Monthly Notices
of the
ROYAL ASTRONOMICAL SOCIETY

Mon. Not. R. Astron. Soc. **424,** 2832–2840 (2012) doi:10.1111/j.1365-2966.2012.21413.x

Application of the Gaussian mixture model in pulsar astronomy – pulsar classification and candidates ranking for the *Fermi* 2FGL catalogue

K. J. Lee,[1⋆] L. Guillemot,[1] Y. L. Yue,[2] M. Kramer[1,3] and D. J. Champion[1]

[1]*Max-Planck-Institut für Radioastronomie, Auf dem Hügel 69, D-53121 Bonn, Germany*
[2]*National Astronomical Observatories, Chinese Academy of Sciences, A20 Datun Road, Chaoyang District, Beijing 100012, China*
[3]*Jodrell Bank Centre for Astrophysics, University of Manchester, Manchester M13 9PL*

Accepted 2012 May 29. Received 2012 May 29; in original form 2011 November 14

ABSTRACT

Machine learning, algorithms designed to extract empirical knowledge from data, can be used to classify data, which is one of the most common tasks in observational astronomy. In this paper, we focus on Bayesian data classification algorithms using the Gaussian mixture model and show two applications in pulsar astronomy. After reviewing the Gaussian mixture model and the related expectation–maximization algorithm, we present a data classification method using the Neyman–Pearson test. To demonstrate the method, we apply the algorithm to two classification problems. First, it is applied to the well-known period–period derivative diagram, where we find that the pulsar distribution can be modelled with six Gaussian clusters, with two clusters for millisecond pulsars (recycled pulsars) and the rest for normal pulsars. From this distribution, we derive an empirical definition for millisecond pulsars as $\frac{\dot{P}}{10^{-17}} \leq 3.23 \left(\frac{P}{100\ \mathrm{ms}}\right)^{-2.34}$. The two millisecond pulsar clusters may have different evolutionary origins, since the companion stars to these pulsars in the two clusters show different chemical compositions. Four clusters are found for normal pulsars. Possible implications for these clusters are also discussed. Our second example is to calculate the likelihood of unidentified *Fermi* point sources being pulsars and rank them accordingly. In the ranked point-source list, the top 5 per cent sources contain 50 per cent known pulsars, the top 50 per cent contain 99 per cent known pulsars and no known active galaxy (the other major population) appears in the top 6 per cent. Such a ranked list can be used to help the future follow-up observations for finding pulsars in unidentified *Fermi* point sources.

Key words: methods: statistical – pulsars: general – gamma-rays: stars.

1 INTRODUCTION

A common and important task in observational astronomy is to find or select a certain type of objects from a sample of candidates according to particular physical properties. Examples include selecting active galaxy (AG) candidates from a multicolour optical photometry survey and selecting good pulsar candidates from a large number of candidates produced in pulsar searches. Such tasks are time consuming, especially when the number of candidates is large. In this situation, computers can offer significant help when the selection rules can be derived based on prior experiences or physical considerations. However, for some applications, it is hard to determine the a priori criteria and one has to search for the selection criterion using the *empirical* knowledge embedded in the data themselves. For example, different types of sources usually form clusters in different regions of parameter space. When a large population of candidates is available, the clustering becomes statistically significant, and one can then use this to determine the selection criterion.

In order to build up the empirical selection criterion, we need to find a method to *learn* the knowledge from the data and apply this to generate the selection rules. Machine learning algorithms are designed to extract empirical knowledge from a sample of data and improve their performance based on the knowledge they learnt. They also contain methods to classify data. There have already been many successful applications of machine learning algorithms since the 1960s, some of which were recently used in the pulsar community (e.g. Eatough et al. 2010). We refer to Theodoridis & Koutroumbas (2009) for the details of such algorithms and their applications in broader fields. Clearly, machine learning and related classification algorithms can help to determine the criteria required to select desirable objects from a large sample of candidates in the context of observational astronomy.

⋆E-mail: kjlee@mpifr-bonn.mpg.de

This paper demonstrates the application of the Gaussian mixture model (GMM) in the context of pulsar astronomy. The GMM, which we use here, is one type of unsupervised learning algorithms based on Bayesian decision theory (Press et al. 2007). It assumes that the data clusters in parameter space follow a superposition of several multivariate Gaussian distributions. The parameters of each cluster are determined from data using the expectation–maximization (EM) algorithm. We use two examples to show the application of the GMM in pulsar astronomy. Our first example is to classify pulsars in the parameter space of pulsar period (P) and period derivative ($\dot{P}$), and the second example is to calculate the likelihood of a gamma-ray point source being a pulsar. Here gamma-ray point-source parameters are from the *Fermi* Gamma-ray Space Telescope Large Area Telescope 2-year point-source catalogue (2FGL catalogue; The Fermi-LAT Collaboration 2011b).

This paper is organized as follows. We introduce the statistical technique, the GMM, in Section 2. We use two problems as examples to show the properties and applications of the GMM in Section 3. The first application in Section 3.1 is to find an empirical definition for millisecond pulsars (MSPs) from the period–period derivative (P–$\dot{P}$) distribution of known pulsars. The second application in Section 3.2 is to describe the 2FGL catalogue point-source distribution, to calculate the likelihood of a particular source being a pulsar and to produce a pulsar candidate list for later confirmation observations. Conclusions and discussions are given in Section 4.

2 GMM AND LIKELIHOOD RANKING

In this section, we introduce the basic concepts of the GMM as well as related techniques to classify the data in parameter space.

The GMM is a probabilistic model to describe the distribution of data with clusters in the parameter space, where each cluster is assumed to follow the Gaussian distribution. For a total of m clusters in an n-dimensional parameter space, the probability distribution $P(\boldsymbol{x})$ of data $\boldsymbol{x}$ is given by a weighted summation of all m Gaussian clusters, i.e.

$$P(\boldsymbol{x}) = \sum_{k=1}^{m} P_k P(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\,, \tag{1}$$

where the mixture weight of the kth Gaussian is P_k and the distribution of each individual Gaussian cluster is

$$P(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \frac{\exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)\cdot{\boldsymbol{\Sigma}_k}^{-1}\cdot(\boldsymbol{x}-\boldsymbol{\mu}_k)\right]}{(2\pi)^{n/2}\sqrt{|\boldsymbol{\Sigma}_k|}}\,. \tag{2}$$

$|\Sigma_k|$ is the determinant of Σ_k, the $\boldsymbol{\mu}_k$ and Σ_k are the mean vector and covariance of the kth Gaussian, respectively.

The parameters of the GMM, i.e. P_k, $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$, can be determined from the data by an unsupervised machine learning technique, namely the EM algorithm (Press et al. 2007), which assumes no prior knowledge about the clustering structures. For N data points $\boldsymbol{x}_i$, here $i = 1, \ldots, N$, the EM algorithm starts from an initial guess and learns GMM parameters from the data. The steps involved are as follows.

(1) Guess starting values for $\boldsymbol{\mu}_k$, $\boldsymbol{\Sigma}_k$ and P_k.

