Conference Milky Way Halo, Bonn 2007, Poster 38

Astrophysical Parameter Estimation from Gaia: Comparative Results of Different Machine Learning Methods

Carola Tiede, Kester Smith, Coryn Bailer-Jones

Max-Planck Institute for Astronomy, Königstuhl 17, 69117 Heidelberg, Germany
Email: tiede@mpia.de

Introduction

Gaia is an all sky astrometric and spectro-photometric survey complete to magnitude G=20 (V=20-22). It will map about 10^9 stars, a million quasars as well as a few million galaxies. Detailed information about Gaia can be found in (Bailer-Jones [2004]) or under http://www.rssd.esa.int/Gaia/.
Gaia works without any input catalogue so that an automatic classification and AP estimation become indispensable. The red and blue low resolution photometers provide the input spectra (RP/BP) in addition to parallax for the classification and AP estimation. Both algorithms are currently based on simulated data.
Here the parameter estimation for single stars is described in which the four APs log(Teff), log(g), [Fe/H] and Av are estimated for magnitude 15 and 17 and including noise.
Figure 1 shows the variation of the spectra related to variations in a) log(Teff) and b) [Fe/H]. The photometric spectra are mostly sensitive to changes in log(Teff) (and Av), whereby variations in [Fe/H] (and log(g)) do not strongly affect the photometric spectra at all which makes the parameter estimation for these so called weak parameters ([Fe/H] and log(g)) much more difficult.

