Prepared for ApJ
Preprint typeset using LaTeX style AASTeX6
With modifications by David W. Hogg, Daniel Foreman-Mackey, Boris Leistedt

SHRINKING STELLAR DISTANCE UNCERTAINTIES WITH COLOR-MAGNITUDE INFORMATION BUT NO USE OF PHYSICAL STELLAR MODELS

Boris Leistedt[1,2], David W. Hogg[1,3,4] *Add your name here*

[1]Center for Cosmology and Particle Physics, Department of Physics,
New York University, 726 Broadway, New York, NY 10003, USA
[2]NASA Einstein Fellow
[3]Center for Data Science, New York University, 60 Fifth Avenue, New York, NY 10011, USA
[4]Flatiron Institute, 162 Fifth Avenue, New York, NY 10010, USA

ABSTRACT

We present a hierarchical probabilistic model for improving parallax-based stellar distance estimates using color–magnitude information. This is achieved with a data driven model of the color–magnitude diagram, not relying on stellar models but instead on the relative abundances of stars in color–magnitude cells, which are inferred from noisy magnitudes and parallaxes. This is equivalent to deconvolving observational errors and constructing a probabilistic, noiseless color–magnitude diagram. While the latter can be useful for a range of applications, we focus on deriving improved stellar distance estimates relying on both parallax and color–magnitude information. We demonstrate the efficiency of this approach on the 1.4 million stars of the Gaia TGAS sample which have APASS magnitudes. Our hierarchical model has 4 million parameters in total, most of which are marginalized out numerically or analytically. We find that distance estimates are significantly improved for the noisiest parallaxes and densest regions of the color–magnitude diagram. In particular, the average distance signal-to-noise and uncertainty improve by 18% and 37%, respectively, with 9% of the objects having the latter reduced by a factor of 2. This demonstration is a first step towards a hierarchical internal analysis of the Gaia data.

Keywords: Stellar distances, probabilistic hierarchical models.

1. INTRODUCTION

Gaia and LSST will soon provide parallaxes for billions of stars inhabiting our Galaxy. However, parallax measurements for most of these objects are very noisy due to parallax errors significantly increasing with magnitude, which affects the accuracy of the resulting distance estimates. Most analyses adopt signal-to-noise cuts, thus

discarding large amounts of hard-won data. In rare cases, uncertainties are explicitly accounted for, via probabilistic hierarchical models, for example. Regardless on how parallax and distance uncertainties are accounted or ignored, reducing them can improve a variety of studies and increase the scientific return of a mission like Gaia.

In addition to direct determination via parallaxes, stellar distances can be indirectly inferred from spectral information. This is because the vast majority of stars live in compact regions of color–magnitude space, such as the main sequence and the red giant branch. This information can be used as a prior for breaking the degeneracy between distances, apparent and absolute magnitudes. In other words, photometric information can complement parallax information and prove crucial to determine distances. The influence of distance priors as well as color–magnitude priors based on stellar models was explored in previous works (*e.g.*, Astraatmadja & Bailer-Jones 2016b,a), which relied on theoretical models or simulations. Yet, stellar models predict narrow tracks in color–magnitude which are reductive of the diversity of real objects, which may introduce biases when used as priors for determining distances. In this paper, we address this problem and present a purely data-driven approach for improving parallax-based distance estimates with color–magnitude information inferred directly from the data, with no use of external models.

Our approach relies a standard property of hierarchical probabilistic models: a large number of noisy observations (here, parallaxes and magnitudes) can be exploited to construct an estimate of the underlying distribution (a color–magnitude diagram), which is in turn applied as a prior to the individual objects to shrink the uncertainties on their individual properties (the distances). This *Bayesian shrinkage* naturally occurs whithin hierarchical models when computing or sampling the full joint posterior distribution. In other words, *the whole is greater than the sum of the parts*, and stellar distance estimates can be improved by simultaneously analysing all the available data. This is particularly relevant for Gaia, which will soon provide hundreds of millions of sources with parallax, proper motion, and multicolor information, offering the opportunity to construct accurate models of the distribution of magnitudes, colors, dust, and stellar densities in the Galaxy, with minimal external data. Those models will in turn significantly improve our estimates of the intrinsic properties of individual stars. The core goal of this paper is to demonstrate how one aspect of this hierarchical inference can be done in practice. We focus on evaluating the shrinkage of distance estimates obtained by constructing a flexible color–magnitude diagram based on the joint analysis of millions of noisy parallaxes and apparent magnitudes.

We make a number of simplifying assumptions to make our demonstration more transparent, such as adopting uninformative distance priors and point estimates of dust corrections. Those assumptions could be relaxed.

This paper is structured as follows: our model and inference framework are described in Sec. 2, and applied to the Gaia data in Sec. 3. Conclusions and perspectives are presented in Sec. 4.

s	object index (the s-th star)
d_s, ϖ_s, M_s, C_s	true distance, parallax, absolute magnitude, and color
$\hat{\varpi}_s, \sigma^2_{\hat{\varpi}_s}$	parallax estimate and its variance
$\hat{m}_s, \hat{C}_s, \sigma^2_{\hat{m}_s}, \sigma^2_{\hat{C}_s}$	apparent magnitude and color estimates, and their variances
b_s	index of the color–magnitude bin of the sth object
b	generic index of color–magnitude bin
n_b	object count in the b-th color–magnitude bin
$\{n_b\}$	set of all galaxy counts n_b, summing to $N_{\rm stars}$
f_b	fractional galaxy count in the b-th color–magnitude bin
$\{f_b\}$	set of all fractional bin counts f_b, summing to 1
$\{d_s, b_s\}$	distances and bins of all stars in the sample
$\{\hat{m}_s, \hat{C}_s\}$	all magnitude and color estimates

Table 1. Summary of our notation.

2. MODEL

We consider a set of stars indexed as $s = 1, \cdots, N_{\rm stars}$, each characterized by a distance d_s, an absolute magnitude M_s, and a color C_s. The magnitude and color are taken with respect to an arbitrary reference band. We consider only one magnitude and one color for simplicity, but it should be noted that the model and method presented below can be straightforwardly extended to multiple magnitudes and colors.

Intrinsic properties like distance and absolute magnitude are not observable, they must be inferred from apparent magnitudes and parallax measurements. The estimate of the parallax is denoted $\hat{\varpi}_s$ and is assumed to have a Gaussian variance $\sigma^2_{\hat{\varpi}_s}$. We will consider two magnitudes, $\hat{m}_s$ and $\hat{m}'_s$, assumed to be uncorrelated and to have Gaussian variances $\sigma^2_{\hat{m}_s}$ and $\sigma^2_{\hat{m}'_s}$. We will use the first one $\hat{m}_s$ as a reference magnitude for inferring the absolute magnitude M_s, and the second one to form a color estimate $\hat{C}_s = \hat{m}'_s - \hat{m}_s$ with Gaussian variance $\sigma^2_{\hat{C}_s} = \sigma^2_{\hat{m}_s} + \sigma^2_{\hat{m}'_s}$.

