Maximum Marginal Likelihood Estimation for Nonnegative Dictionary Learning in the Gamma-Poisson Model

Onur Dikmen, *Member, IEEE*, and Cédric Févotte, *Member, IEEE*

***Abstract*—In this paper we describe an alternative to standard nonnegative matrix factorization (NMF) for nonnegative dictionary learning, i.e., the task of learning a dictionary with nonnegative values from nonnegative data, under the assumption of nonnegative expansion coefficients. A popular cost function used for NMF is the Kullback-Leibler divergence, which underlies a Poisson observation model. NMF can thus be considered as maximization of the *joint likelihood* of the dictionary and the expansion coefficients. This approach lacks optimality because the number of parameters (which include the expansion coefficients) grows with the number of observations. In this paper we describe variational Bayes and Monte-Carlo EM algorithms for optimization of the *marginal likelihood*, i.e., the likelihood of the dictionary where the expansion coefficients have been integrated out (given a Gamma 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 datasets. In particular we present face reconstruction results on CBCL dataset and text retrieval results over the musiXmatch dataset, a collection of word counts in song lyrics. The MMLE approach is shown to prevent overfitting by automatically pruning out irrelevant dictionary columns, i.e., embedding automatic model order selection.**

***Index Terms*—Automatic relevance determination, model order selection, Monte Carlo EM, nonnegative matrix factorization, sparse coding, variational EM.**

I. Introduction

NONNEGATIVE matrix factorization (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, $\boldsymbol{W}$ (dictionary) and $\boldsymbol{H}$ (expansion coefficients), of sizes $F \times K$ and $K \times N$, respectively. These two matrices can be estimated via minimizing a measure of fit between $\boldsymbol{V}$ and $\boldsymbol{WH}$. One such popular measure is the (generalized) Kullback-Leibler (KL) divergence

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

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 $\boldsymbol{A} = \boldsymbol{B}$. Minimization 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] and based on the Richardson-Lucy algorithm [2], [3]. This approach also coincides with maximum joint 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 optimize the marginal likelihood of $\boldsymbol{W}$ given by

$$p(\boldsymbol{V}|\boldsymbol{W}) = \int_{\boldsymbol{H}} p(\boldsymbol{V}|\boldsymbol{W},\boldsymbol{H})p(\boldsymbol{H})\,d\boldsymbol{H}, \tag{1}$$

where $p(\boldsymbol{H})$ is an assumed prior distribution for the expansion coefficients. Our approach is similar in spirit to independent component analysis (ICA), where the likelihood of the "mixing matrix" is obtained through marginalization of the latent independent components, see, e.g., [4].

In order to compute the marginal likelihood in (1), we define a prior distribution for $\boldsymbol{H}$. We choose a Gamma distribution, which is conjugate to the Poisson likelihood, mostly for algorithmic convenience as will be apparent further in the paper. The Gamma distribution takes the sparsity-inducing exponential distribution as a special case. We leave the dictionary $\boldsymbol{W}$ to be a deterministic variable. As such, our model coincides with the Gamma-Poisson (GaP) model of Canny [5], which constitutes the base for many elaborate models for text clustering and image interpolation (e.g., [6]–[9]). In Discrete Component Analysis (DCA) of Buntine & Jakulin [8] a Dirichlet prior is assumed for $\boldsymbol{W}$, whereas in Cemgil's work [9] the prior is a Gamma distribution. In these works, maximum a posteriori (MAP) estimate or full posterior distribution of $\boldsymbol{W}$ (thus taken as a random varianble) have been sought after. In this work we do not wish to make any prior assumption on $\boldsymbol{W}$ and rather look for the maximum likelihood estimator of $\boldsymbol{W}$ in the basic GaP model. Our work bears methodological resemblance to [8] and [9], but we here pursue a different objective.

The main motivation in this paper is to learn the dictionary parameters from the marginal model in order to prevent the overfitting in standard KL-NMF, which stems from the growing number of parameters with the number of observations. At this

Manuscript received July 22, 2011; revised December 04, 2011, March 21, 2012, and June 10, 2012; accepted June 12, 2012. Date of publication July 05, 2012; date of current version September 11, 2012. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Samson Lasaulce. This work was supported by Project ANR-09-JCJC-0073-01 TANGERINE (Theory and applications of nonnegative matrix factorization).

O. Dikmen was with the CNRS LTCI; Télécom ParisTech, Paris 75014, France. He is now with the Department of Information and Computer Science, Aalto University, Espoo 02150, Finland (e-mail: onur.dikmen@aalto.fi).

C. Févotte is with the CNRS LTCI; Télécom ParisTech, Paris 75014, France (e-mail: fevotte@telecom-paristech.fr).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSP.2012.2207117

stage a parallel can be made between our approach and Latent Dirichlet Allocation (LDA) [10], which was proposed as a remedy to the overfitting problem of Latent Semantic Indexing (pLSI) [11]. LDA and pLSI are based on the same model (a discrete observation model in which word f in document n is generated with probability $\sum_k w_{fk}h_{kn}$ under the constraint that $\sum_f w_{fk} = 1$ and $\boldsymbol{h}_n = [h_{1n}, \ldots, h_{Kn}]^T$ has a Dirichlet prior), but LDA seeks marginal likelihood estimation whereas pLSI performs joint likelihood estimation (i.e., the dictionary and the expansion matrix are updated together).

This paper describes two approximate learning algorithms for the maximization of the marginal likelihood. The integration in (1) is not tractable, so the expansion variables $\boldsymbol{H}$ cannot be analytically integrated out of the model. In addition, an exact expectation-maximization (EM) algorithm cannot be pursued, because the exact form of the posterior distribution of $\boldsymbol{H}$ is not available either. Our EM algorithms, variational Bayes (VBEM) and Monte Carlo EM (MCEM), are based on inferring the posterior distribution of $\boldsymbol{H}$ in the E-step and maximizing $\boldsymbol{W}$ in the M-step. VBEM maximizes a lower bound of the marginal likelihood $p(\boldsymbol{V}|\boldsymbol{W})$, whereas the functional that MCEM maximizes at each step is asymptotically exact. In the experiments section, we verify that these two methods perform similarly and the lower bound that VBEM optimizes is a tight approximation of criterion (1), as numerically confirmed using the more computationally intensive but asymptotically optimal Chib's method [12]. Computation of the criterion itself is for example needed for classification tasks based on likelihood ratios. In this paper we also describe novel algorithms for maximum joint likelihood estimation (MJLE), equivalent to standard KL-NMF with a penalty term on $\boldsymbol{H}$ that stems from the assumed prior, that extend previous work by Canny [5].

We compare MJLE and MMLE on synthetical and real data and indeed show that MMLE avoids the problem of overfitting by assigning "unnecessary" columns of the dictionary to zero, i.e., performing automatic model order selection. With MMLE, the estimated dictionaries are more column-sparse (in the sense that many columns become negligible when K is overestimated) than those of MJLE and this makes the dictionaries more interpretable. We will in particular demonstrate this feature on the musiXmatch dataset, a large collection of word counts from song lyrics, which has recently been made available [13].

The rest of this paper is organized as follows. Section II describes the generative model and presents the two dictionary estimators considered in this paper. Sections III and IV describe algorithms proposed for these estimators. Section V reports results on real and synthetical data and in particular illustrates a very desirable feature of the marginal likelihood approach: automatic order selection. Section VI concludes the paper. The VBEM algorithm of Section IV was introduced in [14] and was mainly applied to audio data. Here, we also introduce the asymptotically exact MCEM algorithm and verify the results obtained by VBEM. We describe Chib's method [12] for this model to discuss the tightness of the variational lower bound. We also present novel algorithms for MJLE for a large range of shape parameters. The two estimators, MJLE and MMLE, are compared in face reconstruction and text retrieval applications.

II. Model and Estimators

A. GaP Model

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

$$v_{fn} \sim \mathcal{P}\left(v_{fn}|\sum_k w_{fk}h_{kn}\right) \tag{2}$$

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

$$v_{fn} = \sum_{k=1}^{K} c_{k,fn}, \quad c_{k,fn} \sim \mathcal{P}(c_{k,fn}|w_{fk}h_{kn}) \tag{3}$$

where the components $c_{k,fn}$ act as *latent variables*. In the remainder of the text, $\boldsymbol{C}$ will denote the $K \times F \times N$ matrix consisting of $\{c_{k,fn}\}$ and $\boldsymbol{c}_{fn}$ will represent the $K \times 1$ vector $[c_{1,fn}, c_{2,fn}, \ldots, c_{K,fn}]^T$. The posterior distribution of $\boldsymbol{c}_{fn}$ is analytically available and is a multinomial distribution. Similarly, posterior distribution of each $c_{k,fn}$ is binomial. As can be seen from (3), the introduction of the components allows us to break the coupling $\sum_k w_{fk}h_{kn}$ in the probability density function of the observation model. This fact will be used in the data augmentation algorithms described in Sections IV-A and IV-B.

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 $\mathcal{G}(x|\alpha, \beta) = [\beta^\alpha \Gamma(\alpha)]^{-1}\, x^{\alpha-1}\, \exp\left(\frac{-x}{\beta}\right)$, $x \geq 0$, $\alpha > 0$, $\beta > 0$. 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 Gamma-Poisson (GaP) model of [5], [8] which has been used in text analysis. In the rest of the paper, the scale parameters β_k will be fixed, so as to remedy the scale ambivalence between k^{th} column of $\boldsymbol{W}$ (denoted $\boldsymbol{w}_k$ in the following) and k^{th} row of $\boldsymbol{H}$, as more thoroughly discussed in Section II-D. We will denote $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_K]^T$. The shape parameters α_k are also fixed. The dictionary $\boldsymbol{W}$ is taken as a free deterministic parameter.

B. Maximum Joint Likelihood Estimation (MJLE)

The MJLE estimator of $\boldsymbol{W}$ is obtained by maximization (under nonnegativity of all the parameters) of the joint penalized log-likelihood likelihood of $\boldsymbol{W}$, $\boldsymbol{H}$ and $\boldsymbol{\beta}$, defined by

$$\begin{aligned} C_{JL}(\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{\beta}) &\stackrel{\text{def}}{=} \log p(\boldsymbol{V}|\boldsymbol{W}, \boldsymbol{H}) + \log p(\boldsymbol{H}|\boldsymbol{\beta}) \\ &= -D_{KL}(\boldsymbol{V}|\boldsymbol{W}\boldsymbol{H}) - L_{\boldsymbol{\beta}}(\boldsymbol{H}) + cst \end{aligned}$$

where cst denotes terms constant w.r.t $\boldsymbol{W}$, $\boldsymbol{H}$, $\boldsymbol{\beta}$ and where

$$L_{\boldsymbol{\beta}}(\boldsymbol{H}) \stackrel{\text{def}}{=} \sum_{kn} \frac{h_{kn}}{\beta_k} - (\alpha_k - 1)\log h_{kn} + \alpha_k \log \beta_k.$$

As it appears, MJLE in the GaP model is equivalent to penalized KL-NMF [9], [15], with penalty term $L_{\boldsymbol{\beta}}(\boldsymbol{H})$. In Section III, we will present minorization-maximization (MM) algorithms

for MJLE for different α_k values. For a comprehensive study about MM algorithms for NMF, see [16].

