UNIVERSITY OF BRITISH COLUMBIA-OKANAGAN

DATA 448A: Fitting Truncated Mixture Models

Presented by Emily Medema

April 2021

ABSTRACT

This report will go into the process and development of fitting a truncated gaussian mixture model on the complex spectra in order to classify peaks. We will also propose instances of which this would be supremely helpful in a applicable standpoint, such as fitting on the results of Raman Spectroscopy.

INTRODUCTION

What if we were able to personalize radiation treatment in accordance to a tumors response to radiation and anti-cancer drugs? This could help make the currently most cost-effective cancer treatment more effective to more patients around the world. Currently, we have no way to determine how a tumor will react to radiation nor how radio-sensitive it may be [1].

Recently, studies have shown that Raman Spectroscopy - a non-invasive optical interrogation method where vibrational modes of molecules are identified through inelastic light scattering - can be used to gather information pertaining to radiation induced responses. However, the issue remains as to how to predict and classify radiation response with the complex spectra result of Raman Spectroscopy [2].

Classification and clustering are some of the main objectives of machine learning. If we can classifying peaks from the Raman spectra into a cluster that is as similar as possible with the in different clusters as different as possible; we can determine which chemical responses are induced by radiation and eventually predict how someone might react to radiation treatment [3].

There are many different applications of classification and clustering not only within the statistics community but within the fields of biology, meteorology [4], physics, and more [5]. Despite the prevalence, we do not have an accurate solution without major flaws. Currently, we are estimating and fitting Principal Component Analysis (PCA) and Non-Negative Matrix Factorization (NMF) models. These models are unfortunately unable to deal with the widths of the peaks characteristic to particular molecules expanding when measured in a complex system, such as a cell. Luckily, due to the area under the peaks being normalized to one unit and the peaks assumed to be Gaussian or Lorentzian, we can fit a Mixture Model. This will help us classify peaks as well as fix the issue with peak widths expanding.

FITTING PROCEDURES

In order to gain an understanding of the inner-workings of a Gaussian Mixture Model, we started to create one to fit on regular data points from the bank dataset. We then proceeded to change and adjust the model as needed until we had a model fitting to a simplified, simulated Raman Spectroscopy spectra.

GAUSSIAN MIXTURE MODEL

To be begin with we used the bank dataset. The bank dataset was sourced from the gclus library which itself sourced the data from the "Multivariate Statistics A practical approach" study by Bernhard Flury and Hans Riedwyl [6]. This data consists of multiple different measurements of Swiss banknotes used in order to determine if a note is genuine or counterfeit.

We then began to develop a Gaussian Mixture Model. In general, for a random vector X, a

gaussian mixture model distribution would be defined as:

$$f(\mathbf{x}|\mu_g, \sigma_g^2) = \sum_{g=1}^{G} \pi_g p(\mathbf{x}|\mu_g, \sigma_g^2) \tag{0.1}$$

As we know fitting separate distributions to unknown groups is not possible, we will be using the Expectation-Maximization (EM) algorithm to estimate the group memberships.

The EM Algorithm is the iterative computation of maximum likelihood (ML) estimates used to estimate missing data - in this case group membership [7].

An EM algorithm follows these basic steps [8]:

1. Start the algorithm with random values for $h_i(d_i)$ (there are alternative starting options)
2. Assuming those $h_i(d_i)$ are correct, estimate parameters μ_g and σ_g (via MLEs - hence this is the maximization of EM)
3. Assuming those parameters are correct, find the expected value of group memberships ie. Expectation Step:
$$h_i(d_i) = \frac{\phi(d_t|\mu_i, \sigma_i)}{\sum_{j=1}^{N} \phi(d_t|\mu_j, \sigma_j} \tag{0.2}$$
4. Repeat 2 and 4 until there are minimal changes (the log-likelihood of model is monitored for convergence)

After the algorithm has reached convergence, we have an estimate of the missing data and can now predict to which group each peak belongs. In this case, we were able to fit onto the bank dataset and classify if a note was forged or not.

As seen in Figure 1, it fits fairly well. In fact, there were only two misclassifications (see: Table 0.1).

