# SEQUENTIAL MONTE CARLO SAMPLERS FOR MARGINAL LIKELIHOOD COMPUTATION IN MULTIPLICATIVE EXPONENTIAL NOISE MODELS

*Onur Dikmen*

Department of Information and Computer Science,
Aalto University, Finland

*A. Taylan Cemgil*

Computer Engineering Department,
Boğaziçi University, Turkey

## ABSTRACT

Model scoring in latent factor models is essential for a broad spectrum of applications such as clustering, change point detection or model order estimation. In a Bayesian setting, model selection is achieved via computation of the marginal likelihood. However, this is a typically challenging task as it involves calculation of a multidimensional integral over all the latent variables. In this paper, we consider approximate computation of the conditional marginal likelihood in a multiplicative exponential noise model, which is the generative model for latent factor models with the Itakura-Saito divergence such as the Nonnegative Matrix Factorization (NMF). We show that standard approaches are not accurate and propose two new methods in the sequential Monte Carlo (SMC) samplers framework. We explore the performances of these estimators on two problems.

***Index Terms***— sequential Monte Carlo samplers, Itakura-Saito divergence, Nonnegative Matrix Factorization

## 1. INTRODUCTION

Latent factor models which explain the generation of data through latent variables are widely used in machine learning and data analysis. Topic models (e.g., Probabilistic Latent Semantic Indexing (pLSI) [1], Latent Dirichlet Allocation (LDA) [2]) and nonnegative matrix factorization [3] (NMF) models based on Kullback-Leibler divergence have been applied to text analysis, machine vision, bioinformatics and finance. NMF is not limited to discrete count data and can be used in various applications. For example, NMF using Itakura Saito divergence provides a natural generative model for time-frequency coefficients of audio signals and was successfully used in audio source separation [4].

One central but challenging computational problem in latent factor models is model selection (scoring). In a Bayesian setting, this is achieved via calculation of the *marginal likelihood*, or the model evidence. This quantity provides a measure to assess the generalization power of a given model but involves the calculation of an intractable integral over all or some of the latent factors. Several estimators for the marginal likelihood in topic models were proposed in [5, 6]. In this paper, we investigate conditional marginal likelihood in the relatively less studied multiplicative exponential noise model, which is also the generative model for NMF using Itakura-Saito divergence [4]. Conditional marginal likelihood is an integral of form

$$C_{\mathrm{ML}} \equiv p(\mathbf{x}|\mathbf{W}) = \int p(\mathbf{x}|\mathbf{W},\mathbf{h})p(\mathbf{h})\,d\mathbf{h} \tag{1}$$

where $\mathbf{x}$ is observed data, $\mathbf{W}$ is a fixed dictionary matrix and $\mathbf{h}$ are the latent factors. The calculation of the actual marginal likelihood $p(\mathbf{x})$ is beyond the scope of this paper. Subsequently, we will refer to $p(\mathbf{x}|\mathbf{W})$ as the marginal likelihood and drop the ‘conditional’ qualifier. Whilst not elaborated here, the techniques described are applicable to latent factor models using other divergences (Kullback-Leibler, Euclidean, etc.) and their extensions.

For the estimation of the marginal likelihood, we propose two novel methods based on sequential Monte Carlo (SMC) samplers [7] and compare these to standard methods, such as Laplace approximation, Chib’s method [8], importance sampling (IS) [6] and variational lower bounds. SMC samplers constitute a general and flexible framework for constructing integration methods that includes powerful techniques such as resample and move particle filtering, annealed importance sampling (AIS) [9] or population Monte Carlo methods [10] as special cases. SMC samplers use IS sequentially on an extended space, starting from simple target distributions, moving towards more complicated densities admitting the original target as a marginal. This helps exploring separated modes of a target density, moreover they provide an estimate for the marginal likelihood. We describe AIS as an instance of SMC samplers. Moreover, we propose two SMC methods based on sequential processing of data and/or slowly increasing the dimensionality of the problem (number of components). For comparison, we also adopt left-to-right samplers [5, 6], designed for marginal likelihood estimation in LDA. We test these methods first on a synthetical dataset, then on the problem of interpolative decomposition (ID) [11]. We show that the problem is difficult even in moderate dimensions and most of the standard methods fail to estimate the likelihood accurately. Hence we conclude that computationally heavy samplers are indeed needed.

