CZECH TECHNICAL UNIVERSITY IN PRAGUE

FACULTY OF NUCLEAR SCIENCES AND PHYSICAL ENGINEERING

Department of Mathematics

DISSERTATION

Bayesian Blind Source Separation in Dynamic Medical Imaging

Bayesovská slepá separace signálu v dynamickém medicínském zobrazování

Prague 2015 ONDŘEJ TICHÝ

Bibliografický záznam

Autor	Ing. Ondřej Tichý, České vysoké učení technické v Praze, Fakulta jaderná a fyzikálně inženýrská, Katedra matematiky,
Název práce	Bayesovská slepá separace signálu v dynamickém medicínském zobrazování,
Studijní program	Aplikace přírodních věd,
Studijní obor	Matematické inženýrství,
Školitel	doc. Ing. Václav Šmídl, Ph.D.,
Akademický rok	2014/2015,
Počet stran	120
Klíčová slova	Slepá separace signálu, Variační Bayesova aproximace, Řídkost, Dekonvoluce, Obrazová sekvence

Bibliographic Entry

Author	Ing. Ondřej Tichý, Czech Technical University in Prague, Faculty of Nuclear Sciences and Physical Engineering, Department of Mathematics,
Title of Dissertation	Bayesian Blind Source Separation in Dynamic Medical Imaging,
Degree Program	Applied Natural Sciences,
Field of Study	Mathematical Engineering,
Supervisor	doc. Ing. Václav Šmídl, Ph.D.,
Academic Year	2014/2015,
Number of Pages	120
Keywords	Blind Source Separation, Variational Bayes Approximation, Sparsity, Deconvolution, Image Sequence

Abstract

This work is concerned with the blind source separation (BSS) problem in dynamic medical imaging with focus on dynamic planar renal scintigraphy. A common problem of imaging of a three-dimensional object into an image plane is that the signal arises as a superposition of signals from underlaying sources from different depths of a body. The task is to separate individual sources representing functional tissues in medical imaging, i.e. their images and activities over the time.

The main contribution of this work is introduction of novel models of hierarchical priors for Bayesian BSS, development of BSS algorithms for them, and evaluation of their suitability on clinical data. Common method in this application domain is still manual analysis by an expert. Existing knowledge of the expert in dynamic nuclear medicine was used as inspiration for the proposed hierarchical priors. Two key studied properties of the problem are sparsity of the sources and modeling of each source activity as a convolution of the common input function and a source-specific convolution kernel.

The proposed methods are tested together with selected state-of-the-art BSS algorithms on two large datasets from dynamic renal scintigraphy as well as on representative data from other imaging modalities and the significant improvements of separation using proposed methods are demonstrated.

Abstrakt

Tato práce se zabývá problémem slepé separace signálu (blind source separation, BSS) vzniklém při dynamickém zobrazování v medicíně a je především motivována analýzou signálu z dynamické planární scintigrafie ledvin. Problémem při zobrazování třírozměrných objektů do obrazové plochy je vznik signálu, který představuje superpozici signálů z jednotlivých zdrojů v lidském těle. Úkolem slepé separace signálu je pak tyto zdroje, které reprezentují jednotlivé funkční orgány, rozlišit, tedy určit jejich obraz a křivku časové aktivity.

Hlavním přínosem této práce je vytvoření nových modelů pomocí hierarchických aprioren pro Bayesovskou BSS, odvození příslušných BSS algoritmů a zhodnocení jejich přínosu pro klinickou praxi. Protože standardem v této oblasti je stále ruční analýza získaných dat, existující znalosti z dynamické nukleární medicíny byly využity pro konstrukci hierarchických aprioren. Dvě hlavní oblasti studia jsou modely řídkosti signálu a modely aktivit jednotlivých zdrojů jakožto konvoluce mezi společnou vstupní funkcí a specifickými konvolučními jádry.

Odvozené odhadovací metody jsou aplikovány společně se state-of-the-art metodami na dva rozsáhlé datasety z dynamické scintigrafie ledvin, ale i na další vybraná data z různých zobrazovacích modalit, kde jsou demonstrována vyrázná vylepšení separace pomocí navržených metod.

Contents

Declaration

I, the undersigned, hereby declare that I worked on this thesis on my own and I used only the sources which are listed in the Bibliography.

Ondřej Tichý

Prague, 21.4.2015

Acknowledgement

I am very grateful to my supervisor Václav Šmídl for his guidance in this work and for his support, energy, and inspiration. I would also like thank to Martin Šámal for his support, for the use of his data, and for his valuable help with interpretation of results.

Finally, I would like to thank my family, especially to my beloved wife Barbora.

The financial support of the Czech Science Foundation, grant GA13-29225S, is gratefully acknowledged.

Notation

Linear Algebra and Calculus	
$\mathbf{R}$	Set of real numbers.
$A \in \mathbf{R}^{p \times r}$	A is a real matrix of the size $p \times r$.
A^T	Transpose of the matrix A.
A^{-1}	Inverse of the matrix A.
$\mathbf{a}_k$	The kth column of the matrix A.
$\overline{\mathbf{a}}_i$	The ith row if the matrix A.
$a_{i,k}$	The i,kth element of the matrix A; i denotes row index and k denotes column index.
$\mathbf{b} \in \mathbf{R}^n$	$\mathbf{b}$ is a vector of the size n.
b_j	The jth element on the vector $\mathbf{b}$.
$\mathrm{diag}(A)$	A diagonal vector of a square matrix A.
$\mathrm{diag}(\mathbf{b})$	A matrix with vector $\mathbf{b}$ on its diagonal and zeros otherwise.
$\mathrm{tr}(A)$	The trace of a square matrix $A \in \mathbf{R}^{p \times p}$ defined as $\mathrm{tr}(A) = \sum_{i=1}^{p} a_{i,i}$.
$\mathrm{vec}(A)$	Vector composed of columns of the matrix A.
$\lvert A \rvert$	Determinant of the matrix A.
I_n	Identity matrix of the size n.
$\mathbf{1}_{p,n}$	Matrix of ones of the size $p \times n$.
$\mathbf{0}_{p,n}$	Matrix of zeros of the size $p \times n$.
$A \otimes B$	Kronecker product of matrices $A \in \mathbf{R}^{p \times r}$ and B; $A \otimes B = \begin{bmatrix} a_{1,1}B & \cdots & a_{1,r}B \\ \vdots & \ddots & \vdots \\ a_{p,1}B & \cdots & a_{p,r}B \end{bmatrix}$
$A \circ B$	Hadamard product of matrices A and B of the same size; $A \circ B = \begin{bmatrix} a_{1,1}b_{1,1} & \cdots & a_{1,r}b_{1,r} \\ \vdots & \ddots & \vdots \\ a_{p,1}b_{p,1} & \cdots & a_{p,r}b_{p,r} \end{bmatrix}$
$\ln(a)$	Natural logarithm of argument.
$\exp(a)$	Exponential of argument.
$f(x) \propto f(y)$	$f(x)$ is proportionally equal to $f(y)$.

Probabilistic Calculus	
$\mathrm{E}_{f(x)}(..)$	Expected value of argument with respect to probability density function $f(x)$.
$\widehat{X}$	Point estimate of (multivariate) parameter X.
$\mathcal{N}_{\mathbf{x}}(\mu_{\mathbf{x}}, \Sigma_{\mathbf{x}})$	Multivariate normal distribution, see Appendix A.1.1.
$\mathcal{MN}_X(\mu_X, \Sigma_n \otimes \Sigma_r)$	Matrix normal distribution, see Appendix A.1.2.
$t\mathcal{N}_x(\mu_x, \sigma_x, [0, \infty])$	Truncated normal distribution, see Appendix A.1.3.
$t\mathcal{MN}_X(\mu_X, \Sigma_n \otimes \Sigma_r, [0, \infty])$	Truncated matrix distribution, truncated element-wise according to truncated normal distribution.
$\mathcal{G}_x(\alpha, \beta)$	Gamma distribution, see Appendix A.2.
$\mathrm{tExp}_x(\lambda)$	Truncated exponential distribution, see Appendix A.3.
$\mathcal{U}_x(a, b)$	Uniform distribution, see Appendix A.4.
$\mathcal{W}_x(\Sigma, \nu)$	Wishart distribution, see Appendix A.5.

1. Introduction

Aren't you essentially using Bayes' theorem?
(Irving Good)
I suppose.
(Alan Turing)

Let us assume that the number 6 is a measurement of a sum of two numbers. The task is to find these two numbers. There are infinitely many solutions such as: 3 and 3, 5 and 1, or $\frac{1}{5}$ and $\frac{29}{5}$. Similar problems are common in many signal processing areas where some elements of a measured signal are considered to be a sum of signals from unknown number of sources such as in astronomy [34], hyperspectral imaging [79], electroencephalography [58], scintigraphy [108], positron emission tomography [127], or magnetic resonance imaging [24]. Here, the observed data are in the form of a mixture of signals and the task is to reconstruct the original sources. A classical example is the Cocktail Party Problem [76] where many speakers are recorded by a set of microphones. Theoretically, every microphone can hear every speaker. The task is to extract signals from each speaker (source) which is call source separation. Since weight of each source in each microphone is unknown, the problem is called blind source separation.

1.1. Blind Source Separation

There are many possible models for the blind source separation (BSS) problem but we consider only a linear model, where the data stored in the matrix D are assumed to be a linear combination of sources. The linear combination has specific form where source images, stored in columns of the matrix A, are weighted by their related activities, stored in columns of the matrix X. Then, the BSS model can be written as

$$D = AX^T + E, \tag{1.1}$$

where $D \in \mathbf{R}^{p\times n}$, $A \in \mathbf{R}^{p\times r}$, $X \in \mathbf{R}^{n\times r}$, and $E \in \mathbf{R}^{p\times n}$. The scheme of the BSS model is illustrated in Figure 1.1. Here, p is the size of each source, n is the number of data points, and r is the number of sources. Typically, the sizes are ordered as $p \gg n \gg r$. In this work, we refer to the term A as the source images matrix while the mixing matrix is often used in literature and we refer to the term X as the time-activity curves (TAC) matrix while the source term

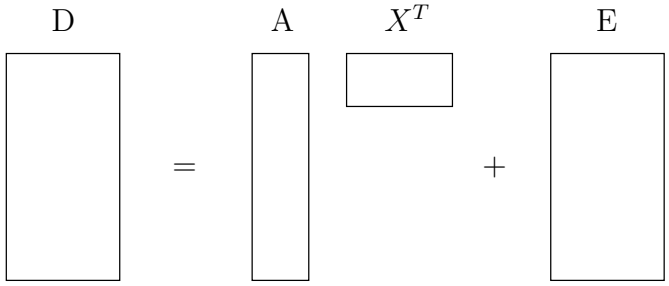

Figure 1.1.: Illustration of the studied blind source separation problem.

or the source values are often used in literature [76]. The term E is the error part in the model which commonly represents the measurement of noise.

There are several ambiguities in the BSS problem addressed in this work: the noise, the number of sources, and the rotation ambiguity.

Noise Noise term in the BSS problem, the equation (1.1), E, is important part of the model which is, however, not always taken into the account.

- Noiseless BSS: A number of approaches assumes noiseless BSS problem, $E = 0$ in the equation (1.1), such as basic versions of the independent component analysis [22] where exact source reconstruction is possible [27]. This possibility is attractive, however, the condition of noise-free data is rarely met in practice.

- Residual Approach: Some approaches assume noiseless model of the BSS, $D \approx AX^T$; however, they incorporate the noise in the approximate solution using, e.g., cost function measuring the quality of approximation such as square of the Euclidean distance between data and its reconstruction

$$\left|\left|D - AX^T\right|\right|^2 = \sum_{i,j} \left(D_{i,j} - \left(AX^T\right)_{i,j}\right)^2, \tag{1.2}$$

  or Kullback-Leibler divergence [68].

- Statistical Approach: The noise model, $f(E)$, can be selected to have an assumed probability distribution. In this work, we assume the BSS model with additive Gaussian noise. The most common case is the uncorrelated noise model

$$f(e_{i,j}) = \mathcal{N}_{e_{i,j}}(0, \omega_{i,j}^{-1}), \tag{1.3}$$

  where each noise element has zero mean and variance $\omega_{i,j}^{-1}$. In this work, we will use the isotropic Gaussian noise model [115] with zero mean and common variance for all pixels, $e_{i,j} \sim \mathcal{N}_{e_{i,j}}(0, \omega^{-1})$, while other versions are studied, e.g., in [96].

Number of sources The number of sources estimation is a bottleneck in many BSS methods. Very often, manual selection of the number of source is a prerequisite step for successful run of a BSS method. However, the selection must be done by an expert or using insight into the problem which is an extremely demanding condition. Many heuristic methods for selection of rank r exist [57], however, none are used as a generally accepted methodology. Hence, dealing with the number of source estimation remain still a challenge.

Separation ambiguity The separation ambiguity arises in the BSS model (1.1) naturally since for every invertible matrix $T \in \mathbf{R}^{r\times r}$, it holds

$$D = AX^T = \Big(AT\Big)\Big(T^{-1}X^T\Big). \tag{1.4}$$

This is known as the rotational ambiguity in literature [4] and T is called the rotation matrix. The equation (1.4) demonstrates that the BSS problem, the equation (1.1), has infinitely many solutions in general. One possible way to restrict the space of possible solutions is to selected some limiting conditions for parameters of the BSS model. One such possible condition can be the constraint of positivity of all elements of matrices A and X. [91] show that under this constraint, the unique solution is guaranteed if each source has at least one non-overlapped pixel; however, even this condition is too restrictive in practice. Further assumption such as sparsity [61, 121] or specific model of dynamics of sources such as convolution [6, 122] were proposed with advantages in specific fields.

1.1.1. Example Source Separation of Medical Image Sequence

For example, the data model from equation (1.1) can be used in medical image sequence analysis from, e.g., the planar scintigraphy. Here, measured signal on each detector is supposed to be a sum of signals from different depths of the body; hence, the resulting image is a superposition of signal from an unknown number of sources. The dynamic image sequence arises when the images, $\mathbf{d}_j$, are taken repetitively; hence, $D = [\mathbf{d}_1, \ldots, \mathbf{d}_n]$. The main advantage of dynamic image sequence is that it provides not only spatial information but also information about source activity over time. The following task is to analyze the sequence; it means to estimate:

- original source images, columns of the matrix A,
- their associated activities, called time-activity curves (TACs) [66], columns of the matrix X.

An example sequence from dynamic renal scintigraphy [29] is given Figure 1.2, on the left, as a demonstration of the typical input of the BSS algorithm. The expected form of the output of the BSS algorithm, i.e. estimates of the source images and related TACs, is in Figure 1.2, right.

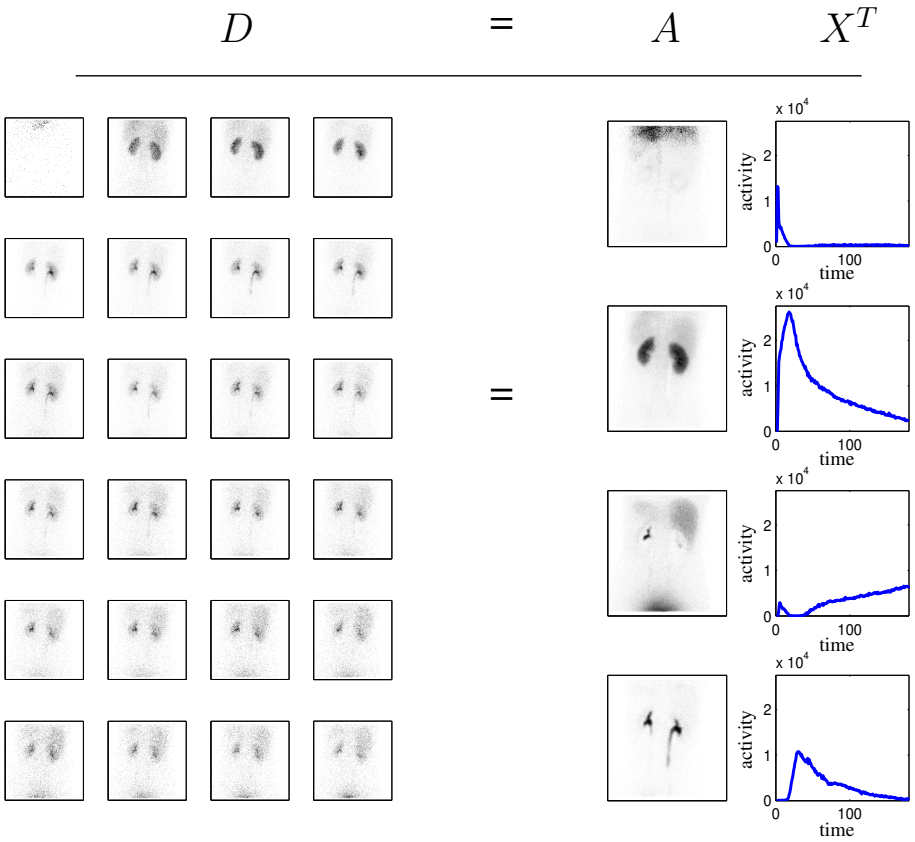

Figure 1.2.: The decomposition $D = AX^T$ is schematically demonstrated. Left: every 7th image from the example scintigraphic sequence. Right: example separation of the sequence into tissue images in the first column and related TACs in the second column.

In scintigraphy, the model (1.1) is also called Factor Analysis of Medical Image Sequences (FAMIS) [29, 10, 19]. The FAMIS model assumes (i) no motion of tissues and (ii) no enlarging of tissues as a simplification.

1.1.2. Requirements of Separation

The separation results (i.e. tissue images, A, and tissue TACs, X) should respect as close as possible physiological expectations of kidneys dynamics which will be briefly described here. In scintigraphy, the dynamics is defined as temporal activity of radioactive tracer in the tissues of interest which is illustrated in Figure 1.2 (right). At first, the radiopharmaceuticals are applied intravenously into the body and the blood background, heart, and lungs are activated, see Figure 1.2 (right, the first row). At second, the blood is cleaned from radiopharmaceuticals in the parenchyma, a spongy tissue of kidney, see Figure 1.2 (right, the second row). At third, waste products including radiopharmaceuticals comes from the parenchyma to the pelvis located on the inner edge of parenchyma, see Figure 1.2 (right, the third row). The pelvis tissue has a characteristic delay in its TAC, and the beginning of its activity roughly corresponds with the peak of activity of the parenchyma [32]. Finally, the activity is excluded from the pelvis to the urinary bladder, see Figure 1.2 (right, the forth row).

1.2. State of the Art

In this Section, we review the most common methods for the BSS. First, we will focus on the general BSS methods. Second, we will discuss the BSS methods in the context of medical image analysis.

1.2.1. Methods for Blind Source Separation

Principal Component Analysis (PCA) The PCA [82, 51, 57], the widely used dimension-reduction method, is based on orthogonal projection of the data space into the linear subspace with a lower dimension. The central idea of PCA is dimensionality reduction while preserving as much variation in a dataset as possible [57]. This is provided by transformation to a new set of variables, called principal components, which are uncorrelated and where the first few components preserve the most variation. The principal components are orthogonal because they are based on eigenvectors of the symmetric covariance matrix. Within the BSS problem, PCA is used for denoising while the final BSS solution can be found, e.g., using calculation of the suitable rotation matrix T from equation (1.4) [91].

