Applying Gaussian mixture models to pulsar classification and candidate selection | A review of Lee et al. (2012)

March 5, 2025

Mark Beyer Stjerne[1], Ingrid Almquist Lien[1]

[1]Niels Bohr Institute, University of Copenhagen, DK-2100 Copenhagen, Denmark

Abstract

Gaussian mixture models (GMMs) provide an unsupervised approach to classifying astrophysical objects without predefined selection criteria. This piece reviews their application to pulsar classification and candidate selection, as demonstrated by Lee et al. (2012). GMMs are used to define an empirical classification for millisecond pulsars (MSPs) from the period-period derivative ($P - \dot{P}$) distribution and to rank gamma-ray sources from the Fermi 2FGL catalog by their likelihood of being pulsars. The Expectation-Maximization algorithm determines optimal cluster parameters, which are then validated via a multidimensional Kolmogorov-Smirnov test. While effective, the assumption of Gaussian-distributed clusters may introduce biases, requiring further observational validation.

1. Introduction

In observational astronomy, a common task is to classify objects by their physical properties without *a priori* assumptions of the required selection criteria to do so. Examples include identifying active galaxy candidates from an optical photometry survey, or selecting desirable pulsar candidates from a pulsar search survey. In order to empirically construct selection criteria, machine learning models (MLMs) can be employed to extract this information from the data.

Lee et al. (2012) presents the use of Gaussian mixture model (GMMs), a form of unsupervised MLM (one that identifies patterns in data without predefined labels or categories) based on Bayesian decision theory, to determine the criteria required to select desirable pulsar candidates in two application examples. The GMM is first used to derive an empirical definition for millisecond pulsars (MSPs) from the period-period derivative ($P - \dot{P}$) distribution of known pulsars. It is then applied to rank the likelihood of a gamma-ray point-source being a pulsar and to generate a ranked pulsar candidate list, using data from the *Fermi* gamma-ray Space Telescope Large Area Telescope 2-year Point Source Catalog (2FGL catalog).

2. Method

The authors introduce the concept of GMM as well as related data classification techniques. The GMM is used to identify a set of m clusters in an n-dimensional parameter space that best represent the observed data distribution, assuming that each cluster consists of a multivariate Gaussian distribution (Press, 2007). The authors describe the probability distribution $P(x)$ of data $\boldsymbol{x}$ (of index $i \dots N$) as the weighted sum of the m Gaussian clusters:

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

where $k = 1 \dots m$. $P(k)$ is defined as the mixture weight, and $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ are the mean and covariance matrices of the k-th Gaussian cluster respectively. It follows that the density distribution of an individual Gaussian cluster is:

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

where $|\boldsymbol{\Sigma}_k|$ is the determinant of $\boldsymbol{\Sigma}_k$. Using Bayes' theorem, the posterior probability (responsibility matrix) is given by:

$$P(k|\boldsymbol{x}_n) = \frac{P(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) P(k)}{P(\boldsymbol{x}_n)}. \tag{3}$$

The likelihood Λ describing the product of the probabilities of finding a point at each observed position is:

$$\Lambda = \prod_{i=1}^{N} P(\boldsymbol{x}_i). \tag{4}$$

The optimal values for the parameters are found by maximising the likelihood Λ. This is akin to maximising the posterior probability of the parameters, given a set of broad priors.

The parameters $\boldsymbol{\mu}_k$, $\boldsymbol{\Sigma}_k$, and $P(k)$ can be determined from the data by means of the Expectation-Maximisation (EM) algorithm, which assumes no prior knowledge of clustering structures in the data:

1. Initial values for $\boldsymbol{\mu}_k$, $\boldsymbol{\Sigma}_k$ and $P(k)$ are guessed.
2. Expectation step (E-step): $P(\boldsymbol{x})$ and $P(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ are calculated using Equations 1 and 2.
3. Maximisation step (M-step): Model parameters are updated using:

$$\boldsymbol{\mu}_{k,\,\text{new}} = \frac{\sum_{i=1}^{N} \boldsymbol{x}_i P(k|\boldsymbol{x}_i)}{\sum_{i=1}^{N} P(k|\boldsymbol{x}_i)} \tag{5}$$

$$\boldsymbol{\Sigma}_{k,\,\text{new}} = \frac{\sum_{i=1}^{N} (\boldsymbol{x}_i - \boldsymbol{\mu}_k) \otimes (\boldsymbol{x}_i - \boldsymbol{\mu}_k) P(k|\boldsymbol{x}_i)}{\sum_{i=1}^{N} P(k|\boldsymbol{x}_i)} \tag{6}$$

$$P_{k,\,\text{new}} = \frac{1}{N} \sum_{i=1}^{N} P(k|\boldsymbol{x}_i) \tag{7}$$

4. The EM steps are repeated until the total likelihood Λ (Equation 4) converges.

The GMM is used to infer the strength of the association between a data point $\boldsymbol{x}$ and a Gaussian cluster. Once the clusters are defined, one can quantify the association of these clusters to a subset $\mathcal{S}$ containing m_0 of the m total clusters. This is done using the likelihood ratio test, supported by the Neyman-Pearson lemma, which stipulates

that one should compare the logarithmic likelihood ratio $\log R_{\mathcal{S}}$ against a statistical decision threshold η, i.e. choose H_0 ($\boldsymbol{x}$ belongs to other clusters) if $\log R_{\mathcal{S}} > \eta$, otherwise choose H_1 ($\boldsymbol{x}$ belongs to subset $\mathcal{S}$). According to the GMM, the logarithmic likelihood ratio $R_{\mathcal{S}}$ is defined as:

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

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

In order to prevent over-fitting, the authors employ a multi-dimensional Kolmogorov-Smirnov (K-S) test to compare the model prediction to the observed distribution of data, by calculating a test statistic $\mathcal{D}$ which describes the maximal difference between the model and the observed data. The statistical threshold and p-values are calculated numerically via Monte-Carlo methods, where mock datasets are generated by resampling from the observed data with replacement. The null-hypothesis distribution of $\mathcal{D}$ is constructed by applying the K-S test to the resampled datasets.

In summary, the authors describe the GMM technique to determine the data distribution as follows:

1. The parameter space is determined and the data vector $\boldsymbol{x}$ is formulated.
2. The number of clusters m is guessed as well as their initial parameters ($\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$).
3. The EM algorithm is used to determine the best fit model parameters for the set of Gaussian clusters.
4. The GMM predicted model is tested using a multi-dimensional K-S test. If the test fails, the number of Gaussian clusters is increased and the GMM is retested.
5. The likelihood ratio test (Equation 8) is applied to classify whether a data point belongs to a specific subset of clusters.

3. Applications

3.1. Classification of millisecond pulsars

The first application is the classification of pulsars by modelling clusters in the $P-\dot{P}$ diagram. The authors employ GMM to model the distribution of pulsars in the 2-dimensional parameter vector space:

$$\boldsymbol{x} = \begin{pmatrix}\log P \\ \log \dot{P}\end{pmatrix}. \quad (9)$$

The authors directly apply the GMM to the dataset. To test the stability of the computed parameters, the authors apply a bootstrap-like resampling method which generates 100 simulated datasets by randomly selecting data points with replacement. The EM algorithm is then re-applied to each dataset to evaluate the robustness of the estimated cluster parameters. The optimal number of clusters is determined using the K-S test, for which a p-value of 95% is chosen. The results indicate that the best-fit model consists of six clusters (shown in Figure 1): Two corresponding to MSPs, and four corresponding to the normal pulsar population.

The likelihood ratio test is used to distinguish MSPs from normal pulsars, by plotting the equal likelihood ratio contours corresponding to $R_{\mathcal{S}} = \eta$, dividing the region of clusters into two subsets, allowing an empirical classification of MSPs. The authors define MSPs as pulsars that satisfy $R_{\mathcal{S}} \geq 0$, which translates into the following boundary condition:

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

This derived relation serves as a novel classification criteria for MSPs based on period and period derivative data, offering an alternative to previous methods that relied on additional astrophysical parameters.