## 2. MODEL

The multiplicative exponential noise model is given by

$$v_f = \hat{v}_f \epsilon_f\,, \tag{2}$$

where $f$ is a feature index, $\epsilon_f$ is an exponential distributed noise variable, $\epsilon_f \sim \exp(-\epsilon_f)$, and $\hat{v}_f$ is a nonnegative variable represented as $\hat{v}_f = \sum_k^K w_{fk}h_k$, i.e., a linear combination of nonnegative dictionary atoms, $w_{fk}$, and their excitations, $h_k$. $K$ is the number of components. Estimating $w_{fk}$ and $h_k$ by maximizing the likelihood of the model, $\log p(\mathbf{v}|\mathbf{W},\mathbf{h}) = \sum_f -\log \hat{v}_f - v_f/\hat{v}_f$, corresponds to minimizing the Itakura-Saito divergence between $\mathbf{v}$ and $\mathbf{Wh}$, as was shown in [4].

With notation $v_f \equiv |x_f|^2$, an equivalent generative model (up to a constant) is the following

$$x_f \sim \mathcal{N}_{\mathrm{u}}(x_f; 0, \sum\nolimits_{k=1}^{K} w_{fk}h_k)\,, \tag{3}$$

OD is funded by The Academy of Finland (Finnish Centre of Excellence in Computational Inference Research COIN, 251170). ATC is funded by Bogazici University research fund BAP 6882.

where u is 0.5 for real, 1 for complex Gaussian distributions. In this work, we are interested in the marginal likelihood of the dictionary, $p(\mathbf{x}|\mathbf{W})$. For this, we introduce an inverse Gamma prior distribution for the excitation variables, $h_k \sim \mathcal{IG}(h_k; a_k, b_k)$, and seek to integrate them out[1]. The model can be equivalently expressed as

$$x_f = \sum_{k=1}^{K} c_{fk}, \quad c_{fk} \sim \mathcal{N}_{\mathrm{u}}(c_{fk}; 0, w_{fk}h_k), \quad h_k \sim \mathcal{IG}(h_k; a_k, b_k)\,,$$

where $c_{fk}$ are the latent variables which will be called *components* throughout the text[2]. These components not only increase the interpretability of the model, but also make some inference (e.g., Gibbs sampling) and optimization procedures easier. Our target is to compute the marginal likelihood as defined in (1).

The full conditional distribution of an excitation variable $h_k$ is an inverse Gamma distribution with the following parameters

$$p(h_k|\mathbf{C}^{(i)}, \mathbf{W}) = \mathcal{IG}(h_k; \alpha_k, \beta_k) \tag{4}$$

$$\alpha_k = a_k + \mathrm{u}F, \quad \beta_k = b_k + \mathrm{u}\sum\nolimits_f |c_{fk}^{(i)}|^2 / w_{fk}\,, \tag{5}$$

where superscript $(i)$ denotes current sample in Gibbs sampling and $F$ is the number of features.

The latent variables $\mathbf{c}_f$ have a full conditional distribution of $K$-dimensional Gaussian with the following mean vectors and covariance matrices

$$p(\mathbf{c}_f|\mathbf{x}, \mathbf{h}^{(i)}, \mathbf{W}) = \mathcal{N}_{\mathrm{u}}(\mathbf{c}_f; \boldsymbol{\mu}_f, \boldsymbol{\Sigma}_f) \tag{6}$$

$$\boldsymbol{\mu}_f = \frac{x_f}{\hat{v}_f}\mathbf{d}_f \qquad \boldsymbol{\Sigma}_f = \mathrm{diag}(\mathbf{d}_f) - \mathbf{d}_f\mathbf{d}_f^T / \hat{v}_f \tag{7}$$

where $\lambda_{fk} = w_{fk}h_k^{(i)}$, $\hat{v}_f = \sum_k \lambda_{fk}$ and $\mathbf{d}_f = [\lambda_{f1} \cdots \lambda_{fK}]^T$.

