MAXIMUM MARGINAL LIKELIHOOD ESTIMATION FOR NONNEGATIVE DICTIONARY LEARNING

Onur Dikmen and Cédric Févotte

CNRS LTCI; Télécom ParisTech
37-39, rue Dareau, 75014, Paris, France
{dikmen, fevotte}@telecom-paristech.fr

ABSTRACT

We describe an alternative to standard nonnegative matrix factorisation (NMF) for nonnegative dictionary learning. NMF with the Kullback-Leibler divergence can be seen as maximisation of the *joint likelihood* of the dictionary and the expansion coefficients under Poisson observation noise. This approach lacks optimality because the number of parameters (which include the expansion coefficients) grows with the number of observations. As such, we describe a variational EM algorithm for optimisation of the *marginal likelihood*, i.e., the likelihood of the dictionary where the expansion coefficients have been integrated out (given a Gamma conjugate prior). We compare the output of both maximum joint likelihood estimation (i.e., standard NMF) and maximum marginal likelihood estimation (MMLE) on real and synthetical data. The MMLE approach is shown to embed automatic model order selection, similar to automatic relevance determination.

Index Terms— Nonnegative matrix factorisation, variational EM, model order selection, automatic relevance determination, sparse coding.

1. INTRODUCTION

Nonnegative matrix factorisation (NMF) [1] is a popular method for nonnegative dictionary learning based on matrix decomposition. The goal is to approximate a $F \times N$ nonnegative matrix $\boldsymbol{V}$ as the product of two nonnegative matrices, $\boldsymbol{W}$ and $\boldsymbol{H}$, of sizes $F \times K$ and $K \times N$, respectively. These two matrices can be estimated via minimising a measure of fit between $\boldsymbol{V}$ and $\boldsymbol{WH}$. One such popular measure is the (generalized) Kullback-Leibler (KL) divergence

$$D_{KL}(\boldsymbol{A}|\boldsymbol{B}) = \sum_{f=1}^{F}\sum_{n=1}^{N}\left(a_{fn}\log\frac{a_{fn}}{b_{fn}} - a_{fn} + b_{fn}\right), \qquad (1)$$

which is always nonnegative, convex with respect to (w.r.t) each factor (but not w.r.t both factors jointly) and is equal to zero if and only if $\boldsymbol{A} = \boldsymbol{B}$. Minimisation of the fit w.r.t the factors can be carried out with a fast, iterative algorithm based on multiplicative updates as described in [1]. This approach also coincides with the maximum likelihood estimation of $\boldsymbol{W}$ and $\boldsymbol{H}$ when $\boldsymbol{V}$ is assumed generated by a Poisson observation model, as will be later recalled. A criticism of NMF for nonnegative dictionary learning is that little can be said about the asymptotical optimality of the learnt dictionary $\boldsymbol{W}$. This is because the total number of parameters $FK + KN$ considered for maximum likelihood estimation grows with the number of observations N. As such, in this paper we seek to optimise the marginal likelihood of $\boldsymbol{W}$ given by

$$p(\boldsymbol{V}|\boldsymbol{W}) = \int_{\boldsymbol{H}} p(\boldsymbol{V}|\boldsymbol{W},\boldsymbol{H})p(\boldsymbol{H})\,d\boldsymbol{H}\,, \qquad (2)$$

where $p(\boldsymbol{H})$ is an assumed prior distribution of the expansion coefficients. Our approach is similar in spirit to independent component analysis (ICA), e.g. [2], where the likelihood of the "mixing matrix" is obtained through marginalisation of the latent independent components. This paper describes a variational EM algorithm for (approximate) maximum likelihood estimation on the marginal likelihood (2). We concentrate on the Poisson observation model and assume a conjugate Gamma prior for $\boldsymbol{H}$, but our approach can be extended to other statistical models employed in NMF, such as additive Gaussian or multiplicative Gamma observation models [3, 4, 5].

