NONNEGATIVE MATRIX FACTORIZATIONS AS PROBABILISTIC INFERENCE IN COMPOSITE MODELS

Cédric FÉVOTTE[1] *and A. Taylan CEMGIL*[2]

[1]CNRS LTCI; Télécom ParisTech
37-39, rue Dareau
75014 Paris, France
fevotte@telecom-paristech.fr

[2] Department of Computer Engineering,
Boğaziçi University
34342 Bebek, Istanbul, Turkey
taylan.cemgil@boun.edu.tr

ABSTRACT

We develop an interpretation of nonnegative matrix factorization (NMF) methods based on Euclidean distance, Kullback-Leibler and Itakura-Saito divergences in a probabilistic framework. We describe how these factorizations are implicit in a well-defined statistical model of superimposed components, either Gaussian or Poisson distributed, and are equivalent to maximum likelihood estimation of either mean, variance or intensity parameters. By treating the components as hidden-variables, NMF algorithms can be derived in a typical data augmentation setting. This setting can in particular accommodate regularization constraints on the matrix factors through Bayesian priors. We describe multiplicative, Expectation-Maximization, Markov chain Monte Carlo and Variational Bayes algorithms for the NMF problem. This paper describes in a unified framework both new and known algorithms and aims at providing statistical insights to NMF.

1. INTRODUCTION

Given a data matrix $\mathbf{V}$ of dimensions $F \times N$ with nonnegative entries, NMF is the problem of finding a factorization

$$\mathbf{V} \approx \mathbf{WH} = \hat{\mathbf{V}} \tag{1}$$

where $\mathbf{W}$ and $\mathbf{H}$ are nonnegative matrices of dimensions $F \times K$ and $K \times N$, respectively. K is usually chosen such that $FK+KN << FN$, hence $\hat{\mathbf{V}}$ becomes a low-rank matrix with reduced number of parameters. In the following, the entries of matrices $\mathbf{V}$, $\mathbf{W}$, $\mathbf{H}$ and $\hat{\mathbf{V}}$ are denoted v_{fn}, w_{fk}, h_{kn} and $\hat{v}_{fn}$ respectively. We use the colon notation ":" to denote all column or row indices so that $\mathbf{W} = [w_{:,1}, \ldots, w_{:,K}]$ and $\mathbf{H} = [h_{1,:}^\top, \ldots, h_{K,:}^\top]^\top$.

NMF has been applied to diverse problems (such as pattern recognition, clustering, mining, source separation, collaborative filtering) in many areas (such as bioinformatics, audio and image processing, and finance). In the literature, the factorization (1) is usually sought after through the minimization problem

$$\min_{\mathbf{W},\mathbf{H} \geq \mathbf{0}} D(\mathbf{V}|\mathbf{WH}) = D(\mathbf{V}|\hat{\mathbf{V}}) \stackrel{\text{def}}{=} \sum_{f=1}^{F} \sum_{n=1}^{N} d(v_{fn}|\hat{v}_{fn}) \tag{2}$$

where $d(x|y)$ is a scalar cost function. Popular choices are the squared Euclidean distance, the (generalized) Kullback-Leibler (KL) divergence, also referred to as I-divergence and the Itakura-Saito (IS) divergence defined as

$$d_{EUC}(x|y) = \frac{1}{2}(x-y)^2 \tag{3}$$

$$d_{KL}(x|y) = x \log \frac{x}{y} - x + y \tag{4}$$

$$d_{IS}(x|y) = \frac{x}{y} - \log \frac{x}{y} - 1 \tag{5}$$

All cost functions are positive and have a single minimum 0 when $x = y$.

The interpretation of NMF as a low rank matrix approximation in the sense of minimizing a given distance metric d may be sufficient for the derivation of useful signal decomposition algorithms. Certainly, many alternative divergence criteria could also be contemplated [4, 3, 12]. However, for many applications it is not clear which distance metric to take or what the dimension of the latent matrices $\mathbf{W}$ and $\mathbf{H}$ should be. Such *model selection* questions are inherently related to the underlying statistical properties of $\mathbf{V}$ and can be approached in a principled manner via a Bayesian treatment.

We recast NMF with the popular Euclidean, KL and IS costs from a statistical perspective. We show in Section 2 how these factorizations are underlain by a well-defined statistical model of superimposed components, either Gaussian or Poisson distributed, and are equivalent to maximum likelihood estimation of either mean, variance or intensity parameters. By treating the components as hidden-variable we derive NMF algorithms in Section 3, based on Expectation-Maximization (EM), Markov chain Monte Carlo (MCMC) and Variational Bayes (VB). We also review standard multiplicative algorithms and elaborate on the connections between cost functions (3), (4), (5) and Bregman and β divergences [3, 4]. Finally, we discuss in Section 4 the potentials of such probabilistic interpretations of NMF. Parts of the statistical analysis and some of the algorithms presented here have already been published in the literature (see subsequent references); this paper aims at describing these related works in a unified statistical setting.