## 3. SEQUENTIAL MONTE CARLO SAMPLERS

SMC samplers [7] are a framework of methods which consider the problem in an extended state space to introduce importance distributions that match better with the target distribution. This target distribution has the original target distribution as its marginal and the goal remains to evaluate expectations under this marginal and/or estimate its normalizing constant. Annealed importance sampling (AIS) [9] is an important special case of SMC, although introduced earlier, and will be explained in Section 3.1. We will describe two SMC methods specifically designed for the multiplicative exponential noise model in Section 3.2.

In SMC, we define a sequence of artificial target distributions, $p_n = \pi_n / Z_n$, moving from a simple distribution $p_1$ towards the actual distribution of interest $p_S = p(\mathbf{h}|\mathbf{x})$. A potentially effective method for constructing good proposals is by drawing samples from a simple importance distribution $q_1$ and moving them using a MCMC transition kernels $K_n$, such as Metropolis-Hastings. The aim is to make the proposal close to the target and apply importance sampling sequentially. For example, at time 2 the importance weights would be given by

$$w_2(\mathbf{h}_2) = \frac{\pi_2(\mathbf{h}_2)}{q_2(\mathbf{h}_2)} = \frac{\pi_2(\mathbf{h}_2)}{\int q_1(\mathbf{h}_1) K_2(\mathbf{h}_1, \mathbf{h}_2) d\mathbf{h}_1}\,. \tag{8}$$

Unfortunately, the marginal importance distribution $q_2$ is difficult to obtain. Some transition kernels (e.g., Metropolis-Hastings) cannot even be evaluated pointwise. In addition, using Monte-Carlo approximation for this integral will be too costly ($O(N_s^2)$). However, this can be overcome by defining alternative weights on extended space

$$w_2(\mathbf{h}_{1:2}) = \frac{\pi_2(\mathbf{h}_2) L_1(\mathbf{h}_2, \mathbf{h}_1)}{q_1(\mathbf{h}_1) K_2(\mathbf{h}_1, \mathbf{h}_2)}\,, \tag{9}$$

where $L_1$ is an arbitrary backward Markov kernel. This trick corresponds to defining the target and importance distributions on extended space

$$\tilde{p}_n(\mathbf{h}_{1:n}) = p_n(\mathbf{h}_n) \prod_{k=1}^{n-1} L_k(\mathbf{h}_{k+1}, \mathbf{h}_k)\,, \tag{10}$$

$$q_n(\mathbf{h}_{1:n}) = q_1(\mathbf{h}_1) \prod_{k=2}^{n} K_k(\mathbf{h}_{k-1}, \mathbf{h}_k)\,. \tag{11}$$

Because $\tilde{p}_n(\mathbf{h}_{1:n})$ has $p_n(\mathbf{h}_n)$ as its marginal, it is possible to estimate this distribution and its normalizing constant $Z_n$ with IS.

The SMC algorithm proceeds as follows: $N_s$ samples are drawn from $q_1(\mathbf{h}_1)$. With these samples $\mathbf{h}_1^{(i)}$ it is straightforward to evaluate $w_1^{(i)}$ because both $q_1$ and $\pi_1$ can be evaluated pointwise. At time $n$, using the weighted particles $\{w_{n-1}^{(i)}, \mathbf{h}_{1:n-1}^{(i)}\}$ from $n-1$, first the particles are extended with $K_n(\mathbf{h}_{n-1}, \mathbf{h}_n)$, then the weights of each particle is updated using

$$w_n(\mathbf{h}_{1:n}) = w_{n-1}(\mathbf{h}_{1:n-1}) \frac{q_n(\mathbf{h}_n) L_{n-1}(\mathbf{h}_n, \mathbf{h}_{n-1})}{q_{n-1}(\mathbf{h}_{n-1}) K_n(\mathbf{h}_{n-1}, \mathbf{h}_n))}\,. \tag{12}$$

As with standard SMC methods, the variance of the unnormalized weights is liable to increase with time due to degeneracy. Resampling can be applied as a remedy whenever effective sample size (ESS) is below a certain threshold. When resampling is used, the sum of current weights is stored as an estimate for $Z_n / Z_{n'}$ where $n'$ is the last time index that resampling was performed and each particle is assigned equal weights ($1/N_s$).

