EFFICIENT MARKOV CHAIN MONTE CARLO INFERENCE IN COMPOSITE MODELS WITH SPACE ALTERNATING DATA AUGMENTATION

C. Févotte[*]*, O. Cappé*

CNRS LTCI; Télécom ParisTech
Paris, France

A. T. Cemgil[†]

Dept. of Computer Engineering
Boğaziçi University, Istanbul, Turkey

ABSTRACT

Space alternating data augmentation (SADA) was proposed by Doucet *et al* (2005) as a MCMC generalization of the SAGE algorithm of Fessler and Hero (1994), itself a famous variant of the EM algorithm. While SADA had previously been applied to inference in Gaussian mixture models, we show this sampler to be particularly well suited for models having a composite structure, i.e., when the data may be written as a sum of latent components. The SADA sampler is shown to have favorable mixing properties and lesser storage requirement when compared to standard Gibbs sampling. We provide new alternative proofs of correctness of SADA and report results on sparse linear regression and nonnegative matrix factorization.

Index Terms— Markov chain Monte Carlo (MCMC), space alternating data augmentation (SADA), space alternating generalized expectation-maximization (SAGE), sparse linear regression, nonnegative matrix factorization (NMF)

1. INTRODUCTION

In many settings the data is modeled as a sum of latent *components* such that

$$\mathbf{x}_n = \sum\nolimits_{k=1}^{K} \mathbf{c}_{k,n} \tag{1}$$

and the individual components are given a statistical model $p(\mathbf{c}_{k,n}|\theta_k)$. Examples of such composite models occur in

- linear regression; scalar data x_n is expressed as a linear combination of explanatory variables $\phi_{k,n}$ such that

$$x_n = \sum_k \underbrace{s_k \phi_{k,n}}_{c_{k,n}} \tag{2}$$

  and the regressors s_n may for example be given a sparse prior,
- source separation; multichannel data $\mathbf{x}_n$ is expressed as a linear combination of unknown sources with unknown mixing coefficients, such that

$$\mathbf{x}_n = \sum_k \underbrace{s_{kn} \mathbf{a}_k}_{\mathbf{c}_{k,n}} \tag{3}$$

  and the sources s_{kn} are typically mutually independent and given an application-specific prior,
- so-called Kullback-Leibler (KL) and Itakura-Saito (IS) nonnegative matrix factorization (NMF) models [1]; multichannel data $\mathbf{x}_n$ is given by Eq. (1) and the components have a model of the form

$$p(\mathbf{c}_{k,n}|\theta_k) = \prod\nolimits_f p(c_{k,fn}|w_{fk}h_{kn}) \tag{4}$$

  where w_{fk}, h_{kn} are nonnegative scalars.

Let us denote by θ the set of parameters $\{\theta_k\}$. In a Bayesian estimation setting, given a prior $p(\theta)$, one wants to characterize the posterior distribution $p(\theta|X)$ of parameters θ given data X through its mode and/or moments. As this often leads to intractable problems, numerical alternatives have to be sought. One such alternative is Markov chain Monte Carlo (MCMC) inference, which aims at sampling values from the posterior distribution, through a Markov chain with transition kernel $K(\theta|\theta')$ having a stationary distribution equal to the distribution of interest $p(\theta|X)$.

A standard MCMC approach for inference in the abovementioned composite models consists in *completing* or *augmenting* the set of data and parameters with the individual components, acting as latent variables, so as to form a Gibbs sampler which samples the components jointly conditioned on X and θ, and subsequently samples each subset θ_k conditioned on k^{th} component. In this paper we show that sampling the components from their individual marginals instead of their joint distribution produces a valid sampler of $p(\theta|X)$. As will be shown in experiments this results in a sampler with improved mixing and with lesser storage requirement. As it appears this alternative sampler is a special case of the space alternating data augmentation (SADA) sampler of Doucet *et al.* [2]. SADA was introduced as a Monte Carlo version of the space alternating generalized expectation-maximization (SAGE) algorithm [3]. Whilst SADA was applied to inference in Gaussian mixtures models in [2], the aim of this paper is to present the relevance of this sampler for inference in composite models, where the results can be spectacular.