C. Maximum Marginal Likelihood Estimation (MMLE)

The MMLE estimator of $\boldsymbol{W}$ is obtained by maximization of the marginal log-likelihood of $\boldsymbol{W}$, defined by

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

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 $\boldsymbol{W}$ as well) have been marginalized. This full Bayesian approach has been considered for example in [9] and [17] for the Poisson and Gaussian additive noise models, respectively. In [8], $\boldsymbol{W}$ 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. In Section IV, we will describe approximate EM algorithms for evaluating and maximizing C_{ML}.

D. Scales

The MJLE and MMLE estimators of $\boldsymbol{W}$ have different behaviors w.r.t scale $\boldsymbol{\beta}$. The MMLE objective function $C_{ML}(\boldsymbol{W},\boldsymbol{\beta})$ is scale-invariant, in the following sense. Let $\boldsymbol{\Lambda}$ be a nonnegative diagonal matrix with coefficients λ_k. Then we have the following property:

$$C_{ML}(\boldsymbol{W}\boldsymbol{\Lambda}^{-1},\boldsymbol{\Lambda}\boldsymbol{\beta}) = C_{ML}(\boldsymbol{W},\boldsymbol{\beta}).$$

This is shown using the change of variable $h_{kn} = \lambda_k \tilde{h}_{kn}$ and the property that if $X \sim \mathcal{G}(\alpha,\beta)$ then $\lambda X \sim \mathcal{G}(\alpha,\lambda\beta)$. More precisely, we have

$$\begin{aligned} C_{ML}(\boldsymbol{W}\boldsymbol{\Lambda}^{-1},\boldsymbol{\Lambda}\boldsymbol{\beta}) &= \int_{\boldsymbol{H}} p(\boldsymbol{V}|\boldsymbol{W}\boldsymbol{\Lambda}^{-1},\boldsymbol{H}) \prod_{kn} \mathcal{G}(h_{kn}|\alpha_k,\lambda_k\beta_k) d\boldsymbol{H} \\ &= \int_{\tilde{\boldsymbol{H}}} p(\boldsymbol{V}|\boldsymbol{W}\boldsymbol{\Lambda}^{-1},\boldsymbol{\Lambda}\tilde{\boldsymbol{H}}) \prod_{kn} \lambda_k \mathcal{G}(\lambda_k\tilde{h}_{kn}|\alpha_k,\lambda_k\beta_k) d\tilde{\boldsymbol{H}} \\ &= \int_{\tilde{\boldsymbol{H}}} p(\boldsymbol{V}|\boldsymbol{W},\tilde{\boldsymbol{H}}) \prod_{kn} \mathcal{G}(\tilde{h}_{kn}|\alpha_k,\beta_k) d\tilde{\boldsymbol{H}} \\ &= C_{ML}(\boldsymbol{W},\boldsymbol{\beta}). \end{aligned}$$

As such, we may fix β_k to any arbitrary value, because of this simple linear mapping that exists between dictionary solutions obtained for different scale parameters.

The MJLE objective function $C_{JL}(\boldsymbol{W},\boldsymbol{H},\boldsymbol{\beta})$ behaves differently w.r.t scale, and as it appears, in a potentially problematic way. Indeed, the following expression holds:

$$C_{JL}(\boldsymbol{W}\boldsymbol{\Lambda}^{-1},\boldsymbol{\Lambda}\boldsymbol{H},\boldsymbol{\Lambda}\boldsymbol{\beta}) = C_{JL}(\boldsymbol{W},\boldsymbol{H},\boldsymbol{\beta}) - N\sum_k \log\lambda_k.$$

This implies that maximization of $C_{JL}(\boldsymbol{W},\boldsymbol{H},\boldsymbol{\beta})$ under mere nonnegativity constraints can only lead to a degenerate solution $(\boldsymbol{W}^\star,\boldsymbol{H}^\star,\boldsymbol{\beta}^\star)$ such that $\boldsymbol{W}\to\infty$, $\boldsymbol{H}\to 0$ and $\boldsymbol{\beta}\to 0$. This can be shown by contradiction: assume that $\boldsymbol{W}^\star$ is finite, then for any nonnegative diagonal matrix $\boldsymbol{\Lambda}$ such that $\lambda_k < 1$ we obtain that $C_{JL}(\boldsymbol{W}^\star\boldsymbol{\Lambda}^{-1},\boldsymbol{\Lambda}\boldsymbol{H}^\star,\boldsymbol{\Lambda}\boldsymbol{\beta}^\star) > C_{JL}(\boldsymbol{W}^\star,\boldsymbol{H}^\star,\boldsymbol{\beta}^\star)$, which contradicts the fact that $(\boldsymbol{W}^\star,\boldsymbol{H}^\star,\boldsymbol{\beta}^\star)$ is the optimum. As such, joint estimation of $\boldsymbol{W}$, $\boldsymbol{H}$, $\boldsymbol{\beta}$ requires to control the norm of $\boldsymbol{W}$ to prevent from degeneracy. In this paper will not try to estimate $\boldsymbol{\beta}$ (either by maximization of C_{JL} or by any other mean such as cross-validation) but we will rather concentrate on the properties of the two estimators (MJLE and MMLE) based on the same generative model, i.e., with both α_k and β_k fixed to arbitrary values. As such, we will remove from now on the dependency of C_{JL} and C_{ML} on $\boldsymbol{\beta}$. Fixing $\boldsymbol{\beta}$ removes the degeneracy problem only when $\alpha_k > 1$. Indeed, it is easy to check that when $\alpha \leq 1$ the penalty term $L_{\boldsymbol{\beta}}(\boldsymbol{\Lambda}\boldsymbol{H})$ can still be made arbitrarily small as λ_k goes to zero, thus encouraging the solution $\boldsymbol{W}\to\infty$. As such, controlling the norm of $\boldsymbol{W}$ is still needed in that case. A set of algorithms for MJLE, depending on the value of α_k and the required constraint on $\boldsymbol{W}$ are presented in the next section. A conclusion of this section is that, besides the question of its asymptotical optimality, MJLE is not an as well-posed problem as MMLE.

III. Algorithms For MJLE

A. Algorithms For $\alpha_k > 1$

1) Canny's Algorithm: Given the discussion of Section II-D, when $\boldsymbol{\beta}$ is fixed and $\alpha_k > 1$, the norm of $\boldsymbol{W}$ does not need to be necessarily controlled. In that setting, an iterative algorithm that guarantees to increase $C_{JL}(\boldsymbol{W},\boldsymbol{H})$ at every iteration is described in [5]. Though not clearly stated as such, it is essentially an EM algorithm based on the set of components $c_{k,fn}$ introduced in Section II-A, that updates $\boldsymbol{H}$ given $\boldsymbol{W}$ and then $\boldsymbol{W}$ given $\boldsymbol{H}$. It can also be seen as a MM algorithm where each update is based on the maximization of a surrogate auxiliary function, i.e., a lower bound of the original objective function which is tight at the current parameter estimate [14]. This strategy ensures the increase of the joint likelihood after every update of $\boldsymbol{W}$ and $\boldsymbol{H}$. Using its concavity w.r.t $\boldsymbol{H}$ given $\boldsymbol{W}$ or $\boldsymbol{H}$ given $\boldsymbol{W}$, the fit to data term $-D_{KL}(\boldsymbol{V}|\boldsymbol{W}\boldsymbol{H})$ can easily be minorized using a Jensen's type inequality, such that, $\forall\tilde{\boldsymbol{H}}$

$$-D_{KL}(\boldsymbol{V}|\boldsymbol{W}\boldsymbol{H}) \geq -\sum_{fkn} \frac{w_{fk}\tilde{h}_{kn}}{[\boldsymbol{W}\tilde{\boldsymbol{H}}]_{fn}} d_{KL}\left(v_{fn}|[\boldsymbol{W}\tilde{\boldsymbol{H}}]_{fn}\frac{h_{kn}}{\tilde{h}_{kn}}\right) \quad (4)$$

with equality when $\boldsymbol{H} = \tilde{\boldsymbol{H}}$, or similarily, $\forall\tilde{\boldsymbol{W}}$

$$-D_{KL}(\boldsymbol{V}|\boldsymbol{W}\boldsymbol{H}) \geq -\sum_{fkn} \frac{\tilde{w}_{fk}h_{kn}}{[\tilde{\boldsymbol{W}}\boldsymbol{H}]_{fn}} d_{KL}\left(v_{fn}|[\tilde{\boldsymbol{W}}\boldsymbol{H}]_{fn}\frac{w_{fk}}{\tilde{w}_{fk}}\right) \quad (5)$$

with equality when $\boldsymbol{W} = \tilde{\boldsymbol{W}}$. Iterative maximization of the lower bound (5) w.r.t $\boldsymbol{W}$ leads to the well known multiplicative algorithm described by [1]–[3]

$$w_{fk} = \tilde{w}_{fk}\frac{\sum_n \frac{h_{kn}v_{fn}}{[\tilde{\boldsymbol{W}}\boldsymbol{H}]_{fn}}}{\sum_n h_{kn}}.$$

The penalty term $-L_{\boldsymbol{\beta}}(\boldsymbol{H})$ needs solely to be added to the right side of (4) to obtain a suitable auxiliary function for $\boldsymbol{H}$, which, when maximized w.r.t $\boldsymbol{H}$, leads to

$$h_{kn} = \frac{\tilde{h}_{kn} \sum_f \frac{w_{fk} v_{fn}}{[\boldsymbol{W}\tilde{\boldsymbol{H}}]_{fn}} + (\alpha_k - 1)}{\frac{1}{\beta_k} + \sum_f w_{fk}}.$$

2) Algorithm for Norm-Constrained $\boldsymbol{W}$: It is also possible to derive an algorithm for maximizing $C_{JL}(\boldsymbol{W}, \boldsymbol{H})$ under the constraint that $\|\boldsymbol{w}_k\| = 1$, where we will take $\|\cdot\|$ as the ℓ_1 norm. As discussed in Section II-D, though not necessary when β_k is fixed (and when $\alpha_k > 1$), this is needed when β_k has to be estimated as well. In that case we want to solve

$$\max_{\boldsymbol{W},\boldsymbol{H}} C_{JL}(\boldsymbol{W}, \boldsymbol{H}) = -D_{KL}(\boldsymbol{V}|\boldsymbol{W}\boldsymbol{H}) - L_{\boldsymbol{\beta}}(\boldsymbol{H}) + cst \quad s.t \quad \boldsymbol{W} \geq 0,\ \boldsymbol{H} \geq 0,\ \|\boldsymbol{w}\|_k = 1. \tag{6}$$

