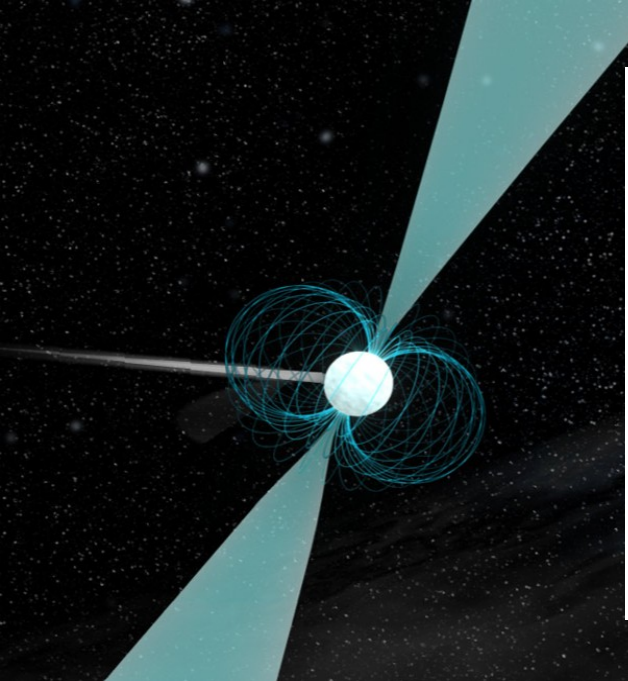


Applications of Gaussian mixture models to pulsar astronomy

A review of Lee et al. (2012)

Mark Beyer Stjerne, Ingrid Almquist Lien
March 6, 2025

UNIVERSITY OF COPENHAGEN

Article

[Submitted on 28 May 2012]

Application of the Gaussian mixture model in pulsar astronomy -- pulsar classification and candidates ranking for {\it Fermi} 2FGL catalog

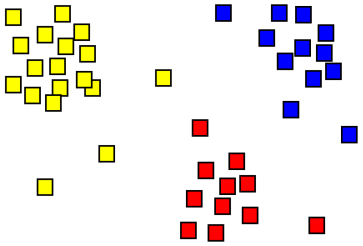
K. J. Lee, L. Guillemot, Y. L. Yue, M. Kramer, D. J. Champion

Machine learning, algorithms 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. Firstly, it is applied to the well known period-period derivative diagram, where we find that the pulsar distribution can be modeled 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 $\frac{\dot{P}}{10^{-17}} \leq 3.23(\frac{P}{100\mathrm{ms}})^{-2.34}$. The two millisecond pulsar clusters may have different evolutionary origins, since the companion stars to these pulsars in the two clusters show different chemical composition. 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 \textit{Fermi} point sources being pulsars and rank them accordingly. In the ranked point source list, the top 5% sources contain 50% known pulsars, the top 50% contain 99% known pulsars, and no known active galaxy (the other major population) appears in the top 6%. Such a ranked list can be used to help the future follow-up observations for finding pulsars in unidentified \textit{Fermi} point sources.

Comments: 9 pages, 4 figures, accepted by MNRAS
Subjects: **Instrumentation and Methods for Astrophysics (astro-ph.IM)**; High Energy Astrophysical Phenomena (astro-ph.HE)
Cite as: arXiv:1205.6221 **[astro-ph.IM]**
(or arXiv:1205.6221v1 **[astro-ph.IM]** for this version)
https://doi.org/10.48550/arXiv.1205.6221
Related DOI: https://doi.org/10.1111/j.1365-2966.2012.21413.x

The problem setup

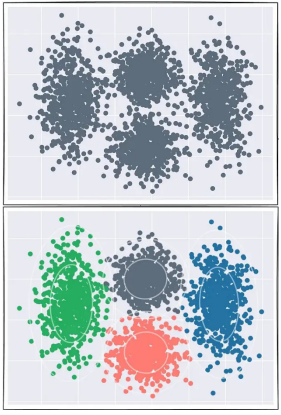
- Idea: Astrophysical objects are described by their properties in multi-dimensional datasets (e.g. position, velocity, brightness).
- Task: Classify these objects into distinct groups based on these properties.
- Issue: What if we don't know the appropriate selection criteria beforehand?



