GAIA Spectroscopy, Science and Technology, U.Munari ed.
ASP Conference Series, Vol. 298, pp. 199–208, 2003

On the classification and parametrization of GAIA data using pattern recognition methods

C.A.L. Bailer-Jones
Max-Planck-Institut für Astronomie, Königstuhl 17, 69117 Heidelberg, Germany

Abstract. I discuss various aspects of source classification and physical parametrization using data from the future Galactic survey mission GAIA. Due to the heterogeneity of the data, the large variety of objects observed and problems of data degeneracy (amongst other things), efficiently extracting physical information from these data will be challenging. I discuss the global and local nature of commonly used pattern recognition algorithms and outline two alternative frameworks for classification – parallel and hierarchical – and describe some aspects of each. A method for calibrating the classification algorithms is proposed which requires only a limited amount of additional (ground-based) data. By way of illustration, an example of stellar parametrization using GAIA-like RVS data is presented.

1. Introduction

The primary scientific goal of GAIA is a detailed study of the composition, structure and formation of our Galaxy. The major contribution which GAIA will make in this area is high precision astrometry of around one billion stars, providing accurate positions, parallaxes and proper motions. Of course, to be able to use this astrometric data for Galactic structure studies, it is essential that the intrinsic properties of the stars so observed are known. For this reason, GAIA will also employ multiband photometry and high resolution spectroscopy (see section 2.).

The classification[1] requirements for GAIA have been outlined in Bailer-Jones (2002), but include: (1) discrete classification of GAIA sources as star, galaxy, quasar, solar system object etc.; (2) determination of stellar astrophysical parameters (APs) ($T_{\rm eff}$, [Fe/H], [α/Fe], $\log g$, $V_{\rm rot}$, mass, age, activity, etc.); (3) accurate determination of interstellar extinction (which is unique for a given star so can effectively be assumed to be a stellar parameter); (4) detection and

[1] I draw a distinction between *classification* and *parametrization*. Classification refers to the allocation of *discrete* classes, such as (1) star or nonstar, (2) star, galaxy, quasar, asteroid or other, or even (3) hot star or cool star. Parametrization, on the other hand, refers to the placing of sources on a *continuous* scale, such as $T_{\rm eff}$, [Fe/H], star formation rate, albedo etc. The distinction is important in terms of the way in which algorithms are used. However, where the distinction is not important, I will use the term *classification* to refer collectively to the process of assigning attributes (classes or parameters) to sources.

description of stellar multiplicity; (5) identification of new types of objects. The goal for GAIA is not "just" to produce a catalogue of astrometric parameters and associated photometry, but also detailed information on source classification and APs.

2. GAIA data

GAIA is an all sky magnitude-limited survey (to V$\sim$20). Due to telemetry limitations from its orbit around the Earth–Sun L2 point, not all data will be transmitted to ground. Instead there will be real time on board detection of all sources above a magnitude limit (V$\sim$20) and only the CCD pixels in patches around each object will be transmitted.

The primary information for classification purposes will come from the Medium Band Photometer (MBP) a set of 10–20 (the system is not yet fixed) medium band filters over the wavelength region 200–1100 nm. This must obtain information on every object down to the GAIA magnitude limit. This will be supplemented by about five broad bands from the astrometric instrument (which are primarily intended to give a chromatic correction to PSF centroiding).

Data relevant to stellar parametrization will also be provided by the Radial Velocity Spectrograph (RVS), the capabilities and optimization of which are the subject of this meeting. The RVS will obtain slitless spectra over the whole sky over the wavelength range 849–874 nm with a resolution (to be decided) of between 5 000 and 10 000 (see contribution by D. Katz in this volume). Due primarily to signal-to-noise considerations, RVS data will only be obtained down to V$\sim$17. For the brighter stars, it will also provide an information on stellar activity (via emission lines), rotational velocities (via line broadening), individual element abundances and permit an independent determination of $T_{\rm eff}$, [Fe/H], [α/Fe] and $\log g$. The RVS instrument has significant implications of parametrization of GAIA data, as it provides different amounts of information (and perhaps even different formats if lossy compression schemes are used) for each object, depending on magnitude (and crowding).

