Performance of Gaussian mixture model clustering algorithms on simulated PI-ICR mass spectrometry data

[redacted][1], [redacted][1,2]

[1] *Physics Division, Argonne National Laboratory, Lemont, Illinois, 60439, United States*

[2] *Department of Physics and Astronomy, [redacted]*

April 30, 2021

Abstract. The development of the phase-imaging ion-cyclotron resonance (PI-ICR) technique for use in Penning trap mass spectrometry (PTMS) increased the speed and precision with which PTMS experiments can be carried out. In PI-ICR, data sets of ion hit locations on the detector, which map linearly to their locations in the trap, are created showing how ions cluster together into spots according to their cyclotron frequency. However, in a given data set there may be multiple spots, distorted spots, or significant noise, all of which can make discerning the spot locations non-trivial. A method for assigning the ions to their respective spots and calculating the spot locations is therefore essential for success in PI-ICR experiments. We present the class of Gaussian mixture model (GMM) clustering algorithms as an effective solution. We show that on simulated PI-ICR data, several types of GMM clustering algorithms perform better than other clustering algorithms over a variety of typical scenarios encountered in PI-ICR.

1 Introduction

Penning trap mass spectrometry (PTMS) provides increasingly high-precision nuclear mass data for studies in nuclear structure, nuclear astrophysics, and fundamental physics.[1] For example, PTMS experiments have been performed to measure the masses of neutron-rich nuclei, many of which participate in the rapid neutron-capture process (r process) of nucleosynthesis via which half of the elements in the universe heavier than iron are created.[2] It has recently been discovered that the r process can occur when neutron stars merge,[3,4] yet other possible astrophysical r-process sites, such as core-collapse supernovae, are being investigated as well.[5] Mumpower *et al.* model masses necessary to produce the observed r-process abundances for a given astrophysical scenario,[6] meaning that measurements of r-process nuclear masses to sufficiently high precision can provide evidence as to where in the universe the r process takes place.

In general, PTMS experiments involve guiding beams of ions into a Penning trap, where a magnetic field provides radial confinement and an electric field provides longitudinal confinement. The trapped ions can then be excited to their different eigenmotions by manipulating the electric field. By measuring the eigenmotions, the cyclotron frequency, and hence the mass, of an ion species can be deduced.

PTMS experiments have become even more precise within the last decade with the development of the phase-imaging ion-cyclotron resonance (PI-ICR) technique at the SHIPTRAP facility,[7–9] allowing for mass measurements of short-lived nuclei to a precision of up to $\delta m/m = 1.4 \times 10^{-9}$.[10] In PI-ICR experiments, an ion's cyclotron frequency ν_c is measured in a trap using the relation

$$\nu_c = \frac{2\pi N + \Delta\phi}{2\pi t_{acc}} \tag{1}$$

where t_{acc} is the duration over which the ion completes N complete revolutions and accumulates an excess phase of $\Delta\phi$ (Fig. 1). An ion's mass can thus be deduced by measuring only its excess phase, which is done by projecting the ions onto a position-sensitive microchannel plate (PS-MCP) detector.[7–9,11–13]

The data analysis procedure in a PI-ICR experiment begins with the locations of the individual ion hits on the detector. Because the locations on the detector are linearly related to locations in the trap, data sets can be thought of as a representation of the locations of ions in the trap (Fig. 2a). Next, manual data cuts are made to highlight specific temporal and spatial regions of the data and obtain a clean signal[12–14]. Finally, because ions of the same cyclotron frequency appear as "clusters" in the data set, the resulting data is fit to a model to determine the centers of the clusters, which can be translated to the necessary phase measurement for ion species.

If the beam is composed of only one species, then the fitting model could be, for example, two univariate Gaussians[8,12], or one two-dimensional multivariate Gaussian[14], although Karthein *et al.* showed that for simulated data the two-dimensional multivariate Gaussian model reduces uncertainty in the ions' phase determination[13]. They also showed with simulated data that fitting a univariate Gaussian model to only the phase dimension of the ion positions in polar coordinates reduces the uncertainty by up to a factor of 10. However, in practice a beam is rarely pure enough to use these methods exclusively. With

