# D3.5 Validated Bayesian Code for Stellar Spectral Fitting

## EXOHOST

Grant Agreement 101079231

Funded by the European Union

UK Research and Innovation

# Document Control Page

| **Document Control Page** | |
|---|---|
| Document Title: | D3.5 Validated Bayesian code for stellar spectral fitting |
| Project Title: | EXOHOST - Building Excellence in Spectral Characterisation of Exoplanet Hosts and Other Stars |
| Document Author: | Colin P. Folsom |
| Description: | The implementation and validation of a Bayesian or machine learning based approach for the Zeeman stellar spectral synthesis code is completed |
| Project Coordinator: | Assoc. Prof. Anna Aret |
| Contributors: | Veronika Mitrokhina, Anna Aret, Mihkel Kama, Luca Fossati, Merili Jauk |
| Date of Delivery: | 13.12.2024 |
| Type: | R – Document, report |
| Language: | EN-GB |
| Rights: | Copyright EXOHOST |
| Sensitivity: | ☒ Public<br>☐ Limited<br>☐ Classified |
| Status: | ☐ In Progress<br>☐ For Review<br>☐ For Approval<br>☒ Approved |

**Document history:**

The Document Author is authorised to make the following types of changes to the document without requiring that the document be re-approved:

- Editorial, formatting, and spelling
- Clarification

To request a change to this document, contact the Document Author or Project Coordinator. Changes to this document are summarised in the following table in chronological order (latest version last).

| **Revision History** | | | |
|---|---|---|---|
| Version | Date | Created by | Short Description of Changes |
| 0.1 | 22.11.2024 | Colin P. Folsom (UTARTU), Veronika Mitrokhina (UTARTU), Anna Aret (UTARTU), Mihkel Kama (UCL), Luca Fossati (OEAW) | First Draft |
| 0.2 | 05.12.2024 | Merili Jauk (UTARTU) | Amendments |
| 1.0 | 05.12.2024 | Merili Jauk (UTARTU) | Finalised for submission |

**Configuration Management: Document Location**

The latest version of this controlled document is stored in **EXOHOST Sharepoint Portal**.

**Document Approver(s) and Reviewer(s)**:

NOTE: All Approvers are required. Records of each approver must be maintained. All Reviewers in the list are considered required unless explicitly listed as Optional.

| **Approvers and reviewers** | | | |
|---|---|---|---|
| **Name** | **Role** | **Action** | **Date** |
| Anna Aret | Member of steering board | Reviewed, approved | 05.12.2024 |
| Luca Fossati | Member of steering board | Reviewed, approved | 05.12.2024 |
| Mihkel Kama | Member of steering board | Reviewed, approved | 12.12.2024 |
| Eric Stempels | Member of steering board | Reviewed, approved | 10.12.2024 |

# Table of Contents

# 1 Introduction

The spectra of stars contain a wealth of information. They contain information about the temperature, density, velocity fields, and chemistry of the star. This can be used to infer information such as the mass, radius, evolutionary state, and space motion of a star. Thus, the detained analysis of stellar spectra is critical for understanding stellar physics. It is also crucial for other areas of astrophysics that rely on understanding stars, such as studies of exoplanets or circumstellar disks.

Within the EXOHOST project, we fine-tuned, tested, and validated an initial testbed software solution for stellar spectroscopic analysis, described below. This software is based on coupling a Python library for Markov Chain Monte Carlo methods with the ZEEMAN stellar spectrum synthesis code.

# 2 Rationale

Several methods are commonly used for deriving stellar parameters from observed stellar spectra, all of which rely on models. The most flexible approaches involve directly fitting model spectra to observations. This can be applied to essentially all classes of star, provided appropriate models, and importantly this can account for cases where lines are heavily blended. However, this approach typically requires determining several interdependent parameters simultaneously. This fitting is usually done using a non-linear least squares approach, in which a $\chi^2$ statistic is minimized iteratively. If only a few parameters are of interest, then a grid of pre-computed models can be used, with interpolation on the grid. If more parameters are of interest, for example a set of chemical abundances, then it is usually necessary to compute new model spectra for each fitting iteration.

