Bayesian non-negative matrix factorization with adaptive sparsity and smoothness prior

Ondřej Tichý, Lenka Bódiová, Václav Šmídl, *Member, IEEE,*

***Abstract*—Non-negative matrix factorization (NMF) is generally an ill-posed problem which requires further regularization. Regularization of NMF using the assumption of sparsity is common as well as regularization using smoothness. In many applications it is natural to assume that both of these assumptions hold together. To avoid ad hoc combination of these assumptions using weighting coefficient, we formulate the problem using a probabilistic model and estimate it in a Bayesian way. Specifically, we use the fact that the assumptions of sparsity and smoothness are different forms of prior covariance matrix modeling. We use a generalized model that includes both sparsity and smoothness as special cases and estimate all its parameters using the variational Bayes method. The resulting matrix factorization algorithm is compared with state-of-the-art algorithms on large clinical dataset of 196 image sequences from dynamic renal scintigraphy. The proposed algorithm outperforms other algorithms in statistical evaluation.**

***Index Terms*—Non-negative matrix factorization, Covariance matrix model, Blind source separation, Variational Bayes method, Dynamic renal scintigraphy**

I. Introduction

The aim of non-negative matrix factorization (NMF) is to find low rank representation of non-negative data matrix, $D \in \mathbf{R}_+^{p\times n}$, as the product of two non-negative matrices, the first, $A \in \mathbf{R}_+^{p\times r}$, comprising basis vectors in data space in columns and the second, $X \in \mathbf{R}_+^{n\times r}$, comprising coefficients scaling the basis vectors in rows [1]. Then, the NMF model is

$$D \approx AX^T, \tag{1}$$

where $(.)^T$ denotes transposition of the vector or matrix and $r \ll \min(p, n)$ is the model order. To tackle this problem efficiently, two algorithms were proposed by Lee and Seung, one minimizing conventional least-square error, and other minimizing the Kullback-Leibler divergence [2]. This allows to apply the approach to various large data problems in data clustering [3], hyperspectral imaging [4], or semi-blind and blind separation of dynamic medical image data [5].

Although the non-negativity itself is an informative assumption, the NMF is in general not unique and additional constraints are often required. While many variants and improvements of NMF such as volume-based regularizations [6] have been proposed, sparsity of the solution is often considered as a natural assumption for the NMF problem since it reduces a total number of coefficients required to encode the data with reasonable proximity between original data matrix and reconstruction low-rank matrices [7]. While sparsity is often associated with L1 norm formulation or relation between L1 and L2 norm [8], we turn our attention to probabilistic formulations using hierarchical priors [9], [10] since they allow to estimate all nuisance parameters. A commonly used form of the prior model is the zero-mean Gaussian prior for basis or coefficient matrix with Gamma prior for diagonal elements of their unknown covariance matrix, which is one form of the automatic relevance determination (ARD) principle [11]. The benefits of probabilistic formulation include quantifying the model order [12], [13] as well as accounting for parameter uncertainties [7]. Although the complexity of these models is higher then that of [2], efficient inferences such as Gibbs sampling or variational Bayes (VB) method have been developed to estimate model parameters in reasonable time [14] and with theoretical guarantees [15].

The sparsity is often the key assumption, however, it may be in a contrast with other fundamental assumption: smoothness. In many natural signals such as defects testing [16], or dynamic medical data [17], it is reasonable to assume that increments between neighboring elements are small. The smoothness may be added to the NMF model e.g. using the Gibbs regularization term [18] or incorporating a dependence term into the iterative NMF algorithm [19], [17]. Various basis functions have been exploited in [20] and also extended to the non-negative tensor factorization problem [21]. In dynamic medical imaging problems, the smoothness has been imposed using convolution modeling [22] and recently by imposing regularization term directly assuming smoothness of organ activities [17] or organ convolution kernels [23]. However, a NMF algorithm with seamless combination of sparsity and smoothness is not available.

In this letter, we propose a NMF model with prior model that combines sparsity and smoothness using flexible covariance matrix model. Priors combining sparsity and smoothness [24] has been proposed based on the fused LASSO formulation [25]. However, they require the use of Gibbs sampler and favor piece-wise constant solution. An alternative prior based on adaptive covariance prior has been proposed in [26] that allows to use faster inference algorithms based on the variational Bayes. While it is possible to impose this prior on both X and A (using 2D prior of [27]), we provide solution to a model with this prior only on matrix X. The prior for matrix A is the sparsity promoting ARD prior that is suitable to our application in dynamic medical imaging [28].