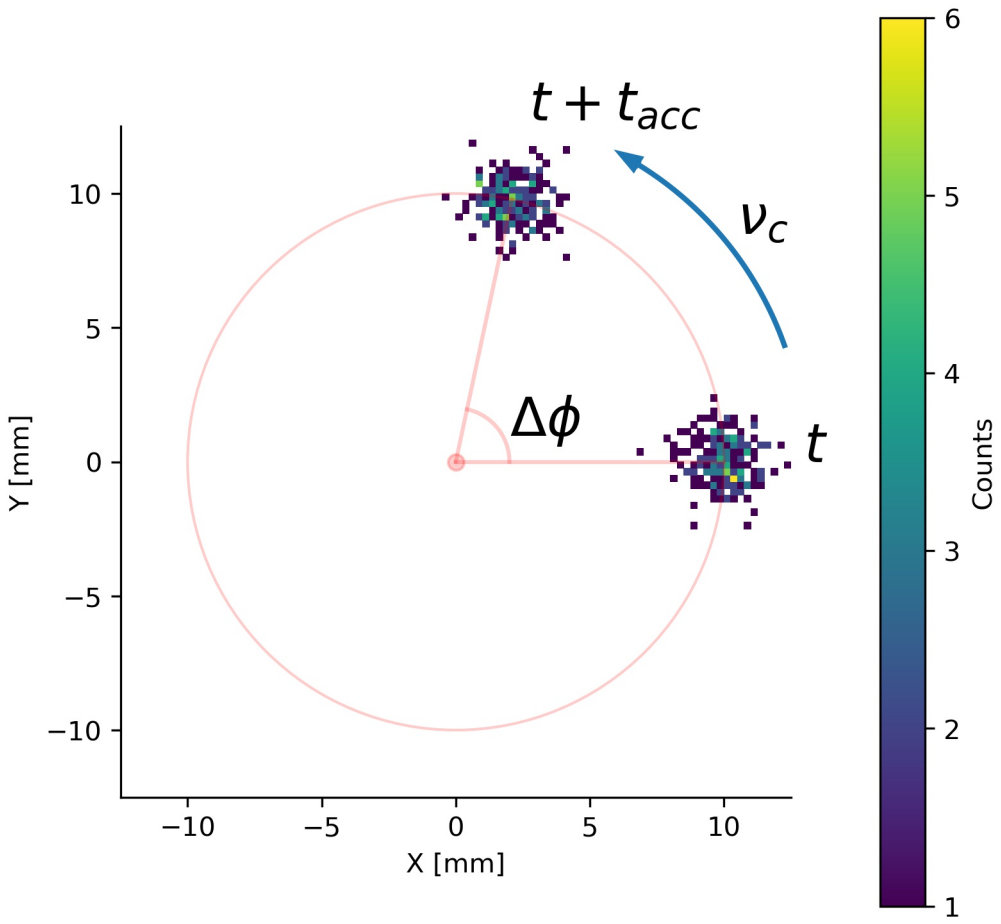
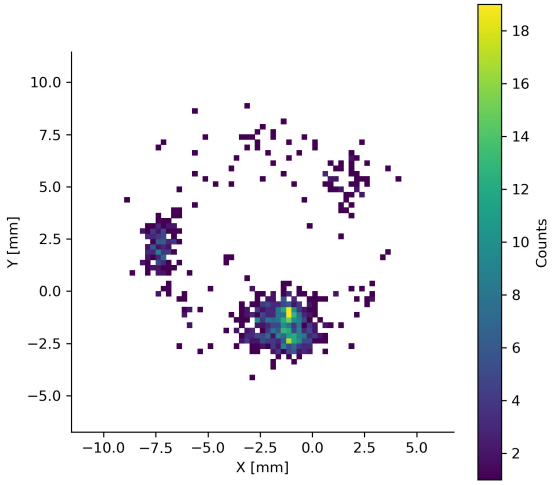
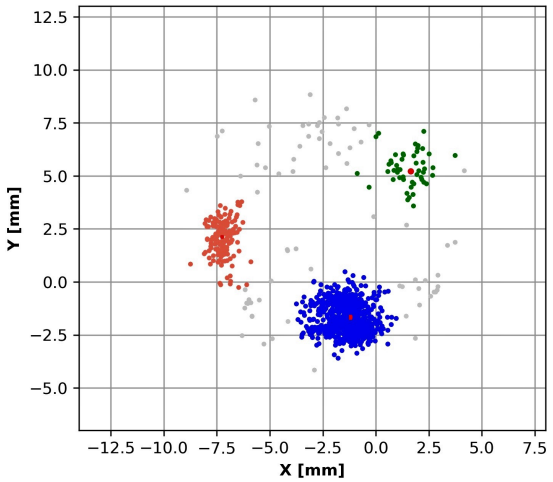


Figure 1: An example of a simulated PI-ICR measurement. An ion species' cyclotron frequency is measured using Eq. 1 by determining its excess phase accumulated during some time t_{acc}. The cyclotron frequency is then used to deduce the species' mass.

empirical data sets, such as the one shown in Fig. 2a, a single multivariate Gaussian is too simple to capture the complexity of the data. Even if a data set were restricted spatially to focus on a single cluster, systematic uncertainties could be introduced by the subjective assignment of boundaries to the subset. These systematic uncertainties could be especially prevalent if other challenging but oft-encountered situations are present, such as if there are multiple or overlapping clusters in the subset, if the subset is distorted, or if there is significant noise remaining in the subset. The need for a method of dividing up a data set into subsets based on a common characteristic, or *clustering*, is therefore critical for improving the precision of PI-ICR experiments.



(a) A raw PS-MCP data set generated from a beam with mass number $A = 168$ and charge state $q = 2+$, in which each pixel represents some number of ions that have hit the detector at that location. Each clustering algorithm begins with a data set like this.



(b) The Mean Shift algorithm discovers regions of the data set in which the ion density is above some threshold value. The ions in these areas are then clustered together into spots, whose radius is a parameter given by the experimenter, and given a color (as in the blue, pink, and green spots above). Ion hits that aren't clustered are identified as noise, and colored gray. Centers and uncertainties are determined using two univariate Gaussian fits to each spot corresponding to the two degrees of freedom of the ions.

Figure 2: An example of the Mean Shift clustering algorithm.

2 Gaussian mixture models: a clustering solution for PI-ICR

Unsupervised machine learning algorithms that "cluster" data automatically, known as *clustering algorithms*, have thus far received little attention with regard to mass spectrometry despite their being adequate tools for determining the boundaries and locations of spots in PI-ICR data. One algorithm that has been used is the Mean Shift algorithm,[12] which clusters ions into a spot if they exceed some pre-determined density threshold (Fig. 2), but no justification is given for its use beyond its simplicity. Furthermore, no methods are described for how to define the density threshold, meaning that an experimenter would have to make a subjective judgement both about the sizes and densities of the spots they expect to find. For these reasons, Mean Shift is an inadequate clustering solution for PI-ICR.