2. STATISTICAL MODELS

2.1 Observation models

The choice of a certain cost function $d(.|.)$ to measure the fit between v_{fn} and $\hat{v}_{fn}$ implies certain statistical assumptions about how v_{fn} is generated from $\hat{v}_{fn}$. It was already pointed in various papers, e.g, [5, 2] that Euclidean, KL and IS NMF

underlie the following generative models :

$$v_{fn} \sim \mathcal{N}(v_{fn}; \hat{v}_{fn}, \sigma^2) \quad \text{EUC-NMF} \quad (6)$$
$$v_{fn} \sim \mathcal{P}(v_{fn}; \hat{v}_{fn}) \quad \text{KL-NMF} \quad (7)$$
$$v_{fn} \sim \mathcal{G}(v_{fn}; a, a/\hat{v}_{fn}) \quad \text{IS-NMF} \quad (8)$$

where $\mathcal{N}$, $\mathcal{P}$, $\mathcal{G}$ refer to the Gaussian, Poisson and Gamma distribution, respectively, defined in the Appendix and where $\hat{v}_{fn}$ obeys the parametrization $\hat{v}_{fn} = \sum_k w_{fk} h_{kn}$. The likelihood of the parameters $\mathbf{W}$ and $\mathbf{H}$ under the latter models can be mapped to the corresponding cost function (2), so that NMF is actually equivalent to maximum likelihood estimation. In other words, EUC-NMF underlies an additive Gaussian noise, KL-NMF underlies a Poisson noise and IS-NMF underlies a multiplicative Gamma noise.

As matter of fact, all three cost functions belong to the family of regular Bregman divergences, which are in one to one correspondence to families of regular exponential distributions [1]. For scalars, a Bregman divergence is defined with respect to a (differentiable) convex function ϕ as follows (see, e.g, [1, 12])

$$d_\phi(x|y) \quad = \quad \phi(x) - (\phi(y) + \phi'(y)(x-y)).$$

We have the following correspondences $d_{EUC}(x|y) \leftrightarrow \phi(y) = y^2/2$, $d_{KL}(x|y) \leftrightarrow \phi(y) = y\log y - y$, $d_{IS}(x|y) \leftrightarrow \phi(y) = -\log y$. NMF with Bregman divergences has been studied in [4] where various multiplicative algorithms are described.

2.2 Composite models

An interesting property of the Gaussian and Poisson distributions is that they are closed under summation; when $x = \sum_k c_k$ and c_k are Poisson (or Gaussian), x is Poisson (or Gaussian). Conversely, any x can be decomposed as $\sum_k c_k$ without changing the underlying model. In the sequel, we will elaborate on these specific models by further pointing and exploiting their composite structure. We here introduce the following generative model

$$x_{fn} \quad = \quad \sum_k c_{k,fn} \quad (9)$$
$$c_{k,fn} \quad \sim \quad p(c_{k,fn}|\theta_k) \quad (10)$$

where $\theta_k = \{w_{:,k}, h_{k,:}\}$. The next paragraphs describe how Euclidean, KL and IS NMF are equivalent to ML estimation of $\theta = \{\theta_1, \ldots, \theta_K\}$ in specific cases of the latter model, with either $v_{fn} = x_{fn}$ or $v_{fn} = |x_{fn}|^2$. We note $\mathbf{C}_k$ and $\mathbf{X}$ the $F \times N$ matrices with entries $\{c_{k,fn}\}_{fn}$ and $\{x_{fn}\}_{fn}$, respectively. In the sequel we will refer to $\mathbf{C}_k$ as *component*.

NMF with the Euclidean distance (EUC-NMF)

The corresponding generative model is

$$c_{k,fn} \quad \sim \quad \mathcal{N}(c_{k,fn}; w_{fk}h_{kn}, \frac{\sigma^2}{K}) \quad (11)$$

It is easily shown that

$$-\log p(\mathbf{X}|\mathbf{W},\mathbf{H},\sigma^2) = \frac{1}{\sigma^2} D_{EUC}(\mathbf{X}|\mathbf{WH}) + \frac{NF}{2}\log(2\pi\sigma^2)$$

Hence, ML estimation of $\mathbf{W}$ and $\mathbf{H}$ is equivalent to NMF of $\mathbf{V} = \mathbf{X}$ into $\mathbf{WH}$ where the Euclidean distance is used. There is however an interpretability ambiguity with the generative model defined by Eqs. (9), (10), (11) as it may produce negative data. As such, even though the resulting optimization problem is in the end the same provided that available data $\mathbf{X}$ is nonnegative, there is a semantic difference between the two points of view given by EUC-NMF and ML estimation in the Gaussian composite generative model. A more suitable approach, would be to assume the components to be generated from a truncated normal distribution, but this would break the formal correspondence between the two approaches due to the necessary re-normalization of the component distributions.

