A simple linear model to aid in analyses of the β Pictoris moving group

Valmir C. Barbosa and Josina O. do Nascimento

1Programa de Engenharia de Sistemas e Computação, COPPE, Universidade Federal do Rio de Janeiro, Rio de Janeiro 21941-914, Brazil
2Observatório Nacional, Rua General José Cristino 77, Rio de Janeiro 20921-400, Brazil

Accepted 2022 November 17. Received 2022 November 15; in original form 2022 June 3

ABSTRACT
We build a 4D linear model of object membership in the β Pictoris moving group (BPMG), using two nested applications of Principal Component Analysis (PCA) for ~1.5 million objects with high-quality data. These data consist of 3D-Galactic space velocities and Gaia G magnitudes. Through PCA, they ultimately result in a 4D straight line, referred to as PC 1′, about which both the bona fide members used to obtain the straight line and the candidate members used to test the model congregate at generally small distances. As a proof of concept, we select bona fide members from a recent, Gaia DR2-based compilation and candidate members mostly from a (pre-Gaia DR2) compilation from 2017. Using a standard procedure to flag groups of outliers in data sets, we find possible outliers flagged on account of their large distances to PC 1′, and find evidence that discrepant radial velocity measurements may indeed bring into question their validity as BPMG members. We propose that PC 1′ be added to the tool set for BPMG analyses and potentially extended to other young stellar moving groups.

Key words: methods: data analysis – open clusters and associations: individual: β Pictoris moving group.

1 INTRODUCTION
Of all the young stellar moving groups discovered and studied to date, the β Pictoris moving group (BPMG; Zuckerman et al. 2001) seems to offer the most tantalizing possibilities for research on planet formation and exoplanets. Not only is it closest to Earth, but also the star that gives BPMG its name has two confirmed planets orbiting it (Lagrange et al. 2010, 2019). These characteristics have motivated substantial effort in the last two to three decades to discover the group’s full suite of members. To the best of our knowledge, the most complete assessment of BPMG membership made before the crucial availability of Gaia DR2 (see, e.g. the discussion in Miret-Roig et al. 2020) is to be found in Shkolnik et al. (2017). This was soon followed by further comprehensive studies (e.g. Gagné et al. 2018; Lee & Song 2019), with Gagné et al. (2018) updating membership lists using data from Gaia DR2 (Carter et al. 2021).

Previous searches for BPMG members have been mostly based on finding candidate young stars whose spatio-kinematics are closely matched with bona fide group members. Particular focus has been placed on identifying candidate members by observing stars whose 3D-Galactic space velocities\(^1\) are closely matched with the bulk kinematics of the group, since the velocity dispersion of moving groups is typically ~2–3 km s\(^{-1}\) (e.g. Zuckerman & Song 2004; Torres et al. 2006, 2008). Since the discovery of the BPMG (Zuckerman et al. 2001), persistent efforts yielded hundreds of candidate members, most of which, in light of unprecedently precise Gaia DR2 astrometry, were found to be false positives. Even in the Gaia era, it remains a formidable task to vet recently identified candidate BPMG members.

Here we demonstrate that, by bringing in a fourth dimension to the data space, a new perspective is gained that can assist in the task of determining cluster membership by resorting to Euclidean distances between points and a straight line in 4D space. With these distances available for all objects of interest (bona fide group members as well as candidate members; see Section 2), outliers can be flagged. In this study, we select our fourth dimension as the Gaia G magnitude (Arenou et al. 2018; Evans et al. 2018; Weiler 2018), though possibly this can be generalized to work with other well-measured physical parameters as well as applied to other co-moving groups.

The possibility of running a straight line through such a 4D representation of a proposed BPMG membership becomes evident as one examines any of the three 2D plots involving dimension G (G versus U, etc.). However, pinning the line down through some generalization of simple linear regression to more than two dimensions would be fraught with difficult decisions, including deciding which variables to take as independent, something that already plagues the 2D case in several domains.

Aiming to make the representation of the data independent of both measurement units and the potentially large variance differences along the various dimensions, and also make it amenable to visualization in a greater number of 2D plots (six rather than three), we work on standardized versions of U, V, W, G (i.e. versions whose means equal 0 and standard deviations equal 1) and moreover rotate the resulting data in such a way that they become represented by uncorrelated coordinates. Such rotation results from the well-known method of Principal Component Analysis (PCA; Jolliffe & Cadima 2016). A prerequisite of PCA in this work is the use of high-quality data to construct an underlying model for transforming the physical dimensions (i.e. U, V, W, G) into uncorrelated coordinates. This holds also for the desired 4D straight line, which can be obtained by a further, restricted use of PCA.

