ILIUM IV: Three-dimensional forward model and demonstration of a strong and ubiquitous T_{eff}-A_V degeneracy in BP/RP

prepared by: Coryn A.L. Bailer-Jones
Max Planck Institute for Astronomy, Heidelberg
Email: calj@mpia.de

approved by:
reference: GAIA-C8-TN-MPIA-CBJ-048
issue: 2
revision: 1
date: 2009-10-11
status: Issued

Abstract

The ILIUM algorithm for estimating astrophysical parameters (APs) from spectral data is extended from the 1D+1D (1 strong AP plus 1 weak AP) to the 2D+1D case. Specifically, I extend it to simultaneously estimate two strong APs ($\mathrm{T_{eff}}$ and $\mathrm{A_V}$) plus one weak AP ($\log g$ or [Fe/H]). In this way, ILIUM is able to estimate the standard four stellar APs. ILIUM reveals the expected but little explored strong degeneracy between $\mathrm{T_{eff}}$ and $\mathrm{A_V}$ when estimated from the BP/RP spectra. The degeneracy is so strong and ubiquitous over the $\mathrm{T_{eff}}$–$\mathrm{A_V}$ grid that it makes little sense to talk of single estimates of $\mathrm{T_{eff}}$ and $\mathrm{A_V}$ unless prior information can be used to suppress ranges of these parameters. This has serious implications for how we report classification results with Gaia.

The degeneracy diminishes the value of simple error statistics as summaries of an algorithm's performance, because the "true" APs can no longer be found even in principle (the spectra are indistinguishable within the noise). If we nonetheless ignore this degeneracy, the mean absolute residuals in the APs on end-of-mission spectra at G=15 are 0.3 dex in $\log g$, 3% in $\mathrm{T_{eff}}$ and 0.07 mag in $\mathrm{A_V}$ for a wide range of APs (4000–15 000 K and 0–10 mag $\mathrm{A_V}$) for a solar-metallicity stars. For cool dwarfs ($\mathrm{T_{eff}} \leq 7000$ K) covering the same range of $\mathrm{A_V}$, the mean absolute error in [Fe/H] is 0.5 dex. $\mathrm{T_{eff}}$ can be more accurately estimated for low extinction stars, and $\mathrm{A_V}$ more accurately for hotter stars (but $\mathrm{A_V}$ accuracy appears not to depend on $\mathrm{A_V}$). The [Fe/H] accuracy for dwarfs also appears to be independent of $\mathrm{A_V}$. At G=18.5 these "accuracies" drop by a factor of about four for all APs (except the error in [Fe/H], which drops to 1.3 dex). Thus [Fe/H] and $\log g$ can no longer be usefully estimated, although this will improve if priors can be used to limit the range of $\mathrm{T_{eff}}$ and/or $\mathrm{A_V}$, as the results in earlier TNs on ILIUM have shown.

Document History

Issue	Revision	Date	Author	Comment
2	1	2009-10-11	CBJ	Corrected error in Fig. 22 and in footnote 3, spotted by Lennart Lindegren. Minor changes to section 4.2 following discussions with Lennart and Anthony Brown (who are not responsible for what remains!)
2	0	2009-09-30	CBJ	Issued to Livelink – no further changes
2	D	2009-09-16	CBJ	Section 4.2 re-written following conversations with Carme Jordi and Anthony Brown and extended to show degeneracy mapping with sensitivity-weighted distance measure and how this can be converted to probabilities. Conclusions and abstract updated. Comments from Vivi Tsalmantza.
1	0	2009-08-27	CBJ	First version

Contents

1 Introduction

In CBJ-042, CBJ-043 and CBJ-046 I introduced ILIUM, a new algorithm for estimating parameters from multidimensional data. There I showed how it could be used to estimate a pair of APs from stellar spectra, the pair comprising one strong AP ($\mathrm{T_{eff}}$) and one weak AP ([Fe/H] or $\log g$) from BP/RP. In this way, ILIUM could be used to estimate three APs (in principle, one strong AP and any number of weak APs). Because the method is based around a forward model – an estimate of the spectrum given the (estimated) APs – we can use it to calculate a goodness-of-fit for the best estimated APs, as well as uncertainty estimates (covariances) on these APs. In these earlier technical notes I also examined the dependence of the performance on the wavelength range, its sensitivity to systematic flux and wavelength errors, and I introduced some extensions, for example a SNR-weighting of the input data.

The performance analyses showed that ILIUM is at least as accurate and precise as more established parametrization methods, such as SVMs and k-nn, and in many cases is superior, and it furthermore provides natural error diagnostics. However, because it was confined to estimate a single strong AP (viz. $\mathrm{T_{eff}}$), the model was limited to stellar grids with zero interstellar extinction. Here I show how ILIUM can be extended to estimate two strong APs ($\mathrm{T_{eff}}$ and $\mathrm{A_V}$) simultaneously, still with a single weak AP. In this way ILIUM becomes a 3D (2D+1D) estimation model. By fitting two models, with [Fe/H] and $\log g$ as the two weak APs in each case, we can estimate the four standard APs. I will also show how the forward model in ILIUM can be used to identify and map the degeneracy between $\mathrm{T_{eff}}$ and $\mathrm{A_V}$.

2 Extension to a 2D+1D forward model

The 1D+1D version of ILIUM introduced in earlier TNs consists of two one-dimensional forward models (see section 2 of CBJ-042). The first models the variation of the flux with a strong AP (e.g. $\mathrm{T_{eff}}$), the second the variation on top of this due to a single weak AP (e.g. [Fe/H] or $\log g$). Both were modelled as one-dimensional smoothing splines which enables us to control their complexity via the degrees-of-freedom (dof). (They also circumvent the need to define the position of the knots, which themselves are of no interest.)

Here I extend the strong forward model to be a 2D model over two APs ($\mathrm{T_{eff}}$ and $\mathrm{A_V}$). I now use a two-dimensional thin-plate spline implemented via the function `Tps` in the package `fields` in R, which is again a smoothing spline with complexity controlled by the dof. Equation 6 in CBJ-042 for combining the strong and weak models still applies. The forward model fitting procedure is as follows (still for each spectral band independently). Given a grid, we identify the unique combinations of the two strong APs. At each of these we calculate the mean flux (i.e. marginalize over the weak AP). The 2D spline is fit to these values. The weak models are then fit in the same way as before. That is, at each point in the 2D grid (each unique $\mathrm{T_{eff}}$/$\mathrm{A_V}$ combination), a 1D spline is fit to the variation of the residual flux (about the mean) with respect

to the weak AP. The combined forward model is applied in the same way as before (section 2.4 of CBJ-042).

Everything else in the ILIUM model remains the same as the original 1D+1D algorithm. In particular, the AP update equation is unchanged (equation 5 in CBJ-042) as is the method of using first differences to calculate the sensitivities (equation 8 in CBJ-042, although this is not fundamental to the algorithm and could be replaced with an analytic method for certain forward model functions).

The forward model is of course much bigger now, as we have many more weak models (330 in the application below, compared to 33 in CBJ-042).

2.1 Further extensions

If we want to estimate four APs (two strong, two weak), then this could be done applying a pair of 2D+1D models, with [Fe/H] and $\log g$ as the weak AP in each case. It would probably be best to first use a $\log g$ model which marginalizes over a range of [Fe/H] (say from -2 to 0 dex) to determine $\log g$, $\mathrm{T_{eff}}$ and $\mathrm{A_V}$, and then use a [Fe/H] model which marginalizes over a range of $\log g$ (e.g. a set of dwarfs and giants) to determine [Fe/H] (and $\mathrm{T_{eff}}$ and $\mathrm{A_V}$ again). Alternatively, we could use the $\log g$ estimate from the first model to select one of several possible second models, each trained on stars with a narrow range of $\log g$. This *might* give a better [Fe/H] estimate. Alternatively, it would be relatively easy to extend also the weak model to be 2D and thus simultaneously estimate [Fe/H] and $\log g$. This will be the subject of future work.

3 Estimation of ($\mathrm{T_{eff}}$, $\mathrm{A_V}$) + ($\log g$ or $[\mathrm{Fe/H}]$)

3.1 Data

The extended model is applied to the same GOG BP/RP star simulations used in earlier TNs, namely simulations of MARCS and BASEL spectra from DPAC cycle 3 (nominal grid). We now include the full range of $\mathrm{A_V}$ values. The $\mathrm{T_{eff}}$–$\mathrm{A_V}$ grid is shown in Fig. 1. There are 33 unique $\mathrm{T_{eff}}$ values and 10 unique $\mathrm{A_V}$ values and all 330 unique combinations are present; this is therefore the number of points in the 2D strong forward model. (There are in fact 340 non-unique combinations, because the BASEL and MARCS libraries are duplicated at 8000 K. See also section 3.3.) The weak AP is either [Fe/H] or $\log g$; their distribution is shown in Fig. 2. The impact of $\mathrm{T_{eff}}$ and $\mathrm{A_V}$ on the spectra is shown in Figs. 3 and 4.

Not all [Fe/H] and $\log g$ values are present at every $\mathrm{T_{eff}}$ value: for the $\mathrm{T_{eff}}$–$\log g$ and $\mathrm{T_{eff}}$–[Fe/H] distributions see CBJ-042 and CBJ-043 respectively. However, each ($\log g$, [Fe/H], $\mathrm{T_{eff}}$) value

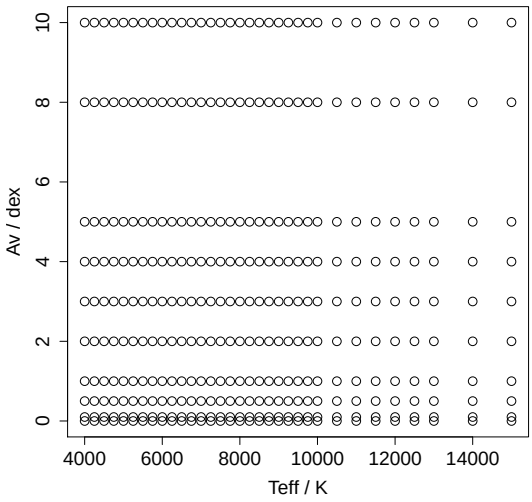

FIGURE 1: The AP grid for the strong APs used for the experiments.

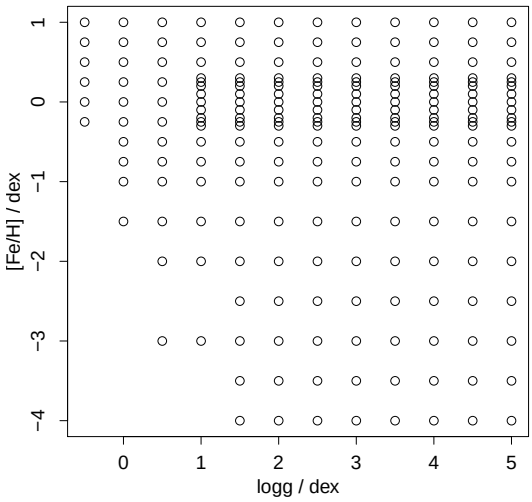

FIGURE 2: The AP grid for the weak APs used for the experiments.

is present at all A_V values. In total there are 41 610 unique stars (AP combinations). The sizes of different subsets (used to fit the forward model) are as follows

- zeromet: [Fe/H] $= 0.0$; 2740 stars
- dwarfs: $\log g \in \{4.0, 4.5, 5.0\}$; 17 160 stars
- giants: $\log g \in \{1.0, 1.5, 2.0, 2.5, 3.0\}$; 18 820 stars

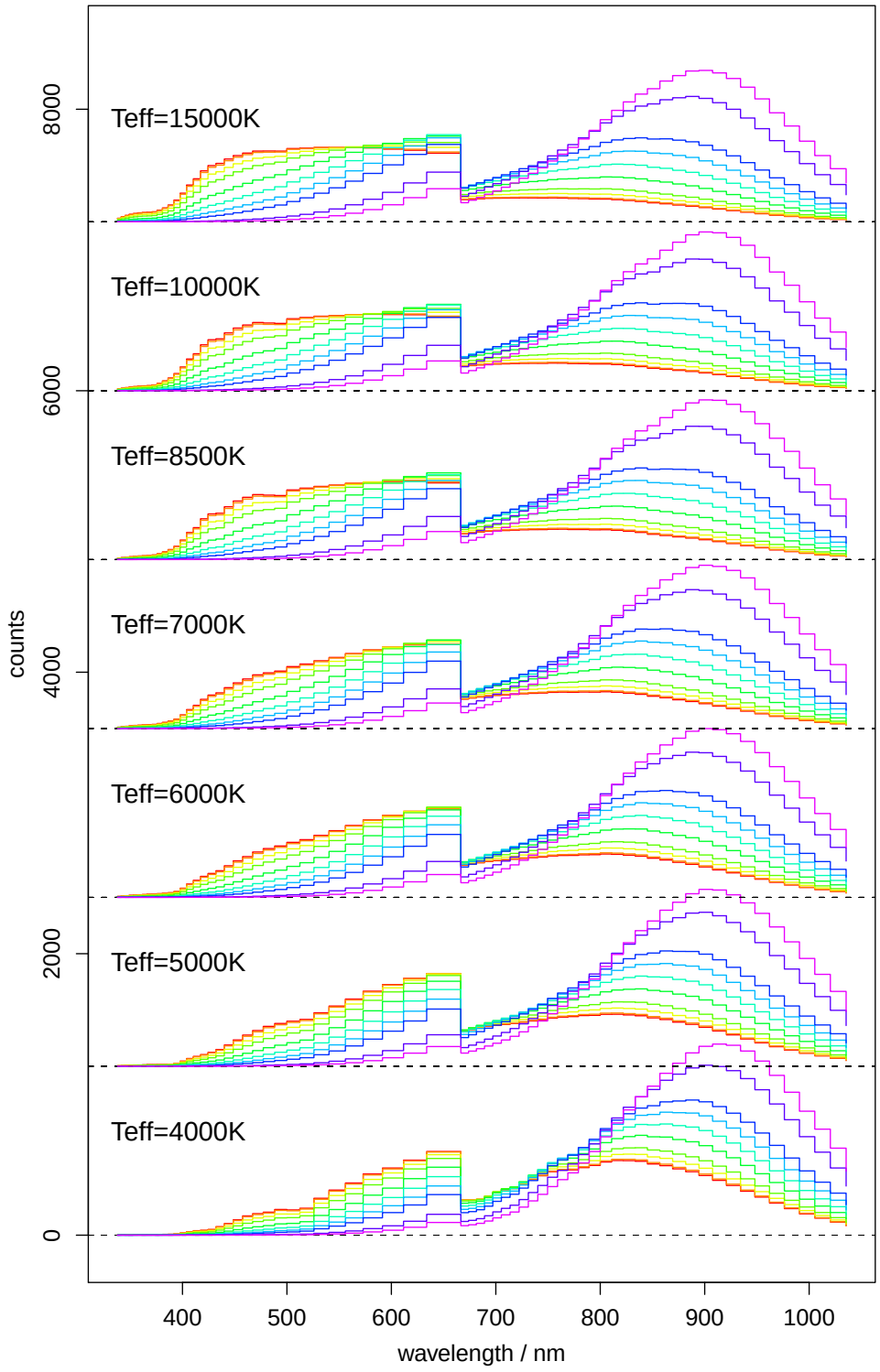

FIGURE 3: Example noise-free BP/RP for zero metallicity dwarfs ($\log g$=4.0) at 7 $T_{\rm eff}$ values for all 10 values of A_V (0.0, 0.1, 0.5, 1, 2, 3, 4, 5, 8, 10) ranging from 0.0 mag (red, lowest line at high wavelengths) to 10.0 mag (violet, highest line at high wavelengths). Each temperature block has been offset by 1200 counts for clarity (the zero levels are shown by the dashed lines).