The rest of this paper is organised as follows. Section 2 describes the generative data model, Section 3 presents the two dictionary estimators considered in this paper and Section 4 describes algorithms. Section 5 reports results on real and synthetical data and in particular illustrates a very desirable feature of the marginal likelihood approach: automatic order selection. Section 6 concludes.

2. MODEL

The generative model assumed for the observations $v_{fn} = [\boldsymbol{V}]_{fn}$ is

$$v_{fn} \sim \mathcal{P}(v_{fn}|\sum_{k} w_{fk}h_{kn})\,, \qquad (3)$$

where $\mathcal{P}$ denotes the Poisson distribution, defined by $\mathcal{P}(x|\lambda) = \exp(-\lambda)\,\lambda^x/x!$, $x = 0, 1, 2, \ldots$ The data is assumed independently distributed conditionally upon $\boldsymbol{W}$ and $\boldsymbol{H}$. Using the superposition property of the Poisson distribution, the generative model can equivalently be written as a *composite model* such that

$$v_{fn} = \sum_{k=1}^{K} c_{k,fn}, \quad c_{k,fn} \sim \mathcal{P}(c_{k,fn}|w_{fk}h_{kn})\,, \qquad (4)$$

where the components $c_{k,fn}$ act as *latent variables* that will be used in the variational EM algorithm described in Section 4.2.

We further take the expansion coefficients h_{kn} to be random variables with Gamma prior, such that $h_{kn} \sim \mathcal{G}(h_{kn}|\alpha_k, \beta_k)$, where $\mathcal{G}(x|\alpha,\beta) = \beta^\alpha/\Gamma(\alpha)\,x^{\alpha-1}\,\exp(-\beta\,x)$, $x \geq 0$. The Gamma distribution is a prior of choice for its conjugacy with the Poisson distribution, and will facilitate some algorithm derivations to be presented next. Under these assumptions our model coincides with the

This work in supported by project ANR-09-JCJC-0073-01 TANGERINE (Theory and applications of nonnegative matrix factorisation).

Gamma-Poisson (GaP) model of [6, 7] which has been used in text analysis. In the rest of the paper the shape parameters α_k are fixed (in particular, we will use the value $\alpha_k = 1$ in the experiments, corresponding to the sparse-inducing exponential distribution). The scale parameters are also fixed, so as to remedy the scale ambivalence between column k of $\boldsymbol{W}$ and row k of $\boldsymbol{H}$. No constraint is imposed on $\boldsymbol{W}$, which is in our setting a free deterministic parameter.

3. ESTIMATORS

Given the model introduced in Section 2, we are interested in the following two estimators.

3.1. Maximum joint likelihood estimation (MJLE)

The joint (penalised) log-likelihood likelihood of $\boldsymbol{W}$ and $\boldsymbol{H}$ writes

$$C_{JL}(\boldsymbol{V}|\boldsymbol{W},\boldsymbol{H}) \stackrel{\text{def}}{=} \log p(\boldsymbol{V}|\boldsymbol{W},\boldsymbol{H}) + \log p(\boldsymbol{H})\,. \quad (5)$$

The log-likelihood term $\log p(\boldsymbol{V}|\boldsymbol{W},\boldsymbol{H})$ is up to irrelevant constants equal to $-D_{KL}(\boldsymbol{V}|\boldsymbol{W}\boldsymbol{H})$ so that MJLE is equivalent to penalised KL-NMF [8, 5]. A majorisation-minimisation (MM) algorithm for minimisation of $C_{JL}(\boldsymbol{V}|\boldsymbol{W},\boldsymbol{H})$ is presented in Section 4.1.