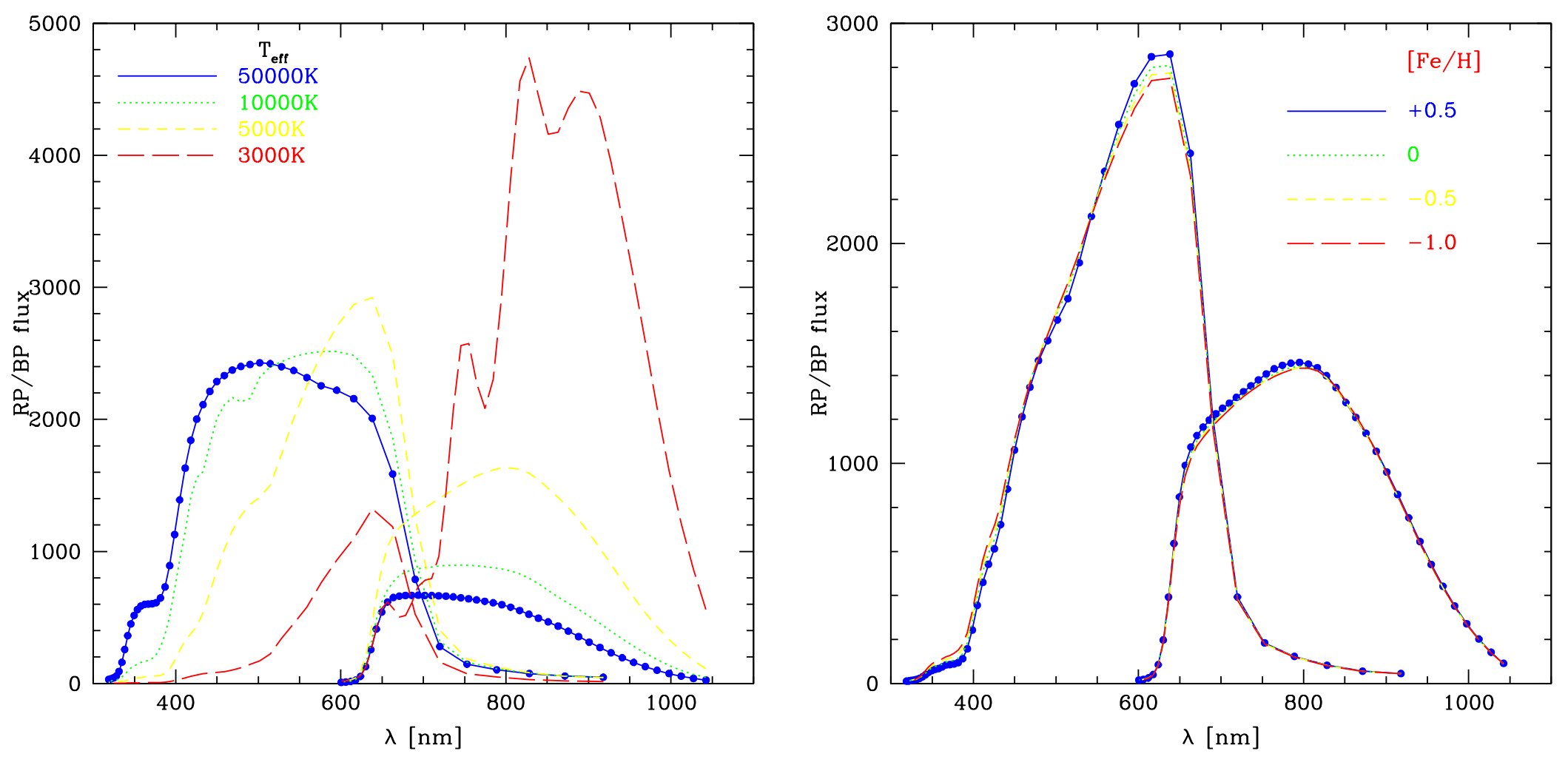

FIGURE 1: Sensitivity of photometric spectra regarding variation in a) log(Teff) and b) [Fe/H]

Parameter Estimation

The parameter estimation is based on 96 bins describing the combined RP/BP photometric spectra and an extra one holding the parallax. 8000 samples are taken from the simulation data at random, consisting of RP/BP spectra, parallax and of a-priori known APs which are used for the training of the model. Each value j of each bin i of the RP/BP spectra and the parallax is normalized by the mean μ_i and the standard deviation σ_i.

$$xn_{i,j} = (x_{i,j} - \mu_i)/\sigma_i \qquad (1)$$

The quality of the classifier performances is given by the root mean square error (RMSE) per AP:

$$RMSE = \sqrt{\frac{\sum_{j=0}^{n}(AP_{est(j)} - AP_{true(j)})^2}{n}} \qquad (2)$$

with n=number of samples (here 8000).
As a benchmark the APs are estimated using k-nearest neighbor with k=1.
In comparison to the local k-nearest neighbor approach, a global fit is computed by training a support vector machine and applying its resulting model to the new data for which the APs are not known a-priori.

K-Nearest Neighbor

K-nearest neighbor searches for the nearest sample in the training data by computing the Euclidean distances between the new sample and the training set samples. It applies all APs of the nearest training set sample to the new data sample. Figure 2 shows the residuals computed for the 4 APs log(Teff), log(g), [Fe/H] and Av plotted versus their true values. The residuals of log(Teff) show the degeneracy between log(Teff) and Av. Furthermore it becomes obvious that k-nearest neighbor is not suitable for the estimation of the so called weak parameters log(g) and [Fe/H].