A. Competing state of the art algorithms

The decomposition (1) becomes very popular especially after Lee and Seung [2] introduced an efficient algorithm that

O. Tichý, L. Bódiová, and V. Šmídl are with the Czech Academy of Sciences, Institute of Information Theory and Automation, Prague CZ-182 08, Czech Republic, e-mail: otichy@utia.cas.cz

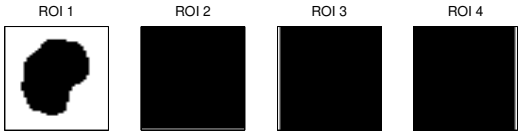

Figure 1. ROIs used to initialize the RUDUR method for a selected data file. ROI1 is the ROI of the parenchyma manually selected by an expert. Black pixels are inside the ROI. Other sources have vague ROIs where all pixels are supposed to be active except for the bottom row (ROI2), the first column (ROI3), and the last column (ROI4).

iteratively optimizes the Euclidean distance between matrices D and AX^T given by

$$f_{\text{NMF}} = ||D - AX^T||_F^2, \tag{2}$$

where $||.||_F$ is the Frobenius norm. The convergence to the local optimum is guaranteed and the algorithm is very fast, however, domain specific assumptions are difficult to incorporate. This algorithm will be denoted simply as NMF.

A version of NMF with sparsity and smoothness assumptions was presented in [17] as robust unmixing of dynamic sequences using regions of interest (RUDUR) method. Region of interest (ROI) is a binary map of each basis vector of A (see Fig. 1 left as an example). In this approach, the basic distance has been changed to

$$f_{\text{WLS}}(A, X) = ||W(D - AX^T)||_F^2, \tag{3}$$

using diagonal weighting matrix W with elements computed as $W_{ii} = \frac{1}{1+\min_k g(R_{i,k})}$, where $R_{i,k}$ is the Euclidean distance of the ith pixel of reconstructed basis vector and the kth pixel of the preselected ROI, and $g(x) = \gamma x^2$ with selected constant $\gamma > 0$. Note that for $\gamma = 0$, the term (3) becomes the ordinary least squares (2).

The basic cost (3) is complemented by two additional penalization terms

$$f_{\text{RUDUR}}(A, X) = f_{\text{WLS}}(A, X) + \alpha f_{\text{ROI}}(A) + \beta f_{\text{Tik}}(X), \tag{4}$$

where $\alpha > 0$ and $\beta > 0$ are selected constants. The second term penalizes distance of the estimate from the given ROI

$$f_{\text{ROI}}(A) = ||(R \circ R) \circ A||_{1,\mu}, \tag{5}$$

where $\circ$ is Hadamard product and the norm $||.||_{1,\mu}$ is defined as $||X||_{1,\mu} = \sum_i \sum_k \sqrt{X_{i,k}^2 + \mu^2} - \mu$ with positive constant μ. It penalizes pixels of A which are not consistent with the preselected ROIs. The last term is the Tikhonov regularization with preference of smooth solution of X defined as

$$f_{\text{Tik}}(X) = ||\Gamma X||_F^2, \tag{6}$$

where Γ is a lower bidiagonal matrix

$$\Gamma = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ -1 & 1 & \ddots & \vdots \\ 0 & \ddots & \ddots & 0 \\ 0 & 0 & -1 & 1 \end{pmatrix}. \tag{7}$$

The minimum of (4) is found numerically under conditions $A \geq 0$ and $X \geq 0$.

The third related method combining sparse prior on A and convolution parametrization of X was proposed in [22] as the sparse blind source separation and vectorized deconvolution (S-BSS-vecDC). It uses probabilistic interpretation of (2) as being negative likelihood of Gaussian distribution and sparsity on the matrix A is imposed using the ARD principle [11]

$$p\left(A_{i,k}|\xi_{i,k}\right) = t\mathcal{N}\left(A_{i,k}|0, \xi_{i,k}^{-1}\right), \tag{8}$$

$$p\left(\xi_{i,k}\right) = \mathcal{G}\left(\xi_{i,k}|\phi_0, \psi_0\right), \tag{9}$$

