Bayesian non-negative matrix factorization

Mikkel N. Schmidt[1], Ole Winther[2], and Lars Kai Hansen[2]

[1] University of Cambridge, Department of Engineering, mns@imm.dtu.dk
[2] Technical University of Denmark, DTU Informatics, {owi,lkh}@imm.dtu.dk

Abstract. We present a Bayesian treatment of non-negative matrix factorization (NMF), based on a normal likelihood and exponential priors, and derive an efficient Gibbs sampler to approximate the posterior density of the NMF factors. On a chemical brain imaging data set, we show that this improves interpretability by providing uncertainty estimates. We discuss how the Gibbs sampler can be used for model order selection by estimating the marginal likelihood, and compare with the Bayesian information criterion. For computing the maximum a posteriori estimate we present an iterated conditional modes algorithm that rivals existing state-of-the-art NMF algorithms on an image feature extraction problem.

1 Introduction

Non-negative matrix factorization (NMF) [1, 2] has recently received much attention as an unsupervised learning method for finding meaningful and physically interpretable latent variable decompositions. The constraint of non-negativity is natural for a wide range of natural signals, such as pixel intensities, amplitude spectra, and occurence counts. NMF has found widespread application in many areas, and has for example been used in environmetrics [1] and chemometrics [3] to find underlying explanatory sources in series of chemical concentration measurements; in image processing [2] to find useful features in image databases; in text processing [4] to find groups of words that constitute latent topics in sets of documents; and in audio processing [5] to separate mixtures of audio sources.

In this paper, we discuss NMF in a Bayesian framework. Most NMF algorithms can be seen as computing a maximum likelihood (ML) or maximum a posteriori (MAP) estimate of the non-negative factorizing matrices under some assumptions on the distribution of the data and the factors. Here, we derive an efficient Markov chain Monte Carlo (MCMC) method for estimating their posterior density, based on a Gibbs sampling procedure. This gives not only an estimate of the factors, but also an estimate of their marginal posterior density, which is valuable for interpreting the factorization, computing uncertainty estimates, etc. This work is related to the Bayesian spectral decomposition (BSD) method of Ochs et al. [6], that uses a (computationally expensive) atomic point-mass prior and is implemented in a modified commercial MCMC toolbox; and to the Bayesian non-negative source separation method of Moussaoui et al. [7] that incorporates a hybrid Gibbs-Metropolis-Hastings sampling procedure. The contributions of this paper are three-fold: 1) We present a fast and direct Gibbs

sampling procedure for the NMF problem that, compared with BSD, reduces computation time by more than an order of magnitude on the same data. 2) We present a marginal-likelihood estimation-method based on the Gibbs sampler, which leads to a novel model order selection method for NMF. 3) We propose an iterated conditional modes algorithm for computing the MAP estimate of the Bayesian NMF, and show that this algorithm rivals current state-of-the-art NMF algorithms. Matlab implementations of the presented algorithms are available at `http://www.mikkelschmidt.dk/ica2009`.

2 Bayesian non-negative matrix factorization

The non-negative matrix factorization problem can be stated as $\boldsymbol{X} = \boldsymbol{A}\boldsymbol{B} + \boldsymbol{E}$, where $\boldsymbol{X} \in \mathbb{R}^{I \times J}$ is a data matrix that is factorized as the product of two element-wise non-negative matrices, $\boldsymbol{A} \in \mathbb{R}_+^{I \times N}$ and $\boldsymbol{B} \in \mathbb{R}_+^{N \times J}$ ($\mathbb{R}_+$ denotes the non-negative reals), and $\boldsymbol{E} \in \mathbb{R}^{I \times J}$ is a residual matrix. In the Bayesian framework, we state our knowledge of the distribution of the residual in terms of a likelihood function, and the parameters in terms of prior densities. The priors are chosen in accordance with our beliefs about the distribution of the parameters; however, to allow efficient inference in the model it is desirable to choose prior densities with a convenient parametric form. In this paper, we choose a normal likelihood and exponential priors, which are suitable for a wide range of problems, while permitting an efficient Gibbs sampling procedure. We assume that the residuals, $E_{i,j}$, are i.i.d. zero mean normal with variance σ^2, which gives rise to the likelihood,