Table 0.1: Classification of Bank Notes

True \ Results	0	1
0	69	2
1	0	69

GMM ON COMPLEX SPECTRA

Nonetheless, while our GMM fit works well, we need to be able to fit onto the complex spectra that results from Raman Spectroscopy not data points. Each Raman acquisition is an observation of $g_i(d)$ where g_i is a univariate mixture model and d is a predetermined grid of

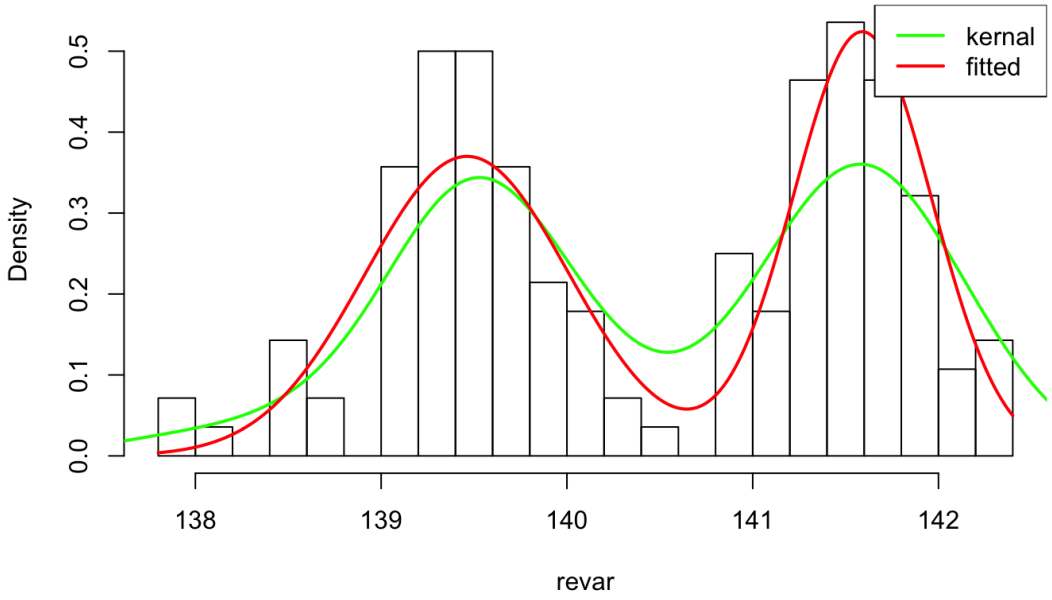

Figure 1: Bank Note GMM Fit

wavenumbers. Therefore, we need to adjust our model to fit onto $g_i(d)$. In particular, we will need to adjust our equation such that we account for the density in the model like so:

$$f(x) = \sum_{g=1}^{G} \pi_g p_g(x) \tag{0.3}$$

where x is a realization from a p-dimensional random vector, G is our number of sub-populations, π_g are the mixing proportions such that $\sum_{g=1}^{G} \pi_g = 1$ and $p_g(x)$ are probability densities.

In terms of adjusting our model to reflect this change, we will need to adjust the log-likelihood calculation as well as the underlying equations within the maximization-step of the EM Algorithm.

Within the Maximization step of the EM Algorithm we make the following calculations:

$$\pi_i = \frac{1}{L}\sum_{t=1}^{L} h_i(d_t)$$
$$\mu = \frac{\sum_{t=1}^{L} h_i(d_t)d_t}{\sum_{t=1}^{L} h_i(d_t)}$$
$$\sigma_i = \frac{\sum_{t=1}^{L} h_i(d_t) \times (d_t - \mu_i)(d_t - \mu_i)^\tau}{\sum_{t=1}^{L} h_i(d_t)}$$

As mentioned above, we will need to adjust these equations to account for the density. We will do so as follows:

$$\pi_i = \frac{\sum_{t=1}^{L} p(d_t)h_i(d_t)}{\sum_{t=1}^{L} p(d_t)}$$
$$\mu = \frac{\sum_{t=1}^{L} p(d_t)h_i(d_t)d_t}{\sum_{t=1}^{L} p(d_t)h_i(d_t)}$$
$$\sigma_i = \frac{\sum_{t=1}^{L} p(d_t)h_i(d_t) \times (d_t - \mu_i)(d_t - \mu_i)^\tau}{\sum_{t=1}^{L} p(d_t)h_i(d_t)}$$

where $p(d_t)$ is the density of grid point d_t. This will now take into account the density in the fit. Fitting with this taken into account results in the following fit on the simulated data [5].

The fit is fairly accurate except for on the ends as seen in Figure 2.

This is due to fitting a model that is assumed infinite onto a finite set of data. In order to fix this, we will need to fit a truncated gaussian mixture model instead of a regular gaussian fit on density.

Truncated GMM on Complex Spectra

Fitting a truncated gaussian mixture model on complex spectra will require a few changes to the code. In fact, it will require us to create a work-around for calculating the close-formed variables of the maximization step in the EM Algorithm as there is no functions for us to use to do so.

