Sparsity in Bayesian Blind Source Separation and Deconvolution

Václav Šmídl, Ondřej Tichý

Institute of Information Theory and Automation, Prague, Czech Republic,
{smidl,otichy}@utia.cas.cz

Abstract. Blind source separation algorithms are based on various separation criteria. Differences in convolution kernels of the sources are common assumptions in audio and image processing. Since it is still an ill posed problem, any additional information is beneficial. In this contribution, we investigate the use of sparsity criteria for both the source signal and the convolution kernels. A probabilistic model of the problem is introduced and its Variational Bayesian solution derived. The sparsity of the solution is achieved by introduction of unknown variance of the prior on all elements of the convolution kernels and the mixing matrix. Properties of the model are analyzed on simulated data and compared with state of the art methods. Performance of the algorithm is demonstrated on the problem of decomposition of a sequence of medical data. Specifically, the assumption of sparseness is shown to suppress artifacts of unconstrained separation method.

1 Introduction

The aim of blind source separation is to recover the original form of signals that can be observed only via their superposition. A classical example of such a situation is the cocktail party problem [9], where multiple speakers are recorded by multiple microphones. The aim is to separate audio signal of the individual speakers. This requires specification of the separability criteria. One such criteria is the assumption of temporal properties of the source, expressed via different convolution kernels [15]. Since the convolution kernels are also unknown, the problem is that of blind deconvolution within blind separation. Algorithms for this problem include optimization of information theoretic measures [4] and the EM algorithm [2]. In the image processing literature, the problem is closely related to the multi-channel blind deconvolution [18].

The presented algorithm was primarily motivated by application in medical image analysis which has the following specific issues: (i) the sources are physiological organs which will be further analyzed by medical experts for final diagnosis, and (ii) poor signal to noise conditions, where a weak signal is hard to separate from the noise. The medical experts expect that the results will respect physiological nature which is very hard to formalize mathematically. The model of source activity by convolution of common input function is one of a few mathematical models that is generally accepted. The assumption of sparsity is also natural in this application. However, such assumptions are not unique to medical imaging and the resulting algorithm may be used in any other application domain.

The poor signal-to-noise conditions of the domain motivated our choice of the Bayesian approach. It has been successfully applied in situations when the number of the sources is lower than the number of the channels, [13]. The ability to marginalize provides an automatic Occam's razor that suppresses the spurious sources and thus provides automatic denoising. Since exact marginalization may not be always possible, approximate methods of Bayesian calculus has been developed. One such formalism is the Variational Bayes method [5,16]. Its use for selection of the number of principal components has been demonstrated in [5], via the use of priors with unknown variance. In connection with the Variational Bayes approximation it favors sparse solutions. Since its introduction, this mechanism has been used in image deconvolution [19], or sparse blind source separation [17]. We introduce this modeling assumption on the convolution kernel and the mixing matrix.

The resulting algorithm is applied to the problem of image sequence decomposition. This problem has been studied independently for many years [3,6], however, it has been recognized as a special case of the blind source separation problem [13,16]. The specific nature of this problem is in interpretation of the resulting components which are further used for medical diagnosis. Even a small improvement in the estimation may have significant impact on the diagnostic quality of the results.

2 Sparsity in Bayesian Analysis

The Bayesian inference is concerned with evaluation of the full posterior density of the parameters θ from the observed data D. It requires a parametric probabilistic model of the data in the form of a probability distribution, $p\left(D|\theta\right)$, conditioned by knowledge of the parameters, θ. The prior state of knowledge of θ is quantified by the *prior* distribution, $p(\theta)$. Our state of knowledge of θ after observing D is quantified by the *posterior distribution*, $p(\theta|D)$. These functions are related via Bayes' rule:

$$p\left(\theta|D\right)=\frac{p\left(\theta,D\right)}{p\left(D\right)}=\frac{p\left(D|\theta\right)p\left(\theta\right)}{\int p\left(D|\theta\right)p\left(\theta\right)d\theta}, \tag{1}$$

where integration in the denominator of (1) is over the whole support of the involved distributions. We will refer to $p\left(\theta,D\right)$ as the joint distribution of parameters and data, or, more concisely, as the *joint distribution.*

2.1 The Variational Bayes Approximation

The Variational Bayes (VB) approximation is a deterministic technique for approximation of the Bayes rule (1), in the sense of the following theorem [16].

Theorem 1. *Let $p(\theta|D)$ be the posterior distribution of multivariate parameter, $\theta = [\theta_1', \theta_2']'$, and $p^*(\theta|D)$ be an approximate distribution restricted to the set of conditionally independent distributions:*

$$p^*(\theta|D) = p^*(\theta_1, \theta_2|D) = p^*(\theta_1|D)\, p^*(\theta_2|D)\,. \tag{2}$$

Any minimum of the Kullback-Leibler divergence from $p^*(\cdot)$ to $p(\cdot)$

$$KL(p^*(\theta|D)\,||p(\theta|D)) = \int p^*(\theta|D) \ln \frac{p(\theta|D)}{p^*(\theta|D)} d\theta, \tag{3}$$

is achieved when $p^*(\cdot) = \tilde{p}$ where

$$\tilde{p}(\theta_i) \propto \exp\left(\mathsf{E}_{\tilde{p}(\theta_{/i})}\left(\ln\left(p\left(\theta, D\right)\right)\right)\right),\ i = 1, 2. \tag{4}$$

Here, $\theta_{/i}$ denotes the complement of θ_i in θ and $\mathsf{E}_{p(\theta)}(g(\theta))$ denotes expected value of function $g(\theta)$ with respect to distribution $p(\theta)$.

Theorem 1 is also known as mean-field approximation [14] and provides a powerful tool for approximation of joint pdfs in *separable form* [16]:

$$\ln p(\theta_1, \theta_2, D) = g(\theta_1, D)'\, h(\theta_2, D)\,. \tag{5}$$

Here, $g(\theta_1, D)$ and $h(\theta_2, D)$ are finite-dimensional vectors. Using (5) in (4),

$$\tilde{p} \propto \exp\left(g(\theta_1, D)'\, \widehat{h(\theta_2, D)}\right), \tag{6}$$

where $\widehat{h(\cdot)} = \mathsf{E}_{\tilde{p}(\theta_2|D)}(h(\cdot))$ are the moments of θ_2, and similarly for θ_1. In cases, where (6) are from exponential family, $h(\cdot)$ form its sufficient statistics [8]. An iterative moment-swapping algorithm is implied [16].