We aim at estimating the distance d_s of each star from the noisy data $\hat{m}_s$, $\hat{C}_s$ and $\hat{\varpi}_s$. While distance is directly connected to the parallax via $\varpi_s = 1/d_s$, it is also informed by the apparent magnitude since $m_s = M_s + 5\log_{10} d_s$ where d_s is expressed in units of 10 pc. Note that when only the apparent magnitude is available, distance and absolute magnitude are degenerate and cannot be disentangled. This degeneracy is partially broken with the parallax information. Here, we seek to incorporate the knowledge that stars do not have arbitrary colors and magnitude. The way this information enters distance estimates is made obvious by writing the posterior probability distribution on the distance,

$$p(d_s|\hat{m}_s, \hat{C}_s, \hat{\varpi}_s) = \int \mathrm{d}M_s \; \mathrm{d}C_s \; p\big(\hat{m}_s, \hat{C}_s, \hat{\varpi}_s \big| M_s, d_s, C_s\big) \; p\big(M_s, d_s, C_s\big). \qquad (1)$$

This integral marginalizes over the true absolute magnitude and color. This might be expensive to perform numerically, but the choices we will make below will allow us to execute it analytically.

The first term of Eq. (1) is a likelihood function, and the second term is the prior. Assuming that the magnitude and parallax estimates are independent, the likelihood function factorizes as the product of two terms,

$$p\left(\hat{\varpi}_s \middle| d_s\right) = \mathcal{N}\left(\hat{\varpi}_s - d_s^{-1}; \sigma^2_{\hat{\varpi}_s}\right), \tag{2}$$

and

$$p\left(\hat{m}_s, \hat{C}_s \middle| M_s, d_s, C_s\right) = \mathcal{N}\left(M_s + 5\log_{10} d_s - \hat{m}_s; \sigma^2_{\hat{m}_s}\right) \mathcal{N}\left(\hat{C}_s - C_s; \sigma^2_{\hat{C}_s}\right). \tag{3}$$

The final term, $p(M_s, d_s, C_s)$, is the prior knowledge about the distances, magnitudes, and colors of stars. It is typically factorized as $p(M_s, C_s|d_s)p(d_s)$, where $p(d_s)$ is the distance prior, which could be based on a three-dimensional model of the Galaxy. The color–magnitude prior is typically made distance-independent, $p(M_s, C_s)$, and based on stellar models (*e.g.*, Astraatmadja & Bailer-Jones 2016b,a).

In this work, we will adopt a uniform distance prior and focus on the magnitude–color term, which we parametrize as $p\left(M_s, C_s \middle| \{f_b\}\right)$. We construct a model of the relative abundance of objects in color–magnitude cells (*i.e.* in two dimensions). The color–magnitude distribution is described as a linear mixture of B components,

$$p\left(M_s, C_s \middle| \{f_b\}\right) = \sum_{b=1}^{B} f_b \; K_b(M_s, C_s), \tag{4}$$

with K_b the bth kernel. The parameters $\{f_b\}$ refer to the relative probabilities of finding objects in the various kernels, and must sum to one ($\sum_b f_b = 1$).

For the kernels, we adopt Gaussian distributions to make the integral of Eq. (1) analytically tractable. The b-th kernel will be centered at $(\mu_{b,0}, \mu_{b,1})$ and have a diagonal covariance $(\sigma^2_{b,0}, \sigma^2_{b,1})$. We take $\mu_{b+1,0} - \mu_{b,0} = \sigma_{b,0}$ and $\sigma_{b,0}$ constant (similarly for the color dimension) to uniformly and contiguously tile a rectangular region of interest in color–magnitude space. With this parameterization, the integral of Eq. (1) is tractable and leads to

$$\begin{aligned} p(d_s \,|\hat{m}_s, \hat{C}_s, \hat{\varpi}_s, \{f_b\}) \;\propto\; & \sum_b \; f_b \, \mathcal{N}\left(\hat{\varpi}_s - d_s^{-1}; \sigma^2_{\hat{\varpi}_s}\right) \\ \times \;\; & \mathcal{N}\left(\mu_{b_s,0} + 5\log_{10} d_s - \hat{m}_s; \sigma^2_{\hat{m}_s} + \sigma^2_{b_s,0}\right) \mathcal{N}\left(\hat{C}_s - \mu_{b_s,1}; \sigma^2_{\hat{C}_s} + \sigma^2_{b_s,1}\right). \end{aligned} \tag{5}$$

Finally, to facilitate parameter inference, we will introduce a latent variable b_s denoting the bin the sth object belongs to. Then, we can equivalently write the color–magnitude model as

$$\begin{aligned} p\left(b_s \middle| \{f_b\}\right) \;&=\; f_{b_s} \\ p\left(M_s, C_s \middle| b_s\right) \;&=\; \mathcal{N}\left(M_s - \mu_{b,0}; \sigma^2_{b,0}\right) \mathcal{N}\left(C_s - \mu_{b,1}; \sigma^2_{b,1}\right). \end{aligned} \tag{6}$$

Our notation is summarized in Table 1. Note that in all of the above we have assumed that all magnitudes are properly dereddenned, *i.e.* that the absorption by interstellar dust has been corrected for. As discussed below, we will use the distance point estimate $1/\hat{\varpi}_s$ for performing this correction. While this is formally incorrect (since

the reddenning correction depends on distance which is a parameter we are inferring), this approximation will be sufficient for the case-study of this paper.

2.1. *Inference*

Since we fix the kernel locations $\{(\mu_{b,0}, \mu_{b,1})\}$ and covariances $\{(\sigma^2_{b,0}, \sigma^2_{b,1})\}$, our color–magnitude model is fully described by the relative amplitudes $\{f_b\}$. Those could be set by prior knowledge (*e.g.*, external data or stellar models), in which case one could use Eq. (6) to infer the distance of each object. Here, we seek to infer $\{f_b\}$ from the data. Thus, the full posterior of interest is $p(\{d_s\}, \{f_b\}|\{\hat{m}_s, \hat{C}_s, \hat{\varpi}_s\})$, which has $B + N_{\rm stars}$ parameters.

