] = −0.50
log g = 1.65	21.26	(17)	21.04	(42)	20.82	(67)	20.59	(92)	20.37	(117)	
log g = 1.80	21.28	(22)	21.06	(47)	20.83	(72)	20.61	(97)	20.41	(122)	
log g = 1.20	21.17	(3)	20.96	(28)	20.74	(53)	20.51	(78)	20.29	(103)	
log g = 1.35	21.21	(8)	21.00	(33)	20.77	(58)	20.55	(83)	20.32	(108)	
log g = 1.50	21.24	(13)	21.02	(38)	20.79	(63)	20.57	(88)	20.35	(113)	[Fe/H] = −0.30
log g = 1.65	21.26	(18)	21.04	(43)	20.81	(68)	20.60	(93)	20.37	(118)	
log g = 1.80	21.28	(23)	21.06	(48)	20.84	(73)	20.78	(98)	20.40	(123)	
log g = 1.20	21.17	(4)	20.96	(29)	20.74	(54)	20.52	(79)	20.29	(104)	
log g = 1.35	21.21	(9)	21.00	(34)	20.77	(59)	20.55	(84)	20.32	(109)	
log g = 1.50	21.24	(14)	21.02	(39)	20.79	(64)	20.58	(89)	20.35	(114)	[Fe/H] = −0.15
log g = 1.65	21.26	(19)	21.04	(44)	20.82	(69)	20.61	(94)	20.38	(119)	
log g = 1.80	21.28	(24)	21.06	(49)	20.84	(74)	20.63	(99)	20.41	(124)	
log g = 1.20	21.17	(5)	20.96	(30)	20.74	(55)	20.52	(80)	20.28	(105)	
log g = 1.35	21.21	(10)	21.00	(35)	20.77	(60)	20.55	(85)	20.33	(110)	
log g = 1.50	21.23	(15)	21.02	(40)	20.79	(65)	20.58	(90)	20.36	(115)	[Fe/H] = 0.00
log g = 1.65	21.25	(20)	21.04	(45)	20.82	(70)	20.61	(95)	20.38	(120)	
log g = 1.80	21.28	(25)	21.07	(50)	20.84	(75)	20.63	(100)	20.41	(125)	

Hence, the minimiser of $T(y, \theta^{(m)})$ should be seen as the OLS estimator for Ω. In practice one can minimise Eq. (2) with a search over a sensitive grid of the parameter vector Ω.

3.2. Application to α Boo

3.2.1. Band 1A

Figure 2 shows the values of $T^{(m)}(y, \theta^{(m)})$ (on a log scale). The vertical lines separate the 125 models by $T_{\rm eff}$.

Model 62 ($T_{\rm eff}$ = 4300 K, log g = 1.50 dex, [Fe/H] = −0.50 dex) has the best goodness-of-fit in band 1A. We note that within one temperature level, the models occur in groups of size 5 (according to the value of the gravity). For example, models 1–5 have the same effective temperature, 4160 K, and the same gravity (log g = 1.2) while for models 6−10 log g = 1.35. Trends in the goodness-of-fit are visualised in Fig. 3. Three patterns are observed: (a) the goodness-of-fit increases with the level of metallicity, (b) a parabolic shape in which the best goodness-of-fit is achieved for models with metallicity between −0.15 to −0.5, and (c) goodness-of-fit decreases with the level of metallicity. For a fixed temperature, the trend changes from trend (a) via trend (b) to trend (c) when the gravity increases. Sometimes, a trend occurs twice or is absent, but the order of trends never changes. The model having the best goodness-of-fit is always situated at the minimum of a parabolic shape, suggesting that we have reached a local minimum – an equilibrium – in the parameter space.

Table 3 shows the 10 models with $T^{(m)}(y, \theta^{(m)})$ having the lowest values, i.e., they have the best goodness-of-fit.

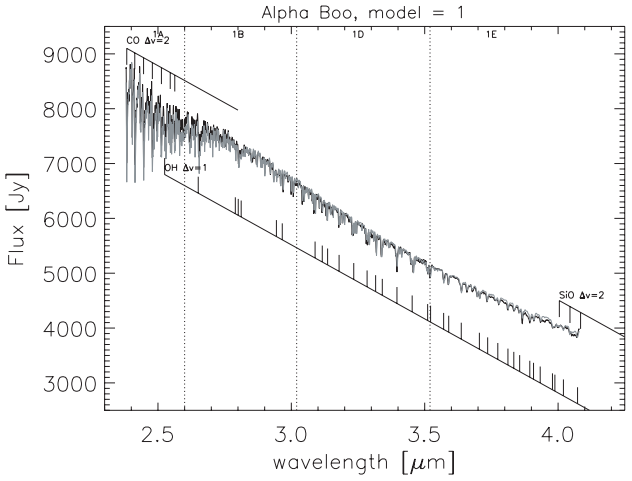

Fig. 1. Comparison between the ISO-SWS data of α Boo (black) and the synthetic spectrum with model number 1, i.e. $T_{\rm eff} = 4160$ K, $\log g = 1.20$, and [Fe/H] $= -0.70$ (grey) in the 4 sub-bands 1A, 1B, 1D, and 1E. The main absorption features caused by the CO 1st overtone lines, the SiO 1st overtone lines, and the OH fundamental lines are indicated.

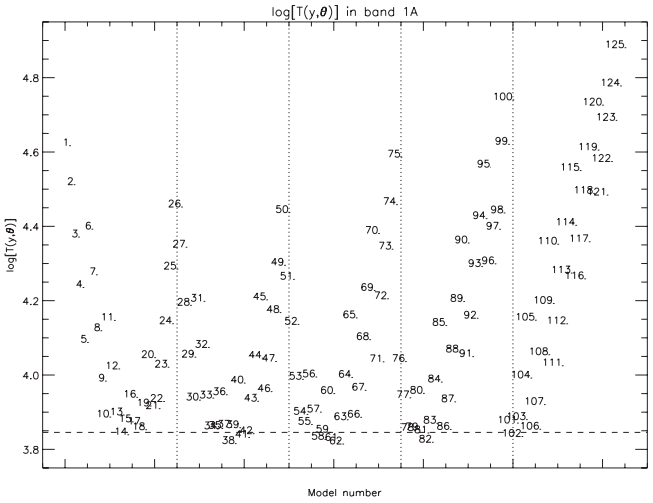

Fig. 2. $\log T^{(m)}(y,\theta^{(m)})$ versus the model numbers for band 1A. The vertical lines separate the models according to the temperature. The 10 models with the best goodness-off-fit are situated below the horizontal dashed line.

For these models $\log g$ is between 1.20 and 1.65 dex, the effective temperature ranges from 4160 K to 4440 K and [Fe/H] is between -0.15 dex and -0.70 dex.

3.2.2. Overall goodness-of-fit

While in the previous subsection we have concentrated on band 1A of the ISO-SWS data of α Boo, we now will take the whole 2.38−4.08 μm wavelength range into account. Band 1A has a very characteristic footprint, determined by the first overtone CO ($\Delta v = 2$) vibration-rotation bands in this wavelength range (Decin et al. 2000). Molecules absorbing in bands 1B, 1D, and 1E are mainly OH and SiO, while also some atomic features are visible. The absorption pattern of these last molecules is however not as pronounced as for CO ($\Delta v = 2$)

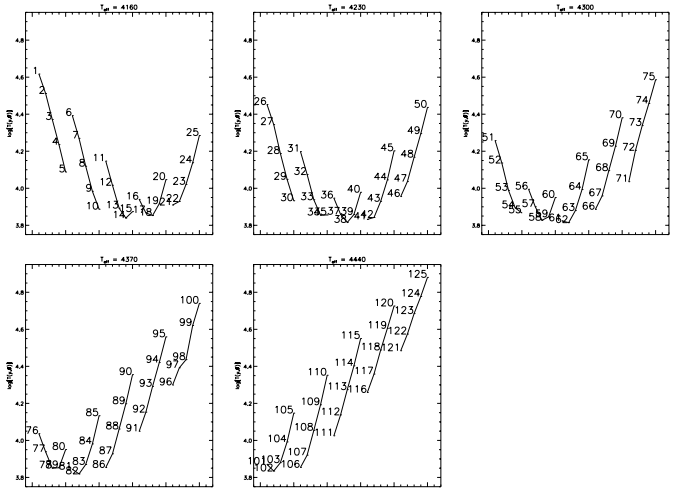

Fig. 3. Trends in the goodness-of-fit condition of $\log T^{(m)}(y,\theta^{(m)})$ for band 1A. The model numbers are as specified in Table 2.