NMF with the generalized KL divergence (KL-NMF)

Assume the following generative model

$$c_{k,fn} \quad \sim \quad \mathcal{P}(c_{k,fn}; w_{fk}h_{kn}) \quad (12)$$

It is easily shown that

$$-\log p(\mathbf{X}|\mathbf{W},\mathbf{H}) \stackrel{c}{=} D_{KL}(\mathbf{X}|\mathbf{WH})$$

where $\stackrel{c}{=}$ denotes equality up to a constant. Hence, ML estimation of $\mathbf{W}$ and $\mathbf{H}$ is equivalent to NMF of $\mathbf{V} = \mathbf{X}$ into $\mathbf{WH}$ where the KL divergence is used. The data $\mathbf{X}$ produced by the generative model defined by Eqs. (9), (10), (12) is nonnegative, but there is still an interpretability ambiguity with real-valued data, as the Poisson process produces integers.

NMF with the IS divergence (IS-NMF)

Assume the following generative model

$$c_{k,fn} \quad \sim \quad \mathcal{N}_c(c_{k,fn}; 0, w_{fk}h_{kn})$$

The data $\mathbf{X}$ generated from this model is complex (but we could also assume a real Gaussian pdf instead of complex). It is easily shown that [5]

$$-\log p(\mathbf{X}|\mathbf{W},\mathbf{H}) \stackrel{c}{=} D_{IS}(|\mathbf{X}|^{\cdot 2}|\mathbf{WH}),$$

where $|\mathbf{X}|^{\cdot 2}$ is the matrix with entries $|x_{fn}|^2$. Hence, ML estimation of $\mathbf{W}$ and $\mathbf{H}$ is equivalent to NMF of $\mathbf{V} = |\mathbf{X}|^{\cdot 2}$ into $\mathbf{WH}$ where the IS divergence is used. This also corresponds to $a = 1$, i.e, exponential multiplicative noise in Eq. (8).

3. ALGORITHMS

3.1 Multiplicative algorithms

The multiplicative gradient descent approach taken in [8, 3] is akin to updating each parameter by multiplying its value at previous iteration by the ratio of the negative and positive parts of the derivative of the criterion w.r.t this parameter, namely $\theta \leftarrow \theta.[\nabla f(\theta)]_-/[\nabla f(\theta)]_+$, where $\nabla f(\theta) = [\nabla f(\theta)]_+ - [\nabla f(\theta)]_-$ and the summands are both nonnegative. This ensures nonnegativity of the parameter updates, provided initialization with a nonnegative value. A fixed point $\theta^\star$ of the algorithm implies either $\nabla f(\theta^\star) = 0$ or $\theta^\star = 0$. This leads to the following updates,

$$\mathbf{H} \quad \leftarrow \quad \mathbf{H}.\frac{\mathbf{W}^T\left((\mathbf{WH})^{\cdot[\beta-2]}.\mathbf{X}\right)}{\mathbf{W}^T(\mathbf{WH})^{\cdot[\beta-1]}} \quad (13)$$

$$\mathbf{W} \quad \leftarrow \quad \mathbf{W}.\frac{\left((\mathbf{WH})^{\cdot[\beta-2]}.\mathbf{X}\right)\mathbf{H}^T}{(\mathbf{WH})^{\cdot[\beta-1]}\mathbf{H}^T} \quad (14)$$

where $\beta = 2$ corresponds to EUC-NMF, $\beta = 1$ to KL-NMF and $\beta = 0$ to IS-NMF, and '.' and './.' denote entrywise operations. Other values of β correspond to performing NMF with the β-divergence $d_\beta(x|y)$ [3, 5], which is actually the Bregman divergence corresponding to $\phi(y) = \frac{1}{\beta(\beta-1)} y^\beta$, for $\beta \notin \{0,1\}$, and which takes the KL and IS cost as limiting cases when β goes to 1 and 0, respectively.