Independent Component Analysis (ICA) In signal processing, the ICA [59] is a powerful tool for solving the BSS problem (1.1). The ICA model assumes that the original components are statistically independent. In its origin, ICA

considers only noise-free model which is, however, often unrealistic. The number of components is the same as the number of observations; however, this is not a necessary condition although the estimation of number of sources is not common within the ICA method.

Factor Analysis (FA) The FA model coincides with the one given in the equation (1.1). Here, observations are expressed as linear combinations of hypothetical variables, A and X, [89, 57, 4], except the error terms, E. The pair containing the source image and TAC is called the factor. Contrary to ICA models, the FA often assumes reduction from p to r parameters for data description.

Optimization Methods The BSS problem (1.1) can be interpreted as an optimization problem of finding matrices A and X, often non-negative, such that

$$D \approx AX^T. \tag{1.5}$$

This is the case of the non-negative matrix factorization (NMF) [68, 52], also known as non-negative matrix approximation. Commonly, the algorithms for NMF are iterative while iterations depend on quality of reached approximation defined by a chosen measure which is optimized [68]. Examples of the measure are, e.g., the square Euclidean distance (1.2) or the Kullback-Leibler divergence [63]. The measure substitutes detailed model of residuals by favoring some values over another; however, the inner dimension r has to be preselected and the results of factorization strongly depends on it.

Clustering Methods Clustering can be seen from different perspective as a search for clusters representing sources in the BSS problem [125]. In addition, a special case of the NMF can be seen from a different perspective as a K-mean clustering where the matrix A denotes cluster centroids and the matrix X denotes features for clusters membership [31].

The decomposition (1.5) has been studied as a problem of identifying of clusters around representative members. These members around which the clusters arise could be, e.g., pure-volume pixels [24] or pure source images in case of near-separable NMF where pure source image within the data is expected for each source in successive projection algorithm [5].

Bayesian Approach Typically, previous methods work well under ideal conditions such as low noise or known number of sources; however, they are limited under demanding conditions involving uncertainties. In such cases, the use of Bayesian methodology could provide more promising results [37, 97, 124].

The key advantages of Bayesian methods are that (i) they provide not only point estimates of parameters but their posterior distributions and (ii) Bayesian model selection properties can be used for selection of the number of sources. On the other hand, the Bayesian inference has limitation in tractability and

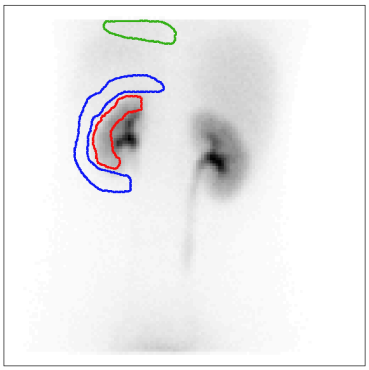

Figure 1.3.: Examples of ROIs: the ROI of parenchyma (red), the ROI of heart (green), and the ROI of background (blue).

full Bayesian solution for complex and realistic models is not possible and approximation has to be considered. In this work, we will take advantage of the Variational Bayes (VB) approximation method [115, 39, 8, 97] which is used to overcome these difficulties. The advantage of the VB method is that there is no need for sampling which is substituted by iterations in the VB inference.

1.2.2. Source Separation in Medical Imaging

Various methods have been used in order to separate the original sources from dynamic medical images. In clinical practice, the separation is done manually by an expert physician. Hence; manual-based methods are described at first and then we will continue to the automatized ones.

1.2.2.1. Manual Source Separation in Nuclear Medicine

In nuclear medicine, the analysis of image sequences is performed in few interconnected steps that will be discussed in the following paragraphs. The separation is typically done manually by an expert physician who defines regions of interests (ROIs) by drawing borders of the tissues of interest or their parts. These ROIs are used for further analysis using, e.g., the Patlak-Rutland plot or deconvolution.

Definition of Region of Interest The region of interest is, in manual medical image analysis context, a selected region with specific tissue, see examples in Figure 1.3: the ROI of parenchyma (red), the ROI of heart (green), and the ROI of background (blue). The problem can be also formalized as

$$\mathrm{ROI} = \begin{cases} 1 & \text{pixel belongs to the tissue of interest,} \\ 0 & \text{pixel does not belong to the tissue of interest.} \end{cases} \tag{1.6}$$

In current practice, the selection of the ROI is usually performed manually by a human operator [126]. The resulting TAC of the source is an integration of the activity over the selected ROI; hence, in general the aim is to select such part of the tissue that does not contain any activity from other tissues. It is easy to imagine that this manually drawn ROIs can suffer from inaccuracy due to overlapping or mixing of the sources and prone to errors [20]. The tendency to define as small ROI as possible in an attempt to exclude other structures fails in the case of weak and very noisy signal and sometimes it is almost impossible.

Many attempts have been made to automatically define ROIs and to avoid manual interaction [73, 86]. Various methodologies are based on the factor analysis model [11, 92] or cluster analysis approach [123, 24]; however, none of them are generally accepted. There were attempts to use regression analysis [74] to subtract the background as well as to use interpolation [35] which can separate the whole organ but not its parts.

Input Function Estimation The input function is the concentration of a radioactive tracer in the arterial blood [53]. The extraction of the input function is one of the step of manual analysis in dynamic nuclear medicine used in the diagnostic coefficients estimation. The input function is traditionally obtained from blood sampling [44] which is very invasive and often inappropriate in practice. A typical alternative practice is to select a ROI where clear blood or vascular structure is present and the TAC estimated from this region (after background correction) is assumed to be the input function. This step can be done manually or automatically [21, 77, 128, 78]; however, such a structure may not be present in the images. Recently, attempts were made to extract image-derived input function as well [117, 3, 65]. All of these methods were reported to provide good results in specific conditions and some of them are used by their authors but, again, none has been generally accepted in general clinical practice.

Quantitative Analysis ROIs selection and input function estimation are prerequisites for quantitative analysis the result of which is a specific parameter representing a functional aspect of a selected tissue. There are two main approaches in dynamic renal scintigraphy: the Patlak-Rutland plot and the deconvolution technique.

Patlak-Rutland plot [81, 66] is based on the assumption that a tissue or its part is an integrator of the input activity. Assume that $R(t)$ is the activity in the tissue ROI in time t and $P(t)$ is the activity in the blood. Then

$$R(t) = a \int P(t)\mathrm{d}t + bP(t), \tag{1.7}$$

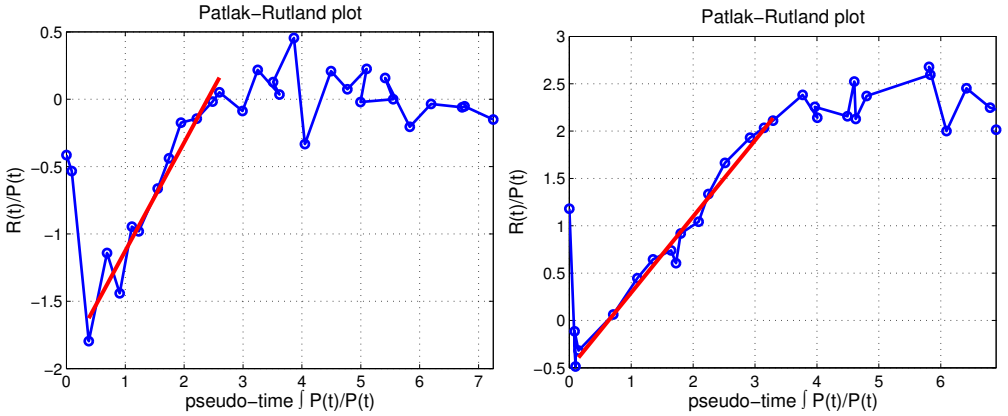

Figure 1.4.: Examples of Patlak-Rutland plots with selected linear parts (red lines).

where coefficient a denotes an ability of absorption of the tissue and coefficient b denotes the influence of the background.

Dividing by $P(t)$, the equation (1.7) can be rewritten as

$$\frac{R(t)}{P(t)} = a\frac{\int P(t)\mathrm{d}t}{P(t)} + b, \tag{1.8}$$

which is a linear model of $\frac{R(t)}{P(t)}$ as a function of $\frac{\int P(t)\mathrm{d}t}{P(t)}$. Estimation of coefficients a and b allows to estimate the proportion of the activity belonging to the tissue (a) and to the background (b). The estimation is performed on a manually selected region of linear part of the plot, see Figure 1.4.

Limitations of the method are sensitivity to the ROI selection, as already discussed, and in selection of the linear part of the Patlak-Rutland plot, see [99].

Deconvolution technique is based on assumption that the TAC of a tissue, $R(t)$, arises as a convolution of the input activity, $I(t)$, and tissue-specific kernel, $H(t)$, [64, 30, 33]. Mathematically,

$$R(t) = \int_0^T I(T-t)H(t)\mathrm{d}t, \tag{1.9}$$

where T is the total time of the measurement. Since the functions $R(t)$ and $I(t)$ can be measured, e.g., using ROI definition, the unknown convolution kernel $H(t)$ can be computed. The convolution kernel $H(t)$ is used for evaluation of diagnostic parameters [32].

However, the direct deconvolution methodology strongly depends on the expert, i.e. on proper tissue ROI selection, correct input function estimation, and correct tissue background and noise subtraction [46].

1.2.2.2. Automatic Blind Source Separation

Many attempts have been made to separate the measured signal automatically or semi-automatically to counteract the difficulties of manual-based methods. Thresholding methods, e.g., [116], are available; however, the thresholding is generally unstable. The automated ROIs detection was studied in [35] based on edge detection and contrast change detection; however, this work is focused only on strong tissues with low distinction of overlap sources and background subtraction using interpolation. The model of fuzzy ROI was established in [93] within the FA model, and the quality of separation of renal scan based on this model was discussed in [11] with promising results.

Clustering was successfully used in analysis of data from imaging modalities such as PET or MRI. In [123], the clustering for segmentation by time-activity curves is proposed. The algorithm segments the dynamic PET data into k clusters, based on the least-square distance measure. The number of clusters, k, has to be set manually. The least square fit is used in case of fMRI for dimension reduction in [25]; however, further analysis is done by clustered component analysis.

Since sources or at least their parts are clearly visible in images with high signal-to-noise ratio, the separation problem can be solved using techniques such as non-negative matrix factorization (NMF) [67, 41] or clustering-based methods [24, 70], where, typically, a pixel which belongs only to a single source need to be identified for each source to achieve uniqueness of solution [91]. However, such conditions are not typically met in real world measurement. In scintigraphy, the sources overlap with each other by their whole volume and the signal-to-noise ratio is intentionally kept low to minimize the radiation dose applied to patient. For this reason, models where the error part is taken into the account are required in the BSS of medical image data. Widely used model is the FA model [19]. The FA was used as a preprocessing for manual methods [2], or as a stand-alone method, e.g., in PET of the heart [62] or in renal scintigraphy [97, 98] where a biologically reasonable model was established and solved using the VB method.

1.3. Aim of the Work

The aim of this work is to study and extend blind source separation methodology with focus on blind source separation of scintigraphic image sequences. For this reason, we select the Variational Bayes methodology as the most, in our opinion, perspective way for inference of a wide range of blind source separation models.

The main aims of this work are:

- to transform the knowledge from nuclear medicine analysis (Section 1.2.2.1) into the BSS methodology, into the form of prior distribution for the Bayesian BSS,

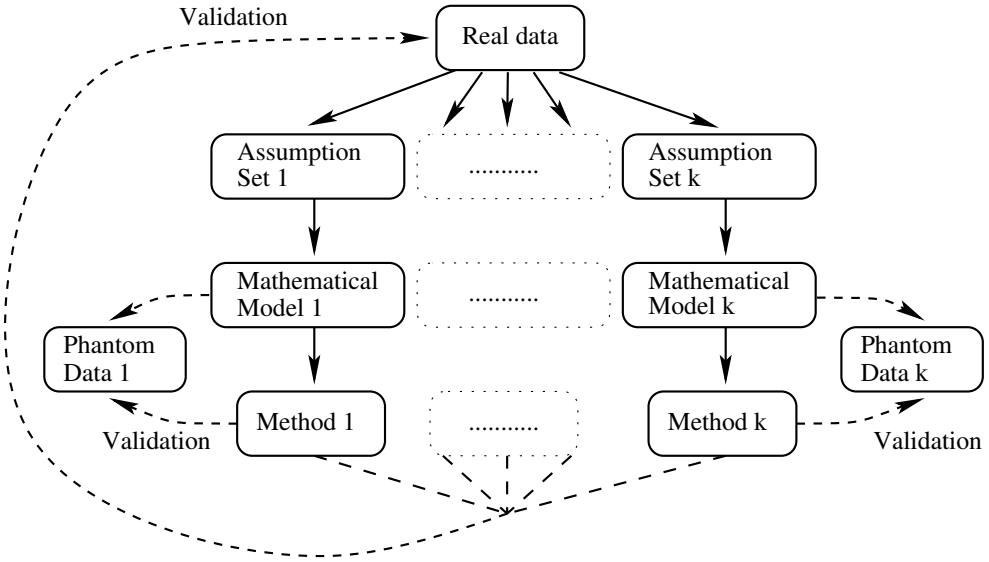

Figure 1.5.: Development-process of each method for medical data analysis.

- to develop a methodology for inference of these models,
- to test and to compare derived methods on real data from dynamic renal scintigraphy,
- to apply derived methods on other problems of BSS.

1.3.1. Methodology of Method Development

Development of methods for analyzing medical data is extremely difficult since the successful method must consider all uncertainty that could occur in medical practice such as noise and artifacts of an imaging method, non-physiological behavior or position of scanned tissues, or even completely rare cases such as three kidneys [105]. Hence, we would like to formalize a general approach how we develop and validate methods for dynamic medical image analysis in this work. The development-process is displayed in Figure 1.5.

Based on the considered application domain and properties of real data acquisition, a set of assumptions can be created. The assumption set should reflect essential attributes of the data such as properties of noise, elasticity and movement of observed sources, the number of sources, the character of TACs, and others. Then, the assumptions are transformed into the mathematical model for which a method of parameter inference is derived.

The derived method will be validated in two stages. First, we validate each method on a synthetically generated data (the phantom data) generated from the assumed mathematical model. Correct inference of the parameters on this

data validates correct implementation of the method and suitability of the required approximations. Second, we use real data to validate the performance of proposed methods in realistic conditions. Performance of the method on this data indicates validity of the selected assumptions for the real application. Very often, the validation on real data is very difficult since no ground truth is available and methodology for this validation should be developed.

1.4. Layout of the Work

The work is organized as follows:

Chapter 2: The basic Bayesian theory that is relevant to blind source separation is introduced. The focus will be given to the Variational Bayes method used through the whole work and to the automatic relevance determination principle for which illustrative examples are introduced. A blind source separation problem is studied and matrix formulation of this problem is introduced.

Chapter 3: Prior models of the noise within the BSS problem are studied. Prior models of the source images are studied with emphasis on models enforcing sparsity of the source images. Prior models of time-activity curves follow with emphasis on convolution parameterization of the TAC. Prior models for the parameters of the convolution parametrization (i.e. the convolution kernels and the input function) are introduced. For all priors, the Variational Bayes method was used to derive equations for shaping parameters of the corresponding posterior distributions.

Chapter 4: Priors from the previous chapter are combined to obtain various blind source separation methods with different purpose and performance. The initialization, estimation of number of sources, and numerical problems are discussed. State of the art methods appropriate for blind source separation of dynamic image data are introduced and tested together with the proposed methods on a synthetic phantom dataset.

Chapter 5: Key application area of this work is dynamic renal scintigraphy. Here, the qualitative evaluation of the separation of data from dynamic renal scintigraphy is discussed and illustrated. Then, quantitative experiments with large datasets are conducted with both, the proposed and state of the art methods.

Chapter 6: Experiments on various image sequence data are conducted. Specifically, data from dynamic positron emission tomography, functional magnetic resonance imaging, and hyperspectral imaging are used to study behavior of the proposed algorithms under various conditions.

Chapter 7: We summarize the main contributions of the work and point some possible and interesting ways for further research which are demonstrated on preliminary studies.

2. Approximate Bayesian Inference

Bayes' theorem has come back from the cemetery to which it has been consigned.

(Jerome Cornfield)

Bayesian probability theory [13, 50] offers a theoretical background that allows to deal with uncertainty and our beliefs and knowledge. In this chapter, we will describe the basics of Bayesian theory. Since the inference is often intractable, we will continue with approximation methods with special focus on the Variational Bayes approximation.

2.1. Bayesian Theory

The probability density function of continuous random vector $\boldsymbol{\theta}$, $f(\boldsymbol{\theta})$, obeys

$$f(\boldsymbol{\theta}) \geqq 0, \quad \forall \boldsymbol{\theta}, \tag{2.1}$$

$$\int_{\Theta} f(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} = 1. \tag{2.2}$$

Assume that D is the measured data and $\boldsymbol{\theta} \in \Theta$ represents parameters of the data model where Θ is the space of parameters $\boldsymbol{\theta}$. A parametric probabilistic model of the data can be given as density function $f(D|\boldsymbol{\theta})$ conditioned by model parameters $\boldsymbol{\theta}$. Our beliefs and knowledge about these parameters $\boldsymbol{\theta}$ is expressed using prior distribution $f(\boldsymbol{\theta})$. Probability distribution over parameters $\boldsymbol{\theta}$ after considering the data D is expressed by posterior density function $f(\boldsymbol{\theta}|D)$. The form of the posterior can be found using the Bayes theorem

$$f(\boldsymbol{\theta}|D) = \frac{f(\boldsymbol{\theta}, D)}{f(D)} = \frac{f(D|\boldsymbol{\theta}) f(\boldsymbol{\theta})}{\int_{\Theta} f(D|\boldsymbol{\theta}) f(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}}. \tag{2.3}$$

The term $\int_{\Theta} f(D|\boldsymbol{\theta}) f(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}$ is the normalizing constant and can be omitted using proportional equality as

$$f(\boldsymbol{\theta}|D) \propto f(D|\boldsymbol{\theta}) f(\boldsymbol{\theta}), \tag{2.4}$$

while the evaluation of the normalizing constant is often very expensive or even intractable.

The moment of a function of parameter $\boldsymbol{\theta}$, $g(\boldsymbol{\theta})$, is defined as

$$\mathrm{E}_{f(\boldsymbol{\theta}|D)}\left(g(\boldsymbol{\theta})\right) = \int_{\Theta} g(\boldsymbol{\theta}) f(\boldsymbol{\theta}|D) \mathrm{d}\boldsymbol{\theta}, \tag{2.5}$$

which will be used in a simple notation (when obvious) as $\widehat{g(\boldsymbol{\theta})}$.

2.1.1. Choice of Prior Distribution

The choice of prior distribution, $f(\boldsymbol{\theta})$, is a subjective and important task when the model is designed [97, 76]. Generally, when the data are expected to be more informative, the prior distribution is chosen as less informative to minimize its influence on posterior distribution $f(\boldsymbol{\theta}|D)$.

In this work, we need to consider the practical impact of chosen prior on tractability of a problem, hence, the prior distributions will be chosen:

1. to supplement the data in cases where the data are demanding or the model is poorly defined which is called regularization via the prior [97],
2. to reflect various restrictions on the model parameters,
3. to express the ignorance about the model parameters, e.g., non-informative prior in case of the information data is expected.

Typically, we will work with conjugate priors. In inference of parametric distributions, all distributions have a known functional form determined by its shaping parameters. It is beneficial when the prior functional form $f_{\mathrm{conjug}}(\boldsymbol{\theta})$ remains the same in posterior distribution after application of the Bayes rule:

$$f_{\mathrm{conjug}}(\boldsymbol{\theta}|D) \propto f(D|\boldsymbol{\theta}) f_{\mathrm{conjug}}(\boldsymbol{\theta}). \tag{2.6}$$

The distribution f_{conjug} is known as self-replicating [13] or conjugate distribution to observation model [97].