where $t\mathcal{N}$ denotes Gaussian distribution truncated to the positive support (representing non-negativity), $\mathcal{G}$ denotes gamma distribution, and ϕ_0 and ψ_0 are prior constant selected as 10^{-10} yielding non-informative prior and improving numerical stability. This prior is closely related to the conventional L1 penalization [29]. Finally, the prior on X was defined using convolution reparametrization of each column of X as $\mathbf{x}_k = \mathbf{b} * \mathbf{u}_k$, where $*$ denotes convolution, $\mathbf{b} \in \mathbf{R}^{n \times 1}$ is an input function common to all sources and $\mathbf{u}_k$ is a source specific convolution kernel with sparsity prior on both, $\mathbf{b}$ and $\mathbf{u}_k$. Due to imposed non-negativeness of the convolution kernel, such prior also promotes smooth solutions. The posterior estimates are obtained using the VB approximation [30] leading to a set of implicit equations which needs to be solved iteratively.

II. Non-negative matrix factorization with adaptive prior covariance

The Bayesian approach to NMF is based on interpretation of (2) as the negative logarithm of isotropic Gaussian likelihood of the data matrix D [31], [22], in the form

$$\log p\left(D|A, X, \omega\right) \propto \frac{pn}{2} \ln \omega - \frac{\omega}{2} ||D - AX^T||_F^2, \tag{10}$$

where symbol $\propto$ denotes equality up to normalizing constant. (10) deviates from (2) and (3) in introduction of the precision parameter ω which is also considered to be unknown. While its value does not change the minimum of (10) it is essential for establishing the weighting of the likelihood with respect to the prior. The prior model of ω is selected to be conjugate to (10), which is the Gamma distribution, $p(\omega) = \mathcal{G}(\vartheta_0, \rho_0)$, with prior constants selected non-informatively as 10^{-10} . The prior model for the matrix A is (8)–(9) favoring sparse solution.

A. Adaptive sparse and smooth prior on coefficient matrix

Let $\mathbf{x}_1, \ldots, \mathbf{x}_r$ be columns of matrix X with truncated Gaussian prior distribution with positive support

$$p\left(\mathbf{x}_k|\Sigma_{\mathbf{x},k}\right) = t\mathcal{N}\left(\mathbf{x}_k|\mathbf{0}_{n,1}, \Sigma_{\mathbf{x},k}\right) \propto -\frac{1}{2}\mathbf{x}_k^T \Sigma_{\mathbf{x},k}^{-1} \mathbf{x}_k. \tag{11}$$

We would like to point out that common assumptions of sparsity and smoothness correspond to different choices of matrix $\Sigma_{\mathbf{x},k}$. The sparsity assumption using ARD principle applied on each element of X corresponds to diagonal covariance matrix $\Sigma_{\mathbf{x},k} = \text{diag}(\mathbf{v}_k)^{-1}$, where elements of vector $\mathbf{v}_k$ are assumed to have Gamma prior model $p(\mathbf{v}_k) = \mathcal{G}(\alpha_0, \beta_0)$. The smoothness assumption is obtained by penalizing differences in $\mathbf{x}_k$, i.e. $\Gamma\mathbf{x}_k$ as in (6)–(7). Interpreting (6) as negative log probability, it is equivalent to (11) with assignment $\Sigma_{\mathbf{x},k}^{-1} = \beta\Gamma\Gamma^T$ which is identical for all columns of matrix X.

In this letter, we propose to model $\mathbf{x}_k$ using (11) with covariance matrix $\Sigma_{\mathbf{x},k} = L_k \mathrm{diag}(\mathbf{v}_k)^{-1} L_k^T$, where lower diagonal matrix

$$L_k = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ l_{1,k} & 1 & \ddots & \vdots \\ 0 & \ddots & \ddots & 0 \\ 0 & 0 & l_{n-1,k} & 1 \end{pmatrix}, \tag{12}$$

has unknown entries on the sub-diagonal vector $\mathbf{l}_k = [l_{1,k}, \ldots, l_{n-1,k}]$. These will be estimated from the data using Gaussian prior distribution for $j = 1, \ldots, n-1, k = 1, \ldots, r$,

$$p(l_{j,k}|\varphi_{j,k}) = \mathcal{N}\left(l_{j,k}|l_0, \varphi_{j,k}^{-1}\right), \tag{13}$$