3.2. Maximum marginal likelihood estimation (MMLE)

The marginal log-likelihood of $\boldsymbol{W}$ writes

$$C_{ML}(\boldsymbol{V}|\boldsymbol{W}) \stackrel{\text{def}}{=} \log \int p(\boldsymbol{V}|\boldsymbol{W},\boldsymbol{H}) p(\boldsymbol{H})\, d\boldsymbol{H}\,.$$

This integral is intractable, i.e., it is not possible to obtain the marginal model analytically. Note that in Bayesian estimation the term *marginal likelihood* is sometimes used as a synonym for the *model evidence*, which would be the likelihood of data given the model, i.e., where all random parameters (including $\boldsymbol{W}$) have been marginalised. This full Bayesian approach has been considered for example in [8] and [3] for the Poisson and Gaussian additive noise models, respectively. In [7], $\boldsymbol{W}$ again has a prior distribution and is estimated with a maximum a posteriori approach. Let us emphasize again that in our setting $\boldsymbol{W}$ is taken as a deterministic parameter and that the term "marginal likelihood" here refers to the likelihood of $\boldsymbol{W}$ where $\boldsymbol{H}$ has been integrated out.

4. ALGORITHMS

4.1. Majorisation-minimisation (MM) for MJLE

We describe an iterative algorithm which sequentially updates $\boldsymbol{W}$ given $\boldsymbol{H}$ and vice versa. The update of $\boldsymbol{W}$ is the standard multiplicative rule derived from MM [1]. The penalty term in $\boldsymbol{H}$ can easily be handled in the same framework. Under our assumptions, criterion C_{JL} is separable in the columns of $\boldsymbol{H}$ so that its maximisation is essentially reduced to the minimisation of

$$C(\boldsymbol{h}) = D_{KL}(\boldsymbol{v}|\boldsymbol{W}\boldsymbol{h}) + L(\boldsymbol{h})\,, \quad (6)$$

where $L(\boldsymbol{h}) \stackrel{\text{def}}{=} \sum_k \beta_k h_k - (\alpha_k - 1)\log h_k$ corresponds to the Gamma prior contribution. By convexity of the KL divergence $d_{KL}(x|y)$ w.r.t y and using Jensen's inequality, the functional

$$G(\boldsymbol{h}|\tilde{\boldsymbol{h}}) = \sum_k \tilde{\lambda}_{kf}\, d\left(v_f|\frac{w_{fk}h_k}{\tilde{\lambda}_{kf}}\right) + L(\boldsymbol{h})\,, \quad (7)$$

where $\tilde{\lambda}_{kf} = w_{fk}\tilde{h}_k/[\boldsymbol{W}\tilde{\boldsymbol{h}}]_f$, is an *auxiliary function* for $C(\boldsymbol{h})$ (i.e., $G(\boldsymbol{h},\boldsymbol{h}) = C(\boldsymbol{h})$ and $G(\boldsymbol{h},\tilde{\boldsymbol{h}}) \geq C(\boldsymbol{h})$). Hence, iterative minimisation of $G(\boldsymbol{h}|\tilde{\boldsymbol{h}})$ leads to the following algorithm, which ensures nonnegativity of the expansion coefficients provided positive initialisation and $\alpha_k \geq 1$:

$$h_{kn} \leftarrow \frac{h_{kn}\sum_f w_{fk}v_{fn}/[\boldsymbol{W}\boldsymbol{H}]_{fn} + (\alpha_k - 1)}{\sum_f w_{fk} + \beta_k}\,. \quad (8)$$

This algorithm is also given in [6], though derived in a different way.

4.2. Variational EM for MMLE

We propose a variational EM algorithm [9] for the maximisation of $C_{ML}(\boldsymbol{V}|\boldsymbol{W})$. The data $\boldsymbol{V}$ is "augmented" with the latent variables $\boldsymbol{H}$ and $\boldsymbol{C}$ and the algorithm is based on the iterative estimation and maximisation of the following functional:

$$Q(\boldsymbol{W}|\tilde{\boldsymbol{W}}) \stackrel{\text{def}}{=} \int \log p(\boldsymbol{V},\boldsymbol{C},\boldsymbol{H}|\boldsymbol{W}) p(\boldsymbol{C},\boldsymbol{H}|\boldsymbol{V},\tilde{\boldsymbol{W}})\, d\boldsymbol{C}\, d\boldsymbol{H}\,.$$

Unfortunately, the computation of the functional (E-step) is intractable, in particular because the analytical form of the latent data posterior is itself intractable. As such, we resort to a variational approximation of $p(\boldsymbol{C},\boldsymbol{H}|\boldsymbol{V},\tilde{\boldsymbol{W}})$ that renders all derivations tractable, though at the cost of approximate inference. The two steps of the variational EM are described next.