Table 3. Top 10 models in band 1A.

Model (rank)	$T_{\rm eff}$ [K]	$\log g$	[Fe/H]
62 (1)	4300	1.50	−0.50
38 (2)	4230	1.50	−0.30
82 (3)	4370	1.35	−0.50
61 (4)	4300	1.50	−0.70
58 (5)	4300	1.35	−0.30
41 (6)	4230	1.65	−0.70
102 (7)	4440	1.20	−0.50
14 (8)	4160	1.50	−0.15
42 (9)	4230	1.65	−0.50
81 (10)	4370	1.35	−0.70

in band 1A. Although these CO features can already give us quite a good idea of the temperature and the gravity of the target, it is essential to use the whole 2.38−4.08 μm wavelength in order to minimise the uncertainties on the stellar parameters being studied. This is due to the fact that all of these molecular and atomic absorption features have their own characteristic dependence on the atmospheric parameters (see Decin et al. 2000). Since, however, each sub-band has its own instrumental characteristics, and since the observational data have their largest uncertainties at the band edges (Leech et al. 2002), we will *not* join the whole 2.38−4.08 μm wavelength range into 1 spectrum, but we will combine the results obtained from the separate bands.

The values of $T^{(m)}(y,\theta^{(m)})$ were ranked at each band, and for each model we have calculated the mean of the ranks. This means that the "best" model is the one with the smallest mean rank. For example, model 38 ($T_{\rm eff} = 4230$ K, $\log g = 1.50$ dex, [Fe/H] $= -0.30$ dex) has a rank of 2 in band 1A, but this model is ranked 9, 28, 25 in band 1B, 1D and 1E respectively (hence, the mean rank is 13.50). Overall, the rank of the mean rank of model 38 is 5. The ranks of the mean ranks of the 125 models are displayed in Fig. 4. The models with the lowest rank are models 62 and 82. Model 62 is ranked 1, 7, 8, and 4 in the 4 bands, respectively, with mean rank being 5.0.

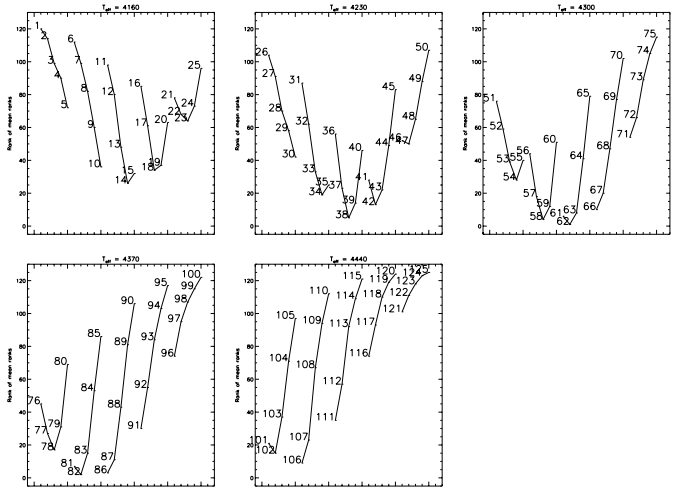

Fig. 4. Rank of the mean ranks for the 125 synthetic spectrum.

Table 4. Stellar parameters of the models having the lowest overall rank. The overall rank is given in between brackets in the first column. In the last column, the rank in band 1A is tabulated.

Model (rank)	$T_{\rm eff}$ [K]	log g	[Fe/H]	Rank in band 1A
62 (1)	4300	1.50	−0.50	1
82 (2)	4370	1.35	−0.50	2
86 (3)	4370	1.50	−0.70	15
58 (4)	4300	1.35	−0.30	5
38 (5)	4230	1.50	−0.30	2
61 (6)	4300	1.50	−0.70	4
81 (7)	4370	1.35	−0.70	10
63 (8)	4300	1.50	−0.30	26
106 (9)	4440	1.35	−0.70	16
66 (10)	4300	1.65	−0.70	28

3.2.3. Conclusions

The models having the best rank of the mean rank are listed in Table 4. Based on the models which rank in the top 10, the range in effective temperature is between 4230 K and 4440 K, in gravity between 1.35 and 1.65, and in metallicity between −0.70 to −0.30 dex.

4. Model selection based on Kolmogorov-Smirnov statistics

In the previous section, $T(y,\theta)$ was used as a measure for the goodness-of-fit. In this section the analysis discussed above was repeated using the Kolmogorov-Smirnov (β) statistics as a measure for the goodness-of-fit.

4.1. Definition

The Kolmogorov-Smirnov statistical test *globally* checks the goodness-of-fit of the observed and synthetic spectra by computing a deviation estimating parameter β (see Eq. (5) in Decin et al. 2000). Without specifying the distribution function of β, we may summarise that

$$\beta = \sqrt{n} \sup_{1\leq k\leq n-1} \left| \frac{\sum_{t=1}^{k} \frac{y(t)}{\theta(t)}}{\sum_{t=1}^{n} \frac{y(t)}{\theta(t)}} - \frac{k}{n} \right| . \quad (3)$$

The lower the β-value, the better the accordance between the observed data and the synthetic spectrum. For more details about the use of Kolmogorov-Smirnov statistics to estimate stellar parameters and their uncertainties see Decin et al. (2000). Hence, the main difference between β and T is that the Kolmogorov-Smirnov parameter β measures a *global* goodness-of-fit, so that local deviations between observations and theoretical data only have a minor influence on the final result, while for $T(y,\theta)$ *local* deviating points are important. Note that a shift in the absolute flux values (e.g. to simulate a change or uncertainty in the angular diameter) influences T a lot, while β remains almost the same.

4.2. Application to α Boo

Since both deviation estimating parameters stress a different point in the goodness-of-fit, a combination of the results based on the two parameters separately can only improve our knowledge on the stellar parameters and their uncertainties. This is illustrated in Fig. 5.

Table 5 shows the best 5 models, which besides appearing in the lower left corner of Fig. 5a also are ranked among the top 30 for both β and $T(y,\theta)$ (see Fig. 5b and Table 4). The combined use of both the scores of β and $T(y,\theta)$ themselves and the ranking of these scores does allow us to determine a set of "best" models! While e.g. model 54 ($T_{\rm eff} = 4300$ K, log g = 1.20 dex, [Fe/H] = −0.15 dex) has very low ranks based on both β and $T(y,\theta)$, the mean β-value is rather high. The advantage of using ranks is that all deviation estimating parameters can be treated in the same magnitude level. The disadvantage of using the ranks is that a sudden increase in the deviation estimating parameter is not translated into a sudden jump in the rank. Using both diagnostics together solves this problem.

Note that one can see a correlation between the ranks of the mean ranks of the T and β parameter, but that there are a few outliers namely in the upper left corner of Fig. 5b where a few models are situated with low T and high β. Inspecting why the deviation estimating parameters do show this trend, shows us that all of these models have a very low rank in T for band 1D and/or band 1E. When zooming into these bands, one indeed sees a resemblance (see e.g. Fig. 6a for model 77: $T_{\rm eff} = 4370$ K, log g = 1.20 dex, [Fe/H] = −0.50 dex), explaining the low T-value. The ratio between the observational and synthetic data does, however, show a trend and is not randomly distributed around 1, explaining the high β value. This is illustrated by the gray line in Fig. 6b with slope −0.02. The few very strong (negative) peaks are due to the underestimation of the OH-lines. Note also that the model having the lowest rank in β (model 40: $T_{\rm eff} = 4230$ K, log g = 1.50 dex, [Fe/H] = 0.00 dex) is only ranked 46 in T. This illustrates once more that a