Lee & Seung [8] showed that criterion (2) is nonincreasing under the latter updates for $\beta = 2$ (Euclidean distance) and $\beta = 1$ (KL divergence) and the proof was extended by Kompass [6] for values $1 \leq \beta \leq 2$, i.e, where $d_\beta(x|y)$ is convex w.r.t y. Solving for the more simple problem $v_{:,n} \approx \mathbf{W} h_{:,n}$ with $\mathbf{W}$ fixed, the proof is simply based on the construction of the functional

$$G(h_{:,n}, \tilde{h}_{:,n}) = \sum_{fk} \lambda_{kfn}\, d(v_{fn} | \frac{w_{fk} h_{kn}}{\lambda_{kfn}}) \quad \text{with} \quad \lambda_{kfn} = \frac{w_{fk}\tilde{h}_{kn}}{[\mathbf{W}\tilde{\mathbf{h}}]_{fn}}$$

which is easily shown to be a suitable auxiliary function for $C(h) = D(v|\mathbf{W}h)$ (i.e, $G(h,h) = C(h)$ and $G(h,\tilde{h}) \geq C(h)$) by convexity of $d(x|y)$ and using Jensen's inequality. A similar auxiliary function can be built to solve for $v_{f,:}^T \approx \mathbf{H}^T w_{f,:}^T$ with $\mathbf{H}$ fixed.

However, the criterion was observed by many authors [4, 3, 5] to be still nonincreasing under updates (13) and (14) for values of β out of the $(1,2)$ interval (and in particular for $\beta = 0$ corresponding to IS divergence), but no proof is available.

Though popularized by Lee & Seung for NMF within the machine learning community in the last decade, the multiplicative updates for each factor in Euclidean and KL NMF corresponds to well-known algorithms for image restoration in the inverse problem community, see [7] and references therein.

3.2 EM algorithms

In Section 2 we have shown how EUC, KL and IS-NMF underlie statistical composite models. The components act as *latent variables* and may be used as complete data in the EM algorithm. In this setting the following functional has to be maximized iteratively

$$Q(\theta|\theta') \stackrel{\text{def}}{=} -\int_{\mathbf{C}} \log p(\mathbf{C}|\theta)\, p(\mathbf{C}|\mathbf{X},\theta')\, d\mathbf{C}.$$

where $\theta = \{\mathbf{W}, \mathbf{H}\}$ and $\mathbf{C}$ is the tensor with slices $\mathbf{C}_k$ and elements $c_{k,fn}$. The convergence of this algorithm to a stationary point is granted. Using conditional independence

$$p(\mathbf{C}|\theta) = \prod_k p(\mathbf{C}_k|\theta_k)$$

the EM functional can be written

$$Q(\theta|\theta') \quad = \quad \sum_k Q_k(\theta_k|\theta'),$$

$$Q_k(\theta_k|\theta') \stackrel{\text{def}}{=} -\int_{\mathbf{C}_k} \log p(\mathbf{C}_k|\theta_k)\, p(\mathbf{C}_k|\mathbf{X},\theta')\, d\mathbf{C}_k. \quad (15)$$

Under suitable i.i.d assumptions the functional is further reduced to

$$Q_k(\theta_k|\theta') = -\sum_{fn} \int_{c_{k,fn}} \log p(c_{k,fn}|\theta_k)\, p(c_{k,fn}|x_{fn},\theta')\, d\, c_{k,fn}. \quad (16)$$

We now explicit the EM algorithm in the specific cases of Euclidean, KL and IS NMF. Note that in the following we are not able to minimize $Q_k(w_{:,k},\, h_{k,:}|\theta')$ jointly in $w_{:,k}$ and $h_{k,:}$, but only to perform coordinate descent, i.e, produce $w_{:,k}^{(i+1)}$ and $h_{k,:}^{(i+1)}$ such that $Q_k(w_{:,k}^{(i+1)}, h_{k,:}^{(i+1)}|\theta^{(i)}) \geq Q_k(w_{:,k}^{(i)}, h_{k,:}^{(i+1)}|\theta^{(i)}) \geq Q_k(w_{:,k}^{(i)}, h_{k,:}^{(i)}|\theta^{(i)})$, which leads strictly speaking to a (converging) generalized EM (GEM) algorithm instead of pure EM. In the following, the apostrophe $'$ will refer to parameter values as of previous iteration (i).

3.2.1 EUC-NMF

$$\begin{aligned} -\log p(c_{k,fn}|\theta_k) &\stackrel{c}{=} \frac{1}{2\sigma^2}(c_{k,fn} - w_{fk}h_{kn})^2 \\ p(c_{k,fn}|x_{fn},\theta) &= \mathcal{N}(c_{k,fn}|\mu_{k,fn}^{post}, \lambda_{k,fn}^{post}) \end{aligned}$$

with

$$\mu_{k,fn}^{post} = w_{fk}h_{kn} + \frac{1}{K}(x_{fn} - \hat{x}_{fn}), \quad \lambda_{k,fn}^{post} = \frac{K-1}{K^2}\sigma^2 \quad (17)$$

