Communications of BAO, Vol. 73, Issue 2, 2026, pp. 346–349

Towards Synthetic Training Samples for Multiple Stars: Structure of the Main Sequence

I.V. Chapaeva *

Faculty of Physics, Lomonosov Moscow State University, Moscow, 119991, Russia

Abstract

The volume of modern observational data on binary and multiple stars makes manual analysis impractical, so it is reasonable to apply artificial-intelligence methods. Obtaining training samples for such methods is not easy: data taken directly from telescopes and databases are unlabeled, while data from numerical simulations do not reproduce observational errors and selection effects. The key task is therefore to create suitable training samples for binary and multiple stars, and as a solution, it is proposed to use synthetic samples that preserve the statistical properties of real observational data. This paper considers the method of obtaining synthetic training samples using one example, as well as a side result produced by this method, namely the inhomogeneous structure of the main sequence. In the longer term, analysing a large number of real distributions will allow a large synthetic training sample to be assembled, and this work may stimulate wider study of the physical causes behind the distributions of multiple stars.

Keywords: *machine learning, training samples, main sequence*

1. Introduction

Astronomical data cannot be obtained under controlled conditions: the observed objects cannot be altered or re-measured, and the results depend on the instruments, the observing conditions, and the processing methods. Catalogues, moreover, often combine results from heterogeneous surveys. For binary and multiple stars, this is particularly significant: their manifestations are more diverse than those of single stars, selection effects are more complex, and the components and the pair they form may appear close to each other in various parameter spaces Kaygorodov et al. (2023).

The goal of the project is to develop a machine-learning model that identifies patterns in multidimensional data. Objects belonging to the same pattern have similar types and evolutionary stages, which allows to classify the data of modern sky surveys automatically, while objects outside any pattern may point to unusual or poorly studied phenomena. Training such a model requires sufficiently large samples with known properties, but real data do not provide them: few objects are labeled, and the interpretation of some of them is ambiguous. Data from numerical simulations, in turn, do not reproduce observational errors and selection effects. It is therefore proposed to generate synthetic samples with known parameters that preserve the statistical properties of real data. Unlike theoretical simulations, this approach accounts for real errors and selection, and it is cheaper.

The first stage of the work is a quantitative description of the distributions observed in real data. This paper demonstrates this stage using the colour–magnitude diagram and discusses a side result, the inhomogeneous structure of the main sequence.

2. Data

The Multiple Star Catalog (MSC; Tokovinin (2024)) is chosen as the source of real observational data. This choice is justified by the fact that the MSC is one of the most complete modern catalogues of multiple systems, collecting astrometric, photometric and spectroscopic data for their components Tokovinin (2018). This makes it possible to describe distributions in a multidimensional parameter space based on real errors

*chapaevaiv@my.msu.ru

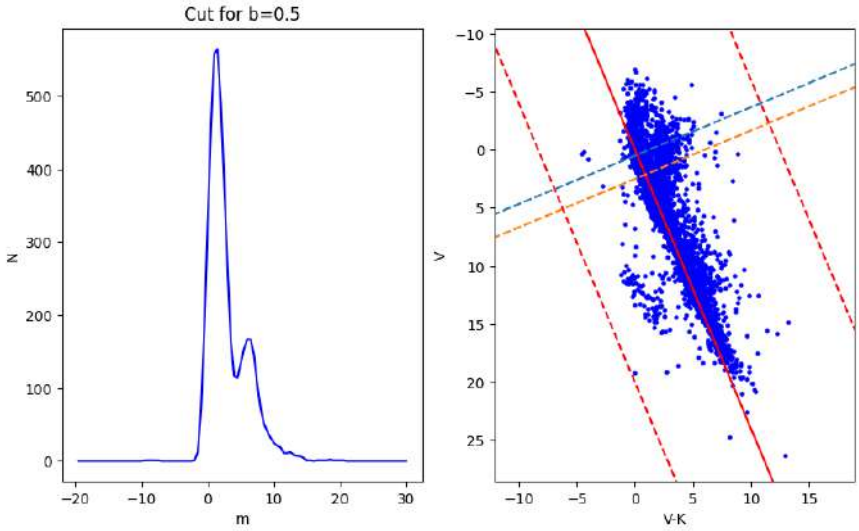

Figure 1. Cross-cut across the main sequence: density profile (left) and position of the cut strip on the colour–magnitude diagram (right).

and selection effects rather than on their model approximation. The colour–magnitude diagram is built from the entire catalogue, using the colour index $V - K$ and the absolute magnitude V. Other catalogues will be used at later stages.

3. Method

A set of local cross-cuts across the main sequence is considered (Fig. 1). The sequence is approximated by a straight line $y = 2.4x$, and the position of a cut is set by the parameter b:

$$V = -\frac{1}{2.4}(V - K) + b \pm 0.25$$

. For each cut, a one-dimensional distribution of star density is built along the direction m perpendicular to the sequence. This makes it possible to trace how the distribution changes along the main sequence and to describe its structure quantitatively. Each distribution is approximated by a sum of three Gaussian components, and for each component, the position μ, the width σ, and the amplitude a are determined.