Approaches for deriving stellar parameters generally have difficulties deriving realistic and reliable uncertainties. The uncertainties are often dominated by systematic uncertainties, either in the model or unexpected distortions in the observation. Furthermore, there are often strong correlations in the uncertainties of some parameters, which most uncertainty estimates fail to capture.

Formal uncertainties can be derived from $\chi^2$ minimization routines, such as from the diagonal of a covariance matrix. However, they are usually only point estimates of the uncertainty. The uncertainties generally don't account for covariances between parameters (i.e. it is only from the matrix diagonal), and they only reflect the random uncertainties provided by the observation. $\chi^2$ minimization also performs poorly for parameters where only upper limits can be derived. A grid search approach using $\chi^2$ can capture covariances between parameters, and upper limits, but with a runtime increasing exponentially with the number of parameters. Alternatively, fitting individual lines, or portions of spectrum, separately and then taking the dispersion of the resulting best fit values can be useful. This can account for some of the more important systematic uncertainties, at least partially. It does not, however, account for covariance between parameters. Additionally, as line blending becomes more severe, or as more parameters need to be determined simultaneously, larger spectral regions need to be used. Fitting larger regions at once means that an observation can be divided into fewer independent regions. If the number of independent regions, or independent lines, becomes small then using the dispersion of those independent values as an uncertainty becomes unreliable, due to small number statistics.

# 3 Software

In order to derive more complete uncertainty estimates we have implemented a Bayesian approach to estimating stellar parameters from observed spectra, using a Markov chain Monte Carlo (MCMC) algorithm. This approach has several advantages. While $\chi^2$ minimization essentially provides a point estimate of the optimal parameters, an MCMC approach provides probability distributions for those parameters. The probability distributions allow for covariance between parameters, asymmetric uncertainties from skewed probability distributions, and upper limits on parameters. Bayesian statistics allow us to incorporate prior information about parameters into the final posterior probability distribution. This provides a mechanism for naturally incorporating physical limits on parameters (such as rotation or turbulent velocities being non-negative), and potentially for including information about a star from other sources (such as a photometric temperature estimate).

To generate synthetic spectra for this analysis we used the ZEEMAN spectrum synthesis code (Landstreet 1988, Wade et al. 2001, Folsom et al 2012). This code generates synthetic spectra under the assumption of local thermodynamic equilibrium, using a model atmosphere and list of atomic line data as inputs. ZEEMAN has successfully been applied to a wide range of stars (e.g. Folsom et al. 2012, 2016). It is a computationally efficient Fortran code, with optional parallelization implemented, which makes it suitable for generating the large volume of synthetic spectra needed for an MCMC analysis. The model atmospheres used here are from a standard grid of ATLAS9 models (Kurucz 1979; Castelli & Kurucz 2003). A grid of model atmospheres in effective temperature ($T_{\rm eff}$) and surface gravity ($\log g$) is used, and models are interpolated between (as in Folsom et al. 2012). A grid of atmospheres was used, rather than computing a new model for each spectrum, to save computation time, since computing model atmospheres is relatively slow. Lists of atomic line data were taken from the Vienna Atomic Line Database (Ryabchokova et al. 2015).

To implement the MCMC parameter estimation routine we used the EMCEE package (Foreman-Mackey et al. 2013). This package provides a Python implementation of the affine invariant sampler of Goodman and Weare (2010). This is an efficient and flexible MCMC sampling routine with a convenient interface. EMCEE requires a user supplied function for the logarithm of the unnormalized posterior probability for a given set of parameters. This incorporates the prior probability (information already known about the model parameters) and the likelihood (a probability that an observation is consistent with a given model). The likelihood function we created takes the form of a Python code, which takes a set of model parameters and runs the ZEEMAN spectrum synthesis code. It then takes the resulting synthetic spectrum, interpolates it onto the observation, and calculates the likelihood assuming uncorrelated Gaussian noise in the observation. While the code currently uses ZEEMAN spectrum synthesis, it is designed to be modular, and other spectrum synthesis codes could be used by replacing one function.