where here $\hat{x}_{fn} = \hat{v}_{fn} = \sum_k w_{fk}h_{kn}$. Hence, the minimization of functional (16) subject to nonnegative constraints leads to

$$h_{kn} = \left\lfloor \frac{\sum_f w_{fk}\left(\frac{1}{K}(x_{fn} - \hat{x}'_{fn}) + w'_{fk}h'_{kn}\right)}{\sum_f w_{fk}^2} \right\rfloor_+ \quad (18)$$

$$w_{fk} = \left\lfloor \frac{\sum_n h_{kn}\left(\frac{1}{K}(x_{fn} - \hat{x}'_{fn}) + w'_{fk}h'_{kn}\right)}{\sum_n h_{kn}^2} \right\rfloor_+ \quad (19)$$

where $\lfloor x \rfloor_+ = \max\{x, 0\}$. These update equations differ from the usual multiplicative updates given from Eq. (13) and (14).

3.2.2 KL-NMF

$$\begin{aligned} -\log p(c_{k,fn}|\theta_k) &\stackrel{c}{=} -w_{fk}h_{kn} + c_{k,fn}\log(w_{fk}h_{kn}) \\ p(c_{k,fn}|x_{fn},\theta) &= \mathscr{B}\left(c_{k,fn}|v_{fn}, \pi_{k,fn}\right) \end{aligned}$$

where $\pi_{k,fn} = w_{fk}h_{kn}/\hat{x}_{fn}$ and here $\hat{x}_{fn} = \hat{v}_{fn} = \sum_k w_{fk}h_{kn}$. This leads to

$$h_{kn} = h'_{kn}\frac{\sum_f w'_{fk}\left(\frac{x_{fn}}{\hat{x}'_{fn}}\right)}{\sum_k w_{fk}}, \quad w_{fk} = w'_{fk}\frac{\sum_n h'_{kn}\left(\frac{x_{fn}}{\hat{x}'_{fn}}\right)}{\sum_n h_{kn}} \quad (20)$$

which coincides with the usual multiplicative updates given by Eq. (13) and (14).

3.2.3 IS-NMF

$$\begin{aligned} -\log p(c_{k,fn}|\theta_k) &\stackrel{c}{=} \log(w_{fk}h_{kn}) + \frac{|c_{k,fn}|^2}{w_{fk}h_{kn}} \\ p(c_{k,fn}|x_{fn},\theta) &= \mathcal{N}(c_{k,fn}|\mu_{k,fn}^{post}, \lambda_{k,fn}^{post}) \end{aligned}$$

with

$$\mu_{k,fn}^{post} = \frac{w_{fk}h_{kn}}{\sum_l w_{fl}h_{ln}}x_{fn}, \quad \lambda_{k,fn}^{post} = \frac{w_{fk}h_{kn}}{\sum_l w_{fl}h_{ln}}\sum_{l\neq k} w_{fl}h_{ln}. \quad (21)$$

Leading to

$$h_{kn} = \frac{1}{F}\sum_f \frac{v'_{k,fn}}{w_{fk}}, \quad w_{fk} \quad = \quad \frac{1}{N}\sum_n \frac{v'_{k,fn}}{h_{kn}}, \qquad (22)$$

with $v'_{k,fn} = |\mu^{post\prime}_{k,fn}|^2 + \lambda^{post\prime}_{k,fn}$. These update equations differ from the multiplicative updates given from Eq. (13) and (14), and are equivalent to the SAGE algorithm described in [5].

3.2.4 Bayesian maximum a posteriori

It is interesting to note that the EM framework readily accommodates Bayesian approaches for which prior information about the parameters $\mathbf{W}$ and $\mathbf{H}$ is available in the form of prior distributions $p(\mathbf{H})$ and $p(\mathbf{W})$. The complete data likelihood term $-\log p(\mathbf{C}_k|\theta_k)$ needs only be changed by $-\log p(\theta_k|\mathbf{C}_k)$ in Eq. (15), leading to the following functional to be maximized

$$Q_k^{MAP}(\theta_k|\theta') = Q_k(\theta_k|\theta') - \log p(w_{:,k}) - \log p(h_{k,:})$$

so that only the M-step is changed.