2.2 Automatic Relevance Determination

The mechanism of automatic relevance determination (ARD) is based on joint estimation of the parameters of the prior (hyper-parameters) with the data [5]. Specifically, the prior of an unknown vector parameter θ that is assumed to have elements redundant for the observed data is chosen as

$$p(\boldsymbol{\theta}|\boldsymbol{\omega}) = \mathcal{N}(0, \mathrm{diag}(\boldsymbol{\omega})), \qquad p(\omega_i) = G(\alpha_0, \beta_0),\ \forall i, \tag{7}$$

where $\boldsymbol{\omega}$ is the vector of unknown precisions (inverse variances) of the prior on the parameter $\boldsymbol{\theta}$ and it is assumed to have conjugate Gamma prior with scalar parameters α_0, β_0. The Bayes rule is then used to estimate both $\boldsymbol{\theta}$ and $\boldsymbol{\omega}$. When the parameter is redundant, the expected value of the prior variance ψ approaches zero. This effect is known as the ARD principle and it is demonstrated on the following example.

Example 1 (Multiplicative scalar decomposition). Consider the following model of scalar measurement d being explained as a product of two unknown parameter, a and x:

$$d = ax + e,\ \ e \sim \mathcal{N}(0, r_e). \tag{8}$$

where variance r_e is assumed to be known. The likelihood function of the model parameters is

$$p(d|a, x) = \mathcal{N}(ax, r_e), \tag{9}$$

and has maximum anywhere on the manifold defined by the signal estimate:

$$\widehat{ax} = d. \tag{10}$$

Separation of the signal from the noise is possible only with additional assumptions. One such assumption is the choice of prior on the x variable as $p(x) = \mathcal{N}(0, r_x)$, with a chosen variance r_x. Maximum of the marginal $p(a|d) = \int p(d|a, x, r_e)p(x)dx$ is then

$$\hat{a}_{marg} = \begin{cases} \frac{\sqrt{d^2 - r_e}}{\sqrt{r_x}} & \text{if } d > \sqrt{r_e}, \\ 0 & \text{otherwise.} \end{cases} \tag{11}$$

Note the inference bound on the signal, $d > \sqrt{r_e}$, i.e. the signal should be higher than the standard deviation of the noise. This bound enforces sparsity of the solution since estimates of the parameters for a weak signal are zeros.

The ARD is based on introduction of the hyper-parameters (7) on any variable, a or x, or both. For example, a fixed prior on a, $p(a) = \mathcal{N}(0, \sigma_a)$, and the ARD prior on x, i.e. $p(x|\omega_x) = \mathcal{N}(0, \omega_x^{-1})$, with unknown precision ω_x with prior $p(\omega_x) = G(\alpha_0, \beta_0)$, yields the variational posteriors of the form

$$\tilde{p}(x|d) = \mathcal{N}(\hat{x}, \sigma_x), \qquad \tilde{p}(\omega_x|d) = G(\frac{3}{2}, \gamma_x), \qquad \tilde{p}(a|d) = \mathcal{N}(\hat{a}, \sigma_a) \tag{12}$$

with shaping parameters satisfying the following set of implicit equations:

$$\begin{aligned} \hat{x} &= \frac{\sigma_x}{r_e} d\hat{a}, & \sigma_x &= ((\hat{a}^2 + \sigma_a)r_e^{-1} + \frac{3}{2\gamma_x})^{-1}, \\ \hat{a} &= \frac{\sigma_a}{r_e} d\hat{x}, & \gamma_x &= \sigma_x + \hat{x}^2. \end{aligned} \tag{13}$$

Numerical solution of this set is achieved by the iterative algorithm [1]. We choose an initial value of $\sigma_x^{(0)}$ and $\hat{a}^{(0)}$ and then iteratively evaluate equations (13) in the order: $\hat{x}$, γ_x, σ_x, $\hat{a}$.

We note the following:

- The choice of σ_a fixes the value of $\hat{a}$ at a constant for all significant values of d, and the free parameter that grows with d is the $\hat{x}$.

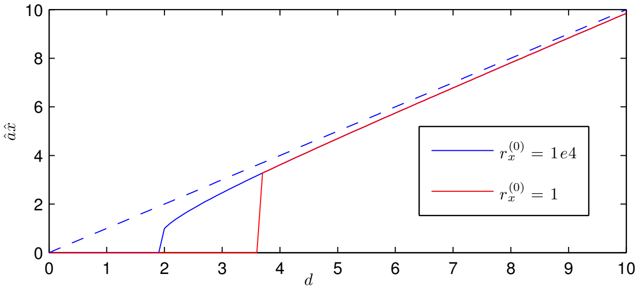

Fig. 1. Product of expected values $\hat{a}$ and $\hat{x}$ of variational posteriors from (12) for two initial conditions of $\sigma_x^{(0)}$. The dashed line denotes maximum likelihood solution (10).

- The product of the $\hat{a}\hat{x}$ is displayed in Fig. 1 for a range of values d. Note the presence of the inference bound similar to (11). In this case, the estimates are zeros for $d < 2\sqrt{r_e}$, i.e. the ARD property of the VB inference enforces sparse estimates more aggressively than the marginalization. We conjecture that this is a consequence of the variance underestimation of the VB approximation [12].
- The converged results are insensitive to the choice of the $\hat{a}^{(0)}$ parameter, and to some extent even to the $\sigma_x^{(0)}$ parameter. The hyper-parameters α, β of both variables were set to zero to yield the Jeffreys' prior. For $\sigma_x^{(0)} > 1$ the results correspond to those of $\sigma_x^{(0)} = 1e4$ in Fig. 1. However, the converged values differ for $\sigma_x^{(0)} \leq 1$, Fig. 1, which illustrates the existence of local minima in the VB procedure [16].
- The Variational PCA [5] is a multivariate extension of this model with ARD applied on the columns of the mixing matrix.