(2) Expectation-step (E-step): calculate $P(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ and $P(\boldsymbol{x})$ using equations (1) and (2).

(3) Maximization-step (M-step): update model parameters $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ using

$$\boldsymbol{\mu}_k,\mathrm{new} = \frac{\sum_{i=1}^{N}\boldsymbol{x}_i P_k P(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)/P(\boldsymbol{x}_i)}{\sum_{i=1}^{N} P_k P(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)/P(\boldsymbol{x}_i)}\,, \tag{3}$$

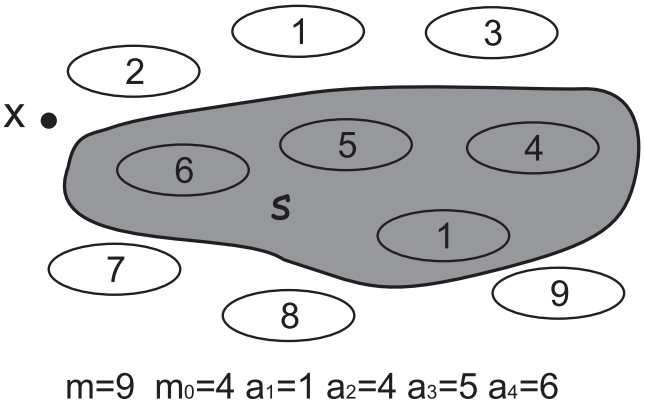

Figure 1. An illustration of the definition of the GMM and data clustering. The $\boldsymbol{x}$ indicates a data point. The index for each Gaussian is labelled. Here the total number of Gaussian clusters is $m = 9$, where four of them belong to the subset $\mathcal{S}$ ($m_0 = 4$), as indicated by the shaded area; thus, $\boldsymbol{x} \notin \mathcal{S}$.

$$\Sigma_{k,\mathrm{new}} = \frac{\sum_{i=1}^{N}(\boldsymbol{x}_i-\boldsymbol{\mu}_k)\otimes(\boldsymbol{x}_i-\boldsymbol{\mu}_k)P_k P(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)/P(\boldsymbol{x}_i)}{\sum_{i=1}^{N} P_k P(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)/P(\boldsymbol{x}_i)}\,, \tag{4}$$

$$P_{k,\mathrm{new}} = \frac{1}{N}\sum_{i=1}^{N} P_k P(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)/P(\boldsymbol{x}_i)\,, \tag{5}$$

where i is the index for data points. $\otimes$ is the symbol for 'outer product', i.e. for any vectors $\boldsymbol{x}$ and $\boldsymbol{y}$, $\boldsymbol{x}\otimes\boldsymbol{y}$ is the matrix, the νth row, μth column element of which is the product of the νth element of $\boldsymbol{x}$ and the μth element of $\boldsymbol{y}$.

(4) Repeat the EM steps, until the total likelihood Λ converges, where

$$\Lambda = \prod_{i=1}^{N} P(\boldsymbol{x}_i)\,. \tag{6}$$

It can be shown that the above iteration of the EM algorithm increases the total likelihood Λ, and the iterations always converge.[1]

With the GMM and its parameters, one can infer the association of any data point $\boldsymbol{x}$ with these clusters. One can also ask if $\boldsymbol{x}$ belongs to a certain subset $\mathcal{S}$, which contains m_0 Gaussian clusters out of a total of m Gaussian clusters. An illustration of the definitions is presented in Fig. 1, where the indices of Gaussian clusters in $\mathcal{S}$ are a_k, $k = 1, \ldots, m_0$. The question of association can be answered via the standard likelihood ratio test. The Neyman–Pearson lemma (Kassam 1988) claims that the most powerful test for the binary hypotheses

$$\begin{cases} H_1\text{: } \boldsymbol{x} \text{ belongs to subset } \mathcal{S}\,; \\ H_0\text{: } \boldsymbol{x} \text{ belongs to other clusters,} \end{cases} \tag{7}$$

is to compare the logarithmic likelihood ratio $\log R_{\mathcal{S}}$ against a statistical decision threshold η, i.e. choose H_0, if $\log R_{\mathcal{S}} \geq \eta$, otherwise choose H_1. According to the GMM, the logarithmic likelihood ratio $R_{\mathcal{S}}$ is

$$\log R_{\mathcal{S}} = \log\left(\frac{\sum_{k\in\mathcal{S}} P_k P(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{k\notin\mathcal{S}} P_{k'} P(\boldsymbol{x}|\boldsymbol{\mu}_{k'}, \boldsymbol{\Sigma}_{k'})}\frac{\sum_{k'\notin\mathcal{S}} P_{k'}}{\sum_{k\in\mathcal{S}} P_k}\right), \tag{8}$$

where summation $\sum_{k\in\mathcal{S}}$ sums over the index k for the clusters in the subset $\mathcal{S}$ and $\sum_{k\notin\mathcal{S}}$ sums over the complementary set of $\mathcal{S}$, i.e. those clusters not in the subset $\mathcal{S}$.

[1] In real situations, due to the finite numerical precision, this is not always true.

The number of Gaussian clusters m, as an input parameter, can be determined using statistical methods. In practice, one usually starts from $m = 1$, and then increases m. As m is increased, one can describe the data better. To avoid overfitting, it is necessary to check the modelling using the multidimensional Kolmogorov–Smirnov test (K-S test). Similar to a 1D K-S test, the multidimensional K-S test is used to test whether two data sets differ significantly from each other or whether a data set differs from a known distribution. The statistic ($\mathcal{D}$) for the K-S test is the maximal difference between the cumulative distribution of two data sets or between the data and the model. However, the cumulative distribution of multidimensional data is not well defined; thus, it was proposed to compute such a 'cumulative distribution' for any possible order and then calculate the $\mathcal{D}$. For example, in order to determine the maximal difference $\mathcal{D}$ between data and data or between data and distribution, it is necessary to check the cumulative probability for all of the four cases, $(x < x_i, y < y_i)$, $(x < x_i, y > y_i)$, $(x > x_i, y < y_i)$ and $(x > x_i, y > y_i)$, for any data point (x_i, y_i) belonging to the 2D data set $\boldsymbol{x} = (x_i, y_i), i \in [1, N]$. For a multidimensional K-S test, the statistical threshold and p values are usually calculated numerically via Monte Carlo simulation, which generates the mock data sets and calculates the null-hypothesis distribution of $\mathcal{D}$ accordingly. We refer the readers to Peacock (1983) and Fasano & Franceschini (1987) for more details of such tests.

In summary, the technique of using the GMM to describe a data distribution and to determine the data association is given in the following recipe.

(1) Determine the parameter space and form the data vector $\boldsymbol{x}$ for the data set.
(2) Guess the number of clusters m and their initial parameters, i.e. $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$, where $k = 1, \ldots, m$.
(3) Use the EM algorithm to find the true model parameters.
(4) Check if the GMM describes the data distribution well enough using the multidimensional K-S test. Increase the number of Gaussian clusters, if the test fails.
(5) Once the model parameters are found, use equation (8) to determine the data association.

We present two examples in the next section to show its applications.