The MCMC parameter estimation routine includes the option of modelling the continuum flux level with a polynomial of arbitrary degree. This is particularly useful for cases where the continuum normalization of the observation is uncertain or unreliable. Uncertainties in the normalization are usually not considered at all in spectral fitting using $\chi^2$ minimization. The Bayesian approach used here allows us to marginalize over nuisance parameters, such as the continuum polynomial coefficients, to derive distributions of the parameters of interest, such as chemical abundances or effective temperature.

Our code has a main function in Python for running the MCMC routine and generating the Markov chain. This allows one to provide an observation file, a list of wavelength regions to fit, a set of parameters whose values are to be inferred along with initial estimates (as a Python dictionary), any parameters with assumed fixed values, and the length of the chain to calculate. The resulting chain is saved to a file, and sample

model spectra can also optionally be generated and saved. There is a second interface function that allows one to add more samples to a chain that has been previously calculated. In this function the parameters used, assumed fixed values, and wavelength ranges, are all read in using the saved file from the previous chain.

The code has been tested on a number of cases and performs well. The first tests were purely theoretical, carried out as a synthetic retrieval self-consistency check. Synthetic spectra were calculated with ZEEMAN, and then converted to synthetic observations by convolving with a typical instrumental broadening profile, resampling onto a typical instrumental pixel size, and adding synthetic Gaussian noise. We used the MCMC parameter inference code to attempt to recover the initial input parameters of the spectrum. This provides a test that the MCMC portion of the code is working properly. In these tests the input parameters were consistently recovered within the uncertainties provided by the posterior distributions.

# 4 Validation of the MCMC software

Within the EXOHOST project, a more complete set of tests was performed using real observations of stars that have been previously studied by Fossati et al. (2011), who performed a standard abundance analysis. We analysed three stars from the open cluster NGC 5460: HD 123226, HD 123269, and CPD-47 6385. The observations were obtained by Fossati et al. (2011) using the FLAMES instrument at the Very Large Telescope (VLT). The observations used the GIRAFFE spectrograph in a low-resolution mode ($R$ = 7500) in the ~4500-5076 Å wavelength range, and GIRAFFE in a medium resolution mode ($R$ = 24000) in the 5140-5355 and 5592-5837 Å ranges.

We re-analysed these observations using the MCMC tool, to derive $T_{\rm eff}$, log$g$, radial velocity ($v_{\rm r}$), projected rotational velocity ($v\sin i$), microturbulence velocity ($v_{\rm mic}$), chemical abundances for up to 12 different elements, and continuum polynomial coefficients. For a comparison with this, we also re-analysed the observations using a standard $\chi^2$ minimization method with ZEEMAN as described by Folsom et al. (2012).

An example of a resulting fit to an observation of HD 123226 is presented in Fig. 1. The wavelength range used here was 4506-5070 Å, although only a portion of this is shown in the figure. The optimal MCMC model was calculated using the median parameter values of the posterior distribution. A set of models sampling the MCMC chain are also plotted, and regions with a larger dispersion in these models correspond to regions depending on more uncertain parameters. Generally, the MCMC result matches the best $\chi^2$ minimization model, and the observation. The MCMC result has the advantage that it provides a range of models that can fit the observation, and the best $\chi^2$ minimization model falls within that range.