3.3 MCMC algorithms

Monte Carlo methods [9] are powerful computational techniques to estimate expectations of form

$$E = \langle \psi(\theta) \rangle_{p(\theta)} \approx \frac{1}{L}\sum_{i=1}^{L} \psi(\theta^{(i)}) = \tilde{E}_L$$

where $\theta^{(i)}$ are samples drawn from $p(\theta)$. Under mild conditions on the test function ψ, the estimate $\tilde{E}_L$ converges to the true expectation for $L \to \infty$. The difficulty here is obtaining independent samples $\{\theta^{(i)}\}_{i=1\ldots L}$ from complicated distributions. MCMC techniques generate subsequent samples from a Markov chain. One particularly convenient and simple procedure is the Gibbs sampler where one samples each block of variables from *full conditional distributions*. In the Bayesian setting for the NMF model, a possible Gibbs sampler is

$\mathbf{C}^{(i)} \sim p(\mathbf{C}|\mathbf{W}^{(i-1)}, \mathbf{H}^{(i-1)}, \mathbf{X})$
for $k = 1 : K$ **do**
  $h^{(i)}_{k,:} \sim p(h_{k,:}|\mathbf{C}^{(i)}_k, w^{(i-1)}_{:,k})$
  $w^{(i)}_{:,k} \sim p(w_{:,k}|\mathbf{C}^{(i)}_k, h^{(i)}_{k,:})$
end for

Denoting $\mathbf{c}_{fn} = [c_{1,fn}, \ldots, c_{K,fn}]^T$, the posterior of the hidden components writes

$$p(\mathbf{C}|\mathbf{W}, \mathbf{H}, \mathbf{X}) = \prod_{fn} p(\mathbf{c}_{fn}|w_{f,:}, h_{:,n}, x_{fn})$$

Next, we derive the full conditionals for the three considered models.

3.3.1 EUC-NMF

The posterior of $\mathbf{c}_{fn}$ is given by

$$p(\mathbf{c}_{fn}|w_{f,:}, h_{:,n}, x_{fn}) \quad = \quad \mathcal{N}(\mathbf{c}_{fn}|\mu^{post}_{fn}, \Sigma^{post}_{fn})$$

with $\mu^{post}_{fn} = [\mu^{post}_{1,fn} \ldots \mu^{post}_{K,fn}]^T$, where $\mu^{post}_{k,fn}$ is defined in Eq. (17), and $\Sigma^{post}_{fn} = \frac{\sigma^2}{K}(\mathbf{I}_K - \frac{1}{K}\mathbf{e}_K\mathbf{e}_K^T)$. The diagonal terms correspond to the posterior variance in Eq. (17). In the unconstrained case conjugate priors for $h_{:,n}$ and $w_{f,:}$ would be Gaussian. However, more sophisticated sampling schemes are required to enforce nonnegativity, typically by using Gamma priors, see, e.g, [10, 11].

3.3.2 KL-NMF

The full conditional of $\mathbf{c}_{fn}$ is given by

$$p(\mathbf{c}_{fn}|w_{f,:}, h_{:,n}, x_{fn}) \quad = \quad \mathcal{M}(\mathbf{c}_{fn}|x_{fn}, \pi_{fn})$$

where $\mathcal{M}$ refers to the multinomial distribution defined in Appendix and $\pi_{fn} = [\pi_{1,fn}, \ldots, \pi_{K,fn}]$, with $\pi_{k,fn} = w_{fk}h_{kn}/x_{fn}$, as defined in Section 3.2.2. Using conjugate priors

$$\begin{aligned} p(w_{fk}) &= \mathcal{G}(w_{fk}|\alpha_w, \beta_w), \\ p(h_{kn}) &= \mathcal{G}(h_{kn}|\alpha_h, \beta_h), \end{aligned}$$

the full conditionals can be derived as [2]

$$\begin{aligned} p(w_{fk}|\mathbf{C}_k, h_{k,:}) &= \mathcal{G}(w_{fk}|\alpha_w + \sum_n c_{k,fn}, \alpha_w\beta_w + \sum_n h_{kn}) \\ p(h_{kn}|\mathbf{C}_k, w_{:,k}) &= \mathcal{G}(h_{kn}|\alpha_h + \sum_f c_{k,fn}, \alpha_h\beta_h + \sum_f w_{fk}) \end{aligned}$$

3.3.3 IS-NMF

Denoting $\lambda_{fn} = [w_{f1}h_{1n} \ldots w_{fK}h_{Kn}]^T$, the posterior of $\mathbf{c}_{fn}$ is given by

$$p(\mathbf{c}_{fn}|w_{f,:}, h_{:,n}, x_{fn}) \quad = \quad \mathcal{N}(\mathbf{c}_{fn}|\mu^{post}_{fn}, \Sigma^{post}_{fn})$$

with $\mu^{post}_{fn} = [\mu^{post}_{1,fn} \ldots \mu^{post}_{K,fn}]^T$, where $\mu^{post}_{k,fn}$ is defined in Eq. (21), and $\Sigma^{post}_{fn} = \mathrm{diag}\left(\lambda_{fn}\right) - \frac{1}{\hat{v}_{fn}}\lambda_{fn}\lambda^T_{fn}$. The diagonal terms correspond to the posterior variance in Eq. (21). Using conjugate inverse-Gamma priors