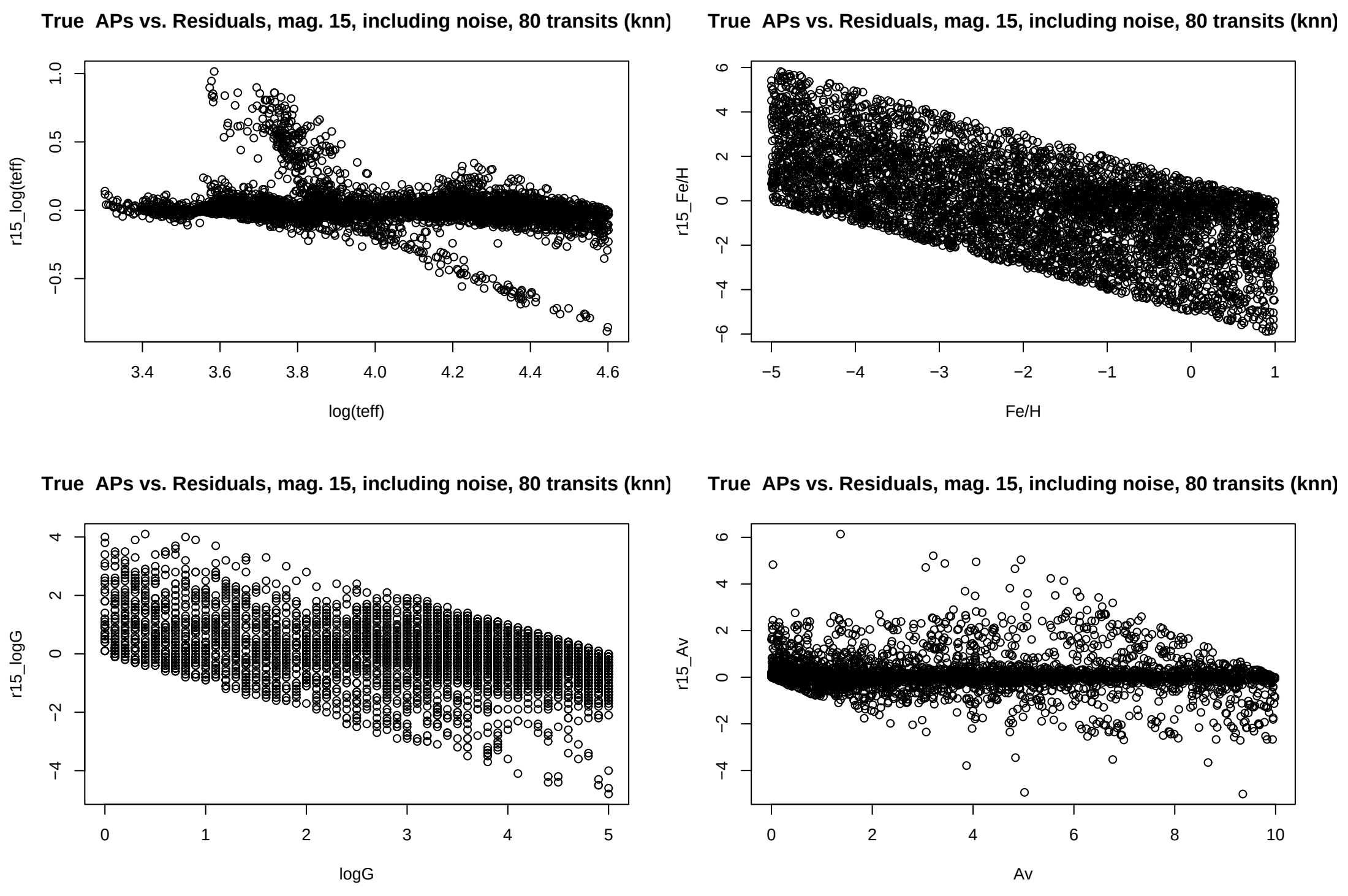

FIGURE 2: Residuals, magnitude 15 of log(Teff), log(g), [Fe/H], Av using k-nearest neighbor.

Support Vector Machines

Support vector machines (svm) after Vapnik [1995] are used. Generally, svms seperate classes by finding an optimal hyperplane with a maximal margin between the different classes. If the input data cannot be seperated in a linear way, the data are mapped into a higher dimensional feature space in which the linear seperation is used. Svms with the inclusion of an RBF kernel are used to treat the nonlinearity of the data. The parameters C (penalty term) and γ (kernel specific) are configured by a grid search.
The residuals of the 4 APs are plotted versus the true values in Figure 3 for magnitude 15 and in Figure 4 for magnitude 17.