Remark 1 (Symmetric ARD). It is possible to introduce ARD on both variables a and x. However, reliable estimation is achieved only with enforced positivity of a and x via truncated Normal prior. In this case, the estimation results are closely similar to those in Fig. 1.

3 Blind Source Separation and Deconvolution

The task of blind source separation arise when the observed signal is assumed to be a superposition of the source signal. In this Section we assume that the source signals are generated via convolution of the common input function with source-specific kernels.

3.1 Signal Superposition and Convolution

The basic formulation of the blind source separation assumes that the vector of observations at time t, $\mathbf{d}_t$ is a linear superposition of all source signals at the same time, $\overline{\mathbf{x}}_t$:

$$\mathbf{d}_t = A\overline{\mathbf{x}}_t, \tag{14}$$

where A is the mixing matrix, $\overline{\mathbf{x}}_t = [x_{t,1}, \ldots, x_{t,r}]$, and r is the number of sources. The number of observations is p and the length of the source signal is n. The full observed sequence can be written in matrix notation as:

$$D = AX^{'}, \tag{15}$$

where $D = [\mathbf{d}_1, \ldots, \mathbf{d}_n]$, and the columns of matrix X are the source signals $X = [\mathbf{x}_1, \ldots, \mathbf{x}_r]$ ($\overline{\mathbf{x}}_t$ are the rows of matrix X). In this case, we assume the number of sources r to be unknown with conditions $r < n$, $r < p$.

The kth source is assumed to be the result of convolution of the common input function $\mathbf{b}$, and source-specific convolution kernels $\mathbf{u}_k$:

$$\mathbf{x}_k = \mathbf{b} * \mathbf{u}_k = B\mathbf{u}_k, \tag{16}$$

where matrix B is defined as follows:

$$B = \begin{pmatrix} b_1 & 0 & 0 & 0 \\ b_2 & b_1 & 0 & 0 \\ \ldots & b_2 & b_1 & 0 \\ b_n & \ldots & b_2 & b_1 \end{pmatrix}. \tag{17}$$

The full model of the data is thus

$$D = AX' = AU'B', \tag{18}$$

where $U = [\mathbf{u}_1, \ldots, \mathbf{u}_k]$ and all parameters A, U, B are unknown. In the sequel, we will assume that all these parameters are positive. This is motivated by the application area in image analysis.

3.2 Probabilistic Model

The deterministic assumptions in the previous Section are valid only approximately. For example, the data vectors $\mathbf{d}_t$ are subject to the observation noise. The observed data $\mathbf{d}_t$ are thus modeled as random realizations from the probability density function. For Gaussian distributed noise, the matrix of observations is assumed to be distributed as

$$p(D|A, B, U, \omega) = \mathcal{N}(AU'B', \omega^{-1}I_p \otimes I_n) = \prod_{t=1}^{n} \mathcal{N}(A\overline{\mathbf{x}}_t, \omega^{-1}I_p), \tag{19}$$

where I_p denotes identity matrix of size $p \times p$, $\mathcal{N}(.,.)$ of matrix argument denotes the matrix normal distribution [16] and symbol $\otimes$ denotes the Kronecker product. Prior distributions for all unknown parameters A, B, U, ω need to be specified.

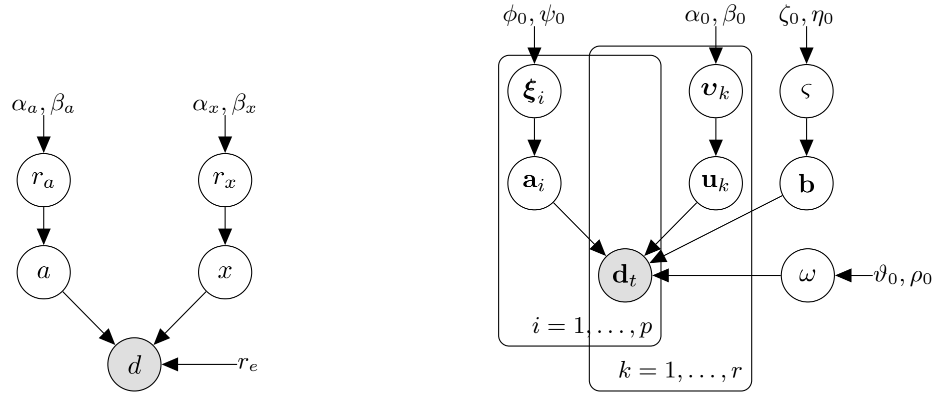

Fig. 2. Graphical model of the scalar multiplicative decomposition from Remark 1 (left) and the proposed sparse blind source separation and deconvolution (right).

The parameter ω is a precision parameter of a Gaussian density and thus it has a conjugate prior in the form of Gamma density

$$p(\omega) = \mathrm{G}(\vartheta_0, \rho_0),$$

with chosen constants ϑ_0, ρ_0. These may be chosen to approach $\vartheta_0 \to 0, \rho_0 \to 0$ yielding an uninformative Jeffrey's prior on the scale parameter [10]. The input function $\mathbf{b}$ is assumed to have all positive elements. This assumption is modeled by Normal distributed prior with its support truncated to positive values (Appendix B.1), the truncated normal distribution is denoted $t\mathcal{N}()$ with the same arguments the Normal distribution since the truncation interval is always $\langle 0, \infty \rangle$. The precision of the prior is also unknown:

$$p(\mathbf{b}|\varsigma) = t\mathcal{N}(0, \varsigma^{-1} I_n), \qquad p(\varsigma) = \mathrm{G}(\zeta_0, \eta_0). \tag{20}$$

