AUTOMATED IDENTIFICATION OF UNRESOLVED BINARIES USING MEDIUM BAND PHOTOMETRY

P.G. Willemsen[1], **T.A. Kaempf**[1], **C.A.L. Bailer-Jones**[2], **K.S. de Boer**[1]

[1]Sternwarte der Universität Bonn, Auf dem Hügel 71, 53121 Bonn, Germany
[2]Max-Planck-Institut für Astronomie, Königstuhl 17, 69117 Heidelberg, Germany

ABSTRACT

We have simulated Gaia medium band photometry in the 1X system for (synthetic) binaries and single stars for G = 15 to 20 mag end-of-mission as well as single shot photometry. Each binary is a system of five independent physical parameters (the components' masses M_1, M_2, age, metallicity and extinction), which translates into six parameters needed to define the binary's spectrum ($T_{\rm eff_1}$, $T_{\rm eff_2}$, $\log g_1$, $\log g_2$, [Fe/H], A_V). To see how the underlying grid of parameters influences the classification, we simulated two sets of binary/single stars. Using an automated classification algorithm known as Support Vector Machine, we show that the capability of correctly identifying binaries is a strong function of the signal to noise ratio and the luminosity ratio of the two components. The maximum of correctly identified binaries is reached at logarithmic luminosity ratios of log(L_r) in the range [0.5;1.0]. For these ratios, correct classification rates better than 70% at G = 15 and 16 mag are possible. The differential analysis for the two different parameter distributions further shows that the underlying grid also strongly influences the classification results.

Key words: Classification; Binary stars; Gaia.

1. SIMULATIONS

The masses of the two binary components are randomly drawn from a mass generating function (with $M_2 \leq M_1$) based on an IMF as given in Kroupa et al. (1991) and Kroupa (2001). Binaries which are calculated based on this IMF approach are referred to as set A. To allow for a test of how the distribution of stellar parameters affects the classification performance, we simulated a second set where masses are drawn from a uniform distribution (always $M_2 \leq M_1$). This is referred to as set B. Since the simulations naturally yield more main-sequence (MS) stars for the two binary components, we artificially increased the number of RG-MS or RG-RG (RG = Red Giant) combinations. The distinction of the two components in a binary is here done by mass: The 'primary' component has a higher mass than the 'secondary' component.

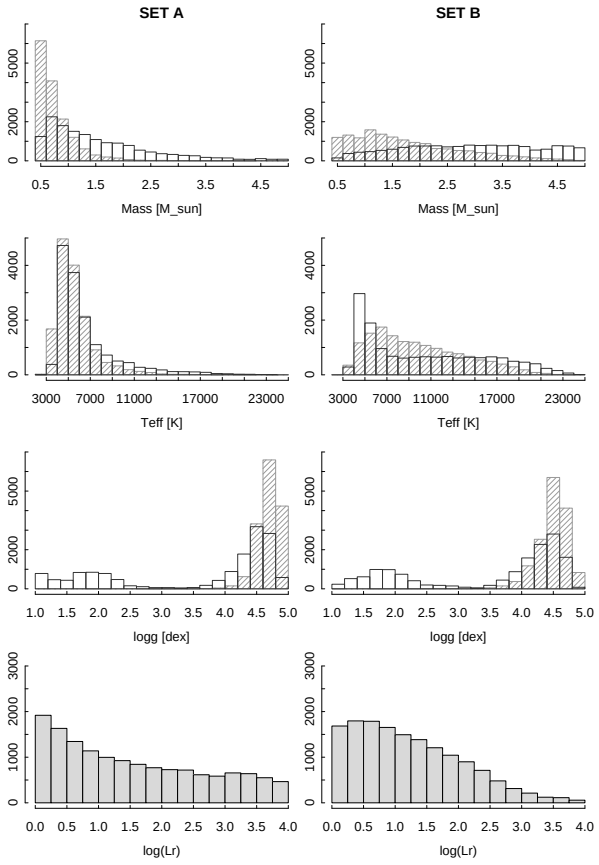

Figure 1. The distributions of the binary stellar parameters for the complete set A (left column, based on an IMF) and set B (right column, based on a uniform mass distribution). Both sets include an excess of 5000 main-sequence / red giant star combinations. From top to bottom are shown the distributions of the masses (in units of solar mass), temperatures $T_{\rm eff_1}$, $T_{\rm eff_2}$, gravities $\log g_1$, $\log g_2$ and logarithmic luminosity ratio log(L_r) = $\frac{L_{\rm prim}}{L_{\rm sec}}$. For the masses, temperatures and gravities, the first component's parameters are given by the solid, white histograms, while those of the second component are represented by shaded histograms. The distributions of [Fe/H] and A_V are essentially uniform and not shown. The corresponding histograms for the simulated single stars are similar to those of the binary's first component.