Guided by our analysis of the Mean Shift algorithm, we can begin to lay down criteria that an optimal PI-ICR clustering algorithm should satisfy in order to avoid the introduction of subjectivity into the data analysis process. For example, an optimal clustering algorithm should have no user-defined parameters that force the experimenter to make assumptions about the size of clusters or the number of clusters K in the data set, and should

perform well with non-spherical data. These criteria eliminate the well-known K-Means and DBSCAN clustering algorithms, as well as many of their variants.[15–22] Of the K-Means and DBSCAN variants that include methods for determining the optimal parameter values, none perform well with variable densities of samples within clusters and overlapping clusters.[23–25]

One set of clustering algorithms that satisfies many of these criteria is the class of Gaussian mixture model (GMM) algorithms. With Gaussian mixture models, the number of user-defined input parameters can be reduced to one, namely, the number of components to use K, and several methods exist to determine this as well as the other model parameters. GMMs also offer intuitive interpretations of the spot centers and uncertainties. We therefore decided to test GMMs against simulated PI-ICR data to determine their aptitude for clustering PI-ICR data. We tested two methods of determining GMM parameters: maximum-likelihood (ML) estimation, corresponding to the classic Gaussian mixture model (GMM), and Variational Bayesian Inference, corresponding to the Bayesian Gaussian mixture model (BGM), both of which we implemented using the Python package *scikit-learn*.[26] Each algorithm was broken down additionally into two flavors corresponding to the two coordinate systems, Cartesian and polar, that are most convenient for PI-ICR data.

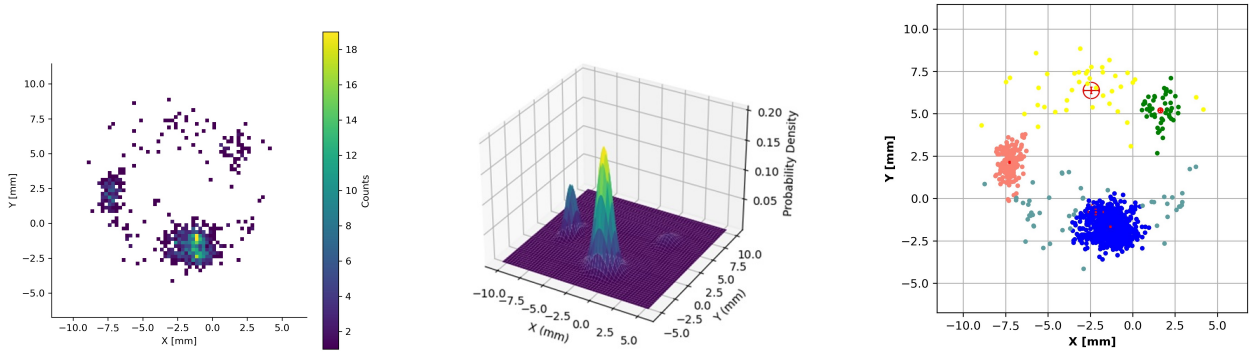
To estimate the ML parameters for a GMM with a given number of n components, we used the expectation-maximization (EM) algorithm, which consists of two steps that repeat until convergence following parameter initialization:

1. Expectation: given the current estimate of parameters, calculate the expected value of the log-likelihood of the data samples to assign each sample to a component of the overall GMM.

2. Maximization: given the expected log-likelihood values and cluster assignments, update the parameters to maximize the complete-data log-likelihood, giving them the value that maximizes the likelihood that the data has the expected labels.[27,28]

To determine K, we fit GMMs to several test values of K and selected the one that minimized the Bayesian Information Criterion (BIC), signifying that it is the most likely model for the data[29–31]. Finally, we assigned each data sample to a cluster based on the Gaussian component with respect to which it had a maximum log-likelihood of occurring, with cluster centers and uncertainties given by the model parameters. A visualization of this procedure is shown in Fig 3.

For the Variational Bayesian Inference method, we follow the procedures as outlined in Bishop[32] and Blei[33] by first assigning *a priori* distributions to the parameters of our BGM model. We used the stick-breaking representation of the Dirichlet process, a Gaussian distribution, and the Wishart distribution as priors for our component weights, means, and precisions, respectively. Next, we again used the EM algorithm to fit our model, except in the maximization step we set the model parameters to their expectation values calculated from their respective posterior distributions, which are updated according to the likelihoods generated in the expectation step. Because this approach allows for the component weights to approach 0 and thus for there to be components to which no samples belong, we initialized our model with an arbitrarily high number of components and determined K-components to be the number of components to which there belongs at least one sample. Data samples, centers, and uncertainties were assigned in the same way as with the GMM methods.



(a) The algorithm begins with an unlabelled data set, such as this one generated from a beam of ions with mass number $A = 168$ and charge state $q = 2+$.

(b) For several test values of K-components, a GMM is fit to the data.

