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 distances 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. The latter are inferred from noisy magnitudes and parallaxes using an efficient sampling method. This approach is equivalent to deconvolving observational errors into a probabilistic, noiseless color–magnitude diagram, which can be useful for a range of applications. We focus on leveraging color–magnitude information to provide more accurate stellar distance estimates. We demonstrate the power of this approach on the Gaia TGAS stars with APASS magnitudes. We find that distance estimates are significantly improved for the noisiest parallaxes and regions of the color–magnitude diagram. In particular, the average distance signal-to-noise and uncertainty improve by 26% and 40%, respectively, with 12% of the objects having the latter reduced by a factor of 2. We make our improved distance estimates publicly available.

Keywords: Stellar distances, parallaxes, hierarchical models.

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.

1. INTRODUCTION

TODO

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 arbitrarily chosen reference band. We only consider 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.

Those intrinsic properties are not directly observable. Instead, all we have at our disposal is a set of apparent magnitude 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 only, $\hat{m}_s$ and $\hat{m}'_s$, assumed to be uncorrelated and 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 infer 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 assumed that all magnitudes are properly dereddenned, i.e. that the absorption by interstellar dust has been corrected for.

In this work, 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,

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

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 perform 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,

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

and

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

The final term, $p\big(M_s, d_s, C_s\big)$, is the prior knowledge about the distances, magnitudes, and colors of stars. [TODO: Cite literature and discuss how this is usually handled.]

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

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

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

While the kernels can be arbitrarily chosen, 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 of the color–magnitude space. With this parameterization, the integral of Eq. (1) is tractable and leads to

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

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

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

Our notation is summarized in Table 1.

2.1. *Inference*

Assuming that the kernel locations $\{(\mu_{b,0}, \mu_{b,1})\}$ and covariances $\{(\sigma^2_{b,0}, \sigma^2_{b,1})\}$ are fixed, our color–magnitude model is fully described by the relative probabilities $\{f_b\}$. If they are fixed by prior knowledge (e.g. , external data or stellar models), then one can use Eq. (6) to infer the distance of each object using both parallax and color–magnitude information. Here, we seek to infer $\{f_b\}$ too. Thus, the full posterior of interest is $p(\{d_s\}, \{f_b\} | \{\hat{m}_s, \hat{C}_s, \hat{\varpi}_s\})$, which has $B + N_\mathrm{stars}$ parameters.

Given the number of parameters and the natural degeneracies between magnitudes and distances, standard sampling techniques may be difficult to apply. Thus, we develop a Gibbs sampling strategy, to draw samples from $p(\{f_b\}, \{d_s, b_s\} | \{\hat{m}_s, \hat{C}_s, \hat{\varpi}_s\})$, including the bins $\{b_s\}$ since it will simplify the inference. At the ith iteration, we will draw new values of the $\{f_b\}$ and $\{d_s, b_s\}$ parameters given the values of the previous iteration, in the following order (the conditional distribution will be made explicit below): First, draw $\{f_b\}^{(i)}$ given $\{d_s, b_s\}^{(i-1)}$ and the data. Second, for each object draw $b_s^{(i)}$ given $\{f_b\}^{(i)}$ and $\{d_s\}^{(i-1)}$ and the data. Finally, for each object draw $d_s^{(i)}$ given $\{f_b\}^{(i)}$ and $\{b_s\}^{(i)}$ and the data. The sequence $\{f_b\}^{(i)}, \{d_s, b_s\}^{(i)}$ for $i = 1, \cdots, N_\mathrm{samples}$ forms a Markov Chain with the target posterior distribution of interest as equilibrium distribution. This allows us to avoid the magnitude–distance degeneracies and parallelize the second and third steps over objects. We now detail how to draw from the correct conditional distributions.

[TODO: Dirichlet-multinomial] The first draw is fairly standard: with the bin locations $\{b_s\}$ fixed, the fractional weights $\{f_b\}$ 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 distribution is

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

which can be sampled from using standard techniques for Dirichlet draws. This first step of the Gibbs sampler is the only one involving all objects; the two subsequent steps can be performed independently over objects (i.e. in parallel).