where l_0 is a chosen prior mean value and $\varphi_{j,k}$ is a precision hyper-parameter with prior $p(\varphi_{j,k}) = \mathcal{G}(\varphi_{j,k}|\zeta_0, \eta_0)$. The prior parameters l_0, ζ_0, η_0 are common for all j, k and good performance was reported for linear problems [26] for choices in range $l_0 \in [-1, 0]$, where 0 favors sparse and -1 favors smooth solution. We select $l_0 = -0.7$ which slightly favors smooth solution. Expected deviation from this value is controlled by ζ_0, η_0. We use values $\zeta_0, \eta_0 = 10^{-2}$ which allows for variation in the sufficient range circa $l_0 \pm 100$. Significantly higher values of ζ_0, η_0 result in posterior estimates closer to l_0. However, significantly lower values of ζ_0, η_0 results in higher sensitivity to local extremes and potentially numerical instability. This setting was found to be a reasonable default for all our experiments and small varying of these parameters do not affect the results significantly, however, it can be arbitrary changed for different applications.

B. Variational Bayes inference

Likelihood (10), together with priors (8)–(9) and (11)–(13) form a probabilistic model with unknown parameters A, X, ω, $\xi_{i,k}$, $\mathbf{v}_k$, $\mathbf{l}_k$, $\boldsymbol{\varphi}_k$, $k = 1, \ldots, r$. Since exact inference of these parameters is intractable, we apply the variational Bayes (VB) method [30] to obtain approximate estimates. Specifically, we seek an approximate posterior distribution in the form of product of factors q:

$$p(\omega, A, \xi_{i,k}, \mathbf{x}_k, \mathbf{v}_k, \mathbf{l}_k, \varphi_{j,k}|D) \approx q(\omega) \prod_{i=1}^{p} q(\overline{\mathbf{a}}_i) \times \prod_{i=1}^{p} \prod_{k=1}^{r} q(\xi_{i,k}) \prod_{k=1}^{r} q(\mathbf{x}_k) \prod_{k=1}^{r} q(\mathbf{v}_k) \prod_{k=1}^{r} q(\mathbf{l}_k) \prod_{k=1}^{r} \prod_{j=1}^{n-1} q(\varphi_{j,k}), \tag{14}$$

where $\overline{\mathbf{a}}_i$ denotes the ith row of matrix A.

By minimizing the Kullback-Leibler divergence between (14) and the true posterior distributions, the best approximate factors of the posterior distributions $\widetilde{q}$ are found using the VB method and are summarized in Tab. I. Together with standard moments of the respective distributions, they form a set of implicit equations. Solution of this set can be found iteratively using Algorithm 1, where one possible initialization is also given. The method will be referred to as the Bayesian non-negative matrix factorization with adaptive prior covariance (NMF-APC) and is available for download from http://www.utia.cz/AS/softwaretools/image_sequences/.

Algorithm 1 VB algorithm for NMF-APC method.

1) Initialization:
   a) Set prior constants: $\vartheta_0, \rho_0, \phi_0, \psi_0, \alpha_0, \beta_0 = 10^{-10}$, $\zeta_0, \eta_0 = 10^{-2}$, and $l_0 = -0.7$.
   b) Set the model order r.
   c) Initialize starting parameters: $\langle\omega\rangle = 1$, $\langle A\rangle = \mathbf{1}_{p,r}$, $\langle L\rangle = I_{nr}$, and $\langle\mathbf{v}\rangle = \mathbf{1}_{nr,1}$.
2) Iterate until convergence is reached:
   a) Eval. $\widetilde{q}(\mathbf{x}_k)$ and subsequent $\widetilde{q}(\mathbf{v}_k), \widetilde{q}(\mathbf{l}_k), \widetilde{q}(\varphi_{j,k})$.
   b) Evaluate $\widetilde{q}(\overline{\mathbf{a}}_i)$ and subsequent $\widetilde{q}(\xi_{i,k})$.
   c) Evaluate $\widetilde{q}(\omega)$.
3) Report resulting estimates $\langle A\rangle$ and $\langle X\rangle$.