E-step: A variational approximation $q(\boldsymbol{C},\boldsymbol{H})$ of the exact posterior $p(\boldsymbol{C},\boldsymbol{H}|\boldsymbol{V},\tilde{\boldsymbol{W}})$ is computed at every iteration of the EM algorithm and plugged in $Q(\boldsymbol{W}|\tilde{\boldsymbol{W}})$. Note that the computation of $q(\boldsymbol{C},\boldsymbol{H})$ requires a few subiterations itself. As fundamental to variational approximations, the computation of $q(\boldsymbol{C},\boldsymbol{H})$ relies on the minimisation of the KL divergence (in *distribution* this time) between $q(\boldsymbol{C},\boldsymbol{H})$ and $p(\boldsymbol{C},\boldsymbol{H}|\boldsymbol{V},\boldsymbol{W})$, given a parametric form of $q(\boldsymbol{C},\boldsymbol{H})$. The variational objective function may be decomposed as

$$\begin{aligned}\mathrm{KL}[q(\boldsymbol{C},\boldsymbol{H})|p(\boldsymbol{C},\boldsymbol{H}|\boldsymbol{V},\boldsymbol{W})] = \\ \log p(\boldsymbol{V}|\boldsymbol{W}) + \mathrm{KL}[q(\boldsymbol{C},\boldsymbol{H})|p(\boldsymbol{V},\boldsymbol{C},\boldsymbol{H}|\boldsymbol{W})]\,. \quad (9)\end{aligned}$$

Because the marginal likelihood $\log p(\boldsymbol{V}|\boldsymbol{W})$ is independent of $q(\boldsymbol{C},\boldsymbol{H})$, the minimisation of the variational objective may be replaced by the (simpler) maximisation of $\mathrm{L}[q(\boldsymbol{C},\boldsymbol{H})] = -\mathrm{KL}[q(\boldsymbol{C},\boldsymbol{H})|p(\boldsymbol{V},\boldsymbol{C},\boldsymbol{H}|\boldsymbol{W})]$, which forms a lower bound of the marginal likelihood $\log p(\boldsymbol{V}|\boldsymbol{W})$ (thanks to nonnegativity of the KL divergence). It can be shown that, given the expression of $p(\boldsymbol{V},\boldsymbol{C},\boldsymbol{H}|\boldsymbol{W})$, the following form of variational distribution appears as a natural choice (in particular for tractability) :

$$q(\boldsymbol{C},\boldsymbol{H}) = \prod_{f=1}^{F}\prod_{n=1}^{N} q(\boldsymbol{c}_{fn}) \prod_{k=1}^{K}\prod_{n=1}^{N} q(h_{kn})\,, \quad (10)$$

where $\boldsymbol{c}_{fn}$ denotes the vector $[c_{1,fn}, c_{2,fn}, \ldots, c_{K,fn}]^T$, $q(\boldsymbol{c}_{fn})$ is multinomial with probabilities $p_{k,fn}$ and $q(h_{kn})$ is a Gamma distribution with shape and scale parameters a_{kn} and b_{kn}. The factors $q(h_{kn})$ and $q(\boldsymbol{c}_{fn})$ can be shown to satisfy the following fixed point equations [9]:

$$\log q(h_{kn}) \stackrel{\text{c}}{=} \langle \log p(\boldsymbol{V},\boldsymbol{C},\boldsymbol{H}|\boldsymbol{W})\rangle_{q(\boldsymbol{H}_{-kn})q(\boldsymbol{C})} \quad (11)$$