Table 1. The ranges of the astrophysical parameters. The maximum allowed age for a star or binary system was chosen to be 13.6 Gyr (Spergel 2003). The mass and metallicity ranges for the two binary components are both defined by the evolutionary tracks which are used to calculate the corresponding $T_{\rm eff}$ and $\log g$ values.

parameter	range
mass	[0.4 ; 5] $M_\odot$
age	]0 ; 13.6] Gyr
[Fe/H]	[−3.6 ; 0.6] dex
A_V	[0 ; 5] mag
$T_{\rm eff}$	[2500 ; 24 000] K
$\log g$	[1 ; 5] dex

With a randomly chosen metallicity and age value we interpolated in a grid of evolutionary tracks taken from Yi et al. (2003) to find $T_{\rm eff}$ and $\log g$ for each component. With these and a randomly chosen extinction value (using an extinction curve from Fitzpatrick 1999), we interpolated in the Basel2.2 grid (Lejeune 1997) of synthetic spectra to find the corresponding stellar energy distributions which are then added up. The spectra of single stars were simulated correspondingly.

To calculate Gaia medium band fluxes in the 1X system (Vansevičius 2002), we used the photometry simulator by Bailer-Jones (Bailer-Jones 2002, version 2) for magnitudes of G = 15 to 20 mag for two cases: single shot (SS), i.e., only one simulated observation and end of mission (EM) photometry. There are 6000 objects (3000 of each class: binary/single star) for training and 24 000 objects (12000 of each class) for validation purposes for each set and magnitude case. The distributions of the major parameters are summarized in Figure 1 and Table 1.

2. SUPPORT VECTOR MACHINES

For the classification (binary/single star) we used an algorithm which is known as Support Vector Machine (SVM, Vapnik 1995).

A SVM separates classes (binary/ single star) by finding an optimal hyperplane which divides the training data by a maximal margin. In general, the SVM performs an implicit mapping of the training data into a high-dimensional feature space where a linear separation should be possible. The problem can then be solved since a linear separating hyperplane in a high-dimensional feature space results in a nonlinear separating hypersurface in the original input space. The solution is found in a quadratic optimization problem with inequality constraints which can be solved by a Lagrange function with well defined properties.

The application of SVMs requires adjusting of two controlling parameters which define the generalization performance. We evaluated the optimum values by five-fold cross validation on the training set. For the present simulations we used the latest version of the R software package $e1071$ (R-project).

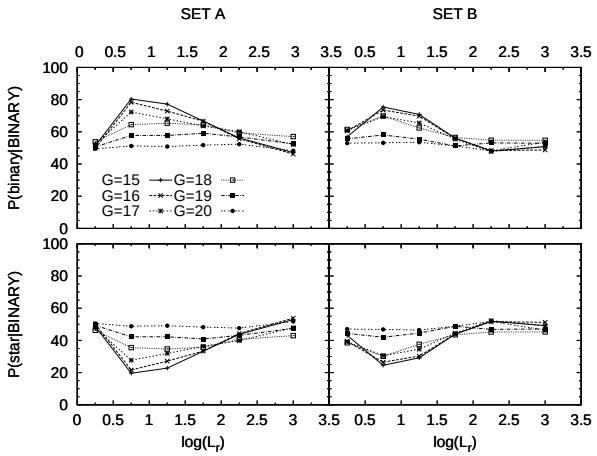

Figure 2. The conditional probability P(binary|BINARY), i.e., the rate at which a true binary is classified as a binary as a function of logarithmic luminosity ratio log(L_r) for different end-of-mission magnitudes. Each data point is the average over a range of logarithmic luminosity ratios and is plotted at the midpoint of that specific range. For example, the points at log(L_r) = 0.25 are for 0.0 < log(L_r) ≤ 0.5. The last interval covers the range 2.5 ≤ log(L_r) ≤ 3.5. The lower plot shows the corresponding probability P(star|BINARY), i.e. the rate at which a binary is falsely classified as a star (100 – P(binary|BINARY)). The number of points averaged in each luminosity range can vary by factors of 2.5 but are always larger than 1000. Note that a classification rate of 50% corresponds to that of a random classifier.

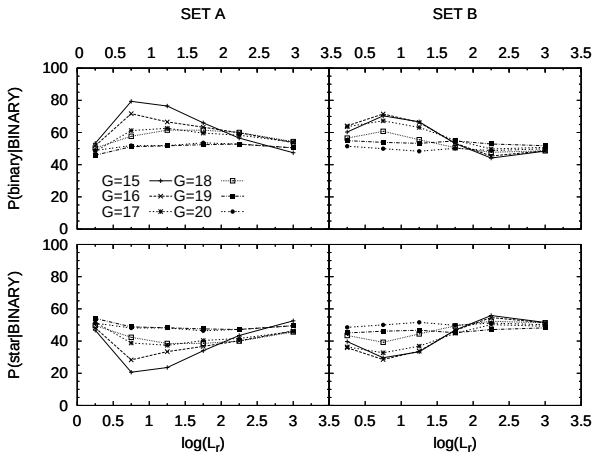

Figure 3. The same as in Figure 2 but for single shot (SS) photometry.