If $K_n$ is chosen to be an MCMC kernel with invariant distribution $p_n$, this backward kernel can be approximated as

$$L_{n-1}(\mathbf{h}_n, \mathbf{h}_{n-1}) = \frac{p_n(\mathbf{h}_{n-1}) K_n(\mathbf{h}_{n-1}, \mathbf{h}_n)}{p_n(\mathbf{h}_n)}\,. \tag{13}$$

with the assumption $p_n \approx p_{n-1}$. This is the reversal Markov kernel associated with $K_n$.

### 3.1. Annealed Importance Sampling

In AIS [9], a sequence of distributions $p_n(\mathbf{h}) = p(\mathbf{h})^{1-\beta_n} p(\mathbf{x}, \mathbf{h}|\mathbf{W})^{\beta_n}$ is chosen, with $0 \leq \beta_1 < \cdots < \beta_S = 1$. The unnormalized density at time $n$, $\pi_n(\mathbf{h})$ is equal to $p(\mathbf{x}|\mathbf{h}, \mathbf{W})^{\beta_n} p(\mathbf{h})$.

Choosing $q_1(\mathbf{h}_1)$ as the prior $p(\mathbf{h})$, the weights at time 1 becomes

$$w_1(\mathbf{h}_1) = \frac{\pi_1(\mathbf{h}_1)}{p(\mathbf{h}_1)} = p(\mathbf{x}|\mathbf{h}_1, \mathbf{W})^{\beta_1}\,. \tag{14}$$

Using an MCMC kernel with invariant distribution $p_n(\mathbf{h})$ and its reverse given in (13), the weights at time $n$ is given by

$$w_n(\mathbf{h}_{1:n}) = w_{n-1}(\mathbf{h}_{1:n-1}) \frac{\pi_n(\mathbf{h}_{n-1})}{\pi_{n-1}(\mathbf{h}_{n-1})} \tag{15}$$

$$= w_{n-1}(\mathbf{h}_{1:n-1}) p(\mathbf{x}|\mathbf{h}_{n-1}, \mathbf{W})^{\beta_n - \beta_{n-1}}\,. \tag{16}$$

[1] $\mathcal{IG}(x|\alpha,\beta) = \frac{x^{-\alpha-1}\exp(-\beta/x)\beta^\alpha}{\Gamma(\alpha)}$, $x > 0$
$\langle \frac{1}{x} \rangle = \frac{\alpha}{\beta}$, $\quad \langle \log x \rangle = \Psi(\alpha) + \log \beta$

[2] In the text, $\mathbf{C}$ denotes a matrix with elements $c_{fk}$ and $\mathbf{c}_f$ is its $f$th row.

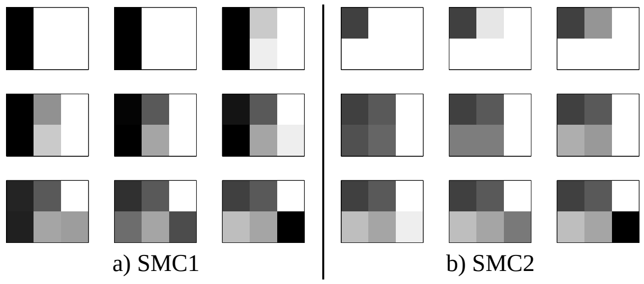

**Fig. 1**. Dictionary sequences constructed by SMC1 and SMC2 from a dictionary of size $2 \times 3$. Initial and final (original) dictionaries are on top-left and bottom-right, respectively. Iterations progress row-wise.

### 3.2. SMC with Dictionary Masking

In this section, we propose two SMC samplers for estimating the marginal likelihood in multiplicative exponential noise model. The main idea lies in the fact that this problem is trivial for $K = 1$, easy for $K = 2$ or $K = 3$ and gets increasingly complicated as K increases. This enables us to design $p_1, \dots, p_S$ with increasing complexity. In the first method (we will call it SMC1 from now on), we collapse all columns of $\mathbf{W}$ into its first column and set the rest of the columns to zero. We define the masked dictionary at time 1, $\mathbf{W}'_1$, as

$$w'_{1,f1} = \sum\nolimits_k w_{fk}, \quad \forall f; \qquad w'_{1,fk} = 0, \quad \forall f, \forall k > 1$$

So, $p_1(\mathbf{h}_1) = p(\mathbf{h}_1|\mathbf{x}, \mathbf{W}'_1)$. Over the next $S/(K-1)$ iterations, we slowly increase the values in the second column of the masked dictionary $\mathbf{W}'_n$, by moving $\beta_n w_{f2}$ from the first column to the second. Here, $\beta_n$ goes from zero to one. This is repeated for each of the other columns in turn. At the last step, the masked dictionary becomes the same as the original, $\mathbf{W}'_S = \mathbf{W}$. The progression of $\mathbf{W}'_n$ in time is illustrated in Fig. 1a.