$$\log q(\boldsymbol{c}_{fn}) \stackrel{\text{c}}{=} \langle \log p(\boldsymbol{V},\boldsymbol{C},\boldsymbol{H}|\boldsymbol{W})\rangle_{q(\boldsymbol{H})q(\boldsymbol{C}_{-fn})}\,, \quad (12)$$

where $\langle . \rangle_\pi$ denotes expectation under probability distribution π and $\boldsymbol{A}_{-ij}$ refers to the set of coefficients of $\boldsymbol{A}$ excluding a_{ij}. The fixed point equations translate into the following parameter updates:

$$a_{kn} \leftarrow \alpha_k + \sum\nolimits_f \langle c_{k,fn} \rangle$$

$$b_{kn} \leftarrow (1/\beta_k + \sum\nolimits_f w_{fk})^{-1}$$

$$p_{k,fn} \leftarrow \frac{w_{fk} \exp(\langle \log h_{kn} \rangle)}{\sum_l w_{fl} \exp(\langle \log h_{ln} \rangle)},$$

where $\langle . \rangle$ denotes expactation w.r.t the variational distribution.

M-step: Given the $\tilde{\boldsymbol{W}}$-dependent variational distribution $q(\boldsymbol{C}, \boldsymbol{H})$ obtained in the E-step, it can be shown that the evaluation and maximisation of $Q(\boldsymbol{W}|\tilde{\boldsymbol{W}})$ lead to the following multiplicative update

$$w_{fk} \leftarrow w_{fk} \frac{\sum_n \exp(\langle \log h_{kn} \rangle) v_{fn} / [\boldsymbol{W} \exp(\langle \log \boldsymbol{H} \rangle)]_{fn}}{\sum_n \langle h_{kn} \rangle}$$

5. EXPERIMENTS

Next we study the performances of MJLE and MMLE on real and synthetical data. The prior hyperparameters are fixed to $\alpha_k = 1$ (exponential distribution) and $\beta_k = 1$, i.e., $h_{kn} \sim \exp(-h_{kn})$. We used 5000 algorithm iterations and nonnegative random initialisations in all cases.

5.1. A piano excerpt

We consider the piano data used in [4]. It is a toy audio sequence recorded in real conditions, consisting of four notes played all together in the first measure and in all possible pairs in the subsequent measures. A magnitude spectrogram of the data was computed, leading to $F = 513$ frequency bins and $N = 676$ time frames. We ran the MM algorithm (for MJLE) and variational EM (for MMLE) for $K = 1 \ldots 10$ and the joint and marginal log-likelihood end values (after the 5000 iterations) are displayed in Fig. 1. The marginal log-likelihood is here approximated by its lower bound, as described in Section 4.2.[1]

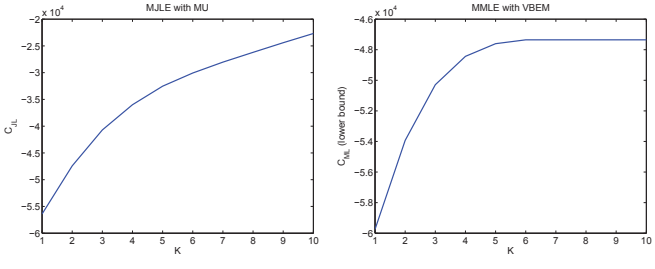


Fig. 1. Joint likelihood C_{JL} (left) and marginal likelihood C_{ML} (right) versus number of components K.

The likelihood values increase with the number of components, as expected from nested models. However, and very interestingly, the marginal likelihood stagnates after $K = 6$. Manual inspection reveals that passed this value of K, the extra columns of $\boldsymbol{W}$ are pruned to zero, leaving the criterion unchanged. Hence, MMLE appears to embed automatic order selection, similar to automatic relevance determination [11, 12]. This is illustrated in Fig. 2 which displays the dictionary columns estimated by MJLE and MMLE with $K = 10$. Reconstruction of the time-domain components associated with the MMLE decomposition reveals that the 6 components correspond to individual notes, note attacks and residual noise, which is the expected result, see [4] for more details about the experimental setup. These components also appear in the reconstruction obtained from MJLE but less accurately and duplicates appear when $K > 6$. With $K = 10$, 5000 iterations of the MM and variational EM algorithms take 92 and 142 seconds of CPU time, respectively, on an average computer.