2.1.2. Model Selection

We might wish to select a model from a set of models for the given data. In Bayesian theory, unknown model is treated as unknown variable and under this designation belongs task such as structure learning, cardinality, or dimensionality inferring [8]. Here, consider a simple task of decision between two models, $M_1(\boldsymbol{\theta})$ and $M_2(\boldsymbol{\theta})$, for the given data D. The probability of data under the given model and its parameters is $f\left(D|M_1(\boldsymbol{\theta})\right)$ or $f\left(D|M_2(\boldsymbol{\theta})\right)$. Then, the Bayes rule provides a way for computing the probability of the model in the light of the observed data as

$$f\left(M_i(\boldsymbol{\theta})|D\right) \propto f\left(D|M_i(\boldsymbol{\theta})\right) f\left(M_i(\boldsymbol{\theta})\right), \quad \forall i, \tag{2.7}$$

where $f\left(M_i(\boldsymbol{\theta})\right)$ is the prior of the model $M_i(\boldsymbol{\theta})$.

2.2. Approximate Bayesian Inference

Bayes rule (2.3) provides the theoretical update of distribution over parameters with incoming data. However, these integrals are analytically intractable in most cases and numerical integration is not suitable for high-dimensional integrals. For this reason, various approximate methods for solving this issue were proposed.

2.2.1. Maximum a Posteriori Method

The simplest way to approximate the posterior distribution is to approximate it by the Dirac delta function δ. Maximum a posteriori method (MAP) is closely related to the maximum likelihood method [13]; however, prior information can be incorporated into an estimate using straightforward application of the Bayes theorem (2.3) in MAP as

$$\widehat{\boldsymbol{\theta}}_{\mathrm{MAP}} = \arg\max_{\boldsymbol{\theta}} f(\mathbf{d}_1, \ldots, \mathbf{d}_n|\boldsymbol{\theta}) f(\boldsymbol{\theta}), \tag{2.8}$$

where $f(\boldsymbol{\theta})$ is prior distribution of parameter $\boldsymbol{\theta}$. Then, MAP approximates $f(\boldsymbol{\theta}|D)$ as

$$f(\boldsymbol{\theta}|D) \approx \delta\left(\boldsymbol{\theta} - \boldsymbol{\theta}_{\mathrm{MAP}}\right). \tag{2.9}$$

2.2.2. Laplace Approximation

The method is useful approximation of the density function $f(\boldsymbol{\theta}|D)$ at the MAP estimate $\widehat{\boldsymbol{\theta}}$ using Normal distribution as

$$f(\boldsymbol{\theta}|D) \approx \mathcal{N}\left(\widehat{\boldsymbol{\theta}}, -H^{-1}\right), \tag{2.10}$$

where the matrix $H(\widehat{\boldsymbol{\theta}})$ is Hessian matrix defined as

$$H\left(\widehat{\boldsymbol{\theta}}\right)_{i,j} = \left.\frac{\partial^2 \log f(\boldsymbol{\theta}|D)}{\partial\theta_i \partial\theta_j}\right|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}}, \quad i,j = 1, \ldots, p, \tag{2.11}$$

for p-dimensional vector $\boldsymbol{\theta}$. See [60, 8] for more details.

2.2.3. Markov Chain Monte Carlo Method

Markov Chain Monte Carlo (MCMC) approximation is a strategy for approximation of a posterior density function using histogram constructed from a sequence of random samples of variable $\boldsymbol{\theta}$, $\left\{\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(n)}\right\}$. This sequence is called Markov Chain if

$$f\left(\boldsymbol{\theta}^{(n)}|\boldsymbol{\theta}^{(n-1)}, \ldots, \boldsymbol{\theta}^{(1)}\right) = f\left(\boldsymbol{\theta}^{(n)}|\boldsymbol{\theta}^{(n-1)}\right), \tag{2.12}$$

i.e. if the nth sample $\boldsymbol{\theta}^{(n)}$ is generated from conditional distribution $f\left(\boldsymbol{\theta}^{(n)}|\boldsymbol{\theta}^{(n-1)}\right)$ depending only on previous state $\boldsymbol{\theta}^{(n-1)}$.

Then, the expectation of the function $g(\boldsymbol{\theta})$ under the density function $f(\boldsymbol{\theta}|D)$ using generating of samples $\boldsymbol{\theta}^{(i)} \sim f(\boldsymbol{\theta}|D)$ can be computed as unbiased estimate using N samples as

$$I_N(g) = \frac{1}{N}\sum_{i=1}^{N} g(\boldsymbol{\theta}^{(i)}) \simeq \int_{\Theta} g(\boldsymbol{\theta}) f(\boldsymbol{\theta}|D) \mathrm{d}\boldsymbol{\theta}. \tag{2.13}$$

The more samples are taken, the more reliable estimates are obtained. This approach has typically computational problems when the dimension of a problem is high.

2.2.4. Expectation Maximization Algorithm

The expectation maximization (EM) algorithm [28] is an iterative approach for estimation of a subset of model parameters of interest while the model also depends on unobserved latent variables. The EM algorithm provides expectation step, which computes posterior distribution over latent variables using estimates of parameters of interest, and maximization step, which computes model parameters that maximizing the log-likelihood found in the expectation step.

Formally, suppose notation (t) for number of iteration and two-dimensional parameter $\boldsymbol{\theta} = [\theta_1, \theta_2]$. Then:

E-step: $\quad f^{(t+1)}\left(\theta_1|D\right) \approx f\left(\theta_1|D, \widehat{\theta_2}^{(t)}\right),$

M-step: $\quad \widehat{\theta_2}^{(t+1)} = \arg\max_{\theta_2} \int_{\theta_1} f^{(t+1)}\left(\theta_1|D\right) \ln f\left(\theta_1, \theta_2, D\right) \mathrm{d}\theta_1.$

2.3. Variational Bayes Approximation

Variational Bayes (VB) approximation is an effective way to design tractable inference for parametric probabilistic models. The key step is to replace the intractable integration in marginalization. We will review the basics of the VB method [97, 8] (also known as ensemble learning [76]).

The task is to find optimal function $\breve{f}(\boldsymbol{\theta}|D) \in \mathbf{F}_c$, where $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_q]$ and $\mathbf{F}_c$ is the space of conditionally independent distributions,

$$\mathbf{F}_c = \left\{ f(\theta_1, \ldots, \theta_q|D) \mid f(\theta_1, \ldots, \theta_q|D) = \prod_{i=1}^{q} f(\theta_q|D) \right\}. \tag{2.14}$$

The number of parameters is assumed to be $q > 1$. Since we try to find approximate function $\breve{f}(\boldsymbol{\theta}|D)$, it is necessary to measure the proximity of the $\breve{f}(\boldsymbol{\theta}|D)$ to the true function $f(\boldsymbol{\theta}|D)$. It was shown that optimal loss function is logarithmic when the task is to extract maximum information from the data [12] which

leads to use of Kullback-Leibler divergence (KLD) [63] (also known as information divergence or relative entropy) between approximative and true function. The KLD is defined as

$$\mathrm{KLD}\left(\breve{f}(\boldsymbol{\theta}|D)||f(\boldsymbol{\theta}|D)\right) = \int_{\Theta} \breve{f}(\boldsymbol{\theta}|D) \ln \frac{\breve{f}(\boldsymbol{\theta}|D)}{f(\boldsymbol{\theta}|D)} \mathrm{d}\boldsymbol{\theta}. \tag{2.15}$$

$\mathrm{KLD}(.||.) \geq 0$ and $\mathrm{KLD}(.||.) = 0$ if arguments are equal almost everywhere. Moreover, $\mathrm{KLD}(\breve{f}||f) \neq \mathrm{KLD}(f||\breve{f})$. In the following text, we will use the version (2.15) in VB approximation.

The VB theorem for distributional approximation is then [97]:

Theorem 1. *Let $f(\boldsymbol{\theta}|D)$ is the posterior distribution of the multivariate parameter $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_q]$. Let $\breve{f}(\theta|D)$ be an approximate distribution form a set of conditionally independent distribution for $\theta_1, \ldots, \theta_q$ such that*

$$\breve{f}(\boldsymbol{\theta}|D) = \prod_{i=1}^{q} \breve{f}(\theta_i|D). \tag{2.16}$$

Then, the minimum of the KLD

$$\tilde{f}(\boldsymbol{\theta}|D) = \arg\min_{\breve{f}(\boldsymbol{\theta}|D)} \mathrm{KLD}\left(\breve{f}(\boldsymbol{\theta}|D)||f(\boldsymbol{\theta}|D)\right) \tag{2.17}$$

is reached for

$$\tilde{f}(\theta_i|D) \propto \exp\left(\mathrm{E}_{\tilde{f}(\boldsymbol{\theta}_{/i}|D)}\left(\ln f(\boldsymbol{\theta}, D)\right)\right), \quad i = 1, \ldots, q, \tag{2.18}$$

where $\boldsymbol{\theta}_{/i}$ denotes the complement of θ_i in $\boldsymbol{\theta}$. The $\tilde{f}(\theta_i|D)$ will be refereed to as the VB-marginal.

The proof of the Theorem 1 can be found e.g. in [97] or in [76]. The approximation (2.18) is deterministic but not unique. The approximation (2.18) forms a set of implicit equations that leads naturally to an iterative algorithm, see iterative VB (IVB) Algorithm 2.1.

The IVB algorithm is closely related to the EM algorithm, Section 2.2.4, and is also known as the Variational EM algorithm [9].

Remark 2. The reverse order of arguments in KLD, equation (2.15), leads to the Expectation-Propagation algorithm, see [75, 12] for details. However, this approach is not suitable for large dimensional problems.

2.3.1. Review of the Variational Bayes Method

We review the systematic way for using the VB approximation (2.18) for Bayesian inference, the VB method [96, 97].

Algorithm 2.1 Iterative Variational Bayes Algorithm.

The following steps monotonically decrease KLD in Theorem 1:

1. Set the initialization $\widetilde{f}^{[0]}(\theta|D)$ and counter $n = 1$.
2. While stopping rule is not met, run the following steps:
   a) Compute update of the VB-marginal $\widetilde{f}(\theta_i|D)$ at the nth iteration for $i = 1, \ldots, q$ using (2.18):

$$\widetilde{f}^{[n]}(\theta_i|D) \propto \exp \int_{\Theta_{/i}} \widetilde{f}^{[n-1]}(\theta_{/i}|D) \ln f(\theta, D) \mathrm{d}\theta_{/i}, \tag{2.19}$$

   b) $n = n + 1$.
3. Report estimates $\widehat{\theta}_i$ for $i = 1, \ldots, q$.

Step 1: Choose a Model

Choice of the joint distribution $f(\boldsymbol{\theta}, D)$; i.e. observations model of the data, $f(D|\boldsymbol{\theta})$, and prior model of parameters, $f(\boldsymbol{\theta})$; have to be done.

Step 2: Parameters Separation

Partition of $\boldsymbol{\theta}$ into q sub-vectors is a crucial test if the VB approximation is applicable to the selected Bayesian model. It has to ensure that it is possible to expand natural logarithm $\ln f(\boldsymbol{\theta}, D)$ as (for, e.g., $q = 2$)

$$\ln f(\boldsymbol{\theta}, D) = f_1(\theta_1, D) f_2(\theta_2, D), \tag{2.20}$$

where functions f_1, f_2 on the right side are vectors of compatible dimensions. If so, the $f(\boldsymbol{\theta}, D)$ is said to be in separable-in-parameters family. If this condition does not hold, the VB approximation is not suitable for the model.

Step 3: VB-marginals

Direct application of the VB theorem to ith parameter, $i = 1, \ldots, q$, is straightforward:

$$\tilde{f}(\theta_i|D) \propto \exp\left[\mathrm{E}_{\tilde{f}(\boldsymbol{\theta}_{/i}|D)}(\ln f(\boldsymbol{\theta}, D))\right] \propto \exp\left[f_i(\theta_i, D) \prod_{j=1, j\neq i}^{q} \widehat{g_j(\theta_j, D)}\right], \tag{2.21}$$

where symbol $\widehat{g(..)}$ denotes the VB-moments.

Step 4: Identification of Standard Distributional Forms

A standard functional form of (2.21) has to be identified. The VB-moments of $\widehat{g_j(\theta_j, D)}$ are taken as constants. The standard forms are

$$\tilde{f}(\theta_i|D) = f_{\text{std}}(\theta_i|\vartheta_i), \tag{2.22}$$

where f_{std} is the standard distributional form (examples are listed in Appendix A) and ϑ_i is a set of parameters of the standard form, called the shaping parameters.

Step 5: VB-moments Formulation

The required VB-moments of $\widehat{g_j(\theta_j, D)}$ are functions of shaping parameters and typically can be listed in standard parametric distributions, see Appendix A.

The equations for the shaping parameters together with those for the VB-moments form a set of VB-equations to be solved.

Step 6: VB-equations Reduction

This step is technical and we present it in order to be consistent with [97]. The set of VB-equations is typically implicit; however, a subset could have an explicit solution or could be simplified.

Step 7: Run IVB Algorithm

The IVB algorithm, Algorithm 2.1, was already described. Special care should be given to the choice of initial shaping parameters. They can be set randomly [15]; however, a good problem-specific initial guess can lead to better and faster solution since this algorithm is gradient-based and converges only to a local minimum. The number of iterations should be also carefully considered and will be discussed in the following text in concrete cases.

Step 8: Report VB-marginals

The results of the IVB algorithm are in the form of shaping parameters of the standard forms and related VB-moments.

Remark 3. We will use the steps of the VB method in this work; however, we will not refer to their numbers in text but we will use the terminology defined here which helps to identify the individual steps.

2.3.2. Message Passing in Variational Bayes Method

The implementation of the IVB algorithm is often presented in the form of message passing [119] which allows to generalize the algorithm to almost arbitrary model. We will discuss a special case of the message passing idea between

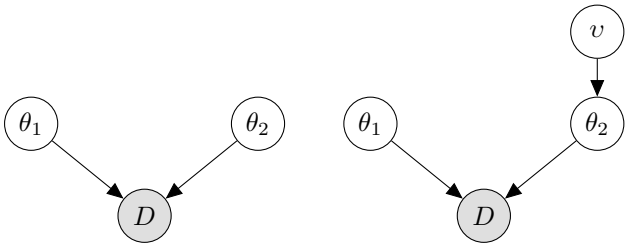

Figure 2.1.: Graphical hierarchical prior model for demonstration of communication flow within the IVB algorithm.

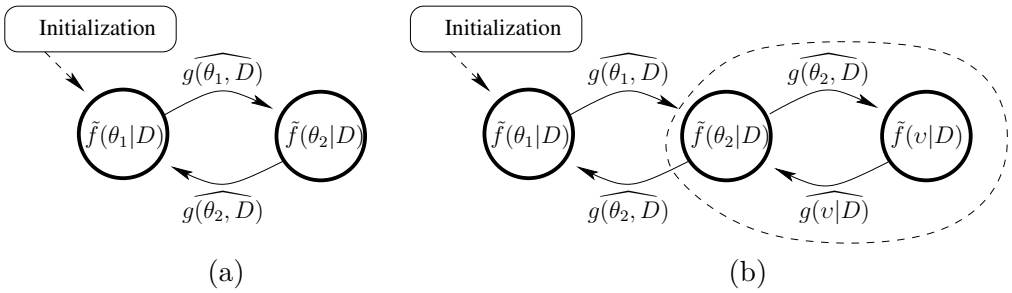

Figure 2.2.: Example scheme of the IVB algorithm for the graphical hierarchical prior models from Figure 2.1.

the VB-marginals within the IVB algorithm. Consider hierarchical prior model where the data D is modeled using 2 parameters: θ_1 and θ_2, see the hierarchical model in Figure 2.1, left. Formally, the likelihood and the prior models are

$$f(D|\theta_1, \theta_2), \quad f(\theta_1), \quad f(\theta_2). \tag{2.23}$$

Following the VB method, the natural logarithm of the joint distribution is expanded as

$$\ln f(D, \theta_1, \theta_2) = \ln f(D|\theta_1, \theta_2) + \ln f(\theta_1) + \ln f(\theta_2). \tag{2.24}$$

We impose the conditional independence between parameters θ_1 and θ_2. Then, the term $\ln f(\theta_2)$ is considered as a constant in the inference for parameter θ_1 while moments of θ_2 are used for VB-marginal of the parameter θ_1 in the term $\ln f(D|\theta_1, \theta_2)$, see the illustration of the IVB algorithm for this example in Figure 2.2a, and vice versa for the parameter θ_2.

Consider now another similar hierarchical model where the data D is modeled using 2 parameters: θ_1 and θ_2, however, the parameter θ_2 has one additional prior hyper-parameter υ, see graphical model in Figure 2.1, right. Formally, the likelihood and the prior models are

$$f(D|\theta_1, \theta_2), \quad f(\theta_1), \quad f(\theta_2|\upsilon)f(\upsilon). \tag{2.25}$$

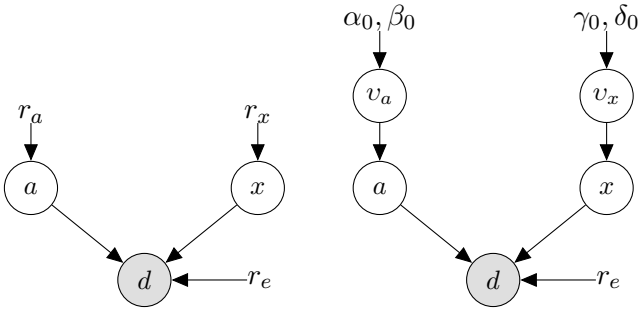

Figure 2.3.: Graphical hierarchical prior model of the scalar multiplicative decomposition (left) and scalar multiplicative decomposition with ARD prior (right).

The natural logarithm of the joint distribution is

$$\ln f(D, \theta_1, \theta_2, \upsilon) = \ln f(D|\theta_1, \theta_2) + \ln f(\theta_1) + \ln f(\theta_2|\upsilon) + \ln f(\upsilon) \tag{2.26}$$

Let us focus on the inference of parameter θ_2. Since hyper-parameter υ appears only in the model of the parameter θ_2, the only VB-marginal that needs the moments of υ is the VB-marginal $\tilde{f}(\theta_2|D)$ and similarly, the VB-marginal $\tilde{f}(\upsilon|D)$ needs only the moments of θ_2 as demonstrated in Figure 2.2b.

The inference of the parameter θ_2 and its hyper-parameter υ provides only moments $g(\widehat{\theta_2, D})$ for the remaining VB-marginals of the IVB algorithm. This is demonstrated using dashed circle in Figure 2.2b. This implies significant simplification of software implementation of the IVB algorithm since various priors for the parameters influence only a small part of the algorithm. Implementation via the message passing idea thus allows to write a modular software with clearly defined interface.

2.3.3. Scalar Example

We will demonstrate the VB method on a simple scalar example as proposed in [97]. Consider scalar model

$$d = ax + e, \tag{2.27}$$

where the term ax is supposed to be the signal and the term e is supposed to be the noise. To separate the signal from the noise, the model must be regularized since infinitely many solutions are possible otherwise. Let us assume that e is normally distributed, see Appendix A.1, $e \sim \mathcal{N}_e(0, r_e)$, then

$$f(d|a, x, r_e) = \mathcal{N}_d(ax, r_e). \tag{2.28}$$

Even with this assumption, the solution holds $\widehat{a}\widehat{x} = d$ is as good as those with rotation l: $(\widehat{a}l)\,(l^{-1}\widehat{x}) = d$. Further selection of prior models for a and x is necessary to achieve unique solution.