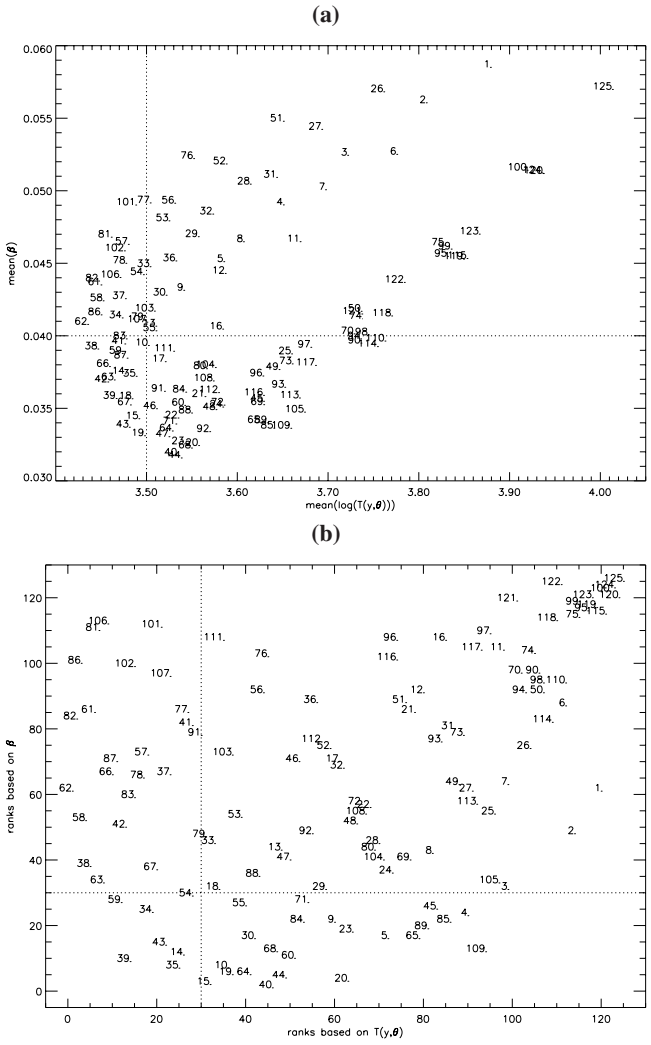

Fig. 5. Panel **a)** mean of the β-values of the 4 sub-bands versus the mean of $T(y,\theta)$ for the 4 sub-bands. Panel **b)** ranks of the mean ranks based on β versus ranks of the mean ranks based on $T(y,\theta)$.

combination of a *local* and *global* deviation estimating parameter enlarges our knowledge on the estimated stellar parameters and their uncertainties, which in this particular case (see Table 5) results in a range in $T_{\rm eff}$ from 4160 K to 4300 K, in log g from 1.35 dex to 1.65 dex, and in [Fe/H] from −0.30 to 0.00 dex. Note that while the local goodness-of-fit parameter T favours the lower range in metallicity, the combined use of T and β clearly indicates a higher range in metallicity.

4.3. Discussion on the estimated parameter ranges

It is important to compare the derived parameter ranges with literature values. We therefore will use Table D.3 as published in Decin et al. (2003a), in which a comprehensive list of derived and assumed parameter values published in the literature is given. These listed parameters have already been compared with the results as deduced from the ISO-SWS data in Sect. 3.5.2 in Decin et al. (2003a). However, the uncertainties as given in Decin et al. (2003a) were *empirical* values estimated from (1) the intrinsic uncertainty on the synthetic spectrum (i.e., the possibility to distinguish different synthetic spectra at a specific resolution, i.e. there should be a significant difference in β-values) which is thus dependent on both the resolving power of the observation and the specific values of the fundamental parameters; (2) the uncertainty on the ISO-SWS spectrum which is directly related to the quality of the ISO-SWS observation; (3) the value of the β-parameters in the KS-test; and (4) the still remaining discrepancies between observed and synthetic spectra. Decin et al. (2003a) obtained $T_{\rm eff} = 4320 \pm 140$ K, log $g = 1.50 \pm 0.15$, and [Fe/H] $= -0.50 \pm 0.20$. Comparing these results with the ones as given in the previous section, we see that a combination of both a global and a local deviation estimating parameter restricts the uncertainty ranges for these 3 fundamental parameters in the case of the ISO-SWS data of α Boo.

Table 5. Overall goodness-of-fit.The 5 models given in this table do belong to the group of "best" models based on both the values of $T(y,\theta)$ and β, and are moreover ranked among the top 30 for both $T(y,\theta)$ and β.

Model	Rank β	Rank $T(y,\theta)$	$T_{\rm eff}$ [K]	log g	[Fe/H]
39	9	14	4230	1.50	−0.15
43	14	22	4230	1.65	−0.30
35	7.5	25	4230	1.35	0.00
14	11	26	4160	1.50	−0.15
59	27	12	4300	1.35	−0.15

Combining the results given in Table 5 with the angular diameter values tabulated in Table 2, we obtain for the angular diameter $20.77 < \theta_{\rm d} < 21.24$ mas, mainly resulting from the uncertainty in the effective temperature (compared to the gravity and the metallicity). However, from Eq. (2) in Decin et al. (2003b), one can see that in the case of the ISO-SWS data the main uncertainty on the angular diameter is determined from the uncertainty in the absolute flux level (being ≈10%, see Sect. 2.1). Taking this last uncertainty into account (as done in Eq. (2) in Decin et al. 2003b), we obtain $\theta_{\rm d} = 21.01 \pm 1.24$ mas.

Uncertainty intervals as listed in Table D.3 in Decin et al. (2003a) are often intrinsic uncertainties or (sometimes) have been propagated from uncertainties in other parameters. However, as in this analysis, observational measurement uncertainties are never taken into account. As commented on in Decin et al. (2003a), we do see that the derived parameters from the ISO-SWS data are in good agreement with other listed values, but it should be noted that our uncertainty on the metallicity is quite large compared to other results. Several causes can be reported for this larger uncertainty range: (1) the used grid is not sensitive enough in the metallicity, and we should diminish the spacing in metallicity; (2) the used low-resolution ISO-SWS data are not that well suited to derive the metallicity; (3) the lack of a proper uncertainty estimate significantly underestimates the derived uncertainty ranges found in other studies.

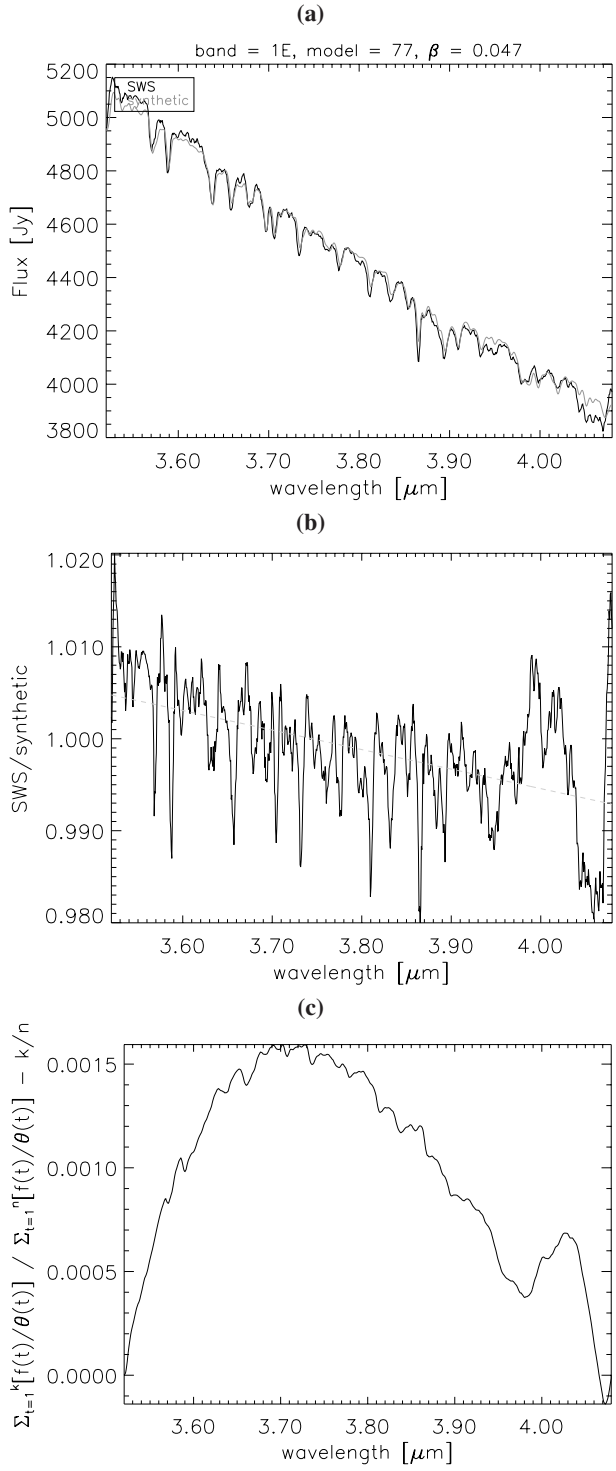

Fig. 6. ISO-SWS observations of α Boo in band 1E versus the synthetic data of model 77 ($T_{\rm eff} = 4370$ K, $\log g = 1.20$ dex, [Fe/H] = −0.50 dex). Panel **a)** ISO-SWS data are plotted in black, synthetic data in grey. Panel **b)** the ratio between observational and synthetic data. Panel **c)** illustration of how the β parameter is calculated: the argument of the absolute value in the right-hand side of Eq. (3) is displayed as a function of the wavelength.