III. Application to dynamic renal scintigraphy

The proposed NMF-APC method will be compared with three competing methods (Section I-A) on a large clinical dataset from dynamic renal scintigraphy [28]. In dynamic scintigraphy, the radio-tracer is applied into the bloodstream and the activity of the tracer in the body is recorded by a scintillation camera. The obtained 2D projections of the activity are stored in the form of matrices with counts of particles recorded at each pixel. This can be done repetitively to obtain a dynamic sequence. For ease of processing, images of the sequence are stored columnwise in the matrix D. Due to 2D projection, each pixel in the recorded scintigraphic image is a linear combination of contributions from several biological tissues weighted by the activity of each tissue in given time [17], [22]. The NMF model(1) is a good representation of this sequence, with the following interpretation: columns of matrix A are images of the sources of the radiation (tissues), and columns of coefficient matrix X are the corresponding time activity curves (TAC) of tissues.

A dataset of 98 dynamic image sequences with non-zero kidney activity is available [28]. Each sequence consists of 180 images ($n = 180$) taken with 10 seconds temporal resolution and with 128×128 image size. From these sequences, 196 sequences of kidney (left and right) were extracted using rectangular window of the size 47×37 ($p = 1739$) and time-activity curve (TAC) was extracted for each of the 196 kidneys by an experienced physician using standard procedures [32]. These curves will serve as the ground truth in our experiment.

Each sequence is analyzed using the proposed NMF-APC method and the RUDUR, NMF, and S-BSS-vecDC methods. The noise in the sequence is Poisson distributed, therefore, scaling [33] is applied before processing. All methods have the same initial conditions. The preselected number of sources is four ($r = 4$) which is a common assumption [17], [22] since each sequence is assumed to accumulate activity from the parenchyma (outer part of kidney), pelvis (inner part of kidney), other tissues activity, and blood background. The tuning parameters of the RUDUR method are selected as recommended by the authors of RUDUR to $\alpha = 1$, $\beta = 10$, and $\gamma = 3$. The regions of interests (ROI) required by the RUDUR method in (3) and (5) were chosen to be the ROI of the parenchyma provided by the expert, and non-informative different ROIs for other tissues, see Fig. 1 for an example.

Table I
APPROXIMATE POSTERIOR DISTRIBUTIONS OF THE PROPOSED METHOD ACCOMPANIED BY THEIR SHAPING PARAMETERS.

$\widetilde{q}(\omega) = \mathcal{G}(\vartheta, \rho)$	$\vartheta = \vartheta_0 + \frac{pn}{2}$	$\rho = \rho_0 + \frac{1}{2}\mathrm{tr}\left(\langle (D - AX^T)^T (D - AX^T)\rangle\right)$
$\widetilde{q}(\overline{\mathbf{a}}_i) = t\mathcal{N}(\mu_{\overline{\mathbf{a}}_i}, \Sigma_{\overline{\mathbf{a}}_i})$	$\mu_{\overline{\mathbf{a}}_i} = \Sigma_{\overline{\mathbf{a}}_i}\langle\omega\rangle \sum_{j=1}^{n} D_{i,j}\langle\overline{\mathbf{x}}_j\rangle$	$\Sigma_{\overline{\mathbf{a}}_i} = \left(\langle\omega\rangle\langle X^T X\rangle + \mathrm{diag}(\left\langle\overline{\boldsymbol{\xi}}_i\right\rangle)\right)^{-1}$, $i = 1, \ldots, p$
$\widetilde{q}(\xi_{i,k}) = \mathcal{G}(\phi_{i,k}, \psi_{i,k})$	$\phi_{i,k} = \phi_0 + \frac{1}{2}$	$\psi_{i,k} = \psi_0 + \frac{1}{2}\left\langle A_{i,k}^2\right\rangle$, $i = 1, \ldots, p$, $k = 1, \ldots, r$
$\widetilde{q}(\mathbf{x}) = t\mathcal{N}(\mu_{\mathbf{x}}, \Sigma_{\mathbf{x}})$	$\mu_{\mathbf{x}} = \Sigma_{\mathbf{x}}\langle\omega\rangle \mathrm{vec}\left(D^T\langle A\rangle\right)$	$\Sigma_{\mathbf{x}} = \left(\langle\omega\rangle\left(\langle A^T A \otimes I_n\rangle\right) + \langle L\mathrm{diag}(\mathbf{v})L^T\rangle\right)^{-1}$
$\widetilde{q}(v_{j,k}) = \mathcal{G}(\alpha_{i,k}, \beta_{j,k})$	$\alpha_{i,k} = \alpha_0 + \frac{1}{2}$	$\beta_{j,k} = \beta_0 + \frac{1}{2}\langle L_k^T \mathbf{x}_k^T \mathbf{x}_k L_k\rangle_{i,i}$, $j = 1, \ldots, n-1$
$\widetilde{q}(l_{j,k}) = t\mathcal{N}(\mu_{l_{j,k}}, \sigma_{l_{j,k}})$	$\mu_{l_{j,k}} = \sigma_{l_{j,k}}\left(-\langle v_{j,k}\rangle\langle x_{j,k}x_{j+1,k}\rangle + l_0\langle\varphi_{j,k}\rangle\right)$	$\sigma_{l_{j,k}} = \left(\langle v_{j,k}\rangle\left\langle x_{j+1,k}^2\right\rangle + \langle\varphi_{j,k}\rangle\right)^{-1}$, $j = 1, \ldots, n-1$
$\widetilde{q}(\varphi_{j,k}) = \mathcal{G}(\zeta_{j,k}, \eta_{j,k})$	$\zeta_{j,k} = \zeta_0 + \frac{1}{2}$	$\eta_{j,k} = \eta_0 + \frac{1}{2}\langle(l_{j,k} - l_0)^2\rangle$, $j = 1, \ldots, n-1$