We can do this by eye;
How do we do it by computer?

Solution: The Gaussian mixture model (**GMM**)

- A probabilistic model that belongs to a family of methods within cluster analysis
- Fits the observed data to 'clusters', each shaped by a multivariate Gaussian distribution
- Unsupervised machine learning model (MLM) $\longrightarrow$ No prior knowledge of structures in data



GMM in a nutshell

- Q: Given N points in an m-dimensional space, can we find K multivariate Gaussians to describe the data?
- We want to know:
  - $\boldsymbol{\mu}_k$: The mean of cluster k (location)
  - $\boldsymbol{\Sigma}_k$: The covariance matrix of cluster k (shape)
  - $P(k|\boldsymbol{x}_i)$: The probabilities of x_i being a datapoint that stems from the k-th cluster
    - The responsibility matrix, i.e. "How *responsible* is a cluster k for a point $\boldsymbol{x}_i$?"
- Bringing this all together:

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

  - $P(k)$ is the *mixture weight* of the k-th cluster (contribution to the whole dataset), and $P(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ is a multivariate Gaussian density
- We can decompose (1) with Bayes' theorem to get the probability contribution from each cluster:

$$P(k|\boldsymbol{x}_n) = \frac{P(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) P(k)}{P(\boldsymbol{x}_n)}. \tag{2}$$

GMM in a nutshell (part deux)

- We define the *likelihood* based on the parameter set $\{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, P(k)\}$:

$$\Lambda = \prod_{i=1}^{N} P(\boldsymbol{x}_i) \tag{3}$$

- When the maximised Λ converges, the best 'fit' parameters for the clusters are found
  - Imagine this as maximizing the posterior probability of the parameters, given broad priors

- Next question: How do we find the best parameters?

Expectation-Maximimisation (**EM**) algorithm

- An iterative algorithm that evaluates the likelihood with the current parameters and updates them in the optimal direction.
- It can be shown that the EM-algorithm always converges to a max likelihood

1. Guess values for $\boldsymbol{\mu}_k$, $\boldsymbol{\Sigma}_k$ and $P(k)$.
2. Expectation step (E-step): Calculate $P(\boldsymbol{x})$ and $P(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$
3. Maximisation step (M-step): Update model parameters $\boldsymbol{\mu}_{k,\text{ new}}$, $\boldsymbol{\Sigma}_{k,\text{ new}}$ and $P(k)_{\text{new}}$
4. Repeat EM steps until maximum Λ converges.

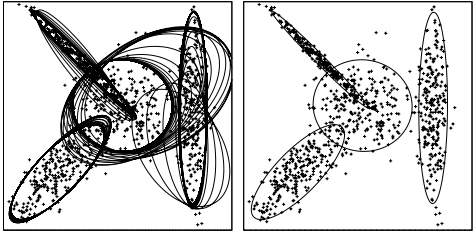


Figure: Convergence of $K = 4$ clusters onto $N = 1000$ datapoints in $m = 2$ dimensions.

Validating identified clusters

- Once the clusters are defined, the likelihood ratio test is used to see if a point falls within a subset of clusters $\mathcal{S}$ (group of objeccts of interest).
  - Done by computing log-likelihood ratio $\log R_{\mathcal{S}}$ and using a threshold η defined by Pearson-Neyman lemma.
  - If $\log R_{\mathcal{S}} > \eta$, it's within the subset (i.e. its object A)
  - If $\log R_{\mathcal{S}} \leq \eta$, it is not (i.e. its object B)
- To check for overfitting, the Kolmogorov-Smirnov (K-S) test is used to compare between the model and observed data, and returns a test statistic to quantify the difference.
  - The authors use a p-value of 95% to test the GMM.

