Bayesian Blind Source Separation with Unknown Prior Covariance

Ondřej Tichý[1,2(✉)] and Václav Šmídl[1]

[1] Institute of Information Theory and Automation, Pod Vodárenskou věží 4, 18208 Prague 8, Czech Republic
otichy@utia.cas.cz
[2] Faculty of Nuclear Sciences and Physical Engineering, Břehová 7, Prague 1, Czech Republic

Abstract. The task of blind source separation (BSS) is to recover original signal sources which are observed only via their superposition with unknown weights. Since we are interested in estimation of the number of relevant sources in noisy observation, we use the Bayesian formulation which automatically removes spurious sources. A tool for this behavior is joint estimation of the unknown prior covariance matrix of the sources in tandem with the sources. In this work, we study the effect of various choices of the covariance matrix structure. Specifically, we compare models using the automatic relevance determination (ARD) principle on the first and the second diagonal, as well as full covariance matrix with Wishart prior. We obtain five versions of the variational BSS algorithm. These are tested on synthetic data and on a selected dataset from dynamic renal scintigraphy. MATLAB implementation of the methods is available for download.

Keywords: Blind source separation · Covariance model · Variational bayes approximation · Non-negative matrix factorization

1 Introduction

The blind source separation (BSS) problem arises in situations where several sources are observed only via their superposition such as in case of audio or medical signal processing [8] or hyperspectral imaging [10]. The task is to separate original sources, e.g. in the form of images and their related weights. The classical separation methods include principal or independent component analysis [6], non-negative matrix factorization [7], or projection methods [1,4].

In this work, we are focused on the Bayesian approach to the BSS problem which has advantages under poor signal to noise conditions and is capable to provides an estimate of the number of relevant sources. Another advantage for further processing of the results is the availability of uncertainty bounds around the estimate in the form of full probability distribution function. The ability to estimate the number of relevant sources is available due to a specific choice of the prior structure, typically unknown covariance matrix [9,12]. In this paper,

E. Vincent et al. (Eds.): LVA/ICA 2015, LNCS 9237, pp. 352–359, 2015.
DOI: 10.1007/978-3-319-22482-4_41

we study various choices of the structure of the prior covariance matrix and their effects on the behavior of the resulting separation algorithm. Specifically, we study three different assumptions leading to different covariance structures: (i) the source weights are most likely sparse which can be modeled using automatic relevance determination (ARD) approach [2,14], (ii) the source weights are smooth with occasional abrupt changes which can be modeled by sparse differences of the weights, and (iii) both weights and their differences can be sparse, which can be modeled by bi-diagonal covariance matrix. Since evaluation of exact posterior densities is not tractable, we apply the Variational Bayes method to obtain approximate posterior densities [11]. The first two structures are standard and the algorithms are well known, however, the last model is computationally intractable even under the Variational Bayes approach. Therefore, we propose to derive the posterior distribution for a full prior covariance matrix of the source weights using Wishart distribution. The introduced overparametrization is mitigated by the use technique known as matrix localization [5]. This heuristics is very successful in atmospheric modeling. Similar approach has been also applied for model of convolution kernels in blind deconvolution [13].

The resulting variants of the variational BSS algorithm are tested on a synthetic dynamic image data where advantages and disadvantages of the tested priors are demonstrated. The advantages of the proposed method were also observed on a real data set from dynamic renal scintigraphy, where the proposed method compares favorably with competing approaches such as the NMF algorithm [7]. Matlab implementations of the algorithms are freely available for download.

2 Bayesian Blind Source Separation

We introduce the Bayesian model of blind source separation. Prior models for all parameters of the model are described here except the prior for source weights which is described in the next section.

2.1 Observation Model

A sequence of recorded data vectors, $\mathbf{d}_t \in \mathbf{R}^{p\times 1}$, $t = 1, \ldots, n$, is stored columnwise in matrix $D \in \mathbf{R}^{p\times n}$. The assumed decomposition is

$$D = AX^T + E, \tag{1}$$

where matrix $A \in \mathbf{R}^{p\times r}$ represents the source images in its columns, matrix $X \in \mathbf{R}^{n\times r}$ represents source weights in its columns, matrix $E \in \mathbf{R}^{p\times n}$ represents noise term of the observation model, and symbol $()^T$ denotes transposition of a vector or a matrix in this paper.

