ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8. Available from http://www.i6doc.com/en/.

Bayesian Semi Non-negative Matrix Factorisation

Albert Vilamala, Alfredo Vellido and Lluís A. Belanche

Universitat Politècnica de Catalunya - Computer Science Department
Omega Building, Jordi Girona 1-3, 08034, Barcelona - Spain

Abstract. Non-negative Matrix Factorisation (NMF) has become a standard method for source identification when data, sources and mixing coefficients are constrained to be positive-valued. The method has recently been extended to allow for negative-valued data and sources in the form of Semi- and Convex-NMF. In this paper, we re-elaborate Semi-NMF within a full Bayesian framework. This provides solid foundations for parameter estimation and, importantly, a principled method to address the problem of choosing the most adequate number of sources to describe the observed data. The proposed Bayesian Semi-NMF is preliminarily evaluated here in a real neuro-oncology problem.

1 Introduction

Non-negative Matrix Factorisation (NMF) has quickly established itself [1] as a reliable source identification method when data, sources and mixing coefficients are constrained to be positive-valued. The method has recently been extended [2] to allow for negative-valued data and sources in the form of Semi- and Convex-NMF. In previous work, we have developed variants of these including Discriminant Convex-NMF [3] and Probabilistic Convex-NMF [4].

Given a matrix of real-valued observations $\mathbf{X} \in \mathbb{R}_{\pm}^{D\times N}$, where N is the number of instances and D the dimensionality, Semi-NMF decomposes this matrix as a linear combination of K D-dimensional sources of mixed sign $\mathbf{S} \in \mathbb{R}_{\pm}^{D\times K}$ and a matrix $\mathbf{H} \in \mathbb{R}_{+}^{K\times N}$ of non-negative mixing coefficients. This decomposition takes the form: $\mathbf{X} = \mathbf{SH} + \mathbf{E}$, where $\mathbf{E} \in \mathbb{R}_{\pm}^{D\times N}$ is the error matrix.

NMF was described within the Bayesian probability paradigm in [5]. Here, we re-elaborate Semi-NMF [2] within a full Bayesian framework. This provides solid foundations for parameter estimation and, importantly, a principled method to address the problem of choosing the most adequate number of sources to describe the observed data, which is of utmost importance in real applications. The proposed Bayesian Semi-NMF is preliminarily evaluated in a real neuro-oncology problem for which NMF methods have shown to yield relevant results over recent years [6].

2 Full Bayesian Semi Non-negative Matrix Factorisation

In this section, we provide a full Bayesian probabilistic formulation for Semi-NMF: elements of the source matrix $\mathbf{S}$ are encoded as samples from a Gaussian distribution, while the values of the mixing matrix $\mathbf{H}$ are conveniently obtained from an Exponential density. Residuals in $\mathbf{E}$ are assumed to be i.i.d.; specifically, normally distributed with zero mean. According to the Bayes rule, the joint posterior for the model is defined as:

$$p\left(\mathbf{S},\mathbf{H},\sigma^2 \mid \mathbf{X}\right) = \frac{p\left(\mathbf{X} \mid \mathbf{S},\mathbf{H},\sigma^2\right) \cdot p\left(\mathbf{S} \mid \boldsymbol{\theta}_S\right) \cdot p\left(\mathbf{H} \mid \boldsymbol{\theta}_H\right) \cdot p\left(\sigma^2 \mid \boldsymbol{\theta}_\sigma\right)}{p\left(\mathbf{X}\right)}. \tag{1}$$