$$p(\boldsymbol{X}|\boldsymbol{\theta}) = \prod_{i,j} \mathcal{N}\left(X_{i,j}; (\boldsymbol{A}\boldsymbol{B})_{i,j}, \sigma^2\right), \tag{1}$$

where $\boldsymbol{\theta} = \{\boldsymbol{A}, \boldsymbol{B}, \sigma^2\}$ denotes all parameters in the model and $\mathcal{N}(x; \mu, \sigma^2) = (2\pi\sigma^2)^{-1/2} \exp(-(x-\mu)^2/(2\sigma^2))$ is the normal density. We assume $\boldsymbol{A}$ and $\boldsymbol{B}$ are independently exponentially distributed with scales $\alpha_{i,n}$ and $\beta_{n,j}$,

$$p(\boldsymbol{A}) = \prod_{i,n} \mathcal{E}\left(A_{i,n}; \alpha_{i,n}\right), \qquad p(\boldsymbol{B}) = \prod_{n,j} \mathcal{E}\left(B_{n,j}; \beta_{n,j}\right), \tag{2}$$

where $\mathcal{E}(x; \lambda) = \lambda \exp(-\lambda x) u(x)$ is the exponential density, and $u(x)$ is the unit step function. The prior for the noise variance is chosen as an inverse gamma density with shape k and scale θ,

$$p(\sigma^2) = \mathcal{G}^{-1}\left(\sigma^2; k, \theta\right) = \frac{\theta^k}{\Gamma(k)} (\sigma^2)^{-k-1} \exp\left(-\frac{\theta}{\sigma^2}\right). \tag{3}$$

Using Bayes rule, the posterior is proportaional to the product of Equations (1–3), and it can be maximized to yield an estimate of $\boldsymbol{A}$ and $\boldsymbol{B}$; however, we are interested in estimating the marginal density of the factors, and because we cannot directly compute marginals by integrating the posterior, we proceed in the next section by deriving an MCMC sampling method.

2.1 Gibbs sampling

In Gibbs sampling, a sequence of samples is drawn from the conditional posterior densities of the model parameters, and this converges to a sample from the joint posterior. We first consider the conditional densities of $\boldsymbol{A}$ and $\boldsymbol{B}$ which are proportional to a normal multiplied by an exponential, i.e., a rectified normal density, which we denote by $\mathcal{R}\left(x;\mu,\sigma^2,\lambda\right) \propto \mathcal{N}(x;\mu,\sigma^2)\mathcal{E}(x;\lambda)$. The conditional density of $A_{i,n}$ is

$$p(A_{i,n}|\boldsymbol{X}, A_{\setminus(i,n)}, \boldsymbol{B}, \sigma^2) = \mathcal{R}\left(A_{i,n}; \mu_{A_{i,n}}, \sigma^2_{A_{i,n}}, \alpha_{i,n}\right), \tag{4}$$

$$\mu_{A_{i,n}} = \frac{\sum_j \left(\boldsymbol{X}_{i,j} - \sum_{n' \neq n} \boldsymbol{A}_{i,n'}\boldsymbol{B}_{n',j}\right)\boldsymbol{B}_{n,j}}{\sum_j \boldsymbol{B}^2_{n,j}}, \quad \sigma^2_{A_{i,n}} = \frac{\sigma^2}{\sum_j \boldsymbol{B}^2_{n,j}}, \tag{5}$$

where $A_{\setminus(i,n)}$ denotes all elements of $\boldsymbol{A}$ except $A_{i,n}$, and due to symmetry, the similar expression for $B_{n,j}$ can easily be derived. The conditional density of σ^2 is an inverse-gamma

$$p(\sigma^2|\boldsymbol{X}, \boldsymbol{A}, \boldsymbol{B}) = \mathcal{G}^{-1}(\sigma^2; k_{\sigma^2}, \theta_{\sigma^2}) \tag{6}$$

$$k_{\sigma^2} = \frac{IJ}{2} + 1 + k, \quad \theta_{\sigma^2} = \frac{1}{2}\sum_{i,j}(\boldsymbol{X} - \boldsymbol{A}\boldsymbol{B})^2_{i,j} + \theta. \tag{7}$$

The posterior can now be approximated by sequentially sampling from these conditional densities.