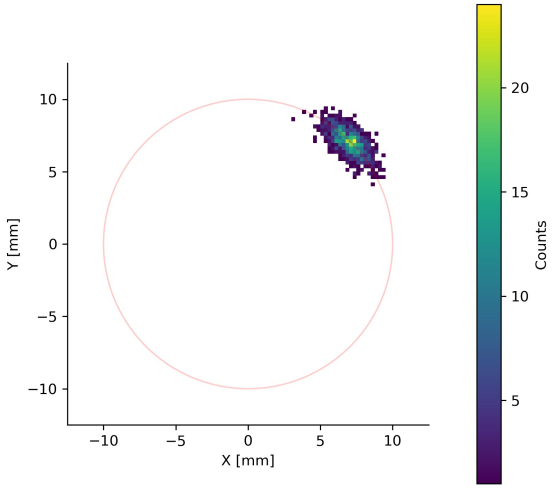
(c) After using the BIC to determine an optimal value for K, the data is assigned to clusters based on the Gaussian component with respect to which it had a maximum log-likelihood of occurring.

Figure 3: An example of how maximum likelihood estimation is used to fit a GMM to PI-ICR data.

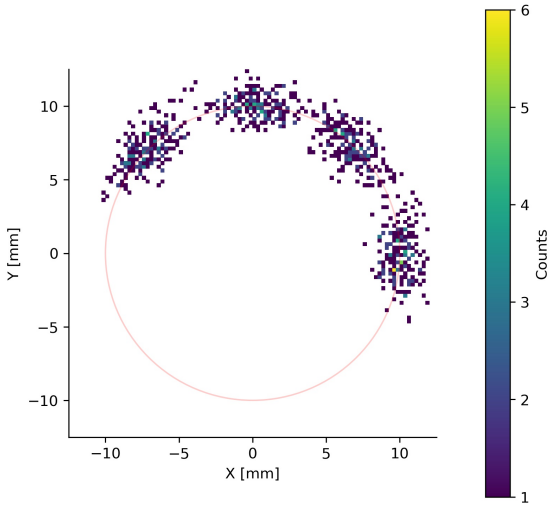
The fifth method we tested, the so-called "Phase-First Gaussian Mixture", is a version of GMM in which we consider that for PI-ICR analysis, we are primarily concerned with the phase dimension of the ions. First, we fit a GMM to the phase dimension of the data, and then used the fit centers and sample assignments as the initialization for a GMM fit to the full polar data set. In the second fit, we fixed the phase dimension of the centers to be the phases given by the phase-only fit, and then determined the cluster assignments, centers, and uncertainties as with the other models.

3 GMM performance evaluation methods

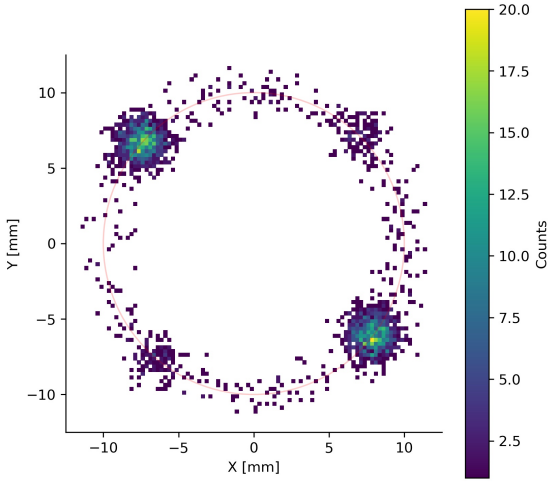
To evaluate the aptitude of GMMs for PI-ICR applications, we systematically measured their accuracy and precision relative to the Mean Shift algorithm on simulated data in Monte Carlo simulations consisting of 1000 repetitions. To generate the simulated data, we defined three broad subsets of PI-ICR data, as well as significant variables within each subset to test the models against. The simplest subset of data consisted of the spectra in which there is only one spot and no noise, which corresponds to the realistic experimental situation of generating a reference spot without applying any trap excitations to separate the ions from each other (Fig. 4a). Within the single-spot subset, we tested data sets where the spot had $10^2, 10^3$, and 10^4 samples. Samples were drawn from polar or Cartesian coordinates, and spots had an ellipticity of 1 or 2. The spot centers were at phases of $0° - 45°$ in $15°$ increments, measured counter-clockwise from the horizontal line $x = 0$. We also tested spots with additional rotations applied relative to the tangent to the ring at the location of the spot center at values of $0° - 45°$ in $15°$ increments. Additional phase locations and rotations were not tested due to the rotational symmetries of the algorithms.



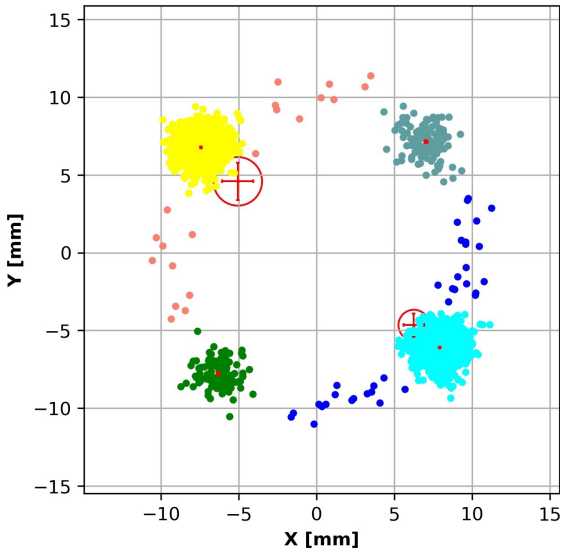
(a) A simulated scenario with a single spot composed of 1000 samples drawn from polar coordinates centered at a phase of 45°with ellipticity 1.