Suppose prior models for a and x to be

$$f(a|r_a) = \mathcal{N}_a(0, r_a), \tag{2.29}$$
$$f(x|r_x) = \mathcal{N}_x(0, r_x), \tag{2.30}$$

with given r_a and r_x, see graphical model in Figure 2.3, left. Then, the VB theorem yields the VB-marginals:

$$\tilde{f}(a|d) \propto \exp\left(-\frac{1}{2}a^2(\widehat{x}^2 r_e^{-1} + r_a^{-1}) - a(d\widehat{x}r_e^{-1})\right), \tag{2.31}$$
$$\tilde{f}(x|d) \propto \exp\left(-\frac{1}{2}x^2(\widehat{a}^2 r_e^{-1} + r_x^{-1}) - x(d\widehat{a}r_e^{-1})\right), \tag{2.32}$$

where standard distributional forms are easily recognized to be Normal distributions:

$$\tilde{f}(a|d) = \mathcal{N}_a(\mu_a, \sigma_a), \tag{2.33}$$
$$\tilde{f}(x|d) = \mathcal{N}_x(\mu_x, \sigma_x), \tag{2.34}$$

with shaping parameters

$$\sigma_a = (\widehat{x}^2 r_e^{-1} + r_a^{-1})^{-1}, \qquad \mu_a = \sigma_a d\widehat{x}r_e^{-1}, \tag{2.35}$$
$$\sigma_x = (\widehat{a}^2 r_e^{-1} + r_x^{-1})^{-1}, \qquad \mu_x = \sigma_x d\widehat{a}r_e^{-1}. \tag{2.36}$$

The required VB-moments of the Normal distribution are

$$\widehat{a} = \mu_a, \qquad \widehat{a^2} = \mu_a^2 + \sigma_a, \tag{2.37}$$
$$\widehat{x} = \mu_x, \qquad \widehat{x^2} = \mu_x^2 + \sigma_x. \tag{2.38}$$

The set of equations (2.35)–(2.38) can be solved in different ways as shown in Figure 2.4.

1. Exact solution can be found for this scalar example, see [97], in Figure 2.4, red cross.
2. MAP solution for this scalar problem, see [97], in Figure 2.4, red star.
3. Using IVB algorithm, Figure 2.4, full line with final estimate by circle.
4. Using a method for solution of linear equation problem, e.g. conjugate gradients (CG):

$$\begin{bmatrix} \mu_a \\ \sigma_a \\ \mu_x \\ \sigma_x \end{bmatrix} = \mathrm{diag}\left(\begin{bmatrix} (x^2 r_e^{-1} + r_a^{-1}) \\ (x^2 r_e^{-1} + r_a^{-1}) \\ (a^2 r_e^{-1} + r_x^{-1}) \\ (a^2 r_e^{-1} + r_x^{-1}) \end{bmatrix}\right) \begin{bmatrix} dxr_e^{-1} \\ 1 \\ dar_e^{-1} \\ 1 \end{bmatrix}, \tag{2.39}$$

   Figure 2.4, dashed line with final estimate by circle.

We conclude that the convergence is reached much faster using the IVB algorithm than using the CG solution for this particular example.

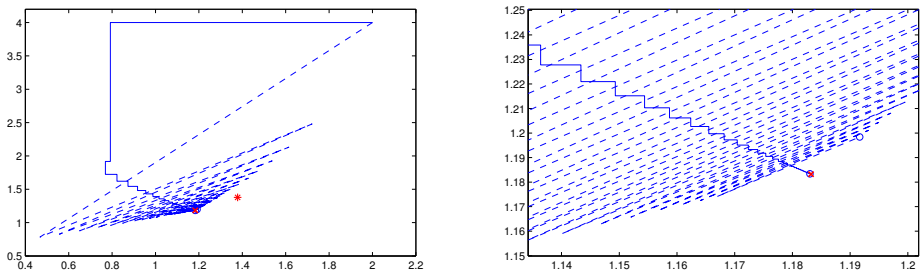

Figure 2.4.: VB-approximation for the scalar decomposition where $d = 2$, $r_e = 1$, $r_a = 10$, $r_x = 10$. The star denotes MAP solution, the full line denotes solution using IVB algorithm, and the dashed line denotes solution using conjugate gradients algorithm.

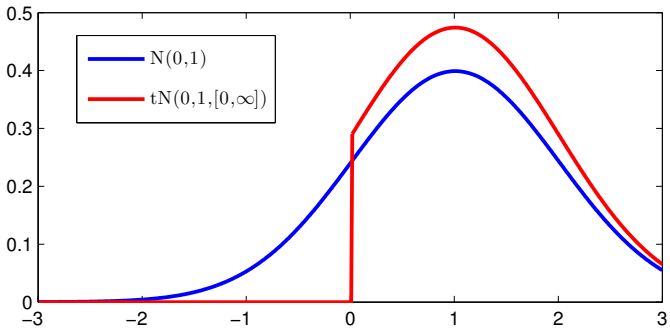

Figure 2.5.: Example of the normal distribution $\mathcal{N}(1, 1)$, blue line, and the truncated normal distribution $t\mathcal{N}(1, 1, [0, \infty])$, red line.

2.3.3.1. Solution in Positive Domain

Consider again scalar model (2.27); however, assume the parameters a and x to be positive. Then, the priors of a and x, (2.29)–(2.30), should be changed in order to reflect positivity. Here, we use the truncated normal distribution for this purpose and reformulate the prior distribution for these parameters as

$$f(a|r_a) = t\mathcal{N}_a(0, r_a, [0, \infty]), \tag{2.40}$$
$$f(x|r_x) = t\mathcal{N}_x(0, r_x, [0, \infty]), \tag{2.41}$$

where $t\mathcal{N}$ denotes the truncated normal distribution defined in Appendix A.1.3 on defined support. An example of the truncated normal distribution is given in Figure 2.5. Then, the VB-marginals have the standard distributional forms

of the prior and are recognized to be

$$\tilde{f}(a|d) = t\mathcal{N}_a(\mu_a, \sigma_a, [0, \infty]), \tag{2.42}$$
$$\tilde{f}(x|d) = t\mathcal{N}_x(\mu_x, \sigma_x, [0, \infty]), \tag{2.43}$$

where the shaping parameters are computed using the same equations as in the previous case, (2.35)–(2.36). The difference is in computation of the VB-moments which are non-trivial and can be found in Appendix A.1.3. Using notation from Appendix A.1.4, the VB-moments are computed as

$$\widehat{a} = \mathrm{M}_1^{\mathrm{tN}}(\mu_a, \sigma_a, 0, \infty), \tag{2.44}$$
$$\widehat{a^2} = \mathrm{M}_2^{\mathrm{tN}}(\widehat{a}, \mu_a, \sigma_a, 0, \infty), \tag{2.45}$$
$$\widehat{x} = \mathrm{M}_1^{\mathrm{tN}}(\mu_x, \sigma_x, 0, \infty), \tag{2.46}$$
$$\widehat{x^2} = \mathrm{M}_2^{\mathrm{tN}}(\widehat{x}, \mu_x, \sigma_x, 0, \infty). \tag{2.47}$$

2.4. Automatic Relevance Determination

Automatic relevance determination (ARD) principle [15, 114, 120] is based on joint estimation of the parameters of the prior together their variances in the VB inference. Specifically, the parameter θ has an unknown precision υ which is assumed to have conjugate Gamma prior

$$f(\theta|\upsilon) = \mathcal{N}_\theta(0, \upsilon^{-1}), \tag{2.48}$$
$$f(\upsilon) = \mathcal{G}_\upsilon(\alpha_0, \beta_0), \tag{2.49}$$

with scalar prior parameters α_0, β_0. The ARD principle is an effect when the expected value of the prior precision $\widehat{\upsilon}^{-1}$ approaches to zero which also tighten the expected value of the parameter of interest, $\widehat{\theta}$, close to zero.

We will demonstrate this principle on scalar decomposition example from Section 2.3.3.

2.4.1. Relation to Model selection

The ARD principle can be seen as a special version of model selection, Section 2.1.2. Consider a problem of selection between two models:

$$M_1 : \boldsymbol{\theta} = [\theta_1, \theta_2], \tag{2.50}$$
$$M_2 : \boldsymbol{\theta} = [\theta_1, 0]. \tag{2.51}$$

Let us focus on parameter θ_2. In the sense of ARD, we would select the prior for θ_2 as

$$f(\theta_2|\upsilon_2) = \mathcal{N}_{\theta_2}\left(0, \upsilon_2^{-1}\right) \tag{2.52}$$

with Gamma prior for υ_2 as $f(\upsilon_2) = \mathcal{G}_{\upsilon_2}(\alpha_0, \beta_0)$. Joint estimation of parameter of interest, θ_2, and its precision, υ_2^{-1}, on the VB method yields $\upsilon_2^{-1} \approx 0$ resulting in selection of the model M_2 or $\upsilon_2^{-1} \gg 0$ resulting in selection of the model M_1.

2.4.2. Scalar Example with ARD

Consider again the scalar example (2.27) with $e \sim \mathcal{N}_e(0, r_e)$, however, the prior model is now the ARD prior for a and x as follows:

$$f(d|a, x, \upsilon_a, \upsilon_x) = \mathcal{N}_d(ax, r_e), \tag{2.53}$$
$$f(a|\upsilon_a) = \mathcal{N}_a(0, \upsilon_a^{-1}), \tag{2.54}$$
$$f(\upsilon_a) = \mathcal{G}_{\upsilon_a}(\alpha_0, \beta_0), \tag{2.55}$$
$$f(x|\upsilon_x) = \mathcal{N}_x(0, \upsilon_x^{-1}), \tag{2.56}$$
$$f(\upsilon_x) = \mathcal{G}_{\upsilon_x}(\gamma_0, \delta_0), \tag{2.57}$$

with given prior parameters $\alpha_0, \beta_0, \gamma_0, \delta_0$, see graphical model in Figure 2.3, right. Then, the VB theorem yields the VB-marginals:

$$\tilde{f}(a|d) \propto \exp\left(-\frac{1}{2}a^2(\widehat{x^2}r_e^{-1} + \widehat{\upsilon_a}) - a(d\widehat{x}r_e^{-1})\right), \tag{2.58}$$
$$\tilde{f}(\upsilon_a|d) \propto \exp\left(\left(\frac{1}{2} + \alpha_0 - 1\right)\ln \upsilon_a - \left(\frac{\widehat{a^2}}{2} + \beta_0\right)\upsilon_a\right), \tag{2.59}$$
$$\tilde{f}(x|d) \propto \exp\left(-\frac{1}{2}x^2(\widehat{a^2}r_e^{-1} + \widehat{\upsilon_x}) - x(d\widehat{a}r_e^{-1})\right), \tag{2.60}$$
$$\tilde{f}(\upsilon_x|d) \propto \exp\left(\left(\frac{1}{2} + \gamma_0 - 1\right)\ln \upsilon_x - \left(\frac{\widehat{x^2}}{2} + \delta_0\right)\upsilon_x\right), \tag{2.61}$$

which are recognized to be in standard distributional forms:

$$\tilde{f}(a|d) = \mathcal{N}_a(\mu_a, \sigma_a), \tag{2.62}$$
$$\tilde{f}(\upsilon_a|d) = \mathcal{G}_{\upsilon_a}(\alpha, \beta), \tag{2.63}$$
$$\tilde{f}(x|d) = \mathcal{N}_x(\mu_x, \sigma_x), \tag{2.64}$$
$$\tilde{f}(\upsilon_x|d) = \mathcal{G}_{\upsilon_x}(\gamma, \delta), \tag{2.65}$$

with shaping parameters

$$\sigma_a = (\widehat{x^2}r_e^{-1} + \widehat{\upsilon_a})^{-1}, \qquad \mu_a = \sigma_a d\widehat{x}r_e^{-1}, \tag{2.66}$$
$$\alpha = \frac{1}{2} + \alpha_0, \qquad \beta = \frac{\widehat{a^2}}{2} + \beta_0, \tag{2.67}$$
$$\sigma_x = (\widehat{a^2}r_e^{-1} + \widehat{\upsilon_x})^{-1}, \qquad \mu_x = \sigma_x d\widehat{a}r_e^{-1}, \tag{2.68}$$
$$\gamma = \frac{1}{2} + \gamma_0, \qquad \delta = \frac{\widehat{x^2}}{2} + \delta_0 \tag{2.69}$$

The required VB-moments are

$$\widehat{a} = \mu_a, \qquad \widehat{a^2} = \mu_a^2 + \sigma_a, \tag{2.70}$$
$$\widehat{x} = \mu_x, \qquad \widehat{x^2} = \mu_x^2 + \sigma_x, \tag{2.71}$$
$$\widehat{\upsilon_a} = \frac{\alpha}{\beta}, \qquad \widehat{\upsilon_x} = \frac{\gamma}{\delta}. \tag{2.72}$$

2.4.2.1. Positive Solution Enforcement

Consider again the scalar example (2.27) with the ARD prior; however, now with positivity restriction using truncated normal distribution

$$f(a|r_a) = t\mathcal{N}_a(0, \upsilon_a^{-1}, [0, \infty]), \tag{2.73}$$

$$f(x|r_x) = t\mathcal{N}_x(0, \upsilon_x^{-1}, [0, \infty]). \tag{2.74}$$

Then, the VB-marginals reflect the form of prior and are recognized as

$$\tilde{f}(a|d) = t\mathcal{N}_a(\mu_a, \sigma_a, [0, \infty]), \tag{2.75}$$

$$\tilde{f}(x|d) = t\mathcal{N}_x(\mu_x, \sigma_x, [0, \infty]), \tag{2.76}$$

where shaping parameters are computed using the same equations as in the previous case, (2.66)–(2.68) with different computation of the VB-moments which are non-trivial and can be found in Appendix A.1.3. Using notation from Appendix A.1.4, the moments are computed as

$$\widehat{a} = \mathrm{M}_1^{\mathrm{tN}}(\mu_a, \sigma_a, 0, \infty), \tag{2.77}$$

$$\widehat{a^2} = \mathrm{M}_2^{\mathrm{tN}}(\widehat{a}, \mu_a, \sigma_a, 0, \infty), \tag{2.78}$$

$$\widehat{x} = \mathrm{M}_1^{\mathrm{tN}}(\mu_x, \sigma_x, 0, \infty), \tag{2.79}$$

$$\widehat{x^2} = \mathrm{M}_2^{\mathrm{tN}}(\widehat{x}, \mu_x, \sigma_x, 0, \infty). \tag{2.80}$$

2.4.3. Influence of Prior Selection on Scalar Decomposition

Performance of the introduced methods for scalar decomposition is studied here. The results for fixed prior, Section 2.3.3, and for ARD prior, Section 2.4.2, are given using both, unrestricted support and support truncated to positive values.

The results are given in Figure 2.6:

1. No positivity is enforced during VB and ARD scenarios. The results are given in Figure 2.6 using the black line (basic VB scenario) and the blue line (ARD scenario). The results well correspond with the inference bound in VB scenario [97],

   $$d > \sqrt{r_e}, \tag{2.81}$$

   and the inference bound in ARD scenario [101],

   $$d > 2\sqrt{r_e}. \tag{2.82}$$

   Note that the ARD property enforces sparse estimates more aggressively then basic solution. This may be a consequence of the variance underestimation of the VB approximation [71].

2. The positivity is enforced during IVB algorithms using truncation described in Appendix A.1.3. Figure 2.6 shows that in case of low signal, the basic VB solution (magenta line) has tendency to overestimation which is suppressed in the case of ARD priors (green line).

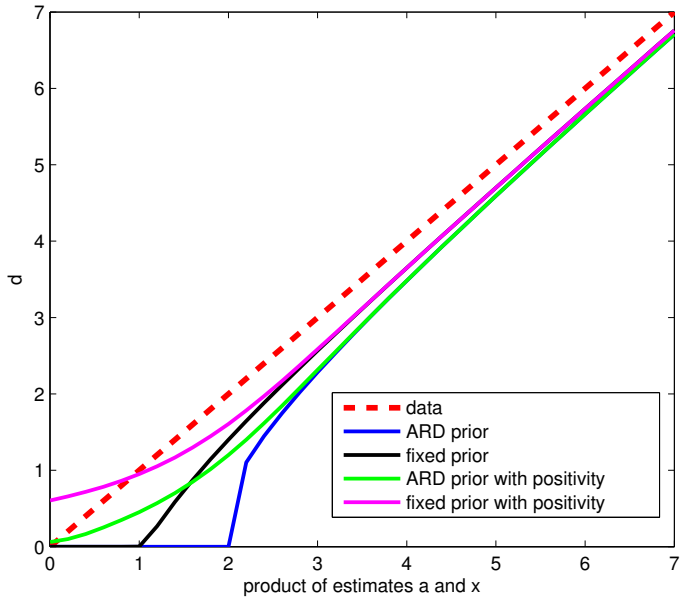

Figure 2.6.: Product of estimates $\widehat{a}$ and $\widehat{x}$ in increasing data term d and constant noise term $r_e = 1$ for scalar examples from Sections 2.3.3 and 2.4.2. The settings for VB scalar example without ARD priors (black and magenta lines) are $r_a = r_x = 10$.

2.5. Extension to Matrix Decomposition

Here, we expand the scalar example from Section 2.3.3 into the matrix form to explain the matrix decomposition which is used in the remainder of this work. Consider observation model (1.1), $D = AX^T + E$, where $D \in \mathbf{R}^{p\times n}$, $A \in \mathbf{R}^{p\times r}$, $X \in \mathbf{R}^{n\times r}$, and $E \in \mathbf{R}^{p\times n}$. For the isotropic Gaussian noise model [115] $e_{i,j} \sim N(0, \omega^{-1})$, the prior data model is

$$f\left(d_{i,j}|A, X, \omega\right) = \mathcal{N}_{d_{i,j}}\left(\sum_{k=1}^{r} a_{i,k}x_{j,k}, \omega^{-1}\right) \tag{2.83}$$

which may be reduced to scalar model (2.28) for $p = n = r = 1$.

Let us focus on modeling of the matrix A while the same approach can be used for the matrix X. The term (2.83) contains combinations of elements of the matrix A; hence, the standard form of the VB-marginals fits to the form of the multivariate normal distribution, see Appendix A.1.1. Here, we need to define an ordering of the elements of the matrix $A = [\mathbf{a}_1, \ldots, \mathbf{a}_r]$ which we do

using vectorization:

$$\mathrm{vec}(A) = \begin{pmatrix} \mathbf{a}_1 \\ \mathbf{a}_2 \\ \vdots \\ \mathbf{a}_p \end{pmatrix} \equiv \mathbf{a} \in \mathbf{R}^{pr \times 1}. \tag{2.84}$$

Then, the prior distribution for the matrix A can be written as the multivariate normal distribution

$$f(A) = f(\mathbf{a}) = \mathcal{N}_{\mathbf{a}}\left(\boldsymbol{\mu}_{\mathbf{a}}, \Sigma_{\mathbf{a}}\right) = \frac{1}{(2\pi)^{\frac{pr}{2}} |\Sigma_{\mathbf{a}}|^{\frac{1}{2}}} \times \\ \times \exp\left(-\frac{1}{2}\left(\mathbf{a} - \boldsymbol{\mu}_{\mathbf{a}}\right)^T \Sigma_{\mathbf{a}}^{-1}\left(\mathbf{a} - \boldsymbol{\mu}_{\mathbf{a}}\right)\right), \tag{2.85}$$

where $\Sigma_{\mathbf{a}} \in \mathbf{R}^{pr \times pr}$ is the covariance matrix.

