ONLINE MAXIMUM-LIKELIHOOD ESTIMATION FOR LATENT FACTOR MODELS

David Rohde, Olivier Cappé

LTCI, Telecom ParisTech & CNRS
46 rue Barrault, 75013 Paris, France

ABSTRACT

The online EM algorithm is a fast variant of the EM algorithm suitable for processing large streams of data. However the online EM algorithm is restricted to models in which an analytical expectation can be computed for the E-step. In this paper, we show that a new algorithm called the simulated online EM algorithm may be applied to a broad class of models used in signal processing and machine learning. These models, which are characterized by the presence of latent (or unobserved) positive factors, include in particular probabilistic variants of Non-Negative Matrix Factorization (NMF). We provide the main convergence properties of the simulated online EM algorithm and detail its application to the Latent Dirichlet Allocation (LDA) model.

1. INTRODUCTION

The Expectation Maximization (EM) algorithm is far and away the most popular method for applying maximum likelihood methods to latent data models and has many desirable properties including versatility, easiness of implementation and convergence to the maximum likelihood estimate. The online EM algorithm is a variant of the EM algorithm suitable for online estimation and fast application to large datasets. The online EM algorithm inherits most of the good properties of its batch counterpart in particular is easy to implement and converges to the maximum likelihood solution [1]. The critical difference between the online EM algorithm compared to its batch counterpart is that only partial E-steps are computed which allows the algorithm proceed quickly to the M-step resulting in faster convergence with lower memory requirements.

The online EM algorithm can be applied to complete-data exponential family models in which it is possible to analytically compute the E-step. In this study we introduce the simulated online EM algorithm which allows us to consider the much wider family of models which are also in complete data exponential family but includes models for which the E-step cannot be computed analytically but can be approximated with simulations.

In Section 2 we discuss a broad class of models that are in complete-data exponential family form but for which it is not possible to analytically compute the E-step. In Section 3 we introduce the simulated online EM algorithm and show that, like the online EM algorithm, it converges to the maximum likelihood solution albeit with a higher variance. In Section 4 we consider how the simulated online EM algorithm can be applied to the Latent Dirichlet Allocation model [2] and outline the steps for the simulated online EM algorithm and consider two possible Markov chain Monte Carlo (MCMC) algorithms for approximately computing the required expectations. The use of the simulated online EM algorithm applied to the LDA model is demonstrated on a benchmark image decomposition task.

2. LATENT FACTOR MODELS

The models we are interested in are of the form

$$Y|H \sim g_{\sum_{k=1}^{K} \theta_k H_k} ,$$

where $\{g_\lambda\}_{\lambda\in\Lambda}$ is an exponential family of distributions, H is a latent random vector of real weights (or loadings, or proportions) with a known prior distribution and $\{\theta_k\}_{1\leq k\leq K}$ is a set of common shared factors. Our focus is on estimating these factors in the maximum-likelihood (or MAP) sense.

Such models are known under very different names such as Bayesian exponential Family PCA [3] but also as probabilistic matrix factorization [4], discrete component analysis [5], Bayesian partial membership [6] or simplicial mixture models depending on the context.

A first example is the *Latent Dirichlet Allocation (LDA)* model where $\{g_\lambda\}_{\lambda\in\Lambda}$ are multinomial distributions (assuming a finite vocabulary size), θ_k correspond to vectors of word occurrence probabilities and the proportions H lie in the probability simplex (i.e., are normalized) and are given a Dirichlet prior distribution. This model has been extensively used to model text documents Y represented as bags of words [2].

A second example is the *Itakura-Saito Non-negative Matrix Factorization (NMF)* of [7] where $\{g_\lambda\}_{\lambda\in\Lambda}$ corresponds to a product of independent exponential distributions, θ_k correspond to positive prototype spectra and the mixture weights H are also positive and given independent inverse Gamma priors. This model correspond to a form of NMF which is more meaningful from a statistical point of view when dealing with power spectra computed from times series. The approach has proven to be very successful in audio applications.

Note that we can view usual finite mixtures as a special case of the above where the weights H are constrained to be vectors of zeros that contain a single one (this is the basis of the "mixed-membership" interpretation). In the following however we only consider the more challenging case where the weights are real.