As a result, a set of parameters is obtained from which the shape of the main sequence can be reconstructed using a random number generator, yielding a set of distributions for training.

4. Results and Discussion

Besides the main sequence, the fits reveal additional components, whose positions on the diagram are shown in Fig. 2:

- $(V - K) \approx 1$–3: a group of points above the main sequence, i.e. giants.
- $(V - K) \approx 5$–8: a small peak above and to the right of the main sequence that makes the central peak asymmetric. These are likely unresolved binaries: their total luminosity is higher and their colour redder than those of a single star of the same luminosity, so they are shifted upward and to the right of the sequence.
- $(V - K) \approx 0$–3, $V \approx 10$–16: a broad weak peak well below the main sequence, corresponding to the white-dwarf region.
- $(V - K) \approx 6.5$–7.5: a small peak forming an asymmetric wing on the opposite side of the sequence, presumably associated with a Population II component.

The most interesting result is given by the amplitude a_1 of the main component as a function of b (Fig. 3). It is clearly inhomogeneous and is well described by a sum of three Gaussians.

doi: https://doi.org/10.52526/25792776-26.73.2-346

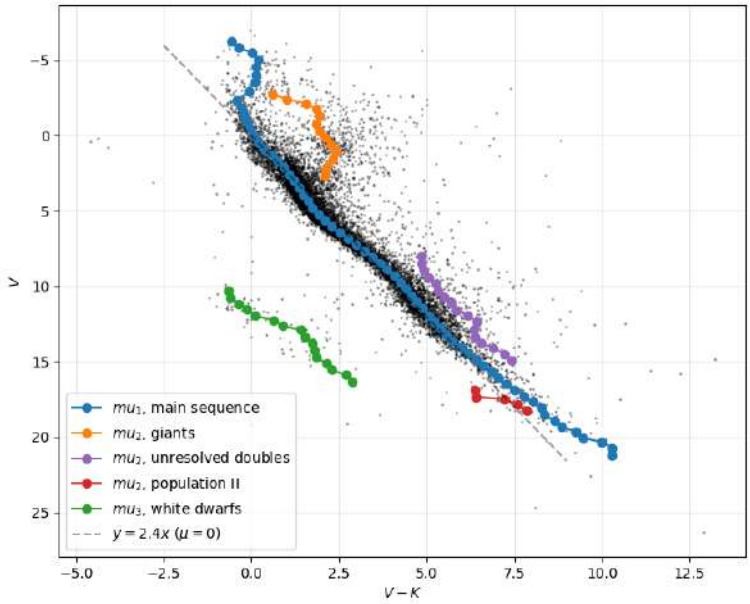

Figure 2. Groups of components on the colour–magnitude diagram (positions μ from fitting the cuts).

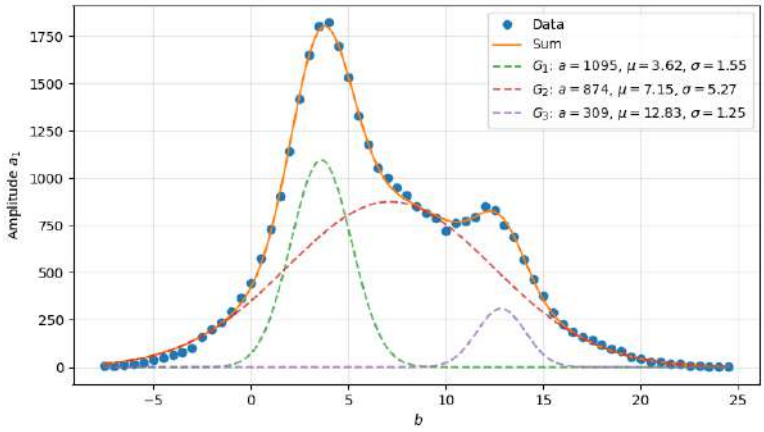

Figure 3. Amplitude of the main component as a function of b.

5. Conclusions and Future Work

Using the colour–magnitude diagram of the MSC as an example, the first stage of obtaining synthetic samples is considered: a quantitative description of the observed distribution. The diagram is divided into local cross-cuts across the main sequence, and the star density distribution in each cut is described by a sum of up to three Gaussians with parameters μ, σ and amplitude. Besides the main sequence, the cuts reveal components interpreted as giants, unresolved binaries, white dwarfs, and Population II . The amplitude of the main component along the sequence is inhomogeneous and is well described by a sum of three Gaussians, i.e. the main sequence has an internal structure. The resulting set of parameters allows to reconstruct the shape of the sequence with a random number generator and to obtain a set of distributions for training. Further work consists of analysing a large number of real distributions to build a large synthetic training sample and of studying the physical causes behind the distributions of multiple stars.

References

Kaygorodov P., Skvortsov N., Kovaleva D., Malkov O., 2023, Open Astronomy, 32, 20220215

Tokovinin A., 2018, The Astrophysical Journal Supplement Series, 235, 6

Tokovinin A., 2024, Multiple Star Catalog, https://www.ctio.noirlab.edu/~atokovin/stars/