(b) A scenario with multiple spots of 200 ions each drawn from Cartesian coordinates, separated by 5σ, and with ellipticity 2.



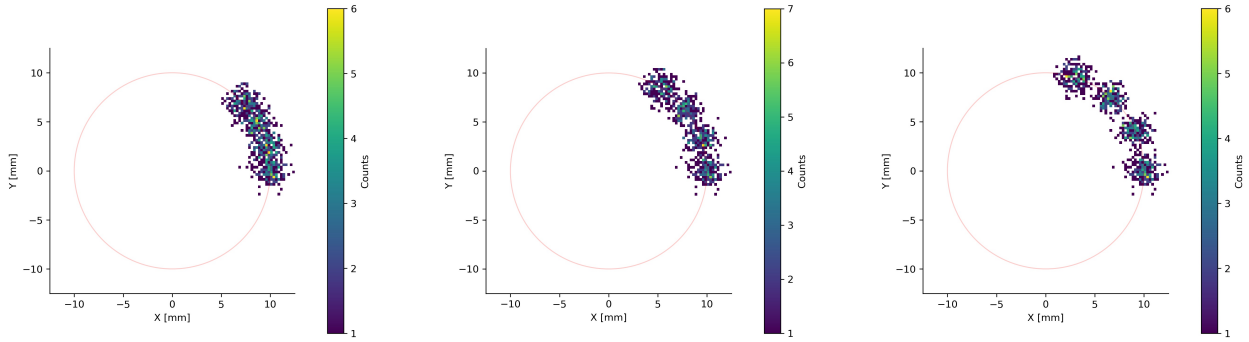
(c) A scenario with noise on the ring in which the ratio of noise spots to the number of spots in the smallest cluster is 1.



(d) The result of clustering Fig. 4c using the EM algorithm in Cartesian coordinates.

Figure 4: Examples of simulated PI-ICR data.

The second subset of data consisted of spectra in which there were at least two spots, and no noise (Fig 4b), corresponding to the experimental scenario in which there are multiple species in the beam and all ions are acted upon equally by the excitations. We tested scenarios in which there were 2, 4, or 8 spots. The spot separation, measured in



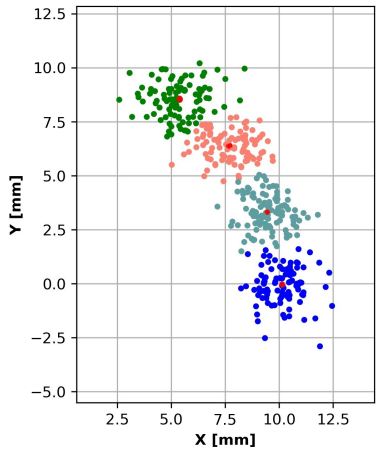
(a) 3σ (b) 4σ (c) 5σ

Figure 5: Various spot separations in terms of the standard deviation σ in the distribution of the ions' phase coordinates.

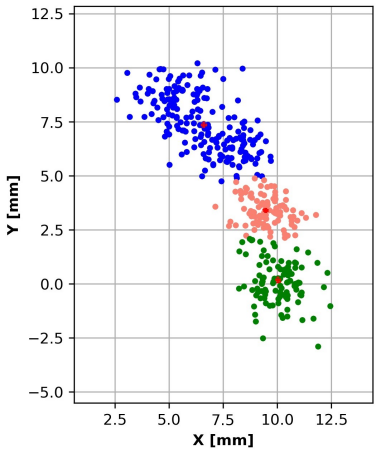
terms of the standard deviation σ of the phase values of the samples in each spot, was set at 3σ, 4σ, or 5σ. For all data sets, we placed the first spot on the horizontal line $x = 0$ and subsequent spots on the main ring at the given phase separation counter-clockwise from the first spot (Fig. 5). We also tested scenarios in which spots had different intensities. The spot intensity variable, which describes the ratio between the number of samples in the more dense spots to the number of samples in the less dense spots, was evaluated at 1 and 10. Beginning with the first spot, all odd-numbered spots were assigned to be less dense and were given 100 samples, whereas the even-numbered spots were given a number of samples based on the spot intensity variable.

The third subset of data consisted of spectra in which there was one spot in each quadrant consisting of 100 samples drawn from Cartesian coordinates with ellipticity one and noise on the ring (Fig 4c), corresponding to experiments in which there are multiple species in the beam and in which manual noise cuts fail to remove all noise from the data set. The remaining noise can originate from several sources, including but not limited to the excitations not acting on all ions equally, electronic noise, and cosmic ray interactions with the detector. One of the two variables considered for these scenarios was noise intensity, which we defined to be the ratio between the number of noise samples to the number of samples in the smallest spot. We tested noise intensity values from 0-15 in whole numbers. The second variable we considered for scenarios with noise on the ring was spot intensity, which we defined the same as with the scenarios with multiple spots. Spot intensity values evaluated were 1 and 10.

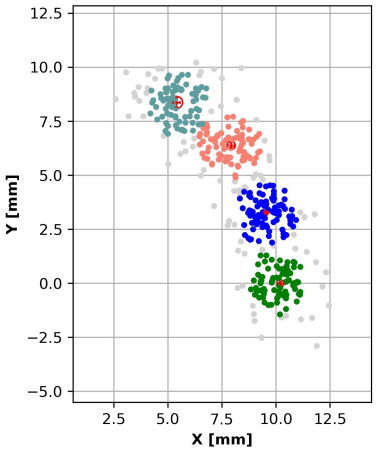
Each combination of variables was used to generate a set of 1000 spectra to be used in the Monte Carlo simulations. Each of the five clustering methods described above was then used to cluster the data, and each method was evaluated for precision and accuracy relative to the well-known Mean Shift clustering algorithm. When using the Mean Shift algorithm, spot centers and uncertainties were calculated by fitting univariate Gaussian curves to the x and y dimensions of the samples in a cluster.