5. Lack-of-fit tests

Within the classical regression framework, estimation is usually followed by inference and model diagnostics. In the previous sections we focused on model selection and estimation of the stellar parameters. In this section we propose a tool for model diagnostics. The term model "diagnostics" refers to any technique that offers evidence of whether a particular model is an adequate description of the data or not. Our main argument is that choosing the "best" model (out of a grid) does not necessarily imply that the model is a "good" representation of the observed data. Hence, when a model is chosen, one can investigate how well the model fits the data. We focus on the variable

$$V_t = \frac{y(t)}{\theta(t)}.$$

Note that V_t was used to construct the Kolmogorov-Smirnov statistics in the previous section. Now, if a specific synthetic spectrum is a "good" model, then we expect that $V_t \approx 1$. The aim of this section is to investigate the behaviour of V_t locally rather than globally. Note that if the synthetic spectrum is a good model, we expect that a non-parametric smoother (see details in the appendix) of V_t will be flat around 1. Contrasting a hypothesized parametric model to a non-parametric model is a key aspect in an omnibus lack-of-fit test.

Lack-of-fit tests are an attractive tool for model diagnostics since they allow us to asses the goodness-of-fit of a proposed model in a formal way. For a comprehensive discussion on lack-of-fit tests we refer to Hart (1997). In our setting, we wish to test the null hypothesis $H_0 : E(V_t) = 1$ against the alternative hypothesis that $H_0 : E(V_t) = \eta(t)$ where $\eta(t)$ is a smooth function which is not necessarily constant along the wavelength. For the remainder of this section we first formulate the hypotheses to be tested in more detail, review the test procedure proposed by Bowman & Azzalini (1997) and apply these methods to our setting.

5.1. Test of hypothesis for the "no effect" model

Let $y(t)$ be the rebinned data at wavelength t and let $\mu(t)$ be the "true" spectrum at the same wavelength. We assume that $\mu(t)$ is determined by the values of the stellar parameters $\Omega = (T_{\rm eff}$, $\log g$, [Fe/H]) and we consider the model

$$y(t) = \mu(t) + \varepsilon(t), \tag{4}$$

with $\varepsilon(t)$ being the observational error on the observed spectrum. We further assume that $E(\varepsilon(t)) = 0$. Since the parameter of interest is Ω, we wish to test the hypotheses

$$H_0 : \Omega = \Omega_0,$$
$$H_1 : \Omega \neq \Omega_0, \tag{5}$$

with Ω_0 representing the null stellar parameters of the target being studied. The hypotheses in Eq. (5) can be reformulated in terms of the synthetic spectrum,

$$H_0 : \mu(t) = \theta_0(t), \quad \text{for all } t$$
$$H_1 : \mu(t) \neq \theta_0(t), \quad \text{for some } t, \tag{6}$$

where $\theta_0(t)$ is the null synthetic spectrum corresponding to the stellar parameters Ω_0. In terms of the observed spectrum the hypotheses in Eq. (6) can be rewritten as

$$H_0 : E(y(t)) = \theta_0(t), \quad \text{for all } t$$
$$H_1 : E(y(t)) \neq \theta_0(t), \quad \text{for some } t. \tag{7}$$

Under the null hypothesis in Eq. (7) we expect that $E(V_t) = 1$. Thus, in terms of V_t, we consider two competing models,

$$H_0 : E(V_t) = 1,$$
$$H_1 : E(V_t) = \eta(t) \text{ with } \eta(t) \neq 1, \quad \text{for some } t. \tag{8}$$

Here, $\eta(t)$ is assumed to be a smooth function. The model under H_0 is called the "no effect" model (see e.g., Hart 1997, p. 148).

To test the hypotheses in Eq. (8), we used the procedure as described by Bowman & Azzalini (1997). One therefore needs to calculate the residual sum of squares under the two hypotheses and compare them. Since the mean of V_t under H_0 is constant, the residual sum of squares under the null hypothesis is

$$RSS_0 = \Sigma_{t=1}^{n} \{V_t - 1\}^2 , \tag{9}$$

and under H_1

$$RSS_1 = \Sigma_{t=1}^{n} \{V_t - \hat{\eta}(t)\}^2 , \tag{10}$$

where $\hat{\eta}(t)$ is a linear smoother of V_t. Note that we do not specify any parametric structure for $\eta(t)$ under the alternative in Eq. (8). The underling assumption that we made is that if a specific synthetic model it not a "good" model, there is a structure in the rebinned data that this specific model cannot capture. This structure can be captured by the non-parametric smooth function $\hat{\eta}(t)$. In practice, we use the Loess method, e.g. see Cleveland (1979) and Chambers & Hastie (1992), to model the relationship between V_t and the wavelength. More details about the Loess method is given in the appendix. Intuitively, it is clear that for a "good" synthetic spectrum RSS_0 and RSS_1 have close values. Therefore we will not reject the null hypothesis if RSS_0 is sufficiently close to RSS_1. Formally, the test statistics which quantifies the difference between the residual sum of squares is given by

$$F = \frac{RSS_0 - RSS_1}{RSS_1}. \tag{11}$$

Note that if H_0 is correct we expect that F will be small. Hence, we reject the null hypothesis for a large value of F.

To proceed further we need to find the distribution of F under the null hypothesis. This can be done using a bootstrap procedure (Davison & Hinkley 1997) which we describe in more detail in the appendix. Briefly, the bootstrap procedure we applied consists of sampling B samples from the original sample while reflecting the null hypothesis. For each bootstrap sample we calculate the value of F. The empirical p-value of the test statistics is simply the proportion of the bootstrap statistics that is larger than the one observed in the original sample. For a given significance level α, one cannot reject the null hypothesis if the p-value $> \alpha$.

To calculate the residual sum of squares under H_1 we smooth V_t with Loess. Thus, the distribution of the test statistics in Eq. (11) depends on the choice of the smoothing parameter. As argued by Hart (1997), the smoothing parameter should be chosen in advance and should be fixed for all bootstrap samples. Therefore, our conclusion of whether to reject the null hypothesis or not depends on our choice of the smoothing parameter. A method to overcome this problem is to calculate the so-called "significance trace" (Hart 1997 p. 160; Bowman & Azzalini 1997 p. 89). In this method, one computes the p-value for a wide range of smoothing parameters and the decision (whether to reject H_0 or not) is based on the significance trace plot. This point will be illustrated in the following section (Sect. 5.2.1).

Table 6. Empirical p-values in band 1A. The first column gives the model number, and the second column the rank of the corresponding model determined from the β-value in band 1A. Empirical p-values based on a bootstrap with $B = 1000$ and smoothing parameter $\gamma = 0.85$ are given in the third and fourth column: the third column shows the empirical p-values calculated under the null hypothesis in Eq. (8), and the fourth column shows the empirical p-values under the null hypothesis in Eq. (12). The last column gives the bias as determined from Eq. (14).