Each point on the sky is observed about 100 times over the course of five years, providing variability information relevant for identifying some types of stars and quasars. Astrometric information will also be very useful for classification, e.g. parallaxes for determining stellar luminosity and radius, (zero) proper motions for identifying quasars. However, we should not use kinematic information to parametrize stars, as this would require a Galaxy model and hence introduce classification biases based on our *current* and limited understanding of Galactic structure.

In summary, classification with GAIA is characterised by the *heterogeneity* of the data: photometry (from two separate instruments), spectroscopy (only for some targets, varying formats) and astrometry. Making full interdependent use of these data is a challenge, and, as discussed by Bailer-Jones (2002), is not something which classification methods used to date in astronomy have had to deal with on this scale. This is further complicated by the wide range objects and astrophysical parameter scales which will be encountered.

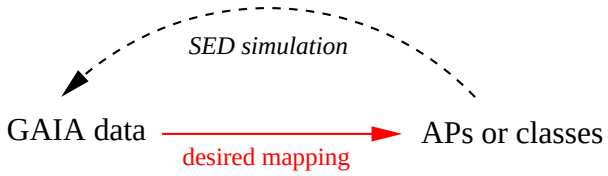

Figure 1. Classification or parametrization is the process of determining the mapping from a data domain to a class or astrophysical parameter (AP) domain. The opposite mapping is equivalent to the simulation of the data, e.g. the emergent stellar spectral energy distribution (SED).

3. Classification principles

Classification is the problem of assigning object classes or APs and generally involves determining some kind of mapping from the data space to the parameter space (Fig. 1).[2] A frequently used approach is the *supervised* or pattern matching approach, in which pre-classified data (*templates*) are used to infer the desired mapping. This mapping is then applied to new data to establish their classes or APs. Perhaps the most familiar such technique is the minimum distance method (MDM), shown schematically in Fig. 2. This is a *local* template matching method, in which only the properties of the local neighbours in the data space influence the APs of the new object. This is in contrast to a neural network, which attempts to do a *global* interpolation of the function APs=f′(data) over the whole data space.

Classification, then, is the process of mapping from the data space to the AP or class space. By contrast, simulation of source SEDs is the opposite mapping, i.e. from the AP space to the data space. This is generally a many-to-one mapping: a given set of APs provides a unique SED but because of photon noise and degeneracies, two sets of APs could produce the same SED (within the noise). Thus the inverse mapping (i.e. classification) is generally one-to-many and not unique.

This is illustrated schematically in Fig. 3. In the left panel we see that there are four templates (those lying within the noise bounds) which give rise to data consistent with the new observation. Confronted with this degeneracy we must decide what to do. Do we quote all results? Do we average the APs? There are in fact whole ranges of the AP which are consistent with the data, so an unweighted average will be biased by the distribution of the nearest templates. Moreover, at large AP, there is actually another solution which we have completely failed to recognise due to the low density of templates in that region. The problem is worse with a lower density template grid (right panel of Fig. 3), or, equivalently, lower noise data. Clearly, a MDM which just assigns the APs of the nearest neighbour or even averages over nearby neighbours will give

[2]The *data space* refers to the data acquired from GAIA, such as fluxes in different filters or the RVS spectrum. The *parameter space* refers to those properties of the sources we wish to determine, such as $T_{\rm eff}$ or extinction, but could also refer to discrete classes (e.g. star, galaxy, quasar etc.).