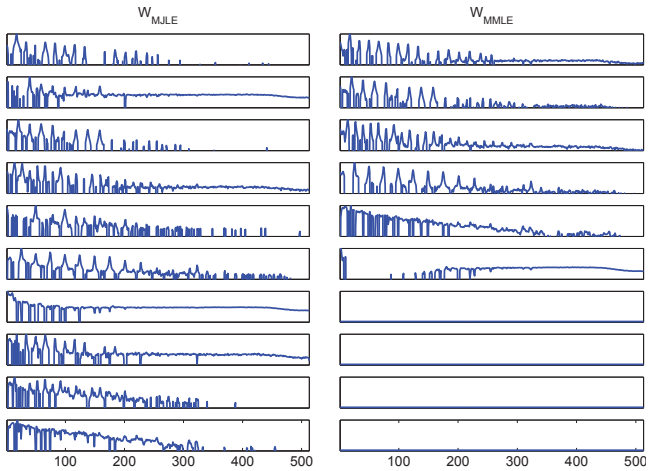


Fig. 2. Dictionaries learnt from the piano excerpt with $K = 10$. The values (y-axis) are in log scale and the x-axis corresponds to the frequency bins.

5.2. Swimmer dataset

To further investigate the automatic model order selection feature of MMLE, we consider the synthetical `Swimmer` dataset [13], for which a ground truth can be defined. The dataset is composed of 256 images of size 32×32, representing a swimmer built of an invariant torso and 4 limbs. Each of the 4 limbs can be in one of 4 positions and the dataset is formed of all combinations (see some samples in Fig. 3). Hence, the ground truth dictionary corresponds to the collection of individual limb positions. As explained in [13] the torso is an unidentifiable component that can be paired with any of the limbs, or even split among the limbs.

The dictionaries learnt from MJLE and MMLE with $K = 20$ components are shown in Fig. 4. As can be seen from Fig. 4 (a), MJLE produces spurious or duplicated components. In contrast, the ground truth is perfectly recovered with MMLE.

6. CONCLUSIONS

In this paper we have challenged the standard NMF approach to nonnegative dictionary learning, based on maximum *joint* likelihood estimation, with a better-posed approach consisting in maximum *marginal* likelihood estimation. The proposed algorithm based on variational inference has comparable computational complexity to standard NMF. Experiments on real and synthetical data have

[1] Let us mention that we checked the validity of the approximation by separately running Chib's method [10] for stochastic approximation of $p(\boldsymbol{V}|\boldsymbol{W})$ and the results, not shown here, confirmed the accuracy of the bound.

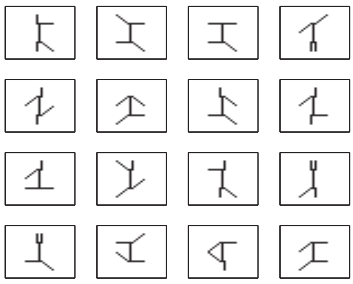
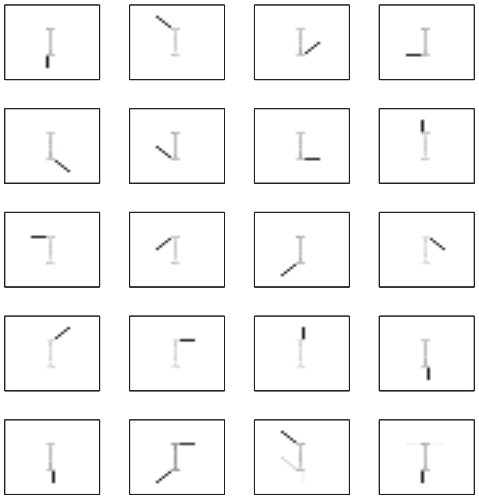
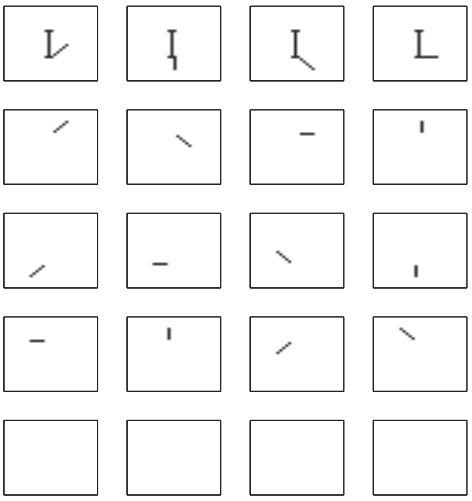