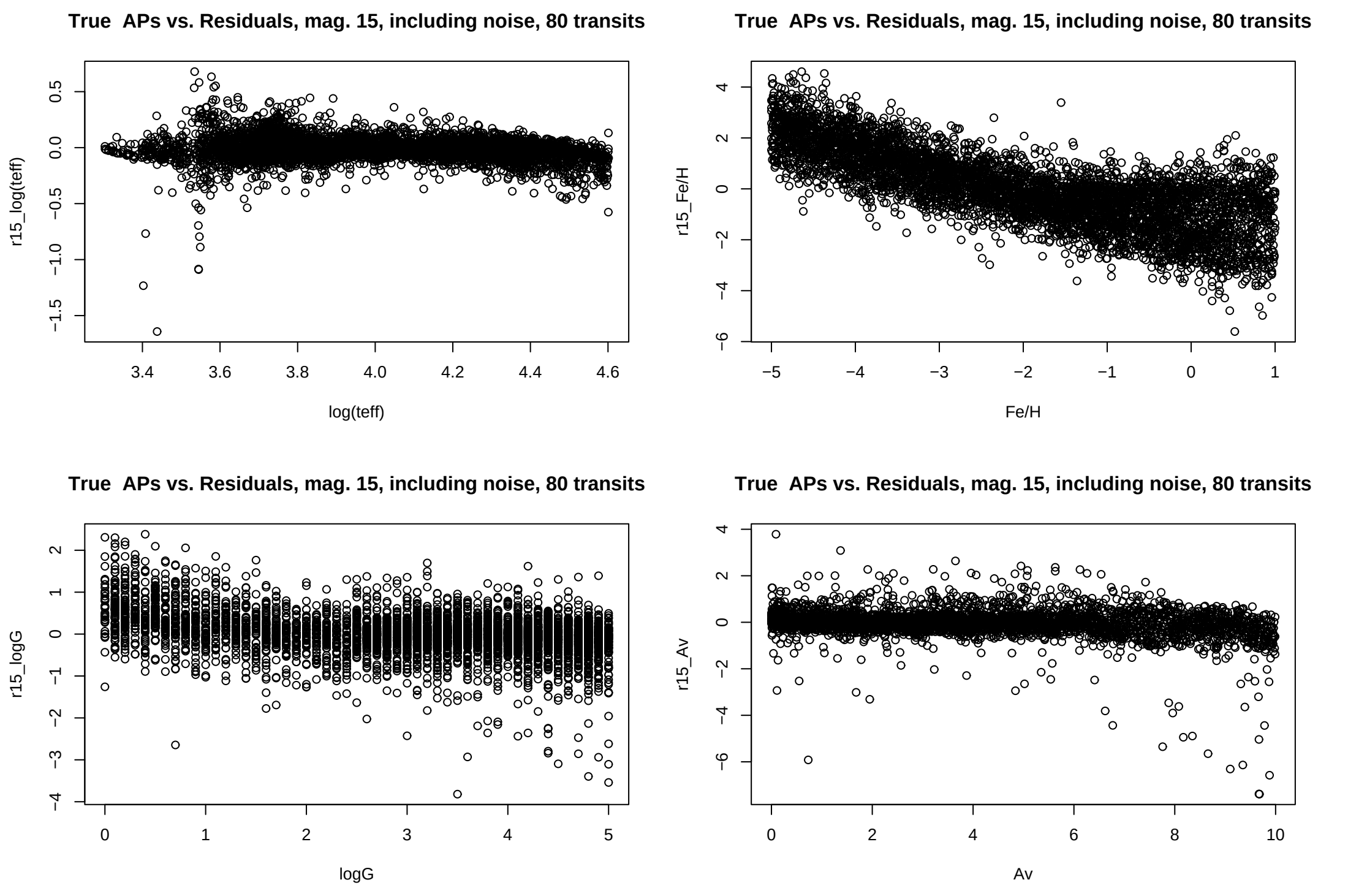

FIGURE 3: Residuals, magnitude 15 of log(Teff), log(g), [Fe/H], Av using svm.