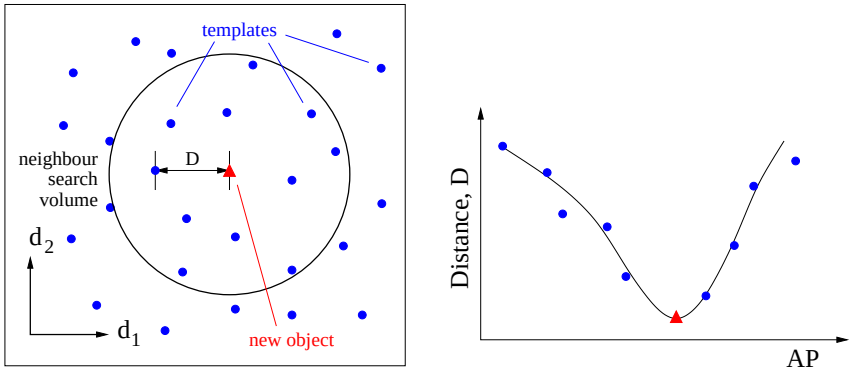

Figure 2. Schematic illustration of the generic minimum distance method (MDM). Left: a two-dimensional data space populated with pre-classified templates. Assigning parameters to a new object involves looking at the APs of the nearest neighbours (with the data dimensions suitably scaled). APs are assigned either by interpolating in the data space (i.e. solving the function APs=f′(data) locally at the new object – in the simplest case this is just an average of one or more neighbours) or in the parameter space (i.e. minimising the function D=g(APs), shown for one AP in the right panel).

biased results. We might want to get around this by having a *very* dense template space, but this will probably become prohibitive if we have a large number of APs. But even this will give rise to biases in the classifications where the mapping function is nonlinear, as one particular neighbour will be preferentially selected. Although the errors would be small for a single star, it could produce a significant systematic error in the average classification of many similar stars.

Thus, any sensible implementation of MDM or other pattern matching method will do some kind of interpolation to provide solutions between templates, i.e. provide us with an approximation to the curve shown in Fig. 3. But this will only be a single-valued function if done in the reverse manner, i.e. data=f(APs). This is the inverse of the function which we would like to have, APs=f′(data), which would enable us to deduce APs given new data. This is important, because it means global interpolation methods for determining the function APs=f′(data), such as neural networks, will give poor interpolants in the presence of data degeneracies. (Think of rotating the left panel of Fig. 3 by 90° and trying to fit a single-valued function through the templates.)

Appropriate design of the GAIA photometric system is a prerequisite to avoiding such degeneracies, but given the relatively few photometric bands and large variety of objects observed, they cannot be avoided entirely. The issue, then, is how to recognise degeneracies and appropriately report multiple solutions without biasing the parameter determinations. This is of course simple in the one dimensional example in Fig. 3 which we can visualise, but it is more complex with 5–10 APs and tens or even hundreds of data dimensions, as will be the case with GAIA. An appropriate solution to this problem is the subject of ongoing work.

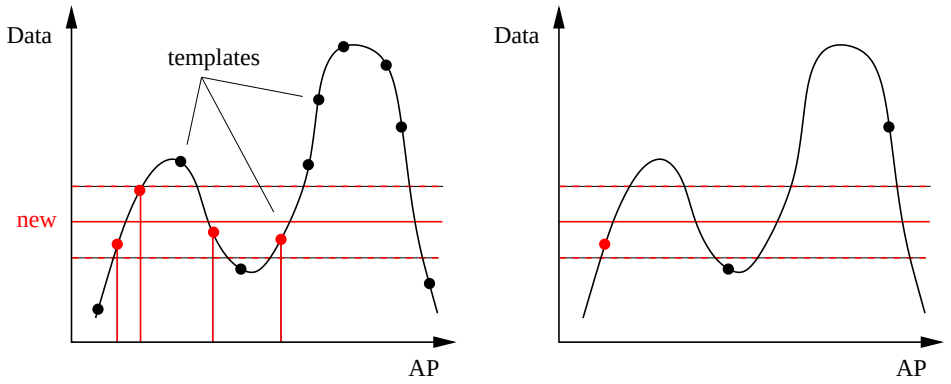

Figure 3. Schematic diagram of the functional relationship between a one-dimensional data and a one-dimensional AP space showing degeneracies, i.e. several AP solutions for a single given data measurement (shown by the horizontal line). The dashed lines show the noise level, so that (in the left diagram) any of the four templates with vertical lines are consistent with the new measurement.