Drawing the bins b_s is also simple, since those are discrete and with the other parameters kept fixed they follow a multinomial distribution with fractional weights given by

$$\begin{aligned} p\left(b_s \middle| \{f_b\}, d_s, \hat{m}_s, \hat{C}_s, \hat{\varpi}_s\right) \;\propto\; & f_{b_s}\, \mathcal{N}\big(\mu_{b_s,0} + 5\log_{10} d_s - \hat{m}_s; \sigma^2_{\hat{m}_s} + \sigma^2_{b_s,0}\big) \\ & \times\; \mathcal{N}\big(\hat{C}_s - \mu_{b_s,1}; \sigma^2_{\hat{C}_s} + \sigma^2_{b_s,1}\big). \end{aligned} \tag{8}$$

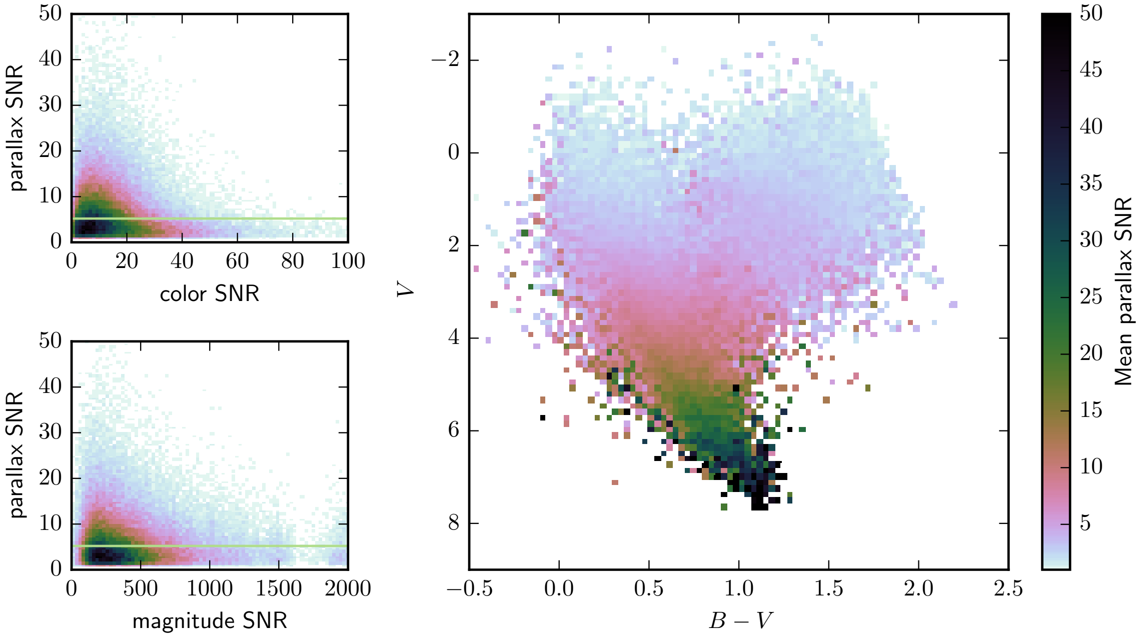

Figure 1. Distributions of the magnitude, color, and parallax signal-to-noise ratios (SNR) of the Gaia TGAS+APASS data we train and validate our model on. The line indicates the parallax SNR level used to split the data into two sub-samples containing the 'best' and 'worst' parallaxes. The right panel 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.

The final step is more complex since the target probability of d_s given the other parameters,

$$p\left(d_s \middle| \{f_b\}, b_s, \hat{m}_s, \hat{C}_s, \hat{\varpi}_s\right) \propto \mathcal{N}\left(\hat{\varpi}_s - d_s^{-1}; \sigma^2_{\hat{\varpi}_s}\right) \\ \times \mathcal{N}\left(\mu_{b,0} + 5\log_{10} d_s - \hat{m}_s; \sigma^2_{\hat{m}_s} + \sigma^2_{b,0}\right), \tag{9}$$

does not follow an analytic law that allows direct sampling. However, it is simple enough that it could be gridded and sampled using an inverse transform method, for example. Yet, we adopt an even more direct method: given that this expression admits trivial gradients and is visibly unimodal, we can use Hamiltonian Monte Carlo, and sample from Eq. (10). We dynamically adjust the stepsize to optimize the exploration of this distribution: we use 10 steps, with step size $\epsilon = 0.1 \times \sigma_{\hat{\varpi}_s}/\hat{\varpi}_s^2$, clipped so that $-5 < \log_{10}\epsilon < -2$.