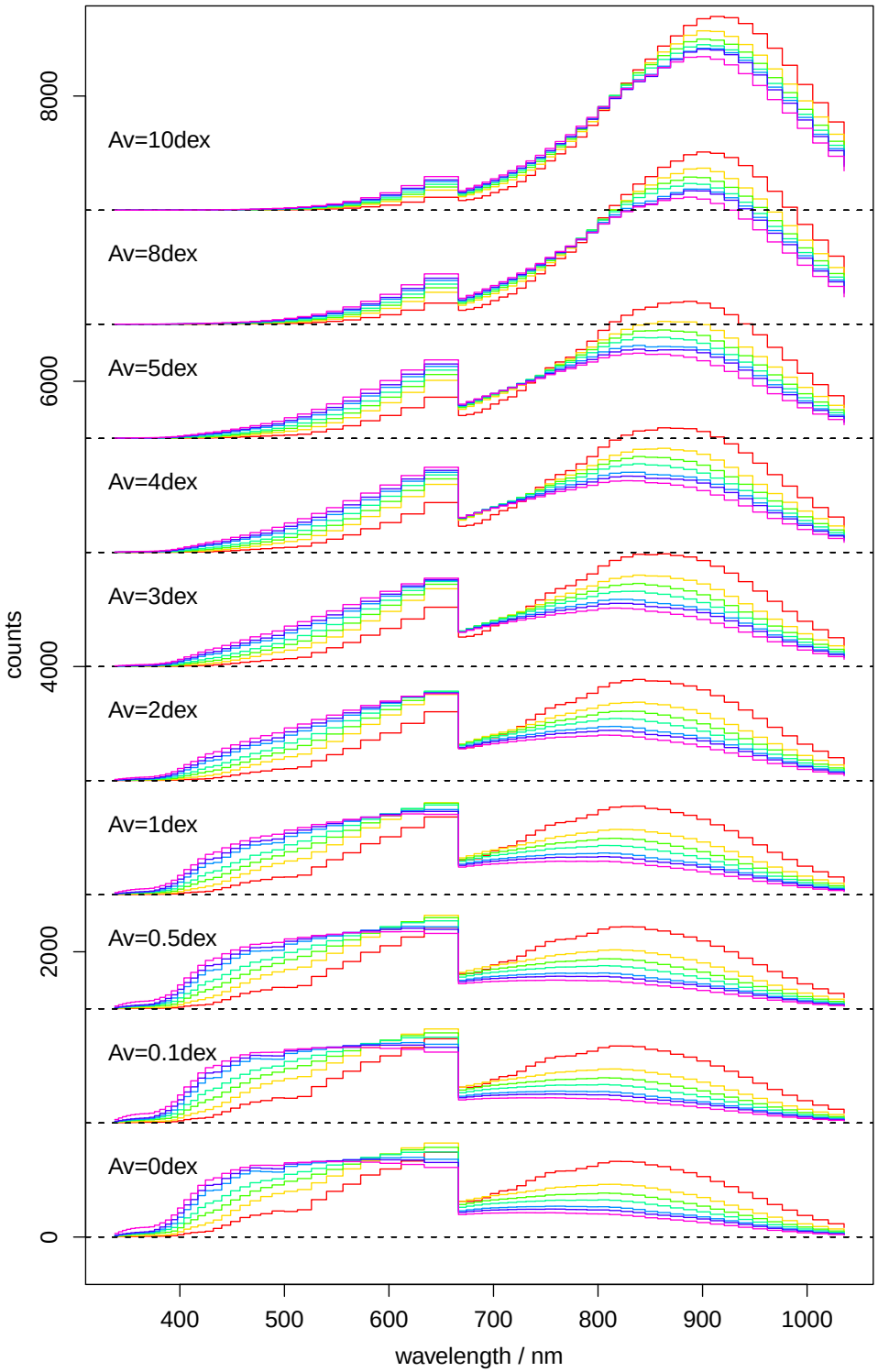

FIGURE 4: The same spectra as in Fig. 3 but now with the spectra of a common A_V value plotted in a single block with an offset of 800 counts between them. T_{eff} ranges from 4000 K (red, highest line at high wavelengths to 15 000 K (violet, lowest line at high wavelengths).

3.2 Internal parameter settings

Based on trial and error for this data set, the number of degrees of freedom (dof) for the 2D smoothing spline was set to be a third of the number of unique $\mathrm{T_{eff}}$–$\mathrm{A_V}$ (strong grid) points, which here is $330/3 = 110$. This is more than adequate to get a reliable fit; the forward model fitting is not very sensitive to the exact number. The dof of the weak model is unchanged (4, or a linear fit is used if four or fewer points are available). The other parameters of ILIUM retain the values given in Table 2 of CBJ-042, the values for $\mathrm{A_V}$ being set to those shown for $\log(\mathrm{T_{eff}})$. The only exception is the value of "delta for sensitivity calculation via rst differences", which is set to 0.01 (in standardized units) for $\mathrm{A_V}$, compared to 0.0034 for $\log(\mathrm{T_{eff}})$. This larger value was adopted based on the impression that the flux does not vary as rapidly with $\mathrm{A_V}$, but it has not been optimized in any way. ILIUM should not be very sensitive to these step sizes.

3.3 Forward model fitting and initialization

Noise-free spectra for the whole grid for a particular problem are used to fit the forward models, as in previous TNs. A random subset of half the grid (noisy spectra) is used for the nearest neighbour initialization. The test set is selected from the remaining half: unless mentioned otherwise below, a random subset of just 1000 stars was used in order to keep the processing times down to a few hours (the R code is not optimized in any way!).

The new 2D+1D forward models fit well. One dimensional cuts of these models are shown in Figs. 5, 6, 7 and 8. We see that this two-component forward model adequately captures the complexity of the variations without overfitting. The strong component marginalizes well over the multiple values of the weak APs (seen in Fig. 7). Recall that the data is a composite of the MARCS and BASEL grids, joined and duplicated at 8000 K. We see discrepancies between these libraries in the plots of the forward model predictions against the weak APs, for example in Fig. 8. Here we see two distinct points for the redder bands (although only at some $\log g$ do both models produce spectra). This is not a result of high extinction, as we see a similar problem at $\mathrm{A_V} = 2$, for example. We also see in Fig. 5 a discontinuity in the first derivative of the forward model at this $\mathrm{T_{eff}}$ for some bands.

3.4 Results

I now apply ILIUM to various data sets and analyse the residuals. In the discussion, I generally refer to the mean absolute residual ($\overline{|\delta\phi|}$) as the error statistic unless otherwise stated. This is a more robust and therefore useful measure than the rms (σ_ϕ). *If* the residuals had a Gaussian distribution, then the corresponding standard deviation ($\pm 1\sigma$ error) is 1.48 times greater. Note that the error, ϵ, in a logged quantity (e.g. $\log(\mathrm{T_{eff}})$) corresponds to a percentage error of 2.3ϵ in the unlogged quantity. The $\mathrm{T_{eff}}$ and $\mathrm{A_V}$ error should not be interpreted too literally, as their significance is limited by the presence of a strong degeneracy (see section 4). In several cases

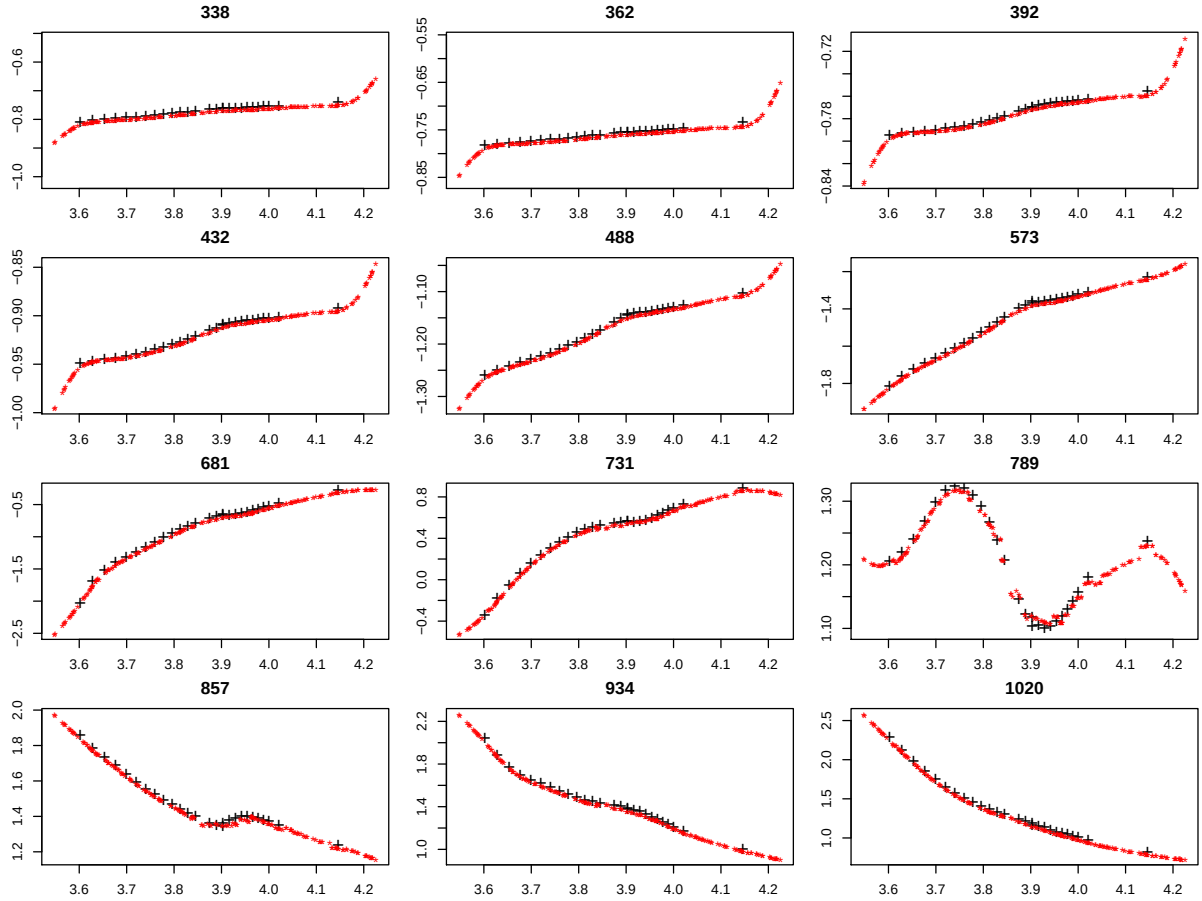

FIGURE 5: Predictions of the full forward model (2D+1D) in 12 different bands (with wavelength in nm at the top of each panel). Here are shown fluxes for highly extinct giants ($\log g = 2$ dex, $A_V = 8$ mag) from the zeromet data set ([Fe/H] = 0) as a function of $\log(T_{eff})$. The black crosses are the (noise-free) grid points, the red stars are the forward model predictions (at randomly selected AP values). The flux plotted on the ordinate is in standardized units.

I fit the forward model on the full range of A_V, T_{eff} and [Fe/H], but often only report summary results on the cooler stars ($T_{eff} \leq 7000$ K) in the test set, because we know from earlier work that [Fe/H] performance on hotter stars is very poor.

3.4.1 G=15

The overall performance of ILIUM fit to the zeromet data set ([Fe/H] = 0) averaged over the full AP ranges is (residuals plotted in Fig. 9).

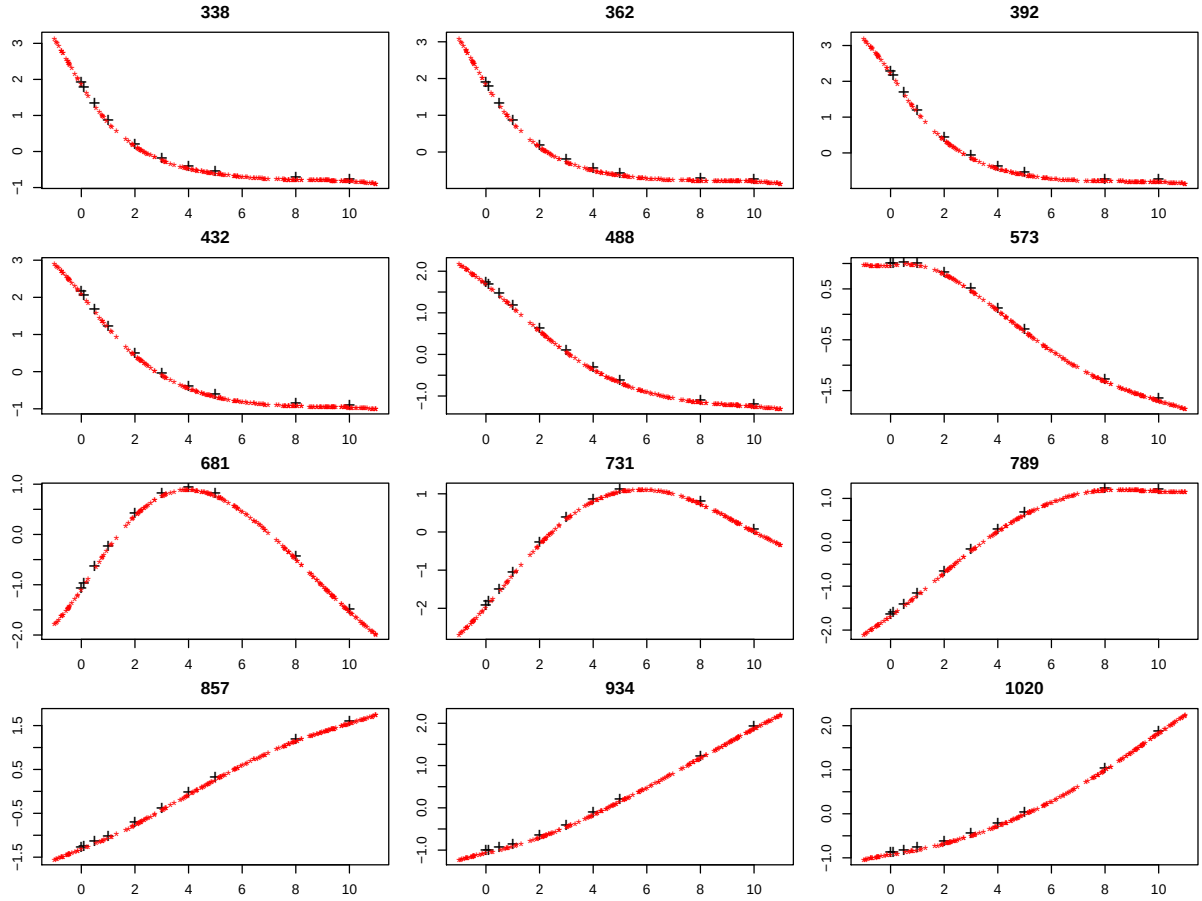

FIGURE 6: As Fig. 5 but now showing variations as a function of A_V with T_{eff} fixed at 10,000 K. (The plots at 5000 K are qualitatively similar)

	$\log g$	$\log(T_{eff})$	A_V
$\overline{\delta\phi}$	−0.052	0.0014	−0.0067
$\overline{\lvert\delta\phi\rvert}$	0.29	0.013	0.072
σ_ϕ	0.57	0.026	0.15

ILIUM, zeromet, G=15, full AP range

The 1D+1D model (i.e. extinction fixed to 0.0 mag) in CBJ-042 yielded mean absolute residuals in $\log g$ and $\log(T_{eff})$ of 0.065 dex and 0.001 dex respectively. The presence of a wide range of extinction has had a significant negative impact on the accuracies of both, as we might expect (note, however, that there is still essentially no systematic error, $\overline{\delta\phi}$). Given the scientific requirements, e.g. dwarf–giant discrimination, the above performance is still quite adequate. That A_V can be estimated to an accuracy of 0.07 mag is encouraging. If we limit the analysis (without changing the fitted model) to just low extinction stars ($A_V \leq 1.0$ mag), then A_V can be estimated slightly better (0.056 mag) but $\log(T_{eff})$ significantly better (0.008 dex). That is, we

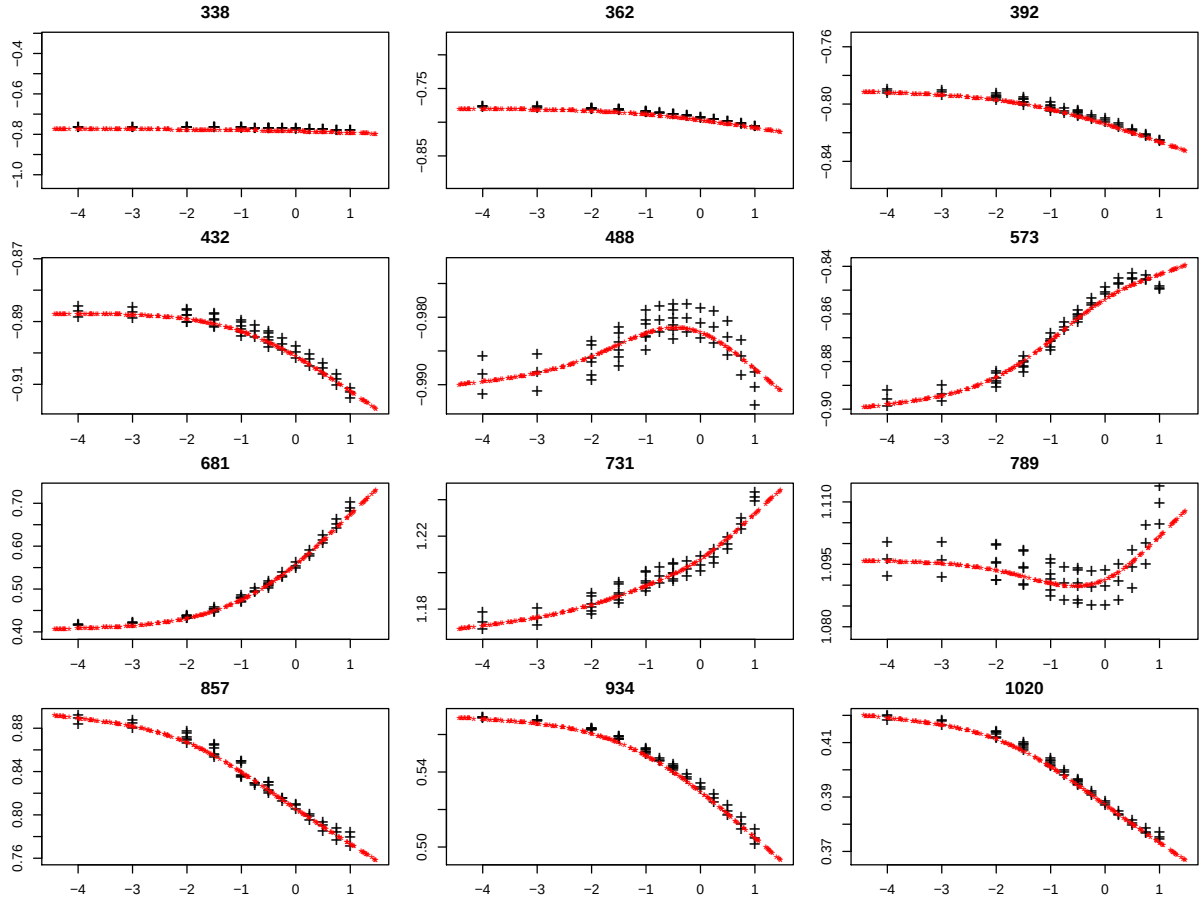

FIGURE 7: Forward model for the dwarfs data set as a function of [Fe/H] with T_{eff} fixed at 6000 K and A_V at 5 mag. The three values at a given T_{eff} correspond to the three values of $\log g$. Note that the strongest metallicity signature is *not* in the bluest bands (cf. zero extinction case in Fig. 4 of CBJ-043).

can estimate T_{eff} more accurately for low extinction stars (averaged over all $\log g$). [1]

I now apply ILIUM to the dwarfs data set ($\log g = 4.0$, 4.5 and 5.0) by fitting the forward model on the full range of the other three APs. The performance of this model on just the cool stars ($T_{eff} \leq 7000$ K) in the test set is (residuals in Fig. 10)

[1] This improvement when limiting A_V is not the same as placing a prior on the A_V value to mitigate the the T_{eff}–A_V degeneracy I discuss later, because we are not placing limits on the *predicted* (posterior) values of T_{eff} or A_V here.