3 APPLICATION

3.1 Application to pulsar classification

As the first example, we apply the GMM to the well-known pulsar P–$\dot{P}$ diagram to find a quantitative description for the

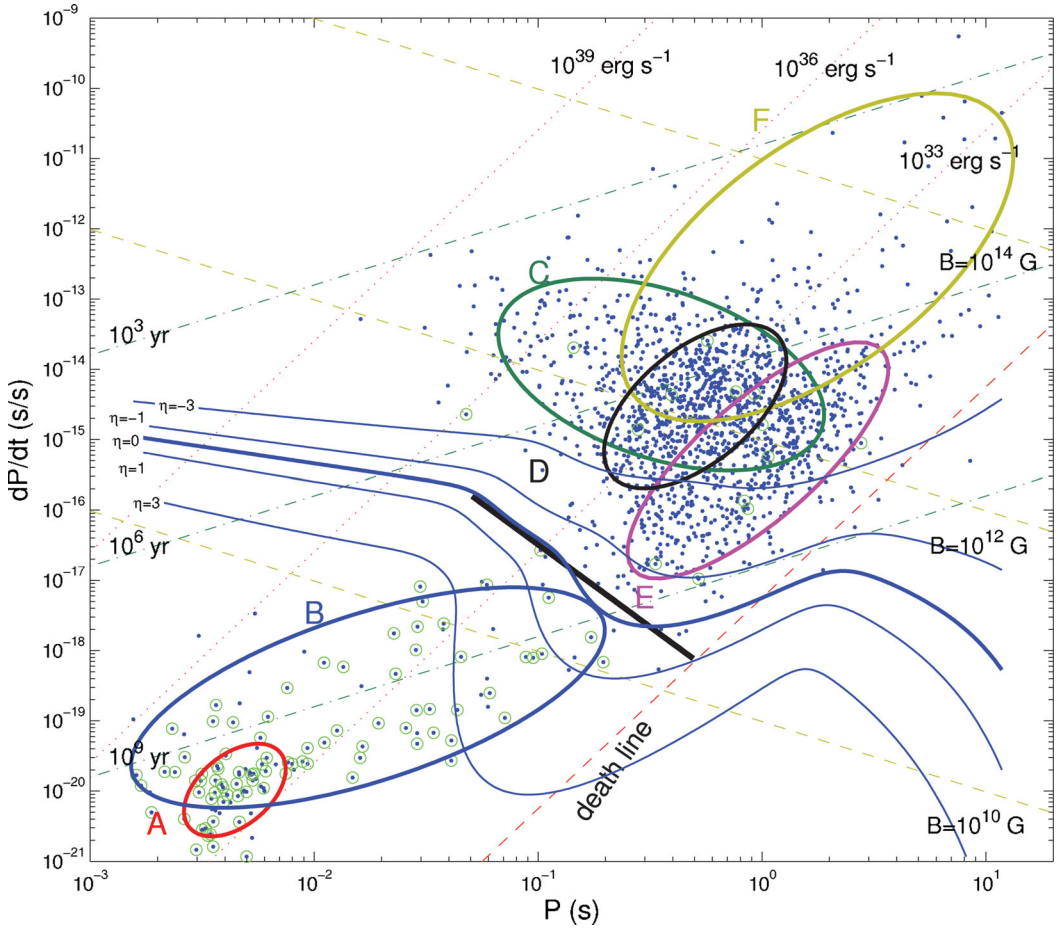

Figure 2. The pulsar P–$\dot{P}$ diagram. Elliptic curves are 2σ contours for the six Gaussian clusters found by the GMM. The dots are all detected pulsars from the ATNF catalogue, and those with circles are binaries. The blue curves with numerical labels on the left ends are contours for equal logarithmic likelihood $\log(R_S)$. The definition of an MSP according to our model is $\log R_S \leq 0$, which is plotted with a thick blue line. The linear approximation of $\log R_S = 0$ with $50\,\mathrm{ms} \leq P \leq 0.5\,\mathrm{s}$ [i.e. equation (10)] is plotted using a thick black straight line. The lines with equal magnetic field strength, spin-down power, characteristic age as well as the death line are also plotted. Data are from the ATNF pulsar catalogue (Manchester et al. 2005, http://www.atnf.csiro.au/research/pulsar/psrcat/).

pulsar distribution. We also seek the 'empirical MSP definition' here, especially because a precise definition for MSPs using their periods and period derivatives is not available yet.

The standard picture for pulsar evolution contains several major stages: (i) the birth of a pulsar in a supernova explosion, (ii) the spin-down of the pulsar due to radiation energy loss, (iii) the pulsar death due to the decrease in radiation power, and possibly for some binary systems, (iv) the pulsar recycling by an accretion-induced spin-up. After birth in the supernova, the *young pulsars* usually have short periods and large period derivatives. They occupy the upper-left part of the P–$\dot{P}$ diagram, and are frequently associated with supernova remnants. As the pulsars age, they slow down, while, for those pulsars with breaking index $n = 2 - \ddot{P}P/\dot{P}^2 > 2$, their $\dot{P}$ also decreases. The pulsars then enter the main population in the centre of the diagram, referred to as the *normal pulsars*. Eventually, such a spin-down causes the *death* of pulsars, i.e. the radiation of pulsar ceases or becomes too weak to detect. Pulsars may also get *recycled* via the accretion process (Bhattacharya & van den Heuvel 1991). The MSPs, which occupy the lower-left corner of the P–$\dot{P}$ diagram, are commonly believed to form due to such a spin-up of a normal pulsar via accretion materials from a companion star. A continuum of pulsars from the MSP population to the normal pulsar population is already observed, where the intermediate population is referred to as *mildly recycled pulsars*. There are also pulsars occupying the upper-right part of the P–$\dot{P}$ diagram. The origin and evolution of these *high magnetic field pulsars* is still unclear.

We use pulsar P and $\dot{P}$ values from the ATNF pulsar catalogue (Manchester et al. 2005). The distribution of pulsars is plotted in Fig. 2. The distribution of the whole pulsar population does not follow a Gaussian distribution, but we can approximate the overall distribution with multiple Gaussian clusters, i.e. the GMM is still valid for the modelling purposes. We apply the GMM in the P–$\dot{P}$ diagram, so our parameter vector $\boldsymbol{x}$ is

$$\boldsymbol{x} = \log \begin{pmatrix} P \\ \dot{P} \end{pmatrix} . \tag{9}$$

We directly apply the GMM to the data set. The EM algorithm for the GMM is stable so the EM algorithm converges for most of the initial values, although the EM algorithm does not guarantee that it converges to the best global maximum of the total likelihood Λ. In order to attain the global maximum, initial GMM parameters are generated randomly from a uniform distribution covering the whole P–$\dot{P}$ diagram (in particular, we choose the range of P from 10^{-4} to 20 s and the range of $\dot{P}$ from 10^{-22} to 10^{-10}). We then use these initial parameters in the EM algorithm. We repeat this process 10^4 times to determine the best global model parameters giving the largest likelihood value. In this case, the probability to converge to the best model is more than 30 per cent for all guesses. The GMM is tested using the multidimensional K-S test (Peacock 1983; Fasano & Franceschini 1987), for which the p value is chosen to be 95 per cent. To pass such a test, we need six different Gaussian components in the GMM.