vec() is the vectorization operator, $\mathrm{vec}(X) = [\boldsymbol{x}_1^T, \boldsymbol{x}_2^T, \ldots, \boldsymbol{x}_r^T]^T$, L is block diagonal matrix $L = \mathrm{diag}(L_1, \ldots, L_r)$, and $\mathbf{v} = [\mathbf{v}_1, \ldots, \mathbf{v}_r]$.

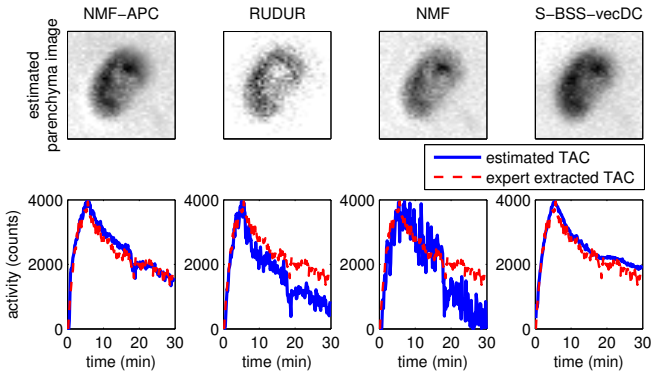

Figure 2. Results of separation of the left parenchyma for a selected sequence using (from the left): the proposed NMF-APC, RUDUR [17], NMF [2], and S-BSS-vecDC [22] methods.

A. Selected result of separation

An example of separation of the parenchyma for a selected sequence is given in Fig. 2. The remaining separated sources can be displayed using example run of the NMF-ACP method available online. In Fig. 2, the estimated parenchyma images are displayed in the first row and the estimated time-activity curves (TAC) are in the second row using blue lines while the ground truth TACs are displayed using dashed red lines.

The estimated TACs in Fig. 2 clearly demonstrate the influence of the selected prior model on the result. While the TAC provided by the original NMF method is rather noisy, the RUDUR method provides smoother results. However, the resulting TAC approaches lower values in case of low signal to noise ratio. The S-BSS-vecDC provides smooth solution which is a result of the convolution model of the TAC. However, it tends to slightly overestimate the ground truth curves. The result of the proposed NMF-APC method shows its ability to reflect abrupt changes, notable at the beginning of the TAC and its peak, as well as to reflect the smooth parts of the TAC. We stress that the NMF-APC follows the ground truth curve closer than the competitors. The estimated parenchyma images are similar for all methods except the RUDUR method which has clearer parenchyma tissue due to availability of the expert selected ROI of the parenchyma as shown in Fig. 1.

B. Statistical evaluation

Statistical comparison of all 196 estimated parenchyma TACs is given in Tab. II using normalized mean absolute error (NMAE) and normalized mean square error (NMSE) with their standard deviations. The normalization ensures that the TACs are scaled to interval $[0; 1]$ so that the results with different maximum of activity can be compared.

Table II
ESTIMATION ERRORS ON 196 PARENCHYMA TIME-ACTIVITY CURVES VIA THE NMAE AND THE NMSE CRITERIA, AND P-VALUES OF T-TEST OF IMPROVEMENT OVER NMF-APC.