2.2. *Discussion*

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

First, we note that we could also infer the distance distribution with this formalism, by adding another kernel mixture and inferring its parameters, for example. Although it is technically trivial to add this layer to our framework, we have not developed it since we focus on how color–magnitude information informs distance estimates. For the same reason, we have adopted uniform distance priors. Similarly, our framework can be extended to other observables such as proper motions and velocities.

Second, our kernel mixture model offers a significant amount of freedom to describe the color–magnitude diagram. In fact, changing the kernels does not affect our inference framework if they are differentiable (for the gradients to exist for the Hamiltonian Monte Carlo draw) and can be integrated with Gaussian likelihood functions (for the analytic marginalization of true magnitude and color). Note that we have not optimized the positions and sizes of the kernels. Compared to a standard Gaussian Mixture model, our tiling of color–magnitude space requires more components (many of which are zero) but is easy to initialize, and also converges quickly.

Third, we have assumed that the magnitudes are dust-corrected. However, dust extinction depends on distance, which is a parameter of our model. Furthermore, reliable 3D dust maps are only available and reliable for a limited region of space. Thus, 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 improving distance estimates. Jointly modelling the dust might improve the accuracy of this process and the uncertainty shrinkage.

3. APPLICATION TO GAIA

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 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 **?**. For large distances when the latter is undefined we use corrections from the 2D dust map of **?**. 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 would bias the results and prevent us from correctly inferring the fainter regions of the color–magnitude space.

We create a validation sample by randomly extracting 10% of the objects. As detailed below, we will add significant amount of noise to the parallax estimates and verify that our framework improves the distances consistently with the original values.