The paper is organized as follows. Section 2 specifies our working assumptions and some notations. Section 3 briefly describes the SAGE algorithm for maximum likelihood (ML) estimation in composite models as it gives the intuition behind the SADA sampler. Section 4 describes and compares Gibbs and SADA samplers, and give alternative proofs of convergence of SADA, in a general case and in the specific case of composite models. Section 5 provides simulation results on sparse linear regression and NMF problems. Section 6 concludes.

[*]Work supported by project ANR-09-JCJC-0073-01 TANGERINE (Theory and applications of nonnegative matrix factorization). Part of this work was done while C. Févotte was visiting Boğaziçi University.

[†]Work supported by scientific and technological research council of Turkey (TUBITAK) by grant 110E292 - BAYTEN (Bayesian matrix and tensor factorisations).

2. NOTATIONS AND WORKING ASSUMPTIONS

In order to ease the notations we will assume scalar data in the following, so that

$$x_n = \sum\nolimits_{k=1}^{K} c_{k,n} \tag{5}$$

and the components $c_{k,n}$ have individual distribution $p(c_{k,n}|\theta_k)$. Despite this simplifying working assumption the results of this paper hold in the general multidimensional case where the single index n is replaced by a tuple of indices, e.g., a pair (f, n) in Section 5.2. We denote by x and c_k the column vectors of dimension N with coefficients $\{x_n\}$ and $\{c_{k,n}\}_n$, respectively, and C the set of coefficients $\{c_{k,n}\}$.[1] Throughout the paper we assume mutual independence of the components conditionally upon θ, i.e.,

$$p(C|\theta) = \prod\nolimits_{k=1}^{K} p(c_k|\theta_k), \tag{6}$$

which is a fundamental assumption for the following results to hold. We will assume for simplicity prior independence of the parameters, i.e., $p(\theta) = \prod_k p(\theta_k)$, though this assumption is not required. Finally, note that model Eq. (5) is not a noiseless model *per se* as one of the components can act as residual noise, which will be the case in one of the two experiments reported in Section 5.

3. EM AND SAGE IN COMPOSITE MODELS

In this section we describe an EM algorithm for ML (or MAP) estimation of θ, which gives the intuition to the forthcoming SADA sampler. The EM algorithm for the maximization of the likelihood $p(x|\theta)$ involves iterative evaluation and maximization of the expected complete data log-likelihood given by[2]

$$Q(\theta|\theta') = \mathrm{E}\{\log p(C|\theta)|x, \theta'\}. \tag{7}$$

Using the factorization implied by the conditional independence $\log p(C|\theta) = \sum_k \log p(c_k|\theta_k)$, the functional may be written as

$$Q(\theta|\theta') = \sum\nolimits_k Q_k(\theta_k|\theta'), \tag{8}$$

$$\text{where} \qquad Q_k(\theta_k|\theta') = \mathrm{E}\{\log p(c_k|\theta_k)|x, \theta'\} \tag{9}$$

$$= \int_{c_k} \log p(c_k|\theta_k)\; p(c_k|x, \theta') dc_k. \tag{10}$$

At this stage it is worth emphasizing that the posterior $p(C|x, \theta)$ of the components is "degenerate", in the sense that the sampled components lie on a hyperplane, because of the constraint $x = \sum_k c_k$. Yet, expectations with respect to this distribution, as required in Eq. (7), are still defined. In contrast, the posterior of the individual components $p(c_k|x, \theta)$, i.e., the marginals of $p(C|x, \theta)$, are not degenerate, so that the integral in Eq. (10) is well defined. Eq. (8) suggests that the task of maximizing $Q(\theta|\theta')$ can be decoupled into k optimization subtasks involving $Q_k(\theta_k|\theta')$ only. The variable θ' can either be refreshed after a full cycle of updates of $\{\theta_1, \ldots, \theta_K\}$ (standard EM), or after every update of θ_k (SAGE). Note that given a prior on θ a MAP estimate can be obtained by changing $p(c_k|\theta_k)$ to $p(c_k, \theta_k)$ in Eq. (10). The decomposition of the EM functional given by Eq. (7) suggests an analogous MCMC approach in which the components c_k would be sampled individually from their marginals $p(c_k|x, \theta)$ instead of being sampled jointly from $p(c_1, \ldots, c_K|x, \theta)$, before sampling each subset θ_k conditionally upon c_k. This is precisely what SADA achieves, while ensuring that the transition kernel $K(\theta|\theta')$ has the correct stationary distribution $p(\theta|x)$, as described in the next section.