4. Classification schemes

It seems unlikely that a single classification algorithm will be able to deal with the large variance of astrophysical sources which GAIA will observe. (Many results in the literature indicate improved performance when a classification problem is broken down into subsets covering smaller ranges of APs.) This can be dealt with in one of two broadly different approaches. The first, which I describe as the *hierarchical scheme*, uses a Global Classification Model (GCM) (a model which can deal with the entire range of sources) to produce a coarse classification. Based on this, one of several refined classifiers (each of which I call a Local Classification Model, or LCM) is used to produce a more precise classification or set of APs. Each of these LCMs only "knows about" (i.e. is capable of producing good results in) a limited part of the parameter space, e.g. a restricted range of effective temperature.[3] This approach is shown schematically in the left of Fig. 4. An example of such a scheme was presented in Bailer-Jones (2002) (section 5 and Fig. 1).

The alternative approach, shown in the right of Fig. 4, is to pass the data to each of many *local* classifiers right from the start. I refer to this as the *parallel scheme*.[4] The key difference is that every single LCM is given the chance to say something about the data, and, crucially, to provide a probability that this source corresponds to a source of its class (or the range of APs which it deals with). A decision regarding which of these classes the source belongs to (i.e.

[3]Generally, these spaces should overlap between the LCMs to obviate the problem of small classification errors from the GCM occuring at the boundaries between the LCMs, which would result in entirely the wrong LCM being chosen.

[4]In Fig. 4 I show coarse and refined levels in the parallel scheme, but this could be reduced to a single refined level.

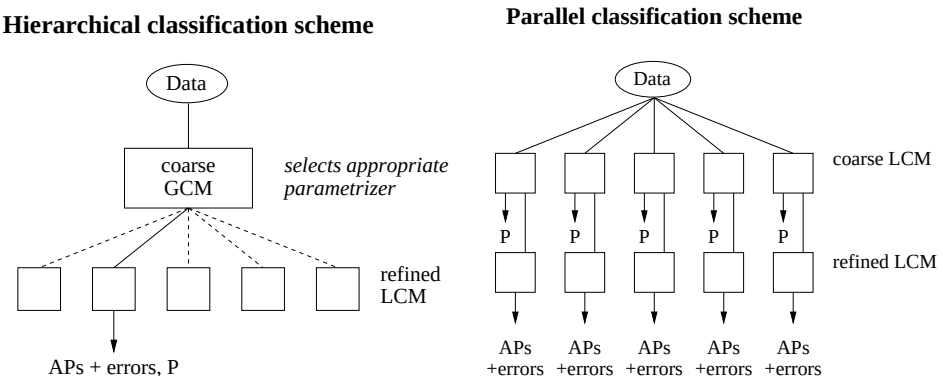

Figure 4. Two alternative classification philosophies. AP = Astrophysical Parameter, P = Probability (that the source conforms to that class or set of APs, GCM = Global Classification Model (one acting on a very wide range of the data space or of source types), LCM = Local Classification Model (one acting on a narrow range of the data space or source types).

which LCM) is then made independently, for example by taking the LCM which yields the largest probability. This approach can be seen from the perspective of Bayesian inference: the determination of the APs within each LCM is the process of determining the posterior probability distribution over the APs, assuming the LCM to be the correct one. The higher level of inference – model comparison – comes about by assessing the *evidence* for each LCM, regardless of the optimal or marginalised APs provided by each LCM, and is provided in this case by the probability, P.

5. A classification framework for GAIA

In the context of the GAIA classification problem, the parallel scheme described in the previous section offers a number of advantages over the hierarchical one.

One of the key advantages relates to the *inhomogeneity of source models*. At some level, classification or parametrization involves matching the observed data to template data of sources with known APs or classifications. A single model which must classify all types of astrophysical sources (even if only coarsely) demands a strong degree of homogeneity of the template data. For example, if a single classifier is to classify both main sequence and pre-main sequence (PMS) stars based on synthetic spectra of such stars, then these synthetic spectra will have to show smooth and self-consistent variations as a function of the APs. In practice, however, different models for different types of stars may be produced independently and according to different assumptions (opacities, treatment of convection and so forth). Thus it may be unrealistic to expect modellers to produce a single, homogeneous grid of synthetic spectra across the full range of APs of stars which GAIA will observe. In the parallel classification scheme, homogeneity is not required. Here, each LCM need only known about a limited set of self-consistent and homogeneous models (e.g. only PMS stars, and perhaps

then only over a small mass range). Each LCM then attempts to classify a data vector according to its own source models, and thus provides the most probable APs, along with error estimates and a probability (P) that this particular data vector is described at all by this ensemble of source models (independently of the specific APs it came up with). So if the data source really is a PMS star, of all the LCMs we would expect the "PMS LCM" to give the highest probability in the parallel scheme, whereas the "evolved giant" and "quasar" LCMs (for example) would give very low probabilities.