In particular, we modified a numeric solver so that we can use weights in non-integer form. Our success in this was largely dependent on two R libraries, truncdist and fitdistrplus. 'truncdist' is a package used to evaluate probability distribution functions for truncated random variables developed by Nadarajah and Kotz (2006)[9]. We can use this package to compute values needed in the e-step. 'fitdistrplus' is the package we will use and modify in order to perform the calculations needed in the m-step [10].

Once the modifications were done, we fit the truncated gaussian mixture model.

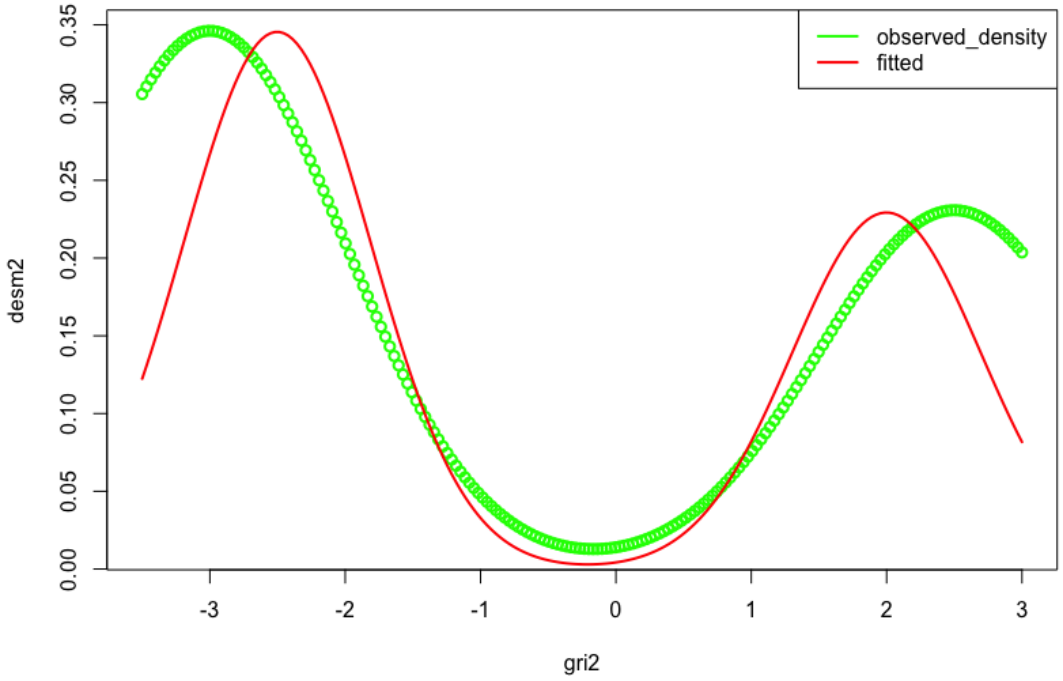

Figure 2: GMM Fit on Density

From Figure 3, we can tell that the fit of the truncated model has little error in comparison to the simulated data. We can also see that the fit of the truncated model is more consistent at the ends of the model.

RESULTS

While all models are applicable to the data we have, the truncated model appears to be more accurate than the others. Not only can we see this visually through Figure 4 but we can also determine the accuracy of the fit through comparing the μ and σ of the models with the generated μ and σ.

Looking at Table 0.2 we can see that the closest fit is the Truncated Gaussian Mixture model as it is the closest to the generated true μ and σ

Due to the accuracy of the fit in the truncated model we can confidently calculate the probability of a peak belonging to each group and thus classify the data.