Given the number of parameters and the natural degeneracies between magnitudes and distances, standard sampling techniques may be difficult to apply. However, the inference is greatly simplified by focusing on $p(\{b_s\}, \{f_b\}|\{\hat{m}_s, \hat{C}_s, \hat{\varpi}_s\})$, where distances are marginalized over. We adopt a Gibbs sampling strategy (see *e.g.*, Casella & George 1992; Wandelt et al. 2004; Levin et al. 2009; Brooks et al. 2011), which consists of alternating draws from the conditional distributions of $\{b_s\}$ and $\{f_b\}$. At the ith iteration, we will draw new values of the $\{f_b\}$ and $\{b_s\}$ parameters given the values of the previous iteration. In other words, we first draw $\{f_b\}^{(i)}$ given $\{b_s\}^{(i-1)}$ (and the data), and for each object then draw $b_s^{(i)}$ given $\{f_b\}^{(i)}$ (and the data). The sequence $\{f_b\}^{(i)}, \{b_s\}^{(i)}$ for $i = 1, \cdots, N_{\rm samples}$ forms a Markov Chain with the target posterior distribution of interest as equilibrium distribution. This allows us to avoid the magnitude–distance degeneracies and exploit other properties of our model, such as conjugate priors. We now detail how to draw from the correct conditional distributions.

The first draw is fairly standard: with the bin locations $\{b_s\}$ fixed, the fractional amplitudes $\{f_b\}$ follow a Dirichlet distribution entirely determined by $\{n_b\}$, with n_b the number of objects in the b-th bin. All the other parameters enter the constant proportionality factor, so the target conditional distribution is

$$p\left(\left\{f_b\right\}\middle|\left\{d_s, b_s, \hat{m}_s, \hat{C}_s, \hat{\varpi}_s\right\}\right) \;=\; p\left(\left\{f_b\right\}\middle|\{n_b\}\right) \;\propto\; \prod_b \frac{f_b^{n_b}}{n_b!}, \tag{7}$$

where we have assumed the simplest uninformative prior for the relative amplitudes, *i.e.* that they are positive and normalized to one, $p(\left\{f_b\right\}) = \delta^D(1 - \sum_b f_b) \prod_b \Theta(f_b)$, where δ^D is the Dirac delta function and Θ the step function. Given the number counts $\{n_b\}$, one can draw $\{f_b\}$ from this distribution using standard tools for Dirichlet distributions. Alternatively, one could draw $\gamma_b \sim \text{Gamma}(n_b + 1)$ and take $f_b = \gamma_b / \sum'_b \gamma'_b$.

The second step of the Gibbs sampler is to draw the bins $\{b_s\}$ given the amplitudes $\{f_b\}$. Those draws can be performed independently (thus, in parallel) over objects. This may not look simple since we have to marginalize over the distances, and the

target conditional distribution to sample is

$$p\left(b_s\middle|\{f_b\},\hat{m}_s,\hat{C}_s,\hat{\varpi}_s\right) \propto \int p\left(b_s,d_s\middle|\{f_b\},\hat{m}_s,\hat{C}_s,\hat{\varpi}_s\right)\mathrm{d}d_s. \tag{8}$$

However, closer examination reveals that the bins can be simply drawn from a multinomial distribution with amplitudes given by

$$p\left(b_s\middle|\{f_b\},\hat{m}_s,\hat{C}_s,\hat{\varpi}_s\right) \propto f_{b_s} g_{b_s} \tag{9}$$

with

$$g_{b_s} = \int \mathcal{N}\left(\mu_{b_s,0} + 5\log_{10} d_s - \hat{m}_s; \sigma^2_{\hat{m}_s} + \sigma^2_{b_s,0}\right) \tag{10}$$

$$\times\, \mathcal{N}\left(\hat{C}_s - \mu_{b_s,1}; \sigma^2_{\hat{C}_s} + \sigma^2_{b_s,1}\right) \mathcal{N}\left(\hat{\varpi}_s - d_s^{-1}; \sigma^2_{\hat{\varpi}_s}\right) \mathrm{d}d_s. \tag{11}$$

Since we keep the kernel locations and covariances fixed, those weights can be precalculated via N_{stars} independent dimensional integrals. We evaluate them numerically given that the integrand is simple and the integration can be tuned easily. If this operation were prohibitive, or if the kernel parameters were not kept fixed, one could sample the distance d_s (jointly or conditionally on the bin b_s). Since the kernels and likelihoods admit simple gradients and Hessians, one could use Hamiltonian Monte Carlo (*e.g.*, Duane et al. 1987; Neal 2012) to efficiently sample all N_{stars} distances simultaneously.

2.2. *Discussion*

We now briefly discuss some of the advantages and limitations of our approach.

First, we restricted our attention to the color–magnitude diagram, and neglected any dependency on other quantities such as galactic latitude. We also adopted uniform distance priors. This can be corrected by adopting or constraining models of the 3D distribution stellar densities, described by kernel mixtures, for example (see *e.g.*, Astraatmadja & Bailer-Jones 2016b,a). Although this extension is technically trivial, we have not developed it since we focus on how color–magnitude information informs distance estimates without the use of stellar models. Similarly, our framework could be extended to other observables such as velocities.

Second, our kernel mixture model offers a significant amount of freedom for describing the color–magnitude diagram, but could be improved in various ways. Changing the kernels would not strongly affect our inference framework, unless they were not differentiable, in which case it would not possible to use Hamiltonian Monte Carlo. Analytic marginalization of true magnitudes and colors was made possible by adopting Gaussian kernels. Furthermore, we did not optimized the positions and covariances of the kernels, unlike in standard Gaussian Mixture models (*e.g.*, Bovy et al. 2011). Compared to those, our tiling of color–magnitude space requires more components (many of which are zero) but is easy to initialize, and also converges quickly.