We start the algorithm by drawing samples from $q_1(\mathbf{h}_1) = p_1(\mathbf{h}_1) = p(\mathbf{h}_1|\mathbf{x}, \mathbf{W}'_1)$. Since $\mathbf{W}'_1$ has only one nonzero column, this distribution is analytically available as

$$p(\mathbf{h}|\mathbf{x}, \mathbf{W}'_1) = p(h_1|\mathbf{x}, \mathbf{W}'_1) \prod_{k=2}^{K} p(h_k)\,, \tag{17}$$

where $p(h_k)$ are the prior distributions and $p(h_1|\mathbf{x}, \mathbf{W}'_1)$ can easily be derived as $\mathcal{IG}(h_1; a_k + \mathrm{u}F, b_k + \mathrm{u}\sum_f |x_f|^2/w'_{1,f1})$. The first weights being constant, $w_1(\mathbf{h}_1) = \pi_1(\mathbf{h}_1)/p(\mathbf{h}_1)$, the weights at time $n$ are given by

$$w_n(\mathbf{h}_{1:n}) = w_{n-1}(\mathbf{h}_{1:n-1}) \frac{p(\mathbf{x}|\mathbf{W}'_n, \mathbf{h}_{n-1})}{p(\mathbf{x}|\mathbf{W}'_{n-1}, \mathbf{h}_{n-1})}\,. \tag{18}$$

In SMC methods which use MCMC kernels, it is crucial to choose $p_{n-1}$ as close as possible to $p_n$. So, when starting to fill a new column, it may be advantageous to select the column which is has the next smallest angle with the sum of the original columns (first column of $\mathbf{W}'_1$).

Our second method (SMC2) starts with one column of $\mathbf{W}$ as the previous one, but only using one observation, $x_1$. Then, iteratively it adds another column *à la* SMC1 and another observation (row). If there is no more columns or rows, it goes in the other direction one by one. When adding a new observation, we choose the closest observation to the last one and copy the dictionary row associated with the last observation as the dictionary row of the new observation and slowly modify it so that it is equal its original value in the end (see Fig. 1b). At time $n$, the weights are evaluated using

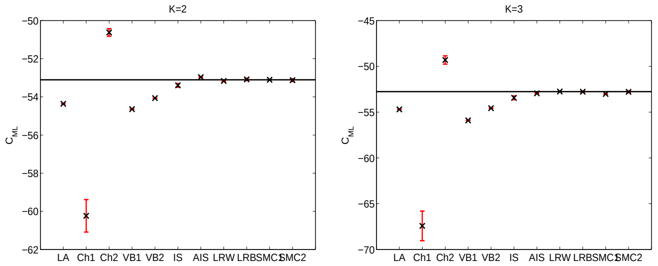

**Fig. 2**. Likelihood estimates on synthetical data for $K = 2$ and $K = 3$. The number of iterations used in VB, VB2 (as well as the VB2 inside IS and MM inside Laplace's method) is 1000, the number of samples in Ch1, Ch2 (Chib-style estimator introduced in [12]), IS, LRW and LRB is 5000. In AIS, SMC1 and SMC2 the number of iterations is 5000, while the number of samples is 100. 10 random initializations are used for each method. The solid line is the exact marginal likelihood.

$$w_n(\mathbf{h}_{1:n}) = w_{n-1}(\mathbf{h}_{1:n-1}) \frac{p(\mathbf{x}'_n|\mathbf{W}'_n, \mathbf{h}_{n-1})}{p(\mathbf{x}'_{n-1}|\mathbf{W}'_{n-1}, \mathbf{h}_{n-1})}\,. \tag{19}$$

Most of the time, the constants in the likelihood term cancel each other, but at the time a new observation is introduced, the cardinality of $\mathbf{x}'_n$ is one more than that of $\mathbf{x}'_{n-1}$, so the weight contains the term $-\mathrm{u}\log \pi/\mathrm{u}$.