[1] By $\{a_{ij}\}_j$ we denote the set $\{a_{ij}\}_{j=1,\ldots,J}$, for a given i. $\{a_{ij}\}$ denotes the set of all coefficients, i.e., for $i = 1, \ldots, I$ and $j = 1, \ldots, J$.

[2] Note that in the general formulation of EM, the complete set may be any set C such that the mapping $C \mapsto x$ is many-to-one, which is the formulation that we use here, and which differs from the more conventional one where the complete set if formed by the union of data and a hidden set.

Algorithm 1 Reference Gibbs sampler

Input : composite data x, initialization $\theta^{(0)}$
for $i = 1, n_{iter}$ **do**
  Choose residual index $r \in \{1, \ldots, K\}$ randomly
  for $k = \{1, \ldots, K\} \backslash r$ **do**
    Sample $c_k^{(i)} \sim p(c_k|x, \{c_j^{'}\}_{j \neq \{k,r\}}, \theta^{'})$ (apostroph ' refers to most recent value)
    Sample $\theta_k^{(i)} \sim p(\theta_k|c_k^{(i)})$
  end for
  $c_r^{(i)} = x - \sum_{k \neq r} c_k^{(i)}$
  $\theta_r^{(i)} \sim p(\theta_r|c_r^{(i)})$
end for
Outp.: samples from $p(C, \theta|x) = p(\theta|x)p(C|x, \theta)$ (after burnin)

4. SADA FOR COMPOSITE MODELS

4.1. From Gibbs to SADA

Let us first discuss Gibbs sampling strategies for $p(\theta|x)$. One iteration of the obvious sampler is based on iteratively sampling C and θ:

(1) $C^{(i)} \sim p(C|x, \theta^{(i-1)})$
(2) $\forall k,\ \theta_k^{(i)} \sim p(\theta_k|c_k^{(i)})$

Because of the sum constraint $x = \sum_k c_k$, sampling the components typically involves reserving one component out, e.g., c_K, acting as a residual noise, sampling from $p(c_1, \ldots, c_{K-1}|x, \theta)$ and then set $c_K = x - \sum_{k=1}^{K-1} c_k$. The components $c_1, \ldots, c_{K-1}$ may be sampled directly from their joint distribution, or conditionally, i.e, from $p(c_k|x, \{c_j\}_{j \neq \{k,K\}}, \theta)$, which corresponds to the following viewpoint:

$$\underbrace{x - \sum\nolimits_{j \neq k,K} c_j}_{\text{observation}} = \underbrace{c_k}_{\text{target}} + \underbrace{c_K}_{\text{residual}}. \tag{11}$$

The component acting as the residual may typically be shuffled at every iteration for improved mixing. The latter strategy will form the basis of our reference Gibbs sampler, which is summarized in Algorithm 1.

SADA essentially consists in sampling each component c_k from its marginal $p(c_k|x, \theta)$ instead of the full conditional $p(c_k|x, \{c_j\}_{j \neq \{k,r\}}, \theta)$, i.e., adopting the following viewpoint:

$$\underbrace{x}_{\text{observation}} = \underbrace{c_k}_{\text{target}} + \underbrace{\sum\nolimits_{j \neq k} c_j}_{\text{residual}}. \tag{12}$$

SADA is summarized in Algorithm 2.