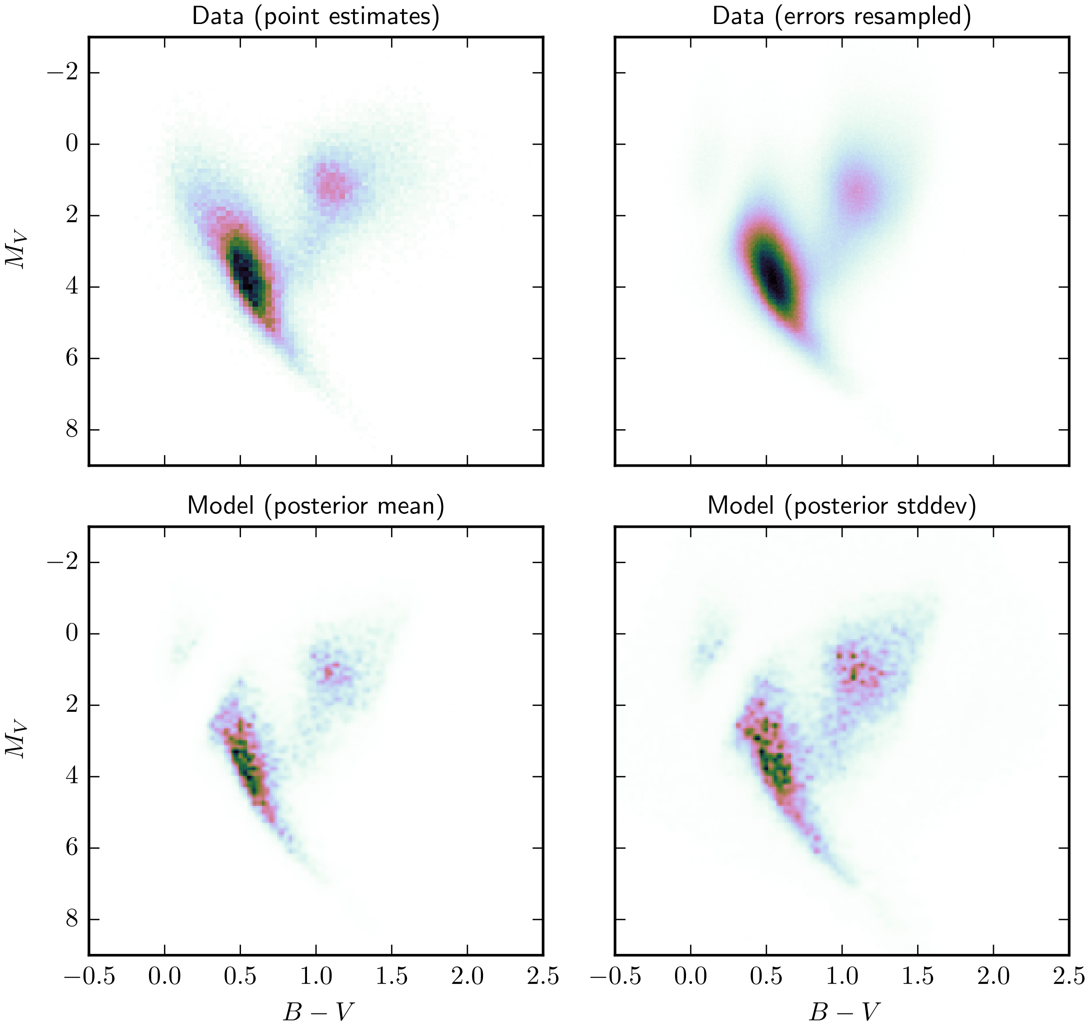

Figure 2. Upper panels: color–diagram based on the noisy data, obtained with magnitude and parallax point estimates (left) and by sampling parallaxes, magnitude and color based on the measurements and their errors (right). Middle and right: mean and standard deviation of our model, which is the result 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.

We also split the main sample according to parallax SNR, into two samples of equal size containing the 'best' and a 'worst' parallaxes. We perform the inference on those two samples as well as the combined one. For each of the three samples, we use the Gibbs sampler presented above to draw 10,000 samples of the fractional bin weights, bins, and distances.

The mean and standard deviation of the resulting color–magnitude diagram (with bins and distances marginalized) are presented in Fig. 2. 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–