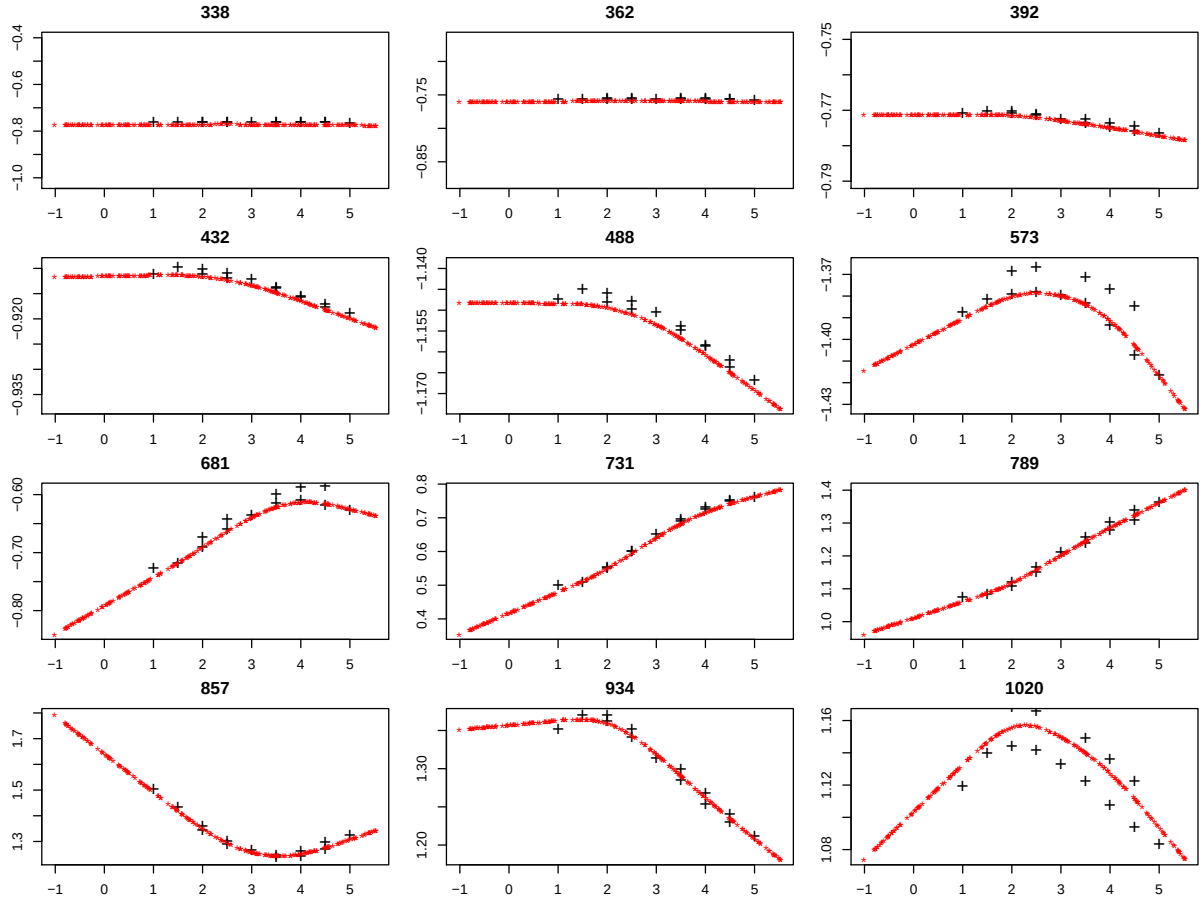

FIGURE 8: Forward model predictions for the zeromet data set at $T_{\rm eff} = 8000\,K$ and $A_V = 8$ mag as a function of $\log g$, showing the discrepancy between the BASEL and MARCS grids (double fluxes at five $\log g$ values).

	[Fe/H]	$\log(T_{\rm eff})$	A_V
$\overline{\delta\phi}$	−0.0034	0.0022	0.032
$\overline{\lvert\delta\phi\rvert}$	0.46	0.018	0.18
σ_ϕ	0.79	0.029	0.32

ILIUM, dwarfs, G=15, $T_{\rm eff} \leq 7000\,K$

The 1D+1D model (i.e. extinction fixed to 0.0 mag) in CBJ-043 yielded mean absolute residuals in [Fe/H] and $\log(T_{\rm eff})$ of 0.14 dex and 0.0017 dex respectively. As for $\log g$, the presence of a large A_V range makes [Fe/H] determination considerably harder. Yet an accuracy of 0.5 dex is still adequate for large population surveys like Gaia. The absence of a systematic in metallicity means that we will achieve more precise results when averaging over a population of objects. Compared to the zeromet sample above, the A_V estimation is worse (0.18 mag compared to 0.07 mag). This is a consequence of having limited the analysis to cool stars, because A_V can be estimated more accurately for hotter stars: the mean absolute residual is only 0.1 mag for the dwarf set for $T_{\rm eff} > 7000\,K$. If we further limit the above analysis to low extinction stars

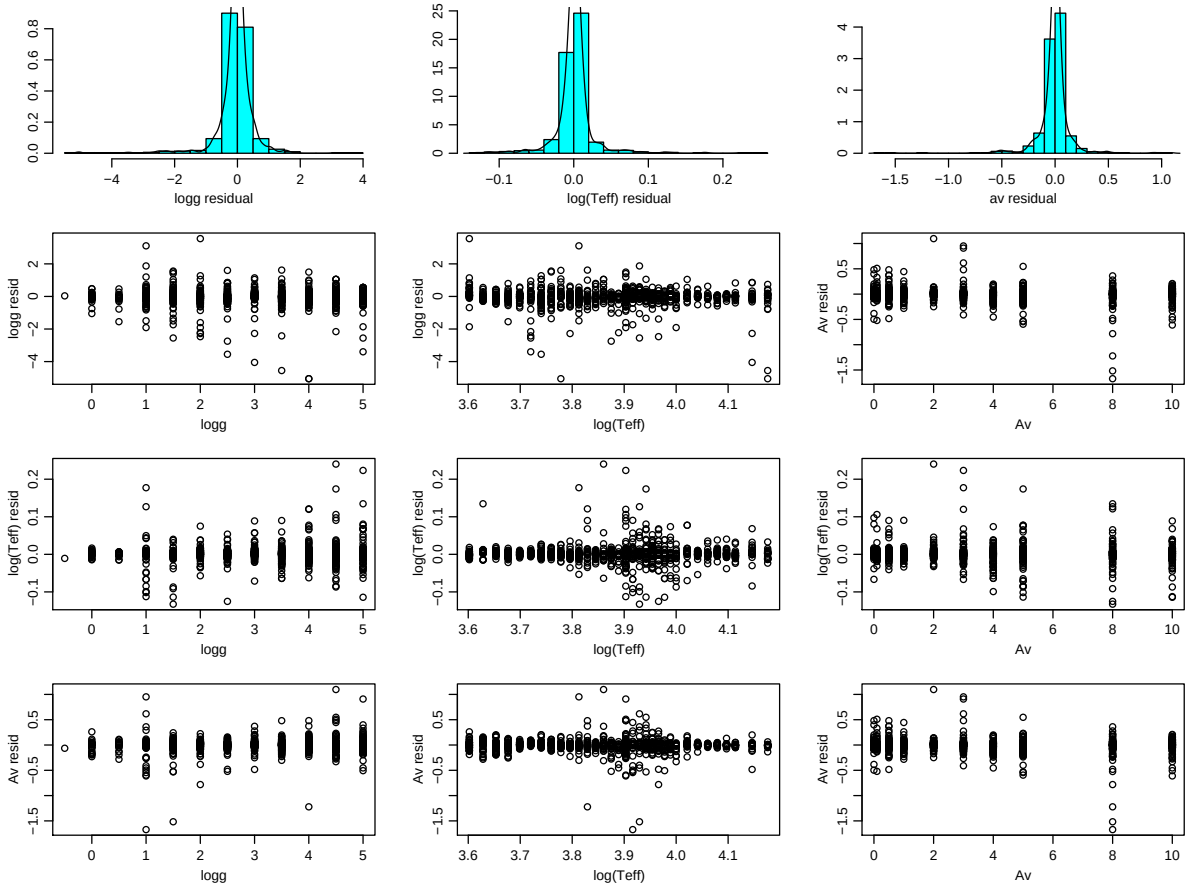

FIGURE 9: AP residuals for the zeromet data set at G=15, plotted as a function of the true APs, for the full range of [Fe/H], T_{eff} and A_V shown in Figs. 1 and 2

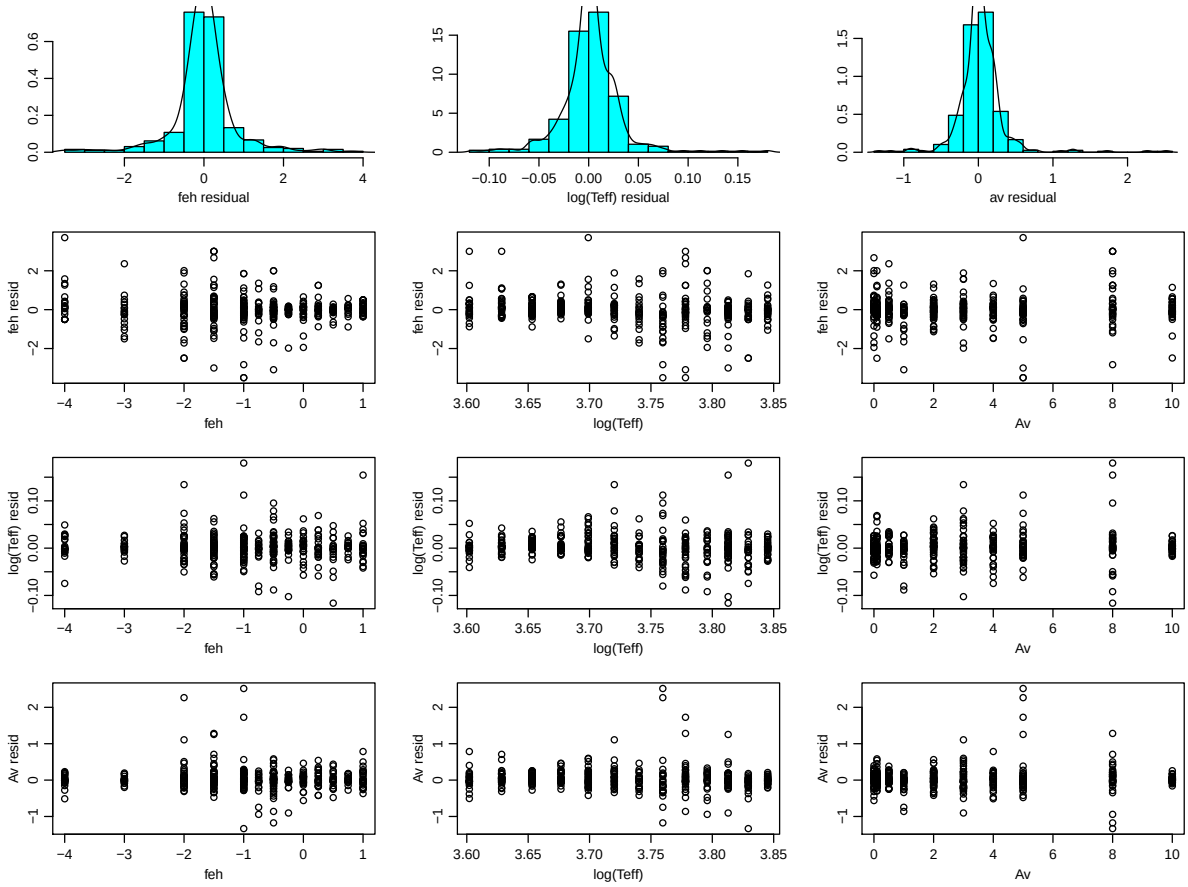

FIGURE 10: AP residuals for the dwarfs data set at G=15, plotted as a function of the true APs, for cool stars ($T_{eff} \leq 7000$ K) and the full range of T_{eff} and A_V.

($A_V \leq 1.0$ mag), then the [Fe/H] precision is essentially unchanged (0.44 dex). In other words, it is not more difficult (on average) to determine the metallicity of stars with large extinction than stars with small extinction. Rather, it must be the fact that we a priori have a large extinction range – and therefore an error in the estimation of the extinction – that degrades the [Fe/H] precision over the zero extinction case in CBJ-043.

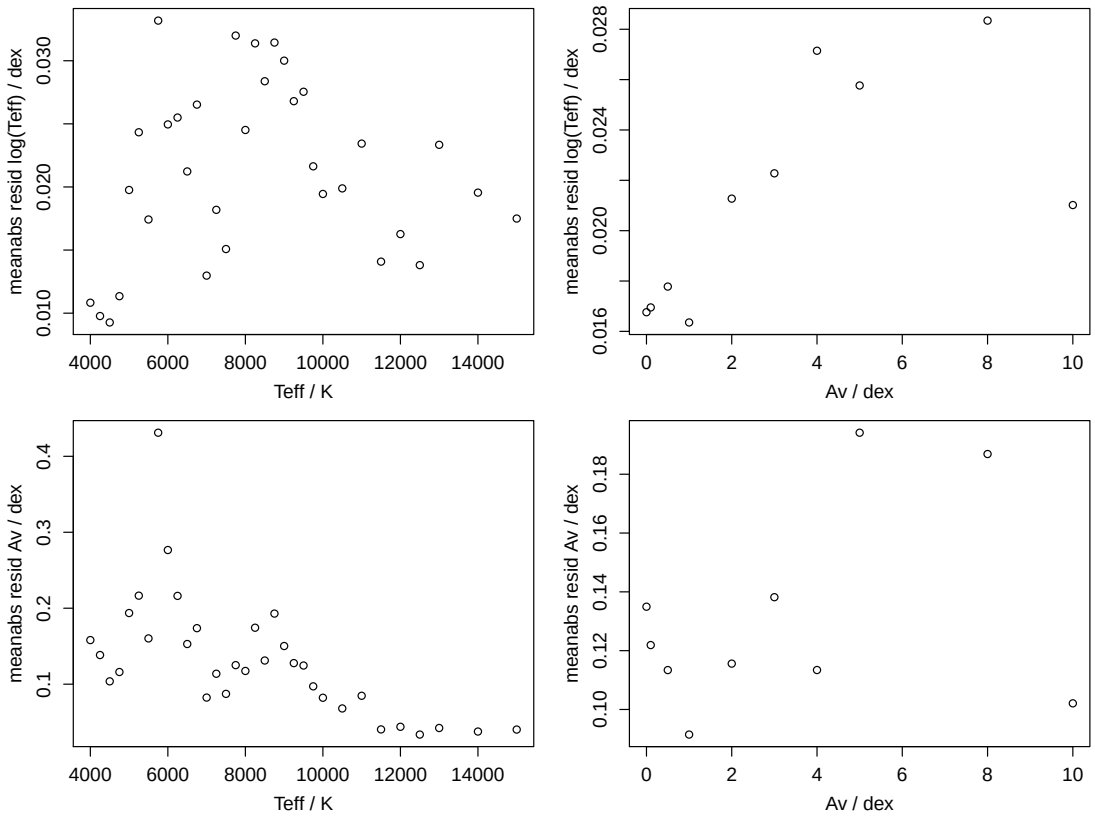

FIGURE 11: The mean absolute residual for $\log(T_{\rm eff})$ and A_V at the individual (discrete) values of $T_{\rm eff}$ and A_V present in the test data set (for all [Fe/H]). (This makes the trend in residuals with the APs easier to see.) For the dwarfs data set at G=15.0.