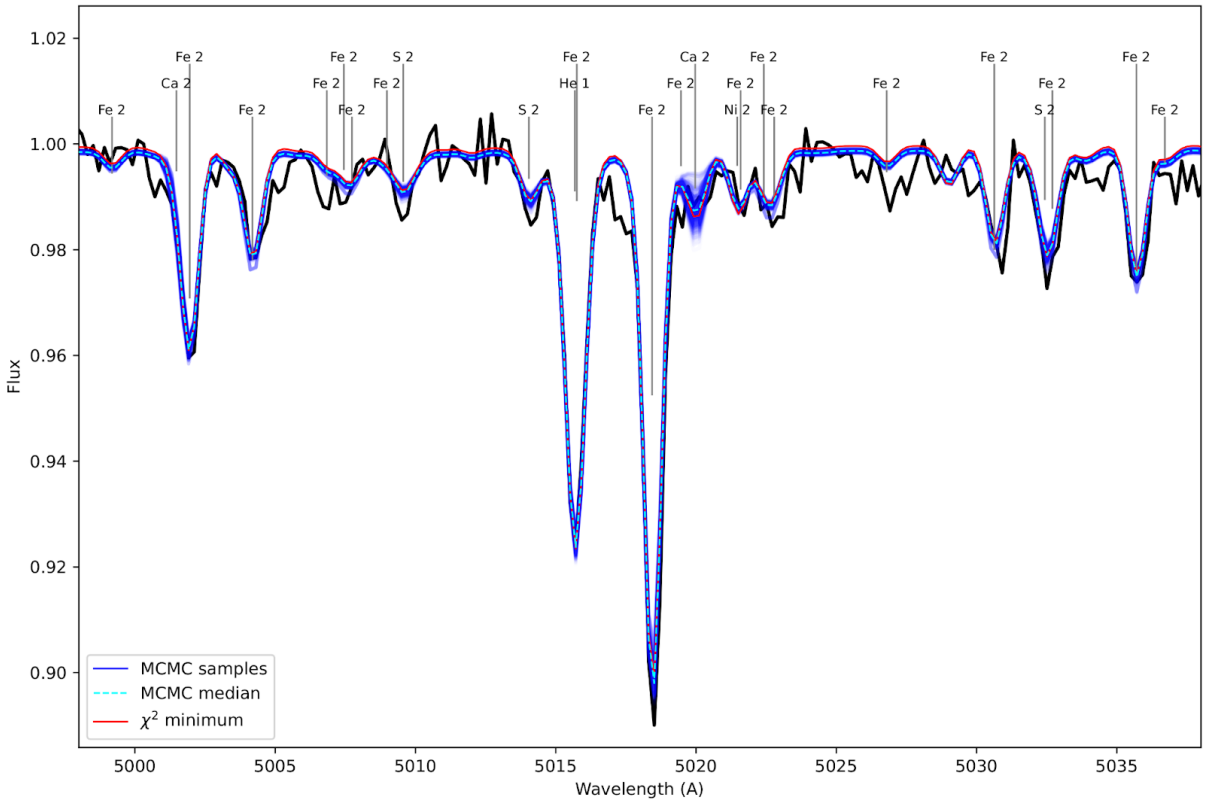


**Figure 1. Observation of HD 123226 (black) compared with model spectra. 1000 sample models from the MCMC analysis are shown, along with the best MCMC model using median values, and the best model from a standard $\chi^2$ minimization analysis. Positions of stronger lines from known species are indicated.**

The results from the MCMC analysis can be plotted as a function of step in the Markov chain, which can be useful for checking if the chain has converged to a stable range of values. An example for HD 123226, fitting the 4506-5070 Å portion of the spectrum, is shown in Fig. 2. The Goodman and Weare (2010) MCMC algorithm used by EMCEE runs multiple parallel chains, as different 'walkers', and these chains are plotted on top of each other. In this case, the chain has converged by around step 800, and the final analysis of samples in the chain only uses samples after step 1000.