Following [18], [19], the latter problem is equivalent to the following surrogate optimization problem, that involves a scale-invariant objective function:

$$\max_{\boldsymbol{W},\boldsymbol{H}} \bar{C}_{JL}(\boldsymbol{W}, \boldsymbol{H}) \stackrel{\text{def}}{=} -D_{KL}(\boldsymbol{V}|\boldsymbol{W}\boldsymbol{H}) - L_{\boldsymbol{\beta}}(\boldsymbol{\Lambda}\boldsymbol{H}) + cst \quad s.t \quad \boldsymbol{W} \geq 0,\ \boldsymbol{H} \geq 0 \tag{7}$$

where $\boldsymbol{\Lambda} = \text{diag}\{\|w_1\|, \ldots, \|w_K\|\}$. The equivalence between (6) and(7) is explained as follows. Let $(\boldsymbol{W}, \boldsymbol{H})$ be a pair of nonnegative matrices and let $(\boldsymbol{W}^\bullet = \boldsymbol{W}\boldsymbol{\Lambda}^{-1}, \boldsymbol{H}^\bullet = \boldsymbol{\Lambda}\boldsymbol{H})$ be their rescaled equivalents. Then, we have $\bar{C}_{JL}(\boldsymbol{W}, \boldsymbol{H}) = C_{JL}(\boldsymbol{W}^\bullet, \boldsymbol{H}^\bullet)$, and $\boldsymbol{W}^\bullet$ satisfies the constraint $\|\boldsymbol{w}_k^\bullet\| = 1$ by construction. As such, one may solve (7), free of scale constraint, and then rescale its solution to obtain a solution to (6). Using same recipe as in previous section, with $L_{\boldsymbol{\beta}}(\boldsymbol{H})$ changed into $L_{\boldsymbol{\beta}}(\boldsymbol{\Lambda}\boldsymbol{H})$, we can obtain the following update for $\boldsymbol{H}$:

$$h_{kn} = \frac{\tilde{h}_{kn} \sum_f \frac{w_{fk} v_{fn}}{[\boldsymbol{W}\tilde{\boldsymbol{H}}]_{fn}} + (\alpha_k - 1)}{\left(1 + \frac{1}{\beta_k}\right) \sum_f w_{fk}}. \tag{8}$$

The penalty term $L_{\boldsymbol{\beta}}(\boldsymbol{\Lambda}\boldsymbol{H})$ also depends on $\boldsymbol{W}$ through λ_k. As such, the dictionary update is also changed. Unfortunately simply adding $L_{\boldsymbol{\beta}}(\boldsymbol{\Lambda}\boldsymbol{H})$ to the right side of (5) does not lead to an auxiliary function for $\boldsymbol{W}$ that can be maximized in close form. The penalty term $-L_{\boldsymbol{\beta}}(\boldsymbol{\Lambda}\boldsymbol{H})$, and more precisely its log part needs to be minorized as well. We may write

$$-L_{\boldsymbol{\beta}}(\boldsymbol{\Lambda}\boldsymbol{H}) = -\sum_{fkn} \frac{w_{fk} h_{kn}}{\beta_k} + N \sum_k (\alpha_k - 1) \log \|\boldsymbol{w}_k\| + cst$$

where cst represents terms constants w.r.t $\boldsymbol{W}$.[1] By concavity of $\log x$ and Jensen's inequality we may write

$$(\alpha_k - 1) \log \|\boldsymbol{w}_k\| \geq (\alpha_k - 1) \sum_f \frac{\tilde{w}_{fk}}{\|\tilde{\boldsymbol{w}}_k\|} \log \left(\|\tilde{\boldsymbol{w}}_k\| \frac{w_{fk}}{\tilde{w}_{fk}} \right)$$

[1]Except when otherwise specified, in the following we will abusively denote by cst any irrelevant term constant w.r.t the variable of the function in which it appears.

with equality when $\boldsymbol{w}_k = \tilde{\boldsymbol{w}}_k$. This minorization of the log part of the penalty term leads to close form maximization of the resulting auxiliary function, which writes

$$w_{fk} = \tilde{w}_{fk} \frac{\frac{(\alpha_k - 1)N}{\|\tilde{\boldsymbol{w}}_k\|} + \sum_n \frac{h_{kn} v_{fn}}{[\tilde{\boldsymbol{W}}\boldsymbol{H}]_{fn}}}{\left(1 + \frac{1}{\beta_k}\right) \sum_n h_{kn}}. \tag{9}$$

B. Algorithm for $\alpha_k \leq 1$

As discussed in Section II-D, when $\alpha_k \leq 1$, the norm of the dictionary needs to be controlled to prevent from degeneracy, β_k being treated either as a fixed or free parameter.

1) Algorithm for $\alpha_k = 1$: Canny's algorithm is not applicable anymore for $\alpha_k = 1$ but the derivations of Section III-A-II still hold. They simplify to the following update rules:

$$h_{kn} = \tilde{h}_{kn} \frac{\sum_f \frac{w_{fk} v_{fn}}{[\boldsymbol{W}\tilde{\boldsymbol{H}}]_{fn}}}{\left(1 + \frac{1}{\beta_k}\right) \sum_f w_{fk}} \tag{10}$$

$$w_{fk} = \tilde{w}_{fk} \frac{\sum_n \frac{h_{kn} v_{fn}}{[\tilde{\boldsymbol{W}}\boldsymbol{H}]_{fn}}}{\left(1 + \frac{1}{\beta_k}\right) \sum_n h_{kn}}. \tag{11}$$

Note the symmetry of the update rules (the roles of $\boldsymbol{W}$ and $\boldsymbol{H}$ are simply exchanged), due to the symmetry of the penalty term, that reduces to $L_{\boldsymbol{\beta}}(\boldsymbol{\Lambda}\boldsymbol{H}) = \sum_{fkn} \frac{w_{fk} h_{kn}}{\beta_k} + cst$.

2) Algorithm for $\alpha_k < 1$: When $\alpha_k < 1$, then $(\alpha_k - 1)$ becomes negative and the minorization of the log part of the penalty term used for the update of $\boldsymbol{W}$ in Section III-A-II does not hold anymore. It is possible to use instead a first order Taylor approximation of $\log \|\boldsymbol{w}_k\|$ (using the property that a concave function is majorized by its tangent), that leads to

$$w_{fk} \leftarrow \tilde{w}_{fk} \frac{\sum_n \frac{h_{kn} v_{fn}}{[\tilde{\boldsymbol{W}}\boldsymbol{H}]_{fn}}}{\frac{(1-\alpha_k)N}{\|\tilde{\boldsymbol{w}}_k\|} + \left(1 + \frac{1}{\beta_k}\right) \sum_n h_{kn}}$$

i.e., the term $\frac{(1-\alpha_k)N}{\|\tilde{\boldsymbol{w}}_k\|}$ that appeared in the numerator of (9) is moved to the denominator (after changing its sign). The main source of difficulty when $\alpha_k < 1$ lies in the update of $\boldsymbol{H}$. The update given by (8) still maximizes the auxiliary function, but not under the nonnegativity constraint. The update may fail to satisfy the nonnegative constraint because of the $(\alpha_k - 1)$ term at the numerator which is now negative. Truncating h_{kn} to zero when it fails to satisfy the nonnegative constraint does provide a valid ascent algorithm, but any coefficient that hits zero will remain zero. Other schemes could be envisaged for $\boldsymbol{H}$ (such as projected gradient descent) but we will not pursue this issue as it is out of main scope of this paper, given that Gamma shape parameters less than one are rarely used in practice.

IV. Algorithms for MMLE

We investigate a set of Expectation-Maximization (EM) [20] algorithms for MMLE. The EM algorithm converges to a stationary point of the likelihood function by iteratively evaluating (E-step) and maximizing (M-step) the expected log-likelihood of the data completed by some latent data. For example, for the

observation model $p(\boldsymbol{V}|\boldsymbol{W},\boldsymbol{H})$ in (2) with the prior distribution $p(\boldsymbol{H})$, an EM algorithm can be built on the complete data set $(\boldsymbol{V},\boldsymbol{H})$ by iteratively evaluating and maximizing the functional defined by

$$Q_1(\boldsymbol{W}|\tilde{\boldsymbol{W}}) = \int_{\boldsymbol{H}} \log p(\boldsymbol{V},\boldsymbol{H}|\boldsymbol{W}) p(\boldsymbol{H}|\boldsymbol{V},\tilde{\boldsymbol{W}})\, d\boldsymbol{H}. \quad (12)$$

As it appears the posterior distribution $p(\boldsymbol{H}|\boldsymbol{V},\tilde{\boldsymbol{W}})$ is not analytically available and the functional can neither be evaluated nor maximized. An other EM algorithm can be built on the larger complete data set $(\boldsymbol{C},\boldsymbol{H})$, exploiting the composite model representation of (3), leading to[2]

$$Q_2(\boldsymbol{W}|\tilde{\boldsymbol{W}}) = \int_{\boldsymbol{C},\boldsymbol{H}} \log p(\boldsymbol{C},\boldsymbol{H}|\boldsymbol{W}) p(\boldsymbol{C},\boldsymbol{H}|\boldsymbol{V},\tilde{\boldsymbol{W}})\, d\boldsymbol{C}\, d\boldsymbol{H}. \quad (13)$$

While this functional is still intractable, the posterior $p(\boldsymbol{C},\boldsymbol{H}|\boldsymbol{V},\tilde{\boldsymbol{W}})$ in the composite model representation is easier to infer using variational Bayes or Markov chain Monte Carlo methods. In Sections IV-A and IV-B, we will describe two EM algorithms where the E-steps are based on these computational methods.

A. Variational Bayes EM (VBEM)

A variational EM algorithm [21] for the maximization of $C_{ML}(\boldsymbol{W})$ based on the functional $Q_2(\boldsymbol{W}|\tilde{\boldsymbol{W}})$ can be constructed as follows. As explained above, we resort to a variational approximation of the posterior $p(\boldsymbol{C},\boldsymbol{H}|\boldsymbol{V},\tilde{\boldsymbol{W}})$ that renders all derivations tractable, though at the cost of approximate inference. The two steps of the variational EM are described next.

1) 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 $Q_2(\boldsymbol{W}|\tilde{\boldsymbol{W}})$. As fundamental to variational approximations, the computation of $q(\boldsymbol{C},\boldsymbol{H})$ relies on the minimization of the KL divergence (in *distribution*) between $q(\boldsymbol{C},\boldsymbol{H})$ and $p(\boldsymbol{C},\boldsymbol{H}|\boldsymbol{V},\tilde{\boldsymbol{W}})$, given a parametric form of $q(\boldsymbol{C},\boldsymbol{H})$. The variational objective function may be decomposed as

$$\begin{aligned}&\mathrm{KL}[q(\boldsymbol{C},\boldsymbol{H})|p(\boldsymbol{C},\boldsymbol{H}|\boldsymbol{V},\tilde{\boldsymbol{W}})]\\ &\quad = \log p(\boldsymbol{V}|\tilde{\boldsymbol{W}}) + \mathrm{KL}[q(\boldsymbol{C},\boldsymbol{H})|p(\boldsymbol{C},\boldsymbol{H}|\tilde{\boldsymbol{W}})]. \quad (14)\end{aligned}$$

Because the marginal likelihood $\log p(\boldsymbol{V}|\tilde{\boldsymbol{W}})$ is independent of $q(\boldsymbol{C},\boldsymbol{H})$, the minimization of the variational objective may be replaced by the (simpler) maximization of

$$\mathrm{L}[q(\boldsymbol{C},\boldsymbol{H})] = -\mathrm{KL}[q(\boldsymbol{C},\boldsymbol{H})|p(\boldsymbol{C},\boldsymbol{H}|\tilde{\boldsymbol{W}})]$$

which forms a lower bound of the marginal likelihood $\log p(\boldsymbol{V}|\tilde{\boldsymbol{W}})$ (thanks to nonnegativity of the KL divergence). It can be shown that, given the expression of $p(\boldsymbol{C},\boldsymbol{H}|\tilde{\boldsymbol{W}})$, the following form of variational distribution appears as a natural choice (in particular for tractability):