Except in simple cases, the resulting complete-data model $p_\theta(Y, H)$ is not in exponential family form and, furthermore, due to the continuous nature of H, computation of $\mathrm{E}_\theta(H|Y)$ or direct simulation from $p_\theta(w|y)$ is not feasible. However, it is possible to introduce further intermediate latent variables Z such that

(i) $p_\theta(Y, Z, H)$ is in exponential family form,

(ii) There exist efficient simulation schemes to produce approximate draws under $p_\theta(z, h|y)$.

The exact nature of the Z variable depends on the considered model and will be examined in detail in Section 4 for LDA. In the following, we assume that both (i)–(ii) hold and ignore the error resulting from approximate simulation, e.g. the bias introduced by the use of MCMC sampling with a finite burn-in period.

For the models of interest here, the only currently available option for online parameter estimation is to resort to online variational Bayes procedures [8, 9]. Specific implementations of these ideas for LDA have been given very recently in [10, 11]. Although a complete comparison of the proposed simulated online EM algorithm with the existing variational Bayes approaches is beyond the scope of the current paper, we note that they have very similar computational complexities: the variational Bayes algorithms require iterating fixed point equations to determine the variational approximation whereas the simulated online EM algorithm uses short sequences of MCMC simulations to approximate the E-step. However, we will see below that the simulated online EM algorithm eventually converges to the same points as the exact maximum-likelihood procedure, which cannot be guaranteed for the variational methods due to the non-vanishing bias caused by the use of the variational approximation.

3. SIMULATED ONLINE EM ALGORITHM

In this section, Y denotes the observation while $X = (Z, H)$ correspond to the latent data. We assume that X is defined in such a way that the complete-data likelihood is in exponential form

$$p_\theta(x, y) \propto \exp\left(\phi'(\theta)s(x, y) - A(\theta)\right), \tag{1}$$

with observed-data likelihood $f_\theta(y) = \int p_\theta(x, y)dx$. Here, $\phi(\cdot)$ is a function that maps θ to natural parameters, $s(x, y)$ is the sufficient statistic, $A(\cdot)$ is the log partition function and the prime denotes transposition.

We focus on the case where the model is well-specified, in the sense that there exists a true parameter value $\theta_\star$ such that the observations $(Y_n)_{n\geq 1}$ are generated independently under $f_{\theta_\star}$. Define

$$I_f(\theta) = -\mathrm{E}_\theta\left[\nabla^2_\theta \log f_\theta(Y_1)\right],$$
$$I_p(\theta) = -\mathrm{E}_\theta\left[\nabla^2_\theta \log p_\theta(X_1, Y_1)\right],$$

which correspond, respectively, to the actual (observed) and complete-data Fisher information matrices for the parameter θ.

3.1. Online EM

We briefly recall the main facts regarding the online EM algorithm for independent observations described in [1]. The algorithm operates by

$$S_n = (1 - \gamma_n)S_{n-1} + \gamma_n \mathrm{E}_{\bar\theta(S_{n-1})}\left[s(X_n, Y_n)|Y_n\right],$$
$$\theta_n = \bar\theta(S_n), \tag{2}$$

where γ_n denotes a sequence of positive step sizes decaying to zero (typically of the form $\gamma_n = n^{-\alpha}$, with $1/2 < \alpha < 1$) and $\bar\theta$ is the solution of the maximum-likelihood equation such that $\bar\theta(S) = \arg\max_\theta\{\phi'(\theta)s(x, y) - A(\theta)\}$.

The recursion on the statistic S_n admits $D(f_{\theta_\star}\|f_{\bar\theta(s)})$ as Lyapunov function, where D denotes the Kullback-Leibler divergence, which is known to correspond to the large-sample limit of the maximum-likelihood criterion. On the other hand, the recursion on θ_n is asymptotically equivalent to the Robins-Monro weighted gradient algorithm[1]

$$\theta_n = \theta_{n-1} + \gamma_n I_p^{-1}(\theta_\star)\nabla_\theta \log f_{\theta_{n-1}}(Y_n),$$

which shows, that $\gamma_n^{-1/2}(\theta_n - \theta_\star) \Rightarrow \mathcal{N}(0, I_p^{-1}(\theta_\star)/2)$ (where the symbol $\Rightarrow$ denotes convergence in distribution).

