Inferring stellar properties using colours, parallaxes and an HRD prior

C.A.L. Bailer-Jones

To appear in proceedings of the JENAM 2010 symposium Star Clusters in the Era of Large Surveys *A. Moitinho and J. Alves (eds.), held in Lisbon in September 2010*

Abstract Stellar parameters – effective temperature, metallicity, interstellar extinction etc. – are typically estimated from a spectrum or multiband photometry. I outline a probabilistic method for estimating stellar parameters which uses not only the spectral energy distribution but also the apparent magnitude, parallax (if available) and the strong constraints provided by the Hertzsprung–Russell Diagram. This (a) improves the accuracy and precision over use of just the spectrum, and (b) ensures that the inferred parameters are both physically realistic and are consistent with the distance, apparent magnitude and stellar physics. The method provides full covariate probability distributions over the parameters, i.e. it provides not just parameter estimates but also confidence intervals and the correlations between the estimates. The latter is particularly important given the degeneracies between some parameters, such as temperature and extinction. These degeneracies are shown to be reduced by use of this method. Here I provide a short summary of the method and and show some results of its application to 85 000 Hipparcos–2MASS stars and to the Hyades clusters. A full description and further results can be found in [1].

1 Introduction

If we are lucky enough to have high resolution spectra of stars, then we can normally measure their parameters with some precision. But obtaining such detailed information on a large number of stars (10^7 or more) is currently out of the question, and we have to resort to low resolution spectroscopy or multiband photometry. This is the case with surveys such as SDSS, Pan-STARRS and LSST (five band photometry), and Gaia (very low resolution spectrophotometry). The parameter accuracy we can achieve with such data alone is limited.

C.A.L. Bailer-Jones
Max Planck Institute for Astronomy, Königstuhl 17, 69117 Heidelberg, Germany
e-mail: calj@mpia.de

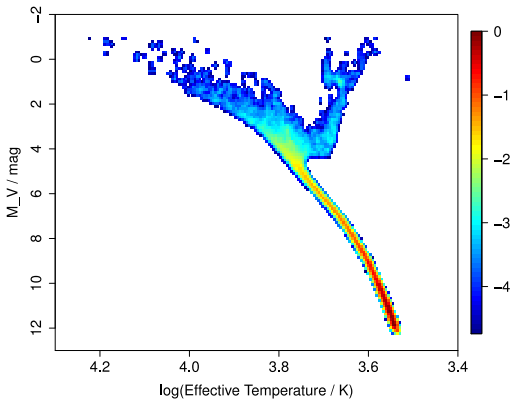

Fig. 1 HRD prior. The colour scale shows $\log P(M_\mathrm{V}, T)$ normalized to have zero at its maximum. Unoccupied areas are shown in white.

What other information is available to help improve performance? The Hertzsprung–Russell Diagram (HRD) describes the location of stars in the (M_V, T) (absolute magnitude, effective temperature) plane, and for virtually any stellar population it is very sparsely and non-uniformly populated (see Fig. 1). That is, a priori we can place strong and plausible constraints on the relative probability of different combinations of the stellar parameters.

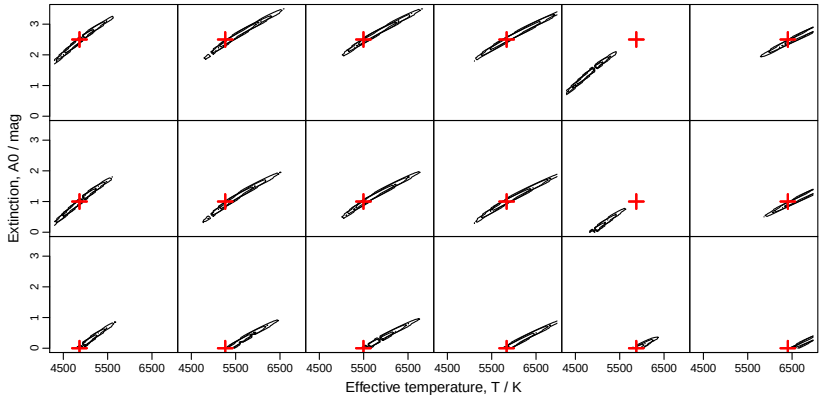

Fig. 2 Posterior probability density function (PDF) from the p-model (colours only) over 18 stars with 6 different true temperatures (columns) and three different true extinctions (rows). Three contours are shown for each star, enclosing 90%, 99% and 99.9% of the total posterior probability. For comparison, the true parameter values are shown with the red cross.