A few remarks on the implementation: Since the elements in each column of $\boldsymbol{A}$ (row of $\boldsymbol{B}$) are conditionally independent, we can samples an entire column of $\boldsymbol{A}$ (row of $\boldsymbol{B}$) simultaneously. When $I \times J \gg (I + J) \times N$ it advantageous to implement equations (5) and (7) in a way that avoids explicitly computing large matrix products of size $I \times J$. The bulk of the computation is then comprised of computing the matrix products $\boldsymbol{X}\boldsymbol{B}^\top$ and $\boldsymbol{A}^\top\boldsymbol{X}$ that can be precomputed in each iteration. Based on this, an efficient NMF Gibbs sampler is given as Algorithm 1, where R and G^{-1} denotes drawing a random sample from the rectified normal and inverse-gamma densities, and the notation $\boldsymbol{A}_{:,\setminus n}$ is used to denote the submatrix of $\boldsymbol{A}$ that consists of all colums except the n'th.

2.2 Estimating the marginal likelihood

An important problem in NMF is to choose the number of factors, N, for a given data set (when this is not evident from the nature of the data). In the Bayesian framework, model selection can be performed by evaluating the marginal likelihood, $p(\boldsymbol{X})$, which involves an intractable integral over the posterior. Several methods exist for estimating the marginal likelihood, including annealed importance sampling, bridge sampling, path sampling, and Chib's method [8]. The latter is of particular interest here since it requires only posterior draws, and can thus be implemented directly using the described Gibbs sampler.

Chib's method is based on the relation $p(\boldsymbol{X}) = \frac{p(\boldsymbol{X}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\boldsymbol{\theta}|\boldsymbol{X})}$. The numerator can easily be evaluated for any $\boldsymbol{\theta}$, so the problem is to evaluate the

denominator, i.e., the the posterior density at the point $\boldsymbol{\theta}$. If the parameters are segmented into K blocks, $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K\}$, using the chain rule we may write the denominator as the product of K terms, $p(\boldsymbol{\theta}|\boldsymbol{X}) = p(\boldsymbol{\theta}_1|\boldsymbol{X}) \times p(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1, \boldsymbol{X}) \times \cdots \times p(\boldsymbol{\theta}_K|\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{K-1}, \boldsymbol{X})$. If these K parameter blocks are chosen such that they are amenable to Gibbs sampling, each term can be approximated by averaging over the conditional density $p(\boldsymbol{\theta}_k|\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{k-1}, \boldsymbol{X}) \approx \frac{1}{M}\sum_{m=1}^{M} p(\boldsymbol{\theta}_k|\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{k-1}, \boldsymbol{\theta}_{k+1}^{(m)}, \ldots, \boldsymbol{\theta}_K^{(m)}, \boldsymbol{X})$, where $\boldsymbol{\theta}_{k+1}^{(m)}, \ldots, \boldsymbol{\theta}_K^{(m)}$ are Gibbs samples from $p(\boldsymbol{\theta}_{k+1}, \ldots, \boldsymbol{\theta}_K|\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{k-1}, \boldsymbol{X})$. Thus, the marginal likelihood can be evaluated by K runs of the Gibbs sampler. This procedure is valid for any value of $\boldsymbol{\theta}$, but the estimation is most accurate when $\boldsymbol{\theta}$ is chosen as a high density point, e.g., the posterior mode.

In the NMF problem, the columns of $\boldsymbol{A}$ and rows of $\boldsymbol{B}$ can be used as the parameter blocks, and the marginal likelihood can thus be estimated by $2N$ runs of the Gibbs sampler, which makes this method attractive especially for NMF models with a small number of components.