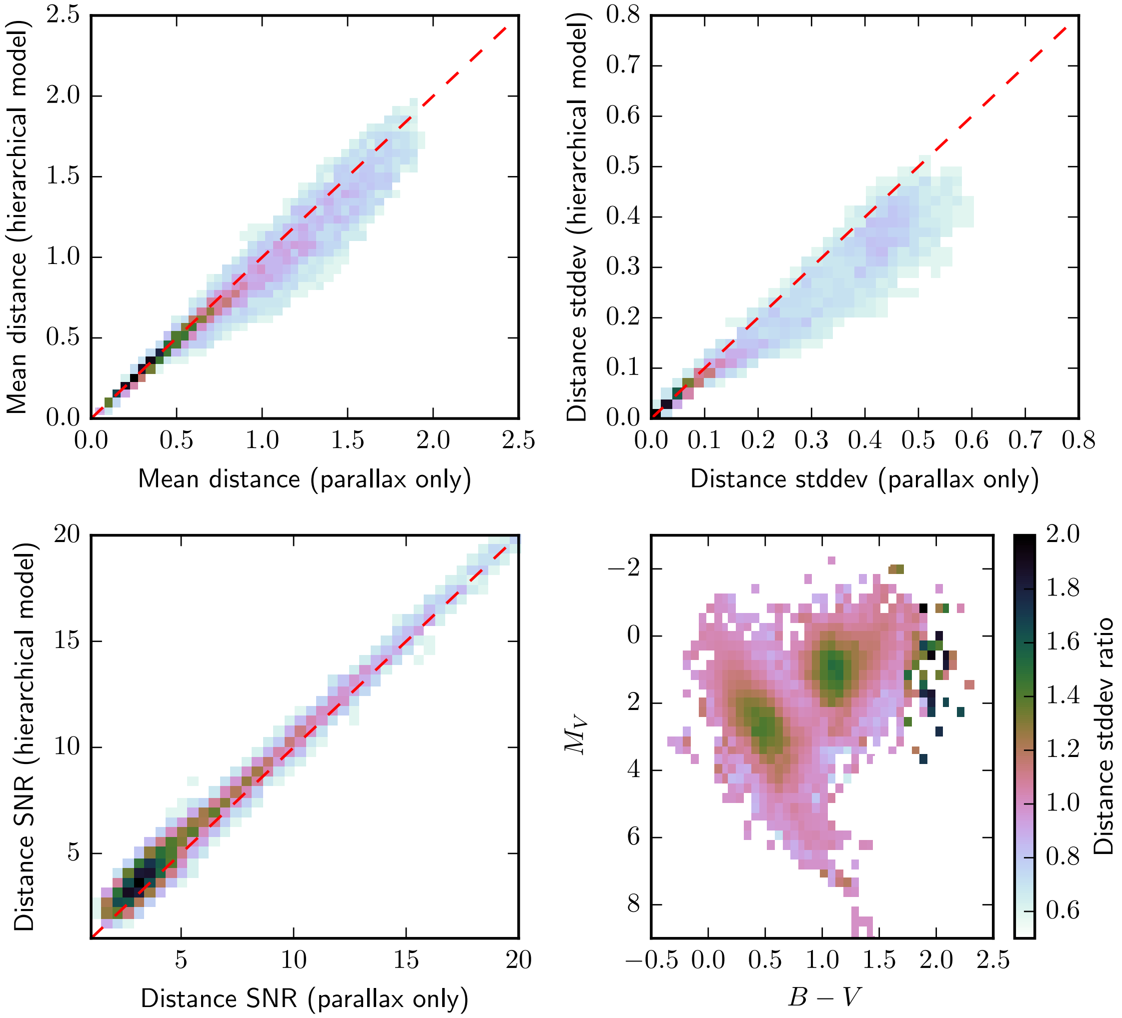

Figure 3. Distances obtained when sampling the hierarchical model which produced the color–diagram shown in Fig. 2. The first three panels shows the change in the mean, standard deviation, and SNR of the distance estimate (based on the posterior distribution), with the number counts in logarithmic scale.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

magnitude diagram. Most classical features are recovered: the main sequence, its turn-over, and the giant branch.

Fig. 3 shows the stellar distances with bins and color–magnitude model marginalized over. We compute the mean and standard deviation using samples of the joint posterior distribution. We also compute mean and standard deviation using samples of the parallax likelihood, 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 standard deviation provides a useful metric. We measure that on average the distance SNR improves by 26% and the