If we broaden the analysis of the dwarf set to all $T_{\rm eff}$ (as well as all A_V and all [Fe/H]; the model is not refit), then the average A_V error is 0.13 mag. If we then limit this analysis to $A_V \leq 1.0$ mag, the A_V precision is still 0.12 mag, i.e. hardly worse. In other words, we can equally well determine extinction for high extinction as we can for low extinction stars. It is easier to see trends in the residuals if we average the error (mean absolute residual) at each unique true value: this is shown in Fig. 11.

We see that the systematic errors in $T_{\rm eff}$ and A_V are negligible: overall the systematic is less than 0.003 dex in $\log(T_{\rm eff})$ over the full range of all APs and shows no trend with $T_{\rm eff}$ or A_V. For A_V it is less than 0.03 mag overall, and less than half this for low extinctions or hot stars.

The overall metallicity systematic error is negligible (-0.003 dex) for $T_{\rm eff} \leq 7000$ K (and even only $+0.1$ dex for hotter stars).

With the forward model fitted to the giants sample (model fitted on all A_V, $T_{\rm eff}$ and [Fe/H]), the performance on the cooler stars in the test set is rather similar to the dwarfs:

	[Fe/H]	$\log(T_{\rm eff})$	A_V
$\overline{\delta\phi}$	-0.016	0.0009	-0.0013
$\overline{\lvert\delta\phi\rvert}$	0.42	0.021	0.20
σ_ϕ	0.64	0.033	0.31

ILIUM, giants, G=15, $T_{\rm eff} \leq 7000$ K

The patterns of the residuals (not shown) are also rather similar.

3.4.2 G=18.5

Fitting ILIUM to the dwarfs data set at G=18.5, we get the following summary results for cool stars

	[Fe/H]	$\log(T_{\rm eff})$	A_V
$\overline{\delta\phi}$	-0.38	0.010	0.047
$\overline{\lvert\delta\phi\rvert}$	1.34	0.056	0.52
σ_ϕ	1.81	0.080	0.74

ILIUM, dwarfs, G=18.5, $T_{\rm eff} \leq 7000$ K

In comparison, the mean absolute residuals for [Fe/H] and $\log(T_{\rm eff})$ at zero extinction from CBJ-043 were 0.26 dex and 0.0024 dex respectively. Although the performance on [Fe/H] is very poor, $T_{\rm eff}$ and A_V still give useful results at this magnitude. The summary residuals for the full range of $T_{\rm eff}$ (and [Fe/H] and A_V) are shown in Figs. 12 and 13. The latter shows that there is only a very small systematic error in $T_{\rm eff}$ and A_V: In $\log(T_{\rm eff})$ it shows a rough decreasing trend from 0.03 dex to -0.05 dex as $T_{\rm eff}$ increases over its whole range. The A_V systematic varies more erratically between about ± 0.2 mag.

Fitting ILIUM to the zeromet data set at G=18.5, we get

	$\log g$	$\log(T_{\rm eff})$	A_V
$\overline{\delta\phi}$	-0.22	0.017	0.039
$\overline{\lvert\delta\phi\rvert}$	1.10	0.061	0.30
σ_ϕ	1.53	0.094	0.45

ILIUM, zeromet, G=18.5, full AP range

Comparing this to the results for the same model (and data) at G=15, we see that the mean absolute residual is larger by a factor of 4–5 for all APs. These results can also be compared to the result from the 1D+1D model in CBJ-042 where extinction was fixed to zero: the mean absolute residuals for $\log g$ and $\log(T_{\rm eff})$ there were 0.35 dex and 0.0057 dex respectively. That is, the precisions are degraded by a factor of 3 and 10 respectively due to the introduction of a wide range of extinctions. This $\log g$ precision is almost useless, although we expect inclusions of parallaxes to be able to improve this. The $T_{\rm eff}$ and A_V precision are quite poor and Fig. 14 shows that their residuals are strongly correlated. The correlation is positive: an overestimation of one corresponds to an overestimation of the other. This is actually a symptom of the $T_{\rm eff}-$

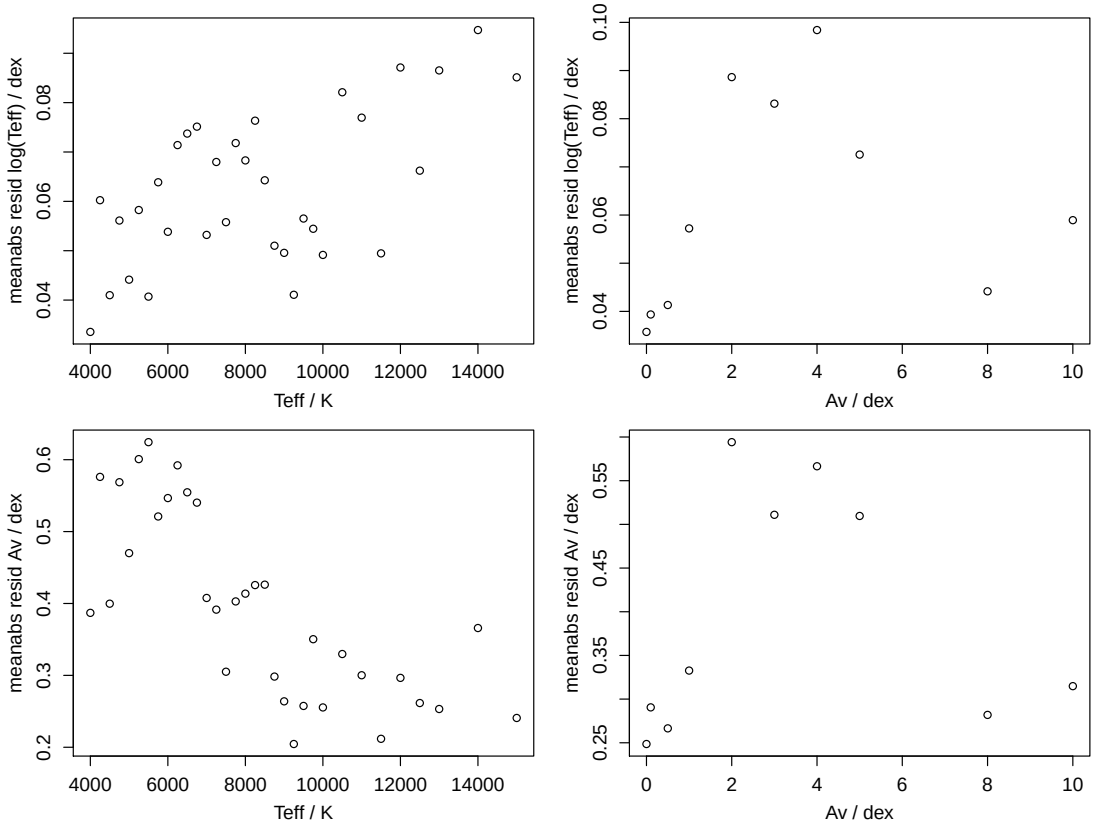

FIGURE 12: The trend in the mean absolute residuals for the dwarfs data set at G=18.5 over the full range of [Fe/H], $T_{\rm eff}$ and $A_{\rm V}$.

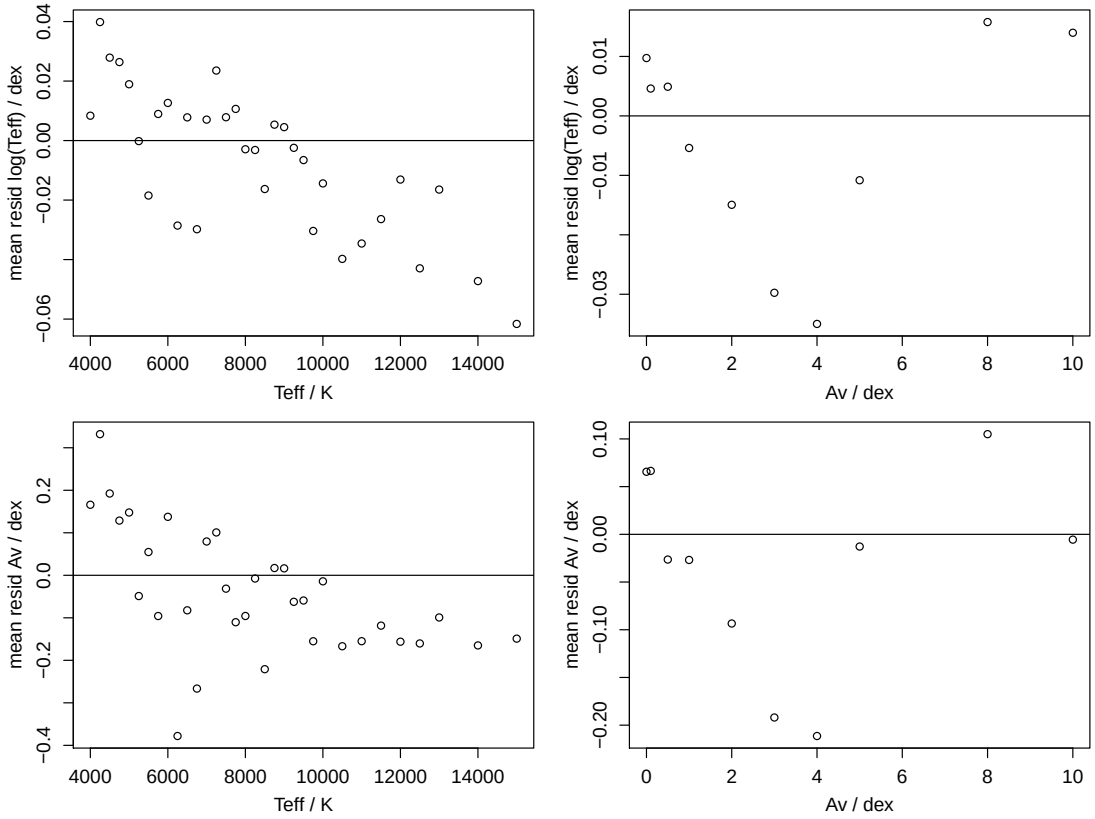

FIGURE 13: As Fig. 12 (G=18.5 dwarfs data set) but now showing the mean residuals (i.e. the systematic errors).

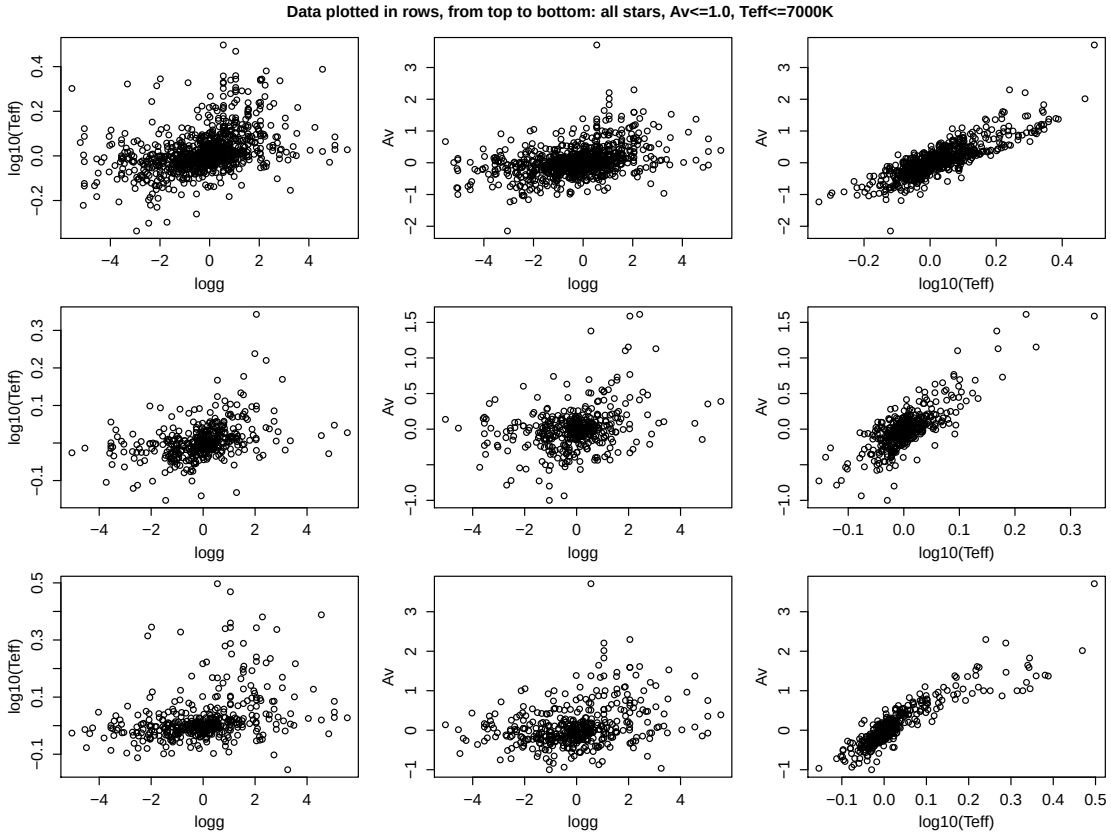

FIGURE 14: The correlations between the residuals from ILIUM for the zeromet data set at G=18.5. The three rows from top-to-bottom are: all stars; stars with $A_V \leq 1.0$ mag; stars with $T_{eff} \leq 7000$ K.

A_V degeneracy which will be discussed in section 4. Essentially, we cannot (even in principle) identify the true T_{eff} and A_V solution, so the particular solution ILIUM converges on will tend to have both T_{eff} and A_V "wrong". We could improve the precision by placing plausible priors on one or both APs.

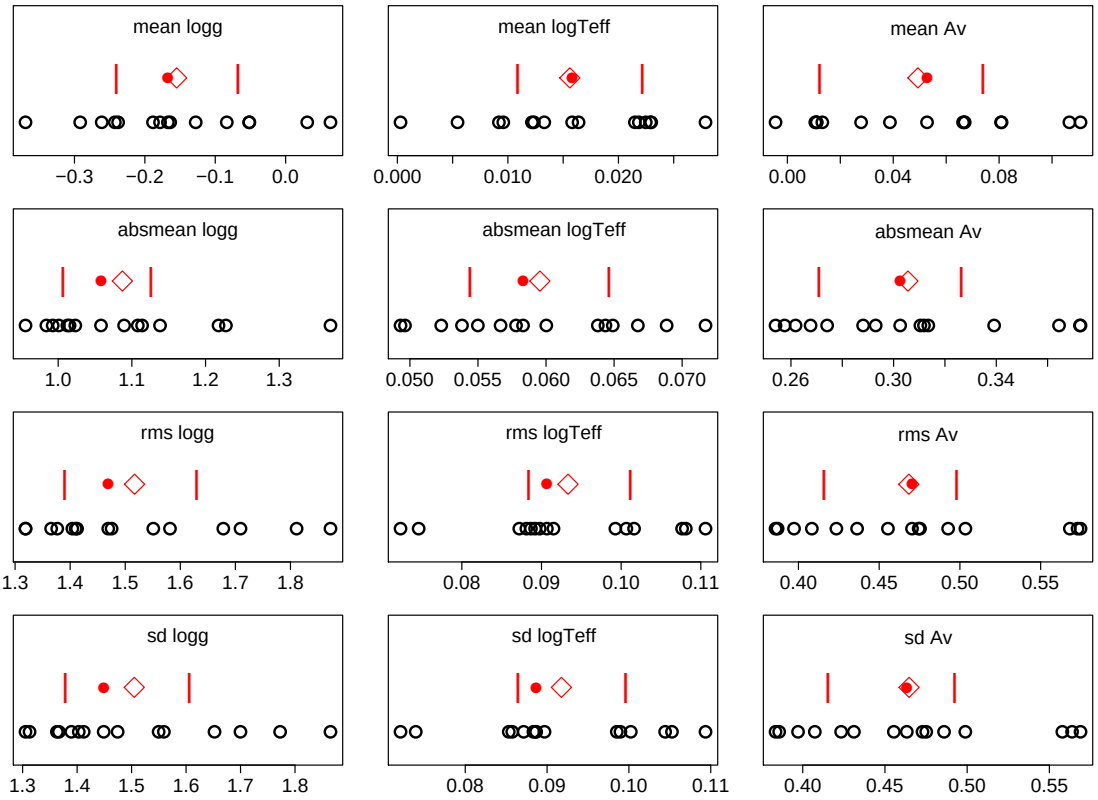

FIGURE 15: ILIUM performance for different randomly selected train and tests on the zeromet data set at G=18.5. The individual runs are shown as open circles. The mean, median and upper/lower quartiles are plotted as a diamond, filled circle and vertical bars respectively. The four rows show four different error statistics, with $\log g$ on the left, $\log(T_{eff})$ in the middle and A_V on the right.