\(^1\) E-mail: valmir@cos.ufrj.br

\(^1\)These are found in the U, V, W coordinate system, where U points toward the Galactic Centre, V in the direction of rotation, and W toward the Galactic North Pole.

© 2022 The Author(s)
Published by Oxford University Press on behalf of Royal Astronomical Society
We proceed by first describing how we acquired and processed all the necessary data in Section 2, then presenting results in Section 3. Concluding remarks follow in Section 4.

## 2 DATA ACQUISITION AND PROCESSING

We focus on the accounts of BPMG membership given in table 4 of Shkolnik et al. (2017) and in table A1 of Carter et al. (2021). The former of these predates the availability of Gaia DR2. The latter, in turn, is based on an update of the methodology of Gagné et al. (2018) to contemplate data from Gaia DR2, with a selection criterion that only admits objects to which the Bayesian, kinematics-based model in Gagné et al. (2018) ascribes a probability greater than 0.9. Thus, despite the relatively short span of time between the two compilations, they stand in strong contrast with each other. Depending on data completeness and quality, we leverage this contrast to compile a list of bona fide members and another of candidate members, as a proof of concept for our PCA-based membership test.

### 2.1 Data acquisition

All our data were acquired from the SIMBAD astronomical data base (http://simbad.u-strasbg.fr; Wenger et al. 2000) between late March and early April 2022. Data curation for SIMBAD involves assigning a quality tag to each measurement, a letter from A through E (highest to lowest quality), and we used these tags for quality control during data acquisition. All data were recorded or transformed into their values at the J2000 epoch.

To calculate a $U$, $V$, $W$ velocity vector for each star we collate right ascension and declination coordinates ($\alpha$, $\delta$), projected proper motions ($\mu_\alpha$, $\mu_\delta$), parallax ($\pi$), and radial velocity ($\rho$). The Gaia $G$ magnitude is selected for the fourth dimension in the subsequent PCA runs. We refer to an object having data of quality tags A, A, A, A, C, respectively for $\alpha$, $\delta$, $\mu_\alpha$, $\mu_\delta$, $\pi$, $\rho$, and $G$, as being of high quality (C is the only quality tag we ever obtained in queries for $G$). Objects having any other combination of quality tags are referred to as being of lower quality.

All parameters consist only of single-epoch values. Should an object turn out to be in a p-variable, tidally locked binary system ($<\sim5$ per cent, e.g. Binks, Jeffries & Maxted 2015; El-Badry, Rix & Heintz 2021), the resulting $U$, $V$, $W$ will vary on short time-scales. This will likely occur for only a small fraction of all objects and will therefore not affect the outcome of the first PCA run, since as detailed next it is based on a very large number of objects. However, such occurrences must be kept in mind as candidate members are flagged as possible outliers.

For use in the first PCA run, we targeted every high-quality star or multiple system available through SIMBAD, amounting to 1 586 412 objects. These and other data are available from the authors, along with information on their provenance. Most data items are reported as coming from Gaia DR2 or EDR3, with further small contributions mostly from the GALAH Survey DR2 (Buder et al. 2018) and from RAVE DR6 (Steinmetz et al. 2020). However, SIMBAD has been in the process of fully upgrading to the Gaia data releases for a while, so it is somewhat uncertain that provenance as reported in our data does indeed reflect the best available sources. For purposes of this proof-of-concept study, we again rely on the sheer number of objects on which the first PCA run is to be based to ensure that data rotation is well calibrated. As we work further on refining the method and expanding its applicability, a more nuanced approach to data acquisition will likely be chosen.

### 2.2 Data processing

The $U$, $V$, $W$ velocities were calculated using the formulation given by Johnson & Soderblom (1987). Along with the Gaia $G$ magnitudes, $Z$-scores for each component were calculated and used as standardized parameters for the PCA runs.