We assume that all elements of the matrices D, A, X, and E are positive; however, modification to full support is straightforward.

2.2 Noise Model

We use the isotropic Gaussian noise model [15] with zero mean and common variance for all pixels, $e_{i,j} \sim \mathcal{N}_{e_{i,j}}(0, \omega^{-1})$. Then, the observation model can be rewritten as

$$f(D|A, X, \omega) = \prod_{t=1}^{n} \mathcal{N}_{\mathbf{d}_t}\left(A\overline{\mathbf{x}}_t^T, \omega^{-1} I_p\right), \; f(\omega) = \mathcal{G}_\omega(\vartheta_0, \rho_0), \tag{2}$$

where symbol $\mathcal{N}$ denotes normal distribution and symbol I_p denotes identity matrix of the given size. In the Bayesian methodology, all unknown parameters have their prior distribution. The prior distribution for the precision of the noise, ω, has a conjugate prior in the form of the Gamma distribution, denoted as $\mathcal{G}$, with selected prior constants ϑ_0, ρ_0.

2.3 Prior Model of Source Images

The prior model of the source images is common for all methods in the paper. Each source image, i.e. column of the matrix A, $\mathbf{a}_k$, has prior in the form of the normal distribution with unknown precision parameter related to each source image as

$$f(\mathbf{a}_k|\xi_k) = t\mathcal{N}_{\mathbf{a}_k}\left(\mathbf{0}_{p,1}, \xi_k^{-1} I_p, [0, \infty]\right), \; f(\xi_k) = \mathcal{G}_{\xi_k}(\phi_0, \psi_0), \tag{3}$$

where $t\mathcal{N}$ denotes truncated normal distribution with given support and ξ_k is an unknown precision parameter with the Gamma prior for $k = 1, \ldots, r$ where ϕ_0, ψ_0 are selected prior constants. This parameter acts as the automatic relevance determination (ARD) term [14].

3 Prior Models of Covariance Matrix of Source Weights

3.1 Isotropic Prior

The only assumption in this case is that the elements of each weights vector are isotropic [11], i.e. that their covariance matrix is identity matrix as $f(\mathbf{x}_k) = t\mathcal{N}_{\mathbf{x}_k}\left(\mathbf{0}_{n,1}, I_n, [0, \infty]\right)$ for $k = 1, \ldots r$.

3.2 Sparse Prior

The key assumption of this prior is that the source weights are most likely sparse. Once again, we employ the ARD principle; however, in a different way than in Sect. 2.3. Here, each element of source weight, $x_{k,j}$, has its own ARD prior with relevance parameter, $\upsilon_{k,j}$, which can be written in vector form as

$$f(\mathbf{x}_k|\boldsymbol{\upsilon}_k) = t\mathcal{N}_{\mathbf{x}_k}\left(\mathbf{0}_{n,1}, \mathrm{diag}(\boldsymbol{\upsilon}_k^{-1}), [0, \infty]\right), \; f(\upsilon_{k,j}) = \mathcal{G}_{\upsilon_{k,j}}(\alpha_0, \beta_0), \tag{4}$$

$\forall j = 1, \ldots, n$, where diag() denotes square matrix with argument vector in its diagonal and zeros otherwise and α_0, β_0 are selected prior constants.

The purpose of this approach is to favor zeros in estimates of the elements of the weights.