The key requirement of the parallel scheme is a robust means of determining the probability (P in Fig. 4) that the data vector is described by that LCM. This could be determined from the distance (in the data space) compared to the data uncertainties between the source position and the different templates in that LCM, with due regard for interpolation errors between the templates. A specific algorithm for this is presently under investigation.

A classification approach using LCMs (rather than a single GCM) not only permits inhomogeneous source models to be used. It also allows very different approaches to classifying different types of astrophysical sources, possibly using different parts of the GAIA data in each case. For example, a Cepheid LCM may want to do a light curve analysis to look for characteristic variability, something which would not be relevant for old G dwarfs.

Another advantage of the parallel scheme is that it naturally provides for multiple solutions in the presence of AP degeneracy (see section 3). This would be evident from several LCMs yielding high probabilities, whereas in the hierarchical scheme only ever one LCM is selected per source (although this condition could of course be relaxed). Thus if more than one LCM provided output probabilities above some threshold, all sets of parameters from these LCMs could be reported.

6. Classification example with GAIA/RVS-like data

Optimization of the GAIA instruments and pre-launch estimates of the AP precision which can be achieved with these must be undertaken using existing real or synthetic data. By way of illustration, I show the results of a stellar parametrization procedure using a neural network applied to RVS-like data obtained and parametrized by Cenarro et al. (2001). The selected dataset consists of 611 spectra covering the wavelength range 849–874 nm, near-critically sampled at a resolution of 5800 (0.15 nm FWHM). The median SNR per resolution is 85, but with a large range (30–170 for 90% of spectra). These spectra were randomly assigned to two nearly equal sized subsets, and a neural network trained on one subset to determine $T_{\rm eff}$, $\log g$ and [Fe/H]. Once training is completed (according to some optimization criterion), the network parameters are frozen and used to determine the APs on the other data subset, from which the performance of the network can be verified. The distribution of the APs in the training and verification sets is shown in Fig. 5.

Figures 6 and 7 summarise the results. $T_{\rm eff}$ can generally be determined to within 5% and shows little trend with $\log g$, although is better determined for near solar metallicity stars. The larger scatter around $\log g$=2.5 may indicate $T_{\rm eff}$ errors in the assignments of Cenarro et al. $\log g$ can be determined to within

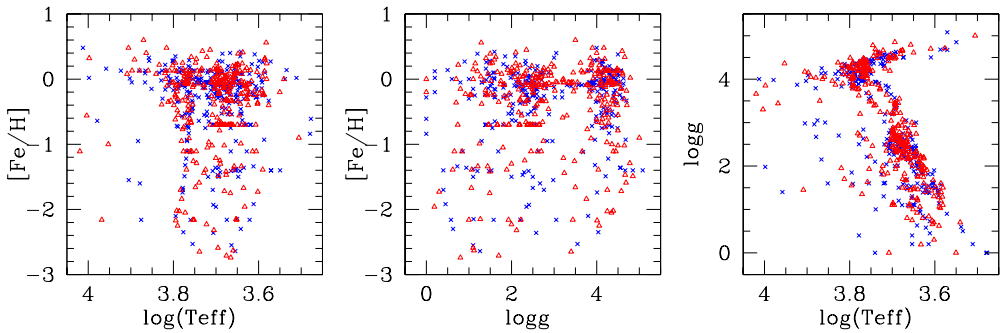

Figure 5. Distribution of the data used from Cenarro et al. (2001) for the training set (blue crosses, 300 spectra) and the verification set (red triangles, 311 spectra).