A caveat of this method is that the six clusters identified may be artifacts resulting from approximating a non-Gaussian distribution. This could arise from selection effects in pulsar surveys, where observational biases may influence the detected population, leading to artificial clustering. However, the authors claim that the clusters highlight features in both pulsar populations which are supported by observational evidence.

3.2. Classification of point sources in *Fermi* catalog

In the second application, the authors use GMM on data from the Fermi 2FGL catalog to assess and rank the likelihood of unidentified gamma-ray point sources being pulsars. A 3-dimensional parameter space is established:

$$\boldsymbol{x} = \begin{pmatrix}\log F_{1000} \\ \log VI \\ Sc\end{pmatrix}, \quad (11)$$

where F_{1000} is the integral gamma ray flux, VI is the variability index (which measures the stability of flux over time), and Sc is the significance of the fit improvement for a curved spectrum. The F_{1000} is included in the space to correct for the correlation between VI and Sc. The GMM is applied to model the distribution of sources in this space, and three clusters are identified (shown in Figure 2), corresponding to pulsars, active galaxies (AGs) and low-flux sources.

To find the probability of a source in the low-flux cluster being a pulsar, the likelihood ratio test is applied. This compares the probability of a source belonging to the pulsar cluster versus the AG cluster. The sources are then sorted by their probability of belonging to a pulsar cluster. A comparison between the ranking results and known population is shown in Figure 3. The results show that the top 5% of sources contain 50% known pulsars, the top 50% of sources contain 99% known pulsars and no AG appears in the top 6% of sources. This ranking provides a prioritized list of pulsar candidates as a help for follow-up searches.

4. Conclusion

Lee et al. (2012) effectively demonstrates the use of Gaussian mixture models for distribution modelling and source classification in pulsar astronomy. By modelling pulsar and gamma-ray source distributions as multivariate Gaussian distributions, the authors distinguish between populations of astrophysical objects without predefined selection criteria.

For pulsar classification, GMM identifies six pulsar population clusters. The Kolmogorov-Smirnov test is used to validate the model, and the likelihood ratio test is applied to determine the empirical definition of millisecond pulsars. For source likelihood ranking, GMM is applied to the *Fermi* 2FGL catalog distinguish pulsars from active galaxies and ranking potential pulsars based on likelihoods. The method may eventually prove successful in furthering the search for new pulsars.

The primary caveat of this method is the possible artifacts associated with assuming a Gaussian distribution for the clusters, which may not represent the real data. Therefore, further comparison with the observational evidence is required to establish a justifiable link between the model prediction and the observed data.

5. References

Lee, K. J., Guillemot, L., Yue, Y. L., Kramer, M., & Champion, D. J. (2012). Application of the gaussian mixture model in pulsar astronomy - pulsar classification and candidates ranking for the fermi 2fgl catalogue. *Monthly Notices of the Royal Astronomical Society*, *424*(4), 2832–2840. https://doi.org/10.1111/j.1365-2966.2012.21413.x

Press, W. H. (2007). *Numerical recipes: The art of scientific computing*. Cambridge University Press.