3.3 Sparse Differences Prior

Sparse prior from Sect. 3.2 could possibly lead to very non-smooth solutions. If smooth solutions are preferred, a model of sparse differences instead of sparse elements could be more appropriate. The differences of $\mathbf{x}_k$ can be expressed using ∇ operator as $\nabla\mathbf{x}_k$, where $\nabla \in \mathbf{R}^{n\times n}$ is the matrix with ones on its diagonal, -1s on its superdiagonal, and zeros otherwise. We employ the ARD principle on each element of $\nabla\mathbf{x}_k$ with relevance parameter $\boldsymbol{\upsilon}_k^{\nabla}$. This can be formulated equally using full vector $\mathbf{x}_k$ as

$$f(\nabla\mathbf{x}_k|\boldsymbol{\upsilon}_k^{\nabla}) \leftrightarrow f(\mathbf{x}_k|\boldsymbol{\upsilon}_k) = t\mathcal{N}_{\mathbf{x}_k}\left(\mathbf{0}_{n,1}, \nabla^{-1}\mathrm{diag}(\boldsymbol{\upsilon}_k^{-1})(\nabla^{-1})^T, [0,\infty]\right), \quad (5)$$

$$f(\boldsymbol{\upsilon}_{k,j}) = \mathcal{G}_{\boldsymbol{\upsilon}_{k,j}}(\alpha_0, \beta_0), \quad \forall j = 1,\ldots,n, \quad (6)$$

with selected prior constants α_0, β_0.

3.4 Wishart Prior

Till this moment, we have modeled only selected diagonals of the covariance matrix. However, it is possible to model the full covariance matrix. For this task, we use vectorized form of the matrix X as $\overrightarrow{\mathbf{x}} = \mathrm{vec}(X) = [\mathbf{x}_1^T, \ldots, \mathbf{x}_r^T]^T \in \mathbf{R}^{nr\times 1}$ where the covariance between all elements is a full covariance matrix $\Upsilon \in \mathbf{R}^{nr\times nr}$. Prior distribution on an unknown full covariance matrix is usually chosen in the form of Wishart matrix distribution,

$$f(\overrightarrow{\mathbf{x}}|\Upsilon) = t\mathcal{N}_{\overrightarrow{\mathbf{x}}}\left(\mathbf{0}_{nr,1}, \Upsilon^{-1}, [0,\infty]\right), \; f(\Upsilon) = \mathcal{W}_{\Upsilon}(\alpha_0 I_n, \beta_0), \quad (7)$$

where $\mathcal{W}()$ denotes the Wishart matrix distribution and α_0, β_0 are selected prior constants.

The weak point of this prior model is that n^2r^2 additional parameters have to be estimated which makes this problem very ill-posed.

3.5 Wishart Prior with Localization

We assume that the most relevant prior knowledge is located only in several diagonals of the covariance matrix and its sub-matrices. This idea originates in data assimilation of atmospheric models [5]. Therefore, we replace the remaining entries in the estimate by zeros. Formally, we use the Hadamard matrix product which is defined between two matrices of the same size as $C = A \circ B$ where $c_{i,j} = a_{i,j}b_{i,j}$. Then, the localization of the posterior estimate of the full covariance matrix from Sect. 3.4 is

$$\widehat{\Upsilon}_{\mathrm{loc}} = \widehat{\Upsilon} \circ L, \quad (8)$$

where $\widehat{\Upsilon}$ denotes estimate of Υ and L is the localization matrix of the same size as the matrix Υ. There could be many possible localization matrices L [3] however their study is out of the scope of this paper. Here, we will show results with two localization matrices (i) matrix of ones, i.e. without any localization (denoted as Wishart), and (ii) localization matrix L with ones on the first and the second diagonals of all sub-matrices and zeros otherwise (denoted as Localized Wishart).

3.6 Approximate Solution Using Variational Bayes Method

The whole probabilistic prior model is formed using equations (2)–(3) and prior model from Sects. 3.1–3.5. Estimation algorithm for each prior model was derived using the Variational Bayes method [11] where equations for shaping parameters of the posterior probability densities of the model parameters are found in the form of a set of implicit equations which has to be solved iteratively. Solutions for the first three models are available from previous publications, equation for the proposed version with Wishart prior and localization are given in the Appendix A. This yields five different versions of the variational BSS algorithm (two versions with Wishart prior, with and without localization). All prior parameters (with subscript 0) are set to $10^{\pm 10}$ in order to yield non-informative priors while all algorithms are not sensitive to this selection.