Since the covariance matrix is extremely large, we can consider some further assumptions. For example, if the pixels of each source image $\mathbf{a}_k, k = 1, \ldots, r$, would be independently identically distributed (i.i.d.), the covariance matrix could be rewritten as

$$\Sigma_{\mathbf{a}} = \Sigma_A \otimes I_p, \tag{2.86}$$

where I_p is the identity matrix of given size, $\Sigma_A \in \mathbf{R}^{r \times r}$ is diagonal matrix, and symbol $\otimes$ denotes Kronecker product defined as

$$A \otimes B = \begin{pmatrix} a_{11}B & \cdots & a_{1r}B \\ \vdots & \ddots & \vdots \\ a_{p1}B & \cdots & a_{pr}B \end{pmatrix} \tag{2.87}$$

for arbitrary matrices A and B. Generalization of (2.86) is model

$$\Sigma_{\mathbf{a}} = \Sigma_A \otimes \Phi_A, \tag{2.88}$$

where $\Phi_A \in \mathbf{R}^{p \times p}$ is symmetric positive definite matrix; hence, two matrices of the size $r \times r$ and $p \times p$ is needed instead of one matrix of the size $pr \times pr$. Using Kronecker structure of the covariance matrix $\Sigma_{\mathbf{a}}$, we can rewrite the prior distribution for the matrix A using matrix normal distribution, see Appendix A.1.2, as

$$\mathcal{N}_A(\mu_A, \Phi_A \otimes \Sigma_A) = \frac{1}{(2\pi)^{\frac{pr}{2}} |\Phi_A|^{\frac{r}{2}} |\Sigma_A|^{\frac{p}{2}}} \times \\ \times \exp\left(-\frac{1}{2}\mathrm{tr}\left[\Phi_A^{-1}(A - \mu_A)(\Sigma_A^{-1})^T(A - \mu_A)^T\right]\right), \tag{2.89}$$

where tr(.) denotes trace of a square matrix defined as

$$\mathrm{tr}(B) = \sum_{j=1}^{n} b_{jj}. \tag{2.90}$$

Note that the order of the covariance matrices in multivariate normal distribution of the vectorized form of the matrix A is $\Sigma_A \otimes \Phi_A$ while the order of these matrices in the matrix normal distribution is $\Phi_A \otimes \Sigma_A$ in this notation.

The following equalities hold for compatible matrices A, B, and C and identity matrix I and will be useful in the following text:

$$\begin{aligned} \mathrm{tr}(A) =& \mathrm{tr}\left(A^T\right), & \mathrm{tr}\left(A^T B\right) =& \mathrm{vec}(A)^T \mathrm{vec}(B), && (2.91)\\ \mathrm{tr}(AB) =& \mathrm{tr}(BA), & \mathrm{vec}(ABC) =& \left(C^T \otimes A\right) \mathrm{vec}(B), && (2.92)\\ (A \otimes B)^T =& A^T \otimes B^T, & \mathrm{vec}(ABC) =& \left(A \otimes C^T\right) \mathrm{vec}\left(B^T\right), && (2.93)\\ (A \otimes B)^{-1} =& A^{-1} \otimes B^{-1}, & \mathrm{vec}(ABC) =& (I \otimes AB)\, \mathrm{vec}(C). && (2.94) \end{aligned}$$

2.5.1. Matrix Decomposition

The observation model (2.83) can be rewritten for the whole data matrix D as

$$f(D|A, X, \omega) = \mathcal{N}_D\left(AX^T, \omega^{-1} I_p \otimes I_n\right), \tag{2.95}$$

where I_p is the identity matrix of the given size and ω^{-1} denotes an unknown common variance of the noise. The observation model must be accompanied with the prior model in the VB method. The parameters A, X, and ω are modeled in the same way as in [97]:

$$\begin{aligned} f(A) =& \mathcal{N}_A\left(\mathbf{0}_{p,r}, I_p \otimes I_r\right), && (2.96)\\ f(X) =& \mathcal{N}_X\left(\mathbf{0}_{n,r}, I_n \otimes I_r\right), && (2.97)\\ f(\omega) =& \mathcal{G}_\omega(\vartheta_0, \rho_0). && (2.98) \end{aligned}$$

Here, the prior parameters ϑ_0 and ρ_0 can be chosen to yield non-informative prior.

The logarithm of the the joint distribution is then

$$\begin{aligned} \ln f(D, A, X, \omega) = \frac{pn}{2} \ln \omega - \frac{1}{2}\mathrm{tr}\left((D - AX^T)(D - AX^T)^T\right) + \\ - \frac{1}{2}\mathrm{tr}(A^T A) - \frac{1}{2}\mathrm{tr}(XX^T) + (\vartheta_0 - 1)\ln\omega + \rho_0\omega + \gamma, \end{aligned} \tag{2.99}$$

where γ stands for all terms that are independent of model parameters.

From this step, we will continue only with inference for the parameter A for simplicity and clarity while inference for remaining parameters is analogical.

Application of the VB theorem (2.18) on the logarithm of the joint distribution results in the following VB-marginal for parameter A:

$$\tilde{f}(A|D) \propto \exp\left[-\frac{1}{2}\widehat{\omega}\mathrm{tr}(-2A\widehat{X}^T D^T) - \frac{1}{2}\widehat{\omega}\mathrm{tr}(A\widehat{(X^T X)}A^T)) - \frac{1}{2}\mathrm{tr}(AA^T)\right], \tag{2.100}$$

The VB-marginal is recognized to has following matrix normal distribution standard form:

$$\tilde{f}(A|D) = \mathcal{N}_A\left(\mu_A, \Phi_A \otimes \Sigma_A\right), \tag{2.101}$$

The shaping parameters of the standard forms are

$$\Phi_A \otimes \Sigma_A = \left(I_p \otimes \widehat{\omega X^T X} + I_p \otimes I_r\right)^{-1}, \tag{2.102}$$

$$\mu_A = \Sigma_A \left(\widehat{\omega} D \widehat{X}\right) \Phi_A, \tag{2.103}$$

Here, we can use the advantage of the Kronecker product and expand the right side of the equation (2.102) as

$$\left(I_p \otimes \widehat{\omega X^T X} + I_p \otimes I_r\right)^{-1} = \left(I_p \otimes \left(\widehat{\omega X^T X} + I_r\right)\right)^{-1} = I_p \otimes \left(\widehat{\omega X^T X} + I_r\right)^{-1}, \tag{2.104}$$

where matrices Φ_A and Σ_A can be easily identify as

$$\Phi_A = I_p, \tag{2.105}$$

$$\Sigma_A = \left(\widehat{\omega X^T X} + I_r\right)^{-1}. \tag{2.106}$$

Here can be seen the advantage of the Kronecker form of the VB-marginal where only inversion of the $r \times r$ size matrix need to be inverted instead of $pr \times pr$ size matrix. The shaping parameters are accompanied with the necessary VB-moments according to the Appendix A.1.2:

$$\widehat{A} = \mu_A, \tag{2.107}$$

$$\widehat{A^T A} = \mu_A^T \mu_A + \mathrm{tr}\left(\Sigma_A\right) \Phi_A. \tag{2.108}$$

2.5.1.1. Truncation to Positive Domain

Consider replacement of the prior (2.96) by the truncated prior with positive support

$$f(A) = t\mathcal{N}_A\left(\mathbf{0}_{p,r}, I_p \otimes I_r, [0, \infty]\right). \tag{2.109}$$

Then, the posterior distribution has also the form of the truncated normal distribution as

$$\tilde{f}(A|D) = t\mathcal{N}_A\left(\mu_A, \Phi_A \otimes \Sigma_A, [0, \infty]\right), \tag{2.110}$$

and moments $\widehat{A}$ and $\widehat{A^T A}$ needs to be evaluated; however, they are not available in closed-form. Hence, in this case, we used approximation using independence [97] as

$$\tilde{f}(A|D) = \tilde{f}(\mathbf{a}|D) \approx t\mathcal{N}_{\mathbf{a}}\left(\boldsymbol{\mu}_{\mathbf{a}}, \mathrm{diag}\left(\boldsymbol{\sigma}_{\mathbf{a}}\right), [0, \infty]\right), \tag{2.111}$$

where $\boldsymbol{\sigma}_{\mathbf{a}} = \operatorname{diag}(\Phi_A \otimes \Sigma_A)^{-1}$. Note that this approximation remove all correlation between pixels in the matrix A, thus, (2.111) can be rewritten as

$$\tilde{f}(\mathbf{a}|D) \approx \prod_{l}^{pr} t\mathcal{N}_{a_l}\left(\mu_{\mathbf{a},l}, \sigma_{\mathbf{a},l}, [0,\infty]\right), \tag{2.112}$$

with moments evaluated according to the scalar truncated normal distribution given in Appendix A.1.3 reaching $\widehat{A}$ and $\widehat{A^T A}$ after rearranging. Using notation from Appendix A.1.4, the moments are computed as

$$\widehat{A} = \mathrm{M}_1^{\mathrm{tN}}\left(\mu_A, \Phi_A \otimes \Sigma_A, 0, \infty\right), \tag{2.113}$$

$$\widehat{A^T A} = \mathrm{M}_2^{\mathrm{tN}}\left(\widehat{A}, \mu_A, \Phi_A \otimes \Sigma_A, 0, \infty\right). \tag{2.114}$$

3. Prior Models in Superposition Problem

Newton spoke of God in his book. I have perused yours but failed to find His name ever once. Why?

(Napoleon)

Sir, I have no need of that hypothesis.

(Pierre Simon Laplace)

In this chapter, we construct a prior model of each parameter in the BSS problem, i.e. incorporate additional informations using hierarchical priors. Recall the BSS problem (1.1), $D = AX^T + E$, where D stands for measured data, A stands for source images, X stands for source TACs, and E stands for error term, mostly noise.

We will use the ability of the VB method to combine various prior models of model parameters in an arbitrary order which is schematically demonstrated in Figure 3.1. Here, the data is modeled hierarchically using the image, weight, and noise priors, i.e. $f(A)$, $f(X)$, and $f(E)$.

Bayesian estimation of all unknown parameters of the BSS model (1.1) requires evaluation of joint posterior densities. However, this is analytically intractable and an approximate evaluation is required. We use the Variational Bayes (VB) approach [76, 97], see Section 2.3, which seeks the best posterior in the form of conditionally independent factors; in this case:

$$f(A, X, E|D) \approx f(A|D)f(X|D)f(E|D). \tag{3.1}$$

The best approximation of this form in the sense of Kullback-Leibler divergence

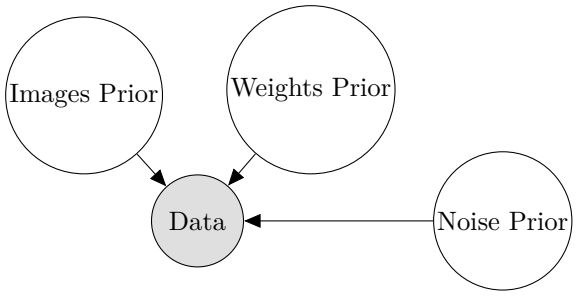

Figure 3.1.: Modular model of blind source separation.

can be found analytically the Variational Bayes (VB) method [97].

The VB methodology allows us to do inference for each parameter separately and then arbitrary combine them as discussed and demonstrated in Section 2.3.2. Hence, we study a prior model for each model parameter A, X, E separately and then combine them in various ways in order to obtain specific separation method.

3.1. Prior of Noise

The observation errors in nuclear medical imaging is Poisson distributed; however, the inference with Poisson noise is analytically nor computationally intractable [97]. Therefore, Poisson noise is substituted by Gaussian noise in Probabilistic Principal Component Analysis (PPCA) [115, 98] which is also the base of our model.

3.1.1. Isotropic Prior Model of Noise

In the model (1.1), the elements of the matrix E are independently identically distributed with unknown common precision ω as

$$f(E|\omega) = \prod_{i=1}^{p}\prod_{j=1}^{n} \mathcal{N}_{e_{i,j}}(0, \omega^{-1}), \tag{3.2}$$

where ω plays a role of precision of noise and is assume to be unknown. In matrix formulation, the matrix E has matrix normal distribution, see Appendix A.1.2,

$$f(E|\omega) = \mathcal{N}_E(\mathbf{0}_{p,n}, \omega^{-1} I_p \otimes I_n), \tag{3.3}$$

known as the isotropic Gaussian noise model [115]. The equations (1.1) and (3.3) can be rewritten together as

$$f(D|A, X, \omega) = \mathcal{N}_D(AX^T, \omega^{-1} I_p \otimes I_n), \tag{3.4}$$

or equally for one time-point t as

$$f(\mathbf{d}_t|A, \overline{\mathbf{x}}_t, \omega) = \mathcal{N}_{\mathbf{d}_t}\left(\sum_{k=1}^{r} \mathbf{a}_k x_{t,k}, \omega^{-1} I_p\right), \tag{3.5}$$

where the bar symbol denotes a row vector, $\overline{\mathbf{x}}_t = [x_{t,1}, \ldots, x_{t,r}]$.

The hierarchical model of the vector version (3.5) is given in Figure 3.2. The task of the further research is to develop a prior models of the noise, the source images, $\mathbf{a}_k$, and prior models of the time-activity curves, $\mathbf{x}_k$.

The prior data model has to be accompanied with prior model of precision parameter ω. The precision parameter ω of the normal density (3.4) or (3.5) has a conjugate prior in the form of Gamma distribution

$$f(\omega) = \mathcal{G}_\omega(\vartheta_0, \rho_0), \tag{3.6}$$

where coefficients ϑ_0, ρ_0 are chosen prior constants.

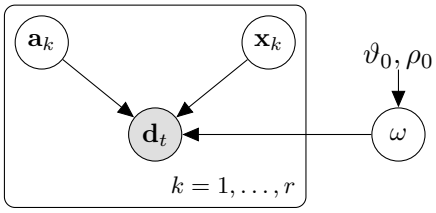

Figure 3.2.: Hierarchical model of superposition.

VB-posterior

Prior noise model (3.6) yields posterior standard distributional form

$$\tilde{f}(\omega|D, r) = \mathcal{G}_\omega(\vartheta, \rho), \tag{3.7}$$

with shaping parameters

$$\vartheta = \vartheta_0 + \frac{np}{2}, \tag{3.8}$$

$$\rho = \rho_0 + \frac{1}{2}\mathrm{tr}\left(DD^T - \widehat{A}\widehat{X}^T D^T - D\widehat{X}\widehat{A}^T\right) + \frac{1}{2}\mathrm{tr}\left(\widehat{A^T A}\widehat{X^T X}\right). \tag{3.9}$$

The associated VB-moment is

$$\widehat{\omega} = \frac{\vartheta}{\rho}. \tag{3.10}$$

Remark 4. The isotropic Gaussian noise model (3.3) can be seen as too simplistic since the whole noise is estimated only using one parameter $\omega \in \mathbf{R}$. Alternatively, the model (3.4) can replaced by more flexible prior model [97] as

$$f\left(D|A, X, \Omega_p, \Omega_n\right) = \mathcal{N}_D\left(AX^T, \Omega_p^{-1} \otimes \Omega_n^{-1}\right), \tag{3.11}$$

$$\Omega_p = \mathrm{diag}\left(\boldsymbol{\omega}_p\right), \tag{3.12}$$

$$\Omega_n = \mathrm{diag}\left(\boldsymbol{\omega}_n\right), \tag{3.13}$$

where vectors $\boldsymbol{\omega}_p \in \mathbf{R}^{p\times 1}$ and $\boldsymbol{\omega}_n \in \mathbf{R}^{n\times 1}$ is additional parameters of the model with their own prior models. Estimates of $\widehat{\Omega_p}$ and $\widehat{\Omega_n}$ can be obtained using VB approximation; however, we conclude that the effect of this more complicated prior model of noise is very low [97].

3.2. Priors of Source Images

The key task is to regularize the separation problem (1.1) using prior models while prior models for source images, the matrix $A = [\mathbf{a}_1, \ldots, \mathbf{a}_r]$, are studied in this Section. After isotropic prior [14, 97], we study incorporation of assumption that the source images are most likely sparse. We propose two different prior models: mixture model and ARD model.

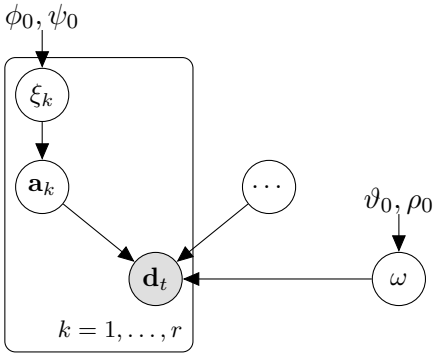

Figure 3.3.: Hierarchical isotropic prior model of source images.

3.2.1. Isotropic Prior

The prior model of $A = [\mathbf{a}_1, \ldots, \mathbf{a}_r]$ is given as follows [14, 97]:

$$f(\mathbf{a}_k|\xi_k) = t\mathcal{N}_{\mathbf{a}_k}(\mathbf{0}_{p,1}, \xi_k^{-1} I_p, [0, \infty]), \tag{3.14}$$

$$f(\xi_k) = \mathcal{G}_{\xi_k}(\phi_0, \psi_0), \tag{3.15}$$

$\forall k = 1, \ldots, r$, where ϕ_0, ψ_0 are chosen prior constants. The parameter $\xi_k \in \mathbf{R}$ is the ARD hyper-parameter, see Section 2.4, measuring significance of the kth source image. This hierarchical prior is shown in Figure 3.3

The VB-marginals are recognized to have standard distributional forms

$$\tilde{f}(A|D, r) = t\mathcal{N}_A(\mu_A, I_p \otimes \Sigma_A, [0, \infty]), \tag{3.16}$$

$$\tilde{f}(\xi_k|D, r) = \mathcal{G}_{\xi_k}(\phi_k, \psi_k), \tag{3.17}$$

with shaping parameters

$$\Sigma_A = \left(\widehat{\omega X^T X} + \mathrm{diag}(\widehat{\boldsymbol{\xi}})\right)^{-1}, \tag{3.18}$$

$$\mu_A = \widehat{\omega} D \widehat{X} \Sigma_A, \tag{3.19}$$

$$\boldsymbol{\phi} = \phi_0 + \frac{p}{2}\mathbf{1}_{r,1}, \tag{3.20}$$

$$\boldsymbol{\psi} = \psi_0 + \frac{1}{2}\mathrm{diag}\left(\widehat{A^T A}\right), \tag{3.21}$$

where $\boldsymbol{\xi} = [\xi_1, \ldots, \xi_r]$. The associated VB-moments are

$$\widehat{A} = \mathrm{M}_1^{\mathrm{tN}}\left(\mu_A, I_p \otimes \Sigma_A, 0, \infty\right), \tag{3.22}$$

$$\widehat{A^T A} = \mathrm{M}_2^{\mathrm{tN}}\left(\widehat{A}, \mu_A, I_p \otimes \Sigma_A, 0, \infty\right), \tag{3.23}$$

$$\widehat{\boldsymbol{\xi}} = \mathrm{diag}\left(\boldsymbol{\phi} \circ \boldsymbol{\psi}^{-1}\right), \tag{3.24}$$

where functions $\mathrm{M}_1^{\mathrm{tN}}()$ and $\mathrm{M}_2^{\mathrm{tN}}()$ are defined in Appendix A.1.4.