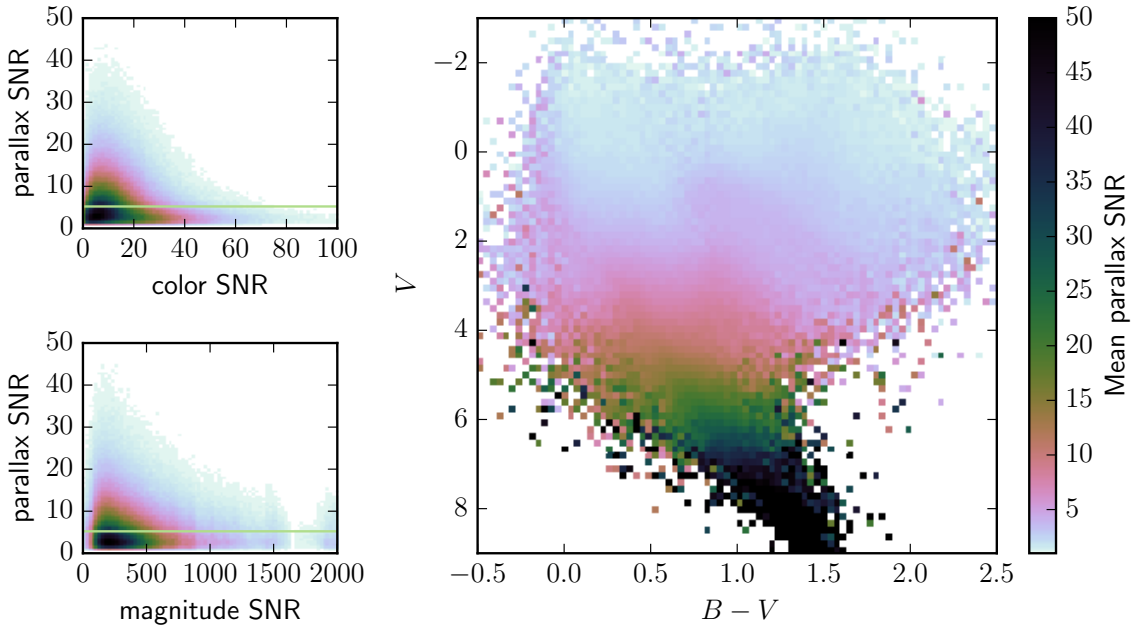

Figure 1. Distributions of the magnitude, color, and parallax signal-to-noise ratios (SNR) of the Gaia TGAS+APASS data we fit and validate our model on. The line indicates the parallax SNR level used to split the data into two sub-samples containing the 'best' and 'worst' parallaxes. The right panel shows the average parallax SNR in color–magnitude cells, illustrating how the upper part of the color–magnitude diagram is dominated by low-SNR objects. Reconstructing the noiseless color–diagram requires an inference framework capable of correctly dealing with uncertainties in colors, magnitudes, and parallaxes.

Third, we assumed that the magnitudes are perfectly dust-corrected. However, dust extinction depends on distance, which is a parameter of our model. Furthermore, reliable 3D dust maps are only available for a limited region of three-dimensional space. In principle, dust corrections should really be inferred jointly with the absolute magnitudes and colors of the data at hand. The approximation we use is sufficient for inferring the color–magnitude diagram and showing the resulting improvement of the distance estimates.

3. APPLICATION TO GAIA TGAS

We consider the Gaia data (Gaia Collaboration et al. 2016), specifically the first data release (DR1) of the Tycho-Gaia astrometric solution (hereafter TGAS, Lindegren et al. 2016). We restrict our attention to the objects with valid B and V magnitudes from the AAVSO Photometric All Sky Survey (APASS) Data Release 9 (Munari et al. 2014; Henden & Munari 2014). We also remove objects with parallax signal-to-noise ratio (SNR) lower than 1. This leads to 1.4 million objects with magnitude, color, and parallax information. We don't apply more stringent parallax or

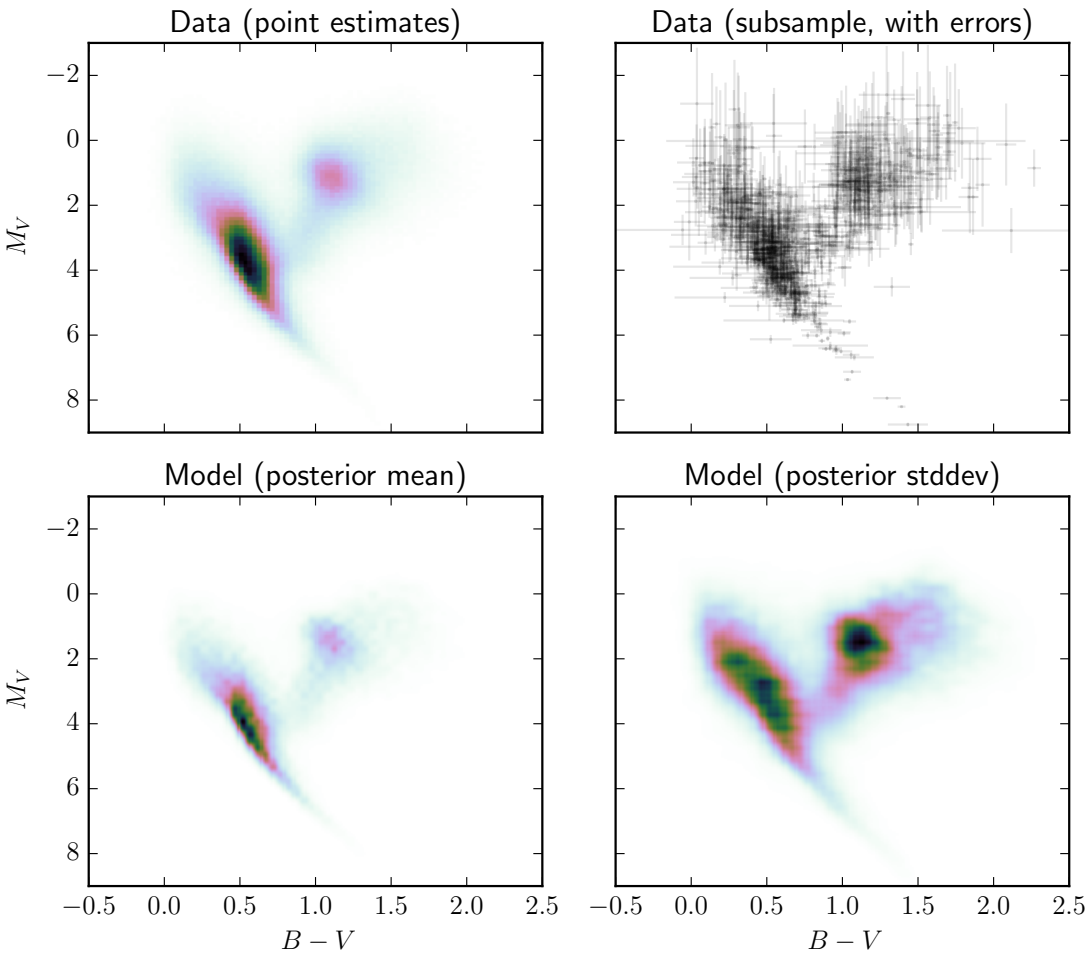

Figure 2. Upper panels: color–magnitude diagram based on the noisy data, obtained with magnitude and parallax point estimates (left) and by sampling parallaxes, magnitude and color based on the measurements and their errors (right). Middle and right: mean and standard deviation of our model, which is the result of deconvolving all observational errors of the data shown in the upper panels and in Fig. 1 into a noiseless color–magnitude diagram described as a mixture of Gaussians tiling the color–magnitude region of interest.