We check the 'robustness' of GMM parameters using an algorithm similar to the bootstrap method. The original P–$\dot{P}$ data are resampled with replacements to form 10^2 simulated data sets, i.e. any data point has the same probability of being sampled at any time. The EM algorithm is then applied to these newly simulated data sets, and the stability of GMM parameters as a function of the total number of data points in each simulated data set is checked. We find that there is no significant change in the structure of the Gaussian clusters if we randomly remove less than 3 per cent of the data. We have also checked the GMM parameter stability with respect to the Shklovskii effect (Shklovskii 1970), an effect that increases the observed period derivative of the pulsar due to its transverse velocity. The differences between GMM parameters for pulsar distributions with and without correcting the Shklovskii effect are less than 1 per cent (the corrected $\dot{P}$ values are also from the ATNF catalogue). Since the Shklovskii correction is not substantial for the GMM, we ignore it in the rest of the discussions. One needs to redo the above analysis to further check the stability and robustness of the model parameters, when more pulsar data become available.

Table 1. Numerical values of the GMM parameters for the pulsar distribution in the P–$\dot{P}$ space. The definitions are in equations (1) and (2). The mixture weights, $P_A, \ldots, P_F$, as probabilities, are dimensionless. The central vectors of clusters, $\boldsymbol{\mu}_A, \ldots, \boldsymbol{\mu}_F$, have the same units with respect to $\boldsymbol{x}$, i.e. (log (s), log (s/s)). The covariance matrices $\Sigma_A, \ldots, \Sigma_F$ are in the linear space of $\boldsymbol{x} \bigotimes \boldsymbol{x}$, so their units are $\begin{pmatrix} \langle XpX \rangle & \log(\mathrm{s})^2 & \log(\mathrm{s})\log(\mathrm{s/s}) \\ \log(\mathrm{s})\log(\mathrm{s/s}) & \log(\mathrm{s/s})^2 \end{pmatrix}$.

Name	Value
P_A	0.0326
P_B	0.0403
P_C	0.2474
P_D	0.3170
P_E	0.3337
P_F	0.0290
$\boldsymbol{\mu}_A$	$(-2.3541 \quad -19.9847)$
$\boldsymbol{\mu}_B$	$(-1.7597 \quad -18.6687)$
$\boldsymbol{\mu}_C$	$(-0.4502 \quad -14.0749)$
$\boldsymbol{\mu}_D$	$(-0.2983 \quad -14.5266)$
$\boldsymbol{\mu}_E$	$(-0.0179 \quad -15.2945)$
$\boldsymbol{\mu}_F$	$(0.2477 \quad -12.4064)$
Σ_A	$\begin{pmatrix} 0.0172 & 0.0229 \\ 0.0229 & 0.1451 \end{pmatrix}$
Σ_B	$\begin{pmatrix} 0.3732 & 0.3388 \\ 0.3388 & 0.8204 \end{pmatrix}$
Σ_C	$\begin{pmatrix} 0.1764 & -0.1382 \\ -0.1382 & 0.6213 \end{pmatrix}$
Σ_D	$\begin{pmatrix} 0.0552 & 0.0886 \\ 0.0886 & 0.4530 \end{pmatrix}$
Σ_E	$\begin{pmatrix} 0.1131 & 0.2524 \\ 0.2524 & 0.9418 \end{pmatrix}$
Σ_F	$\begin{pmatrix} 0.2547 & 0.4106 \\ 0.4106 & 1.8187 \end{pmatrix}$

As plotted in Fig. 2, in order to describe the pulsar distribution in the P–$\dot{P}$ diagram, six Gaussian clusters are required, the parameters of which are listed in Table 1. It is clear that two Gaussian clusters (components 'A' and 'B') are required for describing the MSP distribution and four clusters (components 'C', 'D','E' and 'F') are required for normal pulsars. We calculate the likelihood ratio according to equation (8), where the set $\mathcal{S}$ contains the two MSP clusters. We plot the equal likelihood ratio contours corresponding to $R_{\mathcal{S}} = \eta$ in Fig. 2. Any of these curves divides the P–$\dot{P}$ diagram into two regions, one for MSPs and the other for normal pulsars. We can now empirically define MSPs as pulsars satisfying $R_{\mathcal{S}} \geq 0$. We can derive a linear approximation in $\log P$ to $R_{\mathcal{S}} = 0$. In order to avoid having regions, where only very few pulsars are present, the linear approximation can be confined to the interesting range of P, i.e. the linear approximation is calculated for P from 50 ms to

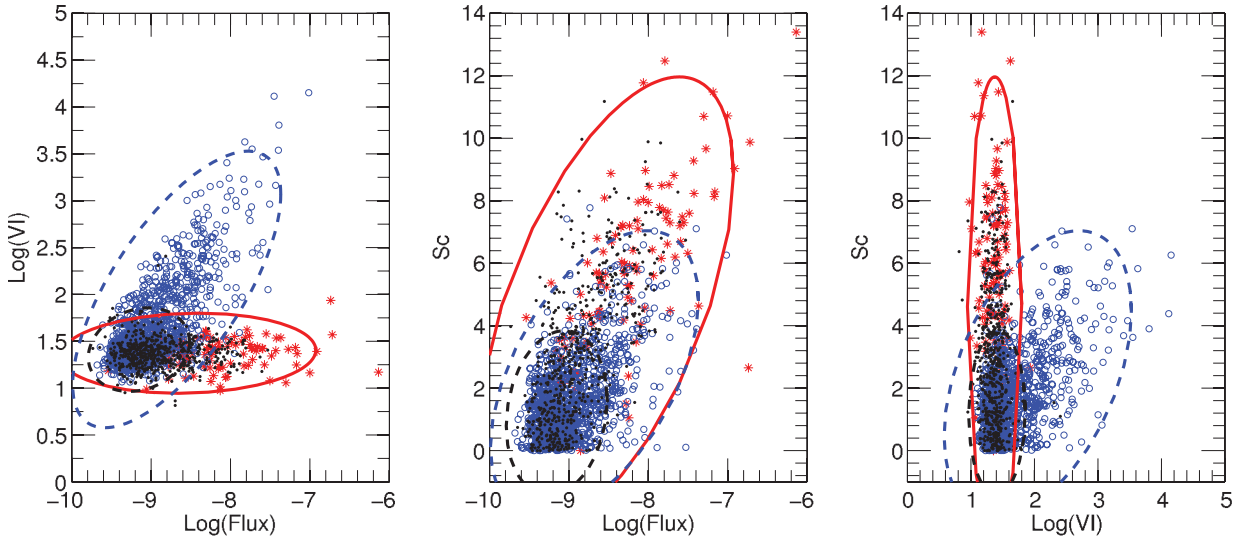

Figure 3. The projected Gaussian clusters and source distribution. The blue '$\circ$' symbols denote AGs, the red '$*$' symbols denote pulsars and unassociated sources are plotted with black solid dots. The units we use for each axis are identical to that in the 2FGL catalogue, where the F_{1000} takes the unit of cm^{-2} s^{-1}. The significance of the curved spectrum and the variability index are dimensionless according to the 2FGL catalogue. There is a 3D movie of this figure available at http://www.mpifr-bonn.mpg.de/staff/kjlee/gmm.html.