While the sampled values of θ have the target distribution $p(\theta|x)$ in both case, a key difference between Gibbs and SADA is the stationary distribution for the components C; the samples from SADA

Algorithm 2 SADA sampler

Input : composite data x, initialization $\theta^{(0)}$
for $i = 1, n_{iter}$ **do**
  for $k = \{1, \dots, K\}$ **do**
    Sample $c_k^{(i)} \sim p(c_k|x, \theta')$ (apostroph ' refers to most recent value)
    Sample $\theta_k^{(i)} \sim p(\theta_k|c_k^{(i)})$
  end for
end for
Output : samples from $p(\theta|x) \prod_k p(c_k|x, \theta)$ (after burnin)

are not from $p(C|X)$, in particular do not satisfy $x = \sum_k c_k^{(i)}$, but still have the correct marginals, i.e., the chain $\{c_k^{(i)}\}_i$ has stationary distribution $p(c_k|x)$.

4.2. A general proof of convergence for SADA

The following theorem states the validity of SADA (under more general assumptions than composite data) and provides an alternative proof of the correctness of SADA to one of [2].

Theorem 1. *Let $\pi(\theta_1, \dots, \theta_K)$ be a target distribution. Assume that for each k there exists a latent variable c_k and a density q_k such that*

$$\int q_k(c_k, \theta_k, \theta_{-k})\, dc_k = \pi(\theta_k, \theta_{-k}) \tag{13}$$

then $c_k \sim q_k(c_k|\theta'_k, \theta_{-k})$, $\theta_k \sim q_k(\theta_k|c_k, \theta_{-k})$ corresponds to a π-reversible move on coordinate θ_k.

Proof. The transition kernel from θ'_k to θ_k writes

$$\begin{aligned} K(\theta_k|\theta'_k) &= \int q_k(\theta_k|c_k, \theta_{-k}) q_k(c_k|\theta'_k, \theta_{-k}) dc_k \qquad (14)\\ &= \int \frac{q_k(c_k, \theta_k, \theta_{-k})}{\int q_k(c_k, \tilde{\theta}_k, \theta_{-k}) d\tilde{\theta}_k} \frac{q_k(c_k, \theta'_k, \theta_{-k})}{\pi(\theta'_k, \theta_{-k})} dc_k \end{aligned}$$

and thus satisfies the detailed balance equation

$$K(\theta_k|\theta'_k)\pi(\theta'_k, \theta_{-k}) = K(\theta'_k|\theta_k)\pi(\theta_k, \theta_{-k}), \tag{15}$$

which indicates that $\pi(\theta_k, \theta_{-k})$ is stationary for $K(\theta_k|\theta'_k)$. □

Convergence of Algorithm 2 is obtained by applying Theorem 1 to the composite model defined in Section 2, with $\pi(\theta) = p(\theta|x)$ and with

$$\begin{aligned} q_k(c_k, \theta) &= \underline{p(c_k|\theta, x)} p(\theta|x) \qquad (16)\\ &= p(\theta_k|c_k, x, \theta_{-k}) p(c_k, \theta_{-k}|x) \qquad (17)\\ &= \underline{p(\theta_k|c_k)} p(c_k, \theta_{-k}|x) \qquad (18) \end{aligned}$$

where the underlined factors correspond to $q_k(c_k|\theta_k, \theta_{-k})$ and $q_k(\theta_k|c_k, \theta_{-k})$, respectively. For further intuition, and along the proof of [2], SADA can also be obtained as a form of partially collapsed Gibbs sampler [4] of $p(C, \theta|x)$, one iteration of which, for a composite model, reads as follows. We take $K = 2$ for simplicity, but the idea holds for any K.

(1) $(c_1^{(i)}, \tilde{c}_2) \sim p(c_1, c_2|x, \theta_1^{(i-1)}, \theta_2^{(i-1)})$
Reduces to $c_1^{(i)} \sim p(c_1|x, \theta_1^{(i-1)}, \theta_2^{(i-1)})$ and $\tilde{c}_2 = x - c_1^{(i)}$
(2) $\theta_1^{(i)} \sim p(\theta_1|x, c_1^{(i)}, \tilde{c}_2, \theta_2^{(i-1)})$
Reduces to $\theta_1^{(i)} \sim p(\theta_1|c_1^{(i)})$
(2) $(\tilde{c}_1, c_2^{(i)}) \sim p(c_1, c_2|x, \theta_1^{(i)}, \theta_2^{(i-1)})$
Reduces to $c_2^{(i)} \sim p(c_2|x, \theta_1^{(i)}, \theta_2^{(i-1)})$ and $\tilde{c}_1 = x - c_2^{(i)}$
(3) $\theta_2^{(i)} \sim p(\theta_2|x, \tilde{c}_1, c_2^{(i)}, \theta_1^{(i)})$
Reduces to $\theta_2^{(i)} \sim p(\theta_2|c_2^{(i)})$