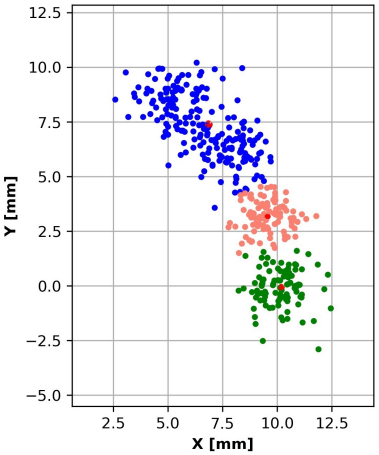
(a) Expectation-maximization algorithm with data in Cartesian coordinates



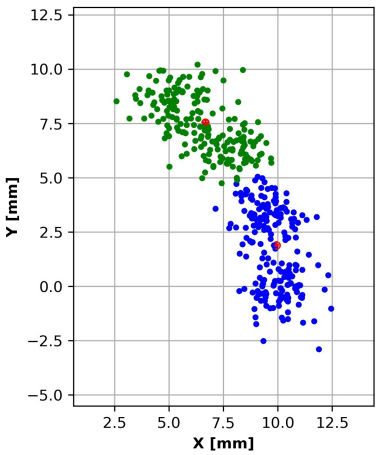
(b) Variational Bayesian algorithm with data in Cartesian coordinates



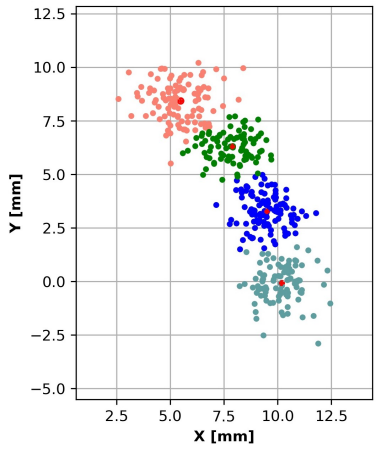
(c) Mean Shift algorithm



(d) Expectation-maximization algorithm with data in polar coordinates



(e) Variational Bayesian algorithm with data in polar coordinates



(f) Phase First algorithm

Figure 6: Examples of clustering results for different algorithms for a simulated data set where spots of ellipticity 1 are separated by 4σ.

4 Results and Discussion

We found that for spots rotated 15° or less relative to the tangent to the ring, GMMs could be used to accurately find spot centers under all tested scenarios, and were more precise than the Mean Shift algorithm. Because we simulated the spots by sampling from a Gaussian distribution, the centers of our sample spots varied based on the standard deviation of the distribution such that 66% of spot centers were within 1σ of the distribution center, 95% of spots were within 2σ of the distribution center, and so on. We therefore expected the clustering results from the GMM algorithms to vary in this way, and our

results confirmed this hypothesis, in some cases exceeding the expected accuracy. The Mean Shift algorithm found spots within a certain distance of the distribution center at rates lower than the GMM algorithms. We speculate that this is due to the Mean Shift algorithm not including samples more than a certain distance away from the cluster center.

All GMMs tested were effective for spots separated by at least 5σ. For lesser spot separations, the Variational Bayesian algorithm and the EM algorithm with polar coordinates failed to fit to all the spot centers for most scenarios (Fig. 6). The EM algorithm in Cartesian coordinates and the Phase First algorithm were able to resolve spots with a separation of 4σ for all scenarios if there were at most 4 spots. This suggests that these two algorithms should be used as spot separation decreases, but for separations less than 4σ, the accuracy of the models deteriorate.

All GMMs became less precise for any amount of noise, but we found that the EM algorithm and Phase First method still performed well for noise intensities less than or equal to one. We note that for scenarios with noise on the ring, it's common for the GMMs to fit to clusters in the noise. Examples of this are the royal blue and salmon clusters in Fig. 4d. These "noise clusters" almost certainly don't represent clusters of ions with identical cyclotron frequency, can be identified by their large uncertainties relative to the real spots, and can be ignored during data analysis following clustering. They should be interpreted as a placeholder that separates noise from the real clusters.

5 Outlook

One spot quality that is sometimes seen in PI-ICR data is a "tail" trailing off the end of spots around the ring opposite the direction of rotation (Fig. 7). These can be indicative of a poorly tuned beam or temporal drift in the system, and are more commonly seen in data that has been accumulated without interruption over a period of hours. In an effort to measure GMM performance on tail spots, we simulated them using an exponentially-modified Gaussian (EMG) distribution for the phase dimension with a probability density function of the form

$$EMG(x;\mu,\sigma,\tau) = Ce^{\frac{1}{2\tau}(2\mu+\frac{\sigma^2}{\tau}-2x)}\text{erfc}\left(\frac{\mu}{\sqrt{2}\sigma}+\frac{\sigma}{\sqrt{2}\tau}-\frac{\sqrt{2}x}{\sigma}\right), \tag{2}$$

where μ and σ are the mean and scale of the normal component of the EMG distribution, τ is the exponential relaxation parameter, and C is a normalization constant. The exponential component is the complementary error function $\text{erfc}(x) = 1 - \text{erf}(x) = \frac{2}{\sqrt{\pi}}\int_x^\infty e^{-y^2}dy$.