In CBJ-046 I showed that the summary statistics ($\overline{\delta\phi}$, $\overline{|\delta\phi|}$, σ_ϕ) for a given model depended on the randomly selected sample of objects used in the test set and in the (distinct) set from which the nearest neighbour is selected (for initialization of ILIUM). There, the variation in $\overline{|\delta\phi|}$ for $\log g$ and $\log(T_{eff})$ on the zeromet sample (with zero extinction) was 5–10% (section 4.2 and Fig. 5 in CBJ-046). Here I repeat the same experiment but now with the full range of extinction, and only for a subset of 100 stars (rather than 1000 used for the test samples in the rest of this TN). The result is shown in Fig. 15. The variance over the inter-quartile range is around 15–20% for the mean absolute error, larger than for the 1D+1D model. But this is at least in part because of the greater variance between the smaller test sets used here.

As a comparison for ILIUM I run a leave-one-out nearest neighbour algorithm on the entire grid (2740 objects) of noisy stars. The results are:

	[Fe/H]	$\log(T_{\rm eff})$	A_V
$\overline{\delta\phi}$	0.072	0.0056	0.059
$\overline{\lvert\delta\phi\rvert}$	1.22	0.073	0.44
σ_ϕ	1.60	0.12	0.81

nearest neighbour, zeromet, G=18.5

This algorithm has the advantage over ILIUM in that all of the APs being tested for occur at exactly those values in the template grid (the template grid is the entire original grid apart from the star being tested), but a disadvantage in the sense that the templates are noisy. (Running the algorithm with noise-free templates is an unfair and inappropriate comparison).

3.4.3 G=20

The performance at G=20 is naturally worse. On the dwarfs data set, the mean absolute error on $\log(T_{\rm eff})$ and A_V is 0.12 dex and 0.84 mag respectively (full range of all APs), i.e. we can only obtain rough estimates of the strong APs (see Fig. 16 for the trends). The error in $\log(T_{\rm eff})$ increases from 0.06 to 0.18 (14% to 40%) from 4000 K to 15 000 K (for dwarfs), but there is now a large systematic error which also shows a negative trend with $T_{\rm eff}$ (Fig. 17). A_V can be estimated to an overall precision of 0.6 to 1.1 mag, with a high negative systematic at large $T_{\rm eff}$. These large errors and especially the systematic errors are at least in part a consequence of the $T_{\rm eff}$–A_V degeneracy. [Fe/H] cannot be estimated at all (the formal mean absolute error is 2 dex at all $T_{\rm eff}$).

Note that at G=18.5 and G=20, $T_{\rm eff}$ can be estimated considerably more accurately (lower random and systematic errors) for cooler stars than for hotter stars.

4 Revelation of the $T_{\rm eff}$–A_V degeneracy

Like essentially all AP estimation algorithms, ILIUM works by finding the best single solution. "Best" is defined in ILIUM as that forward-model predicted spectrum which has the minimum distance from the measured spectrum.[2] However, on account of the low spectral resolution of BP/RP and the presence of noise, it is possible that there are degeneracies between the APs. By degeneracy I mean two or more combinations of APs which give rise to the same measured spectrum (to within some noise-related threshold). In particular we expect that $T_{\rm eff}$ and A_V may be degenerate (as they are in many photometric systems). In this section I explore the extent of this degeneracy using two methods.

[2] Incidentally, "distance" is not the usual Euclidean distance (a quadratic measure, $|x|^2$), but rather its unit power equivalent ($|x|$), the "Manhattan" distance metric. This follows from the definition of the algorithm in section 2.2 and 2.3 of CBJ-042. ILIUM does not explicitly minimize the distance, but makes iterative steps which are expected to reduce it.

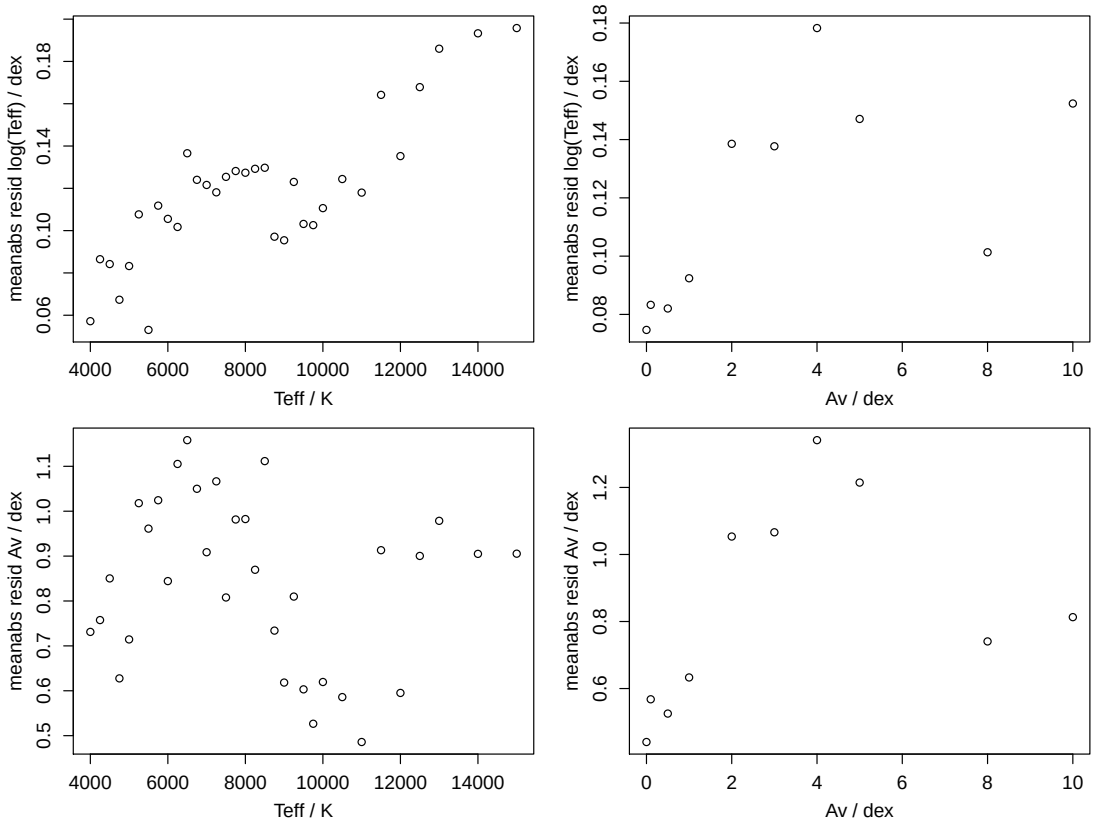

FIGURE 16: The trend in the mean absolute residuals for the dwarfs data set at G=20 over the full range of [Fe/H], $T_{\rm eff}$ and $A_{\rm V}$.

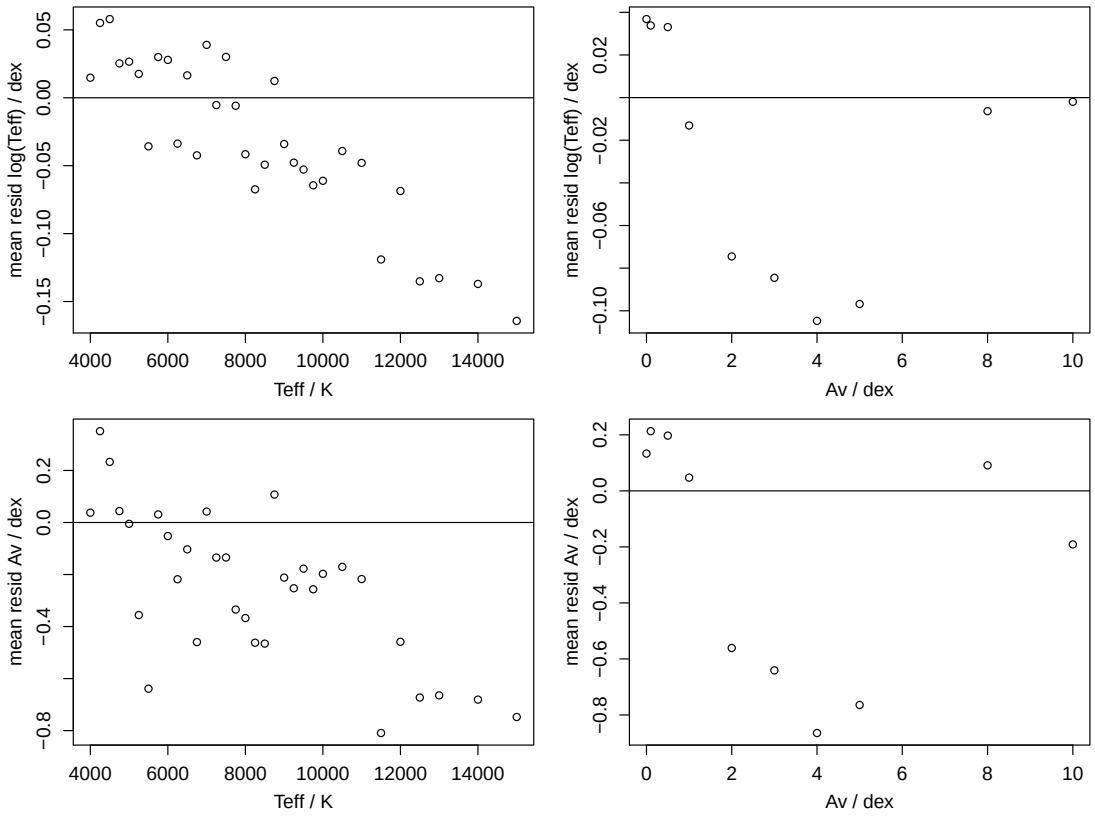

FIGURE 17: As Fig. 16 (G=20 dwarfs data set) but now showing the mean residuals (i.e. the systematic errors).

4.1 Multiple random initializations

ILIUM is usually initialized via the nearest neighbour in a grid of template spectra. One way of investigating the degeneracies is to use a different initialization and examine whether ILIUM converges on a common solution. As ILIUM is only a local search method (see Fig. 1 of CBJ-042), it is a priori possible that ILIUM could converge on different solutions.

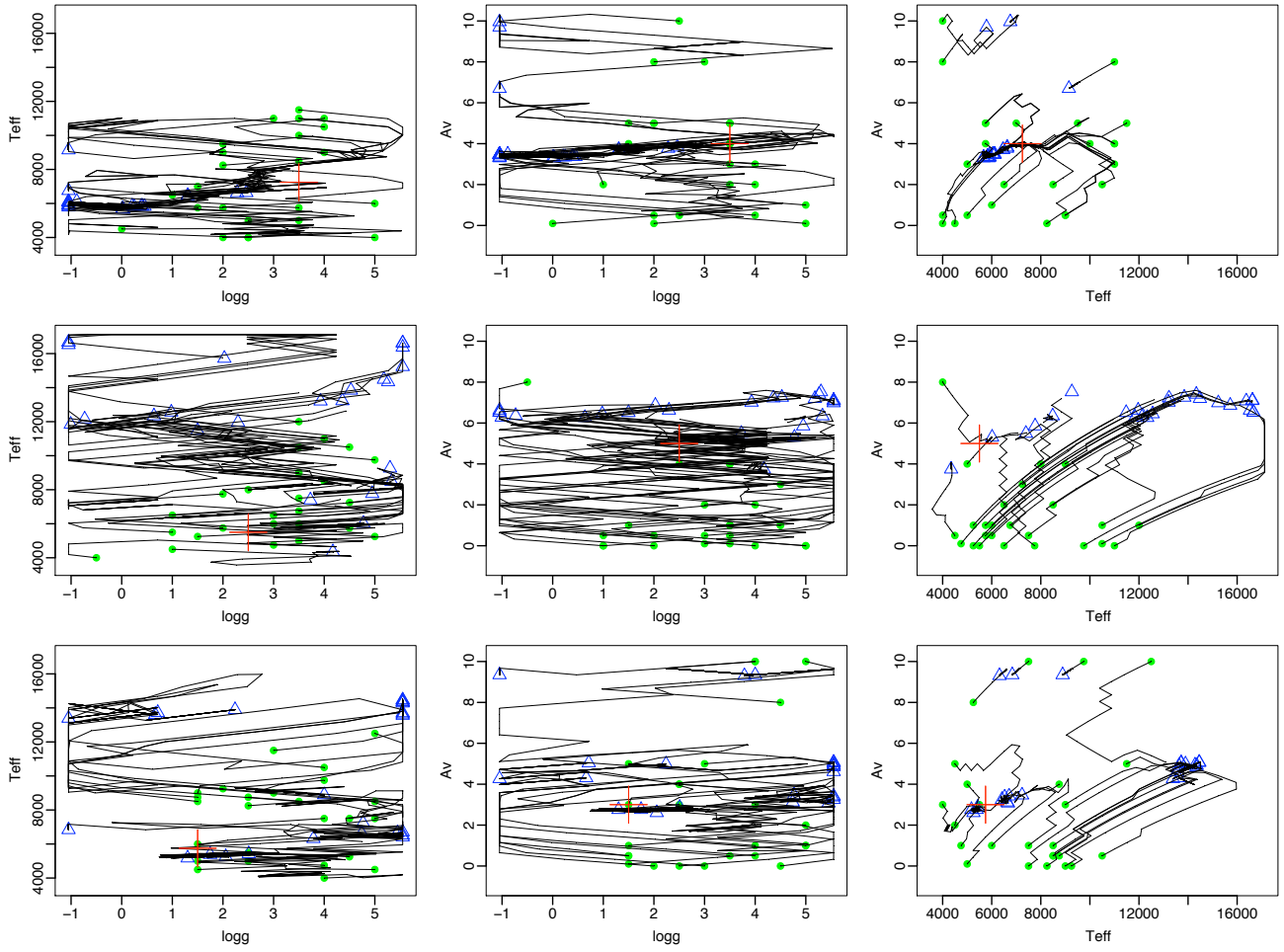

FIGURE 18: Evolution of the three APs over the ILIUM iterations for 3 stars (plotted in the three rows). The three columns just plot the different combinations of the three APs against each other. For each star the true APs are shown with a red cross. The 25 random initializations are shown as green points and the final solutions from each as blue triangles, connected by a black zig-zag line which traces the AP evolution.

I therefore set up ILIUM to classify 10 stars from the zeromet data set at G=18.5, and ran it 25 times randomly selecting a different star each time to initialize it. Fig. 18 shows the evolution of the APs during the 20 ILIUM iterations for three of these stars. Fig. 19 is the same but just

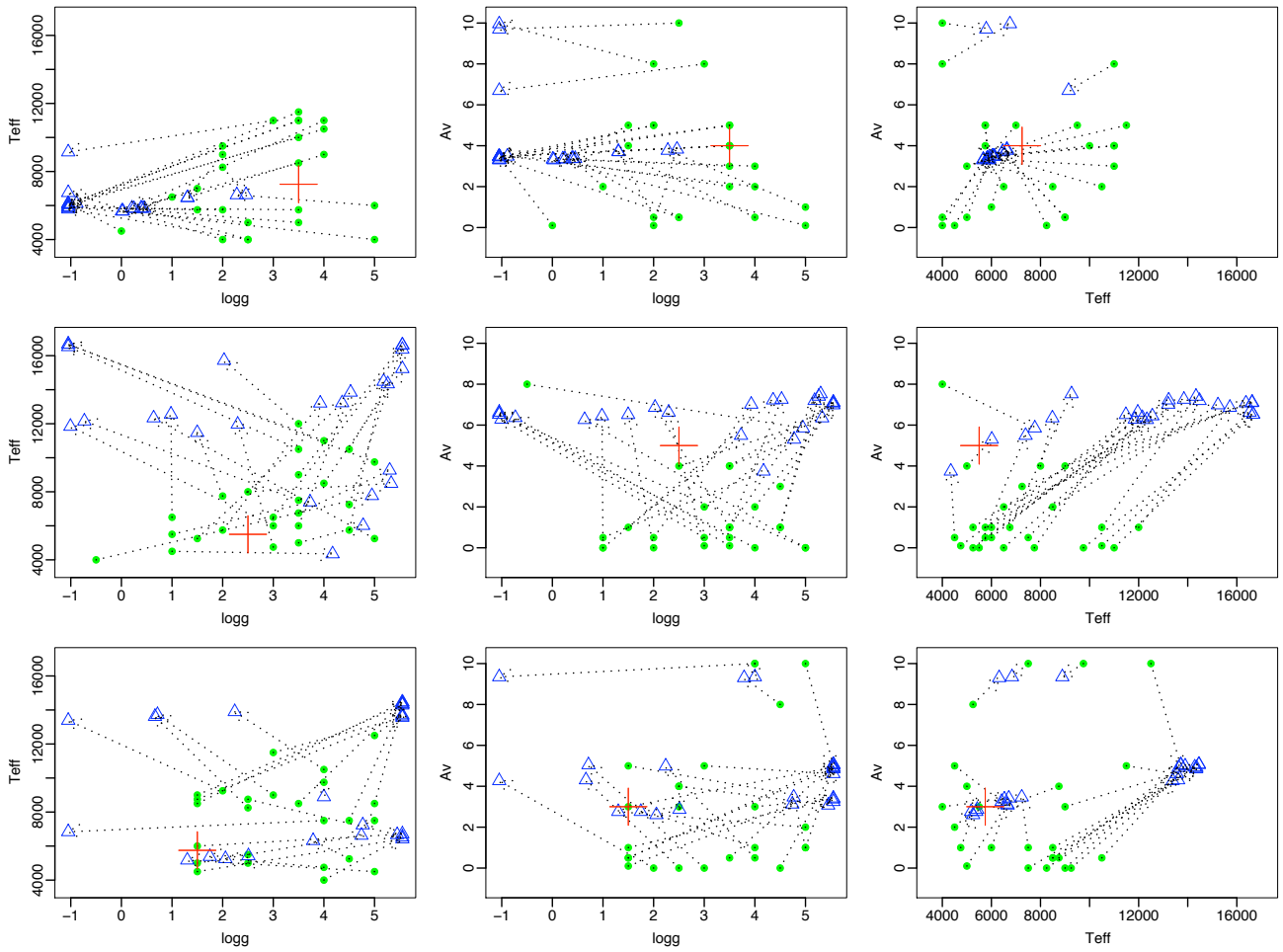

FIGURE 19: As Fig. 18 but now showing just the initializations (green circles) and final solutions (blue triangles) connected with a dashed line.

showing the initializations and the end points in each case. The main panels of interest are those in the right-most column, showing $T_{\rm eff}$ and A_V. For the first star (top row), we see that 22 of 25 initializations over quite a range of $T_{\rm eff}$ and A_V converge on a small region (it's a bit offset from the true solution). For the second star we see a complete lack of convergence on a single solution, and rather a ridge of solutions. For the third star we see two distinct regions of convergence, one of which corresponds to the true APs. In none of these examples does ILIUM get good convergence on $\log g$. As discussed above, $\log g$ performance is generally poor at G=18.5, although there are examples where there is good $\log g$ convergence.