As it appears the variables $\tilde{c}_1$ and $\tilde{c}_2$ are ghost variables in that they never need to be sampled because they are never conditioned upon. Hence the variables $c_1^{(i)}, c_2^{(i)}, \theta_1^{(i)}, \theta_2^{(i)}$ output by the latter Gibbs sampler coincide with the output of SADA.

5. RESULTS

5.1. Sparse linear regression with Student t prior

Let us assume the linear regression model such that

$$x = \sum\nolimits_{k=1}^{K} s_k \phi_k + e \tag{19}$$

where $\{\phi_k\}$ is a given dictionary of column vectors of dimension N and $s = \{s_k\}$ is a set of scalar regressors. As compared to Eq. (2), we here explicitely assume observation/residual Gaussian noise e of variance v_e, which can be thought of as a $(K+1)^{th}$ component. Let us assume the hierarchical prior $s_k|v_k \sim \mathcal{N}(0, v_k)$ and $v_k \sim \mathcal{IG}(v_k|\alpha, \beta)$, where $\mathcal{N}$ and $\mathcal{IG}$ refer to the Gaussian and inverse-Gamma distributions, respectively. The marginal for s_k under this prior is a Student t distribution with 2α degrees of freedom. For low values of α (typically 0.5 to 1) this prior can be considered "sparse" in that it is sharply peaked at zero and exhibits heavy tails. It has been considered for sparse linear regression in [5] and many other subsequent papers, for example in [6] in a MCMC setting. Next we assume α and v_e to be fixed and β to have a (conjugate) Gamma prior $\mathcal{G}(\nu, \lambda)$. We wish to sample from $\prod_k p(s_k|x)$ for variable selection. We may design Gibbs and SADA samplers on space $\{s, v, \beta\}$, where $v = \{v_k\}$. In this model the latent components are $c_k = s_k \phi_k$ (and e), but they will not need to appear in the sampling algorithms as we can sample from s_k directly. The main difference in the samplers is precisely in the update of s, whilst the other variables can be routinely updated as $v_k \sim \mathcal{IG}(1/2 + \alpha, s_k^2/2 + \beta)$ and $\beta \sim \mathcal{G}(\alpha K + \nu, \sum_k 1/v_k + \lambda)$, see [6]. In both samplers the regressors can easily be shown conditionally Gaussian, such that $s_k \sim \mathcal{N}(\bar{\mu}_k, \bar{v}_k)$, with parameters given by

$$\bar{\mu}_k^{\text{Gibbs}} = g_k \phi_k^T (x - \sum\nolimits_{j \neq k} s_j \phi_j), \quad \bar{v}_k^{\text{Gibbs}} = (1 - g_k \phi_k^T \phi_k) v_k$$

where $g_k = v_k / (v_k\, \phi_k^T \phi_k + v_e)$ and

$$\bar{\mu}_k^{\text{SADA}} = \phi_k^T G_k x, \quad \bar{v}_k^{\text{SADA}} = (1 - \phi_k^T G_k \phi_k) v_k$$

where $G_k = v_k (\sum_k v_k \phi_k \phi_k^T + v_e I)^{-1}$.