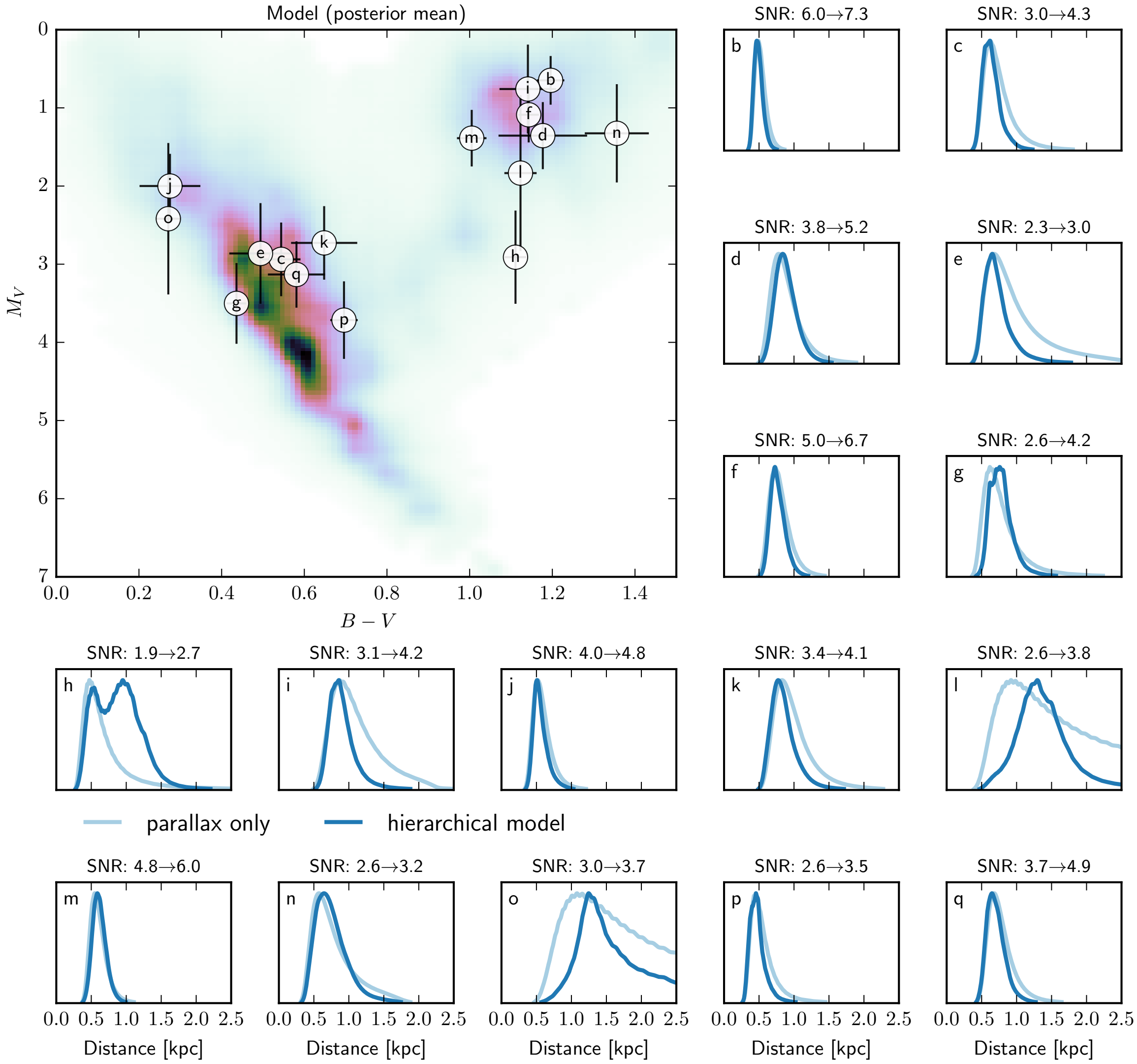

Figure 4. Posterior distributions on the distances of a few objects involved in constraining the model shown in Figs. 2 and 3. The improvement in distance SNR is also shown. The objects are also placed on the inferred color–magnitude diagram, to highlight that the shrinkage is most efficient for low-SNR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

distance uncertainty decreases by 40%. We find that 12% 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 most densely populated regions of color–magnitude space. Fig. 3 shows the distance posterior distributions obtained with our method for a few randomly chosen object, further illustrating the shrinkage of the uncertainties. For clarity we have smoothed the distance samples obtained with the Gibbs sampler. We also place these objects in the color–magnitude diagram.

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 8). Those demonstrate that including the noisiest objects is essential for correctly inferring the fainter 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.

We now turn to the validation sample. Since we do not know the true distances for those objects, we take a different approach: 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) using the parallax likelihood as well as the distance posterior. 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 the stars near each cluster, in a radius corresponding to 3 pc. We then use a Gaussian Mixture model to describe the distribution of proper motions and parallaxes of those objects (using the point estimates and ignoring the errors). We identify the component nearest to the open cluster proper motion and parallax. The stars associated with this component are visually inspected and confirmed as cluster members. Fig. 6 shows the result of applying our framework to those objects, i.e. using the color–diagram inferred above to inform the distance posterior distribution. The distances SNRs are improved and the point estimates change towards the open cluster distance, even though the improvement is relatively modest due to the proximity of those clusters and the good parallax SNR of the cluster members we identified.

4. CONCLUSION

TODO

[TODO: Any acknowledgements missing?]