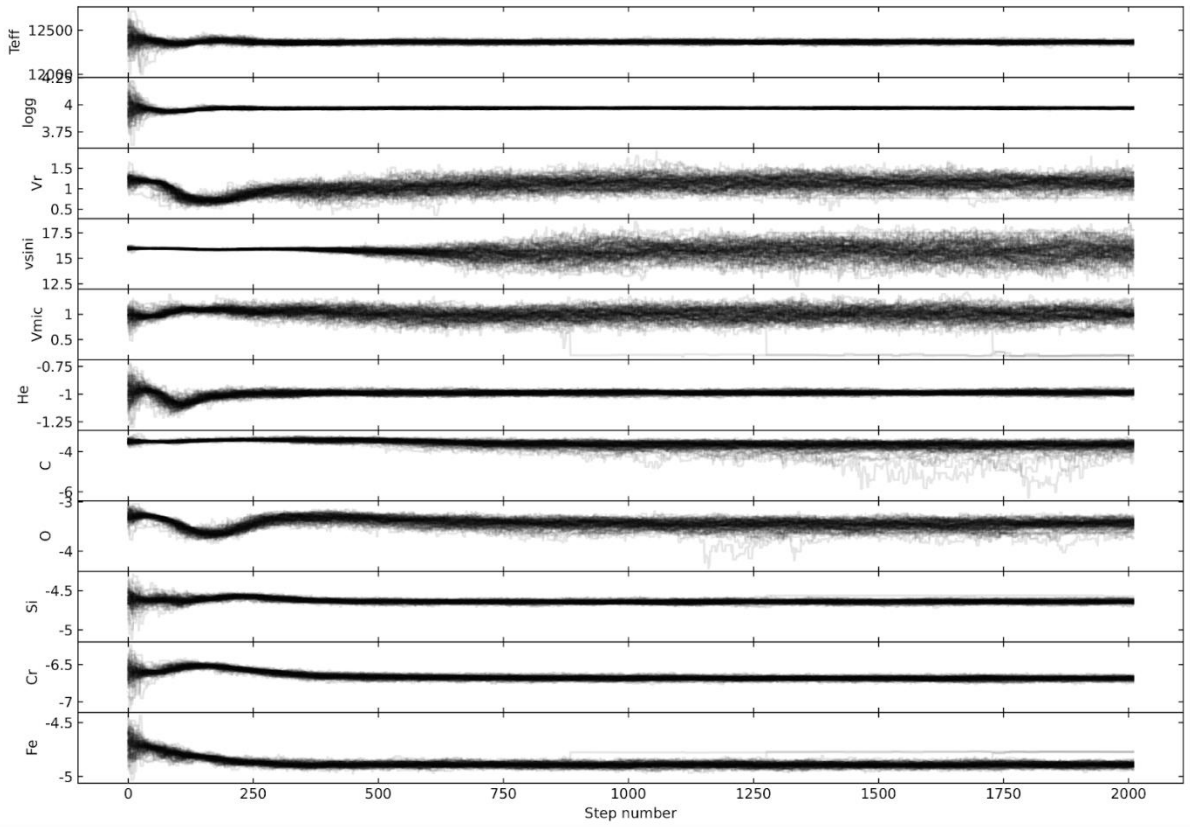


**Figure 2. Parameter values as a function of step number in the MCMC chain, for HD 123226 in the 4506-5070 Å window. Velocities are in km $s^{-1}$, abundances for elements are log number densities relative to H.**

The MCMC results can also be plotted as histograms of the marginalized posterior probability distribution. These histograms can be plotted in one dimension for one parameter, and in two dimensions illustrating correlations between two parameters. A plot of these histograms for the combinations of parameters in the MCMC chain, a 'corner plot', is shown in Fig. 3. While many parameters have little correlation, and approximately follow Gaussian distributions, some important correlations can be seen. There is a strong correlation between $T_{\rm eff}$ and log$g$, which is due to the dependence of these parameters on the wings of the Hβ Balmer line, and those wings depend on $T_{\rm eff}$ and log$g$ in similar fashion. There is an important correlation between $v_{\rm mic}$ and the iron abundance, which is due to the constraint on $v_{\rm mic}$ coming from de-saturation of strong lines, and the strong lines in this wavelength range are dominantly iron. The carbon abundance does not have a strong correlation with other parameters, but it does have an asymmetric distribution with a longer tail to lower abundances. This is because the carbon abundance is based on two very weak lines, and is thus intrinsically rather uncertain, with the lower limit on the abundance being poorly constrained.