We simulated data randomly from the model, using a Gaussian random dictionary, with $N = 100$, $K = 200$, $\alpha = 0.5$, $\nu = \lambda = 1$ and with v_e computed such that the SNR is 50 dB. The simulated samples by Gibbs and SADA of 3 randomly chosen regressors are displayed on Fig. 1. One can see that SADA produces better mixing in such low noise conditions, because it relies on a broader likelihood, as illustrated by Eq. (12). In higher noise scenario, the performances of the samplers are not so contrasted. In this model the computational cost per iteration of SADA is higher than Gibbs because of the matrix inverse involved in the computation of G_k (and despite the inverse can efficiently be refreshed with simple rank-1 updates after every regressor variance update). Hence, SADA may

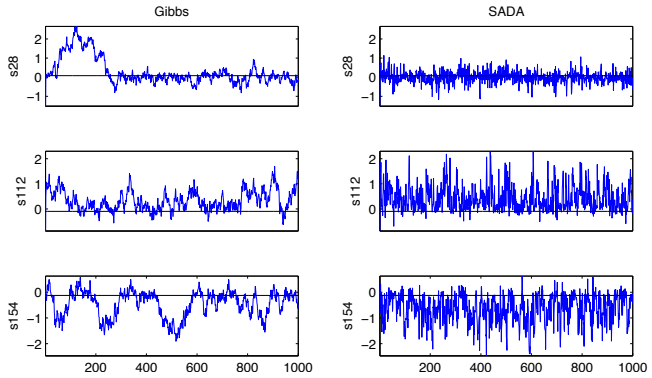

Fig. 1. Samples of three randomly chosen regressors with Gibbs and SADA. Horizontal lines indicate ground truth value. Elapsed times 14 s (Gibbs) and 31 s (SADA) using a MATLAB implementation on a 2.8 GHz Quad-Core Mac with 8 GB RAM.

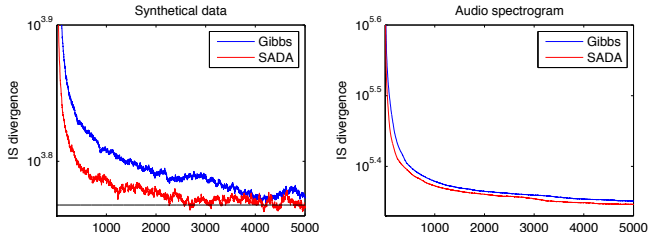

Fig. 2. Itakura-Saito fit $D_{IS}(|X|^2|W^{(i)}H^{(i)})$ (equivalent to minus log-likelihood) from Gibbs and SADA samples on two datasets (one run). Left : synthetical data generated from the model, $F = 100$, $N = 100$, $K = 50$, inverse-Gamma scale and shape prior parameters set to 1. The horizontal line indicates the likelihood of the true parameter. Elapsed times 10 min (Gibbs) and 9 min (SADA). Right : audio data (spectrogram of a short piano sequence), $F = 513$, $N = 674$, $K = 8$. Elapsed times 47 min (Gibbs) and 27 min (SADA).

not be an option of practical use for this model in higher dimension, despite its better mixing properties in low noise. In the next subsection we show an example of model in which SADA comes with lower computational cost than Gibbs, with preserved mixing properties.

5.2. Probabilistic NMF

In [1] we have pointed that ML estimation in models of the form $x_{fn} = \sum_k c_{k,fn}$ with $p(c_{k,fn}|\theta_k)$ chosen as $\mathcal{P}(w_{fk}h_{kn})$ (where $\mathcal{P}$ refers to the Poisson distribution) or $\mathcal{N}_c(0, w_{fk}h_{kn})$ (where $\mathcal{N}_c$ refers to the circular complex Gaussian distribution) leads to the NMF problem $X \approx WH$ under the KL divergence and to the NMF problem $|X|^2 \approx WH$ under the IS divergence, respectively. As described in [1], Gibbs samplers of $\theta = \{W, H\}$ may easily be implemented for these models, with suitable conjugate priors. Denote by $\mathbf{w}_k$ the columns of W and by h_k the rows of H, such that $\theta_k = \{\mathbf{w}_k, h_k\}$, and by C_k the set of coefficients $\{c_{k,fn}\}_{fn}$. In the Gaussian composite model, the posterior distribution $p(c_{1,fn}, \ldots, c_{K,fn}|x_{fn}, \theta)$ is multivariate Gaussian and the posterior distributions $p(w_{fk}|C_k, h_k)$ and $p(h_{kn}|C_k, \mathbf{w}_k)$ are inverse-Gamma.[3] Again, the only difference between Gibbs and SADA lies in how the components are sampled. They are conditionally Gaussian in both case, such that $c_{k,fn} \sim \mathcal{N}(\bar{\mu}_{k,fn}, \bar{v}_{k,fn})$, with