Stellar parametrization in large, deep surveys faces another problem, namely interstellar extinction (A_V). In principle this can also be estimated from the photometry, but it is frequently degenerate with T_{eff} (see Fig. 2, explained further in the next section). This problem is sometimes ignored in survey projects by assuming that the stars have negligible extinction (e.g. at high Galactic latitudes), or by using an extinction map. The first solution is inadmissible for surveys near the Galactic plane or near molecular clouds, and extinction maps often have low spatial resolution or are not three-dimensional (they may only give the integrated extinction to the edge of the modelled Galaxy).

Extinction is a major issue for the all-sky Gaia survey. Yet herein also lies an opportunity. Gaia will measure positions, parallaxes (ϖ) and proper motions with an accuracy of up to 10 microarcseconds for almost all 10^9 stars in our Galaxy brighter than G=20. It will also obtain low resolution optical spectrophotometry and apparent magnitudes in the G band (a "white-light" band much broader than the V band). *If* we knew A_V then we could estimate the absolute stellar magnitude (M_V), a fundamental stellar property, via the relation

$$V + 5\log\varpi = M_V + A_V - 5\,. \tag{1}$$

However, we also have to estimate A_V from the data. How can we do this?

2 Method

The solution is to approach the problem probabilistically. Where we have noise we have uncertainties; these are best represented by probability density functions (PDFs). The Bayesian approach allows one to include all available information as PDFs in a self-consistent manner, and to propagate these PDFs through the calculation to provide not only parameter estimates but confidence inte¡rvals on these estimates.

Let us consider the problem of estimating just the two parameters $\phi = (A_V, T)$. We have three pieces of information:

1. the spectrum (p), which constrains T and A_V;
2. the quantity $q = V + 5\log\varpi$, which constrains $M_V + A_V$ from equation 1;
3. the HRD, which constrains M_V and T (Fig. 1).

The goal is to determine $P(\phi|p,q)$. The spectrum we can predict given ϕ using a *forward model*, which is the result of a fit to a set of labelled data [2]. This is the "training" phase in machine learning speak. Combined with a suitable photometric noise model, the forward model provides $P(p|\phi)$. Adopting a noise model for the apparent magnitude and parallax measurement allows us to write item (2) as $P(q|\phi, M_V)$. Applying Bayes' theorem we can then arrive at an expression for $P(\phi|p,q)$ in terms of these quantities. It involves marginalizing over the unknown M_V to give a (non-parametric) two-dimensional PDF over ϕ for given measurements p and q.

3 Demonstration and application to 85 000 Hipparcos–2MASS stars

This method has been tested by using it to estimate A_V and T for 5280 FGK stars with known "true" parameters using $BVJHK$ photometry and Hipparcos parallaxes. These data are derived from a set of 880 stars with T estimated from high resolution spectroscopy by [4], to which I have applied artificial reddening to provide variance in A_V. First I use just the four colours to determine $P(p|\phi)$ for each star and take the mean of this distribution as the parameter estimates (the *p-model*). The parameter accuracy (mean of absolute residuals) is 5.5% in T and 0.3 dex in A_V. The probabilities of different solutions for these parameters for 18 example stars were shown in Fig. 2. Note the significant degeneracy between the parameters. When introducing the parallax, apparent magnitude and an HRD prior (to give the *pq-model*), these errors are reduced to 3.5% and 0.2 dex respectively, an increase in accuracy of around 40%. (We can also apply the method using just the colours and HRD prior but no measurement of q. Even this improves accuracy by 13% over the p-model.) Posterior probability distributions from this model are plotted in Fig. 3. Note how much smaller the confidence ellipses are, which reflects the increase in the precision of the parameter estimates.