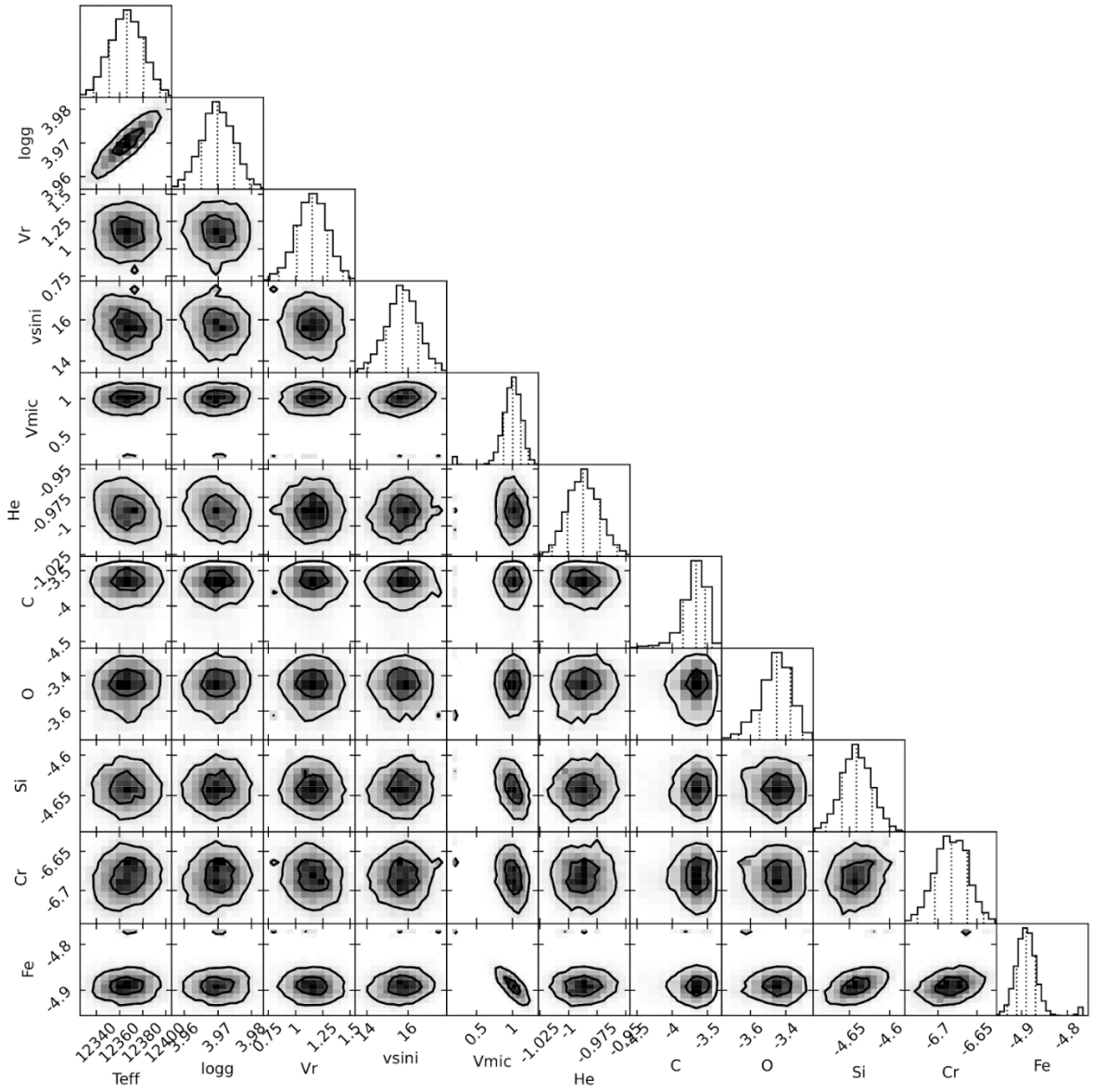


**Figure 3. Posterior probability distributions for parameters from the MCMC analysis, as one- or two-dimensional histograms, for HD 123226 in the 4506-5070 Å window. Velocities are in km $s^{-1}$, abundances for elements are log number densities relative to H. Contours or dotted lines correspond to 1σ and 2σ confidence intervals. Abundances for Mg, Al, S, Ca, Ti, and continuum polynomial coefficients are omitted for clarity.**

The MCMC and $\chi^2$ minimization results are in good agreement, mostly within 1σ of the uncertainties. The differences are largely driven by fitting the continuum with MCMC, and if the continuum is assumed fixed at 1.0 then the results agree to better than 1σ. For well constrained parameters, the uncertainties from the MCMC analysis are similar to the formal uncertainties from the $\chi^2$ minimization. However, for parameters with a strong covariance, or abundances from weak lines where the continuum placement contributes strongly to the uncertainty, the MCMC uncertainties are larger and more realistic. Compared with the results of Fossati et al. (2011), the MCMC results are usually consistent within 1 or 2σ of the joint uncertainties. Although the differences are larger than with the $\chi^2$ results, due to differences in

methodology and likely differences in the atomic line data used. For example, the He, O, and Fe abundances agree within about 1σ, while the S abundances differ by 2 or 2.5σ, likely due to differences in the line data used.

# 5 Conclusion

In conclusion, the MCMC stellar parameter estimation code is working effectively. It produces results that are consistent with previous methods, but provides additional flexibility in model parameters. It also provides additional information about parameters that are weakly constrained with asymmetric uncertainties, upper limits, and correlated uncertainties between parameters. This