$$\bar{\mu}^{\text{Gibbs}}_{k,fn} = g^{\text{Gibbs}}_{k,fn}(x_{fn} - \sum_{j \neq k,r} c_{j,fn}),\ \bar{v}^{\text{Gibbs}}_{k,fn} = (1 - g^{\text{Gibbs}}_{k,fn})(w_{fk}h_{kn})$$

where $g^{\text{Gibbs}}_{k,fn} = w_{fk}h_{kn}/(w_{fk}h_{kn} + w_{fr}h_{rn})$ and r is the residual index as in Algorithm 1, and

$$\bar{\mu}^{\text{SADA}}_{k,fn} = g^{\text{SADA}}_{k,fn}\, x_{fn}, \quad \bar{v}^{\text{SADA}}_{k,fn} = (1 - g^{\text{SADA}}_{k,fn})(w_{fk}h_{kn})$$

where $g^{\text{SADA}}_{k,fn} = w_{fk}\, h_{kn}/\sum_j w_{fj}\, h_{jn}$. One can see that SADA leads to a very simple implementation, which at every iteration (i) requires to store a single matrix of dimension FN for $C_k^{(i)}$, to which the update of θ_k is conditioned, while Gibbs requires to store the whole tensor $C^{(i)}$ of dimension KFN as required for the computation of $\{\bar{\mu}^{\text{Gibbs}}_{k,fn}\}$. The latter also requires an additional $\mathcal{O}(FN)$ operations. This can be very beneficial to SADA in high dimension as the examples reported in Figure 2 show. We also found to SADA to be generally more robust to local convergence (i.e., when the sampler gets stuck in a mode), in particular when K is large.

[3] Sampling directly from $p(\theta_k|C_k)$ is here difficult; we instead make two Gibbs moves, i.e., update $\mathbf{w}_k$ (resp. h_k) conditionally on C_k and h_k (resp. $\mathbf{w}_k$), which still guarantees convergence.

6. CONCLUSIONS

We have discussed an alternative to Gibbs for inference in composite models in the form of a SADA sampler. SADA comes with better mixing properties and potentially lesser storage requirements. Improved mixing is crucial to overcomplete sparse linear regression with low fit to data requirement, though SADA here incurs a computational complexity increase which can be significant in high dimension. In contrast SADA allows reduces complexity and storage requirement in probabilistic NMF models, and improved mixing makes it more robust to local convergence. In the latter problem SADA is a simple and efficient alternative to usual Gibbs sampling.

7. REFERENCES

[1] C. Févotte and A. T. Cemgil, "Nonnegative matrix factorisations as probabilistic inference in composite models," in *Proc. EUSIPCO'09*.

[2] A. Doucet, S. Sénécal, and T. Matsui, "Space alternating data augmentation: Application to finite mixture of gaussians and speaker recognition," in *Proc. ICASSP'05*.

[3] J. A. Fessler and A. O. Hero, "Space-alternating generalized expectation-maximization algorithm," *IEEE Trans. Signal Processing*, vol. 42, no. 10, pp. 2664–2677, Oct. 1994.

[4] A. Van Dyk and T. Park, "Partially collapsed Gibbs samplers : Theory and methods," *Journal of the American Statistical Association*, vol. 103, no. 482, pp. 790–796, 2008.

[5] M. E. Tipping, "Sparse Bayesian learning and the relevance vector machine," *Journal of Machine Learning Research*, vol. 1, pp. 211–244, 2001.

[6] C. Févotte and S. J. Godsill, "A Bayesian approach to blind separation of sparse sources," *IEEE Trans. Audio, Speech and Language Processing*, vol. 14, no. 6, pp. 2174–2188, Nov. 2006.