0.5 dex at solar metallicity, degrading to 1.0 dex or more at lower metallicities. We see that $\log g$ is generally harder to determine at lower metallicities, which we would expect as there the line profiles which the $\log g$ depends upon are weaker. We also see that the $\log g$ of the cooler stars is systematically underestimated for the cooler subset (lower left panel of Fig. 6) across a range of $\log g$. This may indicate a limitation of the algorithm. [Fe/H] precision is 0.3 dex across all temperatures considered (ignoring low number statistics at the extremes), but shows a trend to poorer performance at low metallicity, particularly for evolved stars (lower right panel of Fig. 7. This is expected because the metallicity signature is weaker and small differences are harder to distinguish.

It should be emphasised that these results are based on real data, and therefore include all sources of cosmic scatter. Moreover, the performances assume that the Cenarro et al. parametrizations are true, so only assess the ability of the neural network to reproduce these. Any inconsistencies in that calibration will be reflected by the network. Finally, no attempt has been made to optimize the neural network implementation for this purpose, which furthermore is prone to the degeneracy problem described in section 3..

7. Calibration

The example in the previous section raises the issue of how the parametrization algorithms for GAIA will be calibrated, or, in other words, how the training/template data set(s) will be defined and parametrized. Unfortunately, a homogeneous database of real spectra to serve as the GAIA templates – covering the required wavelengths and APs – does not presently exist. Even if it did, it would presumably consist primarily of ground-based spectra which would need to be processed to remove telluric features (and still the UV data would be lacking), and APs would still have to be assigned to those spectra. Obtaining stellar parameters ultimately requires some kind of stellar model. The emergent spectral energy distributions derived from such models can be used directly in the parametrization process by training pattern recognition methods on such spectra, after suitably processing them with the instrument model to look like

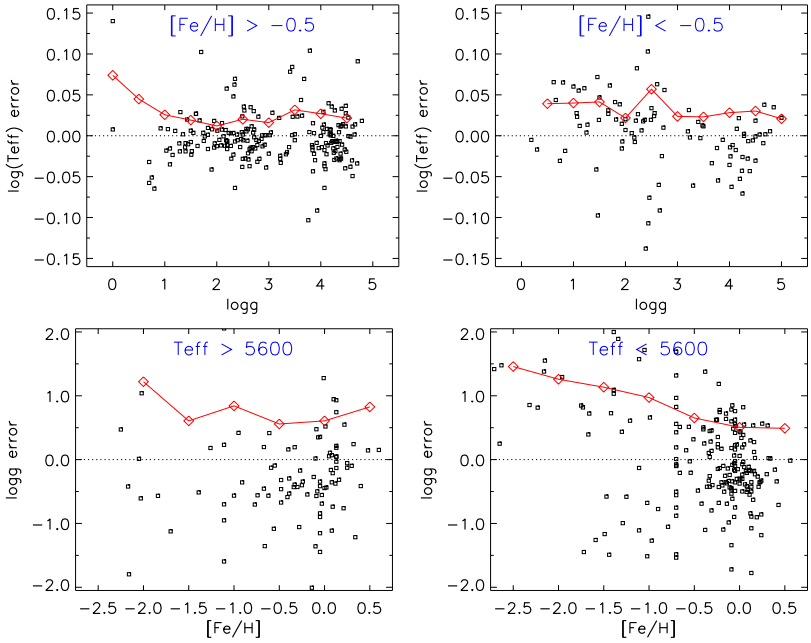

Figure 6. $T_{\rm eff}$ (top row) and $\log g$ (bottom row) parametrization errors on the verification data set. Each point corresponds to a single spectrum and gives the network determination minus the "true" value (as established by Cenarro et al.). The diamonds joined by a line show the RMS error for all spectra in a bin centered on that point. The $T_{\rm eff}$ errors are plotted as a function of $\log g$ for two metallicity ranges. The $\log g$ errors are plotted as a function of [Fe/H] for two temperature ranges.

the data GAIA will obtain (e.g. Bailer-Jones et al. 1997). However, synthetic spectra differ from real spectra in two significant ways. First, they may show systematic differences due to modelling uncertainties (e.g. missing opacities). Second, real spectra show increased cosmic scatter due to unaccounted-for APs (e.g. abundance variations, chromospheres, etc.). So training models on synthetic spectra to apply to real spectra is not ideal.