Model	Rank (β)	p-value ($\gamma = 0.85$) $H_0 : E(V_t) = 1$	p-value ($\gamma = 0.85$) $H_0 : E(V_t) = \mu$	Bias under H_0 (%)
51	1	0	0.011	1.361
26	2	0	0.004	1.880
52	3	0	0.000	1.072
27	4	0	0.001	1.590
76	5	0	0.002	0.819
28	6	0	0.000	1.199
53	7	0	0.000	0.708
77	8	0	0.000	0.571
29	9	0	0.000	0.850
56	10	0	0.000	0.725
54	11	0	0.000	0.399
30	12	0	0.000	0.484
101	13	0	0.000	0.311
1	14	0	0.002	2.379
2	15	0	0.001	2.054
31	16	0	0.000	1.209
3	17	0	0.000	1.658
4	18	0	0.000	1.306
5	19	0	0.000	0.921
78	20	0	0.000	0.227

In addition to the hypotheses in Eq. (8) we test the following hypotheses

$$H_0 : E(V_t) = \mu \quad \text{(any constant)},$$
$$H_1 : E(V_t) = \eta(t). \tag{12}$$

The null hypothesis in Eq. (12) states that the mean of V_t is constant, but not necessarily equal to 1. Under H_0 in Eq. (12) the residual sum of squares is

$$RSS_0 = \Sigma_{t=1}^{n} \left\{V_t - \bar{V}_t\right\}^2 , \tag{13}$$

with $\bar{V}_t$ indicating the mean of V_t. Note that if we reject the null hypothesis in Eq. (12) the null hypothesis in Eq. (8) will be rejected as well but not vice versa.

5.2. Application to the data

5.2.1. Band 1A

Table 6 presents the results for the lack-of-fit tests for the top 20 models in band 1A. For each synthetic spectrum 1000 bootstrap samples ($B = 1000$) were drawn from

the original sample as described in the appendix. Whenever the empirical p-value is greater than 0.05 the null hypothesis cannot be rejected. This means that the relationship between V_t and t is assumed to be constant for all models with p-value greater than 0.05.

(1) Testing $H_0 : E(V_t) = 1$:

The third column in Table 6 presents the empirical p-values calculated under the null hypothesis in Eq. (8). Using a smoothing parameter $\gamma = 0.85$ for the Loess model, the empirical p-values are all 0. Thus, we reject H_0 in Eq. (8) for all models.

(2) Testing $H_0 : E(V_t) = \mu$:

The bias in the last column in Table 6 is defined as

$$\text{bias} = (\bar{V}_t - 1) \times 100, \quad (14)$$

where $\bar{V}_t$ was estimated under the null hypothesis in Eq. (12). Thus, a good synthetic model for the spectrum is one with empirical p-value greater than 0.05 (hence, constant relationship between V_t and t) and small bias (hence, the constant is closed to 1). When the empirical p-value was calculated under H_0 in Eq. (12) (with $\gamma = 0.85$) the null hypothesis is rejected for all models. Figure 7 shows the plot of V_t with Loess smoothers (with several values for the smoothing parameter) for model 51 ($T_{\rm eff} = 4300$ K, $\log g = 1.20$ dex, [Fe/H] $= -0.70$ dex). Note that, for $\gamma = 0.85$ (the value that was used to calculate the empirical p-value in Table 6) the Loess model is quite flat, but lies above 1. This means that, in general, the values of the rebinned observational data are greater than the values of the synthetic spectrum along the wavelength. Figure 8 shows similar patterns for model 26 ($T_{\rm eff} = 4230$ K, $\log g = 1.20$ dex, [Fe/H] $= -0.70$ dex). Model 52 ($T_{\rm eff} = 4300$ K, $\log g = 1.20$ dex, [Fe/H] $= -0.50$ dex) is shown in Fig. 9. Figure 10 shows the results for model 125 ($T_{\rm eff} = 4440$ K, $\log g = 1.80$ dex, [Fe/H] $= 0.00$ dex), which fit the data poorly according to the least squares criterion. Note how the Loess smoother is always below 1 and suggest an increasing trend with the wavelength.

We turn now to the discussion on the effect of smoothing parameters on the estimation procedure which depends on the choice of the smoothing parameter of the Loess. To be able to calculate the p-value one needs to construct the null distribution of the test statistics. This can be done only if the smoothing parameter is held fixed for each bootstrap replication (e.g., see Hart 1997, Sect. 6.4). King et al. (1991) proposed to compute the p-values corresponding to several different choices of the smoothing parameter. The plot in which p-values are plotted versus the smoothing parameter is called a "significance trace". Figure 11 shows the significance trace plot for model 51 for the null hypothesis $E(V_t) = \mu$. For all values of γ the null hypothesis is rejected (the p-value is below the horizontal line of 0.05). This means that the data do not support the null hypothesis. The same conclusion can be drawn for all other models. The fact that, for all models, the significance trace for the null hypothesis $E(V_t) = \mu$ is below the 0.05 line regardless of the choice of the smoothing parameter means that the null hypothesis is rejected for all possible values of the smoothing parameter.

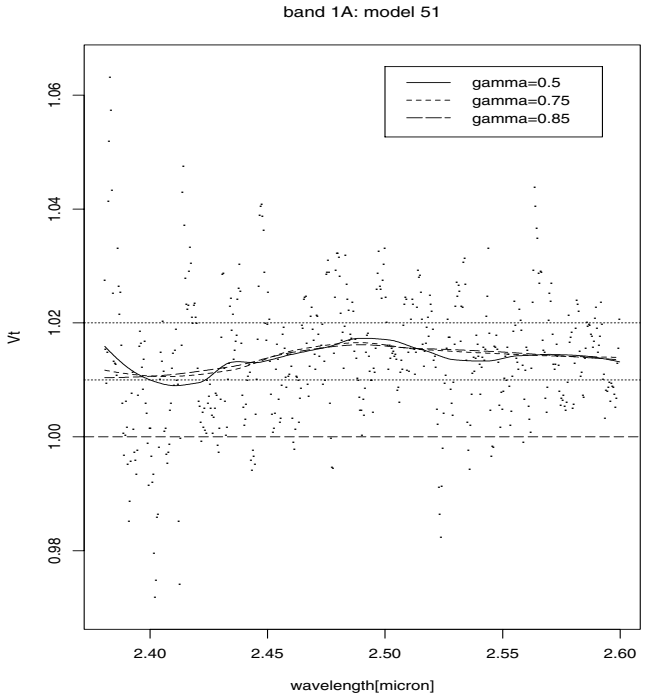

Fig. 7. Band 1A: model 51 ($T_{\rm eff} = 4300$ K, $\log g = 1.20$ dex, [Fe/H] $= -0.70$ dex). V_t and the Loess smoother with three values of γ.

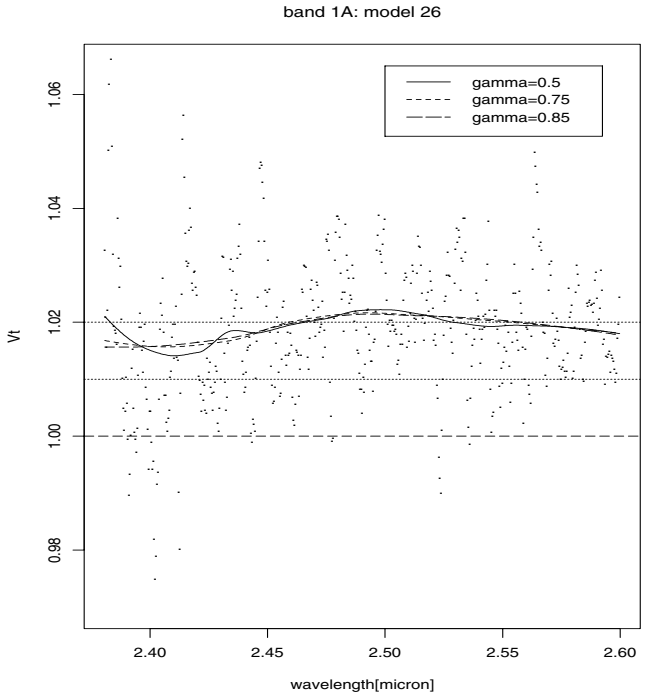

Fig. 8. Band 1A: model 26 ($T_{\rm eff} = 4230$ K, $\log g = 1.20$ dex, [Fe/H] $= -0.70$ dex). V_t and the Loess smoother with three values of γ.