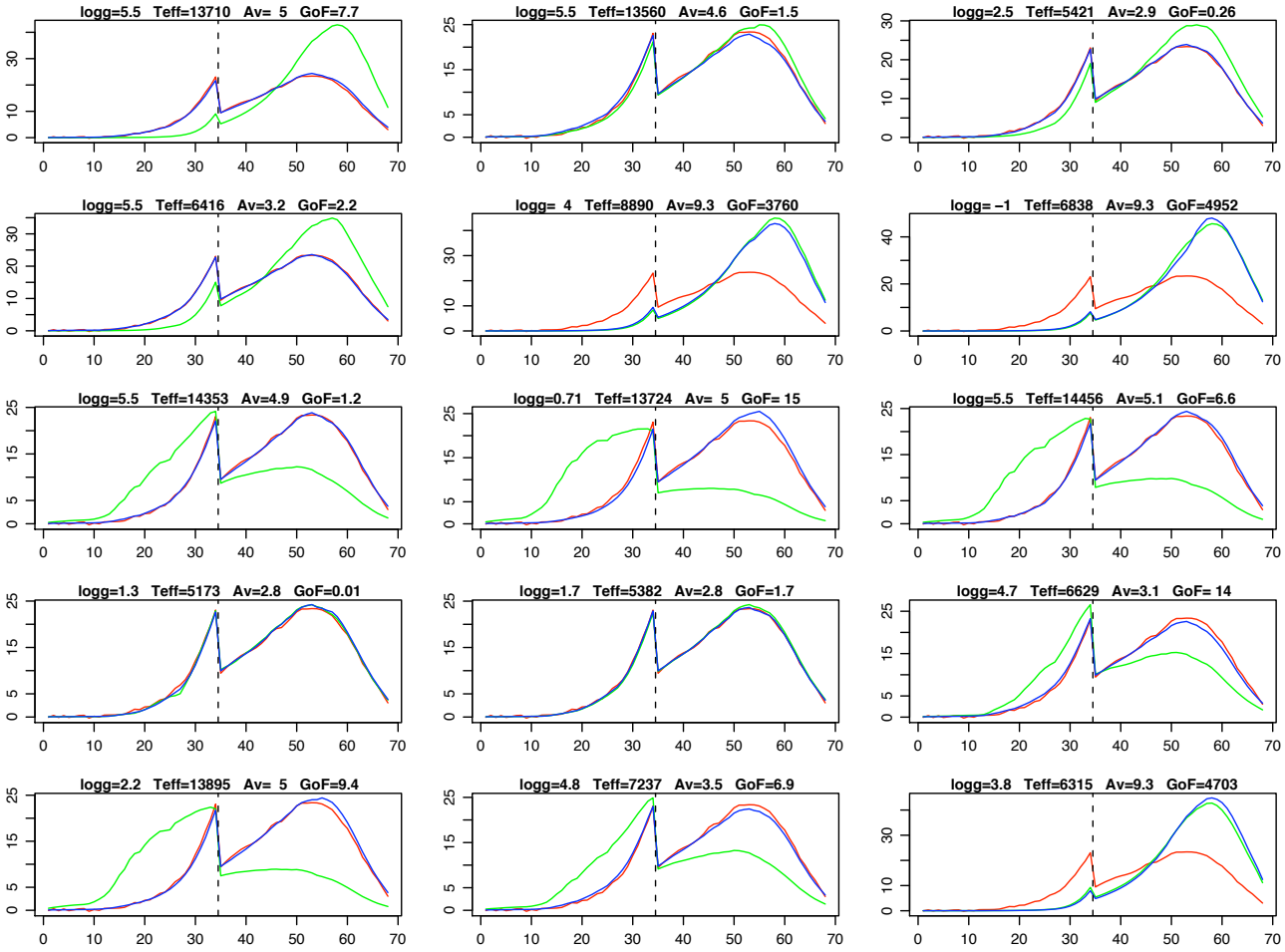

FIGURE 20: The spectra corresponding to the APs plotted for the third star (bottom row) in Fig. 19 (just 15 of the 25 runs are shown). The red line is the true spectrum with $\log g = 1.5$ dex, $T_{\rm eff} = 5750$ K, $A_V = 3.0$ dex (same in all panels), the green and blue lines are the initial and final spectra respectively. At the top of each panel are written the APs of the solution (final spectrum) and its GoF

What has ILIUM done in all these cases? For example, in the bottom row, are the solutions at

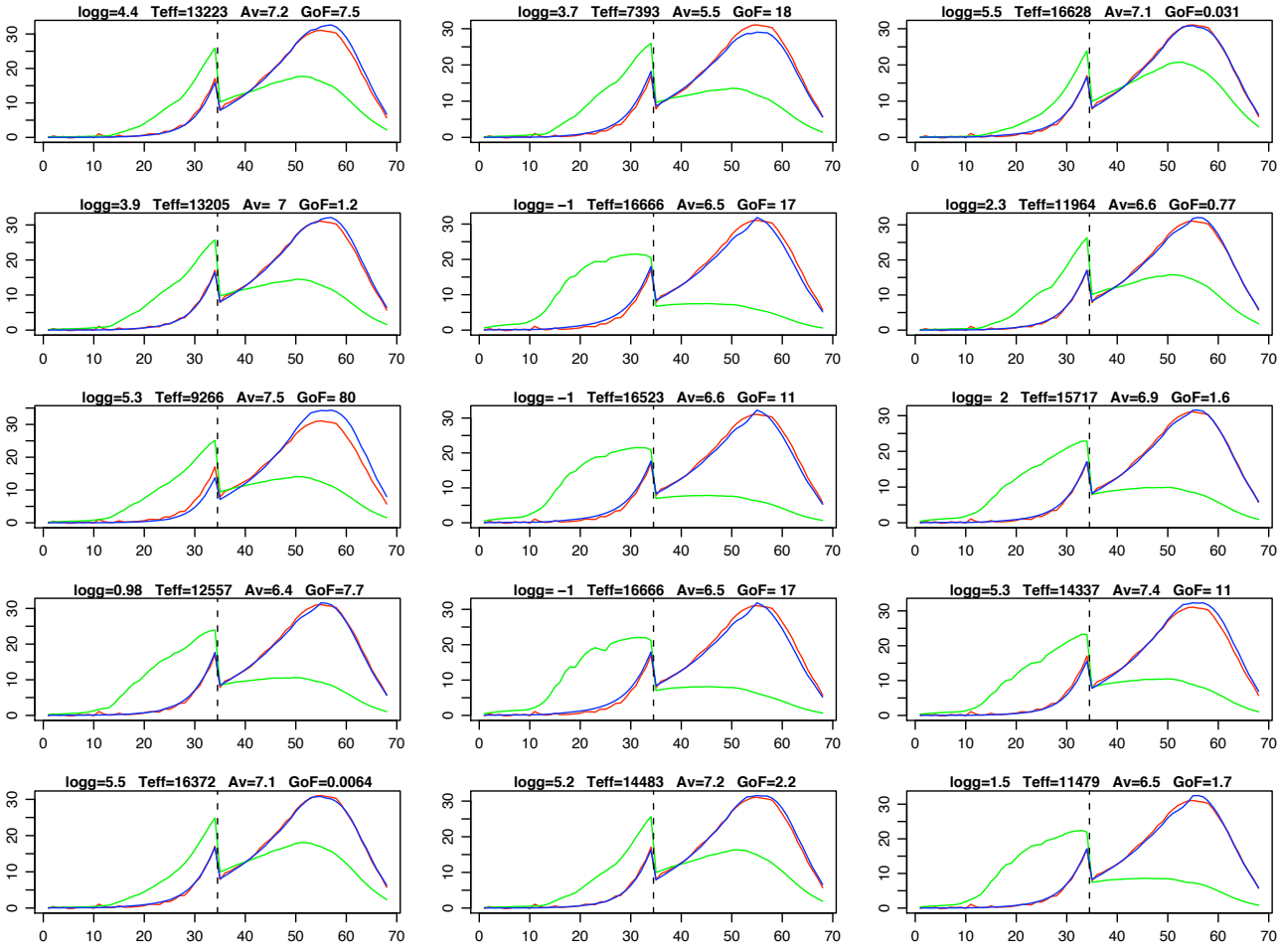

FIGURE 21: As Fig. 20 but for the second star (middle row) in Fig. 19 (ridge of solutions) which has true APs $\log g = 2.5$ dex, $T_{\rm eff} = 5500$ K, $A_V = 5.0$ mag

around $T_{\rm eff} = 13\,000$ K (call these S13) simply poor solutions compared to those around 6000 K (call these S6)? We can answer this by inspecting the corresponding spectra predicted by the forward model, shown in Fig. 20. The predicted spectra for the S13 ("wrong") solutions are as similar to the true spectrum as are the S6 ("correct") solutions. In other words, these solutions are degenerate in the APs $T_{\rm eff}$ and A_V. This is not a problem of ILIUM, but rather an intrinsic degeneracy in the data. We can quantify their similarity by looking at the goodness-of-fit (GoF) values given in the panels of the spectral plots. (GoF is the reduced, defined in section 2.6 of CBJ-042.) We see that we often get equally good GoFs for S13 as we do for S6, with values ranging from 0.01 to 15. Also shown in the spectral plot are the three solutions at high A_V which predict spectra very different from the true spectra. In these cases ILIUM was initialized too far from a good solution to be able to converge on anything like a similar spectrum. However, we can easily identify (and eliminate) these poor solutions from their very high GoF values (over 3000 in all these three cases).

Note the sign of the degeneracy: S13 has both higher $T_{\rm eff}$ and higher A_V, i.e. a positive correlation. Another symptom of this degeneracy was shown in Fig. 14 as a positive correlation between the $T_{\rm eff}$ and A_V residuals. That is, if we overestimate one of these APs we tend to overestimate the other, because this will then correspond to a spectrum which is very close to the observed spectrum. This leads us to suspect that the ridge of solutions for the second star in Fig. 19 is also a consequence of a degeneracy. In Fig. 21 I show 15 of the corresponding spectra, and inspection of these confirms this suspicion. Most of these solutions correspond very closely to the true spectrum, with very low GoF values, but have significantly different APs. Indeed, the closest solution in terms of the strong APs is at $T_{\rm eff} = 6015$ K and $A_V = 5.3$ and has GoF = 1.2, larger than several other solutions. (The closest solution in the $T_{\rm eff}$–$\log g$ plane has GoF = 18 and is plotted in the middle of the top row of Fig. 21.) Identification of similar ridges in plots of multiple solutions for other stars (not shown) suggests that such degeneracy ridges may be widespread.

4.2 Systematic mapping of the $T_{\rm eff}$–A_V degeneracy

How universal is this $T_{\rm eff}$–A_V degeneracy? To systematically map it we can measure some distance between a given spectrum and all other spectra in a grid and identify which are close relative to the distance we would expect just due to photometric noise. To do this we don't need the iterative updating part of ILIUM, just the forward model. I use it to generate spectra, $\boldsymbol{p}$, over the full $\log(T_{\rm eff})$ and A_V ranges in steps of 0.01 dex and 0.2 mag respectively, corresponding to 51 steps and 58 steps respectively, a total grid of 2958 spectra (let's call this the "dgen" grid). The weak APs are fixed at $\log g = 4.0$ and [Fe/H] = 0.0 dex. I then create noisy versions of each of these spectra using the sigma spectra, $\boldsymbol{\sigma}$, from the original BASEL+MARCS grid. (For simplicity I just use the sigma spectrum of the star with the nearest APs to each star in the dgen grid.) Interpreting this as the 1σ of a zero mean Gaussian noise model, I create a noise vector by drawing Gaussian random variables from this, and then add this vector to the noise-free spectrum to create the noisy spectrum $\boldsymbol{n}$. For each of the noise-free stars I calculate

the (squared) distance to each of the noise-free stars as

$$D^2_{cpw} \;=\; \delta\boldsymbol{p}^T\, C_p^{-1}\, \delta\boldsymbol{p} \tag{1}$$

where $\delta\boldsymbol{p} = \boldsymbol{p} - \boldsymbol{n}$ is a $I \times 1$ column vector (I is the number of bands) and C_p is the covariance matrix of $\boldsymbol{p}$. Here it is diagonal, $C_p = \mathrm{diag}\{\sigma^2_{p_i}\}$, so the equation simplifies to

$$D^2_{cpw} \;=\; \sum_{i=1}^{i=I} \left(\frac{p_i - n_i}{\sigma_{p_i}}\right)^2 \tag{2}$$

C_p of course has dimensions of p^2 making D^2_{cpw} dimensionless. Note that D^2_{cpw} is equal to $I - 1$ times the GoF (goodness-of-fit) defined in equation 7 of CBJ-042. "cpw" stands for "covariance-matrix-in-p weighted".

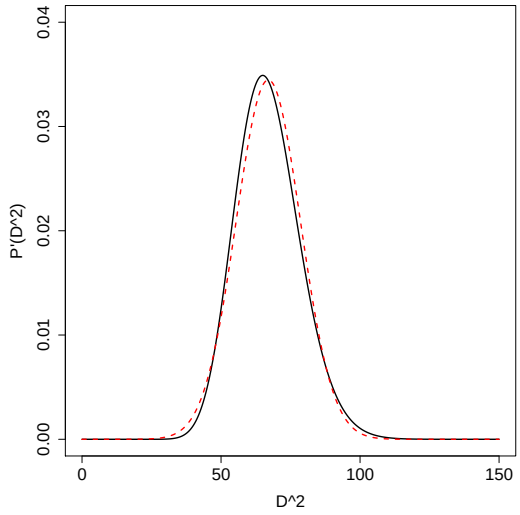

FIGURE 22: The black line is the normalized χ^2 probability density function for $\nu = 67$ degrees of freedom, which is the probability distribution for D^2_{cpw} under the assumption of Gaussian errors. $P(D^2_x) = \int_0^{D^2_x} P' dD^2$ is the distribution function, the probability that $D^2 \leq D^2_x$. The red dashed line is the Gaussian with the same mean (ν) and variance (2ν)

A degeneracy arises between two stars with different APs 2because D^2_{cpw} is sufficiently small that the difference could just be due to photometric noise. We need to quantify "sufficiently small". Under the null hypothesis that the differences between the spectra are only due to Gaussian noise and that each pixel is independent, then D^2_{cpw} follows a χ^2 distribution with $I - 1 = 67$ degrees of freedom.[3] The probability density function of D^2_{cpw} is shown as the black

[3] χ^2 is the distribution followed by a sum of squares of ν independent unit Gaussian variables, $N(0,1)$. For ν degrees-of-freedom its density function is $P'(D^2) = D^{\nu-2}e^{-D^2/2}$ (to within a normalization constant). I use $\nu = I - 1$ degrees of freedom rather than I because the spectra are standardized, a process which "uses up" a degree of freedom by bringing them to a common scale.