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

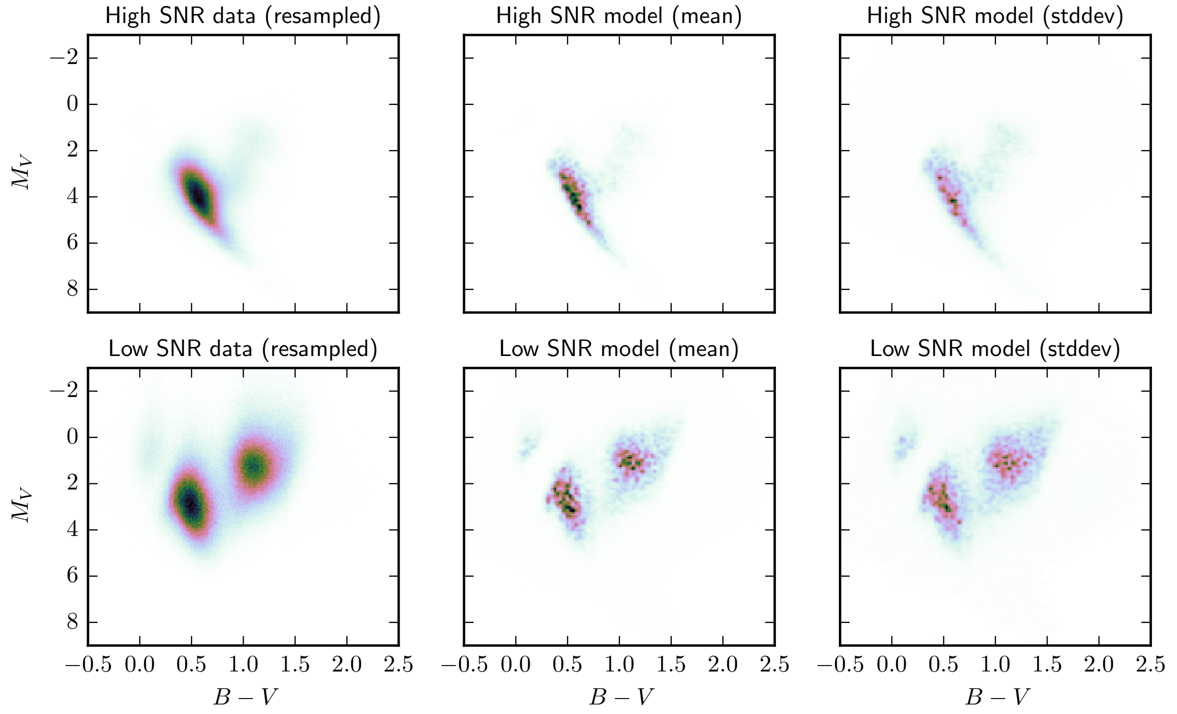

Figure 5. Same as Fig. 2 for with the main sample split based on parallax SNR. This highlights the contributions of the stars with the 'best' and 'worst' parallaxes to the color–magnitude diagram, and the importance of using a correct scheme for inferring the latter in the presence of significant observational errors.

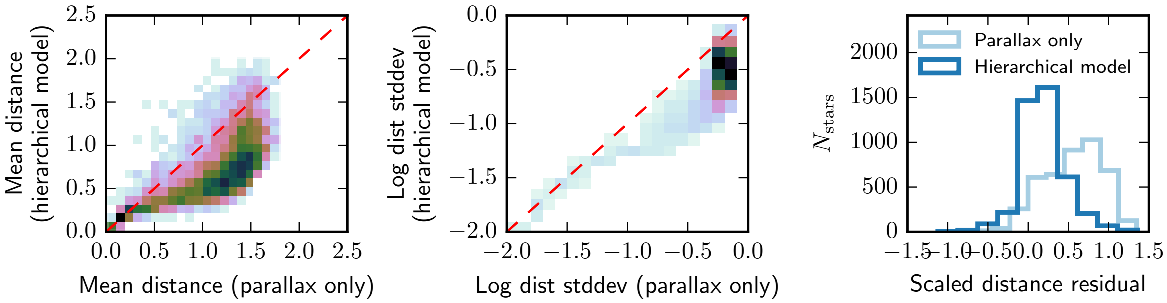

Figure 6. Mean, standard deviation, and scaled residuals (truth - mean estimate, divided by standard deviation) of the distances in our validation sample, based on the posterior distributions. Given the more significant levels of noise the distance are more significantly improved than in our main sample.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

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.

