NON-NEGATIVE MATRIX FACTORIZATION WITH GAUSSIAN PROCESS PRIORS ON SIMULATED RAMAN SPECTROSCOPY DATA

Elika Araghi, Sanaz Behboodi, Christian A. Rasmussen

DTU - Technical University of Denmark

ABSTRACT

Spectrum features of the components in a simulated Raman spectroscopy dataset is to be analysed and extracted by applying non-negative matrix factorization. With raman spectroscopy one is assured that data is non-negative. In addition, we choose Gaussian process priors for D and H, in order to introduce prior knowledge about the data. Prior knowledge is included such that factors are assumed linked to an underlying gaussian process explained by a covariance function that can tell about the features such as sparsity and smoothness. Moreover, Metropolis-Hashings sampling is implemented as a proposal for gaining prior knowledge of the factors.

Index Terms— Non-negative Matrix Factorization, Gaussian Process Priors, Link Functions, Co-variance matrix, Sampling, Metropolis Hastings

1. INTRODUCTION

The aim of this project is to decompose simulated Raman spectroscopy data [1] by applying non-negative matrix factorization (NMF). It is of interest to see if underlying features in the data can be captured by factors yielded by NMF. With respect to the paper [2], it is of interest to see, if prior knowledge about the data can be included to improve the quality of the NMF. This interest will be explained further in the following section.

2. MOTIVATION

NMF is an unsupervised learning tool for analysing high-dimensional data and can be applied to gain knowledge about relevant features in a dataset. NMF is a type of linear dimensionality reduction where basis elements are assumed to be component-wise non-negative [3]. By extracting sparse features, one can learn and examine properties of a dataset. In this case, NMF can be used to learn the underlying variables that constitutes the signals in spectroscopy data. Raman spectroscopy is for instance used to identify and find the fingerprint of molecules in chemistry. Furthermore, the quality of the decompositions can be improved by including prior knowledge of what the data should look like. By assuming the factors can be linked to an underlying Gaussian process, it allows one to compute factors that correspond to the knowledge such as smoothness and sparseness in the data. Hence, the generated data can ultimately be compared to the given non-negative matrix factorization with Gaussian process priors, denoted GPP-NMF, and then be evaluated to see if decompositions agree with the prior knowledge of the factors.

3. METHOD

Notice, essential equations are similar to those given in the paper [2] due to the objectives in both studies being similar. The NMF problem can be written as

$$X = DH + N \tag{1}$$

where $X \in R^{K \times L}$ is a data matrix which is factorized to product of two matrices, $D \in R_+^{K \times M}$ and $H \in R_+^{M \times L}$. The matrix $N \in R^{K \times L}$ is the residual noise.

3.1. Least squares likelihood

Appropriate negative log-likelihood should be used under certain assumption about the distribution of the data. Least squares can be used, together with the function form of a Gaussian distribution, for computing the maximum likelihood, given by:

$$P^{Ls}_{X|D,H}(D,H) = \frac{1}{(\sqrt{2\pi}\sigma_N)^{KL}} \exp\left(-\frac{(\|X - DH\|)_F^2}{2\sigma_N^2}\right) \tag{2}$$

, where F is the Frobenius norm.

The Negative log likelihood, as a cost function for optimization is then:

$$L_{X|D,H}(D,H) \propto \frac{1}{2\sigma_N^2}\|X - DH\|_F^2 \tag{3}$$

Next step for estimating the maximum a posteriori (MAP) of the factors by using Bayes rule that is given by

$$P_{D,H|X}(D,H|X) = \frac{P_{X|D,H}(X|D,H)P_{D,H}(D,H)}{P_X(X)} \tag{4}$$

Since the Numerator is constant, the estimation of the negative log posterior is given by

$$L_{D,H|X}(D,H) \propto L_{X|D,H}(D,H) + L_{D,H}(D,H) \tag{5}$$

3.2. Change of variables for optimization

A change of variables [4] for the factors D and H is a rather practical way for estimating the MAP solution because the limitation of being in the space of positive reals, R_+, is dealt with. A Gaussian distribution is in the domain $[-\infty;\infty]$ and so, it is now possible of introducing Gaussian process priors for describing the factors and allow for optimization. Variables δ and η which are related to D and H is introduced by the change of variables theorem, such that:

$$D = g_d(\delta) = Vec^{-1}(f_d^{-1}(C_d^T\delta)) \tag{6}$$

$$H = g_h(\eta) = Vec^{-1}(f_h^{-1}(C_h^T\eta)) \tag{7}$$

Where the matrices C_d and C_h are the matrix square roots (Cholesky decompositions) [5] of the Covariance matrices of two factors D(Σ_d) and H (Σ_h) such that δ and η have i.i.d Gaussian distribution [6] and function Vec^{-1} maps its vector input into a matrix of suitable size. Furthermore, variables δ and η used for optimization is initialized as:

$$\delta = Vec(D_0) \tag{8}$$

$$\eta = Vec(H_0) \tag{9}$$

, where D_0 and H_0 are in same dimensions as D and H respectively with all elements intialized to zero. Moreover, the covariance function for both δ and η is given by a Gaussian radial basis function (RBF):

$$\phi(i,j) = \exp\big(-\frac{(i-j)^2}{\beta^2}\big) \tag{10}$$

Where i and j are two sample indices and the β is the length-scale parameter, which specifies the smoothness of the factors. The prior distribution of the transformed variable η by using change of variables theorem is given by the equation below. Due to symmetry, the following equations for δ are not listed.

$$P_\eta(\eta) = P_H(g_h(\eta))|J(g_h(\eta))| = \frac{1}{(2\pi)^{\frac{LM}{2}}}\exp(-\frac{1}{2}\eta^T\eta) \tag{11}$$

The negative log prior is thus:

$$L_\eta(\eta) \propto \frac{1}{2}\eta^T\eta \tag{12}$$

Finally, MAP estimate of the transformed variables is given:

$$L_{\delta,\eta|X}(\delta,\eta) = L_{X|D,H}(g_d(\delta), g_h(\eta)) + \frac{1}{2}\delta^T\delta + \frac{1}{2}\eta^T\eta \tag{13}$$

MAP solution is obtained by optimizing over δ and η, so we have

$$\delta_{MAP}, \eta_{MAP} = argmin_{\delta,\eta} L_{\delta,\eta|X}(\delta,\eta) \tag{14}$$

, by using the least squares likelihood for δ and η, one gets:

$$L^{LS-GPP}_{\delta,\eta|X}(\delta,\eta) = \frac{1}{2}\big(\sigma_N^{-2}\|X - g_d(\delta), g_h(\eta)\|_F^2 + \delta^T\delta + \eta^T\eta\big)$$

3.3. Choice of link function

For MAP estimation, it is needed to build prior assumptions into solution of the NMF method such that assumptions are linked to non-negative factors of D and H through a strictly increasing linked function. The choice of link function is made such that the inverse link function maps a distribution described by the underlying Gaussian process. In this case, the desired distribution is to be exponential and smooth. The inverse exponential-to-Gaussian link function with respect to H is given by:

$$f_h^{-1}(h_i) = -\frac{1}{\lambda}\log\big(\frac{1}{2} - \frac{1}{2}\phi(\frac{h_i}{\sqrt{2}\sigma_i})\big) \tag{15}$$

where λ is an inverse scale parameter, ϕ is an error function, and h_i are elements in $C_h^T\eta$ with respect to equation 6.

3.4. Sampling

One approach for gaining prior knowledge of the factors is to use approximate inference by sampling. In this case, the Metropolis Hastings sampling algorithm is chosen which is a class of Markov chain Monte Carlo sampling algorithms [7]. At step τ in which the current state is $z^{(\tau)}$, one draws a sample z^* from the distribution $q_k(z|z^{(\tau)})$, and accept it with probability $A_k(z^*, z^{(\tau)})$, where:

$$A_k(z^*, z^{(\tau)}) = min\Big(1, \frac{\tilde{p}(z^*)q_k(z^{(\tau)}|z^*)}{\tilde{p}(z^{(\tau)})q_k(z^*|z^{(\tau)})}\Big) \tag{16}$$

, where k labels the set of members of possible transitions. In this case, the initial state is the MAP estimate. The proposal distribution is chosen to be a Gaussian distribution.

4. RESULTS

We generate one hotspot explained by three voigts [8] with specified locations of 10, 40, 70 (corresponding to frequencies in Raman spectroscopy) on a 2-D map with size [10x10].