3.2. Simulated Online EM

By analogy with (2), we define the simulated online EM algorithm by

$$S_n = (1 - \gamma_n)S_{n-1} + \gamma_n \frac{1}{m}\sum_{i=1}^m s(\tilde X_n^i, Y_n),$$
$$\theta_n = \bar\theta(S_n), \tag{3}$$

where $\tilde X_n^1, \ldots, \tilde X_n^m$ are independent draws under $p_{\bar\theta(S_{n-1})}(x_n|Y_n)$.

In Section 4 we will require two slight modifications of (3). First, in some cases it is more appropriate to draw only a subcomponent $\tilde H_n^i$ of $\tilde X_n^i$. In this case, we make use of Rao-Blackwellized updates replacing $s(\tilde X_n^i, Y_n)$ by its conditional expectation

$$\mathrm{E}_{\bar\theta(S_{n-1})}\left[s(X_n, Y_n)\left|Y_n, \tilde H_n^i\right.\right].$$

This modification does not change the convergence behavior of the algorithm and can only reduce the variance due to Rao-Blackwell theorem.

Next, for some models it is more appropriate to use MAP estimation so as to allow the specification of a prior on θ. If we select the prior to be in the conjugate family corresponding to the complete-data likelihood (1), it is easily shown that the function $\bar\theta$ should be replaced by $\bar\theta_{\text{MAP}}$ such that

$$\bar\theta_{\text{MAP}}(S_n) = \bar\theta(S_n + \beta/n),$$

where β is the hyperparameter of the prior (see also [12]).

We now state (without proof for reason of space) the main convergence properties of the simulated online EM algorithm.

Proposition 1 *Under the assumptions of [1],*

1. *The simulated online EM recursion (3) on s_n admits the same Lyapunov function $D(f_{\theta_\star}\|f_{\bar\theta(s)})$.*
2. *The recursion on θ_n is asymptotically equivalent to*

$$\theta_n = \theta_{n-1} + \gamma_n I_p^{-1}(\theta_\star)\frac{1}{m}\sum_{i=1}^m \nabla_\theta \log p_{\theta_{n-1}}(\tilde X_n^i, Y_n).$$

3. *The rate of convergence is given by*

$$\gamma_n^{-1/2}(\theta_n - \theta_\star) \Rightarrow \mathcal{N}\left(0, \left\{I_p(\theta_\star)^{-1} + \frac{1}{m}(I_f(\theta_\star)^{-1} - I_p(\theta_\star)^{-1})\right\}/2\right). \tag{4}$$

Note that as $I_p(\theta_\star) \succeq I_f(\theta_\star)$, the asymptotic variance of the recursion is monotonically decreasing from $I_f(\theta_\star)^{-1}/2$, for $m = 1$, to $I_p(\theta_\star)^{-1}/2$, when m increases. In particular, in models where the so-called *fraction of missing information* $I_p(\theta_\star) - I_f(\theta_\star)$ is not too large there is no reason to use large values of the number m of simulated replicas. Note that as we will use in practice MCMC simulations to draw the $\tilde X_n^i$'s, these simulation will be correlated and it is safer to average from slightly longer sections of the chain than is suggested by (4).

[1] This result is obtained assuming convergence to $\theta_\star$, as the gradient $\nabla_\theta D(f_{\theta_\star}\|f_\theta)$ of the limiting criterion may vanish in points other than the true parameter. Proposition 6 of [1] features the weight matrix $J(\theta) = -\mathrm{E}_{\theta_\star}\left[\mathrm{E}_\theta\left(\nabla^2_\theta \log p_\theta(X_1, Y_1)|Y_1\right)\right]$, however for-well specified models this converge to $I_p(\theta_\star)$. In the terminology of [10], the online EM algorithm is a *natural gradient* algorithm but note that the weight is given by the complete-data information matrix $I_p(\theta_\star)$ rather than by its observed counterpart $I_f(\theta_\star)$.

4. APPLICATION TO LATENT DIRICHLET ALLOCATION

4.1. Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is a probabilistic model for categorical count data which has been historically used mostly for modelling text documents. There are two equivalent representations of LDA which arise from different choices of latent data; the corresponding Bayesian networks are depicted in Fig 1.