The prior distribution (3.14), $f(\mathbf{a}_k)$, is chosen with prior parameter ξ_k on diagonal of its covariance matrix, hence we call this prior isotropic.

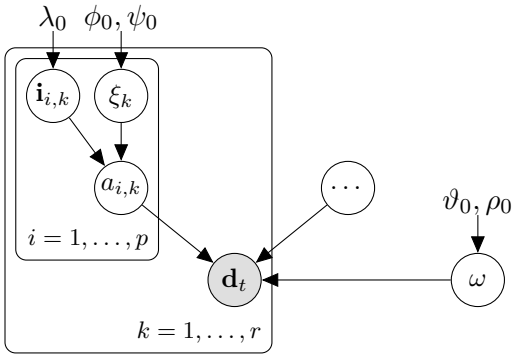

Figure 3.4.: Hierarchical sparse prior model of source images using mixture model.

3.2.2. Sparsity Using Mixture Prior

We have proposed the enforcement of sparsity into the superposition model using indicator variables related to each element (pixel) of the matrix A [100]. Suppose that each element of the matrix A, $a_{i,k}$, is a mixture of uniform distribution $\mathcal{U}_{a_{i,k}}(0,1)$ and $t\mathcal{N}_{a_{i,k}}(0,\xi_k^{-1})$, where the uniform distribution is the model of significant signal and normal distribution with unknown variance is the noise model, which is switched using the indicator $\mathbf{i}_{i,k} \in \{0;1\}$ as

$$f(a_{i,k}|\xi_k, \mathbf{i}_{i,k}) = \begin{cases} \mathcal{U}_{a_{i,k}}(0,1) & \mathbf{i}_{i,k} = 1, \\ t\mathcal{N}_{a_{i,k}}(0, \xi_k^{-1}, [0,\infty]) & \mathbf{i}_{i,k} = 0. \end{cases} \tag{3.25}$$

However, the inference for this model with discrete variable $\mathbf{i}$ would be computationally very costly; therefore, a "soft version" of the model (3.25) is adopted using continuous variables $\mathbf{i}_{i,k} \in [0,1]$. The model is modified as

$$f(a_{i,k}|\xi_k, \mathbf{i}_{i,k}) = \mathcal{U}_{a_{i,k}}(0,1)^{\mathbf{i}_{i,k}} t\mathcal{N}_{a_{i,k}}(0, \xi_k^{-1}, [0,\infty])^{(1-\mathbf{i}_{i,k})} \tag{3.26}$$

$\forall i = 1, \ldots, p$ and $\forall k = 1, \ldots, r$, with extremes given in (3.25). Following the VB methodology, priors for $\mathbf{i}_{i,k}$ and ξ_k have to be selected. We proposed the following in [100]:

$$f(\mathbf{i}_{i,k}) = \text{tExp}_{\mathbf{i}_{i,k}}(\lambda_0, (0,1]), \tag{3.27}$$

$$f(\xi_k) = \mathcal{G}_{\xi_k}(\phi_0, \psi_0), \tag{3.28}$$

where tExp() is truncated exponential distribution with given support, see Appendix A.3. This hierarchical prior is shown in Figure 3.4.

Natural partitioning for the rows of source images, rows $\overline{\mathbf{a}}_i$ of the matrix A,

can be found to match the Gaussian distribution as follows:

$$\operatorname{tr}\left[\left(D-AX^T\right)\left(D-AX^T\right)^T\right]= \\ =\sum_{i=1}^{p}\left[2\overline{\mathbf{a}}_i\sum_{j=1}^{n}(\overline{\mathbf{x}}_j d_{i,j})^T-\overline{\mathbf{a}}_i\sum_{j=1}^{n}(\overline{\mathbf{x}}_j^T\overline{\mathbf{x}}_j)\overline{\mathbf{a}}_i^T\right]+c_A, \quad (3.29)$$

where standard form can be easily seen as

$$\tilde{f}(\overline{\mathbf{a}}_i|D,r)=t\mathcal{N}_{\overline{\mathbf{a}}_i}\left(\boldsymbol{\mu}_{\overline{\mathbf{a}}_i},\Sigma_{\overline{\mathbf{a}}_i},[0,\infty]\right). \quad (3.30)$$

Posterior distributions of other model parameters are recognized from the VB-marginals to have standard distributional forms

$$\tilde{f}(\xi_k|D,r)=\mathcal{G}_{\xi_k}(\phi_k,\psi_k), \quad (3.31)$$
$$\tilde{f}(\mathbf{i}_{i,k}|D,r)=\mathrm{tExp}_{\mathbf{i}_{i,k}}(\lambda_{i,k},(0,1]), \quad (3.32)$$

where the bar symbol denotes row vector, $\overline{\mathbf{a}}_i=[a_{i,1},\ldots,a_{i,r}]$. The shaping parameters are

$$\Sigma_{\overline{\mathbf{a}}_i}=\left(\widehat{\omega}\sum_{k=1}^{n}\widehat{\mathbf{x}_k^T\mathbf{x}_k}+\widehat{\boldsymbol{\xi}}(I_r-\operatorname{diag}(\widehat{\overline{\mathbf{i}}_i}))\right)^{-1}, \quad (3.33)$$

$$\boldsymbol{\mu}_{\overline{\mathbf{a}}_i}=\Sigma_{\overline{\mathbf{a}}_i}\widehat{\omega}\sum_{k=1}^{n}\widehat{\mathbf{x}_k}d_{i,k}, \quad (3.34)$$

$$\phi_k=\phi_0+\frac{1}{2}\sum_{i=1}^{p}\left(1-\widehat{\mathbf{i}_{i,k}}\right), \quad (3.35)$$

$$\psi_k=\psi_0+\frac{1}{2}\sum_{i=1}^{p}\left(1-\widehat{\mathbf{i}_{i,k}}\right)\widehat{a_{i,k}^2}, \quad (3.36)$$

$$\lambda_{i,k}=\lambda_0-\frac{1}{2}\widehat{\ln\xi_k}+\frac{1}{2}\widehat{a_{i,k}\xi_k a_{i,k}}, \quad (3.37)$$

where $\boldsymbol{\xi}=[\xi_1,\ldots,\xi_r]$. The associated VB-moments are

$$\widehat{\overline{\mathbf{a}}_i}=\mathrm{M}_1^{\mathrm{tN}}\left(\boldsymbol{\mu}_{\overline{\mathbf{a}}_i},\Sigma_{\overline{\mathbf{a}}_i},0,\infty\right), \quad (3.38)$$
$$\widehat{\overline{\mathbf{a}}_i^T\overline{\mathbf{a}}_i}=\mathrm{M}_2^{\mathrm{tN}}\left(\widehat{\overline{\mathbf{a}}}_i,\boldsymbol{\mu}_{\overline{\mathbf{a}}_i},\Sigma_{\overline{\mathbf{a}}_i},0,\infty\right), \quad (3.39)$$
$$\widehat{\xi_k}=\frac{\phi_k}{\psi_k}, \quad (3.40)$$
$$\widehat{\mathbf{i}_{i,k}}=\lambda_{i,k}^{-1}, \quad (3.41)$$

where functions $\mathrm{M}_1^{\mathrm{tN}}()$ and $\mathrm{M}_2^{\mathrm{tN}}()$ are defined in Appendix A.1.4.

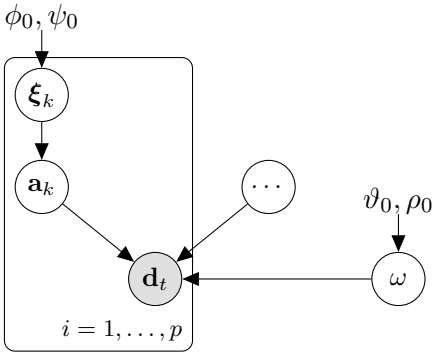

Figure 3.5.: Hierarchical sparse prior model of source images using automatic relevant determination.

Remark 5. We have studied the so called direct ROI model where the ROI is represented using the matrix $\mathcal{I}$ of the same size as the matrix A with source images stored columnwise. The prior model for $\mathcal{I}$ is $\mathcal{I} \in \{0,1\}^{p \times r}$. Then, the model (1.1) is modified using $\mathcal{I}$ as

$$D = (A \circ \mathcal{I})\, X' + E, \tag{3.42}$$

where $\circ$ denoted the Hadamard product. We have selected the prior for $\mathcal{I}$ as

$$f(\mathcal{I}) = \mathrm{tExp}_{\mathcal{I}}\left(\lambda_{pr,0}, (0,1]\right) \tag{3.43}$$

and derived the iterative VB algorithm for this model. In this derivation, the solution for the matrix A is partitioned to solution for each row of the matrix, $\overline{\mathbf{a}}_i$, as in Section 3.2.2.

We did not published this model since it suffers from computational instability and highly depends on starting point of iterations as well as on stopping condition.

3.2.3. Sparsity Using ARD Prior

We have used the sparsity model suggested in 2.4. The ARD property can be adopted not only on the whole sources, Section 3.2.1, but also on each pixel, $a_{i,k}$. Each element (pixel) of the matrix A, $a_{i,k}$, has its own ARD Gaussian prior modeled by its precision $\xi_{i,k}$ suppressing the value $a_{i,k}$ down to zero while being low.

Hence, each pixel has the prior

$$f(\mathbf{a}_k|\boldsymbol{\xi}_k) = t\mathcal{N}_{\mathbf{a}_k}\left(\mathbf{0}_{p,1}, \mathrm{diag}\left(\boldsymbol{\xi}_k\right)^{-1}, [0,1]\right), \tag{3.44}$$

$\forall k = 1, \ldots, r$, while the prior for precision parameter $\boldsymbol{\xi}_k$ is chosen as Gamma distribution such as

$$f(\boldsymbol{\xi}_k) = \prod_{i=1}^{p} \mathcal{G}_{\xi_{i,k}}(\phi_0, \psi_0) \tag{3.45}$$

with selected prior constants ϕ_0, ψ_0. Note that the prior distributions of elements of the matrix A, equation (3.44), are restricted to interval $[0, 1]$. This restriction is not necessary but helps to suppress scaling ambiguity of the final algorithm. The parameter $\xi_{i,k}$ is estimated together with the parameter of interest, $a_{i,k}$. If the $a_{i,k}$ is not significant then the precision $\xi_{i,k}$ will be large, forcing the $a_{i,k}$ closer to zero. This hierarchical prior is shown in Figure 3.5.

Posterior distributions are recognized from the VB-marginals to have standard distributional forms

$$\tilde{f}(\overline{\mathbf{a}}_i|D) = t\mathcal{N}_{\overline{\mathbf{a}}_i}(\boldsymbol{\mu}_{\overline{\mathbf{a}}_i}, \Sigma_{\overline{\mathbf{a}}_i}, [0, 1]), \tag{3.46}$$

$$\tilde{f}(\overline{\boldsymbol{\xi}}_i|D) = \prod_{k=1}^{r} \mathcal{G}_{\xi_{i,k}}(\phi_{i,k}, \psi_{i,k}), \tag{3.47}$$

$\forall i = 1, \ldots, p$, with shaping parameters

$$\Sigma_{\overline{\mathbf{a}}_i} = \left(\widehat{\omega} \sum_{j=1}^{n} (\widehat{\overline{\mathbf{x}}_j^T \overline{\mathbf{x}}_j}) + \mathrm{diag}(\widehat{\overline{\boldsymbol{\xi}}_i}) \right)^{-1}, \tag{3.48}$$

$$\boldsymbol{\mu}_{\overline{\mathbf{a}}_i} = \Sigma_{\overline{\mathbf{a}}_i} \widehat{\omega} \sum_{j=1}^{n} (\widehat{\overline{\mathbf{x}}_j} d_{i,j}), \tag{3.49}$$

$$\boldsymbol{\phi}_i = \phi_{i,0} + \frac{1}{2} \mathbf{1}_{r,1}, \tag{3.50}$$

$$\boldsymbol{\psi}_i = \psi_{i,0} + \frac{1}{2} \mathrm{diag}\left(\widehat{\overline{\mathbf{a}}_i^T \overline{\mathbf{a}}_i} \right), \tag{3.51}$$

The associated VB-moments are

$$\widehat{\overline{\mathbf{a}}_i} = \mathrm{M}_1^{\mathrm{tN}}\left(\boldsymbol{\mu}_{\overline{\mathbf{a}}_i}, \Sigma_{\overline{\mathbf{a}}_i}, 0, 1 \right), \tag{3.52}$$

$$\widehat{\overline{\mathbf{a}}_i^T \overline{\mathbf{a}}_i} = \mathrm{M}_2^{\mathrm{tN}}\left(\widehat{\overline{\mathbf{a}}}_i, \boldsymbol{\mu}_{\overline{\mathbf{a}}_i}, \Sigma_{\overline{\mathbf{a}}_i}, 0, 1 \right), \tag{3.53}$$

$$\widehat{\overline{\boldsymbol{\xi}}}_i = \boldsymbol{\phi}_i \circ \boldsymbol{\psi}_i^{-1}, \tag{3.54}$$

where functions $\mathrm{M}_1^{\mathrm{tN}}()$ and $\mathrm{M}_2^{\mathrm{tN}}()$ are defined in Appendix A.1.4.

3.3. Priors of Time-activity Curves

The same problem of regularization as for the source images arises in case of the source activities, the matrix X. Hence, similar isotropic prior, Section 3.2.1, and ARD prior, Section 3.2.3, as for the matrix A will be proposed here for the matrix X. In addition, we study the convolution prior models of TACs motivated by techniques used in clinical practice from Section 1.2.2.1.

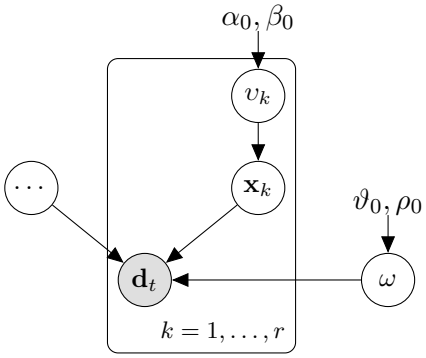

Figure 3.6.: Hierarchical isotropic prior model of source TACs.

3.3.1. Isotropic Prior

The isotropic prior model of $X = [\mathbf{x}_1, \ldots, \mathbf{x}_r]$ is constructed in the same way as in Section 3.2.1:

$$f(\mathbf{x}_k|\upsilon_k) = t\mathcal{N}_{\mathbf{x}_k}\left(\mathbf{0}_{n,1}, \upsilon_k^{-1} I_n, [0,\infty]\right), \tag{3.55}$$

$$f(\upsilon_k) = \mathcal{G}_{\upsilon_k}(\alpha_0, \beta_0), \tag{3.56}$$

$\forall k = 1, \ldots, r$, where α_0, β_0 are chosen prior constants. Again, the parameter $\upsilon_k \in \mathbf{R}$ is the ARD hyper-parameter measuring significance of each TAC $\mathbf{x}_k$. This hierarchical prior is shown in Figure 3.6.

Posterior distributions are recognized from the VB-marginals to have standard distributional forms

$$\tilde{f}(X|D, r) = t\mathcal{N}_X(\mu_X, I_n \otimes \Sigma_X, [0,\infty]), \tag{3.57}$$

$$\tilde{f}(\upsilon_k|D, r) = \mathcal{G}_{\upsilon_k}(\alpha_k, \beta_k), \tag{3.58}$$

with shaping parameters

$$\Sigma_X = \left(\widehat{\omega}\widehat{A^T A} + \mathrm{diag}(\widehat{\boldsymbol{\upsilon}})\right)^{-1}, \tag{3.59}$$

$$\mu_X = \widehat{\omega} D^T \widehat{A} \Sigma_X, \tag{3.60}$$

$$\boldsymbol{\alpha} = \alpha_0 + \frac{n}{2}\mathbf{1}_{r,1}, \tag{3.61}$$

$$\boldsymbol{\beta} = \beta_0 + \frac{1}{2}\mathrm{diag}\left(\widehat{X^T X}\right), \tag{3.62}$$

where $\boldsymbol{\upsilon} = [\upsilon_1, \ldots, \upsilon_r]$. The associated VB-moments are

$$\widehat{X} = \mathrm{M}_1^{\mathrm{tN}}\left(\mu_X, I_n \otimes \Sigma_X, 0, \infty\right), \tag{3.63}$$

$$\widehat{X^T X} = \mathrm{M}_2^{\mathrm{tN}}\left(, \widehat{X}, \mu_X, I_n \otimes \Sigma_X, 0, \infty\right), \tag{3.64}$$

$$\widehat{\boldsymbol{\upsilon}} = \boldsymbol{\alpha} \circ \boldsymbol{\beta}^{-1}, \tag{3.65}$$

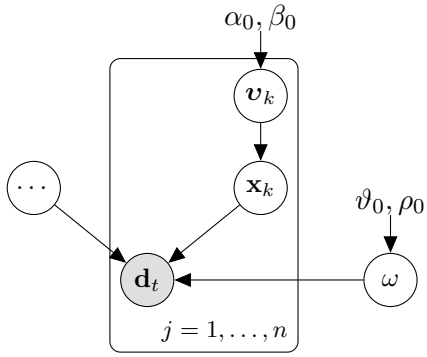

Figure 3.7.: Hierarchical sparse prior model of source TACs using automatic relevant determination.

where functions $\mathrm{M}_1^{\mathrm{tN}}()$ and $\mathrm{M}_2^{\mathrm{tN}}()$ are defined in Appendix A.1.4.

3.3.2. Sparse TACs Using ARD Prior

We propose very similar model for TACs as for source images in Section 3.2.3. Once again, we recall ARD methodology from Section 2.4 for each element of each TAC [109] as

$$f(\mathbf{x}_k|\boldsymbol{\upsilon}_k) = t\mathcal{N}_{\mathbf{x}_k}\left(\mathbf{0}_{n,1}, \mathrm{diag}(\boldsymbol{\upsilon}_k)^{-1}, [0,\infty]\right), \tag{3.66}$$

$$f(\boldsymbol{\upsilon}_k) = \prod_{j=1}^{n} \mathcal{G}_{\upsilon_{j,k}}(\alpha_0, \beta_0), \tag{3.67}$$

where $\upsilon_{j,k}$ is unknown variance parameter to be estimated together with weight $x_{j,k}$ while α_0, β_0 are known prior parameters. This hierarchical prior is shown in Figure 3.7.