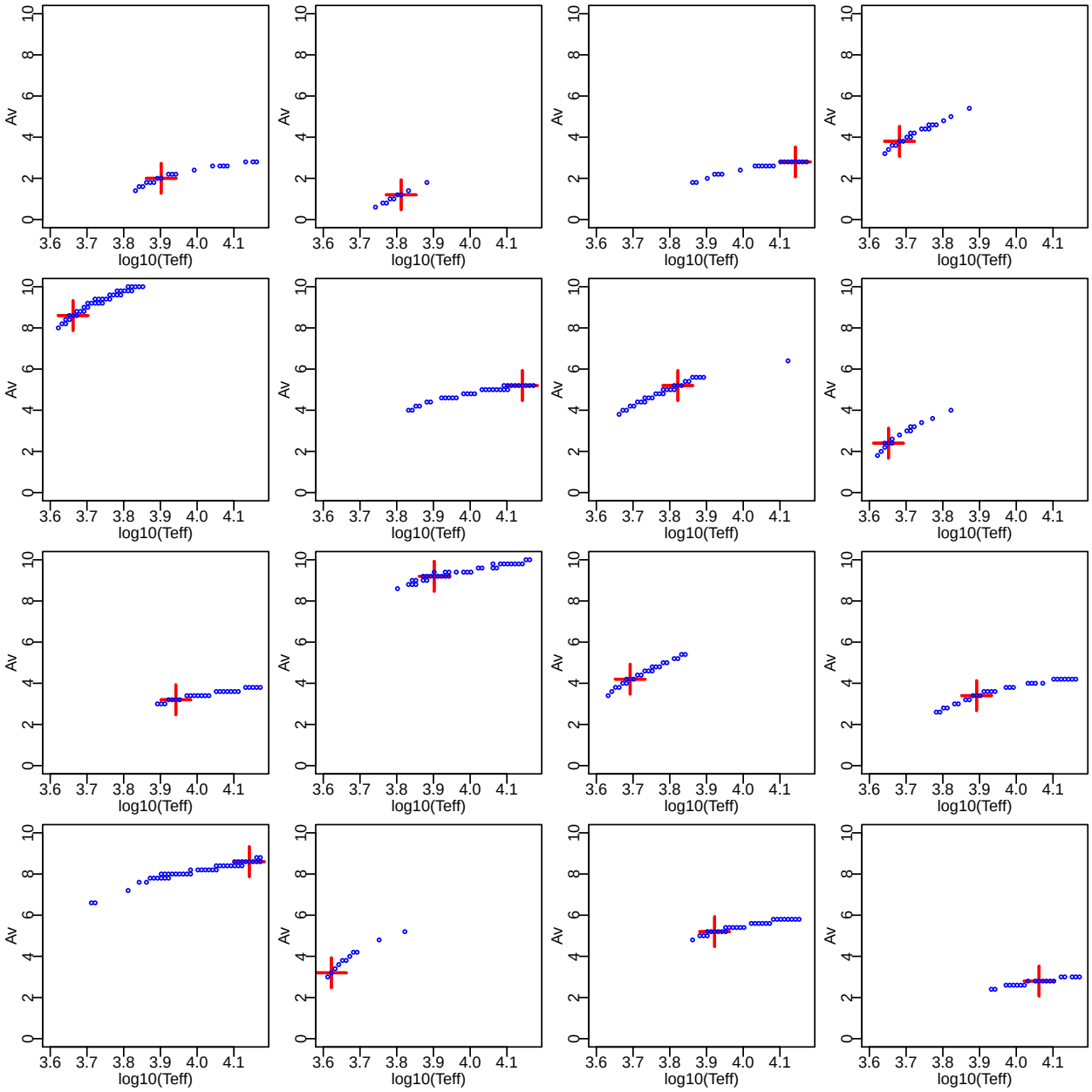

FIGURE 23: Map of the $T_{\rm eff}$–$A_{\rm V}$ degeneracies for 16 stars. In each panel, the stars with APs plotted in blue have a (noisy, G=18.5) spectrum which agrees with the (noise-free) spectrum of the star plotted with the red cross to within a probability of 99.9% (assuming the noise to be Gaussian), which corresponds to their spectra being separated by a distance $D^2_{cpw} \leq 108.5$

line in Fig. 22. Let us define any two stars as degenerate if their (squared) distance is such that the probability of getting this distance or less under the null hypothesis is 0.1% or more. The critical value of D^2_{cpw}, call it H^2_{cpw}, is given by $1 - P(H^2_{cpw}) = 0.001$, where $P(D^2_{cpw})$ is the distribution function, viz. the integral of the curve in Fig. 22 from zero up to D^2_{cpw}. For $\nu = 67$ degrees of freedom, $H^2_{cpw} = 108.5$.

Using this I plot the degeneracy map for each of 16 "target" stars from the grid in Fig. 23. The target star is shown in red, and in blue are plotted all the other stars (the "candidate" stars) which are degenerate with it, i.e. have $D^2_{cpw} \leq H^2_{cpw}$. Note that I use the noise-free spectrum for the target star and the noisy spectra for the candidate stars. C_p is the value for the target star. In each case we see that the degenerate stars lie on a narrow locus in $\mathrm{T_{eff}}$–$\mathrm{A_V}$ space with a common orientation across the whole space. H^2_{cpw} corresponds to a value of the GoF (reduced χ^2) of $H^2_{cpw}/(I-1) = 1.62$. This seems small when we consider the values of the GoF show in Figs. 20 and 20 for apparently very similar spectra. Yet this is a consequence of having so many bands and demanding that noise variations be strictly Gaussian in all of them. Naturally if we use a larger probability threshold (or had larger expected errors) then the band gets wider. [4]

Instead of just plotting the candidate stars closer than some threshold, we can plot all stars and denote their distance using a colour scale. This is shown in Fig. 24 where I actually plot the *inverse* (squared) distance, $1/D^2_{cpw}$. (As a similarity measure it better shows the nearest neighbours than using the distance, although it is equivalent to changing the colour scale.)[5] These plots show that there is a strong degeneracy for all of these stars. By strong, I mean that it extends over a large fraction of the AP range. If we repeat this for all the stars in the grid we find that this $\mathrm{T_{eff}}$–$\mathrm{A_V}$ degeneracy is ubiquitous. The implication is significant: it makes little sense to report a single pair of $\mathrm{T_{eff}}$ and $\mathrm{A_V}$ values for a star. Rather we must report a whole ridge of solutions. As we can map these degeneracies in advance, the purpose of a classifier such as ILIUM would be to identify which of these degeneracy ridges is the appropriate one for the observed spectrum. That is, ILIUM would effectively report which red cross we have, and a look-up table would give the blue ridge. A single nearest neighbour initialization of ILIUM is generally adequate for this, but as we do get some non convergence at fainter magnitudes, we may want to investigate multiple initializations to ensure we don't "miss" the ridge. If we had prior information, we could limit the $\mathrm{T_{eff}}$ and $\mathrm{A_V}$ values we report, or rather, we could assign probabilities across the ridge. Note the orientation of the ridges: they have a low inclination, spanning a large fraction of the $\mathrm{T_{eff}}$ range compared to the $\mathrm{A_V}$ range (although recall that the maximum $\mathrm{T_{eff}}$ used is 15 000 K). This implies that a prior on $\mathrm{T_{eff}}$ would generally be more useful than a prior on $\mathrm{A_V}$. This is good, because $\mathrm{T_{eff}}$ can also be estimated for many stars from higher resolution spectroscopy (e.g. RVS on Gaia for the very bright stars). On the other hand,

[4] In the first version of this TN I plotted the degeneracy map by plotting all candidate star with distances up to 5 times the expected value of D^2_{cpw}, i.e. $5 \times 68 = 340$. We see from Fig. 22 that the probability of getting values this high is essentially zero, so the degeneracy band I plotted there was, for the Gaussian assumption, too wide.

[5] Once we have defined a probability model – as discussed in section 4.2.2 – we can plot the (log) likelihood function in the same way. This is shown in the ILIUM article currently in preparation.

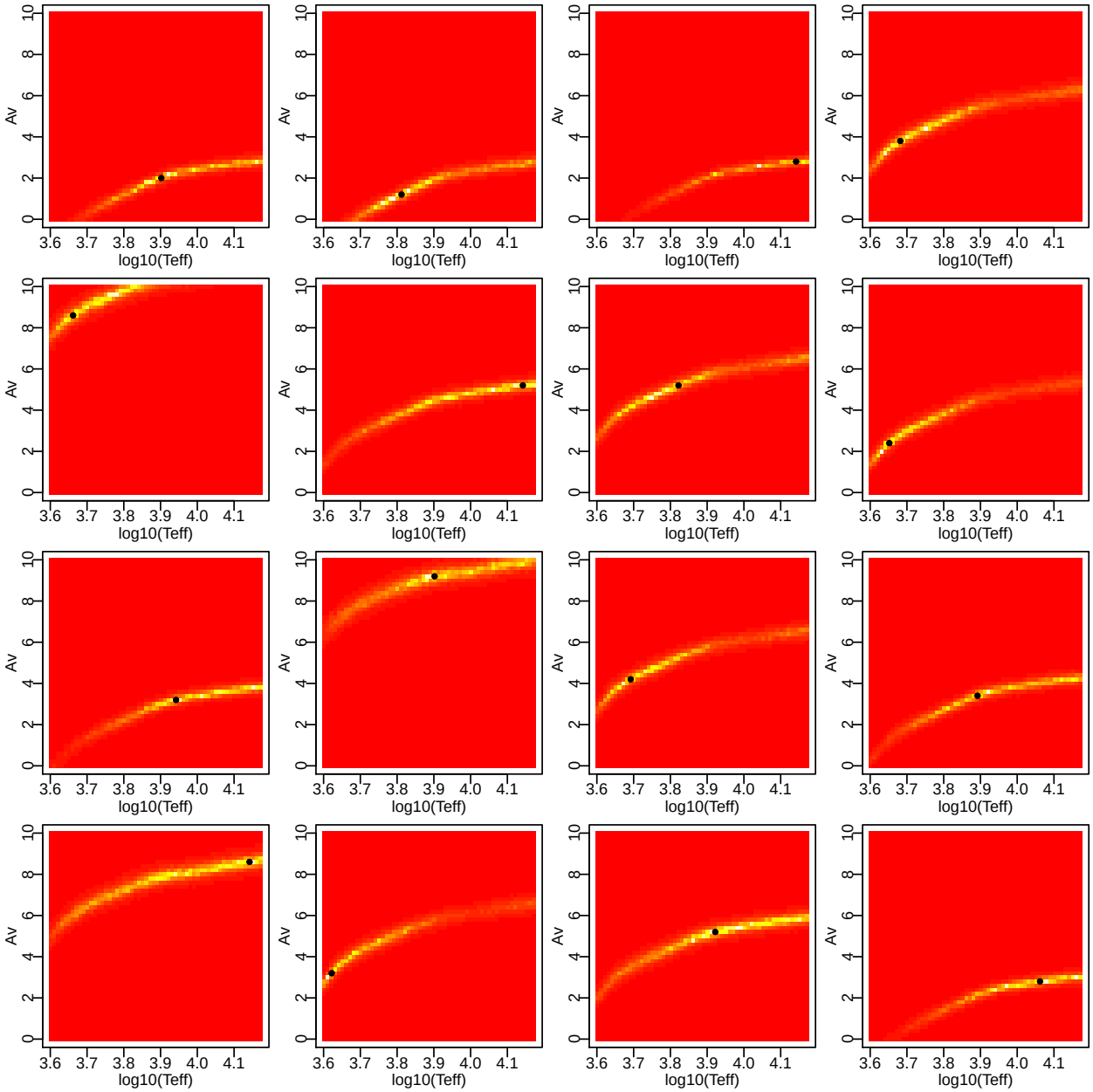

FIGURE 24: Map of $T_{\rm eff}$–A_V degeneracies for the same 16 stars shown in Fig. 23 but now plotting the *inverse* of the distance metric D^2_{cpw}. The colour heat scale is proportional to the inverse (squared) distance, with white the smallest distance and red the largest. The true APs are indicated by the black point.

if we could assume that the extinction is very low (or even zero) for some stars, then we remove most of the degeneracy, and earlier TNs have shown that performance is then very good on $\mathrm{T_{eff}}$, $\log g$ and feh.

4.2.1 Use of a sensitivity-weighted distance estimate

The above degeneracy map simply used the noise-weighted (squared) Euclidean distance as a measure of similarity of two spectra (equation 1). Yet this metric does not discriminate between the effects of $\mathrm{T_{eff}}$ and $\mathrm{A_V}$ on the spectrum. For example, we might distinguish between a reddened hot stars and an unreddened cool stars by the strong TiO bands present in the latter and visible to some degree in RP despite the LSF, depending on the SNR (Lopez Marti et al. 2009). ILIUM of course does precisely this by using sensitivity weighting (one example of this impact can be seen in Figure 10 in CBJ-042). So perhaps we can use the sensivities to define a more discriminative metric and thus reduce the extent of the degeneracies?

We could try to introduce the $(I \times J)$ sensitivity matrix, S, (at $\boldsymbol{p}_0$) defined in equation 3 of CBJ-042, by replacing

$$C_p \;\rightarrow\; S\, C_\phi\, S^T \tag{3}$$

in equation 1. C_ϕ is the covariance matrix of the APs at $\boldsymbol{\phi}_0$, i.e. for the spectrum corresponding to $\boldsymbol{p}_0$. It could be calculated using equation 17 in CBJ-046, which gives the estimated covariance in the ILIUM solution for the APs given the photometric errors. The new expression in equation 3 has the right dimensions, but it is of course singular because $J < I$. (J is the number of APs.) SS^T is likewise singular. We can avoid this by adding to it a positive definite matrix (Lindegren 2003, Brown 2005). Any covariance matrix is strictly non-negative definite (semi-positive definite), and in practice all are positive definite because the modelled uncertainties are always greater than zero. Lindegren (2003) suggests to add C_p, which is a logical choice with the right dimensions. (We could also consider adding any scalar multiple of C_p). With this I define a sensitivity-weighted (squared) distance measure as

$$D^2_{sw} \;=\; \delta\boldsymbol{p}^T\, R\, \delta\boldsymbol{p} \tag{4}$$

where

$$R \;=\; (S\, C_\phi\, S^T \;+\; C_p)^{-1} \tag{5}$$

(In their notes, Brown and Lindegren both swap the meaning of p and ϕ relative to what is used here and in the other ILIUM technical notes. Note also the erroneous inverse (typo) in the definition of the elements of the two covariance matrices in Lindegren 2003.) For the 2D+1D ($J = 3$) model introduced in this TN the sensitivity matrix is normally 3×3. Here we are only interested in the sensitivity to $\mathrm{T_{eff}}$ and $\mathrm{A_V}$ so we just use the 2×2 part of the full matrix with $\log g$ fixed.

There are some issues here: What do we use for C_ϕ? We can't really use equation 17 of CBJ-046 as this is an expression for the uncertainties in the APs *after* we have estimated them. So

here I simply use a diagonal form with 0.05^2 as the two diagonal elements. As I am working with standardized variables this corresponds to assuming an uncertainty in the APs of order 5%, but this is somewhat arbitrary and an arbitariness which is reflected in the degeneracy map. Second, what value do we use for the sensitivity matrix, S, in equation 5? I simply use the value calculated at the target star (rather than the candidate star).

I now calculate D^2_{sw} for the same 16 target stars as shown in Fig. 23. Even when we have Gaussian errors we do not expect D^2_{sw} to have a χ^2 distribution. Nonetheless, to provide a comparison to the previous case I again plot those candidate stars which lie within a distance, H^2_{sw}, which is defined as D^2_{sw} calculated setting $\delta\boldsymbol{p}$ to the expected noise and multiplying by $108.5/I = 108.5/68 = 1.60$. The reason for this factor is that in the limit of $C_\phi = 0$, $D^2_{sw} = D^2_{cpw} = I = 68$, so H^2_{sw} will again be 108.5. These degenerate stars are plotted in Fig. 25. There is clearly some arbitrariness in this selection. The corresponding inverse distance (squared) plot is shown in Fig. 26. These shows some interesting patterns, more complex than those with the noise-weighted distance metric.

But does it give fewer degeneracies? Comparing either plot with the corresponding noise-weighted case, it appears that the degeneracy region is larger when using the sensitivity-weighted distance. This is not what I expected: I had hoped that the sensitivity weighting would be more discriminant and thus reduce the number of degeneracies. In fact it may have, because as already mentioned we don't expect D^2_{sw} to follow a χ^2 distribution, so comparing figures Fig. 23 and Fig. 25 may not be a fair comparison. Based only on these plots we cannot yet say whether use of the sensitivities defines a smaller degeneracy region.

The actual value of the critical value H^2_{sw} ranges from 103.4 to 107.1 for the 16 stars, compared to $H^2_{cpw} = 108.5$. This suggests that the sensitivity part of R in equation 1 makes up between 1% and 5% of the total distance measure when using $C_\phi = \mathrm{diag}(0.05^2)$. The fact that R has a larger magnitude when we include a sensitivity term does not in itself explain why the band is wider when using D^2_{sw} rather than D^2_{cpw}, because in both cases we select the stars within some factor of the expected distance *for that metric*. That is, a larger R gives smaller distances, but the distance limit, H^2, is also smaller. Rather it is the fact that we have a constant term in R in addition to the photometric noise variance which does not "cancel" with the noise part of $\delta\boldsymbol{p}$. If I increase the diagonal elements of C_ϕ to 0.5^2, i.e. assume the errors in the APs are a factor of 10 larger, then the degeneracy bands get wider, even though the critical value H^2_{sw} is recalculated (it now lies between 98.9 and 106.8). This is because there is now an even larger term in R which cannot "cancel". This further suggests that simply defining a critical value which tries to correspond to the 99.9% limit for the χ^2 metric D^2_{cpw} is inappropriate.

It's actually not that clear whether D^2_{sw} is a valid distance metric, partly because it's not clear which values of S and C_ϕ should be used. This will be considered in future work.