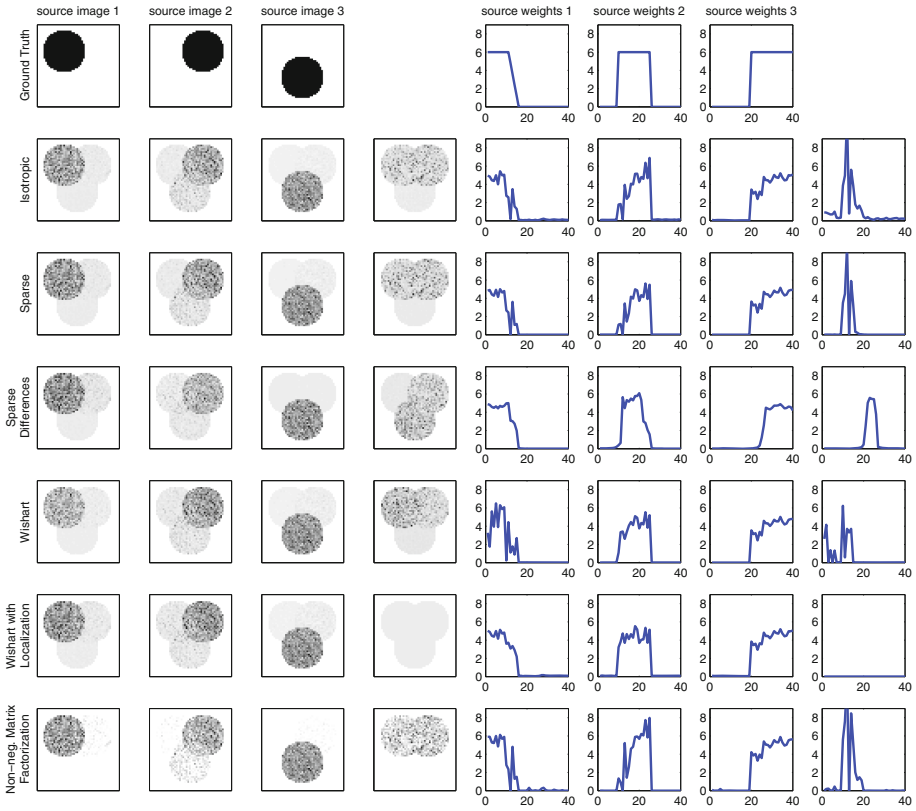

Fig. 1. The results of the five studied methods (the second to the sixth rows) together with NMF algorithm results (the seventh row) in synthetically generated data (the first row).

4 Experiments

4.1 Experiment on Synthetic Dataset

All five derived algorithms are now being tested on synthetic dataset in order to study the impact of covariance matrix models on resulting estimates. The data are generated according to model (1) using three sources with different time-dependent weights as displayed in Fig. 1, top row, degenerated by homogeneous Gaussian noise. All algorithms run with the same conditions such as starting point of iterations and expected number of sources which is set to $r = 4$ in order to study the ability of algorithms to recognize the correct number of sources since the modeled number of sources is 3.

The results from all tested algorithms are given in Fig. 1, rows 2–6, together with the state of the art non-negative matrix factorization (NMF) algorithm [7], row 7. There are estimated source images and source weights in row-wise schema where four images in each row are accompanied with related four weights vectors. It can be seen that all algorithms are capable to correctly estimate source images. The main differences between the algorithms is in estimates of the source weights. The fourth redundant source from BSS with isotropic prior of NMF has been estimated such that its activity is taken from the first and the second source. The same behavior can bee seen on the result of BSS with the Sparse prior; however, the tendency to favor zeros in source weights can be nicely observed here. The BSS with sparse differences prior provides smooth estimates of the source weights; however, the algorithm estimated the fourth source as a combination of the second and the third source. The BSS with the Wishart prior does not penalizes redundant sources and the activity in fourth source is taken from the first. Only the BSS with Wishart prior and localization achieves suppression of the redundant source. It is still estimated, however, with negligible activity which is under the displayed resolution.