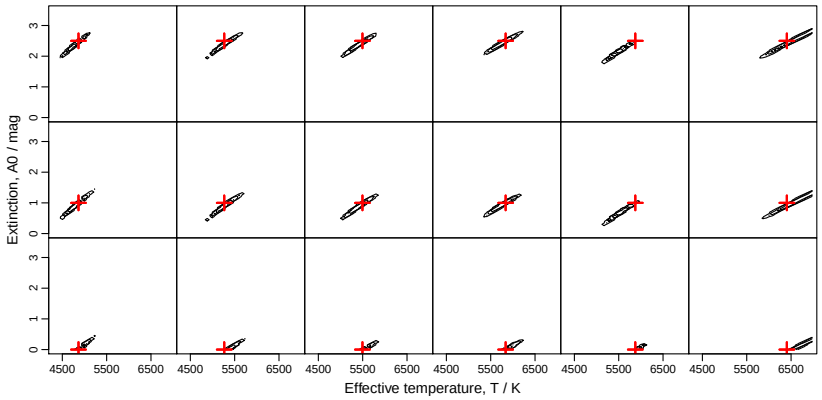

Fig. 3 Posterior probability distribution for the pq-model (i.e. including the HRD prior, parallax and apparent magnitude) for the same stars as shown in Fig. 2.

I then applied the method to a set of 85 000 Hipparcos stars for which I obtained a reliable astrometric cross match with 2MASS (to give $BVJHK$ photometry), but for which the "true" parameters are unknown. Many of these stars (42%, it turns out) are not FGK stars, so their parameters cannot be estimated reliably by this method. (I identify these stars by their inferred PDF peaking at or very close to the edge of the parameter space.) Once we have estimated A_V and T we can estimate M_V (or rather

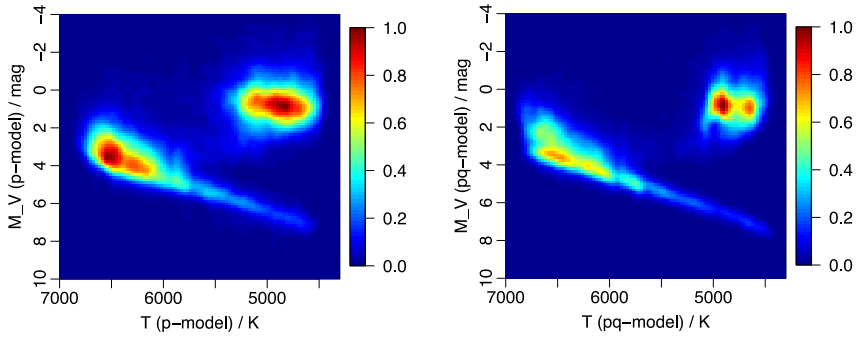

Fig. 4 HRD for the Hipparcos–2MASS stars derived from the p-model (left) and pq-model (right) shown as a density plot (achieved via smoothing with a Gaussian kernel). The number of stars per unit area is normalized to a value of 1.0 at the maximum density (separate normalization in each plot).

a PDF over it) from equation 1 and so plot the stars in an HRD: see Fig. 4. These are discussed in more detail in [1]. As the Hipparcos sample covers the whole sky, we can also combine the individual extinction measurements to produce an extinction map; a 2D map is shown in Fig. 5. The median distance to these stars is 170 pc (90% have distances between 40 and 730 pc).

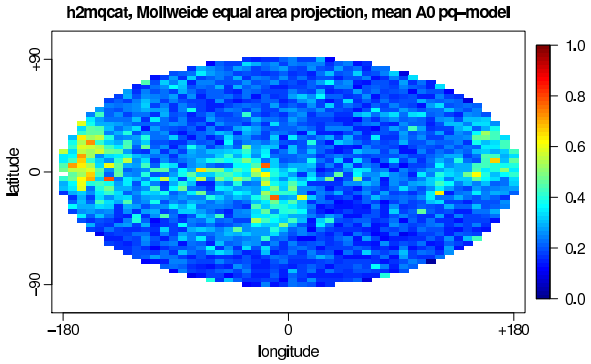

Fig. 5 The mean extinction (A_V) from the pq-model along the line of sight to the Hipparcos stars, plotted in Galactic coordinates.

As an additional test, I identified 137 stars in my sample in the list of 218 Hipparcos Hyades members from [3]. The HRD diagram (pq-model) for these objects is plotted in Fig. 6. As expected, the majority of these have very low extinctions,

yet a significant number of the cooler stars have relatively large extinctions. Further investigation of this is beyond the space available in this paper.