## 4. EXPERIMENTAL RESULTS

### 4.1. Synthetical Data

In order to compare SMC sampler based estimators to standard methods, we generate data from the generative model, for various $K$ with $F = 10$, $a_k = 1$ and $b_k = 1$. $\mathbf{W}_{\text{true}}$ is drawn from the Gamma distribution, $w_{fk} \sim \mathcal{G}(w_{fk}; 1, 1)$. When $K = 1$, the only excitation variable, $h_1$, can be integrated out, i.e., the exact marginal likelihood is analytically available. In the general case, defining normalized excitation variables $\tilde{h}_k = h_k/H$ with $H = \sum_k^K h_k$, applying a change of variables between $h_1 ... h_K$ and $\tilde{h}_1 ... \tilde{h}_{K-1}, H$ and integrating out $H$ analytically, one can arrive at an expression with $K-1$ degrees of freedom. For $K = 1$, the marginal likelihood is readily available with $\tilde{h}_1 = 1$. Using this reduction, one can easily approximate the log likelihood when $K = 2$ or $K = 3$ by a Riemann sum with $\mathbf{h}$ in the range $[0, 1]$ as a ground truth.

In Fig. 2, log likelihood estimates for $K = 2$ and $K = 3$ are displayed. It is evident from the figure that simple approximation methods, such as Laplace's method (LA), Chib's method (Ch1, Ch2), variational lower bounds (VB1, VB2) and importance sampling (IS), cannot estimate the likelihood accurately. On the other hand, SMC methods (including AIS) and left-to-right samplers (LRW [5], LRB [6])[3] perform well. In a more difficult scenario, we compare the estimators in a classification experiment. We construct two dictionaries, $\mathbf{W}_1$ and $\mathbf{W}_2$ with $K = 15$ such that they

[3] For this model, these samplers are not stand-alone, they require $p(x_1|\mathbf{W})$ from any of the other methods.

**Table 1**. Classification errors (the number of errors out of 20 samples) with $K = 15$. All standard methods considered give higher error rates, thus were removed from the table.

| LRW | LRB | AIS | SMC1 | SMC2 |
|---|---|---|---|---|
| $9.6 \pm 1.5$ | $7.8 \pm 1.9$ | $8.8 \pm 1.5$ | $8.4 \pm 3.0$ | $7.6 \pm 1.1$ |

only differ in one column. We generate two datasets from these two dictionaries with $N_1 = N_2 = 10$ data samples. Then, we try to classify these 20 samples by comparing their marginal likelihoods, $p(\mathbf{x}_n|\mathbf{W}_1)$ and $p(\mathbf{x}_n|\mathbf{W}_2)$. The classification errors obtained with five repetitions (different datasets) are given in Table 1. According to these results LRB and SMC2 performs the best when $K$ is large.

### 4.2. Nonnegative Interpolative Decomposition (ID)

We illustrate the accuracy of our marginal likelihood estimates in a model selection scenario. The interpolative decomposition (ID) (e.g., see [11] for an excellent introduction) is a technique for decomposing a data matrix $\mathbf{X}$ as $\mathbf{X}(:, \neg\mathbf{r}) = \mathbf{X}(:, \mathbf{r})\mathbf{H}$ where $r$ denotes the selected columns of $\mathbf{X}$ and $\neg\mathbf{r}$ denotes unselected columns. It is natural to encode $\mathbf{r}$ as a binary vector where the corresponding element is one (zero) when the corresponding column is selected (not selected).

Inspired by the ID, we define the following generative model as a nonnegative interpolative decomposition (NID) model

$$\begin{aligned} w_{fk} &\sim \mathcal{G}(w_{fk}; 1, 1), \;\; f = 1..F, \; k = 1..K \\ h_{kn} &\sim \mathcal{IG}(h_{kn}; a_k, b_k), \;\; k = 1..K, \; n = 1..N - K \\ r_n &\sim \mathcal{BE}(r_n; 0.5), \;\; n = 1..N \\ \mathbf{X}(:, \mathbf{r}) = \mathbf{W}, &\quad\;\; \mathbf{X}(:, \neg\mathbf{r}) \sim p(\cdot|\mathbf{WH}) \,. \end{aligned}$$