2.3 An iterated conditional modes algorithm

With the conditional densities of the parameters in the NMF model in place, an efficient iterated conditional modes (ICM) [9] algorithm can be derived for computing the MAP estimate. In this approach, we iterate over the parameters of the model, but instead of drawing random samples from the conditionals, as in the Gibbs sampler, we set each parameter equal to the conditional mode, and after a number of iterations the algorithm converges to a local maximum of the joint posterior density. This forms a block coordinate ascent type algorithm with the benefit that the optimum is computed for each block of parameters in each iteration. Since the modes of the conditional densities have closed-form expressions, the ICM algorithm has a low computational cost per iteration. The ICM NMF is given as Algorithm 2. In the algorithm, P_+ sets negative elements of its argument to zero.

3 Experimental evaluations

3.1 Analysis of chemical shift brain imaging data

We demonstrate the proposed bayesian NMF method on a chemical shift imaging (CSI) data set [6], that consists of 369-dimensional spectra measured at 256 positions in a human head. The measured spectra are mixtures of different underlying spectra, and the NMF decomposition finds these as the columns of $\boldsymbol{A}$ and the corresponding active positions in the head as the rows of $\boldsymbol{B}$. Ochs et al. [6] demonstrate that the data is well described by two components that correspond to brain tissue and muscle tissue, and present a bilinear MCMC method that provides physically meaningful results but takes several hours to compute. Sajda et al. [10, 3] demonstrate, on the same data set, that a constrained NMF method provides meaningful results, and Schmidt and Laurberg [11] extend the NMF approach by including advanced prior densities.

In our experiments, the priors for $\boldsymbol{A}$ and $\boldsymbol{B}$ were chosen as $\alpha_{i,n} = 1$ and $\beta_{n,j} = 10^{-6}$ to match the amplitude of the data. For the noise variance we used an uninformative prior, $k = \theta = 0$. We sampled $40,000$ points from the posterior and discarded the first half to allow for the sampler to burn in. The computations took 80 seconds on a 2 Ghz Pentium 4 computer. Since the samples were not guaranteed to correspond to the same permutation of the factors, we used the area around 0 ppm to identify the two unique sources. We then computed the mean and the 5'th and 95'th percentile of the marginal distributions of the resolved spectra, and for comparison we computed the MAP estimate using the ICM algorithm, that provides a solutions almost identical to the results of Sajda et al. [3] (see Figure 1).

Uncertainty in NMF can be caused by noise in the data, but since NMF in general is not unique multiple competing solutions may also be reflected in the posterior. Also, because of the bilinear structure of NMF, uncertainties may be correlated between $\boldsymbol{A}$ and $\boldsymbol{B}$, which can not be seen in plots of marginal distributions, but can be assesed through further analysis of the posterior density.

3.2 NMF model order selection

To demonstrate the proposed model order selection method, we generated a data matrix by multiplying two random unit mean i.i.d. exponential distributed matrices with $I = 100$, $J = 20$, and $N = 3$ and adding unit variance zero mean i.i.d. normal noise. Using Chib's method, we computed the marginal likelihood for model orders between 1 and 5. We generated $20,000$ samples per parameter

Algorithm 1: Gibbs sampler

for $m = 1$ to M **do**
- $\boldsymbol{C} = \boldsymbol{B}\boldsymbol{B}^\top$, $\boldsymbol{D} = \boldsymbol{X}\boldsymbol{B}^\top$
- **for** $n = 1$ to N **do**
  - $\boldsymbol{A}_{:,n} \leftarrow \mathrm{R}\left(\boldsymbol{a}_n, \frac{\sigma^2}{C_{n,n}}, \boldsymbol{\alpha}_{:,n}\right)$
- **end**
- $\sigma^2 \leftarrow \mathrm{G}^{-1}\left(\frac{IJ}{2} + k + 1, \chi + \theta + \xi\right)$
- $\boldsymbol{E} = \boldsymbol{A}^\top\boldsymbol{A}$, $\boldsymbol{F} = \boldsymbol{A}^\top\boldsymbol{X}$
- **for** $n = 1$ to N **do**
  - $\boldsymbol{B}_{n,:} \leftarrow \mathrm{R}\left(\boldsymbol{b}_n, \frac{\sigma^2}{E_{n,n}}, \boldsymbol{\beta}_{n,:}\right)$
- **end**
- $\boldsymbol{A}^{(m)} \leftarrow \boldsymbol{A}, \quad \boldsymbol{B}^{(m)} \leftarrow \boldsymbol{B}$