6. Appendix

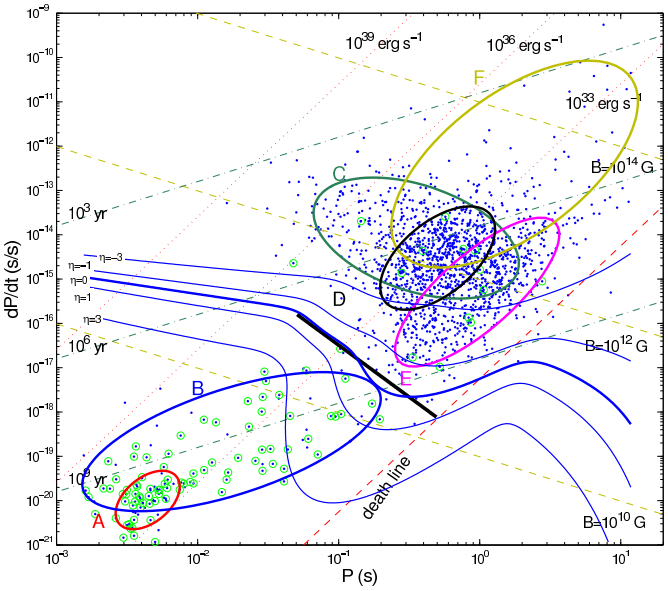

Figure 1: *Pulsar $P - \dot{P}$ diagram with elliptical 2σ contours for Gaussian clusters. Dots are pulsars and those circled in green are binaries. Blue curves with numerical labels on left are isotopic contours of $\log R_{\mathcal{S}}$. The dotted straight lines correspond to equal magnetic field strength, spin-down power, characteristic age, and lastly the death line.* ***Source****: Lee et al. (2012)*

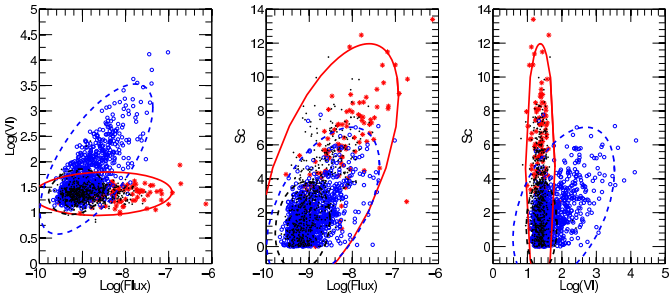

Figure 2: *Projected 3σ contours of Gaussian clusters and source distributions. Blue '$\circ$' symbols are AGs, red '$*$' symbols are pulsars, unassociated sources are plotted with black solid dots.* ***Source****: Lee et al. (2012)*

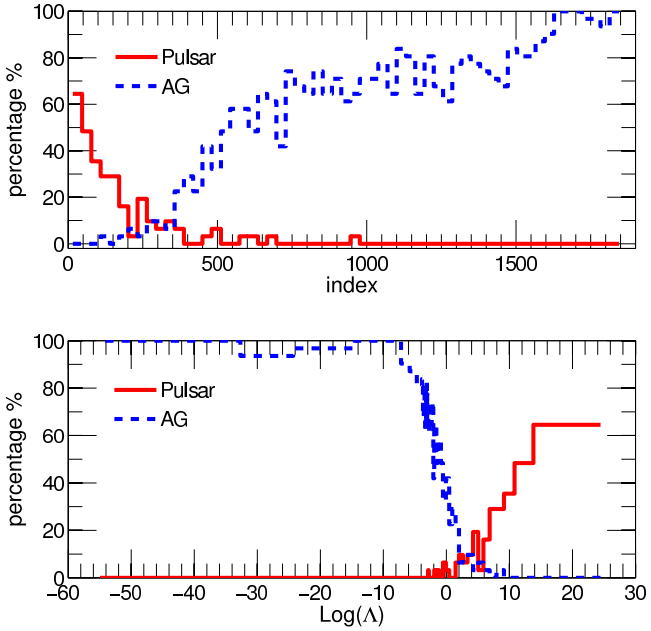

Figure 3: ***Upper panel:*** *Proportion of pulsars and AGs as a function of ranking index, $n_{bins} = 60$. The solid line represents pulsars, while the dashed line represents AGs. The x-axis is the average index value, the y-axis is the proportion of pulsars/AGs in the bin. One can see that the EM-algorithm statistically ranks pulsars higher than AGs.* ***Lower panel:*** *The proportion of pulsars and AGs as functions 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 the author's definition for the likelihood, the proportion of pulsars and AGs equals to each other are at $\log R_{\mathcal{S}} = 0$, as expected.* ***Source****: Lee et al. (2012)*