Posterior distributions are recognized from the VB-marginals to have standard distributional forms

$$\tilde{f}(\overline{\mathbf{x}}_j|D) = t\mathcal{N}_{\overline{\mathbf{x}}_j}\left(\boldsymbol{\mu}_{\overline{\mathbf{x}}_j}, \Sigma_{\overline{\mathbf{x}}_j}, [0,\infty]\right), \tag{3.68}$$

$$\tilde{f}(\overline{\boldsymbol{\upsilon}}_j|D) = \prod_{k=1}^{r} \mathcal{G}_{\upsilon_{j,k}}(\alpha_{j,k}, \beta_{j,k}), \tag{3.69}$$

$\forall j = 1, \ldots, n$ and the shaping parameters computed accordingly to

$$\Sigma_{\overline{\mathbf{x}}_j} = \left(\widehat{\omega} \sum_{i=1}^{p} (\widehat{\overline{\mathbf{a}}_i^T \overline{\mathbf{a}}_i}) + \mathrm{diag}(\widehat{\overline{\boldsymbol{\upsilon}}_j})\right)^{-1}, \tag{3.70}$$

$$\boldsymbol{\mu}_{\overline{\mathbf{x}}_j} = \Sigma_{\overline{\mathbf{x}}_j} \widehat{\omega} \sum_{i=1}^{p} (\widehat{\overline{\mathbf{a}}}_i d_{i,j}), \tag{3.71}$$

$$\boldsymbol{\alpha}_j = \alpha_0 + \frac{1}{2}\mathbf{1}_{r,1}, \quad \boldsymbol{\beta}_j = \beta_0 + \frac{1}{2}\mathrm{diag}\left(\widehat{\overline{\mathbf{x}}_j^T \overline{\mathbf{x}}_j}\right). \tag{3.72}$$

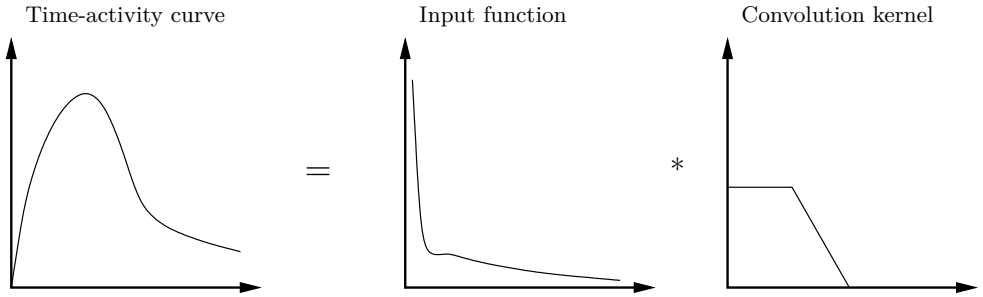

Figure 3.8.: Example deconvolution of TAC to an input function and convolution kernel.

The associated VB-moments are

$$\widehat{\overline{\mathbf{x}}_j} = \mathrm{M}_1^{\mathrm{tN}}\left(\boldsymbol{\mu}_{\overline{\mathbf{x}}_j}, \Sigma_{\overline{\mathbf{x}}_j}, 0, \infty\right), \tag{3.73}$$

$$\widehat{\overline{\mathbf{x}}_j^T \overline{\mathbf{x}}_j} = \mathrm{M}_2^{\mathrm{tN}}\left(\widehat{\overline{\mathbf{x}}_j}, \boldsymbol{\mu}_{\overline{\mathbf{x}}_j}, \Sigma_{\overline{\mathbf{x}}_j}, 0, \infty\right), \tag{3.74}$$

$$\widehat{\overline{\boldsymbol{\upsilon}}_j} = \boldsymbol{\alpha}_j \circ \boldsymbol{\beta}_j^{-1}, \tag{3.75}$$

where functions $\mathrm{M}_1^{\mathrm{tN}}()$ and $\mathrm{M}_2^{\mathrm{tN}}()$ are defined in Appendix A.1.4.

3.3.3. Convolution Priors

Assumption of convolution is common in compartment modeling [66, 6] where sources are modeled based on kinetics of radiopharmaceuticals inside them. Generally, compartment models assume that the object of interest can be modeled as its parts (i.e. compartments) and using kinetic relations between them. Accumulation of radiopharmaceuticals is commonly modeled as a convolution between common input function and source-specific convolution kernel [104, 6, 83].

Since the measurement in dynamic nuclear medical imaging is discrete in time, we assume the discrete version of convolution. We use the following notations: input function is stored in vector $\mathbf{b} \in \mathbf{R}^{n\times 1}$ and convolution kernels of the kth source is stored in vector $\mathbf{u}_k \in \mathbf{R}^{n\times 1}$. Then, the element of the kth TAC in time t, $x_{t,k}$, can be written as

$$x_{t,k} = \sum_{i=1}^{t} b_{t-i+1} u_{i,k}. \tag{3.76}$$

More elegant form of (3.76) can be found using linear algebra by constructing the matrix B from elements of vector $\mathbf{b}$ as

$$B = \begin{pmatrix} b_1 & 0 & 0 & 0 \\ b_2 & b_1 & 0 & 0 \\ \dots & b_2 & b_1 & 0 \\ b_n & \dots & b_2 & b_1 \end{pmatrix}. \tag{3.77}$$

Using this notation, the kth TAC, $\mathbf{x}_k$, can be rewritten as

$$\mathbf{x}_k = B\mathbf{u}_k. \tag{3.78}$$

The matrix form of convolution is straightforward extension of the equation (3.78):

$$X = BU, \tag{3.79}$$

where matrix U stores convolution kernels in columns, $U = [\mathbf{u}_1, \ldots, \mathbf{u}_k] \in \mathbf{R}^{n\times r}$. The shapes of the convolution kernels and the input function are important in analysis and diagnosis in medical practice [32].

3.3.3.1. Prior of Input Function

We select the prior model for input function, vector $\mathbf{b}$, as follows:

$$f(\mathbf{b}|\varsigma) = t\mathcal{N}_{\mathbf{b}}(\mathbf{0}_{n,1}, \varsigma^{-1}I_n, [0,\infty]), \tag{3.80}$$

$$f(\varsigma) = \mathcal{G}_{\varsigma}(\zeta_0, \eta_0), \tag{3.81}$$

where ς plays to role of scaling parameter and ζ_0, η_0 selected prior constants.

Posterior distributions are recognized from the VB-marginals to have standard distributional forms

$$\tilde{f}(\mathbf{b}|D) = t\mathcal{N}_{\mathbf{b}}(\boldsymbol{\mu}_{\mathbf{b}}, \Sigma_{\mathbf{b}}, [0,\infty]), \tag{3.82}$$

$$\tilde{f}(\varsigma|D) = \mathcal{G}_{\varsigma}(\zeta, \eta), \tag{3.83}$$

with shaping parameters

$$\Sigma_{\mathbf{b}} = \left(\widehat{\varsigma} I_n + \widehat{\omega} \sum_{i,j=1}^{r} (\widehat{\overline{\mathbf{a}}_i^T \overline{\mathbf{a}}_j}) \left(\sum_{k,l=0}^{n-1} \Delta_k^T \Delta_l \widehat{u_{k+1,j} u_{l+1,i}} \right)\right)^{-1}, \tag{3.84}$$

$$\boldsymbol{\mu}_{\mathbf{b}} = \Sigma_{\mathbf{b}} \widehat{\omega} \sum_{k=1}^{r} \left(\sum_{j=0}^{n-1} \Delta_j \widehat{u_{j+1,k}} \right)^T D^T \widehat{\mathbf{a}_k}, \tag{3.85}$$

$$\zeta = \zeta_0 + \frac{n}{2}, \tag{3.86}$$

$$\eta = \eta_0 + \frac{1}{2} \mathrm{tr}\left(\widehat{\mathbf{b}^T \mathbf{b}}\right). \tag{3.87}$$

Here, 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 associated VB-moments are

$$\widehat{\mathbf{b}} = \mathrm{M}_1^{\mathrm{tN}}\left(\boldsymbol{\mu}_{\mathbf{b}}, \Sigma_{\mathbf{b}}, 0, \infty\right), \tag{3.88}$$

$$\widehat{\mathbf{b}^T \mathbf{b}} = \mathrm{M}_2^{\mathrm{tN}}\left(\widehat{\mathbf{b}}, \boldsymbol{\mu}_{\mathbf{b}}, \Sigma_{\mathbf{b}}, 0, \infty\right), \tag{3.89}$$

$$\widehat{\varsigma} = \frac{\zeta}{\eta}, \tag{3.90}$$

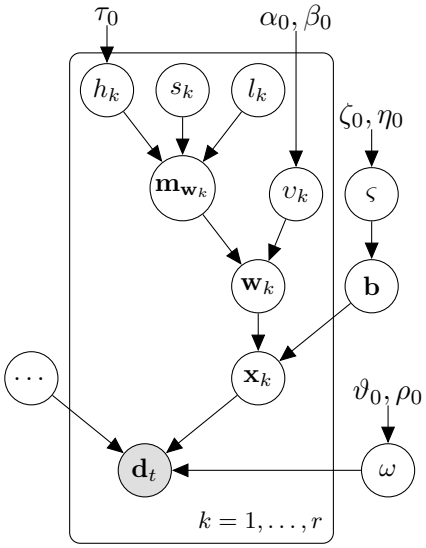

Figure 3.9.: Hierarchical convolution prior model of source TACs using piece-wise linear model of convolution kernels.

In the following text, we need to compute moments $\widehat{B}$ and $\widehat{B^T B}$ which will be prepared here with the help of the auxiliary matrix Δ_k. The moments are

$$\widehat{B} = \sum_{j=0}^{n-1} \Delta_j \widehat{b_{j+1}}, \tag{3.91}$$

$$\widehat{B^T B} = \sum_{j=1}^{n-1} \sum_{l=1}^{n-1} \Delta_j^T \Delta_l \widehat{b_{j+1} b_{l+1}}. \tag{3.92}$$

Remark 6. We have studied other versions of input function prior. Increases of input function have been modeled to ensure the monotonicity of the input function [112] and ARD prior

$$f(b_j|\varsigma_j) = t\mathcal{N}_{b_j}(0, \varsigma_j^{-1}), \tag{3.93}$$

$$f(\varsigma_j) = \mathcal{G}_{\varsigma_j}(\zeta_{j,0}, \eta_{j,0}); \tag{3.94}$$

for each element of input function have been proposed [113]; however, no significant differences from the prior model (3.80)–(3.81) have been observed.

3.3.3.2. Piece-wise Linear Prior of Convolution Kernels

We proposed and studied the piece-wise linear model of convolution kernels in [113, 112, 102]. The main idea is that each convolution kernel is restricted to be composed from a linear plateau following by linear decline to zero, see Figure 3.8, left, as an example. This can be modeled using increments as proposed in [64].

The matrix X is decomposed according to (3.79) to input function, vector $\mathbf{b}$, and the matrix with convolution kernels in column, matrix U. Each element of the matrix U is further modeled using non-negative increments:

$$u_{t,k} = \sum_{j=t}^{n} w_{j,k}, \tag{3.95}$$

where vector $\mathbf{w}_k \in \mathbf{R}^{n\times 1}$ contains increments forming the vector $\mathbf{u}_k$. We select model where h_k is the height of each non-zero increment, s_k is the starting point of increments, and l_k is the length of increments; hence, $\mathbf{w}_k$ has a prior structure such as

$$\mathbf{w}_k = [0, \ldots, 0, h_k, \ldots, h_k, 0, \ldots, 0] \equiv \mathbf{m}_{\mathbf{w}_k}, \tag{3.96}$$

where the positions of cluster-based structure with h_k is fully determined by start s_k and length l_k.

Following the VB methodology, each parameter has its prior distribution:

$$f(\mathbf{w}_k|\upsilon_k) = t\mathcal{N}_{\mathbf{w}_k}\left(\mathbf{m}_{\mathbf{w}_k}, \upsilon_k^{-1} I_n, [0,\infty]\right), \tag{3.97}$$

$$f(\upsilon_k) = \mathcal{G}_{\upsilon_k}(\alpha_0, \beta_0), \tag{3.98}$$

$$f(h_k) = t\mathcal{N}_{h_k}(0, \tau_0, [0,\infty]), \tag{3.99}$$

$$f(s_k) = \mathcal{U}_{s_k}(0, n), \tag{3.100}$$

$$f(l_k|s_k) = \mathcal{U}_{l_k}(0, n - s_k), \tag{3.101}$$

where $\alpha_0, \beta_0, \tau_0$ are selected prior constants. This hierarchical prior is shown in Figure 3.9.

Posterior distributions are recognized from the VB-marginals to have standard distributional forms

$$\tilde{f}(\mathbf{w}|D, r) = t\mathcal{N}_{\mathbf{w}}(\boldsymbol{\mu}_{\mathbf{w}}, \Sigma_{\mathbf{w}}, [0,\infty]), \tag{3.102}$$

$$\tilde{f}(\upsilon_k|D, r) = \mathcal{G}_{\upsilon_k}(\alpha_k, \beta_k), \tag{3.103}$$

where $\mathbf{w} = \mathrm{vec}(W)$, with shaping parameters

$$\Sigma_{\mathbf{w}} = \left(\widehat{(A^T A)}^T \otimes \widehat{\omega} C^T \widehat{B^T B} C) + (\mathrm{diag}(\widehat{\boldsymbol{\upsilon}}) \otimes I_n)\right)^{-1}, \tag{3.104}$$

$$\boldsymbol{\mu}_{\mathbf{w}} = \Sigma_{\mathbf{w}}(\mathrm{diag}(\widehat{\boldsymbol{\upsilon}})\mathrm{vec}\left((C^T \widehat{B^T B} C)^{-1} C^T \widehat{B}^T D^T \widehat{A}(\widehat{A^T A})^{-1}\right) \tag{3.105}$$

$$+ \Sigma_{\mathbf{w}}((\mathrm{diag}(\widehat{\boldsymbol{\upsilon}}) \otimes I_n)\mathrm{vec}(\widehat{M_W}), \tag{3.106}$$

$$\alpha_k = \alpha_0 + \frac{n}{2}, \tag{3.107}$$

$$\beta_k = \beta_0 + \frac{1}{2}(\widehat{W^T W})_{k,k} + \frac{1}{2}(-2\widehat{W}^T \widehat{M_W})_{k,k} + \frac{1}{2}(\widehat{M_W^T M_W})_{k,k}. \tag{3.108}$$

Here, M_W contains prior vectors of $\mathbf{w}_k$, $\mathbf{m}_{\mathbf{w}_k}$, in columns composed of estimates of h_k, s_k, and l_k (obtained using EM algorithm, see [106] for details), auxiliary

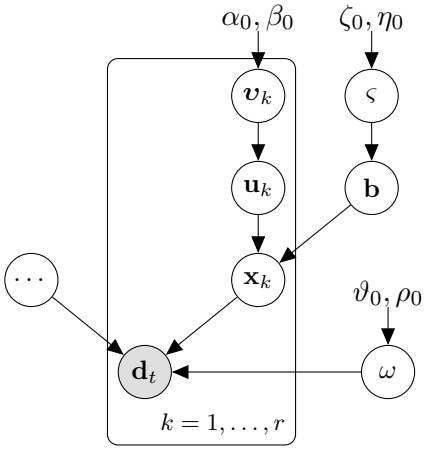

Figure 3.10.: Hierarchical convolution prior model of source TACs using automatic relevant determination model of convolution kernels.

matrix $\Delta_k \in \mathbf{R}^{n\times n}$ is defined as $(\Delta_k)_{i,j} = \begin{cases} 1, & \text{if } i-j=k, \\ 0, & \text{otherwise,} \end{cases}$, and auxiliary matrix $C \in \mathbf{R}^{n\times n}$ is defined as

$$C = \begin{pmatrix} 1 & 1 & \cdots & 1 & 1 \\ 0 & 1 & \cdots & 1 & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 1 \end{pmatrix}. \tag{3.109}$$

The associated VB-moments are

$$\widehat{\mathbf{w}} = \mathrm{M}_1^{\mathrm{tN}}\left(\mu_{\mathbf{w}}, \Sigma_{\mathbf{w}}, 0, \infty\right), \tag{3.110}$$

$$\widehat{\mathbf{w}^T\mathbf{w}} = \mathrm{M}_2^{\mathrm{tN}}\left(\widehat{\mathbf{w}}, \mu_{\mathbf{w}}, \Sigma_{\mathbf{w}}, 0, \infty\right) \tag{3.111}$$

$$\widehat{\xi_k} = \frac{\kappa_k}{\nu_k}, \tag{3.112}$$

where functions $\mathrm{M}_1^{\mathrm{tN}}()$ and $\mathrm{M}_2^{\mathrm{tN}}()$ are defined in Appendix A.1.4.

3.3.3.3. ARD Prior of Convolution Kernels

The convolution assumption of the TACs has been proven to be relevant and helpful [104, 113]; however, too restrictive models of the convolution kernels such as [24], exponential form, or [113], piece-wise linear form, have limited area of usage and work only under ideal conditions that hold their assumptions. This is hardly met in such a complicated system as a living organism. Therefore, we seek a more relaxed parametrization of the convolution kernels that could reflect high variability of dynamics of sources in dynamic medical imaging.

We adopt the general convolution model of TACs, (3.79),

$$X = BU, \tag{3.113}$$

where B is composed of elements of the vector with input function $\mathbf{b}$ as defined in (3.77) and U is composed of convolution kernels $\mathbf{u}_k$ stored columnwise. The only further assumption on the convolution kernels is that, once again, the convolution kernels are sparse which can be modeled using ARD priors in the similar way as in Section 2.4.

Each column of the matrix U; $\mathbf{u}_k$, $k = 1, \ldots, r$, is modeled using the truncated normal distribution as

$$f(\mathbf{u}_k|\boldsymbol{\upsilon}_k) = t\mathcal{N}_{\mathbf{u}_k}\left(\mathbf{0}_{n,1}, \mathrm{diag}(\boldsymbol{\upsilon}_k)^{-1}, [0,\infty]\right), \tag{3.114}$$

while the prior for vector with precision parameters for each convolution kernel $\mathbf{u}_k$, $\boldsymbol{\upsilon}_k \in \mathbf{R}^{n\times 1}$, has conjugate Gamma prior distribution

$$f(\boldsymbol{\upsilon}_k) = \prod_{j=1}^{n} \mathcal{G}_{\upsilon_{j,k}}(\alpha_0, \beta_0) \tag{3.115}$$

with selected prior constants α_0, β_0. This hierarchical prior is shown in Figure 3.10.

We will discuss the VB method for the constructed ARD prior model. In this case, the VB method is not straightforward and several partitioning of the model parameters may by considered.

The log-likelihood function (3.4) with imposed convolution model of TACs, (3.113), is

$$\ln f(D|A, U, \mathbf{b}, \omega) \propto \frac{pn}{2}\ln\omega - \frac{1}{2}\omega\mathrm{tr}\left[\left(D - AU^TB^T\right)\left(D - AU^TB^T\right)^T\right]. \tag{3.116}$$

No partitioning to match standard distribution in form of the matrix normal distribution has been found for the matrices A and U. The partitioning for the matrix A was already derived in (3.29). No such situation happen for the convolution kernels U where issue with the term $\mathrm{tr}\left(AB^TU^TUBA^T\right)$ from the equation (3.116) arises. However, full posterior distributional form can be derived for vectorized matrix U, using multivariate normal distribution rather than using matrix normal distribution since Kronecker structure of the covariance matrix does not preserve here.

Dependent Convolution Kernels Using Vectorization Operator Let the vector $\mathbf{u} \in \mathbf{R}^{nr\times 1}$ without subscript arises as the vectorization of the matrix U, $\mathbf{u} = \mathrm{vec}(U)$. Then, it holds:

$$\mathrm{tr}\left(AB^TU^TUBA^T\right) = \mathbf{u}^T\left(A^TA \otimes B^TB\right)\mathbf{u}, \tag{3.117}$$

Note that all elements of convolution kernels interact with each other lowing the approximation error of the VB method and allowing the estimate the full posterior $f(\mathbf{u}|D)$. We published this partitioning in [108] where it is shown that this version outperforms the parametrization with independent convolution kernels which will be briefly revised in Remark 7 at the end of this section.