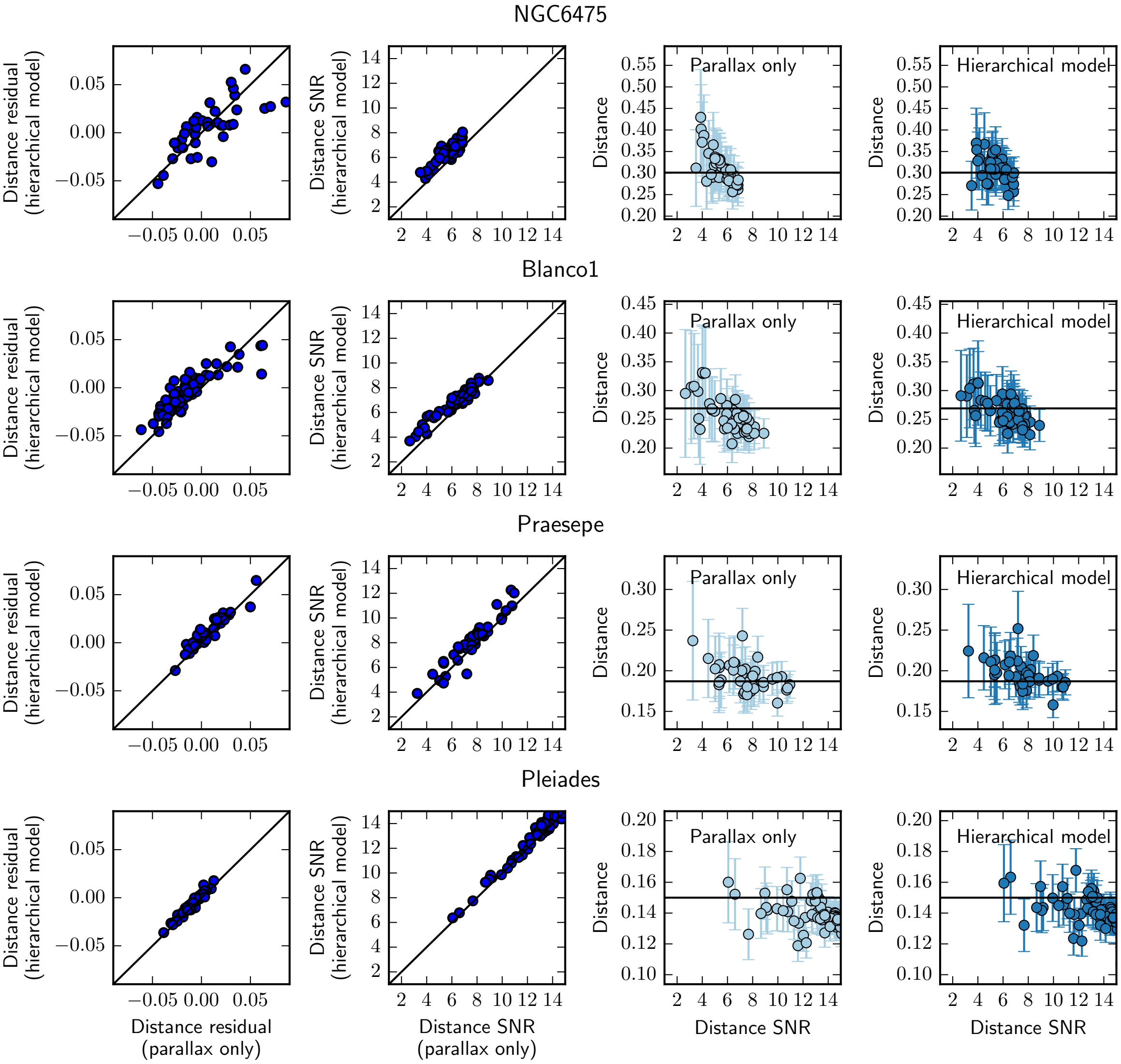

Figure 7. Distances estimates of the members of a few open clusters in our data set. The members firmly identified based on position, proper motion and parallax point estimates.

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

Gaia Collaboration, Prusti, T., de Bruijne, J. H. J., et al. 2016, A&A, 595, A1

Henden, A., & Munari, U. 2014, Contributions of the Astronomical Observatory Skalnate Pleso, 43, 518

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