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

where $\boldsymbol{c}_{fn}$ denotes the vector $[c_{1,fn}, c_{2,fn}, \ldots, c_{K,fn}]^T$ as in Section II-A, $q(\boldsymbol{c}_{fn})$ is multinomial with probabilities $p_{k,fn}$ and $q(h_{kn})$ is a Gamma distribution with shape and scale parameters $\bar{\alpha}_{kn}$ and $\bar{\beta}_{kn}$. The factors $q(h_{kn})$ and $q(\boldsymbol{c}_{fn})$ can be shown to satisfy the following fixed point equations [21]:

$$\log q(h_{kn}) = \langle \log p(\boldsymbol{C},\boldsymbol{H}|\tilde{\boldsymbol{W}})\rangle_{q(\boldsymbol{H}_{-kn})q(\boldsymbol{C})} + cst \quad (15)$$

$$\log q(\boldsymbol{c}_{fn}) = \langle \log p(\boldsymbol{C},\boldsymbol{H}|\tilde{\boldsymbol{W}})\rangle_{q(\boldsymbol{H})q(\boldsymbol{C}_{-fn})} + cst, \quad (16)$$

where $\langle . \rangle_\pi$ denotes expectation under probability distribution π and $\boldsymbol{A}_{-ij}$ refer to the set of coefficients of $\boldsymbol{A}$ excluding a_{ij}. In particular, the optimal variational distribution $q(\boldsymbol{c}_{fn})$ satisfies

$$\begin{aligned}\log q(\boldsymbol{c}_{fn}) = &\sum_k c_{k,fn} \log \tilde{w}_{fk}\\ &+ c_{k,fn}\langle \log h_{kn}\rangle - \log\Gamma(c_{k,fn}+1) + cst,\end{aligned}$$

which lead to the following fixed point update for its probability parameters

$$p_{k,fn} = \frac{\tilde{w}_{fk}\exp(\langle \log h_{kn}\rangle)}{\sum_l \tilde{w}_{fl}\exp(\langle \log h_{ln}\rangle)}$$

where the expectation is w.r.t the variational distribution $q(h_{kn})$ and $\langle \log h_{kn}\rangle = \psi(\bar{\alpha}_{kn}) + \log\bar{\beta}_k$.[3] Similarly, the optimal variational distribution $q(h_{kn})$ satisfies

$$\begin{aligned}\log q(h_{kn}) = &-\left(\sum_f \tilde{w}_{fk} + \frac{1}{\beta_k}\right) h_{kn}\\ &+ \left(\sum_f \langle c_{k,fn}\rangle + \alpha_k - 1\right)\log h_{kn} + cst\end{aligned}$$

from which the updates are found as

$$\begin{aligned}\bar{\alpha}_{kn} &= \alpha_k + \sum_f \langle c_{k,fn}\rangle\\ \frac{1}{\bar{\beta}_{kn}} &= \frac{1}{\bar{\beta}_k} = \frac{1}{\beta_k} + \sum_f \tilde{w}_{fk}\end{aligned}$$

where the expectation is w.r.t $q(c_{k,fn})$ and has the analytical form $\langle c_{k,fn}\rangle = p_{k,fn} v_{fn}$.

2) M-Step: Given the variational distribution $q(\boldsymbol{C},\boldsymbol{H})$ obtained in the E-step, the EM functional can be approximated as

$$\begin{aligned}Q_2(\boldsymbol{W}|\tilde{\boldsymbol{W}}) \approx &\sum_{fn}\sum_k -w_{fk}\langle h_{kn}\rangle\\ &+ \langle c_{k,fn}\rangle(\log w_{fk} + \langle \log h_{kn}\rangle)\\ &- \langle \log\Gamma(c_{k,fn}+1)\rangle\\ &+ \sum_{kn}(\alpha_k - 1)\langle h_{kn}\rangle - \frac{\langle h_{kn}\rangle}{\beta_k}\\ &- \alpha_k \log\beta_k - \log\Gamma(\alpha_k)\end{aligned}$$

where $\langle h_{kn}\rangle = \bar{\alpha}_{kn}\bar{\beta}_{kn}$. Maximization of $Q_2(\boldsymbol{W}|\tilde{\boldsymbol{W}})$ leads to the update rule

$$w_{fk}^{\mathrm{VBEM}} = \frac{\sum_n \langle c_{k,fn}\rangle}{\sum_n \langle h_{kn}\rangle}.$$

[2]Note that $\boldsymbol{V}$ is encompassed in $\boldsymbol{C}$ because of the relation $v_{fn} = \sum_k c_{k,fn}$ and as such needs not to appear in the complete set, see [20].

[3]ψ is the digamma function defined as $\psi(x) = \frac{d}{dx}\log\Gamma(x)$.

If the value of $\langle c_{k,fn}\rangle$ is plugged in, it is easy to see that this is a multiplicative update rule

$$w_{fk}^{\mathrm{VBEM}} = \tilde{w}_{fk}\frac{\sum_n \frac{\exp(\langle \log h_{kn}\rangle)v_{fn}}{[\tilde{\boldsymbol{W}}\exp(\langle \log \boldsymbol{H}\rangle)]_{fn}}}{\sum_n \langle h_{kn}\rangle} \tag{17}$$

which remains nonnegative provided that the initial value of $\tilde{\boldsymbol{W}}$ in nonnegative.

Note that one could contemplate plugging the variational distribution $q(\boldsymbol{H})$ into $Q_1(\boldsymbol{W}|\tilde{\boldsymbol{W}})$ so as to produce an alternative EM algorithm, but the integration incurred in $Q_1(\boldsymbol{W}|\tilde{\boldsymbol{W}})$ is still intractable.

B. Monte Carlo EM (MCEM)

In this section, we describe how the dictionary parameters can be optimized using Monte Carlo EM (MCEM) [22]. The algorithm consists of an E-step where the posterior distribution $p(\boldsymbol{C},\boldsymbol{H}|\boldsymbol{V},\tilde{\boldsymbol{W}})$ is inferred by a Gibbs sampler [23]. This is a Markov chain Monte Carlo (MCMC) method, in which the chain is constructed by drawing samples from full conditional distributions of (blocks of) variables. Then, the EM functional, either Q_1 or Q_2, is approximated using Monte Carlo integration and maximized w.r.t. $\boldsymbol{W}$. Contrary to VBEM, the Monte Carlo approximation is asymptotically exact.

1) E-Step: In our model, the full conditional distributions of component variables, $\boldsymbol{C}$, are multinomial and those of expansion coefficients, $\boldsymbol{H}$, are Gamma. So, it is highly convenient to use the Gibbs sampler that samples $\boldsymbol{H}$ given $\boldsymbol{C}$ and $\boldsymbol{C}$ given $\boldsymbol{H}$. At iteration $i+1$, a new sample $h_{kn}^{(i+1)}$ is drawn from a Gamma distribution $\mathcal{G}(\bar{\alpha}_{kn},\bar{\beta}_k)$ with parameters given by:

$$\bar{\alpha}_{kn} = \alpha_{kn} + \sum_f c_{k,fn}^{(i)}$$
$$\frac{1}{\bar{\beta}_{kn}} = \frac{1}{\bar{\beta}_k} = \frac{1}{\beta_k} + \sum_f \tilde{w}_{fk}$$

and a new sample $\boldsymbol{c}_{fn}^{(i+1)}$ is subsequently drawn from a multinomial distribution with probabilities

$$p_{k,fn} = \frac{\tilde{w}_{fk}h_{kn}^{(i+1)}}{\sum_l \tilde{w}_{fl}h_{ln}^{(i+1)}}.$$

2) M-Step: Once a set of samples from $p(\boldsymbol{C},\boldsymbol{H}|\boldsymbol{V},\tilde{\boldsymbol{W}})$ have been obtained (after a burn in period), the functional $Q_1(\boldsymbol{W}|\tilde{\boldsymbol{W}})$ and $Q_2(\boldsymbol{W}|\tilde{\boldsymbol{W}})$ may be approximated as

$$Q_1(\boldsymbol{W}|\tilde{\boldsymbol{W}}) \approx \frac{1}{N_s}\sum_{i=1}^{N_s}\log p(\boldsymbol{V}|\boldsymbol{W},\boldsymbol{H}^{(i)}) \tag{18}$$

$$Q_2(\boldsymbol{W}|\tilde{\boldsymbol{W}}) \approx \frac{1}{N_s}\sum_{i=1}^{N_s}\log p(\boldsymbol{C}^{(i)},\boldsymbol{H}^{(i)}|\boldsymbol{W}). \tag{19}$$

The maximization of $Q_2(\boldsymbol{W}|\tilde{\boldsymbol{W}})$, approximated as in (19), is available in close form and leads to

$$w_{fk}^{\mathrm{MCEM2}} = \frac{\sum_{in} c_{k,fn}^{(i)}}{\sum_{in} h_{kn}^{(i)}}. \tag{20}$$

The maximization of $Q_1(\boldsymbol{W}|\tilde{\boldsymbol{W}})$, approximated as in (18), is not available in close form and requires an optimization procedure. Equation (18) reduces to

$$Q_1(\boldsymbol{W}|\tilde{\boldsymbol{W}}) \approx -\frac{1}{N_s}\sum_i^{N_s} D_{KL}(\boldsymbol{V}|\boldsymbol{W}\boldsymbol{H}^{(i)}) + cst.$$

By minorizing the individual terms $-D_{KL}(\boldsymbol{V}|\boldsymbol{W}\boldsymbol{H}^{(i)})$ as in (5) and summing up the auxiliary functions corresponding to all samples, one gets a general auxiliary function whose iterative maximization leads to following fixed point equation:

$$w_{fk} \leftarrow w_{fk}\frac{\sum_{in}\frac{h_{kn}^{(i)}v_{fn}}{[\boldsymbol{W}\boldsymbol{H}^{(i)}]_{fn}}}{\sum_{in} h_{kn}^{(i)}}.$$

In practice we perform only one iteration of the fixed point equation (which only increases $Q_1(\boldsymbol{W}|\tilde{\boldsymbol{W}})$ instead of fully maximizing it), starting from the current EM update $\tilde{\boldsymbol{W}}$, so that

$$w_{fk}^{\mathrm{MCEM1}} = \tilde{w}_{fk}\frac{\sum_{in}\frac{h_{kn}^{(i)}v_{fn}}{[\tilde{\boldsymbol{W}}\boldsymbol{H}^{(i)}]_{fn}}}{\sum_{in} h_{kn}^{(i)}}. \tag{21}$$

As it appears the update (21) based on $Q_1(\boldsymbol{W}|\tilde{\boldsymbol{W}})$ is the Rao-Blackwellized version of the update (20) based on $Q_2(\boldsymbol{W}|\tilde{\boldsymbol{W}})$, i.e.,

$$w_{fk}^{\mathrm{MCEM1}} = \frac{\sum_{in}\mathrm{E}[c_{k,fn}|\boldsymbol{V},\tilde{\boldsymbol{W}},\boldsymbol{H}^{(i)}]}{\sum_{in} h_{kn}^{(i)}}.$$

Rao-Blackwellization is known to produce updates with lesser variance [24] and as such, we will use update(21) in practice.