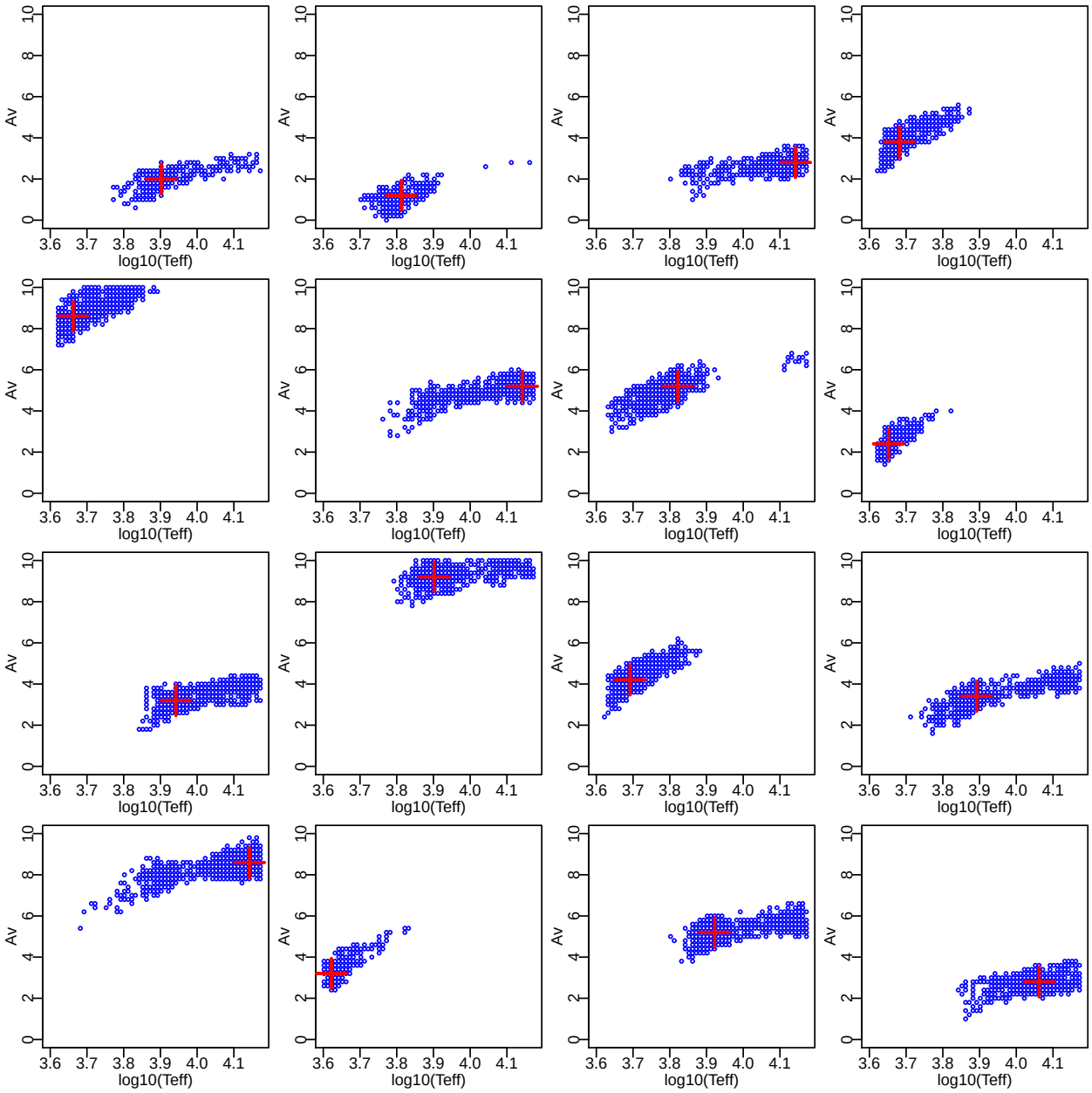

FIGURE 25: As Fig. 23 but now using the sensitivity-weighted distance

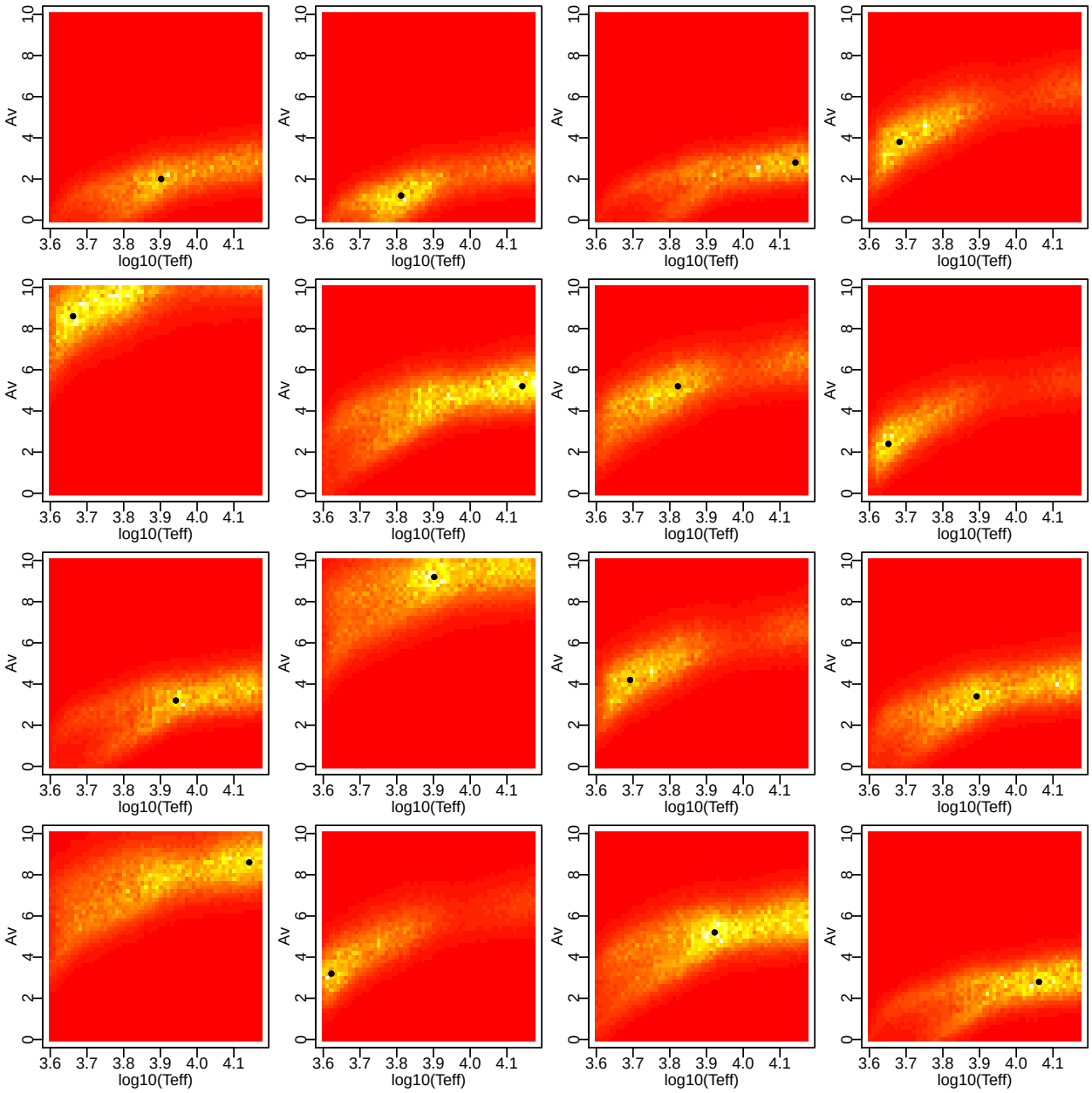

FIGURE 26: As Fig. 24 but now using the sensitivity-weighted distance.

4.2.2 Conversion of distances to probabilities

A key point of this discussion is that the size of the degenerate region depends on what we define as a "close" spectrum, and correspondingly what distance limit we adopt or colour scale we use for the map. The larger the distance the less likely we think there is a degeneracy. To quanity this consistently we must convert the distance into a probability. One of the simplest definitions for an appropriate probability density function is

$$P'(D^2) \propto e^{-D^2/2} \tag{6}$$

The choice of function is motivated by the property that P' decreases monotonically for increasing D^2, has the limit $P'(D^2 \to \infty) \to 0$, is finite at $D^2 = 0$, and has a finite integral over all values of D. (The factor of $1/2$ in the exponent just makes it a multidimensional Gaussian.) We can therefore normalize it

$$P'(D^2) = \frac{e^{-D^2/2}}{\int_0^\infty e^{-D^2/2}\, dD^2} \tag{7}$$

$$= \frac{e^{-D^2/2}}{\sum_{\mathrm{T_{eff}}} \sum_{\mathrm{A_v}} e^{-D^2/2}} \tag{8}$$

the second line giving the practical calculation for a discrete dgen grid. Note that

$$\log_e P'(D^2) = -\frac{D^2}{2} + A \tag{9}$$

for a constant A. Thus a plot of $-D^2/2$ just shows the log probability (the constant A and the factor $-1/2$ can just be absorbed by the colour scale adopted). (Note that this is not a Gaussian over the APs!) If we have Gaussian noise and use $D^2 = D^2_{cpw}$ then this is self-consistent.

Although we could think of other probability models, e.g. $P'(D^2) \propto (A + D^2)^{-1}$, the Gaussian has convenient properties. Of course, there is no reason why we have to use a Gaussian with the *same* mean and standard deviation as the χ^2 model. After all, the χ^2 model is also not theoretically correct for the sensitivity-weighted distance because R in equation 5 is not Gaussian and because the photometric noise is rarely truely Gaussian.

To address this we could introduce a length scale parameter, a, into equation 6 which sets the scale over which changes in distance are significant

$$P'(D^2) \propto e^{-D^2/2a^2} \tag{10}$$

(a could instead be considered as a factor inside R; equation 5.) The parameter a must be determined (the probabilities calibrated) by consideration of the distribution of the noise model and distance measure. In practice we could generate many noisy examples of a given star and construct the distribution of $P'(D^2)$ empirically using pairs of stars. To make $P'(D^2)$ Gaussian-like we then set a so that 68% of the pairs have $D^2 \leq a^2$. Once a is know we can use the dgen

grid to identify which stars with other APs fall within a high probability region of $P'(D^2)$ and from this construct the degeneracy map. a depends on the signal-to-noise level and therefore the G magnitude and in principle also depends on the values of $\mathrm{T_{eff}}$ and $\mathrm{A_V}$ (i.e. the target star) because of the dependence of D^2 on C_ϕ.

We may want to consider other distance metrics or possible modification of D^2_{sw}. It is not clear, for example, what to use for C_ϕ or whether to multiply this by a separate factor in equation 5. The expression for it given in equation 17 of CBJ-046 is an estimate of the uncertainties in the APs given the photometric uncertainties, but this is of course not the only source of uncertainty in the APs. The degeneracy itself is another. The goal is to provide the most discriminative measure which can minimize the size of the degeneracy region.

5 Conclusions and future work

I have developed a 2D+1D extension of the basic 1D+1D ILIUM model to allow it to simultaneously estimate two strong APs and one weak AP. I applied it here to the estimation of stellar $\mathrm{T_{eff}}$ and $\mathrm{A_V}$ (strong APs) and $\log g$ or [Fe/H] (weak APs), but it could equally well be applied to other multidimensional estimation problems.

I have shown that there is a strong and ubiquitous degeneracy between $\mathrm{T_{eff}}$ and $\mathrm{A_V}$. It is smooth and continuous over these APs (we are not talking of a few discrete degeneracy islands) with a positive correlation. Consequently, we need to re-think how we report estimates of these parameters, for example in the final Gaia catalogue. The forward model is a useful way of constructing a degeneracy map, with the classifier used to identify the corresponding degeneracy region. As the weak APs are presumably not involved in this degeneracy, the classifier can still be used to report unique values for these. (The $\log g$ residuals appear uncorrelated with the strong AP residuals, but this need to be analysed more closely.) We should therefore consider seriously the use of additional or prior information and how this can be used to assign probabilities to the solutions. This includes RVS, the distance and external catalogues. We could even use a Galactic model to assign a weak prior on $\mathrm{A_V}$, although we must balance this need against an undesired biasing of the the Gaia results based on our present understanding of the Galaxy.

The existence of this degeneracy also means that we should not interpret the above-reported $\mathrm{T_{eff}}$ and $\mathrm{A_V}$ uncertainties too literally. These uncertainties were calculated assuming that the known APs are "true", but if this truth cannot even be known in principle (based on the BP/RP data), then a single average error isn't a very useful summary of performance. This also has implications for what we report as the accuracy and precision of our classifiers and of the results in the Gaia catalogue.

The performance of ILIUM on the four stellar APs is good at G=15 and adequate for the strong APs at G=18.5 (if we ignore the degeneracy), but at G=18.5 performance on $\log g$ and [Fe/H]

is poor, with mean absolute errors of around 1.1 dex and 1.3 dex respectively. The significant degradation with respect to the 1D+1D model is on account of the wide A_V range and our inability to determine it exactly. That is, a large but perfectly known A_V would not reduce the accuracy in the $\log g$ and [Fe/H] estimates. Prior information on the $\log g$ and [Fe/H] would presumably help here too; its use would be less controversial as the parallax will aid the $\log g$ estimation and we could adopt a plausible and smooth prior on the metallicity distribution. Here I used [Fe/H] ranging from -4 to $+1$ with near equal probabilities, whereas we can be confident that the two extremes are much less common in the Gaia sample. While priors and some tuning of ILIUM may improve the $\log g$ and [Fe/H] performance, these results do suggest that we will not be able to make useful estimates of these APs at around G=18.5 or fainter for individual stars. And all of this of course assumes that the variable dispersion and LSF in BP/RP can be perfecltly corrected for when creating the mean spectrum.

The future priorities for ILIUM development are as follows

- Consider alternative choices for the distant metric for mapping the degeneracies and calibrating the probabilities (section 4.2.2).
- Come up with an appropriate method for encoding the degeneracy regions and implementation of a code for identifying these given an AP solution from ILIUM.
- Incorporation of additional information in the form of probability distributions on the APs, in particular on $T_{\rm eff}$ and A_V to help reduce the size of the degeneracy region. This additional information includes: prior information, perhaps based on the magnitude and latitude of the star; AP estimates from GSP-Spec; AP estimates from external catalogues (esp. infrared data). Provided all AP estimates are given as probability distributions we can simply combine them to give a posterior distribution.
- Incorporation of parallax information (together with the apparent magnitude, this gives an estimate of the intrinsic luminosity at zero extinction), to help improve the AP estimation.
- Replacement of the method of first differences for calculating the sensitivities with a faster and possibly more accurate method. (Note that the current method may nonetheless be reasonably accurate because the forward model is guaranteed to be smooth.) This could include changing the forward model, e.g. replacing the smoothing splines with B-splines, which have analytic derivatives (e.g. Eilers & Marx 1996), although this reintroduces the problem of having to define the knot locations, something which smoothing splines elegantly circumvent.
- Introduce a conversion measure to determine whether ILIUM has converged. If it has not after a fixed number of iteration, then either flag the solution or attempt to re-run ILIUM with different parameters.

- The next logical step is to replace the 1D weak forward model with a 2D one so that we can simultaneously estimate [Fe/H] and $\log g$ in a 2D+2D model. However, we cannot extend spline smoothing to arbitrarily high dimensions. A key concept of ILIUM is its partition of the APs (and thus their variance) into "strong" and "weak". This allows these two categories to be modelled separately and reliably: earlier attempts to forward model strong and weak APs together gave poor results. (An entirely different problem in which there were several important parameters covering a broad range of "strengths" may not be handled well by this particular forward modelling approach.)
- ILIUM is currently written in R without any optimization, so is quite slow (there are many unnecessary recalculations), plus R cannot handle large data sets and is not suited to a software development much larger than the current code (a few hundred lines). We will now write a DPCC-compliant Java version and include it as part of the CU8 software development and delivery.

References

Bailer-Jones C.A.L., 2009, ILIUM*: An iterative local interpolation method for parameter estimation*, GAIA-C8-TN-MPIA-CBJ-042

Bailer-Jones C.A.L., 2009, *Application of* ILIUM *to the estimation of the* $T_{\rm eff}$*–[Fe/H] pair from BP/RP*, GAIA-C8-TN-MPIA-CBJ-043

Bailer-Jones C.A.L., 2009, ILIUM III. *Further observations, tests and developments*, GAIA-C8-TN-MPIA-CBJ-046

Brown A.G.A., 2005, *Linear least squares*, unpublished notes

Eilers P.H.C., Marx B.D., 1996, *Flexible smoothing with B-splines and penalties*, `http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.47.4521`

Lindegren L., 2003, *Optimizing Gaia's photometric system - thoughts on distance measure and figure of merit Functions*, GAIA-LL-047

Lopez Marti, B., Figueras F., Jordi C., Carrasco J.M., Gebran M., 2009, *Sources with spectral features as seen by Gaia BP/RP*, GAIA-C5-TN-UB-BLM-001