$$\begin{aligned} p(h_{kn}) &= \mathcal{IG}(h_{kn}|\alpha_h, \beta_h), \\ p(w_{fk}) &= \mathcal{IG}(w_{fk}|\alpha_w, \beta_w), \end{aligned}$$

the full conditionals of $h_{k,:}$ and $w_{:,k}$ write

$$\begin{aligned} p(w_{fk}|\mathbf{C}_k, h_{k,:}) &= \mathcal{IG}(w_{fk}|\alpha_w + N, \beta_w + \sum_n |c_{k,fn}|^2/h_{kn}) \\ p(h_{kn}|\mathbf{C}_k, w_{:,k}) &= \mathcal{IG}(h_{kn}|\alpha_h + F, \beta_h + \sum_f |c_{k,fn}|^2/w_{fk}) \end{aligned}$$

3.4 Variational Bayes

We finally describe how the composite structure of Euclidean, KL and IS NMF can be exploited to derive a variational Bayes algorithm [13]. The idea is to bound the marginal likelihood from below

$$\begin{aligned} \mathscr{L}_{\mathbf{X}}(\vartheta) &\equiv \log p(\mathbf{X}|\vartheta) \geq \mathscr{B}_{VB}[q] \\ &\equiv \int q \log \frac{p(\mathbf{X}, \mathbf{C}, \mathbf{W}, \mathbf{H}|\vartheta)}{q} d(\mathbf{C}, \mathbf{W}, \mathbf{H}) \\ &= \langle \log p(\mathbf{X}, \mathbf{C}, \mathbf{H}, \mathbf{W}|\vartheta) \rangle_q + H[q] \end{aligned}$$

where ϑ denotes the hyperparameters and q is defined as

$$q \quad = \quad \left(\prod_{fn} q(\mathbf{c}_{fn})\right)\left(\prod_{fk} q(w_{fk})\right)\left(\prod_{kn} q(h_{kn})\right) \equiv \prod_{\alpha \in \mathscr{C}} q_\alpha$$

The integral over $\mathbf{C}$ will be a summation when $\mathbf{C}$ are discrete (i.e, Poisson component in the KL case). Here, $\alpha \in \mathscr{C} = \{\mathbf{C}, \mathbf{W}, \mathbf{H}\}$ denotes the set of disjoint clusters of variables. A local optimum can be attained by the following fixed point iteration:

$$q_\alpha^{(i+1)} \quad \propto \quad \exp\left(\langle \log p(\mathbf{X}, \mathbf{C}, \mathbf{W}, \mathbf{H}|\vartheta)\rangle_{q_{\neg\alpha}^{(i)}}\right)$$

where $q_{\neg\alpha} = q/q_\alpha$. The expectations of $\langle \log p(\mathbf{X}, \mathbf{C}, \mathbf{W}, \mathbf{H}|\vartheta)\rangle$ are functions of the sufficient statistics of q. It turns out that the variational update equations have very similar forms to the full conditionals derived for the Gibbs sampler. Here, due to lack of space we only give the equations for the KL case:

$$q(\mathbf{c}_{fn}) \quad = \quad \mathscr{M}(\mathbf{c}_{fn}|x_{fn}, \pi_{fn})$$

where $\pi_{fn} = [\pi_{1,fn}, \ldots, \ldots \pi_{K,fn}]$ and $\langle c_{k,fn}\rangle = x_{fn}\pi_{k,fn}$ with

$$\pi_{k,fn} \quad \equiv \quad \frac{\exp(\langle \log w_{fk}\rangle + \langle \log h_{kn}\rangle)}{\sum_k \exp(\langle \log w_{fk}\rangle + \langle \log h_{kn}\rangle)}$$

The full conditionals can be derived as [2]

$$q(w_{fk}) \quad = \quad \mathscr{G}(w_{fk}|\alpha_w + \sum_n \langle c_{k,fn}\rangle, \alpha_w\beta_w + \sum_n \langle h_{kn}\rangle)$$
$$q(h_{kn}) \quad = \quad \mathscr{G}(h_{kn}|\alpha_h + \sum_f \langle c_{k,fn}\rangle, \alpha_h\beta_h + \sum_f \langle w_{fk}\rangle)$$

One attractive feature of VB is that the hyperparameters can be optimized by maximizing the variational bound $\mathscr{B}_{VB}[q]$. While this does not guarantee to increase the true marginal likelihood, it leads in this application to algorithms that enables one to do full Bayesian model selection a lot more faster than MCMC based sampling approaches where calculation of the marginal likelihood is trickier. For a detailed discussion see [2].