end

Output: $\left\{\boldsymbol{A}^{(m)}, \boldsymbol{B}^{(m)}\right\}_{m=1}^{M}$

Algorithm 2: ICM

repeat
- $\boldsymbol{C} = \boldsymbol{B}\boldsymbol{B}^\top$, $\boldsymbol{D} = \boldsymbol{X}\boldsymbol{B}^\top$
- **for** $n = 1$ to N **do**
  - $\boldsymbol{A}_{:,n} = P_+(\boldsymbol{a}_n)$
- **end**
- $\sigma^2 = \dfrac{\theta + \chi + \xi}{\frac{IJ}{2} + k + 1}$
- $\boldsymbol{E} = \boldsymbol{A}^\top\boldsymbol{A}$, $\boldsymbol{F} = \boldsymbol{A}^\top\boldsymbol{X}$
- **for** $n = 1$ to N **do**
  - $\boldsymbol{B}_{n,:} = P_+\left(\boldsymbol{b}_n\right)$
- **end**

until convergence

Output: $\boldsymbol{A}$, $\boldsymbol{B}$

Definitions:

$$\chi = \frac{1}{2}\sum_{i,j} \boldsymbol{X}_{i,j}^2, \quad \xi = \frac{1}{2}\sum_{i,n} \boldsymbol{A}_{i,n}\left(\boldsymbol{A}\boldsymbol{C} - 2\boldsymbol{D}\right)_{i,n}$$

$$\boldsymbol{a}_n = \frac{\boldsymbol{D}_{:,n} - \boldsymbol{A}_{:,\backslash n}\boldsymbol{C}_{\backslash n,n} - \boldsymbol{\alpha}_{:,n}\,\sigma^2}{C_{n,n}}, \quad \boldsymbol{b}_n = \frac{\boldsymbol{F}_{n,:} - \boldsymbol{E}_{n,\backslash n}\boldsymbol{B}_{\backslash n,:} - \boldsymbol{\beta}_{n,:}\,\sigma^2}{E_{n,n}}$$

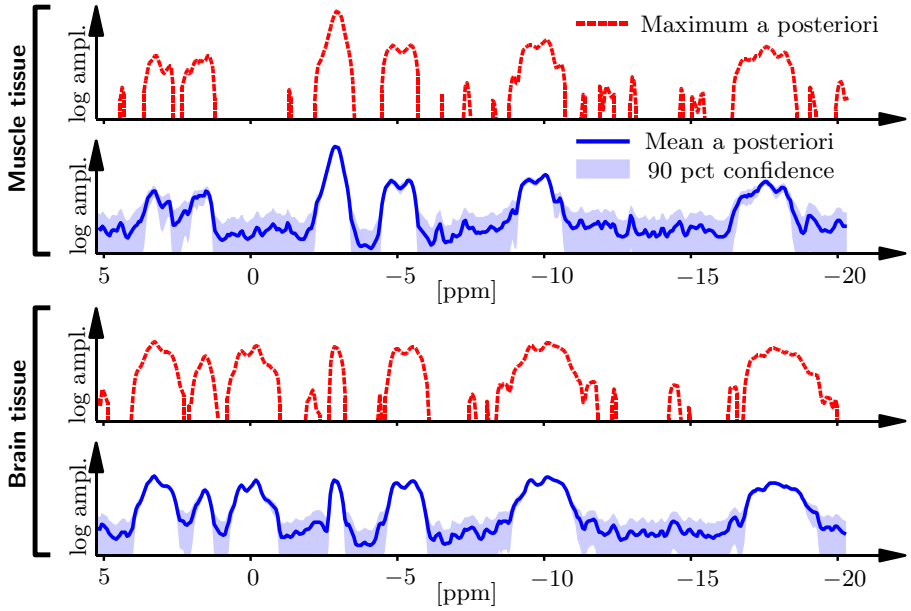

Fig. 1. Analysis of chemical shift imaging data. Two components are identified corresponding to (top) muscle and (bottom) brain tissue. MAP estimation provides a point estimate of the components whereas Gibbs sampling gives full posterior marginals, that provide an uncertainty estimate on the spectra, and leads to better interpretation of the results. For example, the confidence intervals show, that many of the low amplitude peaks in the MAP spectra may not be significant.