Marginal Likelihood From Gibbs Samples (Chib's Method): While in VBEM an estimation of the criterion $C_{ML}(\tilde{\boldsymbol{W}})$ comes as a by-product in the form of the lower bound $\mathrm{L}[q(\boldsymbol{C},\boldsymbol{H})]$, the MCEM approach does not come with such an estimation. It is however possible to "recycle" the samples of $\boldsymbol{C}$ obtained during E-step to form an estimation of the objective, using Chib's method [12]. Chib's approach is based on the following relation (Bayes' theorem)

$$C_{ML}(\tilde{\boldsymbol{W}}) = \log p(\boldsymbol{V}|\tilde{\boldsymbol{W}},\boldsymbol{H}^*) + \log p(\boldsymbol{H}^*) - \log p(\boldsymbol{H}^*|\boldsymbol{V},\tilde{\boldsymbol{W}}) \tag{22}$$

which holds for any $\boldsymbol{H}^*$, but recommendation is to use the posterior mode $\arg\max p(\boldsymbol{H}|\boldsymbol{V},\tilde{\boldsymbol{W}})$ in practice. The first two terms in the right side of (22) are easy to compute, and are given by the generative model. The last term is not readily available, however it can be estimated by the following Monte Carlo integration

$$p(\boldsymbol{H}^*|\boldsymbol{V},\tilde{\boldsymbol{W}}) = \int p(\boldsymbol{H}^*|\boldsymbol{C},\tilde{\boldsymbol{W}})p(\boldsymbol{C}|\boldsymbol{V},\tilde{\boldsymbol{W}})\,d\boldsymbol{C} \approx \frac{1}{N_s}\sum_{i=1}^{N_s} p(\boldsymbol{H}^*|\boldsymbol{C}^{(i)}\tilde{\boldsymbol{W}}),$$

where $\boldsymbol{C}^{(i)}$ are the samples drawn from the posterior $p(\boldsymbol{C},\boldsymbol{H}|\boldsymbol{V},\tilde{\boldsymbol{W}})$ during Gibbs sampling.

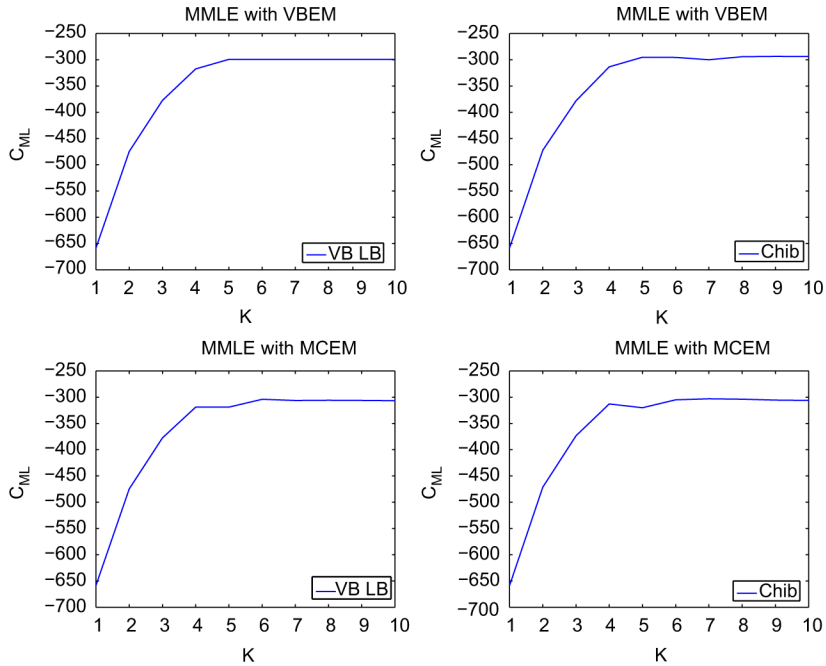

Fig. 1. Synthetical dataset. Marginal log-likelihood $C_{ML}(\hat{\boldsymbol{W}})$ for $K = 1, \ldots, 10$. Top row: $\hat{\boldsymbol{W}}$ is obtained with VBEM (4000 iterations), bottom-row: $\hat{\boldsymbol{W}}$ is obtained with MCEM (100 samples including burn in, 8000 iterations). Left column: $C_{ML}(\hat{\boldsymbol{W}})$ is approximated by the VB lower bound, right column: $C_{ML}(\hat{\boldsymbol{W}})$ is computed with Chib's method.

V. Experiments

We study the performances of MJLE and MMLE on real and synthetical data. We start with a synthetical data of small size to study the performances of the two MMLE methods, VBEM and MCEM. Next, we test MJLE and MMLE on the Swimmer dataset which is a benchmark dataset in dictionary learning problems. We also compare the two approaches on a face reconstruction experiment on CBCL dataset. As the last experiment, we compare MJLE, MMLE and LDA on the topic learning problem on text (lyrics) data. In the experiments, the prior hyperparameters are fixed to $\alpha_k = 1$ (exponential distribution) and $\beta_k = 1$, i.e., $p(h_{kn}) = \exp(-h_{kn})$, unless otherwise stated. In the algorithms, we initialize $\boldsymbol{W}$ and $\boldsymbol{H}$ ($\langle \boldsymbol{H} \rangle$ in VBEM, $\boldsymbol{H}^{(0)}$ in MCEM) as $\texttt{W} = \texttt{abs}(\texttt{randn}(\texttt{F},\texttt{K})) + \texttt{ones}(\texttt{F},\texttt{K})$ and $\texttt{H} = \texttt{abs}(\texttt{randn}(\texttt{K},\texttt{N})) + \texttt{ones}(\texttt{K},\texttt{N})$, in Matlab notations. In MCEM one third of the samples are used as the burn in period and discarded in the estimations.

A. Synthetical Dataset

We fix a $\boldsymbol{W}^*$ matrix of size 10×5, of which columns are linearly independent and consist of zeros and tens. We generate data from model in Section II-A, with $N = 50$.

In Fig. 1, we investigate MMLE with the two possible EM algorithms (VB or MC) and the two possible means of estimating the marginal likelihood C_{ML} (variational lower bound or Chib's evaluation), as we increase the number of components, K. On the top row the dictionaries are estimated with VBEM, and with MCEM on the bottom row. On the left column the marginal likelihood is estimated with the variational lowerbound and on the right it is estimated with Chib's method. A first observation is that the four plots essentially coincides. This illustrates the equivalent performances of algorithms and likelihood evaluation techniques. A second observation is that the marginal likelihood ceases increasing when $K \geq 5$. As a matter of fact, visual inspection of the estimated dictionaries reveals that both VBEM and MCEM push the redundant columns of $\boldsymbol{W}$ to zero when K is greater than the ground truth value. In other words, MMLE performs an intrinsic automatic order selection during inference.

Fig. 2 displays the joint and marginal log-likelihood criteria, $C_{JL}(\hat{\boldsymbol{W}}, \hat{\boldsymbol{H}})$ and $C_{ML}(\hat{\boldsymbol{W}})$, when $\hat{\boldsymbol{W}}$ is learnt with MJLE. Right plot of Fig. 2 verifies that learning $\boldsymbol{W}$ from MJLE *does not* increase the marginal likelihood criterion as K increases, i.e., we verify experimentally that MJLE does not imply MMLE.

On this synthetical dataset, 4000 iterations of MJLE and MMLE with VBEM take 1.28 sec and 2.15 sec, respectively, on an average computer with a 2.66 GHz CPU and 4GB memory. However, 4000 iterations of MCEM which uses 100 samples per iteration including burn in take 673 sec, which means that running MCEM on large datasets can be prohibitive. In this section we showed that the dictionaries $\hat{\boldsymbol{W}}$ estimated with VBEM and MCEM behave similarly as K is increased and the marginal log likelihood values $C_{ML}(\hat{\boldsymbol{W}})$ of these dictionaries are very close. This means that approximate optimization with VBEM does not lead to a significant performance loss. Thus, we will not consider MCEM on the large datasets of the upcoming experiments.

B. Swimmer Dataset

To further investigate the automatic model order selection feature of MMLE, we consider the synthetical Swimmer dataset [25], for which a ground truth can be defined. The

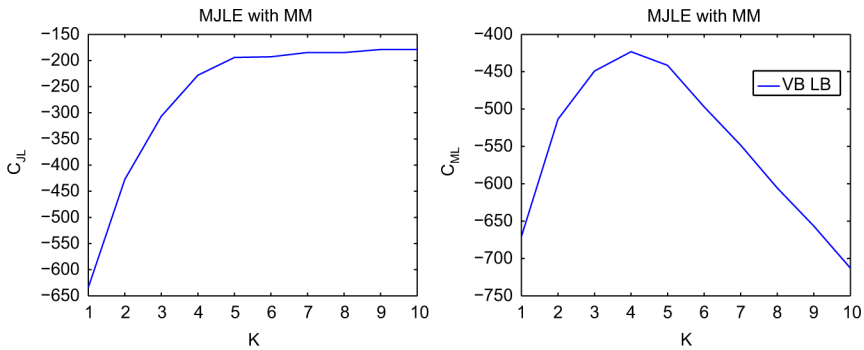

Fig. 2. Synthetical dataset. Joint and marginal log-likelihood criteria $C_{JL}(\hat{\boldsymbol{W}}, \hat{\boldsymbol{H}})$ and $C_{ML}(\hat{\boldsymbol{W}})$ when $\hat{\boldsymbol{W}}$ is learnt with MJLE (4000 iterations), for $K = 1, \ldots, 10$.

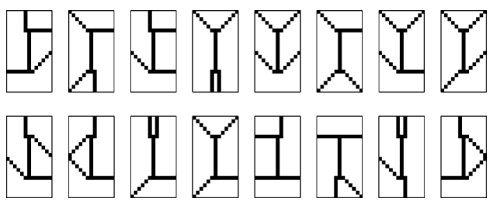

Fig. 3. Data samples consisting of images from the Swimmer dataset with Poisson noise.

dataset is composed of 256 images of size 13×22, 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. Hence, the ground truth dictionary corresponds to the collection of individual limb positions. As explained in [25] the torso is an unidentifiable component that can be paired with any of the limbs, or even split among the limbs. In our experiments, we mapped the binary values onto the range $[1, 100]$ with which we generated the data using the Poisson observation model, see some samples in Fig. 3.

The behavior of the marginal log likelihood, $C_{ML}(\hat{\boldsymbol{W}})$, against K on the noisy swimmer problem is given in Fig. 4. The marginal likelihood increases until $K = 16$, after that value redundant components are assigned to zero and the likelihood does not change. In Fig. 5, we compared the dictionaries learnt by MJLE and MMLE with $K = 20$ components. As can be seen from Fig. 5(a), MJLE produces spurious or duplicated components, i.e., overfits. In contrast, the ground truth is perfectly recovered with MMLE. We do not have a rigorous explanation for the self-ability of MMLE to prune columns of $\boldsymbol{W}$, but in Appendix we present a Laplace approximation of C_{ML} that gives an intuition of why this is happening.

C. Interpolation on CBCL Dataset

We now apply MJLE and MMLE on a missing data prediction task to further investigate overfitting. We used the CBCL face dataset [26] which is composed of 2429 face images. The raw images are 8 bits grayscale images with dimensions 19×19. Similar to [1], the grayscale intensities have been linearly scaled so that the pixel mean and standard deviation were equal to 64, and then clipped to the range $(0, 255]$, so as to homogenize illumination.