For processing by PCA the data must be arranged as an $n \times p$ matrix $X$, where $p$ is the data set’s number of dimensions and $n$ is the number of $p$-dimensional data points. The essence of PCA is the calculation of a $p \times p$ rotation matrix $R$ from the covariance matrix of the columns of $X$ and then using it to project the $n$ data points on to new orthogonal directions, yielding the rotated matrix $XR$. The row-$i$, column-$k$ element of $XR$ is the so-called $k$th principal component (PC $k$) of the $i$th data point. Thus, in the present context of $p = 4$, the standardized $U$, $V$, $W$, $G$ get combined into PC 1 through PC 4. Unlike the columns of $X$, those of $XR$ are uncorrelated with one another, and moreover PC 1 explains more of the variance in the data than does PC 2, this one more than PC 3, etc. These two properties justify the typical use of PCA (which is dimensionality reduction), but here we exploit them differently: uncorrelatedness is used to afford better visual exploration; the variance-wise pricacy of PC 1, in turn, is used in the determination of the PC-space version of the 4D straight line that seems to bind BPMG members together.

The first PCA run operated on all high-quality objects accounted for in Table 1, therefore with $n = 1 586 412$. The resulting $R$ can be viewed as a model, learned from high-quality data, of how to project standardized $U$, $V$, $W$, $G$ points on to PC space. Matrix $R$ is shown in Table 2.

The bona fide members’ projections onto PC space were collected in a new $n \times p$ data matrix, now denoted by $X'$ and with $n = 41$, and used as the basis for the second PCA run. This second run did not aim another rotation of the $n$ data points, but merely the determination of the best orthogonal fit of a 4D straight line to them. Letting $\mu = (\mu_1, \ldots, \mu_4)$ be the means vector of the $n$ points, this

| Object counts in the data set used. | High quality | Lower quality | Total |
|------------------------------------|-------------|---------------|------|
| Background objects                 | 1 586 308   | 0             | 1 586 308 |
| Bona fide members                  | 41          | 0             | 41   |
| Candidate members                  | 63          | 58            | 121  |
| Total                              | 1 586 412   | 58            | 1 586 470 |
was achieved by first subtracting $\mu$ off each row of $X'$, then running PCA to determine the new rotation matrix $R'$. Let $a_1$ be the first column of $R'$, that is, the vector to which the first component (now called PC 1') of the data in $X'$ refers. In parametric form, and for $-\infty < t < \infty$, the straight line going through $\mu$ in the direction of $a_1$ is $f(t) = \mu + ta_1$. Because PC 1' accounts for more variance in $X'$ than any of the other three principal components, $f(t)$ is the straight line to which all $n$ points are closest in the least-squares sense and is therefore the desired 4D straight line. We henceforth view $f(t)$ as a linear model of the BPMG and refer to it simply as PC 1'. Vectors $\mu$ and $a_1$ are shown in Table 3.

### 3 RESULTS

Our results are summarized as the 2D plots in the six panels of Fig. 1. Each panel zooms in on a different 2D projection of the standardized and rotated $U$, $V$, $W$, $G$ vectors for all the objects accounted for in Table 1. The background objects are represented as tiny dots and the bona fide and candidate members as larger dots in the foreground. Note that the latter include all the 104 high-quality objects and the remaining 58, lower-quality ones. One technicality is that these 58 objects had to undergo the same standardization and rotation as those of high quality. Thus, because they did not participate in the first PCA run, their Z-scores had to be calculated using means and standard deviations to which they did not contribute. Visible in all panels is the corresponding projection of PC 1' as well.

#### 3.1 Outlier detection

While the plots in Fig. 1 seem to suggest that the bona fide members are very close to the PC 1' line, to varying degrees this can also be said of the candidate members. A better view of the distances for all these 162 objects is given in Fig. 2. Not only does this figure show that distances are in fact spread more widely than we might suppose by simply examining Fig. 1, it also shows that some of those objects do indeed lie farther apart from PC 1' than do the bona fide ones or those that appear to lie near them. It then makes sense that we should try and set aside the ones that would be flagged as outliers by some criterion. Most such criteria assume the data to follow a normal distribution, though testing them for this is inevitably affected by the presence of the very outliers to be eventually flagged.

Distances to PC 1' are definitely not normally distributed, but visually inspecting the distribution of their natural logarithms does change this significantly, as illustrated in Fig. 3. That is, the distances themselves seem to follow an approximately lognormal distribution. Even this is far from fully reliable, though, as the MLE (maximum-likelihood estimate) normal density plotted in the same figure suggests. In cases such as this, customarily one resorts to outlier-flagging methods that substitute the median for the mean and the MAD (median absolute deviation) for the standard deviation. In general this provides some added robustness in the face of uncertain normality of the data.