3. ANALYSIS

3.1. The Dependence of Classification on the Luminosity Ratio

We analyzed how well the unresolved binaries can be identified as a function of logarithmic luminosity ratio $\log(L_r)$ and magnitude for the two different cases of underlying parameters (set A and B). The results for end-of-mission (EM) photometry are shown in Figure 2 and those for single shot (SS) photometry in Figure 3.

At low luminosity ratios ($\log(L_r) \in]0.0;0.5]$), the classification is almost random for set A ($\sim$50%) even for brighter systems, but not for set B: clearly, the more uniform distribution of temperatures and masses in set B allows a better separation at small luminosity ratios (see Figure 1). In set A, stars with small luminosity ratios are rather similar (MS-MS combinations) so that the spectral differences of the two components will be rather small (the difficult classification cases). In set B, there is a broader range of temperatures/masses so that it is more likely that there are stars with different temperatures in the binary system.

For larger luminosity ratios ($\log(L_r) \in$ [0.5-1.0] and [1.0-1.5]), we see a maximum in the classification performance with a decline again for high luminosity ratios. The performance for set B at the maximum is always below that of set A but the overall shapes of the curves look very similar. From Figure 1 we see that there are more objects at higher luminosity ratios per luminosity interval in set B than there are in set A. Since we plot limited ranges of $\log(L_r)$, the results of set B thus always include binaries with higher luminosity ratios (on average). For higher absolute ratios, this means that a binary is naturally more often falsely classified as a single star.

For high luminosity ratios, the performance to recognize the object as a binary declines to that of a random classifier (50%) for very large $\log(L_r)$.

3.2. The Dependence of Classification on Temperatures

Figures 4 and 5 show the classification performance as a function of the first component's temperature and the overall luminosity ratio for end-of-mission photometry and single shot photometry, respectively.

We observe large differences in the classification results for set A and B for low temperatures (only temperature of first component), while for higher temperature ranges the performance is rather similar for different luminosity intervals. For set A at low temperatures, $G = 15$ mag (EM) objects we find a difference of almost 60 percentiles between the classification performance for very different luminosity ratio intervals, while for higher temperatures (8000 K $< T_{\rm eff_1} \leq$ 12 000 K) the difference is only $\simeq 30$ percentiles. The same trend but even more pronounced is observed for set B. Here we find that binaries where the first component has a low temperature (first and second row) in combination with high luminosity ratios (L4 to L6) are consistently misclassified: almost 90% of the binaries are classified as stars and this trend only stops for faint objects where the classification becomes random.

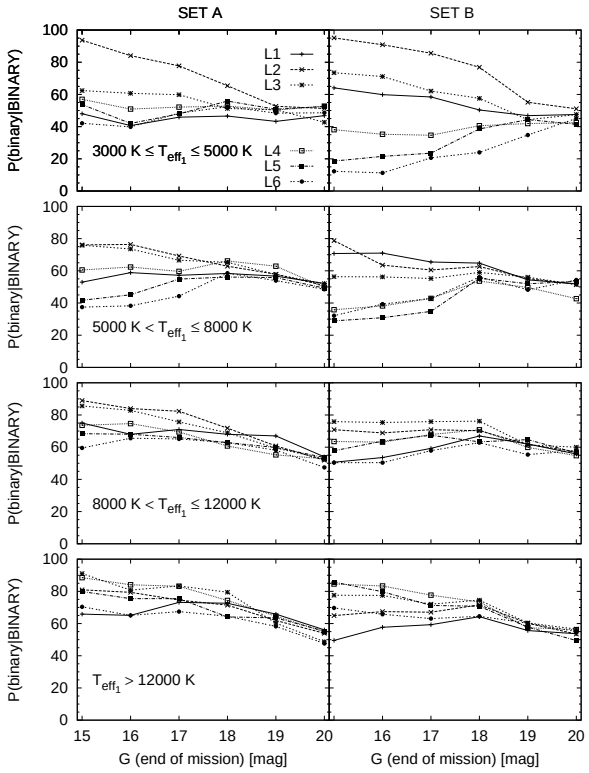

Figure 4. The conditional probability P(binary|BINARY) as a function of magnitude for different luminosity ratios and first component's temperature ranges for set A (left column) and set B (right column), only for end-of-mission photometry. There are four temperature intervals (temperature of the first component) from top to bottom and six (logarithmic) luminosity ranges for each temperature range. ***L1****: $0.0 < log(L_r) < 0.5$,* ***L2****: $0.5 \leq log(L_r) < 1.0$,* ***L3****: $1.0 \leq log(L_r) < 1.5$,* ***L4****: $1.5 \leq log(L_r) < 2.0$,* ***L5****: $2.0 \leq log(L_r) < 2.5$ and* ***L6****: $2.5 \leq log(L_r) \leq 3.5$.*