Given that the marginal likelihood $p\left(\mathbf{X}\right)$ is constant with respect to the model parameters, $p\left(\mathbf{S},\mathbf{H},\sigma^2 \mid \mathbf{X}\right) \propto p\left(\mathbf{X} \mid \mathbf{S},\mathbf{H},\sigma^2\right) \cdot p\left(\mathbf{S} \mid \boldsymbol{\theta}_S\right) \cdot p\left(\mathbf{H} \mid \boldsymbol{\theta}_H\right) \cdot p\left(\sigma^2 \mid \boldsymbol{\theta}_\sigma\right)$, where:

$$\begin{aligned}
p\left(\mathbf{X} \mid \mathbf{S},\mathbf{H},\sigma^2\right) &= \prod_{d=1}^{D}\prod_{n=1}^{N} \mathcal{N}\left(\mathbf{X}_{d,n}; (\mathbf{SH})_{d,n}, \sigma^2\right), \\
p\left(\mathbf{S} \mid \boldsymbol{\theta}_S\right) &= \prod_{d=1}^{D}\prod_{k=1}^{K} \mathcal{N}\left(\mathbf{S}_{d,k}; \mu_o, \sigma_o^2\right), \\
p\left(\mathbf{H} \mid \boldsymbol{\theta}_H\right) &= \prod_{k=1}^{K}\prod_{n=1}^{N} \mathcal{E}\left(\mathbf{H}_{k,n}; \lambda_o\right);
\end{aligned}$$

$\mathcal{N}\left(x;\mu,\sigma^2\right) = (2\pi\sigma^2)^{-1/2}\exp\{-\left(x-\mu\right)^2/(2\sigma^2)\}$ and $\mathcal{E}\left(x;\lambda\right) = \lambda\exp\left\{-\lambda x\right\}$ being the Normal and Exponential densities, respectively. In addition, $\boldsymbol{\theta}_S = \{\mu_o, \sigma_o^2\}$ and $\boldsymbol{\theta}_H = \{\lambda_o\}$ are the hyperparameters for the source and mixing priors. Finally, the prior for the noise variance is appropriately chosen to be an Inverse Gamma of the form:

$$p\left(\sigma^2 \mid \boldsymbol{\theta}_\sigma\right) = \mathcal{IG}\left(\sigma^2; \alpha_o, \beta_o\right) = \frac{\beta_o^{\alpha_o}}{\Gamma(\alpha_o)}(\sigma^2)^{-\alpha_o-1}\exp\left(-\frac{\beta_o}{\sigma^2}\right),$$

with $\boldsymbol{\theta}_\sigma = \{\alpha_o, \beta_o\}$ as hyperparameters. From this joint posterior, we aim at estimating the marginal density of each $\mathbf{S}$ and $\mathbf{H}$ factor, but this involves the computation of an intractable integral. This intractability is overcome by deriving the following Markov Chain Monte Carlo (MCMC) sampling method.

2.1 Gibbs sampling approach

A Gibbs sampling method for our model is here derived as a particular instance of MCMC. It is of special interest when the calculation of either the joint posterior distribution, the marginal distribution of any subset of factors, or the expected value of any of the factors becomes intractable. Assuming that sampling from the full conditional posterior distribution is feasible, drawing a set of instances from this density converges to a sample from the joint posterior. If samples from the marginal distribution of a subset of factors are required, only the samples for that subset are kept; finally, the expected value of any factor can be computed by averaging over all its samples. For our problem, we are interested in the second output and, hence, we formulate the conditional density of $\mathbf{S}$, which is proportional to a Normal distribution multiplied by a Normal prior; that is:

$\mathcal{N}\left(x;\mu_p,\sigma_p^2\right) \propto \mathcal{N}\left(x;\mu,\sigma^2\right)\mathcal{N}\left(x;\mu_o,\sigma_o^2\right)$. Let $\mathbf{A}_{\setminus(i,j)}$ represent all elements of $\mathbf{A}$ except $\mathbf{A}_{i,j}$; the full conditional density of $\mathbf{S}_{d,k}$ is

$$p(\mathbf{S}_{d,k} \mid \mathbf{X}, \mathbf{S}_{\setminus(d,k)}, \mathbf{H}, \sigma^2) = \mathcal{N}\left(\mathbf{S}_{d,k};\mu_p,\sigma_p^2\right), \tag{2}$$

where

$$\mu_p = \sigma_p^2\left(\frac{\mu_o}{\sigma_o^2} + \frac{\sum_{n=1}^{N}\left(\mathbf{X}_{d,n} - \sum_{k'\neq k}\mathbf{S}_{d,k'}\mathbf{H}_{k',n}\right)\mathbf{H}_{k,n}}{\sigma^2}\right),\ \sigma_p^2 = \frac{\sigma^2\cdot\sigma_o^2}{\sigma^2 + \sigma_o^2\sum_{n=1}^{N}\mathbf{H}_{k,n}^2}.$$