0.5 s as

$$\frac{\dot{P}}{10^{-17}} \leq 3.23 \left(\frac{P}{100\ \mathrm{ms}}\right)^{-2.34}. \quad (10)$$

This can be seen as an empirical 'definition' for MSPs.

Individual Gaussian clusters may be artefacts, due to an intrinsic pulsar distribution that is non-Gaussian. In this case, multiple components are required to describe the distribution. Such a non-Gaussian distribution may come from selection effects in pulsar searching or pulsar evolutionary mechanisms. However, we would like to mention a few interesting features here, which may agree with other evidence.

There are two possible MSP clusters. As shown in Fig. 2, the principal axes with maximal eigenvalue for component 'A' are parallel to constant energy losing rate lines [defined by $\dot{E} = 4\pi^2 I \dot{P} P^{-3} = 3.9 \times 10^{46} (P/\mathrm{s})^{-3}\ \dot{P}$ erg], while the principal axes with maximal eigenvalue of component 'B' are parallel to the equal characteristic age lines (defined by $\tau_c = 0.5\, P \dot{P}^{-1}$). Such different directions of eigenvectors may indicate that the MSPs have two different origins or evolutionary tracks, which is supported by the fact that nearly all component-'A' pulsars are He white dwarf binaries, while the major population in component 'B' is of CO white dwarf binaries (Tauris 2011).

Component 'B' contains mildly recycled pulsars, which have larger period and higher magnetic field than other MSPs. According to the GMM, there is no statistical evidence for a separate cluster of mildly recycled pulsars in the P–$\dot{P}$ diagram. This may indicate a 'smooth spectrum' for the amount of accreted mass for fully recycled and mildly recycled MSPs; otherwise, we would expect a more complex structure in component 'B'.

There are possibly four normal pulsar clusters. One each for high magnetic field pulsars (component 'F'), old pulsars close to the death line (component 'E'), young pulsars (component 'C') and middle age pulsars (component 'D'). The principal axis directions of these ellipses for middle age pulsars, old pulsars and high magnetic field pulsars are nearly parallel (within $\sim$3°). Such agreement indicates certain common mechanisms among these pulsars,

Table 2. GMM parameters for the point-source distribution in the 2FGL catalogue. Similar to Table 1, $P_1, \ldots, P_3$ are dimensionless. The vectors $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_3$ have the same units with respect to $\boldsymbol{x}$, i.e. (log (cm^{-2} s^{-1}), dimensionless, dimensionless). The matrices $\boldsymbol{\Sigma_1}, \ldots, \boldsymbol{\Sigma_3}$ take units of $\begin{pmatrix} \log(\mathrm{cm}^{-2}\,\mathrm{s}^{-1})^2 & \log(\mathrm{cm}^{-2}\,\mathrm{s}^{-1}) & \log(\mathrm{cm}^{-2}\,\mathrm{s}^{-1}) \\ \log(\mathrm{cm}^{-2}\,\mathrm{s}^{-1}) & \text{dimensionless} & \text{dimensionless} \\ \log(\mathrm{cm}^{-2}\,\mathrm{s}^{-1}) & \text{dimensionless} & \text{dimensionless} \end{pmatrix}$.

Name	Value
P_1	0.2856
P_2	0.2290
P_3	0.4855
$\boldsymbol{\mu}_1$	$\begin{pmatrix} -8.5330 & 1.3732 & 4.6445 \end{pmatrix}$
$\boldsymbol{\mu}_2$	$\begin{pmatrix} -8.6773 & 2.0533 & 2.4719 \end{pmatrix}$
$\boldsymbol{\mu}_3$	$\begin{pmatrix} -9.1534 & 1.4110 & 1.2071 \end{pmatrix}$
$\boldsymbol{\Sigma_1}$	$\begin{pmatrix} 2.8686\mathrm{e}-01 & 5.9415\mathrm{e}-03 & 7.5103\mathrm{e}-01 \\ 5.9415\mathrm{e}-03 & 2.0322\mathrm{e}-02 & 1.9244\mathrm{e}-03 \\ 7.5103\mathrm{e}-01 & 1.9244\mathrm{e}-03 & 5.9513\mathrm{e}+00 \end{pmatrix}$
$\boldsymbol{\Sigma_2}$	$\begin{pmatrix} 1.9069\mathrm{e}-01 & 1.4785\mathrm{e}-01 & 3.3875\mathrm{e}-01 \\ 1.4785\mathrm{e}-01 & 2.4183\mathrm{e}-01 & 3.2970\mathrm{e}-01 \\ 3.3875\mathrm{e}-01 & 3.2970\mathrm{e}-01 & 2.3165\mathrm{e}+00 \end{pmatrix}$
$\boldsymbol{\Sigma_3}$	$\begin{pmatrix} 4.5017\mathrm{e}-02 & 5.9696\mathrm{e}-03 & 2.9135\mathrm{e}-02 \\ 5.9696\mathrm{e}-03 & 2.2089\mathrm{e}-02 & 3.3940\mathrm{e}-03 \\ 2.9135\mathrm{e}-02 & 3.3940\mathrm{e}-03 & 7.7004\mathrm{e}-01 \end{pmatrix}$

because the probability of three random Gaussian clusters having parallel principal axes within 3° is by chance as small as $(3/90)^2 \simeq 10^{-3}$.

The young pulsar ellipse 'C' is rotated clockwise by $\sim$15° with respect to the average principal axis directions of ellipses 'D', 'E' and 'F'.[2] Such a rotation may come from the selection effect that it is easier to find bright pulsars. Bearing this selection effect in

[2] Because of the unequal X–Y scales, the 15° rotation is distorted in Fig. 2 visually.

Table 3. A sample of the source list sorted according to our likelihood. In this list, 2FGL name and corresponding 1FGL name of the sources are given. The source associations are from the 2FGL and $\log(R_S)$ is our logarithmic likelihood. The whole sorted source list is available as supporting information with the online version of this paper.