4. DISCUSSION

In this overview paper, we have discussed the probabilistic interpretation of various NMF models in maximum likelihood, MAP and full Bayesian setting. In all the algorithms we discuss, we are exploiting the closure under summation property of the observation model and the closed form availability of all the full conditionals. It should be noted that this is not the case for all divergence measures. In other cases other optimization techniques need to be employed.

Prior structures are needed to control the decompositions for exploratory data analysis or various problems in signal processing. There is an emphasis on optimization strategies for maximum likelihood or MAP estimation in NMF models but less research on efficient Bayesian integration methods (with a few exceptions such as [14, 2, 11]). Moreover, as the number of alternatives for data modelling increases (for example consider the number of factorization options with increasing data dimension in tensor factorization) there is a need to do model order selection and model averaging in a principled manner for which ML approaches are known to be inappropriate. Due to lack of space, we are not giving in this paper simulation results with the developed algorithms but refer the reader to other work, such as [2, 5]. A detailed and exhaustive comparison of the algorithms in terms of effectiveness for various signal decomposition is a natural next step and is currently under progress.

A. STANDARD DISTRIBUTIONS

Multivariate Gaussian, with $c = 1/2$ or 1 (real/complex case)
$\mathscr{N}(\mathbf{x}|\mu, \Sigma) = |\pi\Sigma/c|^{-c} \exp -c(\mathbf{x}-\mu)^T \Sigma^{-1}(\mathbf{x}-\mu)$
Poisson $\mathscr{P}(x|\lambda) = \exp(-\lambda)\frac{\lambda^x}{x!}$
Binomial $\mathscr{B}(x|n,p) = \binom{n}{x} p^x (1-p)^{n-x}$
Multinomial
$\mathscr{M}(\mathbf{c}|n,\mathbf{p}) = \binom{n}{c_1\, c_2 \ldots c_K} p_1^{c_1} p_2^{c_2} \cdots p_K^{c_K}\, \delta(n - \sum_k c_k)$
Gamma $\mathscr{G}(u|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} u^{\alpha-1} \exp(-\beta u),\ u \geq 0$
inv.-Gamma $\mathscr{IG}(u|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} u^{-(\alpha+1)} \exp(-\frac{\beta}{u}),\ u \geq 0$

Acknowledgements

We wish to thank the reviewers for many very helpful comments and suggestions, as well as O. Cappé for discussions related to this work.

REFERENCES

[1] A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh. Clustering with Bregman divergences. *Journal of Machine Learning Research*, 6:1705–1749, 2005.

[2] A. T. Cemgil. Bayesian inference in non-negative matrix factorisation models. Technical Report CUED/F-INFENG/TR.609, University of Cambridge, July 2008. Accepted for publication in Computational Intelligence and Neuroscience.

[3] A. Cichocki, R. Zdunek, and S. Amari. Csiszar's divergences for non-negative matrix factorization: Family of new algorithms. In *Proc. 6th International Conference on Independent Component Analysis and Blind Signal Separation (ICA'06)*, pages 32–39, Charleston SC, USA, Mar. 2006.

[4] I. S. Dhillon and S. Sra. Generalized nonnegative matrix approximations with Bregman divergences. *Advances in Neural Information Processing Systems (NIPS)*, 19, 2005.

[5] C. Févotte, N. Bertin, and J.-L. Durrieu. Nonnegative matrix factorization with the Itakura-Saito divergence. With application to music analysis. *Neural Computation*, 21(3), Mar. 2009.

[6] R. Kompass. A generalized divergence measure fon nonnegative matrix factorization. *Neural Computation*, 19(3):780–791, 2007.

[7] H. Lantéri, M. Roche, O. Cuevas, and C. Aime. A general method to devise maximum-likelihood signal restoration multiplicative algorithms with non-negativity constraints. *Signal Processing*, 81(5):945–974, May 2001.

[8] D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In *Advances in Neural and Information Processing Systems 13*, pages 556–562, 2001.

[9] J. S. Liu. *Monte Carlo strategies in scientific computing*. Springer, 2002.

[10] S. Moussaoui, D. Brie, A. Mohammad-Djafari, and C. Carteret. Separation of non-negative mixture of non-negative sources using a Bayesian approach and mcmc sampling. *IEEE Trans. on Signal Processing*, 54(11):4133–4145, Nov. 2006.

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

[12] A. P. Singh and G. J. Gordon. A unified view of matrix factorization models. In *Proc. European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD 2008), Part II*, number 5212 in LNAI, pages 358–373. Springer, 2008.

[13] M. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning,*, 1:1–305, 2008.

[14] O. Winther and K. B. Petersen. Bayesian independent component analysis: Variational methods and non-negative decompositions. *Digital Signal Processing*, 17(5):858–872, Sep. 2007.