Method	NMAE	p-value	NMSE	p-value
NMF-APC	0.108±0.097	–	0.029±0.058	–
RUDUR [17]	0.162±0.075	$< 10^{-5}$	0.049±0.043	$< 10^{-5}$
NMF [2]	0.203±0.081	$< 10^{-5}$	0.074±0.049	$< 10^{-5}$
S-BSS-vecDC [22]	0.118±0.084	0.0416	0.028±0.045	0.5075

Tab. II demonstrate that the NMF-APC method outperforms all other methods in proximity of the parenchyma TACs to the ground truth TACs in the sense of NMAE criterion, although with higher standard deviations. Therefore, we test also statistical significance using two-sided t-tests (with null hypothesis that the differences between results of the NMF-APC method and other methods comes from a normal distribution with zero mean and unknown variance) where improvement is proved with a p-value less then 0.05 in all cases. The NMF-APC method is comparable with the S-BSS-vecDC method in the sense of NMSE criterion; however, it is much simpler and computationally cheaper (one sequence: NMF-APC 28.6 s, RUDUR 13.7 s, NMF 0.3 s, S-BSS-vecDC 290.2 s).

IV. CONCLUSION

In this contribution, we proposed to complement the probabilistic model of non-negative matrix factorization by a sparsity and smoothness prior based on general covariance matrix structure. Since the prior allows to use the variational Bayes approximation, it is computationally faster than methods based on Monte Carlo sampling. We have shown the ability of the prior to represent smooth curves with abrupt changes on the example from dynamic renal scintigraphy where the time activity curves have such structure. On a clinical dataset from dynamic renal scintigraphy consisting of 196 image sequences, we have demonstrated that the novel method outperforms other state-of-the-art methods in the task of blind separation of a clinically important source.

ACKNOWLEDGMENT

Support of the Czech Science Foundation project GA18-07247S is gratefully acknowledged.

REFERENCES

[1] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[2] D. Lee and H. Seung, "Algorithms for non-negative matrix factorization," in *Advances in neural information processing systems*, pp. 556–562, 2001.

[3] A. Adler, M. Elad, and Y. Hel-Or, "Probabilistic subspace clustering via sparse representations," *IEEE Signal Processing Letters*, vol. 20, no. 1, pp. 63–66, 2013.

[4] C.-H. Lin, F. Ma, C.-Y. Chi, and C.-H. Hsieh, "A convex optimization-based coupled nonnegative matrix factorization algorithm for hyperspectral and multispectral data fusion," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 3, pp. 1652–1667, 2018.

[5] Y. Cavalcanti, T. Oberlin, N. Dobigeon, S. Stute, M. Ribeiro, and C. Tauber, "Unmixing dynamic PET images with variable specific binding kinetics," *Medical Image Analysis*, vol. 49, pp. 117–127, 2018.

[6] C.-H. Lin, R. Wu, W.-K. Ma, C.-Y. Chi, and Y. Wang, "Maximum volume inscribed ellipsoid: A new simplex-structured matrix factorization framework via facet enumeration and convex optimization," *SIAM Journal on Imaging Sciences*, vol. 11, no. 2, pp. 1651–1679, 2018.

[7] J. Hinrich and M. Mørup, "Probabilistic sparse non-negative matrix factorization," in *International Conference on Latent Variable Analysis and Signal Separation*, pp. 488–498, Springer, 2018.

[8] P. Hoyer, "Non-negative matrix factorization with sparseness constraints," *The Journal of Machine Learning Research*, vol. 5, pp. 1457–1469, 2004.

[9] Q. Sun, J. Lu, Y. Wu, H. Qiao, X. Huang, and F. Hu, "Non-informative hierarchical Bayesian inference for non-negative matrix factorization," *Signal Processing*, vol. 108, pp. 309–321, 2015.

[10] H. Chung, E. Plourde, and B. Champagne, "Discriminative training of NMF model based on class probabilities for speech enhancement," *IEEE Signal Processing Letters*, vol. 23, no. 4, pp. 502–506, 2016.

[11] C. Bishop, "Variational principal components," *IET Conference Proceedings*, pp. 509–514(5), 1999.

[12] V. Šmídl and A. Quinn, "On Bayesian principal component analysis," *Computational statistics & data analysis*, vol. 51, no. 9, pp. 4101–4123, 2007.

[13] V. Tan and C. Fevotte, "Automatic relevance determination in nonnegative matrix factorization with the β-divergence," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 7, pp. 1592–1605, 2013.