The only assumption on the mixing matrix and the convolution kernels is sparsity. In both cases it will be achieved by the ARD property (Section 2.2). Specifically, the ARD prior (7) is used for all elements of matrices A and U. In vector notation, the ARD corresponds to a variance with unknown diagonal:

$$p(\mathbf{u}_k|\boldsymbol{\upsilon}_k) = t\mathcal{N}(0_{n,1}, \mathrm{diag}(\boldsymbol{\upsilon}_k)^{-1}), \tag{21}$$

$$p(\upsilon_{j,k}) = \mathrm{G}(\alpha_{jk,0}, \beta_{jk,0}), \;\; j = 1, \dots, n. \tag{22}$$

Here, diag(.) denotes a matrix with the argument vector in its diagonal and zeros otherwise, and $\upsilon_{j,k}$ are elements of $\boldsymbol{\upsilon}_k$. For notational convenience, we define prior on the rows of matrix A, $\overline{\mathbf{a}}_i, i = 1, \dots p$.

$$p(\overline{\mathbf{a}}_i|\boldsymbol{\xi}_i) = t\mathcal{N}(\mathbf{0}_{1,r}, \mathrm{diag}(\boldsymbol{\xi}_i)^{-1}), \qquad p(\xi_i) = \prod_{k=1}^{r} \mathrm{G}(\phi_{ik,0}, \psi_{ik,0}). \tag{23}$$

The joint distribution of the data is then

$$p(D,A,\mathbf{b},U,\omega)=p(D|A,\mathbf{b},U,\omega)\prod_{i=1}^{p}\left[p(\overline{\mathbf{a}}_i|\boldsymbol{\xi}_i)p(\boldsymbol{\xi}_i)\right] \\ \prod_{k=1}^{r}\left[p(\mathbf{u}_k|\boldsymbol{\upsilon}_k)p(\boldsymbol{\upsilon}_k)\right]p(\mathbf{b}|\varsigma)p(\varsigma)p(\omega). \tag{24}$$

Graphical model of (24) is displayed in Fig. 2. The model differs from the Variational PCA [5] and its positive version [13] in the form where the ARD is applied. While one ARD parameter is common to the whole column of matrix A in the former, every element of the matrices A and U has its own relevance determination parameter in our model. The model has thus much more parameters to estimate from the data.

3.3 The Variational Bayes Posterior

We seek the variational solution in the same form as in (24). The variational posterior distributions (4) for model (24) are found to have functional form:

$$\tilde{p}(\mathbf{u}_k|D,r)=t\mathcal{N}\left(\mu_{\mathbf{u}_k},\mathit{\Sigma}_{\mathbf{u}_k}\right), \qquad \tilde{p}(\boldsymbol{\upsilon}_k|D,r)=\prod_{j=1}^{n}\mathrm{G}(\alpha_{jk},\beta_{jk}), \tag{25}$$

$$\tilde{p}(\mathbf{b}|D,r)=t\mathcal{N}\left(\mu_{\mathbf{b}},\mathit{\Sigma}_{\mathbf{b}}\right), \qquad \tilde{p}(\varsigma|D,r)=\mathrm{G}(\zeta,\eta), \tag{26}$$

$$\tilde{p}(\overline{\mathbf{a}}_i|D,r)=t\mathcal{N}\left(\mu_{\mathbf{a}_i},\mathit{\Sigma}_{\mathbf{a}_i}\right), \qquad \tilde{p}(\boldsymbol{\xi}_i|D,r)=\prod_{k=1}^{r}\mathrm{G}(\phi_{ik},\psi_{ik}), \tag{27}$$

$$\tilde{p}(\omega|D,r)=\mathrm{G}(\vartheta,\rho). \tag{28}$$

The shaping parameters of the posterior distributions are given in Appendix Appendix A:. Together with moments of distributions (25)–(28) they form a set of implicit equations that needs to be solved.

3.4 Iterative Solution

Solution of the implicit set of equations in Appendix A is found using the variation iterative algorithm [1,16]. The algorithm is based on sequential evaluation of the shaping parameters in Appendix A in the following order: 1) image sources $\mathbf{a}_i,\boldsymbol{\xi}_i$, 2) convolution kernels $\mathbf{u}_k,\boldsymbol{\upsilon}_k$, 3) input function $\mathbf{b},\varsigma$, 4) noise precision ω. This order was found to yield the fastest convergence.

Since the Variational Bayes approximation contains local minima, initialization of the iterative algorithm is critical. Similarly to the scalar decomposition example, we set values of all Gamma hyper-parameters, $\phi_0,\psi_0,\alpha_0,\beta_0,\zeta_0,\eta_0,\vartheta_0,\rho_0$, to 10^{-10} to yield uninformative prior. The most sensitive parameter to initialization is the input function $\mathbf{b}$. We propose to initialize the iterative algorithm at $\mathbf{b}^{(0)}=[1,0,\ldots,0]$, for which the matrix B is the identity matrix and the convolution kernels have the role of the sources. This point is known to be an important local extrema in the image deconvolution problems. The convolution kernels $\mathbf{u}_k$ were initialized randomly.

Care is needed with numerical implementation of the iterative algorithm. Specifically, when eigenvalues of the inverted matrices in (29) and (31) are almost equal, the resulting estimates of the convolution kernels contain artifacts (jagged curves). This have been prevented by the use of pseudo-inverse with removal of the smallest eigenvalues. The jagging effect is also suppressed by using (41) to estimate the second moment of $p(U)$.

The resulting algorithm will be denoted as Sparse Blind Source Separation and DeConvolution (S-BSS-DC). It is implemented in matlab and can be downloaded from: `http://www.utia.cas.cz/AS/softwaretools/image_sequences`

4 Results

In this Section, we study properties of the proposed algorithm on a simulated data and demonstrate its practical use on the data from dynamic medical imaging. In both cases, we use a non-standard interpretation of blind source separation which is now briefly introduced.