color cuts since the purpose of our method is exactly to construct a color–magnitude model from both low- and high-SNR objects. Finally, we apply dust corrections based on position and distance point estimate ($1/\hat{\varpi}$) with the three-dimensional dust map of Green et al. (2015). For large distances when the latter is undefined we use corrections from the 2D dust map of Schlegel et al. (1998). Our data sample is summarized in Fig. 1, which shows the magnitude, color, and parallax SNR distributions. The bulk of the objects has parallax SNR lower than 10 and is at $M_V < 4$, in the upper part of the color–magnitude diagram. This highlights the need for a correct inference framework that exploits all objects, since focusing on high-SNR objects

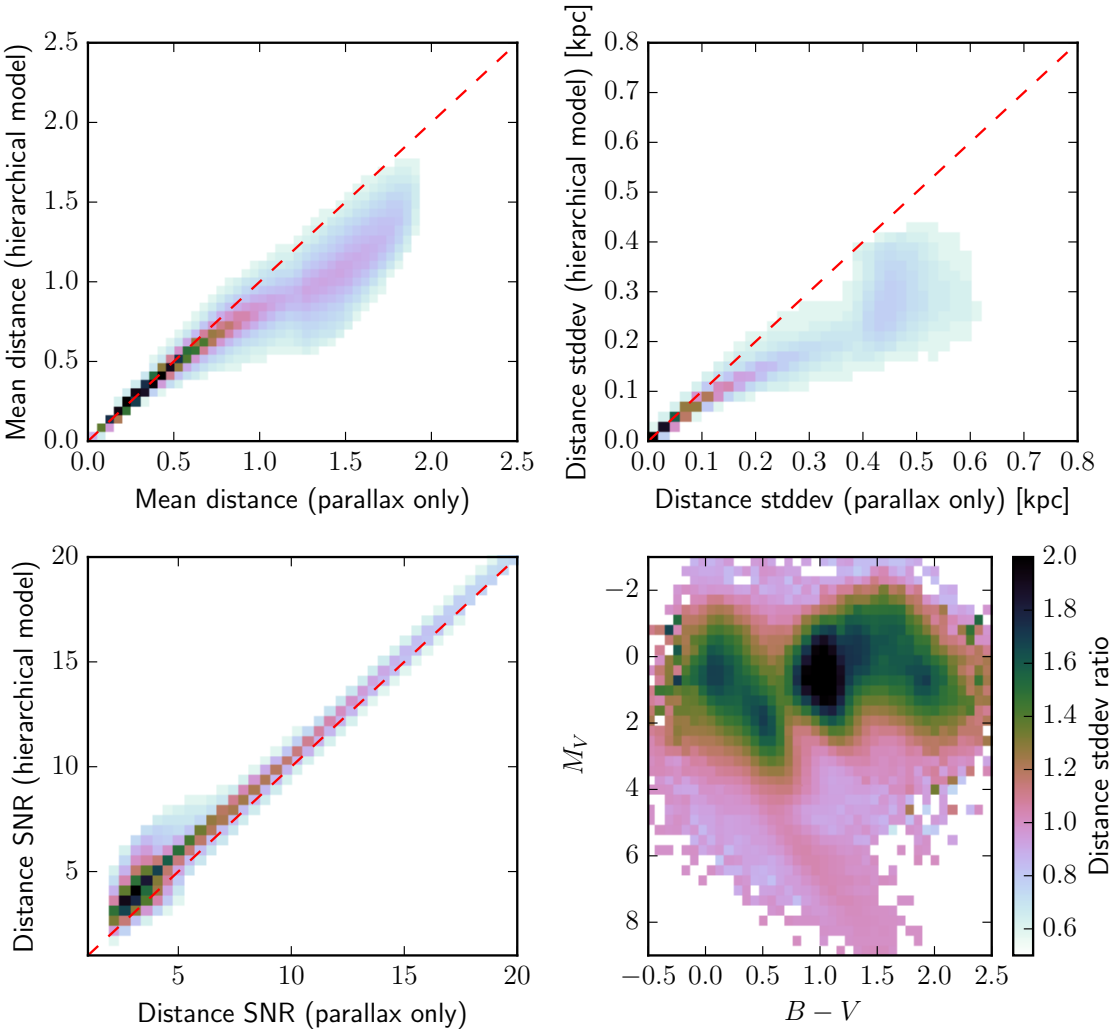

Figure 3. Distances obtained when sampling the hierarchical model which produced the color–diagram shown in Fig. 2. The first three panels shows the change in the mean, standard deviation, and SNR of the distance estimate (based on the posterior distribution), with the number counts in logarithmic scale. The final panel shows the ratio of standard deviations placed in the color–diagram (standard method over hierarchical model). The shrinkage of the uncertainties is a consequence of the hierarchical nature of the model, and is most efficient for low-SNR objects and the densest parts of the color diagram.

would bias the results and prevent us from correctly inferring the fainter regions of the color–magnitude space. This is illustrated below. Note that we did not add the recommended parallax error offset of 0.3 because overestimated errors typically cause excessive deconvolution. For the type of hierarchical inference we are considering, underestimated errors lead to more conservative results.