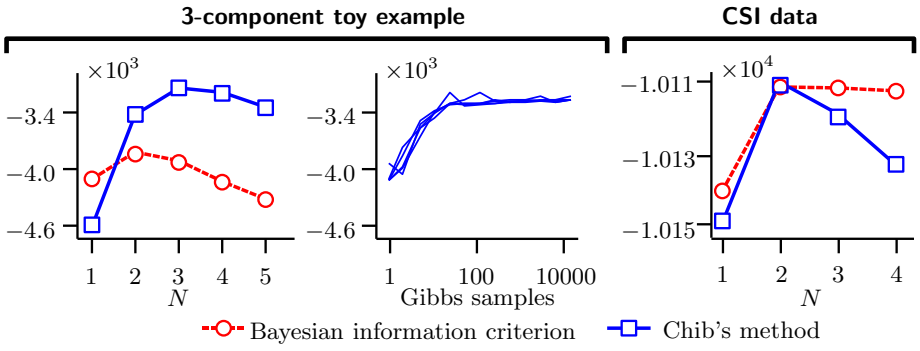

Fig. 2. NMF model order selection using Chib's method and the Bayesian information criterion (BIC). Left: A three-component toy example demonstrates that the method finds the correct number of components where BIC fails due to the small sample size. Center: Several runs of the algorithm suggest that the estimate of the marginal likelihood is stable after a few thousand Gibbs samples. Right: Analysis of the chemical shift imaging data confirms that it contains two spectral components. In this experiment, Chib's method and BIC were in agreement.

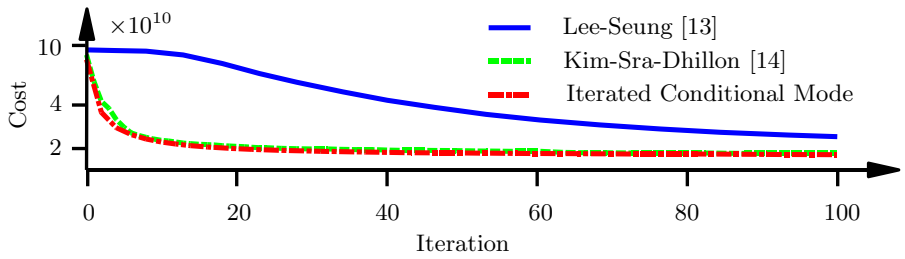

Fig. 3. Convergence rate for three least squares NMF algorithms on an image feature extraction problem. The ICM algorithm converges much faster than the Lee-Seung algorithm and with similar rate per iteration as the Kim-Sra-Dhillon algorithm.

block, and discarded the first half to allow burn-in; several runs of the algorithm suggested that this was sufficient for the sampler to stabilize. For comparison, we computed the Baysian information criterion (BIC), where the number of effective parameters was chosen as the number of non-zero parameters in the MAP estimate. The marginal likelihood attained its maximum at the correct model order, $N = 3$, whereas BIC favored a simpler model, $N = 2$. The reason why BIC fails in this case is the small number of samples in the data set, and our results suggest that Chib's method is more robust. Next, we applied the marginal likelihood estimation technique to the CSI data set described in the previous section, and here Chib's method and BIC agreed in confirming that the data contains two spectral components (see Figure 2).