Summary of methodology

1. Formulate the data vector $\boldsymbol{x}$ from the parameter space.
2. Guess the number of clusters m as well as their initial parameters.
3. Determine the best fit model parameters for the set of Gaussian clusters using the EM-algorithm.
4. Test the model predicted by the GMM using a multi-dimensional K-S test. If the test fails, increase the number of Gaussian clusters and retry.
5. Apply the likelihood ratio test classify whether a data point belongs to a specific subset of clusters.

Application 1: Classifying pulsar populations

- The GMM is applied to the $P - \dot{P}$ diagram to identify distinct groups in pulsar populations.
  - Identifying between millisecond pulsars (MSPs) and normal pulsar populations.
- Approximation of overall distribution.
- Distribution of pulsars defined in a 2-dimensional vector space:

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

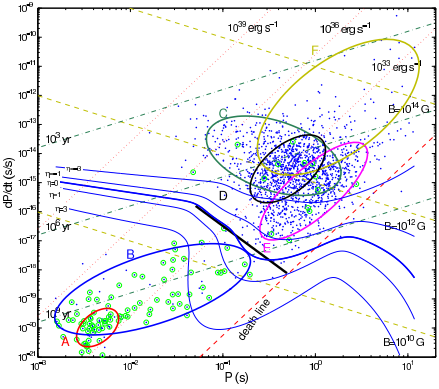


Figure: The pulsar $P - \dot{P}$ diagram with 2σ cluster contour lines.

- Expectation-Maximization algorithm to robustly estimate the cluster parameters
- Kolmogorov-Smirnov test to validate fitting number of clusters
- Likelihood ratio test used to define empirical formula for MSPs:

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

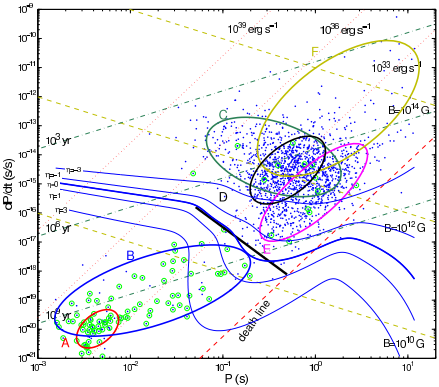


Figure: The pulsar $P - \dot{P}$ diagram with 2σ cluster contour lines.

Application 2: Classification of *Fermi* gamma-ray sources

- GMM applied to *Fermi* 2FGL catalog data to distinguish pulsars from active galaxies (AGs).
- Data vector in a 3-dimensional space:

$$\boldsymbol{x} = \begin{pmatrix} \log F_{1000} \\ \log VI \\ Sc \end{pmatrix}$$

- Identifies three clusters: pulsars, AGs, and low-flux sources.
- Likelihood ratio test ranks sources by pulsar probability.
- The top 5% of sources contain 50% known pulsars, and the top 50% contain 99% known pulsars. The top 6% contain no AGs.

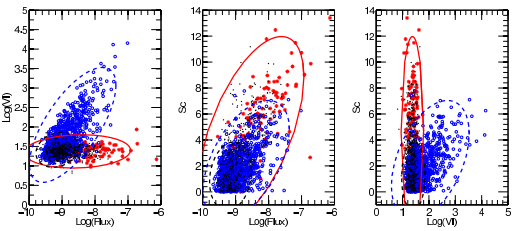


Figure: Source data distribution and projected 3σ Gaussian cluster contours.

Discussion

- Major challenge: Clusters may be artifacts from approximating a non-gaussian distribution.
  - Data may appear Gaussian due to observational selection effects.
  - In both applications the authors discussed the GMM results in the context of observational evidence.
- Benefit of an unsupervised method is that it can be clustered with predefined knowledge of the data's structure.
  - Allows for complex cluster shapes and overlapping distributions to be captured.

Conclusion

- GMMs were applied to distribution modelling and classification in pulsar astronomy.
- They proved efficient in classifying pulsars by physical properties, distinguishing pulsars from other objects, and ranking potential pulsars based on likelihoods.
- Downside: Clusters may not be real Gaussian clusters, but a mirage
  - Further consultation with external observations needed to establish link between GMM prediction and observed data.