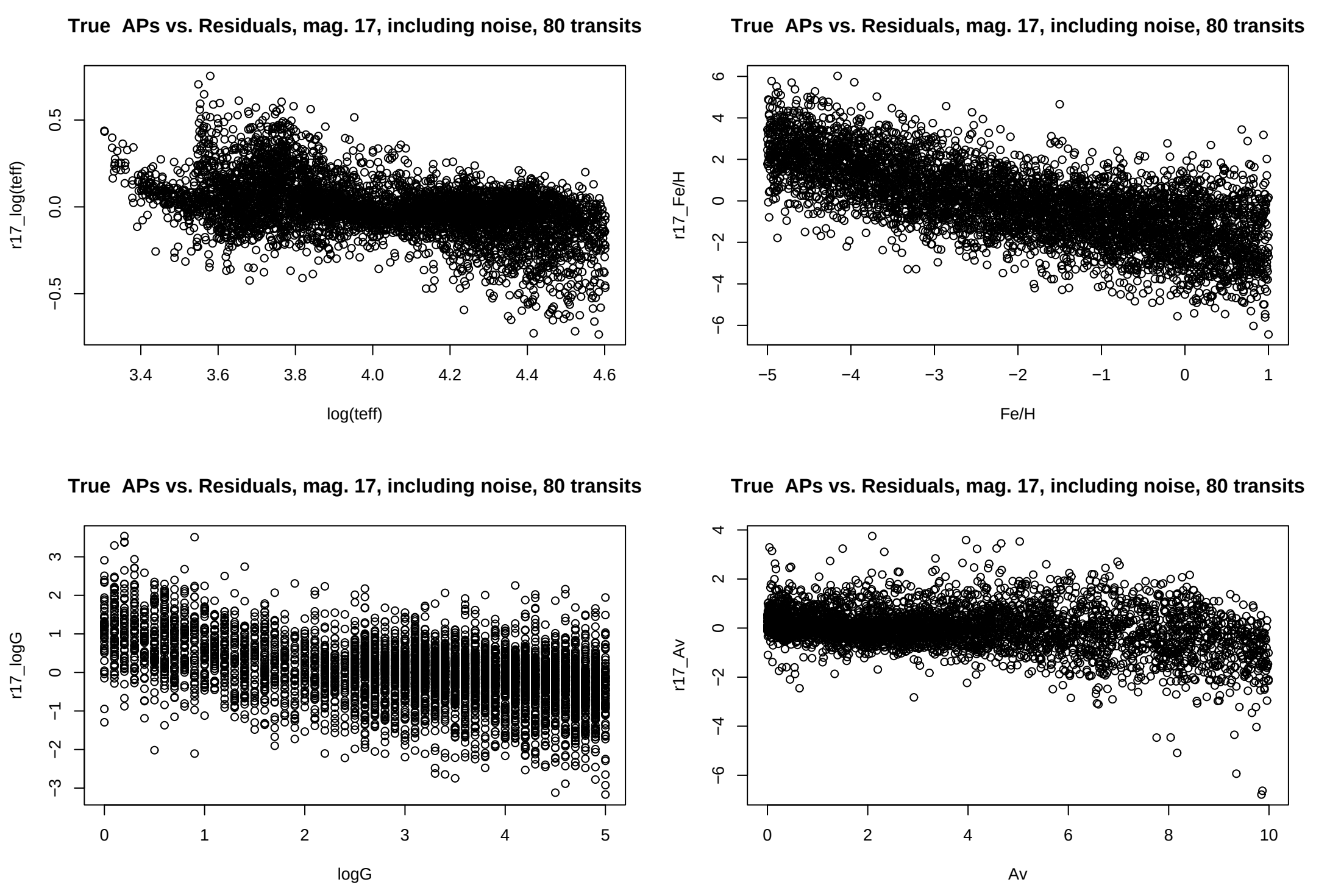

FIGURE 4: Residuals, magnitude 17 of log(Teff), log(g), [Fe/H], Av using svm.

It can be seen that especially the weak parameters log(g) and [Fe/H] show a trend in the residuals.

Results

The weak parameters [Fe/H] and log(g) can be estimated more precisely by using svm whereby the results for the strong parameters log(Teff) and Av remain similar, see Table 1. Analysing the sensitivity of the training data (RP/BP photometric spectra + parallax) regarding changes in the APs, Figure 1 shows that nearly all variation of the photometric spectra is caused by variation in the strong parameters. By using k-nearest neighbor as classifier, the Euclidian distance between the new sample and the training samples is estimated whereby this distance is dominated by effects in the strong APs so that the weak APs cannot be estimated precisely with k-nearest neighbor. When using svm, the model for each AP is trained separately so that the whole variation in the photometric spectra is expressed by the variation of each single AP. The residuals given in Figure 3 and Figure 4 (and Figure 2) include linear trends which shows that the model does not correctly fit the data. This trend (linear trend of form $y = Ax + b$) is estimated by a least squares fit and is subtracted from the residuals. The improvement is shown in the RMSE in Table 1 and can be seen especially for the weak parameters log(g) and [Fe/H] and for larger magnitude (17).