4.2 Experiment on Dynamic Scintigraphy Dataset

In this experiment, we will use a selected data from dynamic renal scintigraphy[1] to demonstrate the performance of the methods on real data. The data has original resolution 128×128 pixels; however, we select a region with one kidney of the size 37×47 where medically relevant sources (kidney pelvis and parenchyma) are located. The whole sequence is composed of 100 images with sampling period of 10 s.

We compare only BSS algorithms based on the Sparse differences prior and the Wishart prior with localization with the NMF algorithm. The $r = 3$ for the tested algorithms. The results are summarized in Fig. 2. Estimated source images and source weights are in a row-wise schema. Two methods, BSS with Sparse differences prior and the NMF, estimate threes significant sources where sources 2 and 3 correspond to biological activity of the pelvis. Only the BSS with

[1] www.dynamicrenalstudy.org.

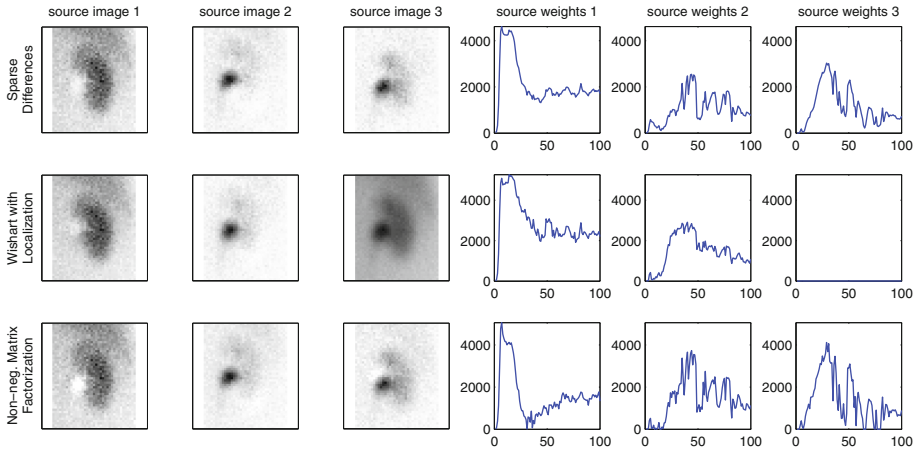

Fig. 2. Results of selected BSS algorithms on dynamic renal scintigraphy data. Source images are in the first three columns and related TACs are in the second three columns.

Wishart prior and localization estimates only two sources which correspond very well with the expected biological function of pelvis and parenchyma.

5 Discussion and Conclusion

The problem of blind source separation (BSS) is generally ill-posed, especially under the conditions such as noisy observations or unknown number of sources. Bayesian approach is generally valuable for its ability to estimate the number of relevant sources using hierarchical priors. In this work, we study various choices of prior covariance structure of the source weights. Covariance structures with ARD and ARD principle of the differences were already published. We propose another model using Wishart prior and develop Variation Bayes estimation algorithm with non-standard step of covariance localization. The proposed algorithm was found to have superior ability to suppress redundant sources in blind source separation of noisy image sequences. All versions of the variational BSS algorithm are implemented in Matlab and freely available for download from http://www.utia.cz/AS/softwaretools/image_sequences/.

Acknowledgement. This work was supported by the Czech Science Foundation, grant No. 13-29225S, and by the Grant Agency of the Czech Technical University in Prague, grant No. SGS14/205/OHK4/3T/14.

A Shaping Parameters of Posterior Distributions

Posterior distributions are $\tilde{f}(A|D) = t\mathcal{N}_A(\mu_A, I_p \otimes \Sigma_A)$, $\tilde{f}(\xi_k|D) = \mathcal{G}_{\xi_k}(\phi_k, \psi_k)$, $\tilde{f}(\mathbf{x}|D) = t\mathcal{N}_{\mathbf{x}}(\mu_{\mathbf{x}}, \Sigma_{\mathbf{x}})$, $\tilde{f}(\Upsilon|D) = \mathcal{W}_{\Upsilon,nr}(\Sigma_\Upsilon, \beta)$, $\tilde{f}(\omega|D) = \mathcal{G}_\omega(\vartheta, \rho)$,