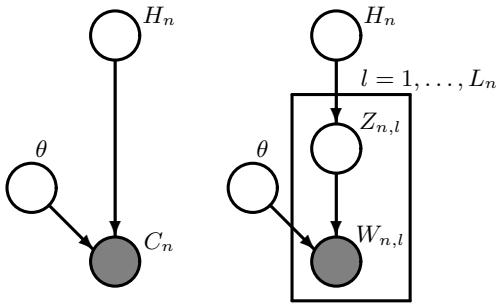


Fig. 1. Bayesian network representations of LDA. Left: minimal completion; right: exponential family completion.

In the first representation, the nth document is represented by the counts $C_{n,1}, \ldots, C_{n,V}$ of all of the possible V words which are assumed to follow a multinomial distribution with probabilities $\sum_{k=1}^{K} \theta_{vk} H_{n,k}$ given H_n. The complete-data likelihood of a document is

$$p_\theta(C_n, H_n) = \frac{L_n!}{\prod_{u=1}^{V} C_{n,u}!} \prod_{v=1}^{V} \left(\sum_{k=1}^{K} \theta_{vk} H_{n,k} \right)^{C_{n,v}} p(H_n), \tag{5}$$

where L_n denotes the length of the document. H_n is then endowed with an exchangeable $\mathrm{Dirichlet}(H_n|\alpha)$ prior.

The second representation enumerates every instances of a word and its topic individually i.e. the lth word instance of the nth document is $W_{n,l}$ and $Z_{n,l}$ is the corresponding latent topic assignment. The LDA model is then[2]

$$p_\theta(W_n, Z_n, H_n) = \prod_{l=1}^{L_n} p_\theta(W_{n,l}|Z_{n,l}) p(Z_{n,l}|H_n) p(H_n), \tag{6}$$

where $p_\theta(W_{n,l} = v|Z_{n,l} = k) = \theta_{vk}$ and $p(Z_{n,l} = k|H_n) = H_{n,k}$. Although involving many more latent variables, this second representation is equivalent to the first one as the observed likelihood $f_\theta(W_n)$ resulting from (6) is equal to that of $p_\theta(C_n)$ in (5), up to the combinatorial factor that measures the number of different word occurrences W_n that correspond to a given count vector H_n. One important use for this representation is for *collapsed Gibbs sampling* [13] which works by sampling the latent data $Z_{n,l}$ (see below). This second representation is in exponential form for θ and the complete-data sufficient statistic can be thought of as a two dimensional histogram $s_{vk}(W_n, Z_n) = \sum_{l=1}^{L_n} \mathbf{1}\{W_{n,l} = v\}\mathbf{1}\{Z_{n,l} = k\}$. Hence, we use this second representation to determine the M-step of the simulated online EM algorithm.

[2]Readers who are familiar with LDA should be made aware that the notation adopted here differs from that of [2, 13]. To convert our notation to that in [2], substitute β for θ and θ for H.

4.2. MCMC Methods for LDA

Collapsed Gibbs sampling, which is the preferred MCMC method for LDA in the literature, makes use of the availability of analytical marginalization for H_n and samples in turn each of the indicator variables $Z_{n,l}$ for $l = 1, \ldots, L_n$. The probability of an indicator variable $Z_{n,l}$ conditional on all the words in the document and all the other indicator variables becomes the following (discrete) probability distribution:

$$p_\theta\left(Z_{n,l}|W_n, \{Z_{n,l'}\}_{l' \neq l}\right) \propto \theta_{W_{n,l} Z_{n,l}} (S^{(-l)}_{Z_{n,l}} + \alpha),$$

where $S^{(-l)}_k = \sum_{l' \neq l} \mathbf{1}\{Z_{n,l'} = k\}$. In words, the probability is proportional to the relevant entry of θ multiplied by the count of words in topic $Z_{n,l}$ within the document, adding the hyperparameter α of the prior Dirichlet distribution on H_n.