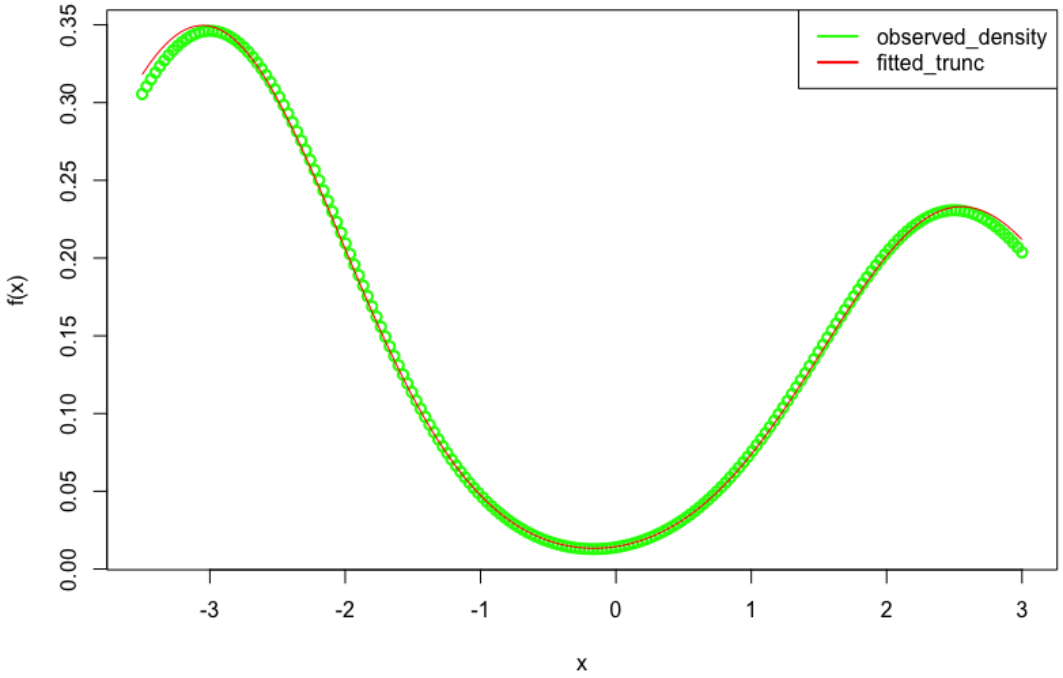

Figure 3: Truncated GMM Plot

DISCUSSION

This development can be extremely helpful in multiple fields, but in particular in use with Raman Spectroscopy. A common problem with the current model fit can be seen in Figure 5.

On the left of Figure 5 we can see that the fit assumes that the model will approach zero, however, it actually rises. This introduces unneeded inaccuracy and error to the model that the truncated model will fix.

Introducing a truncated gaussian mixture model can only improve the classification of biochemical responses to radiation and in turn advance the development of personalized cancer treatments for patients.

There are also multiple other uses for this model. It could be incredibly useful in the analysis of any peak-based classification problem provided the peaks can be assumed Gaussian and the area under the curve is normalized to one. For example, it could be applied to results from other methods such as mass spectroscopy for cancer patients or it could be used to extract geophysical information from topological maps of the earth's surface.

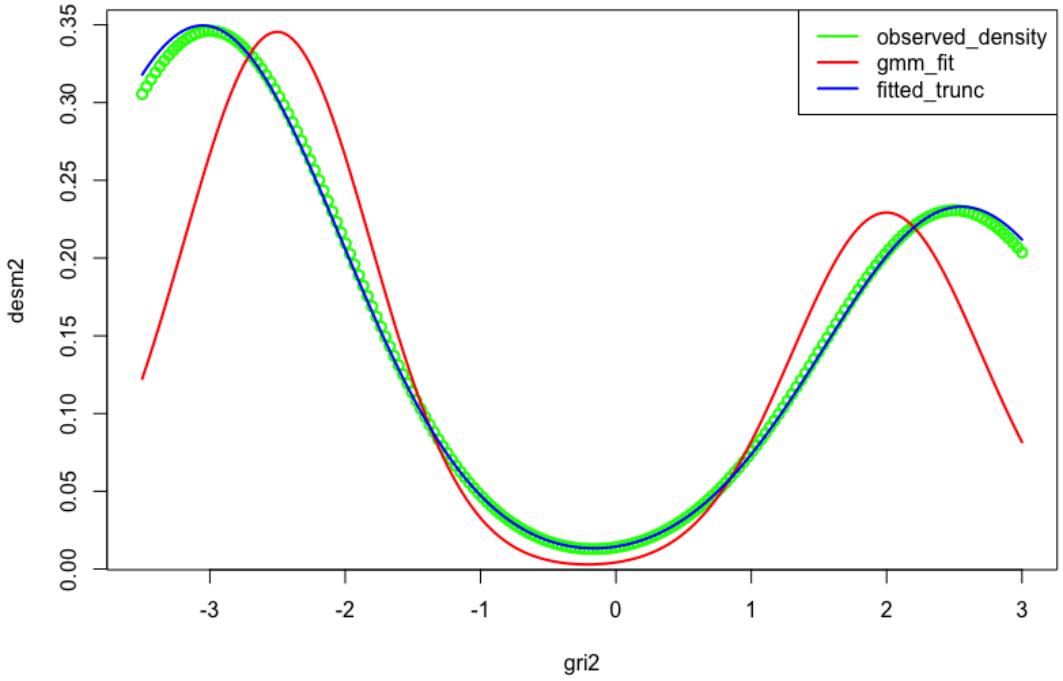

Figure 4: GMM vs. Truncated GMM Plot

CONCLUSION

Truncated Gaussian Mixture Models for peak-based classification are an useful tool that is needed in a lot of fields in order to further advance classification and research. In this study, we successfully developed an accurate truncated gaussian mixture model that can be used within Raman Spectroscopy and other spectra analysis for classification purposes. In the future, we will continue to improve upon the fit and expand towards fitting on other probability distributions such as Lorenztian curves.