with shaping parameters $\Sigma_A^{-1} = \left(\omega\widehat{X^TX} + \widehat{\Xi}\right)$, $\mu_A = \left(\omega D\widehat{X}\right)\Sigma_A$, $\phi = \phi_0 + \frac{p}{2}\mathbf{1}_{r,1}$, $\psi = \psi_0 + \frac{1}{2}\mathrm{diag}\left(\widehat{A^TA}\right)$, $\Sigma_{\mathbf{x}}^{-1} = \left((\widehat{\omega}\widehat{A^TA}) \otimes I_n + \widehat{\Upsilon} \circ L\right)$ $\mu_{\mathbf{x}} = \Sigma_{\mathbf{x}}\left(\widehat{\omega}\mathrm{vec}\left(D^T\widehat{A}\right)\right)$ $\Sigma_{\Upsilon}^{-1} = \left(\widehat{\mathbf{x}\mathbf{x}^T} + \alpha_0^{-1}I_{nr}\right)$ $\beta = \beta_0 + 1$ $\vartheta = \vartheta_0 + \frac{pn}{2}$, $\rho = \rho_0 + \frac{1}{2}\mathrm{tr}\left((D - \widehat{A}\widehat{X^T})(D - \widehat{A}\widehat{X^T})^T\right)$.

References

1. Araújo, M.C.U., Saldanha, T.C.B., Galvão, R.K.H., Yoneyama, T., Chame, H.C., Visani, V.: The successive projections algorithm for variable selection in spectroscopic multicomponent analysis. Chemometr. Intell. Lab. Syst. **57**(2), 65–73 (2001)
2. Bishop, C.M.: Variational principal components. In: IET Conference Proceedings, pp. 509–514(5), January 1999
3. Gaspari, G., Cohn, S.E.: Construction of correlation functions in two and three dimensions. Q. J. Roy. Meteorol. Soc. **125**(554), 723–757 (1999)
4. Gillis, N.: Successive nonnegative projection algorithm for robust nonnegative blind source separation. SIAM J. Imaging Sci. **7**(2), 1420–1450 (2014)
5. Hamill, T.M., Whitaker, J.S., Snyder, C.: Distance-dependent filtering of background error covariance estimates in an ensemble kalman filter. Mon. Weather Rev. **129**(11), 2776–2790 (2001)
6. Hyvärinen, A., Karhunen, J., Oja, E.: Independent Component Analysis, vol. 46. Wiley, New York (2004)
7. Lee, D.D., Seung, H.S.: Algorithms for non-negative matrix factorization. In: Proceedings of Advances in neural information processing systems, pp. 556–562 (2001)
8. Margadán-Méndez, M., Juslin, A., Nesterov, S.V., Kalliokoski, K., Knuuti, J., Ruotsalainen, U.: ICA based automatic segmentation of dynamic cardiac PET images. IEEE Trans. Inf. Technol. Biomed. **14**(3), 795–802 (2010)
9. Miskin, J.W.: Ensemble learning for independent component analysis. Ph.D. thesis, University of Cambridge (2000)
10. Moussaoui, S., Hauksdottir, H., Schmidt, F., Jutten, C., Chanussot, J., Brie, D., Douté, S., Benediktsson, J.A.: On the decomposition of mars hyperspectral data by ica and bayesian positive source separation. Neurocomputing **71**(10), 2194–2208 (2008)
11. Šmídl, V., Quinn, A.: The Variational Bayes Method in Signal Processing. Springer, Heidelberg (2006)
12. Šmídl, V., Quinn, A.: On bayesian principal component analysis. Comput. Stat. Data Anal. **51**(9), 4101–4123 (2007)
13. Tichý, O., Šmídl, V.: Non-parametric bayesian models of response function in dynamic image sequences. Pre-print submitted to Computer Vision and Image Understanding (arXiv:1503.05684 [stat.ML]) (2015)
14. Tipping, M.E.: Sparse bayesian learning and the relevance vector machine. J. Mach. Learn. Res. **1**, 211–244 (2001)
15. Tipping, M.E., Bishop, C.M.: Probabilistic principal component analysis. J. Roy. Stat. Soc. B (Stat. Methodol.) **61**(3), 611–622 (1999)