3.3 Image feature extraction

To compare our ICM algorithm to other methods, we computed an $N = 32$ components factorization of the cropped UMIST Face Database [12] that consists of 564 images of size 92×112 that were vectorized to form a data matrix of size $I = 10304 \times J = 564$. To be able to directly compare with existing least squares NMF methods, we used a flat prior, $\alpha_{i,n} = \beta_{n,j} = k = \theta = 0$. We compared with two state-of-the-art methods: Lee and Seung's multiplicative update algorithm [13] and Kim, Sra, and Dhillon's fast Newton algorithm (FNMA$^{\mathrm{I}}$) [14].

The results (see Figure 3) show that the ICM algorithm converges much faster than the Lee-Seung algorithm and with approximately the same rate per iteration as the Kim-Sra-Dhillon algorithm. Since all three algorithms are dominated by the computation of the same matrix products they have approximately the same computational cost per iteration.

4 Conclusions

We have taken a Bayesian approach to NMF and presented a fast MCMC sampling procedure for approximating the posterior density, and we have showed

that this can be valuable for the interpretation of the non-negative factors recovered in NMF. The sampling procedure can also directly be used to estimate the marginal likelihood, which is useful for model order selection. Finally, we have presented an iterated conditional modes algorithm for computing the MAP estimate, that rivals existing state-of-the-art NMF algorithms.

Acknowledgments We thank Paul Sajda and Truman Brown for making the chemical shift brain imaging data set available to us.

References

1. Paatero, P., Tapper, U.: Positive matrix factorization: A nonnegative factor model with optimal utilization of error-estimates of data values. Environmetrics **5**(2) (Jun 1994) 111–126
2. Lee, D.D., Seung, H.S.: Learning the parts of objects by non-negative matrix factorization. Nature **401**(6755) (1999) 788–791
3. Sajda, P., Du, S., Parra, L.: Recovery of constituent spectra using non-negative matrix factorization. In: Wavelets: Applications in Signal and Image Processing, Proceedings of SPIE. Volume 5207. (2003) 321–331
4. Gaussier, E., Goutte, C.: Relation between PLSA and NMF and implication. In: Research and Development in Information Retrieval, Proceedings of the International SIGIR Conference on. (2005) 601–602
5. Schmidt, M.N., Olsson, R.K.: Single-channel speech separation using sparse non-negative matrix factorization. In: Spoken Language Processing, ISCA International Conference on (INTERSPEECH). (2006)
6. Ochs, M.F., Stoyanova, R.S., Arias-Mendoza, F., Brown, T.R.: A new method for spectral decomposition using a bilinear bayesian approach. Journal of Magnetic Resonance **137** (1999) 161–176
7. Moussaoui, S., Brie, D., Mohammad-Djafari, A., Carteret, C.: Separation of non-negative mixture of non-negative sources using a Bayesian approach and MCMC sampling. Signal Processing, IEEE Transactions on **54**(11) (Nov 2006) 4133–4145
8. Chib, S.: Marginal likelihood from the Gibbs output. Journal of the American Statistical Association **90**(432) (Dec 1995) 1313–1321
9. Besag, J.: On the statistical analysis of dirty pictures. Royal Statistical Society, Journal of the **48**(3) (1986) 259–302
10. Sajda, P., Du, S., Brown, T.R., Stoyanova, R., Shungu, D.C., Mao, X., Parra, L.C.: Nonnegative matrix factorization for rapid recovery of constituent spectra in magnetic resonance chemical shift imaging of the brain. Medical Imaging, IEEE Transactions on **23**(12) (Dec 2004) 1453–1465
11. Schmidt, M.N., Laurberg, H.: Nonnegative matrix factorization with gaussian process priors. Computational Intelligence and Neuroscience **2008** (Feb 2008)
12. Graham, D.B., Allinson, N.M.: Characterizing virtual eigensignatures for general purpose face recognition. In: Face Recognition: From Theory to Applications. Volume 163 of Computer and Systems Sciences. (1998) 446–456
13. Lee, D.D., Seung, H.S.: Algorithms for non-negative matrix factorization. In: Neural Information Processing Systems, Advances in (NIPS). (2000) 556–562
14. Kim, D., Sra, S., Dhillon, I.S.: Fast Newton-type methods for the least squares nonnegative matrix approximation problem. In: Data Mining, Proceedings of SIAM Conference on. (2007)