In other words, the selected (but unknown) columns give directly the $\mathbf{W}$ matrix. By calculating the marginal likelihood for each configuration of $\mathbf{r}$, we can estimate the order of the nonnegative ID decomposition. We seek the optimal indicator vector $\hat{\mathbf{r}}$ which maximizes $p(\mathbf{X}(:, \mathbf{r}), \mathbf{X}(:, \neg\mathbf{r})) = p(\mathbf{X}(:, \mathbf{r}))p(\mathbf{X}(:, \neg\mathbf{r})|\mathbf{X}(:, \mathbf{r}))$.

For small $N$ we can evaluate the marginal likelihood for all $2^N - 1$ (excluding all zero) configurations of $\mathbf{r}$. We observe that when we generate data from the true model we are able to recover the true decomposition, e.g., with $F = 10$, $K = 3$, $N = 5$, $a_k = 1$ and $b_k = 1$. We performed 10 independent trials with the most promising methods from the previous section LRB and SMC2 and compared their performances to IS, which was the most accurate among the fast estimators. The number of errors in bits of $\hat{\mathbf{r}}$ and $\mathbf{r}_{\text{est}}$ out of these 10 runs (thus, total number of bits compared is 50) are zero for LRB and SMC2 and three for IS. These results tell us that we need sampler-based estimators for high accuracy, but this restricts us in the size of the problem, because of their high computational complexity. Still, we have to keep in mind that the inference method used here is a brute force technique which requires very high number of likelihood computations.

## 5. CONCLUSION AND DISCUSSION

We proposed two new SMC samplers for estimating the conditional marginal likelihood of the multiplicative exponential noise model, which is widely used in applications such as music transcription and source separation. In the experiments, we showed that standard approaches estimate the marginal likelihood poorly and more elaborate methods are indeed needed for more accurate estimates. We observe that the SMC samplers (AIS, SMC1 and SMC2) and left-to-right samplers, designed originally for LDA and adopted here for NMF, perform significantly better. SMC2 and LRB[4] give consistently accurate results but are computationally the most demanding ones. The results for SMC samplers suggests that carefully choosing the tempering schedule is key in accurate inference. Currently, this fact renders the use of these methods difficult in large scale data processing applications and parallelization seems to be the direction for further investigation.

## 6. ACKNOWLEDGMENTS

The authors would like to thank Cédric Févotte for discussions on the model and the significance of the problem.

## 7. REFERENCES

[1] T. Hofmann, “Probabilistic latent semantic indexing,” in *Proc. 22nd International Conference on Research and Development in Information Retrieval (SIGIR)*, 1999.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent Dirichlet allocation,” *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[3] D. D. Lee and H. S. Seung, “Learning the parts of objects with nonnegative matrix factorization,” *Nature*, vol. 401, pp. 788–791, 1999.

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

[5] H. Wallach, I. Murray, R. Salakhutdinov, and D. Mimno, “Evaluation methods for topic models,” in *Proceedings of the 26th International Conference on Machine Learning (ICML 2009)*, 2009.

[6] W. L. Buntine, “Estimating likelihoods for topic models,” in *Asian Conference on Machine Learning*, 2009.

[7] P. Del Moral, A. Doucet, and A. Jasra, “Sequential Monte Carlo samplers,” *Journal of the Royal Statistical Society B*, vol. 68, no. 3, pp. 1–26, 2006.

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

[9] R. M. Neal, “Annealed importance sampling,” *Statistics and Computing*, vol. 11, no. 2, pp. 125–139, 2001.

[10] O. Cappé, A. Guillin, J. M. Marin, and C. P. Robert, “Population Monte Carlo,” *Journal of Computational and Graphical Statistics*, vol. 13, pp. 907–930, 2004.

[11] N Halko, P. G. Martinsson, and J. A. Tropp, “Finding structure with randomness: stochastic algorithms for constructing approximate matrix decompositions,” Tech. Rep., Caltech, 2009.

[12] I. Murray and R. Salakhutdinov, “Evaluating probabilities under high-dimensional latent variable models,” in *Advances in Neural Information Processing Systems*, 2009, vol. 21.

[4] LRB is not a real alternative here, because it requires an initial estimate from another method.