MCMC approach does not solve the problematic case of systematic errors dominating the uncertainty. But it does allow for systematic errors in the continuum level to be incorporated into the modelling, leading to an improved characterization of those uncertainties. Future developments may allow for more sources of systematic error to be incorporated into the model. The MCMC analysis comes at a substantial additional computational cost. In a complex case with many abundances to determine, the $\chi^2$ minimization approach may require computing on the order of 100 model spectra, while the MCMC approach may require 10,000 or 100,000 spectra. While this is tractable with modern computers it is not yet trivial. Thus, a practical approach in the short term could involve an initial 'quick look' analysis with standard $\chi^2$ minimization tools, followed by a final analysis with MCMC.

# 6 References

Castelli, F. & Kurucz, R. L. 2003, Modelling of Stellar Atmospheres, Proceedings of the 210th Symposium of the International Astronomical Union, 210, A20.

Folsom, C. P., Bagnulo, S., Wade, G. A., et al. 2012, Monthly Notices of the Royal Astronomical Society, 422, 2072.

Folsom, C. P., Petit, P., Bouvier, J., et al. 2016, Monthly Notices of the Royal Astronomical Society, 457, 580.

Foreman-Mackey, D., Hogg, D. W., Lang, D., et al. 2013, Publications of the Astronomical Society of the Pacific, 125, 306.

Fossati, L., Ryabchikova, T., Shulyak, D. V., et al. 2011, Monthly Notices of the Royal Astronomical Society, 417, 495.

Goodman, J. & Weare, J. 2010, Communications in Applied Mathematics and Computational Science, 5, 65.

Kurucz, R. L. 1979, Astrophysical Journal, Suppl. Ser., 40, 1.

Landstreet, J. D. 1988, Astrophysical Journal, 326, 967.

Ryabchikova, T., Piskunov, N., Kurucz, R. L., et al. 2015, Physica Scripta, 90, 054005.

Wade, G. A., Bagnulo, S., Kochukhov, O., et al. 2001, Astronomy and Astrophysics, 374, 265.