5.2.2. Bands 1B, 1D, and 1E

The results in bands 1B, 1D, and 1E are similar: the empirical p-values for all model were either zero or very close to zero.

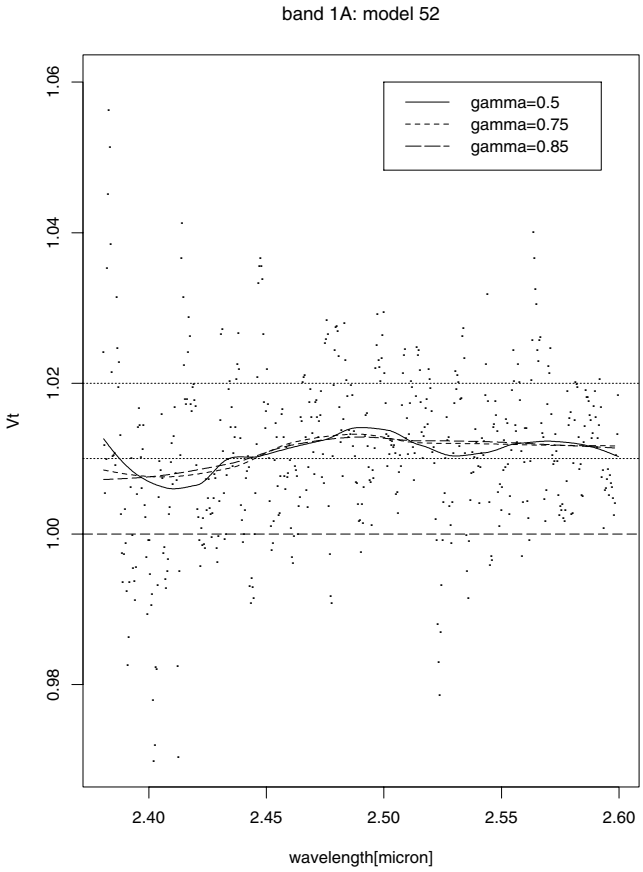

Fig. 9. Band 1A: model 52 ($T_{\rm eff}$ = 4300 K, log g = 1.20 dex, [Fe/H] = −0.50 dex). V_t and the Loess smoother with three values of γ.

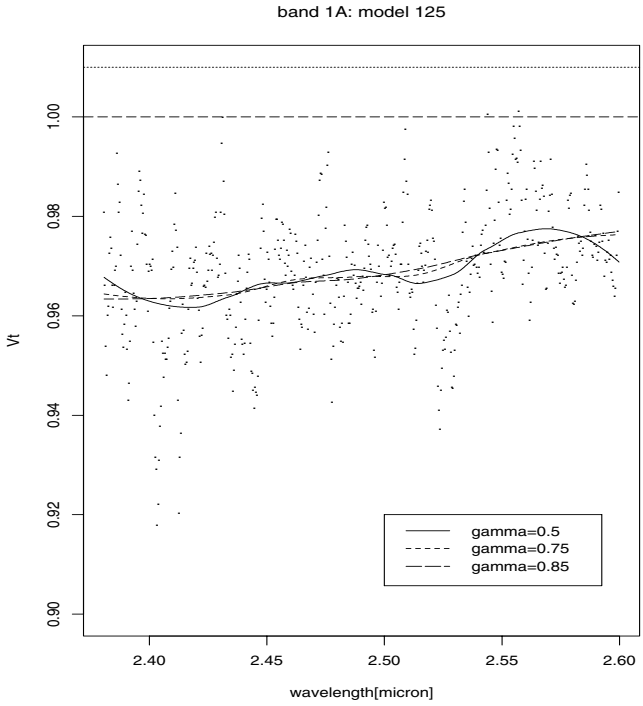

Fig. 10. Band 1A: model 125 ($T_{\rm eff}$ = 4440 K, log g = 1.80 dex, [Fe/H] = 0.00 dex). V_t and the Loess smoother with three values of γ.

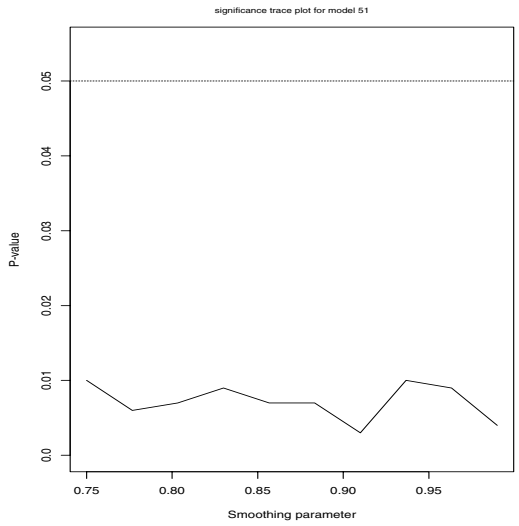

Fig. 11. Significance trace for model 51 for the null hypothesis (H_0 : $V_t = \mu$) in band 1A. The dotted horizontal line represents a significance level of 0.05. Whenever the significance trace plot is above the dotted line, the null hypothesis cannot be rejected.

Hence, the null hypothesis in Eq. (8) was rejected for all models in the three bands which indicates that the synthetic spectra do not follow the same pattern as the rebinned observational data.

5.3. Discussion

What are the lessons learned from the rejection of the null hypothesis in so many cases? It may be clear that this failure cannot be solved by relaxing the criteria, e.g. by lowering the level of significance α. These lack-of-fit tests are an objective tool to demonstrate that there is still too much structure left in V_t. This is illustrated, e.g., in Figs. 12−15 where model 39 with a very good goodness-of-fit is depicted in bands 1A, 1B, 1D, and 1E. As can be seen from the upper panel in Fig. 12 the low-excitation ^{12}CO lines are often predicted as being too strong, while it is clearly visible in the upper panel in Fig. 13 that the low-excitation OH-lines are often predicted as being too weak. This *systematic* discrepancy between observations and theory is captured in V_t and its Loess smoother, explaining why the lack-of-fit test rejects the null hypothesis. This systematic problem is *not* solved by one of the other models in the grid. Neither it is possible to solve this problem by reducing the carbon abundance ε(C) or enhancing the oxygen abundance ε(O), since then other molecular features are mispredicted. The described problem may be an outcome of three possible reasons. (1) We should enlarge our vector parameter Ω including not only the effective temperature, the gravity, and the metallicity, but also the carbon, nitrogen and oxygen abundance and the microturbulence, thus enlarging our grid to a 7-dimensional grid. However, some first tests done in the framework of the study in Decin (2000) do show that this inflation of the parameter range does not solve the problem in the case of α Boo. (2) Secondly, we have to consider that inaccuracies may occur in the temperature distribution in the outermost layers of the model photosphere (Decin et al. 2003b), indicating that some assumptions, on which the theoretical models are based, are questionable for cool stars. One of the assumptions in the MARCS-code is that

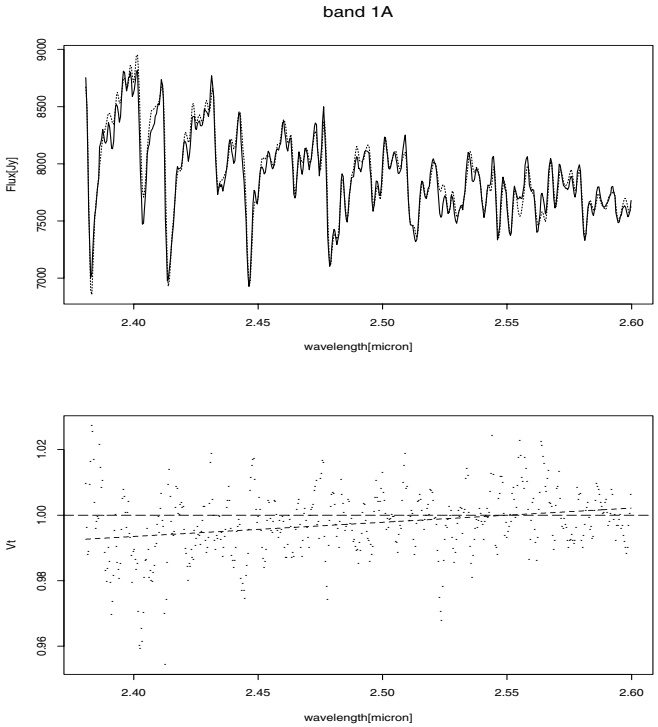

Fig. 12. Band 1A. *Upper panel*: comparison between the rebinned data (solid line) and theoretical data of model 39 (dashed line). *Lower panel*: V_t and the Loess smoother.