4.1 Image Sequence Decomposition

The blind source separation model (14) has been used in image sequence analysis for a long time, usually as a model of principal components [3]. The interpretation of the model parameters is slightly different from the cocktail party problem. The observation $\mathbf{d}_t$ is a vector of pixels of the image observed at time t, where the pixels are stored column-wise. The columns of matrix A are images of activity (e.g. measured by PET, SPECT or fMRI) of the underlying biological organs stored in the same form as pixels of $\mathbf{d}_t$. The elements of $\overline{\mathbf{x}}_t$ are activities of the underlying images at time t. The columns $\mathbf{a}_k$ of the matrix A and the source vectors $\mathbf{x}_k$ are thus considered to belong to each other, where the $\mathbf{a}_k$ is the image of the biological organ and $\mathbf{x}_k$ its activity in time. These will be denoted as *source images* and *source curves*, respectively.

This problem has been addressed by the Variational Bayes approach e.g. in [13,16]. The Variational Bayes method of image decomposition with positivity constraints and ARD on image sources was proposed in [13] and will be used for comparison under label BSS+. Sparsity of the image has been modeled by mixture priors, where the parameters of the mixture had to be selected [13], or discrete hidden variable [17].

In medical applications, the sources $\mathbf{x}$ correspond to the flow of biological fluids in the organism. This flow can be modeled by a compartment model, which yields model of the source as convolution of the activity of the blood stream and the tissues specific kernels [11,7]. However, the parametric convolution kernels are typically used [21,7]. Parameters of the convolution kernels are very important for estimation of diagnostic coefficients [11]. We will study the use of general convolution kernels with the ARD prior.

4.2 Phantom Study

A synthetic phantom study for the sparse blind source separation and deconvolution was proposed in [7]. The data are generated using three sources with parametric convolution kernel in the parametric form assumed in [7], so that their CAM-CM algorithm can be used to estimate them. The original phantom data are displayed in Fig. 3, top right. Each source curve generated as a convolution between a common input function $\mathbf{b} = \exp(-\frac{t}{3})$ and source-specific convolution kernels, $\mathbf{u}_1 = \exp(-\frac{t}{10})$, $\mathbf{u}_2 = 100\exp(-4t)$, and $\mathbf{u}_3 = \frac{1}{2}\exp(\frac{t}{100})$. Each source image has resolution 50×50 pixels, i.e. $p = 2500$, and the sequence contains 50 images, i.e. $n = 50$.

The generated data intentionally contain many overlapping regions. The common assumption in many image decomposition techniques is that there is at least one pixel in each image that do not overlap with others [7]. Many decomposition techniques thus separate the unique areas well, but struggle with assignment of the overlaps.

The results of the proposed algorithm are displayed in Fig. 3, bottom, via the estimated variances of each pixel, $\hat{\xi}_{i,k}$ displayed in the same order of pixels as in the estimated image, image estimates $\hat{\mathbf{a}}_k$, source estimates $\hat{\mathbf{x}}_k = \hat{B}\hat{\mathbf{u}}_k$, and estimated convolution kernels $\hat{\mathbf{u}}_k$, respectively. Note that the first convolution kernel is a pulse, hence the corresponding source curve is the estimated input function. Both the CAM-CM solution and the estimated pixel variances $\hat{\xi}_{i,k}$ (ARD on elements of A), Fig 3. bottom left, tend to select the areas where the images are unique. However, the resulting estimates of the images $\hat{\mathbf{a}}_k$ of the S-BSS-DC have correctly assigned the overlapping parts.

The default starting points of the iterative algorithm (Section 3.4) were used in analysis of the sequence with the following observations of sensitivity:

- Initialization of the input function by the impulse is not a local minima due to the sparsity prior on the convolution kernels. The ARD prior favors sparse kernels and thus the impulse function is typically recovered in one of the convolution kernels. However, initialization of the input function by random values is unreliable and often converges to a local minima.
- Initialization of the convolution kernels by random starts is rather reliable and no local minima were observed.
- The initial estimate of the precision of the observation noise was selected using the mean of the eigenvalues of matrix $D'D$, see [16] for justification. The same results were obtained even with minimum and maximum of the eigenvalues. Local minima were observed only with extreme values of $\hat{\omega}^{(0)}$.
- The results are sensitive to the selected maximum number of sources r. When the number of sources is greater than the simulated, the strongest source is split into two factors with complementary convolution kernel.

4.3 Real Data Experiment

Validity of the model assumptions is now tested on real clinical data from renal scintigraphy. The tested dataset is a selection of dataset 28 from [20] where