Turning to the mixing matrix, the full conditional density of $\mathbf{H}$ is proportional to a Normal multiplied by an Exponential, which turns out to be a rectified Normal density of the form $\mathcal{R}\left(x;\mu_p,\sigma_p^2,\lambda_p\right) \propto \mathcal{N}\left(x;\mu,\sigma^2\right)\mathcal{E}\left(x;\lambda_o\right)$; that is:

$$p(\mathbf{H}_{k,n} \mid \mathbf{X}, \mathbf{S}, \mathbf{H}_{\setminus(k,n)}, \sigma^2) = \mathcal{R}\left(\mathbf{H}_{k,n};\mu_p,\sigma_p^2,\lambda_p\right), \tag{3}$$

where

$$\mu_p = \frac{\sum_{d=1}^{D}\left(\mathbf{X}_{d,n} - \sum_{k'\neq k}\mathbf{S}_{d,k'}\mathbf{H}_{k',n}\right)\mathbf{S}_{d,k}}{\sum_{d=1}^{D}\mathbf{S}_{d,k}^2},\ \sigma_p^2 = \frac{\sigma^2}{\sum_{d=1}^{D}\mathbf{S}_{d,k}^2},\ \lambda_p = \lambda_o.$$

Finally, the full conditional density of σ^2 is proportional to the product of a Normal and an Inverse-Gamma: $\mathcal{IG}\left(x;\alpha_p,\beta_p\right) \propto \mathcal{N}\left(x;\mu,\sigma^2\right)\mathcal{IG}\left(x;\alpha_o,\beta_o\right)$. Specifically:

$$p\left(\sigma^2 \mid \mathbf{X}, \mathbf{S}, \mathbf{H}\right) = \mathcal{IG}\left(\sigma^2;\alpha_p,\beta_p\right), \tag{4}$$

where

$$\alpha_p = \frac{DN}{2} + \alpha_o,\ \beta_p = \frac{\sum_{d=1}^{D}\sum_{n=1}^{N}\left[\mathbf{X}_{d,n} - (\mathbf{S}\mathbf{H})_{d,n}\right]^2}{2} + \beta_o.$$

The resulting Gibbs sampler procedure for the Bayesian Semi-NMF formulation is depicted in Algorithm 1. Details of derivations leading to the calculations of the full conditional densities and their parameterisation can be found in [7].

Algorithm 1 Bayesian Semi-NMF Gibbs sampler

```
1) Normalise data $\mathbf{X}$ (L$_2$-norm)
2) Randomly initialise $\mathbf{S}$, $\mathbf{H}$ and $\sigma^2$
3) For each sample $m \in \{1,\dots,M\}$
   a) For each $d \in \{1,\dots,D\}$ and $k \in \{1,\dots,K\}$:
      i) Sample $\mathbf{S}_{d,k}$ according to Eq. 2
   b) For each $k \in \{1,\dots,K\}$ and $n \in \{1,\dots,N\}$:
      i) Sample $\mathbf{H}_{k,n}$ according to Eq. 3
   c) Sample $\sigma^2$ according to Eq. 4
   d) Store $\mathbf{S}^{(m)} = \mathbf{S}; \mathbf{H}^{(m)} = \mathbf{H}; \sigma^{2(m)} = \sigma^2$
4) Return $\{\mathbf{S}^{(m)}, \mathbf{H}^{(m)}, \sigma^{2(m)}\}_{m=1}^{M}$
```