Index	2FGL name	1FGL name	RA (J2000)	Dec. (J2000)	Association	$\log R_S$
1	2FGL J0633.9+1746	1FGL J0633.9+1746	06:33:55	+17:46:26	PSR J0633+1746	55.8
2	2FGL J0835.3−4510	1FGL J0835.3−4510	08:35:21	−45:10:45	PSR J0835−4510	52.6
3	2FGL J1801.3−2326e	–	18:01:22	−23:26:24	SNR G006.4−00.1	38.2
4	2FGL J1836.2+5926	–	18:36:16	+59:26:01	LAT PSR J1836+5925	33.0
5	2FGL J0007.0+7303	1FGL J0007.0+7303	00:07:06	+73:03:16	LAT PSR J0007+7303	30.0
1869	2FGL J0238.7+1637	1FGL J0238.6+1637	02:38:42	+16:37:26	AO 0235+164	−112
1870	2FGL J1229.1+0202	1FGL J1229.1+0203	12:29:06	+02:02:30	3C 273	−119
1871	2FGL J1512.8−0906	1FGL J1512.8−0906	15:12:50	−09:06:12	PKS 1510−089	−139
1872	2FGL J1224.9+2122	1FGL J1224.7+2121	12:24:54	+21:22:48	4C +21.35	−176
1873	2FGL J2253.9+1609	1FGL J2253.9+1608	22:53:59	+16:09:09	3C 454.3	−181

mind, this rotation may indicate that the old pulsars and the young pulsars have different radiation or evolution mechanisms, which is supported by both the timing and polarization properties. Timing results (Espinoza et al. 2011a) show that pulsars with glitch behaviours correlate with clusters 'C', while the polarization measurements (Weltevrede & Johnston 2008) indicate that the polarization behaviours for high- and low-$\dot{E}$ pulsars are different.

We also note that the ellipse for high magnetic field pulsars is quite extended instead of being localized to only high magnetic field pulsars. It covers the regions of both young pulsars and middle age pulsars. This indicates potential links between the high magnetic field populations and normal pulsars, for which some observational evidence already suggests that high magnetic field pulsars may evolve from young pulsars (Lin & Zhang 2004; Lyne 2004; Espinoza et al. 2011b).

3.2 Application to the classification of *Fermi* point sources

We now apply the GMM to the 2FGL catalogue to demonstrate the use of the GMM to rank candidates according to likelihood. As already pointed out by Abdo et al. (2010) and The Fermi-LAT Collaboration (2011a), one can use the variability and the spectrum information to classify *Fermi* point sources. In the 2FGL catalogue, the Variability_Index (VI) and Signif_Curve (significance of the fit improvement using a curved spectrum, Sc) are used to describe the variability and spectral shape. Fig. 3 shows that *Fermi* point sources form two distinctive classes, the pulsars and the AGs, where pulsars usually have smaller VI but larger Sc compared to AGs.

Although it is straightforward to classify these sources, one still needs to be careful when using parameters VI and Sc, since both VI and Sc correlate with the test statistics (see Abdo et al. 2010 for the definition). Such a correlation smears clustering structures at low detection significances. One way to deal with this situation is redefining VI and Sc to reduce the correlation (The Fermi-LAT Collaboration 2011a). Here, we take an alternative approach, in which the gamma-ray flux is included as an additional parameter to directly correct the correlation. In other words, instead of using only VI and Sc, we use three variables, VI, Sc and integral gamma-ray flux from 1 to 100 GeV, 'Flux1000' (F_{1000}) to classify data. Fig. 3 shows that pulsars and AGs have different dependences in gamma-ray flux, which helps us in data classification.

We now turn to the details of our setup for the GMM. The parameter space we used in the GMM algorithm is spanned by *Sc* and base-10 logarithms of VI and F_{1000}, i.e. the components of the data vector $\boldsymbol{x}$ are

$$\boldsymbol{x} = \begin{pmatrix} \log F_{1000} \\ \log \mathrm{VI} \\ \mathrm{Sc} \end{pmatrix} . \tag{11}$$

The distribution of $\boldsymbol{x}$ for the all 2FGL catalogue sources is plotted in Fig. 3. The GMM is then applied to the whole population to determine the model parameters. To test the robustness of the algorithm and to check whether the result is sensitive to initial values, we run the EM algorithm with randomly generated initial values as in the previous example. We get three Gaussian clusters, where one of the Gaussian clusters overlaps with the AGs, one overlaps with pulsars and one is for sources with low fluxes. The GMM parameters

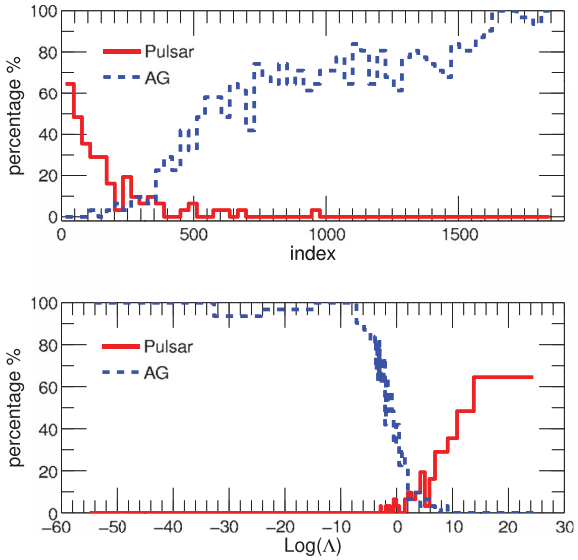

Figure 4. Upper panel: the proportion of pulsars and AGs as a function of the ranking index. We bin all the 2FGL catalogue sources into 60 bins according to the ranking index. The *x*-axis is the average index value and the *y*-axis is the proportion of pulsars/AGs in the bin. The solid line is for pulsars, while the dashed line is for AGs. One can see that our algorithm statistically ranks pulsars higher than AGs. Lower panel: the proportion of pulsars and AGs as a function of logarithms of pulsar likelihood. The *x*-axis is the average logarithm of pulsar likelihood, and the *y*-axis is the proportion of pulsars/AGs in each bin. Due to our definition for the likelihood, the proportion of pulsars and AGs equals each other at $\log R_S = 0$, as expected.

Table 4. A sample of pulsar candidates sorted according to the pulsar likelihood. The 2FGL name with † indicates that there is a known pulsar within the error ellipses of the pointing according to the ATNF pulsar catalogue. The SemiMajor and SemiMinor axes are the source positions at 68 per cent confidence in units of degrees. The column 'Class' denotes the possible association of the source, where 'SNR' denotes supernova remnant, 'GLC' denotes globular cluster and 'SPP' denotes special cases which may be associated with SNR or pulsar wind nebula (The Fermi-LAT Collaboration 2011b).