We define a Bernoulli mask variable m_{fn} with probability p. m_{fn} indicates whether v_{fn} is observed or not. In the presence of missing data, the observation model becomes

$$p(\boldsymbol{V}, \boldsymbol{C}|\boldsymbol{W}, \boldsymbol{H}) = \prod_{fn} \left(\left(\mathbf{1}_{v_{fn} = \sum_k c_{k,fn}} \right) \times \prod_k p(c_{k,fn}|w_{fk}, h_{kn}) \right)^{m_{fn}}.$$

The algorithms of the previous sections can easily be derived using this observation model and the equations only slightly change. For example, in (17), one only needs to sum over the observed entries at the numerator and denominator of the update rule. This leads to more general algorithms for which the algorithms described in Sections III and IV for the completely observed data become special cases. We generated a mask $\boldsymbol{M} = \{m_{fn}\}$ with $p = 0.5$ and estimated $\hat{\boldsymbol{W}}$ and $\hat{\boldsymbol{H}}$ by MJLE and MMLE[4] with $K = 300$, $\alpha_k = 1$, $\beta_k = 1$. We used the same stopping criteria for both methods: the algorithms were exited when the relative increase in the likelihood is less than a threshold (10^{-6}) or the maximum number of iterations (4000) is reached. We repeated each experiment 5 times and reported the results of the run with the highest likelihood end value. We reconstructed the missing values using $\hat{\boldsymbol{V}} = \hat{\boldsymbol{W}}\hat{\boldsymbol{H}}$ and computed the Peak Signal to Noise Ratio (PSNR) between original and reconstructed images which is defined as $20 \log_{10} \left(\frac{FP}{\|\boldsymbol{V} - \hat{\boldsymbol{V}}\|_2} \right)$, where P is the maximum possible pixel value (255 in this experiment). In Fig. 6, we present some examples of the faces in the dataset, their masked versions which were used as data in this experiment and the reconstructions with MMLE and MJLE, respectively. PSNR values for the reconstructions are displayed on top of them. The average PSNR values obtained are 27.2 dB with MMLE and 26.6 dB with MJLE. For 1640 faces (67% of the total number) the PSNR values of MMLE are higher. In general, the reconstructions obtained with MMLE are smoother and more pleasant to look at. This is because MJLE overfitted the data with $K = 300$ while in MMLE 230 out of 300 dictionary columns were assigned to very low values ($\|\boldsymbol{w}_k\| < 10^{-14}$).

We also investigated the effect of α_k on the performances of the methods. We repeated the same experiment with 11 values of α_k between 1 and 100. The PSNR values obtained with 10

[4] $\hat{\boldsymbol{H}}$ in MMLE was taken to be the mean of $q(\boldsymbol{H})$ at convergence.

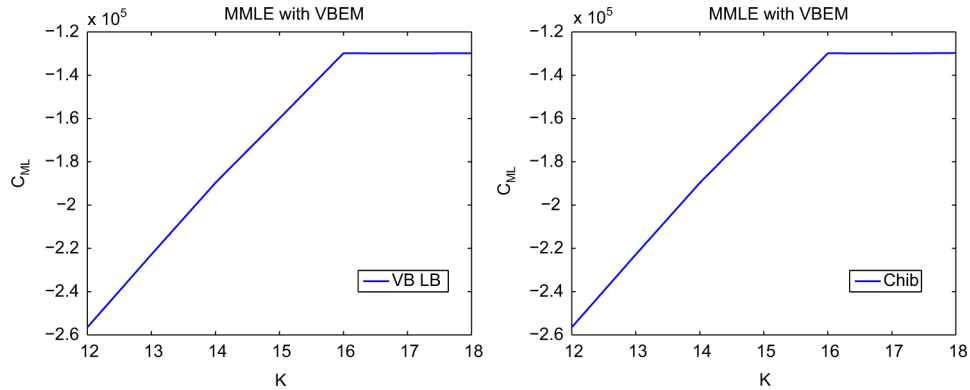

Fig. 4. Swimmer dataset. Marginal log-likelihood $C_{ML}(\hat{\boldsymbol{W}})$ for $K = 12, \ldots, 18$ estimated by VB lower bound (left), Chib's method (right). $\hat{\boldsymbol{W}}$ is obtained with VBEM (4000 iterations).

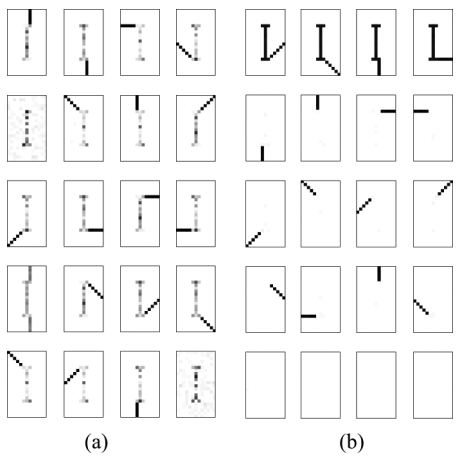

Fig. 5. Dictionaries learnt from the noisy Swimmer dataset with $K = 20$ (a) W_{MJLE} (b) W_{MMLE}.

Fig. 6. Original faces from the CBCL dataset, masked data and reconstructions with MMLE and MJLE using 300 components. PSNR values (dB) are displayed on top of reconstructions.

repetitions of each method are presented in Fig. 7. As α_k increases, the means (and the variances) of the prior distributions

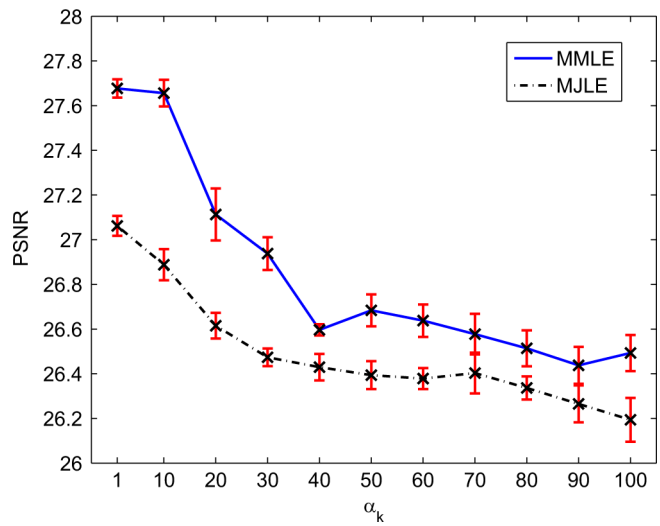

Fig. 7. Average PSNR values obtained with MMLE (solid) and MJLE (dotted) on the CBCL dataset with missing entries for various values of α_k in $[1, 100]$. Error bars represent standard deviations of 10 repetitions.

get bigger and more components (columns of $\boldsymbol{W}$) tend to converge to zero. The algorithm becomes more sensitive to initialization. Still, as can be seen from Fig. 7, the performance does not drop too drastically and MMLE gives consistently better results than MJLE. As a rule of thumb, $\alpha_k \in [1, 10]$ results in optimal degree of pruning in all of our experiments. Optimization of the hyperparameter α_k can be advantageous, but will make the algorithm slower than fixing it to a "good" value.

D. Musixmatch Lyrics Dataset

MusiXmatch (Million Song Dataset) [13] is a lyrics database with more than 230,000 songs. Each song constitutes a column of the data in a bag-of-words representation. The number of occurrences of the most common 5,000 words are used as the feature set. These 5,000 words are stemmed, i.e., related words are mapped onto their roots and cover 92% of all words that appear in the lyrics. The dataset also contains lyrics in other languages like Spanish, German, French, etc. We estimated dictionaries from a random subset of the dataset ($N = 10{,}000$) using MJLE, MMLE and LDA [10] with $K = 200$ components and $N_{iter} = 1000$ iterations. After estimating the dictionaries, in order to reconstruct components we estimated the expansion

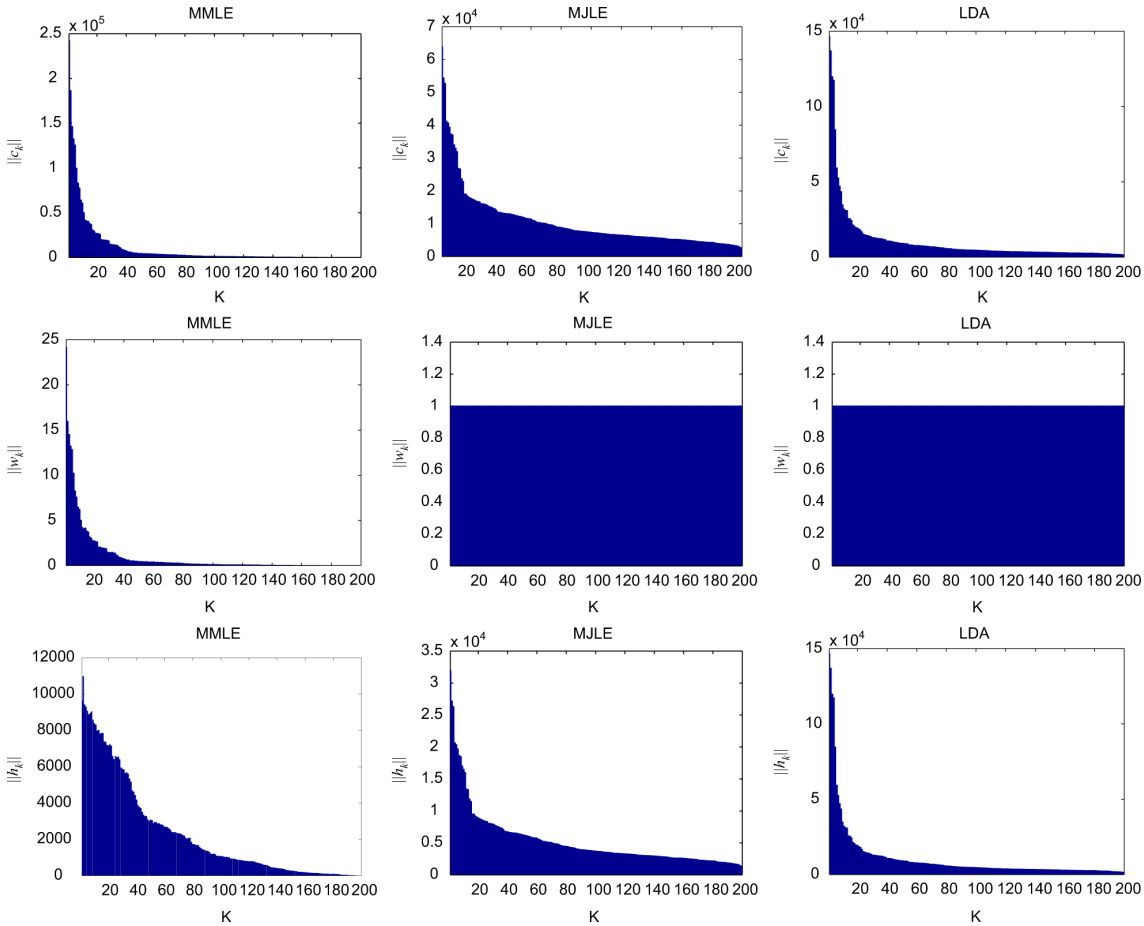

Fig. 8. MusiXmatch dataset analysis. Norms of components $||\boldsymbol{C}_k||$ (top), dictionary vectors $||\boldsymbol{w}_k||$ (middle) and activation vectors $||\boldsymbol{h}_k||$ (bottom), estimated by MMLE/VBEM (left) and MJLE (center) and LDA (right). $N_{iter} = 1000$, $K = 200$.

coefficients $\hat{\boldsymbol{H}}$ using the MAP estimator from corresponding models. The MAP estimator for $\boldsymbol{H}$ is a by product of the MJLE algorithm. For MMLE and LDA, we learnt them from GaP and LDA models with dictionaries set to their previously estimated values. The norms of dictionary vectors, $||\hat{\boldsymbol{w}}_k||$, expansion vectors, $||\hat{\boldsymbol{h}}_k||$, and components, $||\hat{\boldsymbol{C}}_k||$, are presented in Fig. 8.[5] The components are reconstructed using the posterior mean

$$\hat{c}_{k,fn} = \frac{\hat{w}_{fk}\hat{h}_{kn}}{\sum_j \hat{w}_{fj}\hat{h}_{jn}} v_{fn} \tag{23}$$

and components in all of these plots are sorted in descending order of $||\hat{\boldsymbol{C}}_k||$.