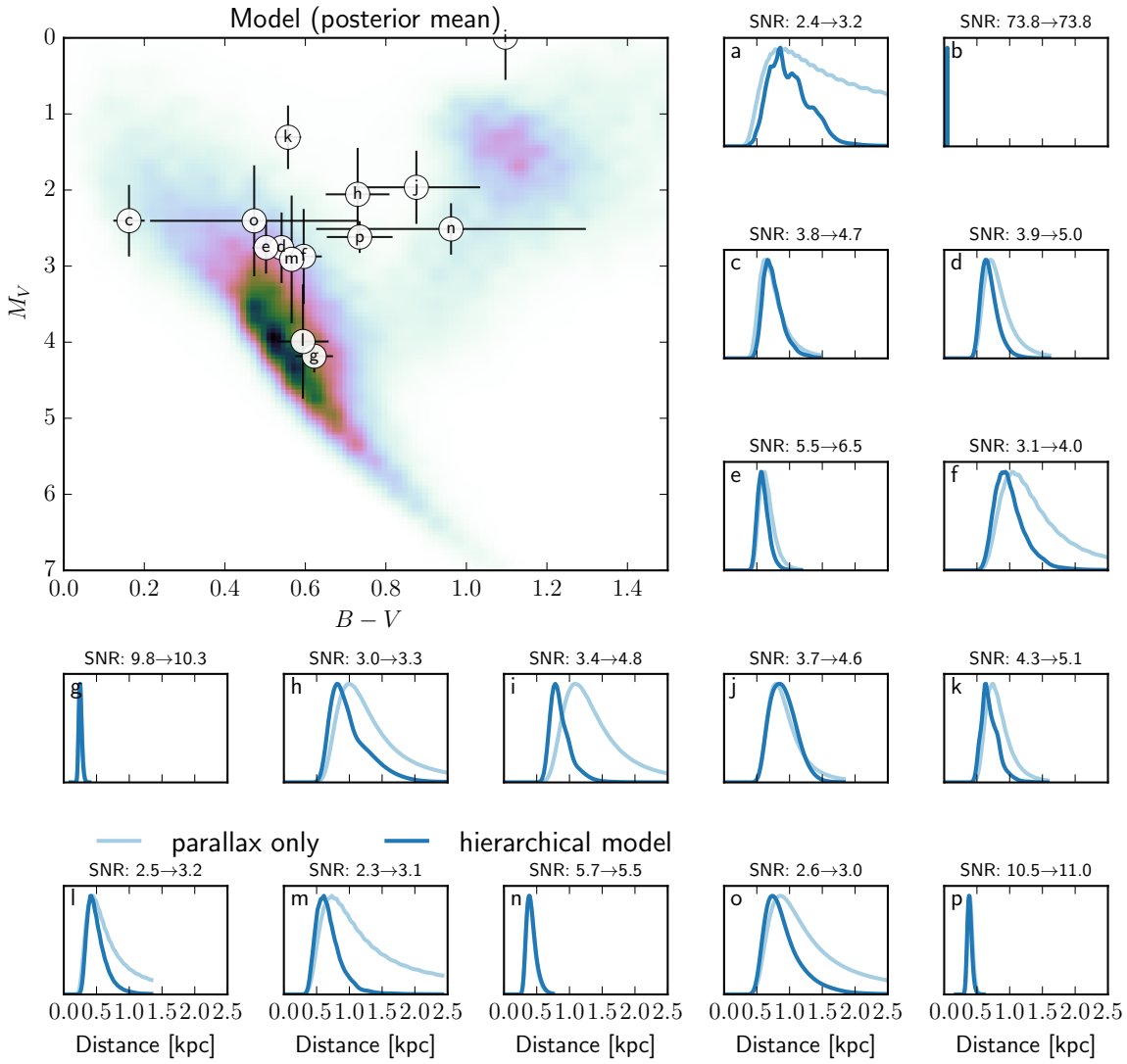

Figure 4. Posterior distributions on the distances of a few objects involved in constraining the model shown in Figs. 2 and 3. The improvement in distance SNR is also shown. The objects are also placed on the inferred color–magnitude diagram, to highlight that the shrinkage is most efficient for low-SNR objects and the densest parts of the color diagram.

We create a validation sample by extracting 10% of the objects, adding significant amount of noise to the parallax estimates, and verifying that our framework improves the distances consistently with the original values. Further details about this process are given below. We also split the main sample according to parallax SNR, into two samples of equal size containing the 'best' and 'worst' parallaxes. We perform the inference on those two samples as well as their union.

For each of the three samples, we use the Gibbs sampler presented above to draw 10,000 samples of the fractional amplitudes and bins. Thus, our model has $B + N_{\rm stars}$ sampled parameters, and $3N_{\rm stars}$ parameters marginalized over (the true magnitudes, colors, and distances), either numerically or analytically. The positions and widths of the kernels are obtained by uniformly tiling (with 60 points) the color and magnitude lines in the ranges $[0, 1.5]$ and $[0, 7]$, respectively.

The mean and standard deviation of the resulting color–magnitude diagram (with bins marginalized over) are presented in Fig. 2. Those were obtained by computing the mean and standard deviation of the set of models obtained by applying Eq. (4) to all the posterior samples we drew. The top panels also show the input data, with and without resampling according to the estimates and their errors. As expected, the recovered models are significantly narrower than the data since we are effectively deconvolving observational errors to produce a noiseless color–magnitude diagram. As expected, classic components are recovered: the main sequence, its turn-over, and the giant branch.

Fig. 3 shows a comparison of the stellar distances with and without the extra information provided by the color–magnitude diagram we constructed. For each object we compute the distance posterior distribution via Eq. (1) on a grid, and we average over samples of the bin allocations and relative amplitudes of the color–magnitude model, which is equivalent to marginalizing them out. Our point of comparison is the set of parallax-only distance estimates, *i.e.* not using our hierarchical model but only the parallax information of Eq. (2). Note that the posterior distributions are not Gaussian, as expected and also shown below. Nevertheless, the mean and standard deviation provides a useful metrics. We measure that on average the distance SNR improves by 18% and the distance uncertainty decreases by 37%. We find that 9% of the objects have their distance uncertainty halved after the inclusion of color–magnitude information via our hierarchical model. This shrinkage of the distance uncertainties is most efficient in the densely populated regions of color–magnitude space. Fig. 4 shows the distance posterior distributions obtained with our method for a few randomly chosen objects, illustrating the uncertainty shrinkage. We also place these objects in the color–magnitude diagram (the errors on the absolute magnitude are obtained by resampling).

Fig. 5 shows the mean and standard deviation of the color–magnitude diagram resulting from performing the inference on the subsamples with parallax SNR cuts (*i.e.* splitting our main sample at parallax SNR of 5). Those demonstrate that including the noisiest objects is essential for correctly inferring the faint regions of magnitude space. The main sequence is well recovered with the high-SNR objects, while the red giant branch is barely detected. By contrast, it is well recovered with the low-SNR objects, but the main sequence is then partially erased. This is a natural consequence of the SNR increasing with absolute magnitude. This highlights the importance of a correct probabilistic framework, capable of correctly exploiting data with heterogeneous noise to reconstruct the noiseless color–magnitude diagram. To

illustrate the effect of the SNR cut we also resampled the data and kept objects that satisfied the cuts. This provides a first-order estimate of the detection probability or selection function of those two subsamples, shown in the right panels of Fig. 5, illustrating what regions of the color–magnitude diagram are unexplored.

We now turn to the validation sample. Since we do not know the true distances for most objects in our data set, we add noise to the parallax estimate, at a level equal to ten times the parallax error. We then compute the posterior distribution on the distance (on a distance grid) with and without color–magnitude information, as in Fig. 4. We simply use the mean model shown in Fig. 2 as a color-magnitude prior. The results are shown in Fig. 6. Given that those objects have significant amounts of noise, causing the distance posterior distribution to be highly non-Gaussian, the mean distance overestimates the true distance (the original parallax-based estimate). The hierarchical model significantly decreases this effect, *i.e.* improves the distance estimates both in terms of mean and uncertainty, demonstrating the validity of our inference scheme.

Finally, we perform an additional test of our method on open clusters. We retrieve the coordinates, distances, and proper motions, of known open clusters from the WEBDA database[1]. From our TGAS-APASS sample, we select cluster member candidates for each cluster in a radius corresponding to 3 pc. Fig. 6 shows the result of applying our framework to those objects for a few selected clusters, *i.e.* using the color–diagram inferred above to inform the distance estimates. Even though not all objects selected by this procedure are true cluster members, most distance estimates undergo a visible improvement toward the cluster distance, with the SNRs improved by a few points.