In the classical context of Bayesian batch parameter inference, the collapsed Gibbs sampler is also favored for its ability to marginalize with respect to the unknown parameter θ. For online estimation however, θ is fixed to the value θ_{n-1} estimated when processing the previous observation and one only requires conditional simulation of the latent variables pertaining to the new observation Y_n. Hence, working with the full collection of indicator variables Z_n can be avoided by resorting to a direct Metropolis-Hastings algorithm on H_n using only the minimal form of the LDA model (left graph in Figure 1). Given the domain constraint on H_n, the Dirichlet distribution is used as a proposal with the following form

$$H^*_n|H_n \sim \mathrm{Dirichlet}(H^*_n|\kappa H_n + 1).$$

Here κ is a constant; $1/\kappa$ being interpreted as a 'step size'. This distribution has an expectation of $(\kappa H_{n,k} + 1)/(\kappa + K)$, which for large κ is approximately H_n. Unlike the most common random-walk Metropolis-Hastings algorithm this proposal distribution is not symmetric and it is necessary to use the full Metropolis-Hastings test for acceptance or rejection, i.e., the probability of acceptance is given by:

$$\min\left(1, \frac{\prod_v^V (\sum_{k=1}^K \theta_{vk} H^*_{n,k})^{C_{n,v}} \mathrm{Dirichlet}(H_n|\kappa H^*_n + 1)}{\prod_v^V (\sum_{k=1}^K \theta_{vk} H_{n,k})^{C_{n,v}} \mathrm{Dirichlet}(H^*_n|\kappa H_n + 1)}\right),$$

where

$$\frac{\mathrm{Dirichlet}(H_n|\kappa H^*_n + 1)}{\mathrm{Dirichlet}(H^*_n|\kappa H_n + 1)} = \frac{(\prod_k \Gamma(\kappa H_{n,k} + 1)) \prod_k {H^*_{n,k}}^{\kappa H^*_{n,k}}}{(\prod_k \Gamma(\kappa H^*_{n,k} + 1)) \prod_k {H^*_{n,k}}^{\kappa H_{n,k}}}.$$

4.3. Simulated online EM algorithm for LDA

The pseudo code for the simulated online EM algorithm is given in Algorithm 1 below.

Algorithm 1 Simulated Online EM Algorithm for LDA

Initialize θ_0.
for $n = 1, \ldots$ **do**
  Compute $\tilde{S}^1_n, \ldots, \tilde{S}^m_n$ using MCMC simulations of Z_n or H_n.
  $S_n = (1 - \gamma_n) S_{n-1} + \gamma_n \frac{1}{m} \sum_{i=1}^m \tilde{S}^i_n$.
  $\theta_{n,vk} = \frac{S_{n,vk} + \beta/n}{\sum_v^V S_{n,vk} + \beta/n}$.
end for

To compute the approximations $\tilde{S}^i_n$ of $E_{\theta_{n-1}}[s(W_n, Z_n)|W_n]$ for $i = 1, \ldots, m$, one typically needs to run the MCMC sampler

for slightly longer than m steps so as to discard an initial burn-in period. When using the collapsed Gibbs sampler, $\tilde{S}_n^i$ can be directly computed from the simulated configuration of indicator variables $\tilde{Z}_n^i$ by

$$\tilde{S}_{n,vk}^i = \sum_{l=1}^{L_n} \mathbf{1}\{W_{n,l} = v\}\mathbf{1}\{\tilde{Z}_{n,l}^i = k\}.$$

When using the Metropolis-Hastings algorithm to simulate $\tilde{H}_n^i$, we use the following Rao-Blackwellized estimate:

$$\tilde{S}_{n,vk}^i = E_{\theta_{n-1}}[s_{vk}(W_n, Z_n)|C_n, \tilde{H}_n^i] = \frac{\theta_{n,vk}\tilde{H}_{n,k}^i}{\sum_{k=1}^K \theta_{n,vk}\tilde{H}_{n,k}^i} C_{n,v}.$$

An important practical consideration, in particular for text documents, is that $\tilde{S}_{n,vk}^i$ needs only be computed for the words v that are present in the nth document, as it is zero elsewhere.

4.4. Simulations

To illustrate the potential of the approach on a simple example, we show how LDA trained with online EM can be used to process image data to achieve an NMF-like decomposition. In our image processing setup, an image is represented as a collection of pixels defined by their integer-valued grey level. To make the connection with the text processing setup, the document is now an image, the words are pixels and word counts correspond to grey levels. We apply the model to a noisy version of the benchmark *swimmer* corpus [14] which consists of 256 32-by-32 pixel binary images. The swimmer is a stick figure with 4 limbs each of which has 4 positions. It is a useful dataset as a standard test of the ability of NMF-like models to decompose the 16 limbs. To make the learning task more realistic, each observed image is drawn from a Poissonized version of a randomly selected binary image of the swimmer collection to which a spatially-independent Poisson noise of intensity 0.2 is added. A few example of the resulting images are shown in Figure 2. In this experiment, V is thus equal to $32^2 = 1024$ and L_n is in the range $200 - 300$ (the background contribution is on average equal to $1024 \times 0.2 \approx 205$).