A widely used method that does this is the modified Z-score method (Iglewicz & Hoaglin 1993). For $ln(d)$ the vector of natural logarithms of all 162 distances to PC 1' and $ln(d_i)$ its ith element, this method recommends flagging the $i$th object as an outlier for further investigation if the absolute difference of $ln(d_i)$ and the median of $ln(d)$ surpasses the MAD of $ln(d)$ by more than a factor of 3.50/0.6745 (MAD/0.6745 is an approximately 1 $\sigma$ value in the normal case). This leads to a lower threshold of about $-6$, below which flagging is recommended (that is, for distances below $e^{-6} \approx 0.002$) and an upper threshold of about $1.41$, above which flagging is recommended as well (that is, for distances above $e^{-1.41} \approx 0.244$). These thresholds are marked as grey lines in both Figs 2 and 3, setting apart no objects lying strictly below the lower threshold and only two lying strictly above the upper threshold. These possible outliers are both candidate members and listed in Table 4 in decreasing order of distance to PC 1' with their radial velocities.

#### 3.2 Possible outliers in the BPMG

Except for $\rho$, the spatial and kinematic data we downloaded for the two possible outliers listed in Table 4 all come from Gaia EDR3. These two objects, however, were included in table 4 of Shkolnik et al. (2017) from earlier literature, therefore prior to the availability of any data from the Gaia mission. Neither one appears in table A1 of Carter et al. (2021). 2MASS J00325584-4405058, a brown dwarf, is one of our lower-quality candidate members and was brought in from Gagné et al. (2015), where it is listed as a candidate member despite the absence of a reliable $\rho$ measurement at the time. It was moreover not tested by the proposed selection process in Shkolnik et al. (2017), having therefore undergone no further confirmation. 2MASS J2035152-2806020, also a brown dwarf, is one of our high-quality candidate members. It was brought in from Liu, Dupuy & Allers (2016), where the authors caution that membership confirmation depends on further verification, ‘typically with radial velocity measurements.’ It passed the test in Shkolnik et al. (2017), based in part on $\rho = -5.81 \pm 0.5 \text{ km s}^{-1}$, which is marginally within 1 $\sigma$ of the value used in our analysis ($\rho = -6.53 \pm 0.24 \text{ km s}^{-1}$; see Table 4 and Faherty et al. 2016).

It must also be noted that, though not listed as a possible outlier in Table 4, 2MASS J18141047-3247344 (a spectroscopic binary hosting a protoplanetary disc) misses the upper threshold by only $9.46 \times 10^{-5}$. It too was brought into Shkolnik et al. (2017) from elsewhere (Torres et al. 2008) and passed the authors’ selection process. However, this seems to have taken into account $\rho = -13.3 \pm 7.7 \text{ km s}^{-1}$ while what we have in our data is markedly different ($\rho = -51.01 \pm 0.59 \text{ km s}^{-1}$; see Miret-Roig et al. 2020). This notwithstanding, this object’s absence from Table 4 is of course amply supported by its presence amid our bona fide members, which we recall comprises all high-quality objects in table A1 of Carter et al. (2021).

As a final remark, we note that in all three cases (the two candidate members in Table 4 and the marginal bona fide member) it seems advisable to keep monitoring the measurements of $\rho$. This is not only on account of some of the inconsistencies pointed out above, but is...
Figure 1. 2D projections of the $U, V, W, G$ vectors of all objects in Table 1 after standardization and rotation, for all six pairs of PC axes. Background objects are coloured in shades of blue to indicate object density (0 is sparsest, 1 is densest). PC $1'$ is projected as well. One curious feature of the PC 1 versus PC 4 panel is the ‘stream’ of points with PC 1 between 2 and 3, PC 4 between −1 and 0. They correspond to the 91 members of the NGC 3201 globular cluster in the data.

also important in regard to the data-provenance and $\rho$-variability issues mentioned in Section 2.1.

4 CONCLUDING REMARKS

Deciding on an object’s membership status in the BPMG (and, in general, in any other young stellar moving group) must rely on the measurement of several astrophysical properties that can be hard and expensive to acquire. In this letter, we have introduced a new tool that can be used in the process. At the heart of this tool is PCA, one of the staple techniques in data science, here used for its ability to rotate variables so they become uncorrelated and to automatically identify the one resulting variable to which more variance in the data can be ascribed than to any other. The outcome of interest is the 4D straight line we have called PC $1'$. The essential computational tasks involved in the process take only minutes to complete on any current CPU core. Code for them in the R language is available online as supplementary material.