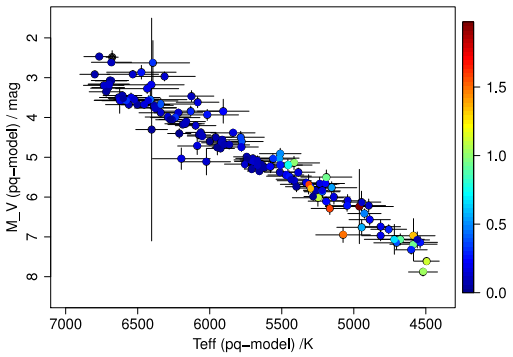

Fig. 6 HRD for 137 Hyades stars with parameters determined using colours, parallaxes and apparent magnitudes. Individual stars are coloured according to their estimated extinction, A_V.

A catalogue of parameter estimates (plus uncertainties) from both the p-model and pq-model for 46 900 stars is available online[1] or from CDS, Strasbourg. More results and discussion of the method can be found in [1].

4 Why you shouldn't use conventional methods

I finish this brief article with some arguments against using conventional machine learning methods (e.g. neural networks, support vector machines) for estimating stellar parameters. By "conventional" I mean multidimensional, nonlinear regression algorithms which attempt to model the parameters as a function of the input data, i.e. fit a function $f(\phi|p)$. These methods *can* give overall good performance in some applications – and I have published work using them – but here are some drawbacks

- $f(\phi|p)$ is an inverse function and so may not be unique. Especially at low spectral resolution or low SNR, a single p may correspond to a broad range of ϕ, or even isolated islands of parameter space.
- this function is likely to be cumbersome and difficult to fit when p is heterogeneous, i.e. includes not only colours/fluxes but also a parallax or something based on it.
- support vector machines, neural networks and related methods are fundamentally non-probabilistic for continuous parameter estimation, so they cannot recog-

[1] http://www.mpia.de/homes/calj/qmethod.html

nise degeneracies, multiple solutions or naturally deliver meaningful error bars. (Techniques exist for forcing probabilities out of these methods, but these are convenient fixes rather rigourous solutions.)
- these methods cannot naturally or explicitly include domain knowledge or prior information. This could lead to inconsistent or non-physical solutions, plus misses an opportunity to include additional information which could improve performance. Probabilistic methods are much more flexible, for example allowing a simple combination of independent solutions based on different pieces of information.
- these methods are not robust to missing information: setting an input to zero is not the same thing! With a probabilistic method, on the other hand, you can usually marginalize over missing input data.

The only real advantage of conventional method is that they are generally much faster, i.e. are cheaper.

Acknowledgements I would like to thank Jo Bovy, Ron Drimmel, David Hogg, Dustin Lang and Antonella Vallenari for useful discussions of this work, and Antonella Vallenari for providing me with output from her stellar population models. I am grateful to Chris Stubbs and the LPPC group at Harvard for support during a sabbatical stay. This work makes use of Hipparcos and 2MASS data and has used Simbad and Vizier at CDS, Strasbourg and IRSA at NASA.

References

1. C. A. L. Bailer-Jones. Bayesian inference of stellar parameters and interstellar extinction using parallaxes and multiband photometry. *MNRAS*, in press (arXiv:1009.2766), 2010.
2. C. A. L. Bailer-Jones. The ILIUM forward modelling algorithm for multivariate parameter estimation and its application to derive stellar parameters from Gaia spectrophotometry. *MNRAS*, 403:96, 2010.
3. M. A. C. Perryman, A. G. A. Brown, Y. Lebreton, A. Gomez, C. Turon, G. Cayrel de Strobel, J. C. Mermilliod, N. Robichon, J. Kovalevsky, and F. Crifo. The Hyades: distance, structure, dynamics, and age. *A&A*, 331:81–120, March 1998.
4. J. A. Valenti and D. A. Fischer. Spectroscopic properties of cool stars (SPOCS). I. 1040 F, G, and K dwarfs from Keck, Lick, and AAT planet search programs. *ApJS*, 159:141, 2005.