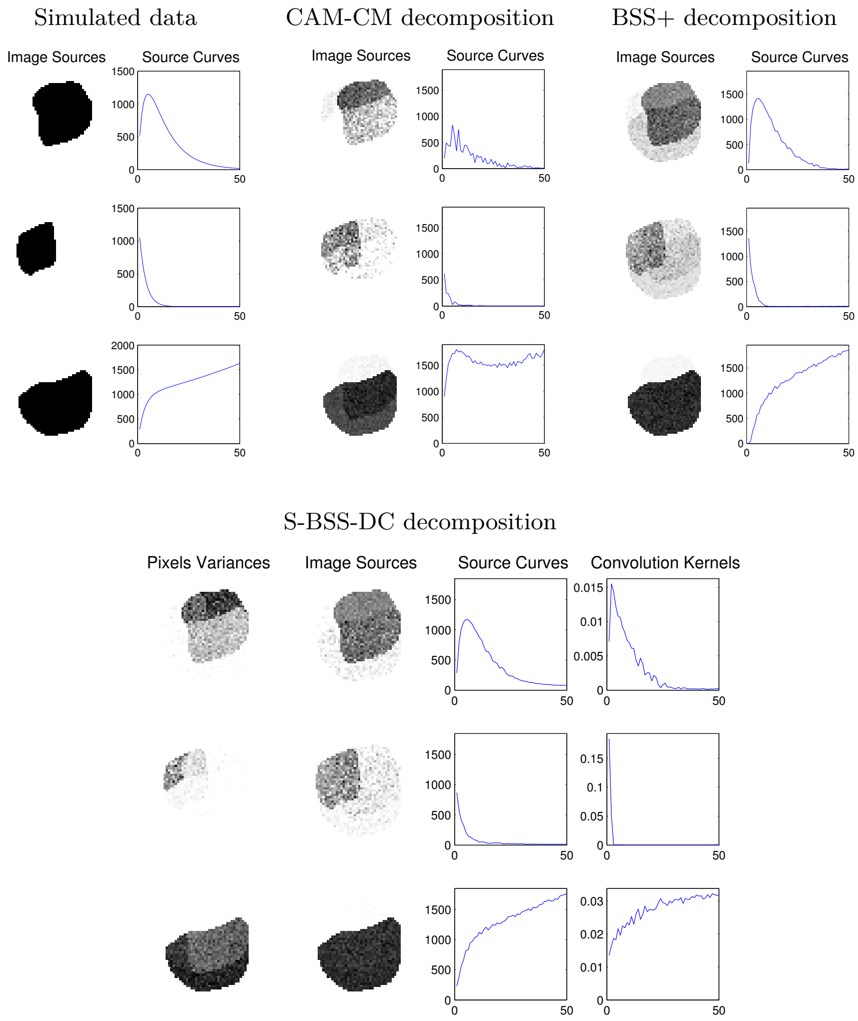

Fig. 3. Generated synthetic dataset using model [7]. **Top left**: simulated source images and curves. **Top middle**: decomposition of the data using the CAM-CM algorithm [7]. **Top right**: decomposition of the data using the BSS+ algorithm [13] **Bottom**: decomposition using the proposed S-BSS-DC algorithm.

a rectangular region of left part of the body and 99 time steps were selected. The images are obtained by counting radioactive particles, hence the observation noise is assumed to be Poisson-distributed. Therefore, we use the correspondence analysis which was found to be optimal conversion of this kind of noise to the homogeneous Gaussian noise [16]. In this application, we have good knowledge of the typical shapes of the input function and the convolution kernels. Thus, we initialize the iterative algorithm by the expected convolution kernels of a typical healthy patient. The convergence of the algorithm is thus significantly faster.

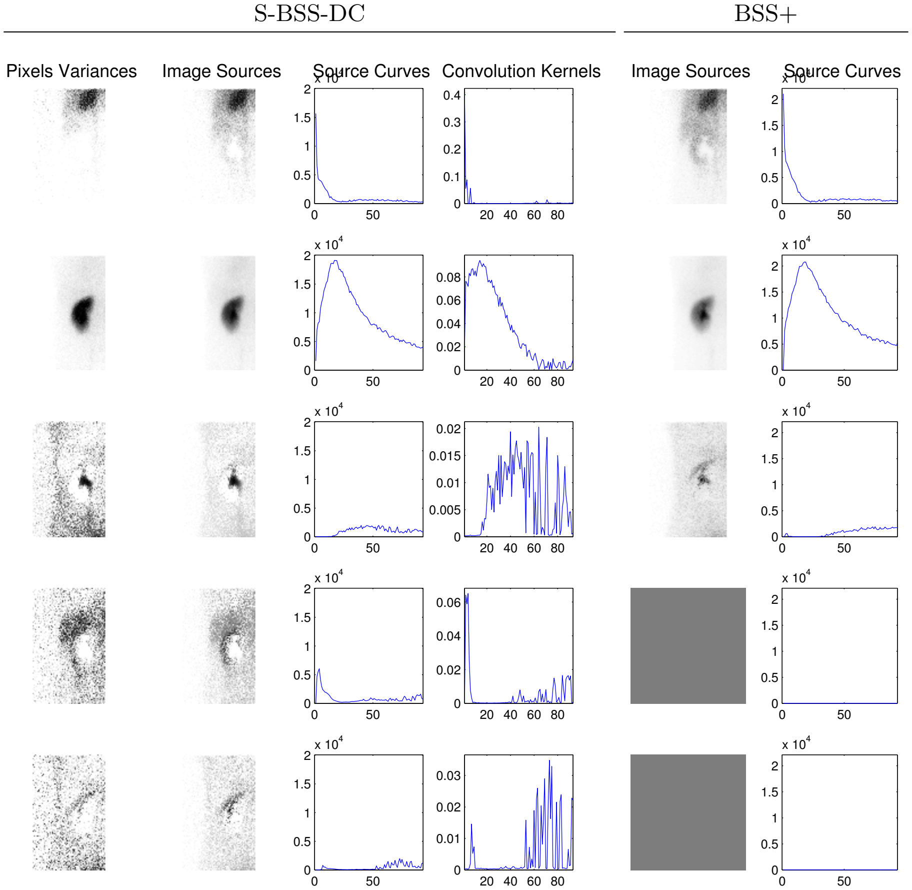

Fig. 4. Results of the proposed algorithm on a real dataset from renal scintigraphy. The columns of the S-BSS-DC algorithm are: the estimate of the variance of the image source prior, the estimate of the source image, the source curve, and its convolution kernel, respectively. The columns of the common BSS+ method are the estimates of the source images and the source curves respectively.

The results of the proposed S-BSS-DC algorithm on this data set are displayed in Fig. 4, via the estimated source pixel variance $\hat{\xi}_{i,k}$, image estimates $\hat{\mathbf{a}}_k$, source estimates $\hat{\mathbf{x}}_k = \hat{B}\hat{\mathbf{u}}_k$, and estimated convolution kernels $\hat{\mathbf{u}}_k$, respectively. Once again, the first convolution kernel was estimated to be a pulse, hence the first estimated source curve is equal to the estimated input function. For comparison, the results of blind source separation with positivity constraints (BSS+)are displayed in Fig. 4, right. Comparison with CAM-CM is omitted since its parametric form of the convolution kernels does not correspond to the data.

Note that S-BSS-DC decomposed the observed sequence to 5 sources, the BSS+ method found only 3 meaningful sources (the remaining were removed by the ARD property). All sources recovered by S-BSS-DC has very good medical interpretation as follows: 1) vascular structure, 2) parenchyma, 3) pelvis, 4) liver, 5) unspecific movement. The results of BSS+ can be interpreted as being superposition of sources discover by S-BSS-DC, namely: 1+4), 2+3) and 3+5), respectively. The results of BSS+ are not diagnostically relevant, due to their inability to separate pelvis and parenchyma.