With MMLE/VBEM (left column in Fig. 8), we observe that 58 out of $K = 200$ dictionary columns have a norm less than 0.05, which is similarly reflected in the component norms. In MJLE and LDA, $||\hat{\boldsymbol{w}}_k||$ are not very informative because all dictionary columns sum to one by design and norms of the activation vectors, $||\hat{\boldsymbol{h}}_k||$, are also equal to those of the components, $||\hat{\boldsymbol{C}}_k||$. Comparing the components of MMLE, MJLE and LDA, we can see that MMLE explains the data making use of less components than MJLE and LDA. LDA also has this tendency but it is impossible to prune out components because every column of $\boldsymbol{W}$ is a discrete probability distribution and sum up to one.

For illustration, we now describe the results of decomposition with MMLE. In order to find the most representative songs of a topic, we propose to investigate the contribution of each topic in songs. We define the contribution of a topic as the number of words generated from that topic divided by the total number of words in that song

$$l_{kn} = \frac{\sum_f \hat{c}_{k,fn}}{\sum_k \sum_f \hat{c}_{k,fn}}.$$

The contributions estimated by MMLE are presented in Fig. 9. The components (rows) are again sorted in descending order of $||\hat{\boldsymbol{C}}_k||$. First 25–30 components mainly contain stop words and are found in most of the songs. The sparser rows among these (e.g., $k = 7, 13, 16\ldots$) belong to languages other than English. Manual inspection on the dictionary reveals that columns of the dictionary actually may correspond to themes of the songs, such as love, sex, religion, war, etc. In Table I, we present the most representative songs for a selection of components, i.e., songs for which l_{kn} are highest for a given k. We also present the most important words of the components and the most frequently used words in the songs. The importance, z_{fk}, is related

[5] $\hat{\boldsymbol{w}}_k = \{\hat{w}_{fk}\}_f$ and $\hat{\boldsymbol{h}}_k = \{\hat{h}_{kn}\}_n$ denote vectors of size F and N, respectively. $\hat{\boldsymbol{C}}_k = \{\hat{c}_{k,fn}\}_{fn}$ is a matrix of size $F \times N$.

TABLE I
MOST REPRESENTATIVE SONGS FOR FOUR OF THE COMPONENTS AND WORDS THAT APPEAR MOST FREQUENTLY.

($k = 2$) get nigga the ya shit like fuck em got hit bitch up off yall ass they that cmon money and	
UGK (Underground Kingz) - Murder	i the to nigga my a you got murder and it is am from we so with they yo cuz
Big Punisher - Nigga Shit	shit that nigga the i and my what to out am in on for love me with gettin you do
E-40 - Turf Drop [Clean]	gasolin the my i hey to a it on you some fuck spit of what one ride nigga sick gold
Cam'Ron - Sports Drugs & Entertainment	a the you i got yo stop shot is caus or street jump short wick either to on but in
Foxy Brown - Chyna Whyte	the nigga and you shit i not yall to a on with bitch no fuck uh it money white huh
($k = 8$) god of blood soul death die fear pain hell power within shall earth blind human bleed scream evil holi peac	
Demolition Hammer - Epidemic Of Violence	of pain death reign violenc and a kill rage vicious the to in down blue dead cold
Disgorge - Parallels Of Infinite Torture	of the tortur by their within upon flow throne infinit are no they see life eye befor
Tacere - Beyond Silence	silenc beyond a dark beauti i the you to and me it not in my is of your that do
Cannibal Corpse - Perverse Suffering	to my pain of i me for agoni in by and from way etern lust tortur crave the not be
Showbread - Sampsa Meets Kafka	to of no one die death loneli starv i the you and a me it not in my is your
($k = 26$) she her girl beauti woman & queen sex sexi cloth herself doll shes pink gypsi bodi midnight callin dress hair	
Headhunter - Sex & Drugs & Rock'N Roll	& sex drug rock roll n is good veri inde and not my are all need dead bodi brain i
Holy Barbarians - She	she of kind girl my is the a littl woman like world and gone destroy tiger me on an
X - Devil Doll	devil doll her she and a the in is of eye bone & shoe rag batter you to on no
Kittie - Paperdoll	her she you i now soul pain to is down want eat fit size and not in all dead bodi
Ottawan - D.I.S.C.O.	is she oh disco i o s d c super incred a crazi such desir sexi complic special candi
($k = 13$) je et les le pas dan pour des cest qui de tout mon moi au comm ne sur jai	
Veronique Sanson - Feminin	cest comm le car de bien se les mai a fait devant heur du et une quon quelqu etre
Nevrotic Explosion - Heritage	quon faut mieux pour nous qui nos ceux de la un plus tous honor parent ami oui
Kells - Sans teint	de la se le san des est loin peur reve pour sa sang corp lumier larm
Stille Volk - Corps Magicien	de les ell dan la se le du pass est sa par mond leur corp vivr lair voyag feu
Florent Pagny - Tue-Moi	si plus que un tu mon mes jour souvenir parc

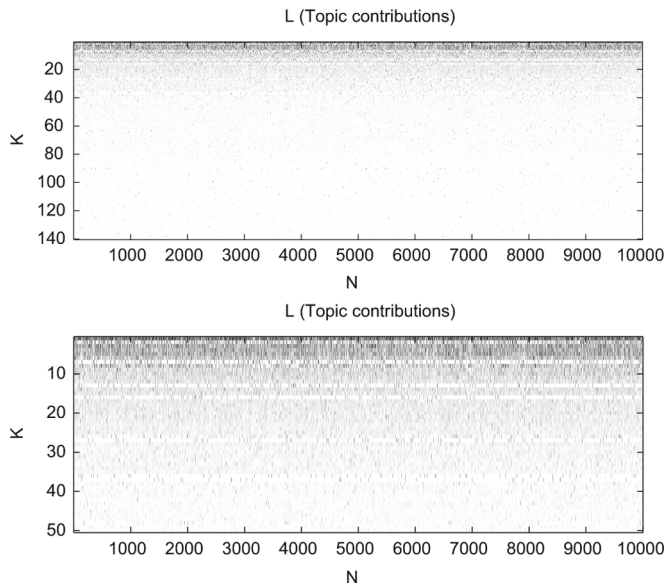

Fig. 9. Contribution matrix $\boldsymbol{L}$ estimated with MMLE on a the MusiXmatch dataset. Whole matrix (top), first 50 rows zoomed in (bottom). l_{kn} represents the contribution of kth topic in nth song. Higher values are represented by darker colors.

to how well that word is represented in that column and is defined by

$$z_{fk} = \frac{(\hat{w}_{fk} - \bar{w}_f)^2}{\bar{w}_f} \mathbf{1}_{\{\hat{w}_{fk} > \bar{w}_f\}} \tag{24}$$

where $\bar{w}_f$ denotes the average over components, i.e., $\bar{w}_f = \frac{1}{K}\sum_{k=1}^{K} \hat{w}_{fk}$.

Finally, we analyze lyrics of two songs by considering the topics of which contribution, l_{kn} for fixed n, are the highest. In Table II, we display ten highest contributing dictionary columns for the song "Do You Love Me?" by Nick Cave and the Bad Seeds. Each column is represented with the most important words according to (24). The same information for "California Love" by 2pac is presented in Table III. It is interesting to notice how different the significant dictionaries are in two songs which both contain "love" in the title.

VI. DISCUSSION AND CONCLUSION

In this paper we have challenged the standard NMF approach to nonnegative dictionary learning, which is based on maximum *joint* likelihood estimation (MJLE) and is ill-posed because the number of parameters to be learned increases with the data. Our approach here is maximum *marginal* likelihood estimation (MMLE), for which we proposed two EM algorithms, VBEM and MCEM. While both of these EM algorithms depend on approximating the functional in the E-step, VBEM maximizes a lower bound of the marginal likelihood, whereas MCEM is asymptotically exact. The same is true for the evaluation of the marginal likelihood, i.e., the approximation with Chib's method is asymptotically exact. Our experiments on synthetical data showed that the dictionaries estimated by these two EM algorithms show similar characteristics. In addition, the variational lower bound was tight with the results obtained with Chib's method.

Experiments on real and synthetical data have 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 is not by design as in other works of automatic relevance determination (e.g., [27], [28]), but stems from the objective function. The dictionaries estimated with MMLE lead to more accurate and interpretable components with no overfitting. In contrast with other model selection approaches in fully Bayesian settings (e.g., [9], [17]), which are 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 VBEM algorithm

TABLE II
TEN HIGHEST CONTRIBUTING DICTIONARY COLUMNS FOR "DO YOU LOVE ME?" BY NICK CAVE AND THE BAD SEEDS. EACH COLUMN IS REPRESENTED BY 15 MOST IMPORTANT WORDS. THE LAST ROW DISPLAYS THE CONTRIBUTION VALUE, l_{kn}, FOR THE CORRESPONDING DICTIONARY COLUMN. THESE 10 COLUMNS COVER 75% OF THE SONG.

the	you	love	not	i	me	she	god	was	so
in	your	buy	do	am	give	her	of	would	to
and	can	liar	wanna	myself	tell	girl	blood	could	for
of	if	tender	care	like	call	beauti	soul	were	now
world	know	dear	bad	know	mmm	woman	death	said	here
with	want	instrument	nobodi	need	show	&	die	had	again
they	make	mood	anyth	want	beg	queen	fear	thought	wait
from	when	treasur	want	feel	rescu	sex	pain	wish	long
as	see	emot	worri	caus	teas	sexi	hell	knew	too
by	yourself	untru	ai	and	squeez	cloth	power	came	home
at	need	surrend	treat	out	everytim	herself	within	made	and
out	with	deeper	but	sorri	knee	doll	shall	told	much
to	feel	sparkl	know	see	strife	shes	earth	took	alon
into	that	sweetest	money	in	contempl	pink	blind	saw	still
sky	how	diamond	hurt	swear	guarante	gypsi	human	then	how
0.15	0.10	0.09	0.08	0.08	0.08	0.06	0.04	0.04	0.03

TABLE III
TEN HIGHEST CONTRIBUTING DICTIONARY COLUMNS FOR "CALIFORNIA LOVE" BY 2 PAC. EACH COLUMN IS REPRESENTED BY 15 MOST IMPORTANT WORDS. THE LAST ROW DISPLAYS THE CONTRIBUTION VALUE, l_{kn}, FOR THE CORRESPONDING DICTIONARY COLUMN. THESE 10 COLUMNS COVER 87% OF THE SONG.

get	the	shake	it	you	we	come	yeah	around	i
nigga	in	motion	is	your	our	babi	five	goe	am
the	and	bump	take	can	us	magic	four	summer	myself
ya	of	groov	doe	if	togeth	til	woo	melt	like
shit	world	booti	make	know	both	lovin	summertim	wors	know
like	with	shakin	easi	want	higher	im	girlfriend	california	need
fuck	they	thigh	matter	make	ourselv	shi	wow	jone	want
em	from	oon	real	when	each	bodi	lala	scheme	feel
got	as	shiver	game	see	divid	sweat	grip	texa	caus
hit	by	panic	possibl	yourself	nation	cant	pine	dreamin	and
bitch	at	dick	play	need	unit	your	engin	screw	out
up	out	claw	chanc	with	other	birthday	feather	darker	sorri
off	to	opportun	give	feel	noel	wont	clap	careless	see
yall	into	collid	harder	that	standard	there	mornin	consol	in
ass	sky	ness	quit	how	rule	bella	gotta	giant	swear
0.49	0.09	0.08	0.07	0.03	0.03	0.02	0.02	0.02	0.02

once. This property is not shared by LDA although it is similar to MMLE in spirit. This is because in LDA the dictionary columns are constrained to sum to one.