2.2 Marginal likelihood for model selection

Importantly, a Bayesian formulation allows model selection as informed choice of the most adequate number of sources, through the estimation of $p(\mathbf{X})$. In this section, Chib's method [8] is used to estimate $p(\mathbf{X})$ by using only posterior draws provided by the Gibbs sampler. By isolating $p(\mathbf{X})$ in Eq. 1, the computation of the resulting equation for any value (i.e., $\mathbf{\Phi}$) will result in a specific evaluation of the marginal likelihood at the $\mathbf{\Phi}$ point (selected to be a high density point for the most accurate estimation). Comparison among models (each using different numbers of sources) will entail comparing their $p(\mathbf{X})$ estimates at $\mathbf{\Phi}$: $p\left(\mathbf{X} \mid \mathbf{\Phi}\right)$. Obtaining the density at $\mathbf{\Phi}$ for any of the factors in the numerator is easy; difficulties arise when calculating $p\left(\mathbf{S}, \mathbf{H}, \sigma^2 \mid \mathbf{X}\right)$. Chib's method provides a solution by segmenting the parameters into B blocks and applying the chain rule to write $p\left(\mathbf{S}, \mathbf{H}, \sigma^2 \mid \mathbf{X}\right)$ as the product of B terms. That is:

$$p\left(\mathbf{\Phi} \mid \mathbf{X}\right) = p\left(\mathbf{\Phi}_1 \mid \mathbf{X}\right) \times p\left(\mathbf{\Phi}_2 \mid \mathbf{\Phi}_1, \mathbf{X}\right) \times \ldots \times p\left(\mathbf{\Phi}_B \mid \mathbf{\Phi}_1, \ldots, \mathbf{\Phi}_{B-1}, \mathbf{X}\right). \quad (5)$$

The blocks of parameters are chosen to be amenable to Gibbs sampling, such that each term is approximated by averaging over the conditional density:

$$p\left(\mathbf{\Phi}_b \mid \mathbf{\Phi}_1, \ldots, \mathbf{\Phi}_{b-1}, \mathbf{X}\right) \approx \frac{1}{M} \sum_{m=1}^{M} p\left(\mathbf{\Phi}_b \mid \mathbf{\Phi}_1, \ldots, \mathbf{\Phi}_{b-1}, \mathbf{\Phi}_{b+1}^{(m)}, \ldots, \mathbf{\Phi}_B^{(m)}, \mathbf{X}\right),$$

where M is the number of samples and $\left\{\mathbf{\Phi}_{b+1}^{(m)}, \ldots, \mathbf{\Phi}_B^{(m)}\right\}$ are Gibbs samples from $p\left(\mathbf{\Phi}_{b+1}, \ldots, \mathbf{\Phi}_B \mid \mathbf{\Phi}_1, \ldots, \mathbf{\Phi}_{b-1}, \mathbf{X}\right)$. In our setting, each column of $\mathbf{S}$, each row of $\mathbf{H}$ and σ^2 are selected to be the blocks in Eq. 5. Therefore, letting $\mathbf{A}^*$ represent a matrix of high density points; $\mathbf{A}_{:,i}$ correspond to all the values in the i-th column and $\mathbf{A}_{j,:}$ to all the values in the j-th row, and in logarithmic scale:

$$\begin{aligned} \log\left\{p\left(\mathbf{S}^*, \mathbf{H}^*, \sigma^{2^*} \mid \mathbf{X}\right)\right\} &= \log\left\{p\left(\mathbf{S}^*_{:,1} \mid \mathbf{X}\right)\right\} + \log\left\{p\left(\mathbf{S}^*_{:,2} \mid \mathbf{S}^*_{:,1}, \mathbf{X}\right)\right\} + \ldots \\ &+ \log\left\{p\left(\sigma^2 \mid \mathbf{S}^*_{:,1}, \ldots, \mathbf{S}^*_{:,K}, \mathbf{H}^*_{1,:}, \ldots, \mathbf{H}^*_{K,:}, \mathbf{X}\right)\right\}. \end{aligned}$$