An undesired artifact of the S-BSS-DC algorithm is its tendency to estimate non-smooth convolution kernel, see Fig. 4 right. This tendency is increasing with decreasing signal-to-noise ratio. More detailed modeling of the structure of the convolution kernels (e.g. via two unknown diagonals of the precision matrix in (21)) is required to allow reliable performance in these conditions.

5 Discussion and Conclusion

The problem of blind source separation and deconvolution is in general ill-posed and needs to be regularized by additional assumptions. In this paper, we proposed to use a hierarchical probabilistic model with unknown variance of all elements of the mixing matrix, and the convolution kernels. In effect, this prior promotes sparse estimates of these parameters. Since the proposed model does not allow for analytical solution, we applied the Variational Bayes method to find approximate solution. All other hyper-parameters are chosen to yield uninformative prior, hence the only additional parameter that needs to be chosen is the number of sources to recover. However, the number of sources needs to be chosen carefully, since the algorithm does not posses the ability to recover their number correctly.

Since the Variational Bayes is known to suffer from local minima, we proposed a general-purpose initialization of the implied iterative algorithm. The algorithm was tested in simulation and the proposed initialization was found to be robust and reliable. In specific applications, more appropriate choices can be made to speed up convergence of the algorithm.

The algorithm was also applied to the problem of decomposition of sequence of medical images. The proposed algorithm was able to identify diagnostically relevant sources better than conventional blind source separation methods.

Acknowledgment. This work was supported by the Czech Science Foundation, grant No. 13-29225S. The authors would like to thank Prof. Šámal from the 1st Faculty of Medicine, Charles University Prague for medical data and support with their evaluation.

References

1. Attias, H.: A Variational Bayesian framework for graphical models. In: Leen, T. (ed.) Advances in Neural Information Processing Systems, vol. 12. MIT Press (2000)
2. Attias, H.: New em algorithms for source separation and deconvolution with a microphone array. In: Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on. vol. 5, pp. V–297. IEEE (2003)
3. Barber, D.C.: The use of principal components in the quantitative analysis of gamma camera dynamic studies. Physics in Medicine and Biology 25, 283–292 (1980)
4. Bell, A.J., Sejnowski, T.J.: An information-maximization approach to blind separation and blind deconvolution. Neural computation 7(6), 1129–1159 (1995)
5. Bishop, C.M.: Variational principal components. In: Proc. of the Ninth Int. Conference on Artificial Neural Networks. ICANN (1999)
6. Buvat, I., Benali, H., Paola, R.: Statistical distribution of factors and factor images in factor analysis of medical image sequences. Physics in medicine and biology 43, 1695 (1998)
7. Chen, L., Choyke, P., Chan, T., Chi, C., Wang, G., Wang, Y.: Tissue-specific compartmental analysis for dynamic contrast-enhanced mr imaging of complex tumors. IEEE Transactions on Medical Imaging 30(12), 2044–2058 (2011)
8. Ghahramani, Z., Beal, M.: Graphical models and variational methods. In: Opper, M., Saad, D. (eds.) Advanced Mean Field Methods. The MIT Press (2001)
9. Haykin, S., Chen, Z.: The cocktail party problem. Neural computation 17(9), 1875–1902 (2005)
10. Jeffreys, H.: Theory of Probability. Oxford University Press, 3 edn. (1961)
11. Lawson, R.: Application of mathematical methods in dynamic nuclear medicine studies. Physics in medicine and biology 44, R57–R98 (1999)
12. MacKay, D.: Information theory, inference, and learning algorithms. Cambridge University Press (2003)
13. Miskin, J.W.: Ensemble Learning for Independent Component Analysis. Ph.D. thesis, University of Cambridge (2000)
14. Opper, M., Saad, D.: Advanced Mean Field Methods: Theory and Practice. The MIT Press, Cambridge, Massachusetts (2001)
15. Pedersen, M.S., Larsen, J., Kjems, U., Parra, L.C.: A survey of convolutive blind source separation methods. Multichannel Speech Processing Handbook pp. 1065–1084 (2007)
16. Šmídl, V., Quinn, A.: The Variational Bayes Method in Signal Processing. Springer (2005)
17. Šmídl, V., Tichý, O.: Automatic Regions of Interest in Factor Analysis for Dynamic Medical Imaging. In: 2012 IEEE International Symposium on Biomedical Imaging (ISBI). IEEE (2012)

18. Sroubek, F., Flusser, J.: Multichannel blind deconvolution of spatially misaligned images. Image Processing, IEEE Transactions on 14(7), 874–883 (2005)
19. Tzikas, D.G., Likas, A.C., Galatsanos, N.P.: Variational Bayesian sparse kernel-based blind image deconvolution with student's-t priors. Image Processing, IEEE Transactions on 18(4), 753–764 (2009)
20. VFN Praha: Database of dynamic renal scintigraphy (June 2013), `www.dynamicrenalstudy.org`
21. Šmídl, V., Tichý, O., Šámal, M.: Factor analysis of scintigraphic image sequences with integrated convolution model of factor curves. In: Proceedings of the second international conference on Computational Bioscience. IASTED (2011)

Appendix A: Shaping Parameters of Posterior Distributions

The shaping parameters of posterior distributions (25) - (28) are as follows:

$$\Sigma_{\mathbf{u}_k} = \left(\left(\widehat{\omega}\widehat{B'B}(\widehat{\mathbf{a}_k'\mathbf{a}_k})\right) + \mathrm{diag}(\widehat{\upsilon_k})\right)^{-1}, \tag{29}$$

$$\mu_{\mathbf{u}_k} = \Sigma_{\mathbf{u}_k}\left(\left(-\widehat{\omega}\sum_{l=1,l\neq k}^{r}\widehat{B'B}\widehat{\mathbf{u}}_l(\widehat{\mathbf{a}_k'\mathbf{a}_l})\right) + \widehat{\omega}\widehat{B'}D'\widehat{\mathbf{a}_k}\right), \tag{30}$$