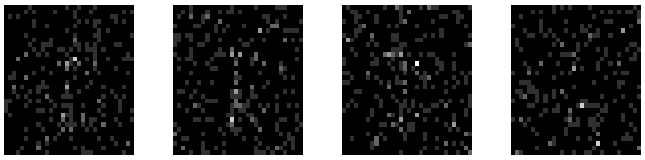


Fig. 2. Sample of training data.

Simulated online EM is applied to the LDA model with 16 topics or 'components' (one for each position of each limb) and after 10000 iterations the 'components' are shown to recover the different positions of each limb see Fig 3 (for this simulation the E-step was computed with the random walk Metropolis algorithm with a 400 sample burn in and averaging over 100 samples with $\kappa = 300$). The fraction of samples accepted varies for different records, but is around 0.4. We see that the algorithm has correctly separated all 16 components. Note that the location of the body is not identifiable under this model.

The convergence properties of the algorithm are shown in Fig 4 which shows the algorithm learning an individual component over time.

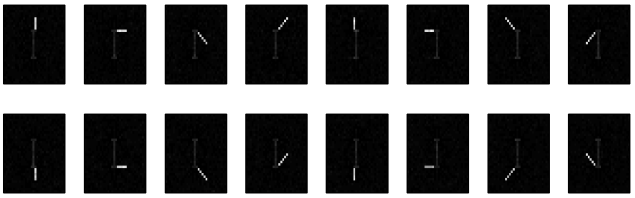
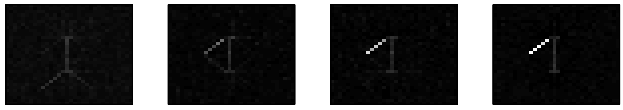


Fig. 3. The 16 components of θ_{10000} plotted as images.



(a) 100 iterations (b) 1000 (c) 2000 (d) 10000

Fig. 4. Convergence of the algorithm for a single component.

5. REFERENCES

[1] O. Cappé and E. Moulines, "On-line expectation-maximization algorithm for latent data models," *J. Roy. Statist. Soc. B*, vol. 71, no. 3, pp. 593–613, 2009.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *JMLR*, vol. 3, pp. 993–1022, 2003.

[3] S. Mohamed, K. Heller, and Z. Ghahramani, "Bayesian exponential family PCA," in *NIPS 21*, MIT Press, 2009.

[4] R. Salakhutdinov and A. Mnih, "Probabilistic matrix factorization," in *NIPS 20*, MIT Press, 2008.

[5] W. Buntine and A. Jakulin, "Discrete component analysis," in *Subspace, Latent Structure and Feature Selection Techniques*. C. Saunders, M. Grobelnik, S. Gunn, and J. Shawe-Taylor, Eds. Springer-Verlag, 2006.

[6] K. A. Heller, S. Williamson, and Z. Ghahramani, "Statistical models for partial membership," in *ICML 25*.

[7] 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, 2009.

[8] M. Sato, "Online model selection based on the variational Bayes," *Neural Comput.*, vol. 13, no. 7, pp. 1649–1681, 2001.

[9] A. Honkela and H. Valpola, "On-line variational Bayesian learning," in *Proc. 4th Int. Symp. on Independent Component Analysis and Blind Signal Separation (ICA2003)*, 2003.

[10] M. Hoffman, D. Blei, and F. Bach, "Online learning for Latent Dirichlet Allocation," in *NIPS 23*, , MIT Press, 2010.

[11] I. Sato, K. Kurihara, and H. Nakagawa, "Deterministic single-pass algorithm for LDA," in *NIPS 23*, , MIT Press, 2010.

[12] O. Cappé, "Online Expectation-Maximization," in *Mixtures*, K Mengersen, M. Titterington, and C. P. Robert, Eds. Wiley, 2011.

[13] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proc. Natl. Academy of Sciences*, vol. 101, no. 1, pp. 5229–5235, 2004.

[14] D. Donoho and V. Stodden, "When does non-negative matrix factorization give a correct decomposition into parts?," in *NIPS 16*, MIT Press, 2004.