Index	2FGL name	1FGL name	RA J2000	Dec. J2000	SemiMajor (10^{-2} deg)	SemiMinor (10^{-2} deg)	log R_S	Class
1	2FGL J1801.3−2326e		18:01:22	−23:26:24	–	–	38.2	SNR
2	2FGL J1745.6−2858		17:45:42	−28:58:42	1.0	1.0	29.0	SPP
3	2FGL J1855.9+0121e		18:55:58	+01:21:18	–	–	27.5	SNR
4	2FGL J0617.2+2234e	1FGL J0617.2+2233	06:17:14	+22:34:47	–	–	27.1	SNR
5	2FGL J1906.5+0720	1FGL J1906.6+0716c	19:06:35	+07:20:33	3.5	2.9	18.2	
6	2FGL J1923.2+1408e		19:23:16	+14:08:42	–	–	18.0	SNR
7	2FGL J1045.0−5941	1FGL J1045.2−5942	10:45:00	−59:41:31	1.5	1.4	17.7	
8	2FGL J1704.9−4618		17:04:59	−46:18:14	15.9	11.1	16.6	
9	2FGL J0848.7−4324		08:48:45	−43:24:24	9.4	6.9	16.5	
10	2FGL J1738.9−2908		17:38:57	−29:08:24	15.1	6.6	15.7	SPP
11	2FGL J1819.3−1523	1FGL J1819.4−1518c	18:19:21	−15:23:29	7.4	5.5	15.5	
12	2FGL J1747.3−2825c		17:47:24	−28:25:52	3.6	3.2	15.4	
13	2FGL J1805.6−2136e		18:05:38	−21:36:42	–	–	15.4	SNR
14	†2FGL J1748.0−2447	1FGL J1747.9−2448	17:48:00	−24:47:04	2.3	2.2	14.8	GLC
15	2FGL J2018.0+3626		20:18:03	+36:26:54	3.7	3.1	14.3	
16	2FGL J1839.0−0539		18:39:04	−05:39:21	1.7	1.6	14.2	
17	2FGL J1901.1+0427		19:01:11	+04:27:27	7.1	5.7	14.1	
18	2FGL J1748.6−2913	1FGL J1748.3−2916c	17:48:39	−29:13:52	4.1	3.5	13.8	
19	2FGL J1932.1+1913	1FGL J1932.1+1914c	19:32:10	+19:13:25	4.2	3.8	13.8	SPP
20	2FGL J1847.2−0236	1FGL J1846.8−0233c	18:47:14	−02:36:40	7.0	4.6	13.4	
21	2FGL J1856.2+0450c		18:56:14	+04:50:16	8.0	6.4	13.3	
22	2FGL J0858.3−4333		08:58:20	−43:33:34	9.1	8.7	12.6	
23	2FGL J1521.8−5735	1FGL J1521.8−5734c	15:21:50	−57:35:53	3.4	3.1	12.4	SPP
24	2FGL J1625.2−0020	1FGL J1625.3−0019	16:25:13	−00:20:04	3.5	3.2	12.2	
25	2FGL J0224.0+6204	1FGL J0224.0+6201c	02:24:06	+62:04:35	3.5	3.1	12.2	
26	2FGL J1857.8+0355c	1FGL J1857.9+0352c	18:57:53	+03:55:29	10.0	6.9	11.9	
27	2FGL J1739.6−2726		17:39:40	−27:26:03	12.3	7.0	11.8	
28	2FGL J1619.0−4650		16:19:04	−46:50:48	26.4	15.3	11.5	
29	2FGL J1405.5−6121	1FGL J1405.1−6123c	14:05:30	−61:21:51	3.5	2.9	11.5	
30	2FGL J0842.9−4721		08:42:58	−47:21:53	7.4	7.1	11.4	SPP
31	2FGL J1814.1−1735c	1FGL J1814.0−1736c	18:14:09	−17:35:31	4.9	4.3	11.0	
32	2FGL J1636.3−4740c	1FGL J1636.4−4737c	16:36:22	−47:40:58	4.3	3.3	10.9	
33	2FGL J2022.8+3843c		20:22:50	+38:43:21	8.2	7.7	10.6	SNR
34	2FGL J1714.5−3829	1FGL J1714.5−3830c	17:14:31	−38:29:32	2.8	2.3	10.5	SPP
35	2FGL J1056.0−5853		10:56:00	−58:53:16	8.4	7.2	10.5	
36	2FGL J1911.0+0905	1FGL J1910.9+0906c	19:11:03	+09:05:38	1.6	1.5	10.4	SNR
37	2FGL J1638.0−4703c		16:38:03	−47:03:10	3.9	3.2	10.3	
38	2FGL J1536.4−4949	1FGL J1536.5−4949	15:36:30	−49:49:45	1.7	1.6	10.2	
39	2FGL J1628.1−4857c		16:28:11	−48:57:36	11.8	6.7	10.1	SPP
40	2FGL J1311.7−3429	1FGL J1311.7−3429	13:11:46	−34:29:19	2.1	2.0	10.1	
41	2FGL J1650.6−4603c	1FGL J1651.5−4602c	16:50:36	−46:03:16	3.5	3.1	10.1	
42	2FGL J1112.1−6040	1FGL J1112.1−6041c	11:12:07	−60:40:17	2.1	2.0	9.9	SPP
43	2FGL J0608.3+2037	1FGL J0608.3+2038c	06:08:20	+20:37:55	7.0	6.4	9.9	
44	2FGL J1615.0−5051		16:15:02	−50:51:06	5.4	4.6	9.9	SPP
45	†2FGL J1740.4−3054c	1FGL J1740.3−3053c	17:40:25	−30:54:41	10.1	6.2	9.8	SPP
46	2FGL J0340.5+5307	1FGL J0341.5+5304	03:40:36	+53:07:52	9.0	7.6	9.7	
47	2FGL J2033.6+3927	1FGL J2032.8+3928	20:33:39	+39:27:05	7.2	5.6	9.7	
48	2FGL J1914.0+1436		19:14:05	+14:36:15	9.6	7.8	9.6	
49	2FGL J1027.4−5730c		10:27:27	−57:30:39	5.4	4.7	9.6	
50	2FGL J1843.7−0312c		18:43:43	−03:12:56	8.1	5.8	9.5	

turn out to be insensitive to initial values. We note that extra Gaussian components are needed to model the details of the low-flux branch to pass the multidimensional K-S test. Such a requirement is simply due to the fact that Sc $\geq$ 0, where such a boundary needs more Gaussian components to approximate the distribution. We have also checked that including these extra components will not significantly alter our results of source classification; thus, for simplicity we prefer to use only three Gaussian clusters here. The projected 3σ contours of clusters are shown in Fig. 3. The GMM parameters, i.e. $\boldsymbol{\mu}$ and Σ, are given in Table 2.

Our pulsar likelihood is calculated using equation (8) with $\boldsymbol{\mu}_k$ and Σ_k from Table 2. We identify the pulsar subset $\mathcal{S}$ as the pulsar cluster; the complementary sets are the AG and low-flux clusters. We exclude the low-flux cluster from subset $\mathcal{S}$ to suppress the candidates with low detection significances.

We sort the 2FGL catalogue point sources according to the pulsar likelihood. The whole sorted source list is available as supporting information with the online version of this paper. The first five and the last five sources are given as examples in Table 3.

In order to verify the validity of our ranking technique, we checked the ranking index of known pulsars and other known sources in our list. To identify pulsars, we used the 'Public List of LAT-Detected Gamma-Ray Pulsars',[3] which lists the pulsars that have been detected as pulsed gamma-ray sources with the Fermi-LAT up to now. For AGs, the association information is from the 2FGL catalogue, i.e. the 'CLASS1' entry. Such a comparison is legitimate, since we do not use any information about known pulsar and AG populations in the process of finding the model parameters. The comparisons between the ranking results and known populations are shown in Fig. 4. We can see that the ranking technique separates the pulsar population and AG population. The proportion of pulsars decreases and the proportion of the AGs increases with the ranking index. The top 5 per cent sources contain 50 per cent known pulsars, the top 50 per cent contain 99 per cent known pulsars and no known AG appears in the top 6 per cent. As predictions, we list the potential pulsar candidates in Table 4, which according to our likelihood are highly pulsar-like objects with no associated sources in the 2FGL catalogue. We hope future pulsar searching will benefit from this information.