4. CONCLUSION

Stellar distances are ubiquitous in astronomy and are most directly estimated with parallaxes. However, parallax measurements are unreliable for the faintest, most distant objects, which inhabit most of the Galaxy. We presented a framework for obtaining improved distance estimates with both parallax and color–magnitude information without external data or priors such as stellar models. This exploits a well-known property of probabilistic hierarchical models: by utilizing the wealth of information concealed in millions of data points (the noisy and less noisy ones), one can model the underlying distributions and in turn improve the quality of the individual parameter estimates. We presented a version of this idea where the color–magnitude distribution of stars is directly inferred from noisy parallax and magnitude measurements via an efficient parameter inference scheme. We applied this methodology to Gaia+APASS data and demonstrated that it leads to significant improvements in the distance estimates, specifically for the faintest, most distant objects.

[1] www.univie.ac.at/webda

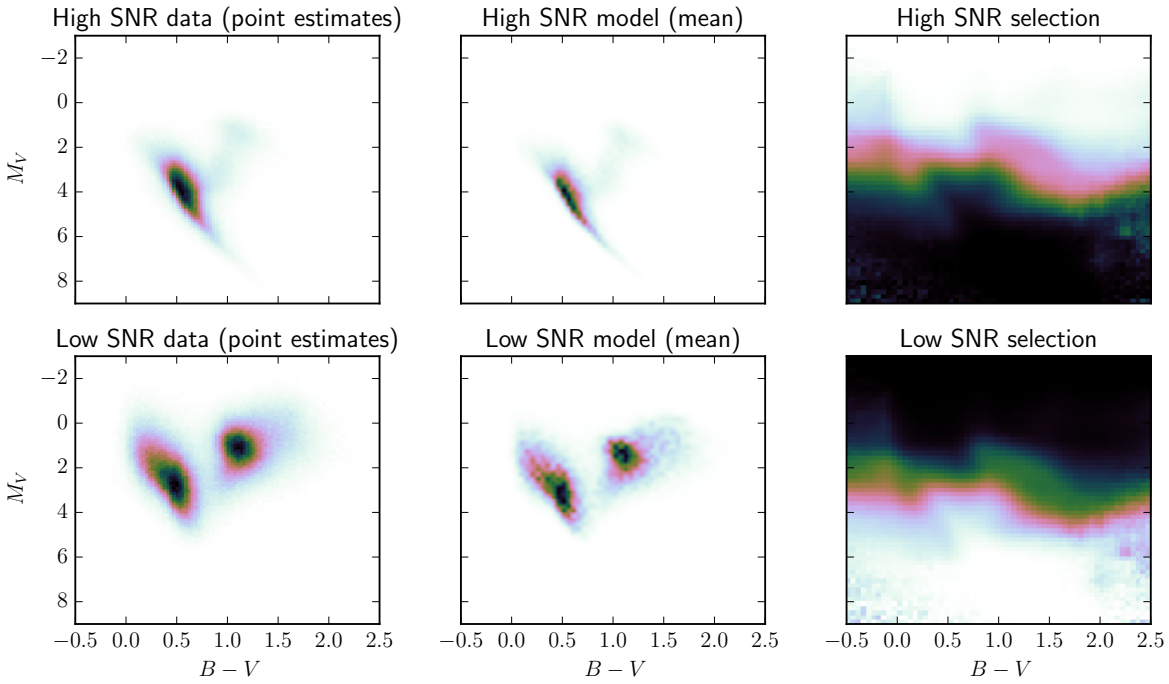

Figure 5. Same as Fig. 2 for with the main sample split based on parallax SNR. This highlights the contributions of the stars with the 'best' and 'worst' parallaxes to the color–magnitude diagram, and the importance of using a correct scheme for inferring the latter in the presence of significant observational errors. The right panels show the probability of detecting sources in both subsamples (approximated by resampling the data), which demonstrates how one misses a significant fraction of color–magnitude space by applying parallax SNR cuts. It is by using all of the objects, not only high-SNR ones, that a satisfactory color–magnitude diagram can be inferred.

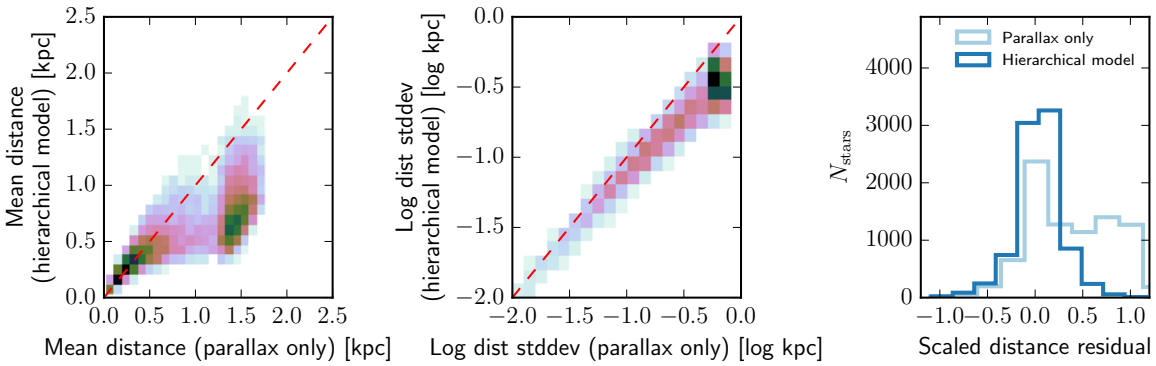

Figure 6. Mean, standard deviation, and scaled residuals (truth - mean estimate, divided by standard deviation) of the distances in our validation sample, based on the posterior distributions. Given the more significant levels of noise the distance are more significantly improved than in our main sample. The mean residuals are not zero due to the non-Gaussianity of the posterior distributions.