$$\Sigma_{\mathbf{b}} = \left(\widehat{\varsigma}I_n + \widehat{\omega}\sum_{i,j=1}^{r}(\widehat{\mathbf{a}_i'\mathbf{a}_j})\left(\sum_{k,l=0}^{n-1}\Delta_k'\Delta_l(\widehat{\mathbf{u}_{k+1,j}\mathbf{u}_{l+1,i}})\right)\right)^{-1}, \tag{31}$$

$$\mu_{\mathbf{b}} = \Sigma_{\mathbf{b}}\widehat{\omega}\sum_{k=1}^{r}\left(\sum_{j=0}^{n-1}\Delta_j\widehat{\mathbf{u}_{j+1,k}}\right)' D'\widehat{\mathbf{a}_k}, \tag{32}$$

$$\alpha_k = \alpha_{k,0} + \frac{1}{2}\mathbf{1}_{n,1}, \quad \beta_k = \beta_{k,0} + \frac{1}{2}\mathrm{diag}\left(\widehat{\mathbf{u}_k\mathbf{u}_k'}\right), \tag{33}$$

$$\zeta = \zeta_0 + \frac{n}{2}, \qquad \eta = \eta_0 + \frac{1}{2}\mathrm{tr}\left(\widehat{\mathbf{b}'\mathbf{b}}\right), \tag{34}$$

$$\vartheta = \vartheta_0 + \frac{pn}{2}, \qquad \rho = \rho_0 + \frac{1}{2}\mathrm{tr}\left(DD' - 2\widehat{A}\widehat{X}'D'\right) + \frac{1}{2}\mathrm{tr}\left(\widehat{A'A}\widehat{X'X}\right), \tag{35}$$

$$\Sigma_{\mathbf{a}_i} = \left(\widehat{\omega}\sum_{j=1}^{n}(\widehat{\mathbf{x}_j'\mathbf{x}_j}) + \mathrm{diag}(\widehat{\xi_i})\right)^{-1}, \quad \mu_{\mathbf{a}_i} = \left(\Sigma_{\mathbf{a}_i}\left(\widehat{\omega}\sum_{j=1}^{n}(\widehat{\mathbf{x}_j}d_{i,j})'\right)\right)', \tag{36}$$

$$\phi_i = \phi_{i,0} + \frac{1}{2}\cdot\mathbf{1}_{r,1}, \qquad \psi_i = \psi_{i,0} + \frac{1}{2}\mathrm{diag}\left(\widehat{\mathbf{a}_i'\mathbf{a}_i}\right). \tag{37}$$

The auxiliary matrix $\Delta_k \in \mathbf{R}^{n\times n}$ is defined as

$$(\Delta_k)_{i,j} = \begin{cases} 1, & \text{if } i - j = k, \\ 0, & \text{otherwise.} \end{cases}$$

The moments of variables are computed using expectations of their probability density function, Appendix B.1.

Appendix B: Required Probability Distributions

B.1 Truncated Normal Distribution

Truncated normal distribution is defined for scalar random variable $x = N_x(\mu, \sigma)$ on interval $a < x \leq b$ as follows:

$$x \sim t\mathcal{N}(\mu, \sigma, a, b) = \frac{\sqrt{2}\exp((x-\mu)^2)}{\sqrt{\pi\sigma}(erf(\beta) - erf(\alpha))}\chi_{(a;b]}(x), \tag{38}$$

where $\alpha = \frac{a-\mu}{\sqrt{2\sigma}}$, $\beta = \frac{b-\mu}{\sqrt{2\sigma}}$, $\chi_{(a,b]}(x)$ is a characteristic function of interval $(a, b]$ defined as $\chi_{(a,b]}(x) = \begin{cases} 1 & x \in (a,b] \\ 0 & x \notin (a,b] \end{cases}$, and $\mathrm{erf}(t) = \frac{2}{\sqrt{\pi}}\int_0^t e^{-u^2}\mathrm{d}\mathrm{d}u$.

Moments of the truncated normal distribution are given as

$$\widehat{x} = \mu - \sqrt{\sigma}\frac{\sqrt{2}[\exp(-\beta^2) - \exp(-\alpha^2)]}{\sqrt{\pi}(erf(\beta) - erf(\alpha))}, \tag{39}$$

$$\widehat{x^2} = \sigma + \mu\widehat{x} - \sqrt{\sigma}\frac{\sqrt{2}[b\exp(-\beta^2) - a\exp(-\alpha^2)]}{\sqrt{\pi}(erf(\beta) - erf(\alpha))}. \tag{40}$$

B.2 Multivariate Truncated Normal Distribution

Truncation of the multivariate Normal distribution $\mathbf{x} \sim \mathcal{N}(\mu, \Sigma)$ is formally simple, however, its moments can not be expressed analytically. Therefore, we approximate the moments of $\mathbf{x}$ of the truncated Normal distribution by the moments of

$$\tilde{\mathbf{x}} \sim t\mathcal{N}(\mu, \mathrm{diag}(\boldsymbol{\sigma})),$$

where $\boldsymbol{\sigma}$ is a vector of diagonal elements of Σ. This corresponds to approximation of the posterior by a product of marginals (38) with mean value $\hat{\mathbf{x}}$ with elements given by (39) and $\widehat{\mathbf{x}\mathbf{x}^T} = \widehat{\mathbf{x}}\widehat{\mathbf{x}}^T + \mathrm{diag}(\hat{\boldsymbol{\sigma}})$, where $\hat{\sigma}_i = \widehat{x_i^2} - \hat{x}_i\hat{x}_i$. However, it may be too coarse approximation since it ignores covariance of the elements. An alternative is to approximate

$$\widehat{\mathbf{x}\mathbf{x}^T} = \widehat{\mathbf{x}}\widehat{\mathbf{x}}^T + \mathrm{diag}(\mathbf{o})\Sigma\mathrm{diag}(\mathbf{o}), \tag{41}$$

where $\mathbf{o}$ is a vector of elements $o_i = \hat{\sigma}_i^{1/2}\sigma_i^{-1/2}$. Heuristics (41) is motivated by the observation that for a Normal distribution with the main mass far from the truncation lines, $o_i \to 1$ and (41) becomes equivalent to the moment of the non-truncated Normal distribution.