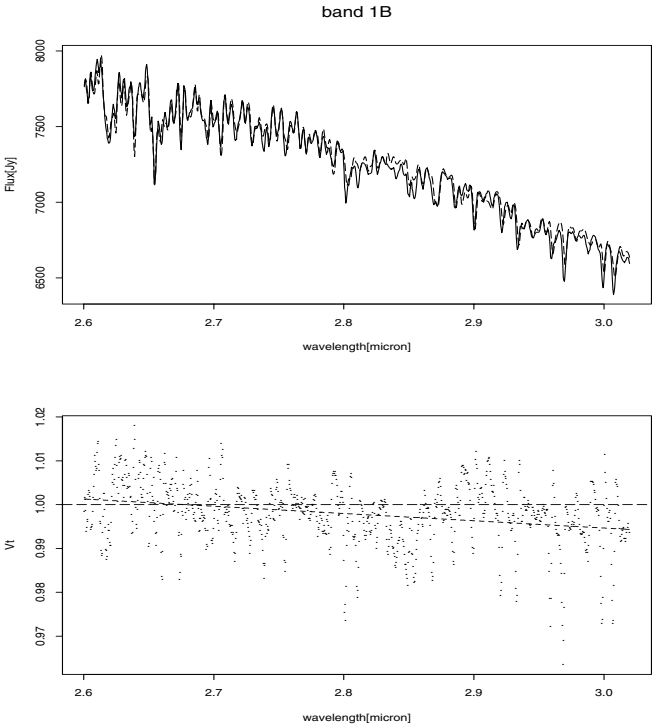

Fig. 13. Band 1B. *Upper panel*: comparison between the rebinned data (solid line) and theoretical data of model 39 (dashed line). *Lower panel*: V_t and the Loess smoother.

radiative equilibrium is required, also for the outermost layers. This implies that temperature bifurcation, caused by e.g. effects of convection and convective overshoot with inhomogeneities in the upper photosphere, cannot be allowed for. Consequently the cores of e.g. the satured CO and OH lines are not predicted with full success, resulting in a systematic pattern in V_t and so to a rejection of the null hypothesis. At least for α Boo, recent studies done by Ryde et al. (2002) show that the outermost surface layers may be a few hundreds of Kelvin cooler than predicted by the MARCS code. (3) Inaccuracies in the (satellite) data-reduction process result in (non)-rebinned data being systematically off from the "true" stellar spectrum. A problem with the Relative Spectral Response Function (RSRF) at the shorter wavelenghts of band 1A has already been reported by Decin et al. (2003b). Since the data are divided by the RSRF, a small problem with the RSRF at these places may introduce a pronounced error at the band edge. This kind of data-reduction problem can never be captured by the synthetic predictions, thus resulting in a systematic rejection of the null hypothesis. Using lack-of-fit tests for a sample of standard stars covering a broad parameter space, one can trace calibration problems.

In general, we may conclude that a systematic rejection of the null hypothesis in the lack-of-fit tests is an indication of a still incomplete modelling of all the physical mechanisms determining the spectral footprint in the wavelength range considered or of problems with the data reduction.

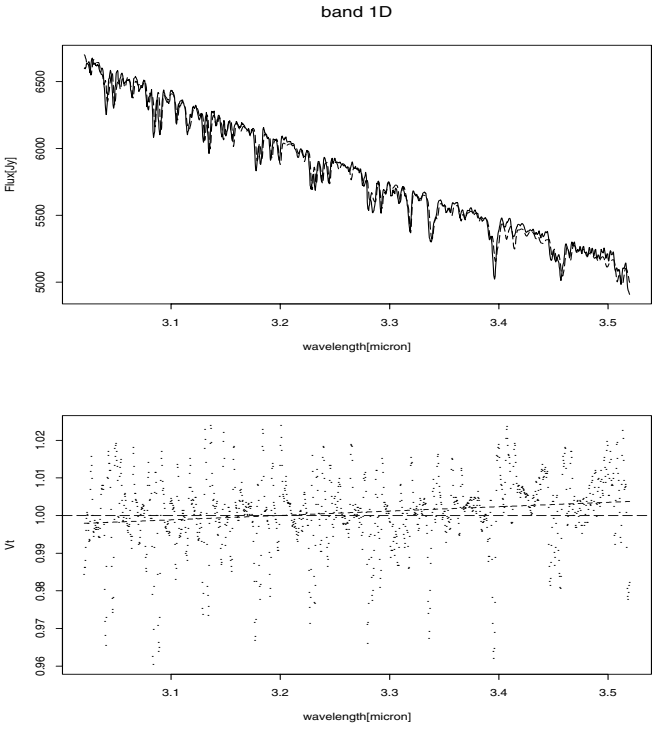

Fig. 14. Band 1D. *Upper panel*: comparison between the rebinned data (solid line) and theoretical data of model 39 (dashed line). *Lower panel*: V_t and the Loess smoother.

6. Summary and conclusions

In the first part of this article (Sects. 3 and 4) we have demonstrated that the use of either a *local* or a *global* goodness-of-fit

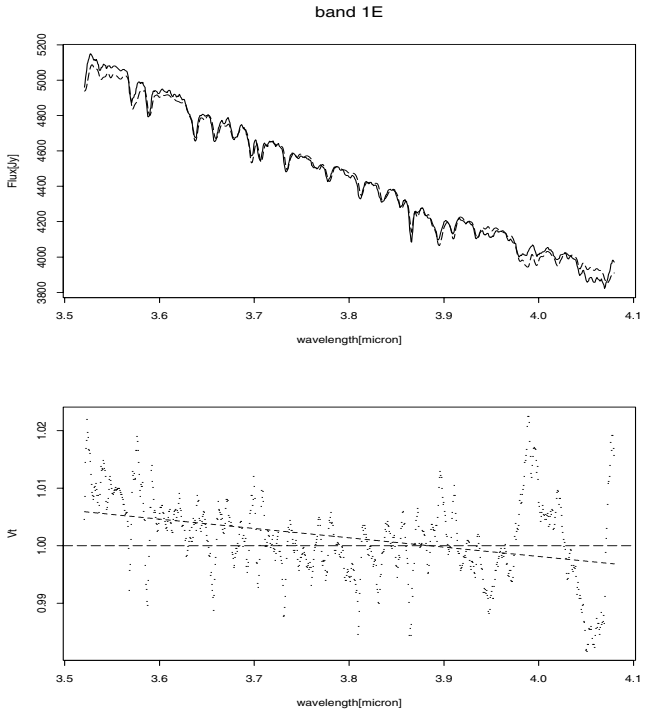

Fig. 15. Band 1E. *Upper panel*: comparison between the rebinned data (solid line) and theoretical data of model 39 (dashed line). *Lower panel*: V_t and the Loess smoother.

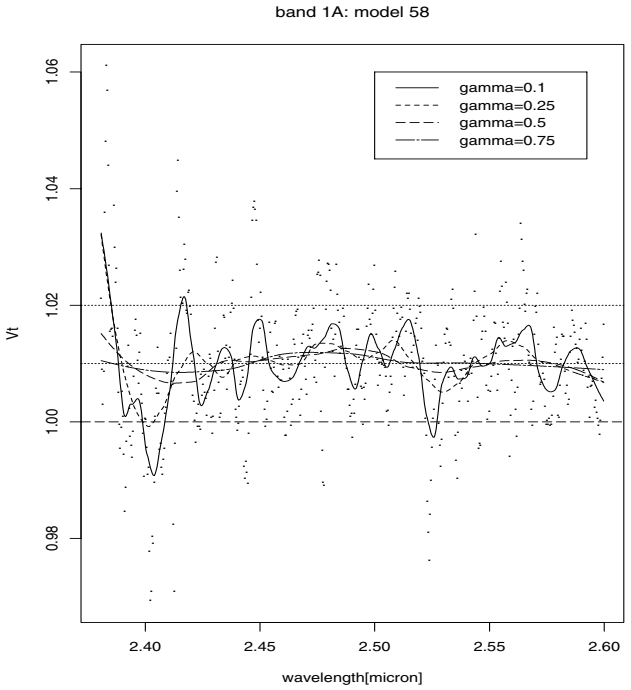

Fig. A.1. Band 1A: model 58 ($T_{\rm eff} = 4300$ K, $\log g = 1.35$ dex, [Fe/H] = −0.30 dex). V_t and 4 Loess smoothers with γ equals to 0.1 (solid line), 0.25 (short dashed line), 0.5 (long dashed line) and 0.75 (dotted-dashed line).

parameter is an important first step to objectively determine the uncertainty range on the estimated parameters. But a very important message is that combining *both* a local and a global deviation estimating parameter allows us to pin down the parameters with even more certainty. In the test-case of the ISO-SWS data of α Boo, we estimate the stellar parameters $T_{\rm eff}$, $\log g$ and metallicity as ranging respectively from 4160 K to 4300 K, from 1.35 to 1.65 dex, and from −0.30 to 0.00 dex using synthetic spectra calculated from MARCS model atmospheres.