Fig. 3. Sample images from the `Swimmer` dataset.



(a) W_{MJLE}



(b) W_{MMLE}

Fig. 4. Dictionaries learnt from the swimmer dataset with $K = 20$.

brought up a very attractive feature of MMLE, the self-ability of discarding "irrelevant" columns from the dictionary, i.e., performing automatic model order selection. This property results in more accurate and interpretable components. In contrast with other model selection approaches in fully Bayesian settings, e.g., [8, 3], based on the evaluation of the model evidence for every candidate value of K, our approach only requires to set K to a sufficiently large value and run the variational EM algorithm once.

As for perspective we intend to confront MMLE with other statistical models, such as the Gaussian composite variance model of [4], which underlies Itakura-Saito NMF and was shown to provide a more natural generative model of audio spectrograms than KL-NMF.

7. REFERENCES

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

[2] M. S. Lewicki and T. J. Sejnowski, "Learning overcomplete representations," *Neural Computation*, vol. 12, pp. 337–365, 2000.

[3] M. N. Schmidt, O. Winther, and L. K. Hansen, "Bayesian non-negative matrix factorization," in *Proc. 8th International Conference on Independent Component Analysis and Signal Separation (ICA'09)*, Paraty, Brazil, Mar. 2009.

[4] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence. With application to music analysis," *Neural Computation*, vol. 21, no. 3, pp. 793–830, Mar. 2009.

[5] C. Févotte and A. T. Cemgil, "Nonnegative matrix factorisations as probabilistic inference in composite models," in *Proc. 17th European Signal Processing Conference (EUSIPCO'09)*, Glasgow, Scotland, Aug. 2009, pp. 1913–1917.

[6] J. F. Canny, "GaP: A factor model for discrete data," in *Proc. of the 27th ACM international Conference on Research and Development of Information Retrieval (SIGIR)*, 2004, pp. 122–129.

[7] W. L. Buntine and A. Jakulin, "Discrete component analysis," in *Subspace, Latent Structure and Feature Selection Techniques*, 2005, pp. 1–33.

[8] A. T. Cemgil, "Bayesian inference for nonnegative matrix factorisation models," *Computational Intelligence and Neuroscience*, vol. 2009, no. Article ID 785152, pp. 17 pages, 2009.

[9] M. J. Beal and Z. Ghahramani, "The variational Bayesian EM algorithm for incomplete data: with application to scoring graphical model structures," in *Bayesian Statistics 7*, Oxford University Press, 2003.

[10] S. Chib, "Marginal likelihood from the Gibbs output," *Journal of the American Statistical Association*, vol. 90, no. 432, pp. 1313–1321, 1995.

[11] D. J. C. MacKay, "Probable networks and plausible predictions – a review of practical Bayesian models for supervised neural networks," *Network: Computation in Neural Systems*, vol. 6, no. 3, pp. 469–505, 1995.

[12] C. M. Bishop, "Bayesian PCA," in *Advances in Neural Information Processing Systems (NIPS)*, pp. 382–388, 1999.

[13] D. Donoho and V. Stodden, "When does non-negative matrix factorization give a correct decomposition into parts?," in *Advances in Neural Information Processing Systems (NIPS)*, 2004.