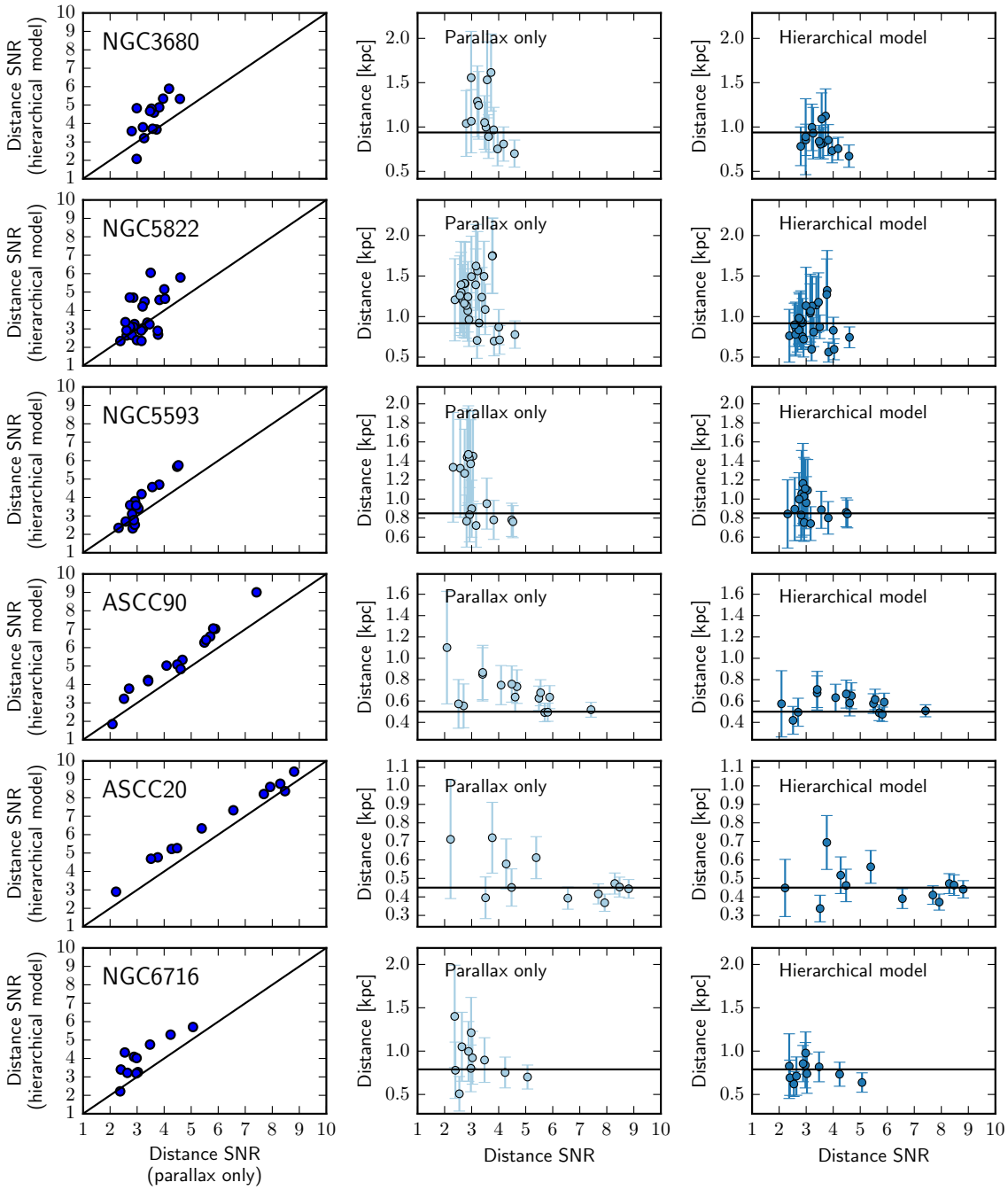

Figure 7. Distances estimates of candidate members of a few open clusters in our data set.

The framework described here includes a number of simplifying assumptions. First, we used uniform distance priors and described the color–magnitude diagram as a mixture of Gaussian kernels with fixed positions and widths. Those two assumptions could easily be relaxed without affecting our inference methodology or the quantitative results presented here. Second, we ignored selection effects that may distort the color–magnitude diagram and the distance distribution. Those could be incorpo-

rated but require external data and modifications to the sampling method. Third, we neglected the dependency of the color–magnitude on other properties (*e.g.*, galactic latitude), and we performed the magnitude reddenning corrections using existing three-dimensional dust maps evaluated at the parallax-based distance point estimates. Our framework could also support those extensions with moderate technical changes. However, in order to fully take advantage of the available data, one should jointly *infer* the three-dimensional distribution of dust and stellar density, and compare with results based on physical models and priors. This requires more substantial technical developments due to the large number of interlaced parameters. We intend to explore this avenue in future work.

We thank Lauren Anderson, Jo Bovy, Adrian Price-Whelan, Semyeong Ho, and Keith Hawkins, for stimulating discussions and useful comments on this manuscript.

BL was supported by NASA through the Einstein Postdoctoral Fellowship (award number PF6-170154). DWH was partially supported by the NSF (AST-1517237) and the Moore–Sloan Data Science Environment at NYU.

This project was developed in part at the 2016 NYC Gaia Sprint, hosted by the Center for Computational Astrophysics at the Simons Foundation in New York City.

This work has made use of data from the European Space Agency (ESA) mission Gaia (http://www.cosmos.esa.int/gaia), processed by the Gaia Data Processing and Analysis Consortium (DPAC, http://www.cosmos.esa.int/web/gaia/dpac/consortium). Funding for the DPAC has been provided by national institutions, in particular the institutions participating in the Gaia Multilateral Agreement.

This research was made possible through the use of the AAVSO Photometric All-Sky Survey (APASS), funded by the Robert Martin Ayers Sciences Fund.

REFERENCES

Astraatmadja, T. L., & Bailer-Jones, C. A. L. 2016a, ApJ, 832, 137
—. 2016b, ApJ, 833, 119
Bovy, J., Hogg, D. W., & Roweis, S. T. 2011, Annals of Applied Statistics, 5, arXiv:0905.2979
Brooks, S., Gelman, A., Jones, G. L., & Meng, X.-L. 2011, Handbook of markov chain monte carlo, Chapman and Hall/CRC
Casella, G., & George, E. I. 1992, The American Statistician, 46, 167
Duane, S., Kennedy, A. D., Pendleton, B. J., & Roweth, D. 1987, Physics Letters B, 195, 216
Gaia Collaboration, Prusti, T., de Bruijne, J. H. J., et al. 2016, A&A, 595, A1
Green, G. M., Schlafly, E. F., Finkbeiner, D. P., et al. 2015, ApJ, 810, 25
Henden, A., & Munari, U. 2014, Contributions of the Astronomical Observatory Skalnate Pleso, 43, 518
Levin, D. A., Peres, Y., & Wilmer, E. L. 2009, Markov chains and mixing times (Providence, R.I. American Mathematical Society), with a chapter on coupling from the past by James G. Propp and David B. Wilson.
Lindegren, L., Lammers, U., Bastian, U., et al. 2016, A&A, 595, A4
Munari, U., Henden, A., Frigo, A., et al. 2014, AJ, 148, 81
Neal, R. M. 2012, ArXiv e-prints, arXiv:1206.1901
Schlegel, D. J., Finkbeiner, D. P., & Davis, M. 1998, ApJ, 500, 525
Wandelt, B. D., Larson, D. L., & Lakshminarayanan, A. 2004, Phys. Rev. D, 70, 083511