Analyzing the results of Monte Carlo simulations for values of τ between 2 and 20 gave evidence that the EMG distribution is an incorrect model for empirical tail spots, as we see a difference in fit centers depending on algorithm that is not present when analyzing real tail spots. Further work is needed on understanding the origins and nature of the tails so that we can formulate a better model for their simulation.

Tail spots aside, we conclude that Gaussian mixture models are a viable class of clustering algorithm for use in PI-ICR clustering analysis given sufficient spot separation, quantity, and noise level. Because of this result, we published a Python package called "GMMClusteringAlgorithms"[34] specifically for use in PI-ICR experiments. The package

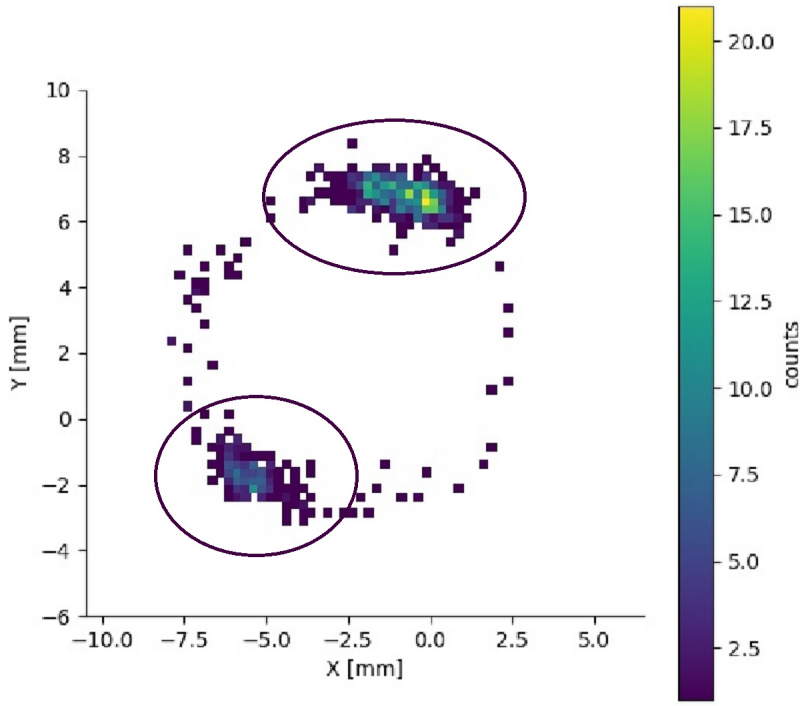


Figure 7: Empirical "tail" spots generated from a beam of ions with mass number $A = 168$ and charge state $q = 2+$.

includes functions to read data in from .lmf files where the raw PS-MCP data is stored, cluster the data using any of the GMMs tested for this report, and produce images like each shown in Fig. 3.

6 Acknowledgements

This work was funded by the Department of Energy through the Science Undergraduate Laboratory Internship (SULI) program and the Department of Energy, Office of Science, Office of Nuclear Physics under contract No. DE-AC02-06CH11357. I would also like to recognize the support of my supervisor, [redacted] for facilitating this work, and [redacted] for providing the initial nudge in the right direction. Finally, this project wouldn't have been possible without the contributions of ideas, guidance, feedback, and computing resources from my primary collaborators, [redacted]

References

[1]J. Dilling, K. Blaum, M. Brodeur, and S. Eliseev, "Penning-trap mass measurements in atomic and nuclear physics", Annu. Rev. Nucl. Part. Sci. **68**, 45 (2018).

[2]M. Arnould, S. Goriely, and K. Takahashi, "The r-process of stellar nucleosynthesis: Astrophysics and nuclear physics achievements and mysteries", Phys. Rep. **450**, 97 (2007).

[3]B. P. Abbott et al., "Multi-messenger Observations of a Binary Neutron Star Merger", Astrophys. J. Lett. **848**, L12 (2017).

[4]R. Chornock et al., "The Electromagnetic Counterpart of the Binary Neutron Star Merger LIGO/Virgo GW170817. IV. Detection of Near-infrared Signatures of r -process Nucleosynthesis with Gemini-South", Astrophys. J. Lett. **848**, L19 (2017).

[5]J. J. Cowan, C. Sneden, J. E. Lawler, A. Aprahamian, M. Wiescher, K. Langanke, G. Martínez-Pinedo, and F.-K. Thielemann, "Origin of the heaviest elements: The rapid neutron-capture process", Rev. Mod. Phys. **93**, 015002 (2021).

[6]M. R. Mumpower, G. C. McLaughlin, R. Surman, and A. W. Steiner, "The link between rare-earth peak formation and the astrophysical site of the r-process", Astrophys. J. **833**, 282 (2016).

[7]G. Eitel, M. Block, A. Czasch, M. Dworschak, S. George, O. Jagutzki, J. Ketelaer, J. Ketter, S. Nagy, D. Rodríguez, C. Smorra, and K. Blaum, "Position-sensitive ion detection in precision penning trap mass spectrometry", Nucl. Instrum. Methods Phys. Res., Sect. A **606**, 475 (2009).

[8]S. Eliseev, K. Blaum, M. Block, C. Droese, M. Goncharov, E. Minaya Ramirez, D. A. Nesterenko, Y. N. Novikov, and L. Schweikhard, "Phase-imaging ion-cyclotron-resonance measurements for short-lived nuclides", Phys. Rev. Lett. **110**, 082501 (2013).