In order to compute the Bayes Factor between two models, namely $\mathcal{M}_i$ and $\mathcal{M}_j$, each built using different number of sources K, we proceed to evaluate the marginal likelihood at $\left\{\mathbf{S}^*, \mathbf{H}^*, \sigma^{2*}\right\}$ for both models and compare them using $\hat{B}_{ij} = \exp\{\log \hat{p}\left(\mathbf{X} \mid \mathcal{M}_i\right) - \log \hat{p}\left(\mathbf{X} \mid \mathcal{M}_j\right)\}$. This Bayes Factor allows us to select the most adequate model out of a pool of models.

Matlab code of the proposed algorithms can be downloaded from `http://www.cs.upc.edu/~avilamala/resources/BayesianSNMF_Toolbox.zip`

3 Empirical evaluation

A preliminary evaluation of the method was performed using a real neuro-oncology problem. It entails tissue identification of different brain tumours from single-voxel proton magnetic resonance spectroscopy (SV-^{1}H-MRS) data with coherent mixed-sign signal amplitudes. Chib's method was used to estimate the most appropriate number of MRS sources and each of them was individually assessed. A confidence measure of the model estimations (90% interval around the signal) is also supplied (calculated as the 5^{th} to 95^{th} percentiles interval of the Gibbs samples). Data belong to the online-accessible and curated INTERPRET repository [9]. The 195 most clinically relevant spectral frequencies were selected for several types of tumour (78 glioblatomas -*gbm*-, 31 metastases -*met*- and 20 astrocytomas grade II -*ac*2-) as well as healthy tissue (15 cases -*nom*-).

Given that all data points were normalised (L_2-norm) prior to any treatment, the parameters for the priors were chosen to match data amplitude. These include $\mu_o = 0.01$ and $\sigma_o^2 = 0.2$ to limit the values of the sources $\mathbf{S}_{d,k}$ between -1 and 1 with $p > 0.95$; setting the $\lambda_o = 3$ to bound the values of the mixing matrix $\mathbf{H}_{k,n}$ to the $[0, 1]$ interval ($p > 0.95$); and $\alpha_o = 1; \beta_o = 0.001$ as flat priors for the noise variance σ^2. The number of samples M generated at each Gibbs sampler run was set to 100,000; the first 50,000 were discarded to allow *burn-in*.

	1	2	3	4	5
nom	**4.55**	3.82	2.94	2.70	1.95
gbm	24.31	**26.56**	25.89	26.32	26.40
met	**9.71**	8.68	8.67	8.62	8.58
*ac*2	6.49	**6.60**	6.15	5.73	5.44

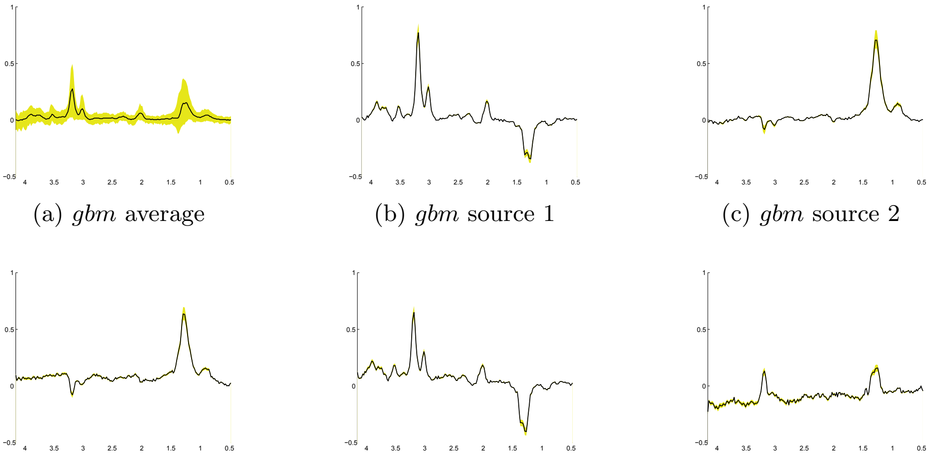
Table 1: Log of the marginal likelihood ($\times 10^3$) for different number of sources (1 to 5) for each tumour type and normal tissue. Best values highlighted in bold.