In the case of the BPMG, we demonstrated that PC $1'$ can function as a linear model of the group, since distances from candidate members to it can be used as a kind of proxy for group membership. We reached this conclusion by selecting bona fide and candidate members from the literature, and then using the bona fide members to construct PC $1'$. This was preceded by an initial selection of more than 1.5 million objects for standardization and visualization of the data. For the case of the BPMG these data were $U, V, W, G$, but
alternatives can certainly be considered. In this regard, we note that we conducted some trials with the 3D-Galactic coordinates \( X, Y, Z \), and also with magnitudes \( B \) and \( V \), but only the latter seemed nearly as promising as \( G \). Those trials were not thorough, however, so further consideration may be in order.

Of course, one downside of the overall approach is that complete data are needed for all participating objects. But then again, obtaining such data is a constantly pursued goal. In general, it is a matter of time before they become available.

Figure 3. Normalized histogram and MLE normal density of the natural logarithms of distances to PC 1’ for the bona fide and candidate members. The corresponding original object counts are given on the vertical axis on the right for context. The vertical grey lines mark the lower and upper thresholds (about −6 and −1.41) for outlier flagging.

Table 4. Possible outliers relative to distances to PC 1’.

| 2MASS J          | Distance to PC 1’ | Value of \( \rho \) in the data |
|------------------|-------------------|---------------------------------|
| 00325584-4405058 | 0.696             | 12.95 ± 1.92 km s\(^{-1}\)     |
| 20135152-2806020 | 0.285             | −6.53 ± 0.24 km s\(^{-1}\)     |

**ACKNOWLEDGEMENTS**

We thank Vladimir Ortega for introducing us to the BPMG, Joel Kastner for commenting on an early version, and Alexander Binks for his careful and detailed reviews. We acknowledge partial support from Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), and a BBP grant from Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ).

**DATA AVAILABILITY**

The data underlying this article are available from the corresponding author’s website (https://www.cos.ufrj.br/~valmir/bpic.tar.gz).

**REFERENCES**

Arenou F. et al., 2018, A&A, 616, A17
Binks A. S., Jeffries R. D., Maxted P. F. L., 2015, MNRAS, 452, 173
Buder S. et al., 2018, MNRAS, 478, 4513
Carter A. L. et al., 2021, MNRAS, 501, 1999
El-Badry K., Rix H.-W., Heintz T. M., 2021, MNRAS, 506, 2269
Evans D. W. et al., 2018, A&A, 616, A4
Faherty J. K. et al., 2016, ApJS, 225, 10
Gagné J., Lafrenière D., Doyon R., Malo L., Artigau É., 2015, ApJ, 798, 73
Gagné J. et al., 2018, ApJ, 856, 23
Iglewicz B., Hoaglin D., 1993, How to Detect and Handle Outliers. The ASQC Basic References in Quality Control: Statistical Techniques, Vol. 16, ASQC Quality Press, Milwaukee, WI
Johnson D. R. H., Soderblom D. R., 1987, AJ, 93, 864
Jolliffe I. T., Cadima J., 2016, Philos. Trans. R. Soc. A, 374, 20150202
Lagrange A.-M. et al., 2010, Science, 329, 57
Lagrange A.-M. et al., 2019, Nat. Astron., 3, 1135
Lee J., Song L. 2019, MNRAS, 489, 2189
Liu M. C., Dupuy T. J., Allers K. N., 2016, ApJ, 833, 96
Miret-Roig N. et al., 2020, A&A, 642, A179
Shkolnik E. L., Allers K. N., Kraus A. L., Liu M. C., Flagg L., 2017, AJ, 154, 69
Steinmetz M. et al., 2020, AJ, 160, 83
Torres C. A. O., Quast G. R., da Silva L., de la Reza R., Melo C. H. F., Sterzik M., 2006, A&A, 460, 695
Torres C. A. O., Quast G. R., Melo C. H. F., Sterzik M. F., 2008, in Reipurth B., ed., Handbook of Star Forming Regions, Volume II: The Southern Sky, Astronomical Society of the Pacific, San Francisco, CA, p. 757
Weiler M., 2018, A&A, 617, A138
Wenger M. et al., 2000, A&AS, 143, 9
Zuckerman B., Song I., 2004, ARA&A, 42, 685
Zuckerman B., Song I., Bessell M. S., Webb R. A., 2001, ApJ, 562, L87

**SUPPORTING INFORMATION**

Supplementary data are available at MNRASL online.

Table S1. Identifiers of the bona fide members.
Table S2. Identifiers of the candidate members.

Please note: Oxford University Press is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

This paper has been typeset from a TeX\&\LaTeX file prepared by the author.