4 CONCLUSION AND DISCUSSION

In this paper, we reviewed the GMM and its application in data modelling and classification. As examples, we apply it to pulsar classification in the P–$\dot{P}$ diagram as well as modelling and ranking the 2FGL catalogue point sources.

In the application of the GMM to model pulsar populations, we find that the pulsar distribution in the P–$\dot{P}$ diagram should be described by six Gaussian clusters. Based on the GMM, we present a rule to separate the MSP population from the normal pulsar population. As caveats, the six Gaussian clusters from the GMM algorithm may be artefacts due to the requirement of approximating a non-Gaussian distribution of pulsars. Such a non-Gaussian distribution can be induced by selection effects, pulsar evolution, star formation history, etc. However, these six Gaussian components coincide with other observational evidence, e.g. chemical composition of companion, radiation and polarization properties, timing behaviours, etc. Although it is still far from drawing any solid conclusion, those clusters found by the GMM algorithm may imply the following: (i) there are two different MSP populations with different evolution scenarios, which are supported by the evidence that one cluster contains mainly CO white dwarf companions and the other mainly He white dwarfs companions. (ii) There are two possible different groups of normal pulsars. (iii) High magnetic field pulsars may come from the evolution of normal pulsars. (iv) Although there are different formation channels, recycled pulsars appear to form a continuum in the P–$\dot{P}$ diagram. Hence, the spectrum of accreted mass for both fully recycled and mildly recycled MSPs should be smooth; otherwise we would identify more clusters.

In the application to the 2FGL catalogue, we use a 3D parameter space spanned by VI, Sc and F_{1000}. We found that the distribution can be well described by three Gaussian clusters, two of which correspond to pulsar and AG populations. The remaining cluster contains sources with lower detection significances. Using the GMM, we calculate the pulsar likelihood for each source and sort the 2FGL catalogue accordingly. In the sorted list, we find that the top 5 per cent sources contain 50 per cent known pulsars, the top 50 per cent contain 99 per cent known pulsars and no known AG appears in the top 6 per cent. Clearly this algorithm has the ability to prioritize the follow-up searching observation scheme to find new pulsars. The statistical behaviour of the sorted list is given in Fig. 4 and a sample with the first and the last five sources of the sorted list is presented in Table 3. In Table 4, we also present a list of unassociated 2FGL catalogue sources with high pulsar likelihood for future pulsar searching projects.

We include the gamma-ray flux as one dimension of our parameter space, although most of the discrimination comes from VI and Sc. The reason to do so is that both VI and Sc correlate with the test statistics. By introducing the gamma-ray flux, we can, to a certain degree, correct such a correlation. Instead of Flux1000 (the integral flux from 1 to 100 GeV), we have also tried using other flux measurements available in the 2FGL catalogue. In particular, we have experimented using Energy_Flux100 (the energy flux from 100 MeV to 100 GeV obtained by spectral fitting), which yields a ranking result with a slightly lower detection rate compared to the result using Flux1000. We have tried with 4D GMMs. In these experiments, various photon fluxes, energy fluxes, test statistics of different energy bands and Galactic coordinates were tried as the fourth dimension. The classification results show little differences. This is mainly due to the fact that most other variables available in the 2FGL catalogue are correlated with the Flux1000, and thus do not provide much extra information to improve the classification.

There are alternatives to the unsupervised machine learning techniques, where one uses supervised machine learning (The Fermi-LAT Collaboration 2011a; Miraball et al. 2012). One particular benefit of being unsupervised is that, since we use no prior information of source associations in our ranking algorithm, we can investigate the statistics quality of the algorithm easily by comparing the results with known pulsar and AG populations, as shown in Fig. 4.

We note that the GMM algorithm detects three clusters in the F_{1000}–VI–Sc parameter space, one of which is in the confused region with low gamma-ray fluxes. Such a component is likely to be an artefact due to the low signal-to-noise statistic.

ACKNOWLEDGMENTS

We gratefully acknowledge support from ERC Advanced Grant 'LEAP', grant agreement number 227947 (PI: MK). YLY is supported by the National Natural Science Foundation of China (grant 11103045), the National Basic Research Program of China (grant 2012CB821800) and the Young Researcher Grant of National Astronomical Observatories, Chinese Academy of Sciences. We thank Ramesh Karuppusamy and Thomas Tauris for reading the manuscript and their very helpful comments. We also appreciate the valuable inputs and help from the anonymous referee.

[3] The list is available at https://confluence.slac.stanford.edu/display/GLAMCOG/Public+List+of+LAT-Detected+GammaRay+Pulsars

REFERENCES

Abdo A. A. et al., 2010, ApJS, 188, 405
Ackermann M., 2012, ApJ, 753, 83
Bhattacharya D., van den Heuvel E. P. J., 1991, Phys. Rep., 203, 1
Eatough R. P., Molkenthin N., Kramer M., Nartos A., Keith M. J., Stappers B. W., Lyne A. G., 2010, MNRAS, 407, 2443
Espinoza C. M., Lyne A. G., Stappers B. W., Kramer M., 2011a, MNRAS, 414, 1679
Espinoza C. M., Lyne A. G., Kramer M., Manchester R. N., Kaspi V. M., 2011b, ApJ, 741, L13
Fasano G., Franceschini A., 1987, MNRAS, 225, 155
Kassam S. A., 1988, Signal Detection in Non-Gaussian Noise. Springer-Verlag, New York
Lin J. R., Zhang S. N., 2004, ApJ, 615, L133
Lyne A. G., 2004, in Camilo F., Gaensler B. M., eds, Proc. IAU Symp. 218, Young Neutron Stars and Their Environments. Astron. Soc. Pac., San Francisco, p. 257
Manchester R. N., Hobbs G. B., Teoh A., Hobbs M., 2005, AJ, 129, 1993
Nolan P. L., 2012, ApJS, 199, 31
Peacock J. A., 1983, MNRAS, 202, 615
Press W. H., Teukolsky S., Vetterling W. T., Flannery B. P., 2007, Numerical Recipes: The Art of Scientific Computing. Cambridge Univ. Press, Cambridge
Shklovskii I. S., 1970, SvA, 13, 562
Tauris T. M., 2011, in Schmidtobreich L., Schreiber M. R., Tapperb C., eds, ASP Conf. Ser. Vol. 447, Five and a Half Roads to Form a Millisecond Pulsar. Astron. Soc. Pac., San Francisco, 285
Theodoridis S., Koutroumbas K., 2009, Pattern Recognition. Academic Press, Burlington, MA
Weltevrede P., Johnston S., 2008, MNRAS, 391, 1210

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this paper:

Table 3. The source list sorted according to our likelihood.

Please note: Wiley-Blackwell are 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 of the paper.

This paper has been typeset from a TEX/LATEX file prepared by the author.