This systematic trend can also be seen in set A for intermediate temperatures (second row), but to a much lesser extent. A possible explanation for this is that binaries with high luminosity ratios, where one component dominates the light, are seen as single stars. However, this is only part of the answer as can be seen from a comparison of set A and set B (EM) performances. Set B shows stronger systematic misclassifications than set A, at least for low- and mid-temperature regimes (first and second row). If the first component has a low temperature and there is a high $\log(L_r)$, then the second component is almost always a lower MS star for set A, given the constraint that $M_2 \leq M_1$ and the steep slope of the IMF. As a result, the total system will be very similar to that of a single Red Giant. Since there are many red objects in the training/validation set (see the temperature distributions in Figure 1), and since no object type is preferred over the other, the decision boundary probably runs randomly

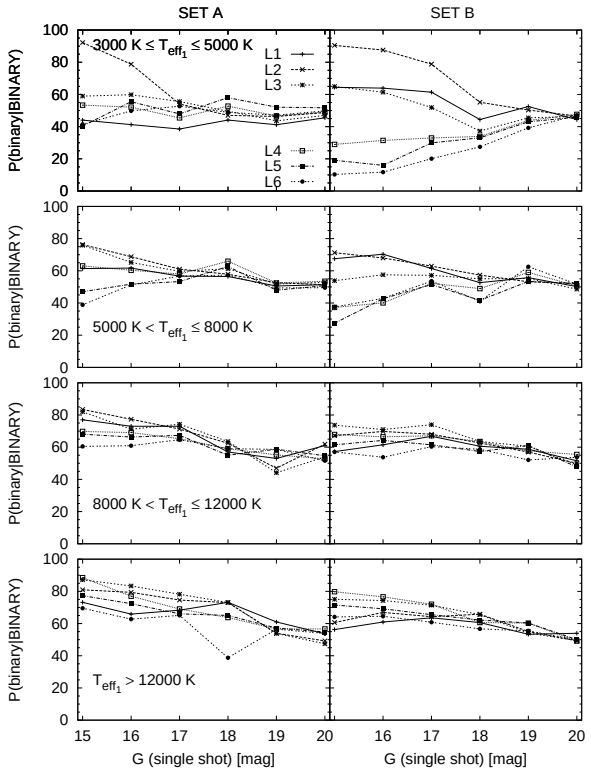

Figure 5. The same as in Figure 4 but for single shot photometry.

between these objects in the 11 dimensional filter space.

In contrast, for set B, there are more cases for high log(L_r), where the second component is a high temperature object on the upper MS if the first component is a Red Giant. Since there are fewer red objects and since both object types cover a larger range of temperatures, it is thus possible that the Red Giant is indeed seen and misclassified as a single star, i.e., the decision boundary at low temperatures (and high log(L_r)) is biased towards stars.

The overall similarity of the figures for the different photometries (end of mission and single shot) at least for higher temperatures of the first component are probably due to systematic misidentifications of the classifier which are independent of the overall S/N.

4. CONCLUSIONS

The results show that:

1. the maximum of correctly identified binaries is at logarithmic luminosity ratios of log(L_r)$\in$ [0.5;1.0].
2. the classification performance depends on the signal to noise andon the underlying grid of physical parameters.
3. the overall correct classification rate is rather low, thus indicating that the identification of these objects by medium band photometry alone is difficult.

The complete study, including the determination of stellar parameters from simulated unresolved binaries can be found in Willemsen et al. (2004).

REFERENCES

Bailer-Jones, C.A.L. 2002, Gaia techical report ICAP-CBJ-004

Fitzpatrick, E.L., 1999, PASP, 111, 63

Kroupa, P., Gilmore, G., Tout, C. A. 1991, MNRAS, 251, 293

Kroupa, P. 2001, MNRAS, 322, 231

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

The R project for Statistical Computing, 2004, http://www.r-project.org/

Spergel, D.N., Verde, L., Peiris, H.V., et al., 2003, ApJS, 148, 175

Vansevičius, V., Bridžius A., 2002, Gaia technical note Gaia-VIL-008

Vapnik, V.N., 1995, The nature of statistical learning theory, New York: Springer - Verlag

Willemsen, P.G, Kaempf, T.A, Bailer-Jones, C.A.L. 2004, Gaia techical report, ICAP-PW-003

Yi, S. K., Kim, Y., Demarque, P. 2003, ApJS, 144, 259