[9]S. Eliseev, K. Blaum, M. Block, A. Dörr, C. Droese, T. Eronen, M. Goncharov, M. Höcker, J. Ketter, E. M. Ramirez, D. A. Nesterenko, Y. N. Novikov, and L. Schweikhard, "A phase-imaging technique for cyclotron-frequency measurements", Appl. Phys. B **114**, 107 (2014).

[10]J. Karthein, D. Atanasov, K. Blaum, S. Eliseev, P. Filianin, D. Lunney, V. Manea, M. Mougeot, D. Neidherr, Y. Novikov, L. Schweikhard, A. Welker, F. Wienholtz, and K. Zuber, "Direct decay-energy measurement as a route to the neutrino mass", Hyperfine Interact. **240**, 61 (2019).

[11]D. A. Nesterenko, T. Eronen, A. Kankainen, L. Canete, A. Jokinen, I. Moore, H. Penttilä, S. Rinta-Antila, A. de Roubin, and M. Vilen, "Phase-imaging ion-cyclotron-resonance technique at the JYFLTRAP double Penning trap mass spectrometer", Eur. Phys. J. A **54**, 154 (2018).

[12]R. Orford, J. Clark, G. Savard, A. Aprahamian, F. Buchinger, M. Burkey, D. Gorelov, J. Klimes, G. Morgan, A. Nystrom, W. Porter, D. Ray, and K. Sharma, "Improving the measurement sensitivity of the Canadian Penning trap mass spectrometer through PI-ICR", Nucl. Instrum. Methods Phys. Res., Sect. B **463**, 491 (2020).

[13]J. Karthein, D. Atanasov, K. Blaum, D. Lunney, V. Manea, and M. Mougeot, "Analysis methods and code for very high-precision mass measurements of unstable isotopes", arXiv:2102.10413 (2021).

[14]J. Karthein, "Next-Generation Mass Spectrometry of Exotic Isotopes and Isomers", PhD (Heidelberg University, 2020).

[15]S. Lloyd, "Least squares quantization in PCM", IEEE Trans. Inf. Theory **28**, 129 (1982).

[16]J. C. Bezdek, R. Ehrlich, and W. Full, "FCM: The fuzzy c-means clustering algorithm", Comput. Geosci. **10**, 191 (1984).

[17]N. R. Pal and J. C. Bezdek, "On cluster validity for the fuzzy c-means model", IEEE Trans. Fuzzy Syst. **3**, 370 (1995).

[18]M. Ester, H.-P. Kriegel, J. Sander, X. Xu, et al., "A density-based algorithm for discovering clusters in large spatial databases with noise.", in KDD '96: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, Vol. 96 (1996), p. 226.

[19]A. Hinneburg and D. Keim, "An efficient approach to clustering in large multimedia databases with noise", in KDD '98: Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (1998), p. 58.

[20]M. Dash, H. Liu, and X. Xu, "'$1 + 1 > 2$': merging distance and density based clustering", in Proceedings Seventh International Conference on Database Systems for Advanced Applications (2001), p. 0032.

[21]J. Ghosh and A. Liu, "K-Means", in *The Top 10 Algorithms in Data Mining*, edited by X. Wu and V. Kumar (Chapman & Hall/CRC, 2009), p. 21.

[22]K. M. A. Patel and P. Thakral, "The best clustering algorithms in data mining", in 2016 International Conference on Communication and Signal Processing (ICCSP) (2016), p. 2042.

[23]X. Xu, M. Ester, H. Kriegel, and J. Sander, "A distribution-based clustering algorithm for mining in large spatial databases", in 2013 IEEE 29th International Conference on Data Engineering (ICDE) (1998), p. 324.

[24]M. Ankerst, M. M. Breunig, H.-P. Kriegel, and J. Sander, "Optics: ordering points to identify the clustering structure", Sigmod Rec **28**, 49 (1999).

[25]S. Jahirabadkar and P. Kulkarni, "Algorithm to determine ϵ-distance parameter in density based clustering", Expert Syst. Appl. **41**, 2939 (2014).

[26]F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: machine learning in Python", J. Mach. Learn Res. **12**, 2825 (2011).

[27]D. Reynolds, "Gaussian mixture models", in *Encyclopedia of Biometrics*, edited by S. Li and A. Jain (Springer, 2015), p. 827.

[28]R. B. Millar, *Maximum likelihood estimation and inference with examples in R, SAS, and ADMB* (Wiley & Sons, Chichester, Sussex, U.K., 2011).

[29]G. Schwarz, "Estimating the dimension of a model", Ann. Stat. **6**, 461 (1978).

[30]R. A. Stine, "Model selection using information theory and the MDL principle", Sociol. Methods Res. **33**, 230 (2004).

[31]J. J. Dziak, D. L. Coffman, S. T. Lanza, R. Li, and L. S. Jermiin, "Sensitivity and specificity of information criteria", Brief Bioinform. **21**, 553 (2020).

[32]C. M. Bishop, *Pattern recognition and machine learning* (Springer, New York, 2006).

[33] D. M. Blei and M. I. Jordan, "Variational inference for Dirichlet process mixtures", Bayesian Anal. **1**, 121 (2006).

[34] [redacted], "GMMClusteringAlgorithms", http://pypi.org/project/GMMClusteringAlgorithms, Accessed: 2021-30-04, 2021.