4.1. GPP-NMF

Figure 1 shows the generated dataset that is used for implementation. Figure 2 displays the reconstruction which is gained by multiplying the calculated factors D and H after optimization. As shown, the model was able to find the three hotspots in their given locations. The reconstruction in general looks similar to the underlying data.

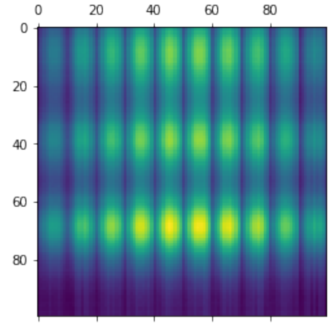

Fig. 1. Generated data simulating a Raman spectrum

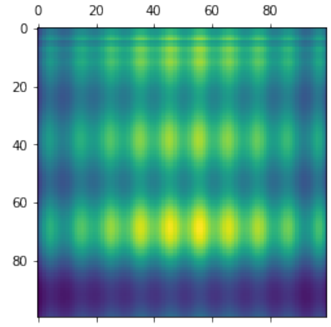

Fig. 2. Reconstructed data of the Raman spectrum

The columns of the matrix, D, shows The three spectrum's component found by the algorithm and seen in figure 3. The locations of the peaks are discovered correctly (10, 40, 70) although, two peaks are detected as one component (Orange graph) and with one feature looking more like noise. It is also observed that the underlying gaussian process prior was somewhat succesful in construcing smooth and exponential peaks that is in agreement with prior knowledge.

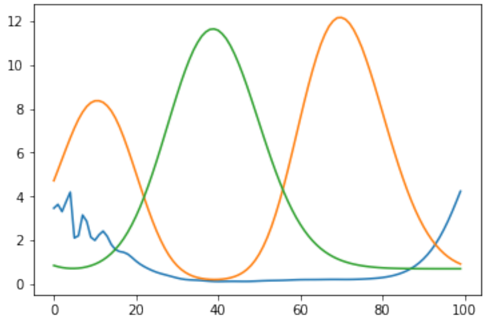

Fig. 3. Features of D. Notice how peaks are smooth and exponential as given by the underlying gaussian process prior

4.2. Sampling - Metropolis Hastings

The metropolis Hastings algorithm for sampling has been implemented in order to make approximate inference and hence, examine the prior distribution of δ and D respectively. For initialization, the first sample is chosen to be the MAP estimate of the change of variable. The proposal distribution is assumed Gaussian with mean μ and standard deviation σ equal to the given δ_{MAP}. Figure 4 is a histogram combined with a KDE visualizing the distribution of the δ_{MAP} (red) and sampled values (blue) using Metropolis Hastings sampling. Figure 5 visualizes the true and sampled distributions when the factor D is calculated from δ_{MAP}.

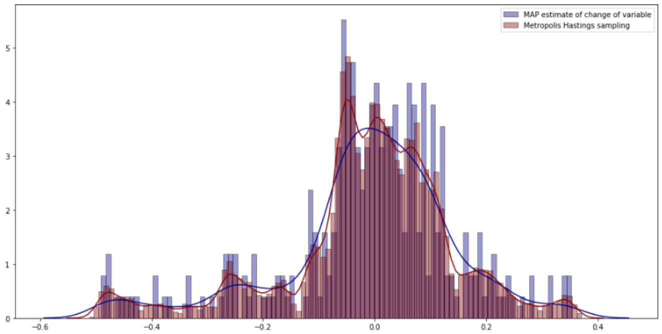

Fig. 4. Histogram comparing the distributions of δ_{MAP} (red) and samples (blue)

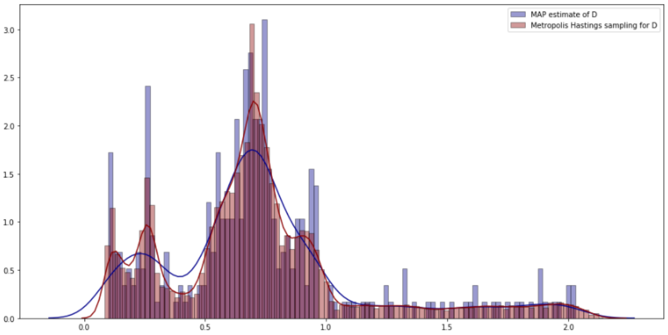

Fig. 5. Histogram comparing the distributions of D (red) and samples (blue)