Table 0.2: Comparing μ and σ

	μ_1	μ_2	σ_1	σ_2
Generated	-3	2.5	1	1
GMM Density	-2.502675	1.998836	0.6922663	0.6968473
GMM Truncated	-3.053944	2.553153	1.023668	1.022966

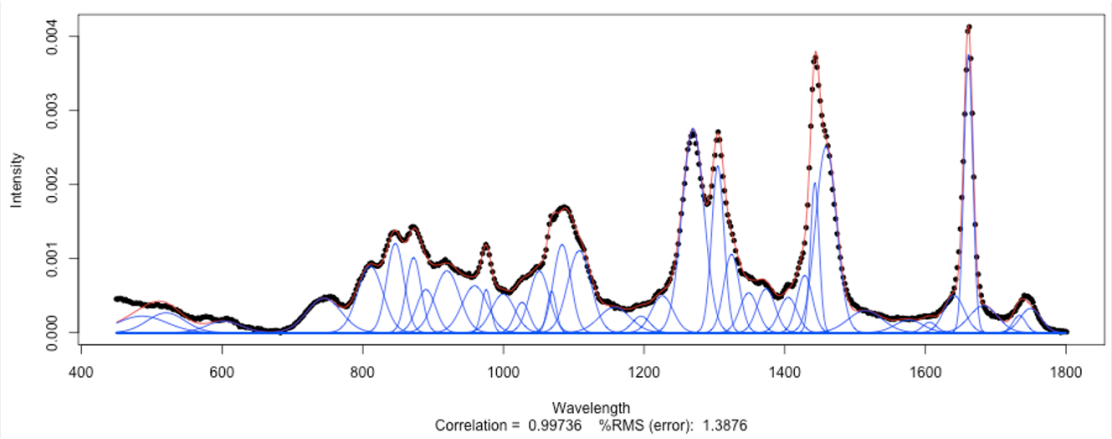

Figure 5: Raman Spectroscopy Fit Issue

REFERENCES

[1] X. Deng, R. Ali-Adeeb, J. L. Andrews, P. Shreeves, J. J. Lum, A. Brolo, and A. Jirasek, "Monitor ionizing radiation-induced cellular responses with raman spectroscopy, non-negative matrix factorization, and non-negative least squares," *Applied spectroscopy*, vol. 74, no. 6, pp. 701–711, 2020.

[2] K. Milligan, X. Deng, P. Shreeves, R. Ali-Adeeb, Q. Matthews, A. Brolo, J. J. Lum, J. L. Andrews, and A. Jirasek, "Raman spectroscopy and group and basis-restricted non negative matrix factorisation identifies radiation induced metabolic changes in human cancer cells," *Scientific reports*, vol. 11, no. 1, pp. 3853–3853, 2021.

[3] Z. Di, Q. Kang, D. Peng, and M. Zhou, "Density peak-based pre-clustering support vector machine for multi-class imbalanced classification," in *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*, pp. 27–32, 2019.

[4] W. Z. H. Z. J. L. K. L. D. Y. H. Tian, "Weather prediction with multiclass support vector machines in the fault detection of photovoltaic system," *IEEE/CAA journal of automatica sinica*, vol. 4, no. 3, pp. 520–525, 2017.

[5] T. Kawabata, "Multiple subunit fitting into a low-resolution density map of a macromolecular complex using a gaussian mixture model," *Biophysical journal*, vol. 95, no. 10, pp. 4643–4658, 2008.

[6] B. Flury and H. Riedwyl, *Multivariate statistics: a practical approach.* London;New York;: Chapman and Hall, 1988.

[7] G. J. McLachlan and T. Krishnan, *The EM algorithm and extensions*, vol. 382. John Wiley & Sons, 2007.

[8] P. D. McNicholas, *Mixture Model-Based Classification.* Oakville: CRC Press, 1 ed., 2017;2016;.

[9] F. Novomestky and S. Nadarajah, *truncdist: Truncated Random Variables*, 2016. R package version 1.0-2.

[10] M. L. Delignette-Muller and C. Dutang, "fitdistrplus: An R package for fitting distributions," *Journal of Statistical Software*, vol. 64, no. 4, pp. 1–34, 2015.