Parameter	Svm	Svm + lsq	knn=1
a) magnitude 15			
log(Teff)	0.1059	0.0978	0.1110
log(g)	0.5496	0.4709	0.8910
[Fe/H]	1.5303	0.8023	1.9422
Av	0.5667	0.5348	0.5465
b) magnitude 17			
log(Teff)	0.1661	0.1452	0.2214
log(g)	0.7708	0.6469	1.2968
[Fe/H]	1.6766	0.9497	2.2035
Av	0.7433	0.6987	0.9721

TABLE 1: AP estimation accuracy measured in RMSE. The first column shows the result computed by using svm, the second column shows the results using svm and correcting the residuals by a least squares fit, the third column shows the results using k-nearest neighbor with k=1 as classifier for a) magnitude 15 and b) magnitude 17.

For analysing the degeneracy of log(Teff) and Av shown in Figure 2 in more detail the residuals of log(Teff) are plotted versus the residuals of Av in Figure 5 with the color-coding regarding the true log(Teff) (black-red: log(Teff)=3.3...4.6) using a) k-nearest neighbor and b) svm as classifier. The large degeneracy between log(Teff) and Av is shown in the k-nearest neighbor result. Hot stars are estimated too cold with a too small extinction, cool stars are estimated too hot with a too large extinction. This degeneracy is much smaller in the svm results.

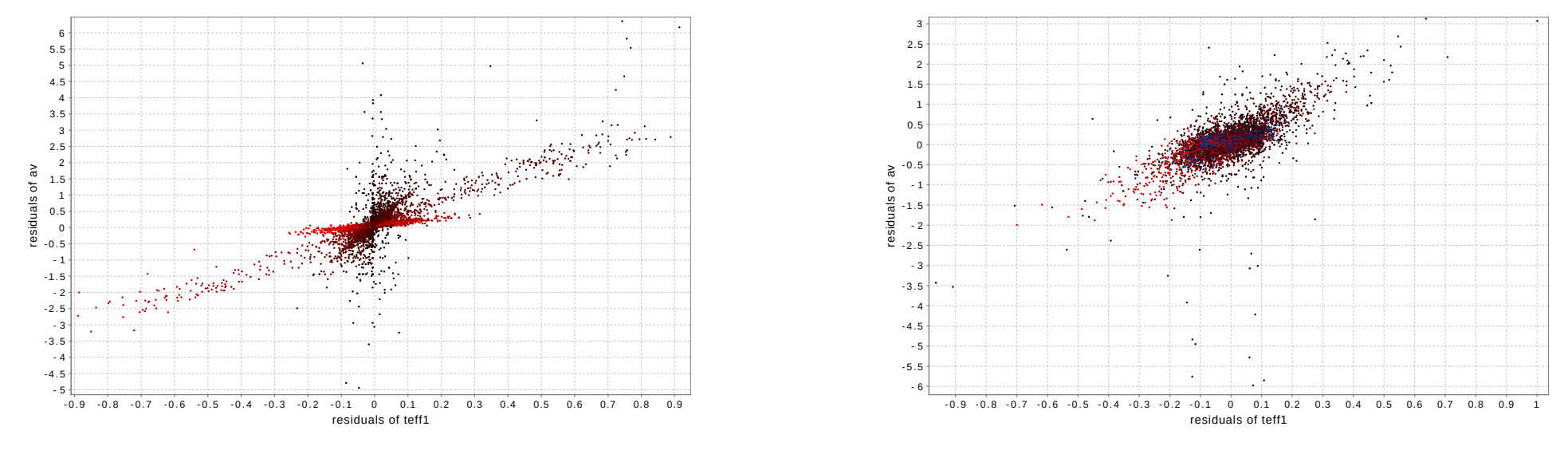

FIGURE 5: Residuals of log(Teff) versus residuals of Av, magnitude 15, using a) k-nearest neighbor and b) svm

References

C.A.L Bailer-Jones. *Microarcsecond astrometry with Gaia: The solar system, the Galaxy and beyond*, pages 429–443. in Transit of Venus: New Views of the Solar System and Galaxy, Proc. IAU Colloquium 196, Cambridge University Press, 2004.

http://www.rssd.esa.int/Gaia/

V. Vapnik. *The Nature of Statistical Learning Theory*. Springer Verlag, New York, 1995.