The computational costs of MJLE and MMLE/VBEM are comparable, whereas MMLE/MCEM is very computationally demanding. Since two MMLE algorithms perform similarly, it is natural to use only VBEM on large datasets. In addition, VBEM can be made even faster by not updating the components which are already zero.

APPENDIX
LAPLACE APPROXIMATION OF $C_{ML}(\boldsymbol{W})$

In this section we give a Laplace approximation of the integral involved in $C_{ML}(\boldsymbol{W})$ whose expression provides an intuition for the pruning effect of MMLE. We can write that

$$C_{ML}(\boldsymbol{W}) = \sum_{n=1}^{N} \log \int_{\boldsymbol{h}_n} p(\boldsymbol{v}_n|\boldsymbol{W}\boldsymbol{h}_n) p(\boldsymbol{h}_n) d\boldsymbol{h}_n. \quad (25)$$

Let $\hat{\boldsymbol{H}}$ be the MAP estimation of $\boldsymbol{H}$ given $\boldsymbol{W}$, i.e.,

$$\hat{\boldsymbol{H}} = \arg\max_{\boldsymbol{H}} \log p(\boldsymbol{V}, \boldsymbol{H}|\boldsymbol{W}).$$

Then, a Laplace approximation of $C_{ML}(\boldsymbol{W})$ around its mode (which essentially consists in replacing the integrand in each term of the sum in (25) by a quadratic function with same mode and curvature at the mode) is given by

$$C_{ML}(\boldsymbol{W}) \approx C_{JL}(\boldsymbol{W}, \hat{\boldsymbol{H}}) - \frac{1}{2}\sum_{n} \log\det \boldsymbol{A}_n + \frac{KN}{2}\log 2\pi \quad (26)$$

where

$$\begin{aligned}\boldsymbol{A}_n &= -\nabla^2_{\boldsymbol{h}_n} \log p(\boldsymbol{v}_n, \boldsymbol{h}_n|\boldsymbol{W})\big|_{\boldsymbol{h}_n=\hat{\boldsymbol{h}}_n} \\ &= \boldsymbol{W}^T \boldsymbol{\Gamma}_{1,n} \boldsymbol{W} + \boldsymbol{\Gamma}_{2,n}\end{aligned}$$

where $\boldsymbol{\Gamma}_{1,n}$ and $\boldsymbol{\Gamma}_{2,n}$ are the diagonal matrices defined by

$$\begin{aligned}\boldsymbol{\Gamma}_{1,n} &= \mathrm{diag}[\boldsymbol{v}_n.(\boldsymbol{W}\hat{\boldsymbol{h}}_n^{.-2})] \\ \boldsymbol{\Gamma}_{2,n} &= \mathrm{diag}[(\boldsymbol{\alpha}-1).\hat{\boldsymbol{h}}_n^{.-2}]\end{aligned}$$

and where the '.' denotes MATLAB-like entry-wise operations. The penalty term $L(\boldsymbol{W}) = \sum_n \log\det \boldsymbol{A}_n$ in (26) will favor solutions such that $\det \boldsymbol{A}_n$ is small, ideally zero. A detailed analysis of $L(\boldsymbol{W})$, not presented here, reveals that it induces group-sparsity at the column level. This for example evident when $\alpha_k = 1$ and thus $\boldsymbol{\Gamma}_{2,n} = \boldsymbol{0}$. In this case, any zero column in $\boldsymbol{W}$ leads to $\det \boldsymbol{A}_n = 0$. While not giving a rigorous explanation, the Laplace approximation of $C_{ML}(\boldsymbol{W})$ hence suggests why MMLE induces self-regularization by pruning.

ACKNOWLEDGMENT

The authors would like to thank O. Cappé, A. Taylan Cemgil, and J. Le Roux for inspiring discussions related to this work.

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] W. H. Richardson, "Bayesian-based iterative method of image restoration," *J. Opt. Soc. Amer.*, vol. 62, pp. 55–59, 1972.
[3] L. B. Lucy, "An iterative technique for the rectification of observed distributions," *Astron. J.*, vol. 79, pp. 745–754, 1974.
[4] M. S. Lewicki and T. J. Sejnowski, "Learning overcomplete representations," *Neural Comput.*, vol. 12, pp. 337–365, 2000.
[5] J. F. Canny, "GaP: A factor model for discrete data," in *Proc. 27th ACM Int. Conf. Res. Develop. Inform. Retrieval (SIGIR)*, Sheffield, U.K., 2004, pp. 122–129.
[6] R. Nallapati, W. W. Cohen, S. Ditmore, J. Lafferty, and K. Ung, "Multiscale topic tomography," in *Proc. 13th ACM SIGKDD Int. Conf. Knowledge Discovery Data Mining (KDD'07)*, San Jose, CA, 2007.
[7] M. Titsias, "The infinite gamma-poisson feature model," presented at the Adv. Neural Inform. Process. Syst. (NIPS'07), Vancouver, BC, Canada, 2007.
[8] W. L. Buntine and A. Jakulin, "Discrete component analysis," *Lecture Notes in Comput. Sci.*, vol. 3940, pp. 1–33, 2006.
[9] A. T. Cemgil, "Bayesian inference for nonnegative matrix factorisation models," *Comput. Intell. Neurosci.*, p. 17, 2009, Article ID 785152, doi: 10.1155/2009/785152.
[10] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn.Res.*, vol. 3, pp. 993–1022, 2003.
[11] T. Hofmann, "Probabilistic latent semantic indexing," presented at the 22nd Int. Conf. Res. Develop. Inform. Retrieval (SIGIR), Berkeley, CA, 1999.
[12] S. Chib, "Marginal likelihood from the Gibbs output," *J. Amer. Statist. Assoc.*, vol. 90, no. 432, pp. 1313–1321, 1995.
[13] T. Bertin-Mahieux, D. P. Ellis, B. Whitman, and P. Lamere, "The million song dataset," presented at the 12th Int. Soc. Music Inform. Retrieval Conf. (ISMIR'11), Miami, FL, 2011.
[14] O. Dikmen and C. Févotte, "Maximum marginal likelihood estimation for nonnegative dictionary learning," in *Proc. Int. Conf. Acoust. Speech Signal Process.(ICASSP'11)*, Prague, Czech Republic, 2011.
[15] C. Févotte and A. T. Cemgil, "Nonnegative matrix factorisations as probabilistic inference in composite models," in *Proc. 17th Eur. Signal Process. Conf. (EUSIPCO)*, Glasgow, Scotland, Aug. 2009, pp. 1913–1917.
[16] C. Févotte and J. Idier, "Algorithms for nonnegative matrix factorization with the beta-divergence," *Neural Comput.*, vol. 23, no. 9, Sep. 2011.
[17] M. N. Schmidt, O. Winther, and L. K. Hansen, "Bayesian non-negative matrix factorization," in *Proc. 8th Int. Conf. Independent Component Anal. Signal Separation (ICA'09)*, Paraty, Brazil, Mar. 2009.
[18] J. Eggert and E. Körner, "Sparse coding and NMF," in *Proc. IEEE Int. Joint Conf. Neural Netw.*, Budapest, Hungary, 2004, pp. 2529–2533.
[19] A. Lefèvre, F. Bach, and C. Févotte, "Itakura-Saito nonnegative matrix factorization with group sparsity," in *Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Prague, Czech Republic, May 2011.
[20] A. P. Dempster, N. M. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Statist. Soc.*, vol. 1, no. 39, pp. 1–38, 1977.
[21] 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*, J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, and M. West, Eds. London, U.K.: Oxford Univ. Press, 2003.
[22] G. C. G. Wei and M. A. Tanner, "A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms," *J. Amer. Statist. Assoc.*, vol. 85, no. 411, pp. 699–704, 1990.
[23] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 6, no. 6, pp. 721–741, Jun. 1984.
[24] E. Lehmann, *Theory of Point Estimation*. New York: Wiley, 1983.
[25] D. Donoho and V. Stodden, "When does non-negative matrix factorization give a correct decomposition into parts?," in *Advances in Neural Information Processing Systems 16*, S. Thrun, L. Saul, and B. Schölkopf, Eds. Cambridge, MA: MIT Press, 2004.
[26] B. Heisele, T. Poggio, and M. Pontil, "Face detection in still gray images," Center for Biological and Computational Learning, MIT, Cambridge, MA, A.I. Memo 1687, 2000.
[27] V. Y. F. Tan and C. Févotte, "Automatic relevance determination in nonnegative matrix factorization," in *Proc. Workshop Signal Process. Adaptative Sparse Structured Representations (SPARS)*, St-Malo, France, 2009.
[28] S. D. Babacan, M. Luessi, R. Molina, and A. K. Katsaggelos, "Low-rank matrix completion by variational sparse bayesian learning," in *Proc. Int. Conf. Acoust., Speech. Signal Process. (ICASSP'11)*, Prague, Czech Republic, 2011, pp. 2188–2111.
[29] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence. With application to music analysis," *Neural Comput.*, vol. 21, no. 3, pp. 793–830, Mar. 2009.

Onur Dikmen (M'09) received the B.Sc., M.Sc., and Ph.D. degrees in computer engineering from Bogaziçi University, Istanbul, Turkey.

He worked at Télécom ParisTech, France, as a CNRS Research Associate. He is currently with the Department of Information and Computer Science at Aalto University, Finland. His research interests include statistical signal processing, Bayesian statistics, and approximate inference. He works on Bayesian source modeling and nonnegative matrix factorization for source separation.

Cédric Févotte (M'09) received the State Engineering degree and the Ph.D. degree in control and computer science from École Centrale de Nantes, Nantes, France, in 2000 and 2003, respectively.

As a PhD. student he was with the Signal Processing Group at Institut de Recherche en Communication et Cybernétique de Nantes (IRCCyN) where he worked on time-frequency approaches to blind source separation. From 2003 to 2006, he was a Research Associate with the Signal Processing Laboratory at University of Cambridge (Engineering Deptartment) where he worked on Bayesian approaches to sparse component analysis with applications to audio source separation. He was then a Research Engineer with the start-up company Mist-Technologies (now Audionamix) in Paris, designing mono/stereo to 5.1 surround sound upmix solutions. In Mar. 2007, he joined Télécom ParisTech, first as a Research Associate and then as a CNRS Tenured Research Scientist in November, 2007. His research interests generally concern statistical signal processing and unsupervised machine learning and, in particular, applications to blind source separation and audio signal processing. He is the scientific leader of project TANGERINE (Theory and applications of nonnegative matrix factorization) funded by the French research funding agency ANR.

Dr. Févotte is a Member of the IEEE "Machine learning for signal processing" technical committee.