Having determined the "best" models is however not the end of the story. The use of lack-of-fit tests enables us to detect systematic patterns in the difference between observed and theoretical data. For the case-study of α Boo, we obtained that in all the 4 sub-bands the closest synthetic spectra to the observational data are not capable of capturing all the structure from the data, i.e. the "best" models are not "good" enough. Both gaps in our knowledge of the physical mechanisms taking place during the life of a star, too simple assumptions in the theoretical modelling, uncertainties in additional stellar parameters – which are now kept fixed – and satellite data reduction problems may result in the rejection of the null hypothesis in the lack-of-fit tests.

As was illustrated by the example of the ISO-SWS data of α Boo, the statistical methods presented in this paper for comparing observational and synthetic data provide useful, practical and *general* tools: (1) to estimate objectively the stellar parameters *and* their uncertainties from observational data and a grid of synthetic spectra; (2) to refine the uncertainty intervals by combining a local and a global goodness-of-fit parameter; and (3) to trace if our knowledge of the physical mechanisms in a star, of the theoretical (numerical) modelling of the stellar photosphere or of the calibration process still need considerable refinement. The main limitation of this methodology is that measurement errors are still not included. In the following paper in this series (Shkedy et al., submitted) we will use hierarchical Bayesian models for the spectrum. In this approach, the measurement errors will be incorporated in the model as well.

Acknowledgements. L.D. acknowledges support from the Science Foundation of Flanders. Z.S. and G.M. gratefully acknowledge support from the Belgian IUAP/PAI network "Statistical Techniques and Modeling for Complex Substantive Questions with Complex Data".

Appendix A: Smooting using Loess models

Non-parametric regression models aim to describe the relationship between a response variable y and a predictor x. The model has the general form of

$$y_i = \eta(x_i) + \varepsilon_i, \quad i = 1 \ldots, n,$$

where η is considered to be a smooth function. The local linear approach for non-parametric models is based on solving the local weighted least square problem

$$\min_{\alpha,\beta} \sum_{i=1}^{n} \left\{ y_i - \alpha - \beta(x_i - x) \right\}^2 w(x_i - x; \gamma).$$

Here, w is a weight function symmetric around zero and γ is a smoothing parameter which controls the width of w and therefore the amount of smoothing in the model. Local quadratic models can be fitted by including the term $(x_i - x)2$ into the

model. In our setting we applied a local linear model. The effects of the smoothing parameter on the fitted model is illustrated in Fig. A.1 which shows 4 Loess smoothers with smoothing parameters increasing from 0.1 to 0.75. Clearly, as the smoothing parameter increases, the fitted model becomes "more" smooth.

For a further discussion about *Loess*, which stands for "local regression", we refer to Cleveland (1979) and the book of Bowman & Azzalini (1997). An intuitive introduction about Loess can be found in Cleveland (1993). The book of (Hart 1997) gives a comprehensive discussion about smoothing and data driven choice of the smoothing parameter.

Appendix B: The bootstrap procedure

We applied the following bootstrap procedure in order to calculate the empirical p-values.

- Construct the residuals $\hat{e}_i = V_{t_i} - \hat{\eta}(t_i)$ where $\hat{\eta}(t_i)$ is an initial estimator for $\eta(t_i)$.
- Create a new set of normalised residuals, $\tilde{e}_i = \hat{e}_i - \sum_i \hat{e}_i/n$.
- Create the bootstrap observation by sampling a value with replacement form $\tilde{e}_1, \ldots, \tilde{e}_n$ and define $V^{\star}_{t_i} = 1 + \tilde{e}^{\star}_i$.
- Calculate the value of $F^{\star}_b$ for the bootstrap sample.
- Repeat the procedure above B times.

The empirical p-value is the proportion of the bootstrap statistics, $F^*_1, \ldots, F^*_B$, that are greater or equal to the value of F which is calculated from the original sample. For further discussion of the bootstrap procedure we refer to the books of Davison & Hinkley (1997) and Bowman & Azzalini (1997).

References

Bailer-Jones, C. A. L. 2000, A&A, 357, 197

Bowman, A. W., & Azzalini, A. 1997, Applied smoothing techniques for data analysis (Oxford: Clarendon press)

Bracewell, R. N. 1985, The Fourier Transform and its Applications (McGraw-Hill, Inc.)

Chambers, J. M., & Hastie, T. J. 1992, Statistical models in S (California: Wadsworth & Brooks)

Cleveland, W. S. 1979, JASA, 74, 829

Cleveland, W. S. 1993, Visualizing data (New-Jersey: AT&T)

Davison, A. C., & Hinkley, D. V. 1997, Bootstrap Methods and Their Application (Cambridge University Press)

De Bruyne, V., Vauterin, P. S. D., & Dejonghe, H. 2003, MNRAS, in press

de Graauw, T., Haser, L. N., Beintema, D. A., et al. 1996, A&A, 315, L49

Decin, L. 2000, Ph.D. Thesis, University of Leuven, Belgium

Decin, L., Waelkens, C., Eriksson, K., et al. 2000, A&A, 364, 137

Decin, L., Vandenbussche, B., Waelkens, C., et al. 2003a, A&A, 400, 709

Decin, L., Vandenbussche, B., Waelkens, C., et al. 2003b, A&A, 400, 679

Decin, L., Vandenbussche, B., Waelkens, C., et al. 2003c, A&A, 400, 695

Edvardsson, B., Andersen, J., Gustafsson, B., et al. 1993, A&A, 275, 101

Erspamer, D., & North, P. 2002, A&A, 383, 227

Farmer, C. B., & Norton, R. H. 1989, A High-Resolution Atlas of the Infrared Spectrum of the Sun and Earth Atmosphere from Space, I, The Sun (Washington, D.C.: NASA, Scientific and Technical Information Division)

Geller, M. 1992, A High-Resolution Atlas of the Infrared Spectrum of the Sun and Earth Atmosphere from Space, III, Key to Identification of Solar Features (Washington, D.C.: NASA, Scientific and Technical Information Division)

Gustafsson, B., Bell, R. A., Eriksson, K., & Nordlund, Å. 1975, A&A, 42, 407

Hart, J. D. 1997, Nonparametric smoothing and lack-of-fit tests (Springer)

Jørgensen, U. G., Johnson, H. R., & Nordlund, Å. 1992, A&A, 261, 263

Katz, D., Soubiran, C., Cayrel, R., Adda, M., & Cautain, R. 1998, A&A, 338, 151

Kessler, M. F., Steinz, J. A., Anderegg, M. E., et al. 1996, A&A, 315, L27

King, E., Hart, J. D., & Wehrly, T. E. 1991, Statist. Probab. Lett., 12, 239

Leech, K., Kester, D., Shipman, R., et al. 2002, in The ISO Handbook, V, SWS – The Short Wavelength Spectrometer, ed. J. Müller, & T. G. Blommaert

Plez, B., Brett, J. M., & Nordlund, Å. 1992, A&A, 256, 551

Ryde, N., & Eriksson, K. 2002, A&A, 386, 874

Ryde, N., Lambert, D. L., Richter, M. J., & Lacy, J. H. 2002, ApJ, 580, 447

Valenti, J. A., & Piskunov, N. 1996, A&AS, 118, 595