The results reported in Table 1 indicate that the values of the marginal likelihood for different number of sources obtained by the Chib's method clearly favour low complexity models with one or two sources. Note that this estimate of the *best* number of sources to represent the observed data from a source extraction viewpoint might not necessarily be the most adequate for interpretability purposes in the current application context. For instance, the homogeneity of healthy tissue makes the 1-source choice sensible. This is not the case for metastases, though, in which the 1-source choice might just reflect the characteristic necrotic tissue of this tumour type obscuring the relevance of lesser-intensity sources. Note also that this estimation does not preclude alternative choices, given that the marginal likelihood provides a real-valued measure, not a binary one; in other words, a relative -not an absolute- measure of relevance.

Results are especially interesting for *gbm*. Chib's method suggests two sources, which, in Fig.1 (*b*-*c*), clearly reflect necrotic (top row, right) and active (top row, centre) tumour tissue. When three sources (*d*-*f*) are arbitrarily extracted, the third one (bottom row, right) clearly accounts for little more than noise.

4 Conclusions

The derived full Bayesian Semi-NMF is proposed as the method of choice for NMF-based source extraction from mixed-sign data, particularly for problems



(a) *gbm* average (b) *gbm* source 1 (c) *gbm* source 2

(d) *gbm* source 1 (e) *gbm* source 2 (f) *gbm* source 3

Fig. 1: Sources identified by Bayesian Semi-NMF for *gbm*: Top row shows the average *gbm* spectrum, together with the two sources captured according to the model selection. Bottom row arbitrarily decomposes the signal into 3 sources. The black solid line represents the mean, while the shadowed region conforms the 90% credible interval (5^{th} to 95^{th} percentiles of Gibbs samples). Y-axes represent unit-free metabolite concentrations and X-axes represent frequency as measured in parts per million (ppm).

in which the choice of the most adequate number of sources is relevant from the point of view of knowledge discovery.

References

[1] D.D. Lee and H.S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

[2] C. Ding, T. Li, and M.I. Jordan. Convex and semi-nonnegative matrix factorizations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32:45–55, 2010.

[3] A. Vilamala et al. Discriminant convex non-negative matrix factorization for the classification of human brain tumours. *Pattern Recognition Letters*, 34(4):1734–1747, 2013.

[4] A. Vilamala, L.A. Belanche, and A. Vellido. A MAP approach for convex non-negative matrix factorization in the diagnosis of brain tumors. In *PRNI*, pages 1–4, 2014.

[5] M.N. Schmidt, O. Winther, and L.K. Hansen. Bayesian non-negative matrix factorization. In *Independent Component Analysis and Signal Separation*, LNCS 5441, pages 540–547. Springer, 2009.

[6] T. Laudadio et al. NMF in MR spectroscopy. In *Non-negative Matrix Factorization Techniques*, pages 161–177. Springer, 2016.

[7] A. Vilamala. *Multivariate Methods for Interpretable Analysis of Magnetic Resonance Spectroscopy Data in Brain Tumour Diagnosis.* PhD thesis, UPC BarcelonaTech, 2015.

[8] S. Chib. Marginal likelihood from the gibbs output. *Journal of the American Statistical Association*, 90(432):1313–1321, 1995.

[9] M. Julià-Sapé et al. A multi-centre, web-accessible and quality control-checked database of in vivo MR spectra of brain tumour patients. *Magnetic Resonance Materials in Physics, Biology and Medicine (MAGMA)*, 19:22–33, 2006.