[14] T. Brouwer, J. Frellsen, and P. Lió, "Comparative study of inference methods for Bayesian nonnegative matrix factorisation," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 513–529, Springer, 2017.

[15] M. Kohjima and S. Watanabe, "Phase transition structure of variational Bayesian nonnegative matrix factorization," in *International Conference on Artificial Neural Networks*, pp. 146–154, Springer, 2017.

[16] B. Gao, H. Zhang, W. Woo, G. Tian, L. Bai, and A. Yin, "Smooth nonnegative matrix factorization for defect detection using microwave nondestructive testing and evaluation," *IEEE Transactions on Instrumentation and Measurement*, vol. 63, no. 4, pp. 923–934, 2014.

[17] M. Filippi, M. Desvignes, and E. Moisan, "Robust unmixing of dynamic sequences using regions of interest," *IEEE Transactions on Medical Imaging*, vol. 37, no. 1, pp. 306–315, 2018.

[18] R. Zdunek and A. Cichocki, "Blind image separation using nonnegative matrix factorization with Gibbs smoothing," in *International Conference on Neural Information Processing*, pp. 519–528, Springer, 2007.

[19] V. Cheung, K. Devarajan, G. Severini, A. Turolla, and P. Bonato, "Decomposing time series data by a non-negative matrix factorization algorithm with temporally constrained coefficients," in *Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE*, pp. 3496–3499, IEEE, 2015.

[20] T. Yokota, R. Zdunek, A. Cichocki, and Y. Yamashita, "Smooth nonnegative matrix and tensor factorizations for robust multi-way data analysis," *Signal Processing*, vol. 113, pp. 234–249, 2015.

[21] T. Yokota, Q. Zhao, and A. Cichocki, "Smooth PARAFAC decomposition for tensor completion," *IEEE Transactions on Signal Processing*, vol. 64, no. 20, pp. 5423–5436, 2016.

[22] O. Tichý and V. Šmídl, "Bayesian blind separation and deconvolution of dynamic image sequences using sparsity priors," *IEEE Transaction on Medical Imaging*, vol. 34, pp. 258–266, January 2015.

[23] O. Tichý and V. Šmídl, "Non-parametric Bayesian models of response function in dynamic image sequences," *Computer Vision and Image Understanding*, vol. 151, pp. 90–100, 2016.

[24] M. Kyung, J. Gill, M. Ghosh, and G. Casella, "Penalized regression, standard errors, and Bayesian lassos," *Bayesian Analysis*, vol. 5, no. 2, pp. 369–411, 2010.

[25] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight, "Sparsity and smoothness via the fused lasso," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 1, pp. 91–108, 2005.

[26] O. Tichý, V. Šmídl, R. Hofman, and A. Stohl, "LS-APC v1.0: a tuning-free method for the linear inverse problem and its application to source-term determination," *Geoscientific Model Development*, vol. 9, no. 11, pp. 4297–4311, 2016.

[27] J. Ševčík, V. Šmídl, and F. Šroubek, "An adaptive correlated image prior for image restoration problems," *IEEE Signal Processing Letters*, vol. 25, no. 7, pp. 1024–1028, 2018.

[28] M. Šámal and J. Valoušek, "Clinically documented data set of dynamic renal scintigraphy for clinical audits and quality assurance of nuclear medicine software," in *European journal of nuclear medicine and molecular imaging*, vol. 39, pp. S170–S171, Springer, 2012.

[29] D. Wipf and S. Nagarajan, "A new view of automatic relevance determination," in *Advances in neural information processing systems*, pp. 1625–1632, 2008.

[30] V. Šmídl and A. Quinn, *The Variational Bayes Method in Signal Processing*. Springer, 2006.

[31] M. Tipping and C. Bishop, "Probabilistic principal component analysis," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 61, no. 3, pp. 611–622, 1999.

[32] I. Gordon, A. Piepsz, and R. Sixt, "Guidelines for standard and diuretic renogram in children," *European journal of nuclear medicine and molecular imaging*, vol. 38, no. 6, pp. 1175–1188, 2011.

[33] H. Benali, I. Buvat, F. Frouin, J. Bazin, and R. Paola, "A statistical model for the determination of the optimal metric in factor analysis of medical image sequences (FAMIS)," *Physics in medicine and biology*, vol. 38, p. 1065, 1993.