5. DISCUSSION

We made the sparsity assumption for both D and H as:

1. D should contains only one Gaussian like peak in each column and the rest should be zeros.
2. We want to represent each column of the dataset with few combinations of these peaks

In short, the distribution is desired to be smooth and exponential. As is was seen from figures 1 and 2, the NMF with least squares captured the essential features such as hotspots and voigt locations. Furthermore, seen in figure 3 by applying a gaussian process prior, some features of the factor, D, are smooth and exponential however, one feature looks like noise while another seem to combine multiple peaks. The choice of optimization algorithm might be one reason why features of D are not entirely distinct. Another choice of optimization algorithm such as a gradient based approach for computing the maximum likelihood can be considered for yielding estimates that agrees with prior knowledge. The chosen sampling method, namely Metropolis-Hastings, is a class of Markov chain Monte Carlo sampling algorithms. One problem with the Metropolis Hastings algorithm is that it takes a random walk in the distribution and thus, large variance will impact the rejection rate of the algorithm which leads to a slow convergence. However, initialization by the MAP estimate while furthermore applying knowledge about the least squares estimate as acceptance criterion yielded decent sampling estimates. An implication with the choice of sampling method is that points are correlated and will at times not be similar to the actual distribution. Another problem is that the acceptance criterion for samples are based on the least squares likelihood *of* the MAP estimate and hence, the sampling will capture the distribution of the MAP estimate rather than the underlying distribution.

6. CONCLUSION

We implemented the NMF with Gaussian Process Priors as stated in [2] for simulated Raman spectroscopy data set by defining a link function and covariance matrix with respect to the distribution of factors D and H which were both assumed sparse and exponential. It was possible to factorize the data and reconstruct it by calculated components D and H that yields hotspots and voigt locations close to the underlying data. Furthermore, it was observed from the factor D that features did somewhat agree with assumptions made about sparsity and smoothness defined by the underlying Gaussian process. Moreover, a Metropolis-Hastings algorithm was implemented to learn about prior distributions of the factors. The acceptance criterion was based around the maximum likelihood estimate which as a consequence was mainly useful in terms of sampling from the MAP estimate. In this case, the data was simulated and hence, the prior knowledge of the data was already known.

7. FUTURE WORKS

2D co-variance matrices can be defined for D and H to improve the model as Raman dataset is in 2-dimensions and by defining a 1D covariance matrix we are not forcing points to be correlated correctly. Furthermore, a gradient based optimization method with respect to the least squares negative log likelihood can be of interest to implement to see if factors are able to better capture prior knowledge of the data such as sparsity and smoothness. Further investigation of sampling algorithms such as adaptive rejection sampling [9] can be looked at. This method is capable of generating independent samples whereas Metropolis-Hastings sampling will have to rely on throwing away samples in this matter.

8. ACKOWLEDGEMENT

Special thanks to Tommy Sonne Alstrm as supervisor of this project as well as assistants Juan Jos Rubio Guillamn and Maximillan Fornitz Vording. We also thank the people responsible for the paper introducing GPP-NMF [2]

9. REFERENCES

[1] Jim Clark, "The fingerprint region of an infra-red spectrum," .

[2] Mikkel N. Schmidt and Hans Laurberg, "Non-negative matrix factorization with gaussian process priors," in *Computational Intelligence and Neuroscience*, 2008.

[3] Nicolas Gillis, *The Why and How of Nonnegative Matrix Factorization.*

[4] Chistopher M. Bishop, *Pattern Recognition and Machine Learning*, p. 18, 2006.

[5] Saul A. Teukolsky et. al., "Numerical recipes in c: The art of scientific computing," Cambridge University England, 1992, vol. ISBN 0-521-43108-5, p. 994.

[6] Aaron Clauset, "A brief primer on probability distributions," Santa Fe Institute, 2011.

[7] Chistopher M. Bishop, *Pattern Recognition and Machine Learning*, p. 542, 2006.

[8] N. M. et. al. Temme, "Voigt function," in *Computational Intelligence and Neuroscience*. Cambridge University Press, 2010, vol. ISBN 978-0521192255.

[9] Chistopher M. Bishop, "Pattern recognition and machine learning," Springer, 2006, p. 530.