Fortunately, there is a way around these problems. To assign parameters to GAIA observed objects we need to know: (1) how these objects will appear in the GAIA multidimensional data space; (2) what the "required" APs for these objects are (where "required" means just those APs which can be derived in principle from the GAIA data). However, we do not have to determine the APs of the templates from the same data that we use in the training. We may define a grid of real star ("calibration stars") on the sky which GAIA will observe, covering the full range of APs at some suitable AP density. Ground-based spectra of each calibration star is obtained with whatever resolution and wavelength coverage is required to determine its APs (probably from detailed line fitting) to limits set by our physical knowledge of stars and the quality of data we can obtain. In some cases existing data or catalogues can be used

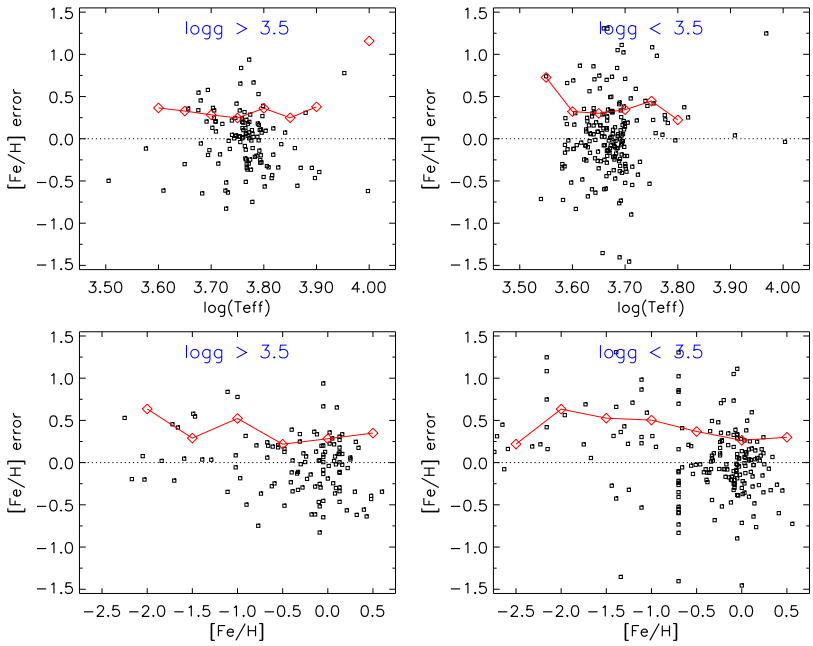

Figure 7. [Fe/H] errors on the Cenarro et al. data (see Fig. 6).

for this purpose. The GAIA observations form the data for these calibration stars, so at the end of the GAIA mission we have both the data and APs on a set of stars which can serve as templates to train our classification algorithms, unaffected by real-synthetic data mismatch.

Prior to and during the mission, GAIA data can be simulated using synthetic spectra, permitting approximate parametrizations. If it turns out that the sky grid of calibration stars is not dense enough in some regions of the AP space, it can be supplemented with synthetic spectra, with broad corrections applied to them to account for systematic differences, in a manner similar to that described by Lejeune et al. (1997). This calibration method would require a ground-based observing program. But it would be on a modest scale, requiring of order 1000 high resolution spectra. Accurate flux calibration is not required, and depending on the wavelength coverage required, might be obtainable with a multi-object fibre or slit spectrograph. Of order 10 nights on 4m and 8m class telescopes would probably suffice.

References

Bailer-Jones, C.A.L., 2002, Ap&SS, 280, 21

Bailer-Jones, C.A.L., Irwin, M., Gilmore, G., von Hippel, T., 1997, MNRAS, 292, 157

Cenarro, A.J., Cardiel, N., Gorgas, J., Peletier, R.F., Vazdekis, A., Prada, F., 2001, MNRAS, 326, 959

Lejeune, T., Cuisinier, F., Buser, R., 1997, A&AS, 125, 229