After constructing the VB-marginals, the recognized standard distributional forms are

$$\tilde{f}(\mathbf{u}|D) = t\mathcal{N}_{\mathbf{u}}(\boldsymbol{\mu}_{\mathbf{u}}, \Sigma_{\mathbf{u}}, [0, \infty]), \tag{3.118}$$

$$\tilde{f}(\upsilon_{j,k}|D) = \mathcal{G}_{\upsilon_{j,k}}(\alpha_{j,k}, \beta_{j,k}), \tag{3.119}$$

with shaping parameters

$$\Sigma_{\mathbf{u}} = \left(\widehat{A^T A} \otimes \widehat{\omega} \widehat{B^T B} + \widehat{\Upsilon}\right)^{-1}, \tag{3.120}$$

$$\boldsymbol{\mu}_{\mathbf{u}} = \Sigma_{\mathbf{u}} \left(\widehat{A^T A} \otimes \widehat{\omega} \widehat{B^T B}\right) \mathrm{vec}\left(\widehat{B^T B}^{-1} \widehat{B}^T D^T \widehat{A} \widehat{A^T A}^{-1}\right), \tag{3.121}$$

$$\alpha = \alpha_0 + \frac{1}{2}\mathbf{1}_{nr,1}, \tag{3.122}$$

$$\beta = \beta_0 + \frac{1}{2}\mathrm{diag}\left(\widehat{\mathbf{u}\mathbf{u}^T}\right), \tag{3.123}$$

where $\Upsilon \in \mathbf{R}^{nr \times nr}$ is a diagonal matrix with $\boldsymbol{\upsilon}_k, k = 1, \ldots, r$, on its diagonal.

The associated VB-moments are

$$\widehat{\mathbf{u}} = \mathrm{M}_1^{\mathrm{tN}}\left(\boldsymbol{\mu}_{\mathbf{u}}, \Sigma_{\mathbf{u}}, 0, \infty\right), \tag{3.124}$$

$$\widehat{\mathbf{u}^T \mathbf{u}} = \mathrm{M}_2^{\mathrm{tN}}\left(\widehat{\mathbf{u}}, \boldsymbol{\mu}_{\mathbf{u}}, \Sigma_{\mathbf{u}}, 0, \infty\right), \tag{3.125}$$

$$\widehat{\Upsilon} = \mathrm{diag}\left(\alpha \circ \beta^{-1}\right), \tag{3.126}$$

where functions $\mathrm{M}_1^{\mathrm{tN}}()$ and $\mathrm{M}_2^{\mathrm{tN}}()$ are defined in Appendix A.1.4.

Remark 7. It is possible to consider computational simplification: to partition the matrix U as its columns, i.e. to assume convolution kernels $\mathbf{u}_k$ to be conditionally independent. The problematic term from the equation (3.116) can be rewritten for kth convolution kernel $\mathbf{u}_k$ as follows

$$\mathrm{tr}\left(AB^T U^T U B A^T\right) = \mathrm{tr}\left(2 \sum_{l=1, l \neq k}^{r} B^T B \mathbf{u}_l \mathbf{a}_l^T \mathbf{a}_k \mathbf{u}_k^T\right) + \mathrm{tr}\left(\mathbf{u}_k^T B^T B \mathbf{a}_k^T \mathbf{a}_k \mathbf{u}_k\right). \tag{3.127}$$

From this partitioning, $\mathbf{u}_k$ can be easily recognized as normally distributed and $\mu_{\mathbf{u}_k}$ and $\Sigma_{\mathbf{u}_k}$ can be computed.

We derived this partitioning in [101] and shown that the model is promising for given true number of sources, r. However; the algorithm has tendency to split the strongest source when the initial number of sources is overestimated due to the independence of the convolution kernels, see Section 4.5.3.4 for demonstration of this issue.

4. Blind Source Separation Methods

As long as you are set that the probability is going to be zero, then nothing's going to change your mind. If you have decided that the sun rises each morning because it has always done so in the past, nothing is going change your mind except one morning when the sun fails to appear.

(Albert Madansky)

The blind source separation (BSS) model (1.1) was studied with different assumptions using hierarchical prior models for all three unknown matrices A, X, and E in Chapter 3. In Variational Bayes (VB) inference, we assume that these three variables are conditionally independent. This allow us to arbitrary combine priors for these parameters, see Section 2.3.2, to obtained specific BSS methods. Thus, we can combine the priors without additional computational of logical difficulties.

Prior models described in Chapter 3 are reviewed in the Table 4.1 and combinations of prior models are suggested. We will not reach all possible combinations but only those that were tested and published. We will discuss potential other combinations; however, they are mainly a subset of the published models.

In the second part of this chapter, we will discuss computational aspects of the algorithms such as initialization, estimation of the number of sources, or numerical stability. The next section will be devoted to state-of-the-art algorithms for the BSS problem related to medical image sequences. Example results on synthetic phantom are given in the end of this chapter.

Priors	TACs			
Images		Isotropic	Conv. P.-wise linear	Conv. ARD
	Isotropic	BSS+ (4.1)	BCMS (4.3)	–
	Mixture	FAROI (4.2)	–	–
	ARD	–	–	S-BSS-vecDC (4.4)

Table 4.1.: Reviews of the prior models described in Chapter 3 with selected combinations.

Algorithm 4.1 Blind Source Separation with Positivity Constraints (BSS+) algorithm.

1. Initialization:
   a) Set prior parameters $\alpha_0, \beta_0, \vartheta_0, \rho_0$.
   b) Set initial values for $\widehat{A}, \widehat{A^T A}, \widehat{X}, \widehat{X^T X}, \widehat{\boldsymbol{\upsilon}}, \widehat{\omega}$.
   c) Set the initial number of sources r_{max}.
2. Iterate until convergence is reached using computation of shaping parameters (and related moments) of:
   a) Source images μ_A, Σ_A using (2.113)–(2.114).
   b) Time-activity curves μ_X, Σ_X using (3.59)–(3.60).
   c) Variance of TACs $\boldsymbol{\alpha}, \boldsymbol{\beta}$ using (3.61)–(3.62).
   d) Variance of noise ϑ, ρ using (3.8)–(3.9).
3. Report estimates $\widehat{A}$ and $\widehat{X}$.

4.1. Blind Source Separation with Positivity (BSS+)

Blind Source Separation with Positivity Constraints (BSS+) [76, 97] combines prior model of source images from Section 2.5.1.1 and isotropic prior model of TACs from Section 3.3.1 where relevance of each source has been modeled. The resulting algorithm is summarized in Algorithm 4.1.

The key assumption of the algorithm is that the priors of the source images as well as that of the TACs are isotropic. This assumption serves as the separability criterion; however, the criterion with no physical or biological meaning. It means that although the general decomposition (1.1), $D = AX^T$, could be fulfilled, the separation results could not correspond to the expected biological sources.

The BSS+ algorithm can be downloaded from `http://www.utia.cz/AS/softwaretools/image_sequences`.

4.2. Factor Analysis with ROI (FAROI)

Factor analysis with integrated regions of interest (FAROI) method [100] models the source images as mixtures as proposed in Section 3.2.2 while the TACs are modeled using isotropic priors from Section 3.3.1. The resulting algorithm is summarized in Algorithm 4.2.

The idea to model sparsity of source images using mixtures (3.26),

$$f(a_{i,k}|\xi_k, \mathbf{i}_{i,k}) = \mathcal{U}_{a_{i,k}}(0,1)^{\mathbf{i}_{i,k}} t\mathcal{N}_{a_{i,k}}(0, \xi_k^{-1}, [0,\infty])^{(1-\mathbf{i}_{i,k})}, \tag{4.1}$$

Algorithm 4.2 The factor analysis with integrated regions of interest (FAROI) algorithm.

1. Initialization:
   a) Set prior parameters $\alpha_0, \beta_0, \vartheta_0, \rho_0, \phi_0, \psi_0$.
   b) Set initial values for $\widehat{A}, \widehat{A^T A}, \widehat{\mathbf{i}}, \widehat{\boldsymbol{\xi}}, \widehat{X}, \widehat{X^T X}, \widehat{\boldsymbol{\upsilon}}, \widehat{\omega}$.
   c) Set the initial number of sources r_{max}.
2. Iterate until convergence is reached using computation of shaping parameters (and related moments) of:
   a) Source images μ_A, Σ_A using (3.33)–(3.34).
   b) Variance of noise parts of images $\phi_k, \psi_k \; \forall k$.(3.35)–(3.36).
   c) Indicator $\lambda_{i,k} \; \forall i, \forall k$ (3.37).
   d) Time-activity curves μ_X, Σ_X using (3.59)–(3.60).
   e) Variance of TACs $\boldsymbol{\alpha}, \boldsymbol{\beta}$ using (3.61)–(3.62).
   f) Variance of noise ϑ, ρ using (3.8)–(3.9).
3. Report estimates $\widehat{A}$ and $\widehat{X}$.

as a normal distributed noise part and uniform distributed signal part is proven to be relevant [100]; however, the main issue with this model is numerical instability and unreliable convergence.

The FAROI algorithm can be downloaded from `http://www.utia.cz/AS/softwaretools/image_sequences`.

4.3. Blind Compartment Model Separation (BCMS)

Blind compartment model separation (BCMS) was introduced in [106] and applied to dynamic renal scintigraphy in [102, 112, 113]. The model combines the isotropic prior of source images, Section 3.2.1, and convolution model of TACs with piece-wise linear parametrization of convolution kernels, Section 3.3.3.2. The BCMS algorithm is summarized in Algorithm 4.3. The parametrization of convolution kernels as a linear plateau and then linear decline to zero, see Figure 3.8 (right), was motivated by specific application in dynamic renal scintigraphy: differential renal function (DRF) estimation. The DRF is relative ratio of activity in left and right kidney, which is computed from the uptake part of a sequence, i.e. the initial part of the sequence when the activity is only accumulated with no excretion from a kidney. It is assumed that the assumption of the piece-wise linear convolution kernels is valid on this uptake part.

However; the practical usage of the BCMS algorithm show that the selection

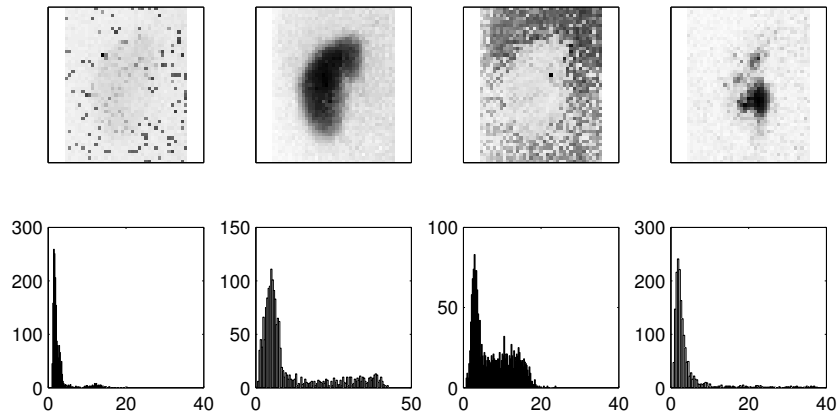

Figure 4.1.: Example results of the FAROI algorithm on data from dynamic renal scintigraphy using source images (top) and their histograms (bottom).

Algorithm 4.3 The blind compartment models separation (BCMS) algorithm.

1. Initialization:
   a) Set prior parameters $\alpha_0, \beta_0, \vartheta_0, \rho_0, \phi_0, \psi_0, \tau_0, \zeta_0, \eta_0$.
   b) Set initial values for $\widehat{A}, \widehat{A^T A}, \widehat{\boldsymbol{\xi}}, \widehat{\mathbf{w}_k}, \widehat{\mathbf{w}_k^T \mathbf{w}_k}, \widehat{\boldsymbol{\upsilon}}, \widehat{\mathbf{b}}, \widehat{\mathbf{b}^T \mathbf{b}}, \widehat{\varsigma}, \widehat{\omega}$, and $\mathbf{m}_{\mathbf{w}_k}$.
   c) Set the initial number of sources r_{max}.
2. Iterate until convergence is reached using computation of shaping parameters (and related moments) of:
   a) Source images μ_A, Σ_A using (3.18)–(3.19).
   b) Variance of source images $\boldsymbol{\phi}, \boldsymbol{\psi}$ using (3.20)–(3.21).
   c) Increases forming convolution kernels $\mu_{\mathbf{w}}, \Sigma_{\mathbf{w}}$ using (3.104)–(3.105).
   d) Variance of increases $\boldsymbol{\alpha}, \boldsymbol{\beta}$ using (3.107)–(3.108).
   e) Compute prior means $\mathbf{m}_{\mathbf{w}_k}$ $\forall k$ using EM algorithm, see [106] for details.
   f) Input function $\mu_{\mathbf{b}}, \Sigma_{\mathbf{b}}$ and their variance ζ, η using (3.84)–(3.87).
   g) Variance of noise ϑ, ρ using (3.8)–(3.9).
3. Report estimates $\widehat{A}$ and $\widehat{X}$.

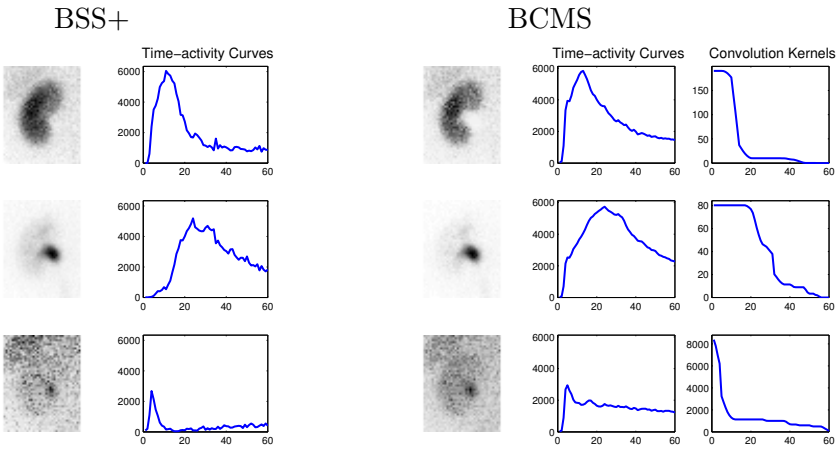

Figure 4.2.: Example results from the BSS+ algorithm (left) and BCMS algorithm (right) on selected sequence from dynamic renal scintigraphy.

of the uptake part of the sequence is a crucial step in DRF analysis and the sensitivity to this selection is unacceptably high. Moreover, the usage of the BCMS is strongly limited to the problems were all sources are activated at the beginning of a sequence; otherwise, the BCMS algorithm provides results where the activity of delayed sources are overestimated in the beginning or where these sources are not separated. This issue is demonstrated in Figure 4.2 on a selected sequence from dynamic renal scintigraphy. The representative results from the BSS+ algorithm are shown on the left with the second source, pelvis, the part of kidney with delayed activity. The results from the BCMS algorithm is given on the right. See that each estimate of convolution kernel starts from the beginning of the sequence which results in overestimated TAC in the second source, pelvis, in comparison with the BSS+ results.

In sum, the BCMS algorithm is appropriate for specific tasks in dynamic medical imaging; however, general usage is limited due to the strict parametrization of the convolution kernels using prior piece-wise linear model.

The BCMS algorithm can be downloaded from `http://www.utia.cz/AS/softwaretools/image_sequences`.

4.4. Sparse BSS and Deconvolution (S-BSS-vecDC)

Sparse blind source separation and deconvolution (S-BSS-vecDC) model [101, 108] is trying to suppress the disadvantages of parametrization in modeling of source images in FAROI model, Section 4.2, and convolution kernels in BCMS model, Section 4.3. The mixture model of pixels and the piece-wise linear parameterization of convolution kernels are replaced by the only assumption that both, source images and convolution kernels, are most likely sparse. The spar-

Algorithm 4.4 The sparse blind source separation and vectorized deconvolution (S-BSS-vecDC) algorithm.

1. Initialization:
   a) Set prior parameters $\alpha_0, \beta_0, \vartheta_0, \rho_0, \phi_0, \psi_0, \zeta_0, \eta_0$.
   b) Set initial values for $\widehat{A}, \widehat{A^T A}, \widehat{\bar{\boldsymbol{\xi}}}_i, \widehat{\mathbf{u}}, \widehat{\mathbf{u}^T \mathbf{u}}, \widehat{\Upsilon}, \widehat{\mathbf{b}}, \widehat{\mathbf{b}^T \mathbf{b}}, \widehat{\varsigma}, \widehat{\omega}$.
   c) Set the initial number of sources r_{max}.
2. Iterate until convergence is reached using computation of shaping parameters (and related moments) of:
   a) Source images $\boldsymbol{\mu}_{\bar{\mathbf{a}}_i}, \Sigma_{\bar{\mathbf{a}}_i}$ and their variances $\boldsymbol{\phi}_i, \boldsymbol{\psi}_i \ \forall i$ using (3.48)–(3.51).
   b) Convolution kernels $\boldsymbol{\mu}_{\mathbf{u}}, \Sigma_{\mathbf{u}}$ and their variances $\boldsymbol{\alpha}, \boldsymbol{\beta}$ using (3.120)–(3.123).
   c) Input function $\boldsymbol{\mu}_{\mathbf{b}}, \Sigma_{\mathbf{b}}$ and its variance ζ, η using (3.84)–(3.87).
   d) Variance of noise ϑ, ρ using (3.8)–(3.9).
3. Report estimates $\widehat{A}$ and $\widehat{X}$.

sity is modeled using the ARD principle introduced in Section 2.4 and applied to the prior model of source images, Section 3.2.3, and the prior model of TACs, Section 3.3.3.3. The S-BSS-vecDC algorithm is summarized in Algorithm 4.4. The term 'vec' refers to the version of the VB inference where the convolution kernels are dependent (vectorized) as described in Section 3.3.3.3.

We will demonstrate the behavior of the S-BSS-vecDC algorithm on an example run on the same data as in Section 4.3. The results of the S-BSS-vecDC algorithm are given in Figure 4.3 using (from the left) pixels variances, source images, TACs, and convolution kernels. It can be seen that the resulting TACs do not suffer from any imposed curve parametrization and delayed source (pelvis, the second one) is correctly estimated with no activity in the beginning. An artifact of the method is its tendency to estimate to non-smooth convolution kernels in cases when the input function does not start from the beginning of the sequence, see the first convolution kernel in Figure 4.3. This behavior and its cause will be discussed in Section 4.5.3. However, this non-smoothness does not propagate to TACs due to the convolution operation.

The S-BSS-vecDC algorithm can be downloaded from `http://www.utia.cz/AS/softwaretools/image_sequences`.

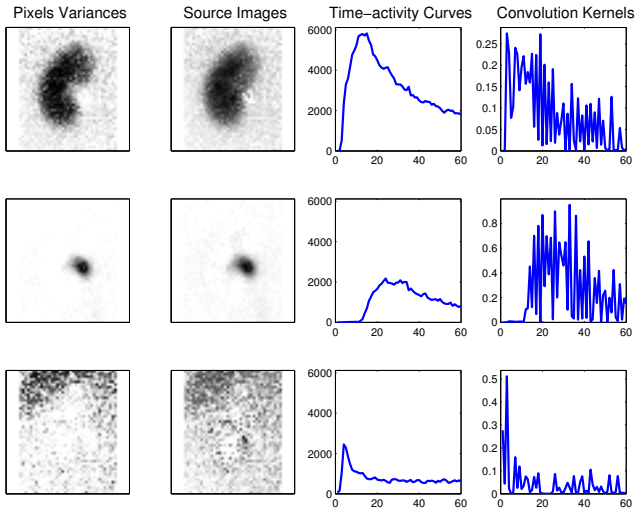

Figure 4.3.: Example results from the S-BSS-vecDC algorithm on selected sequence from dynamic renal scintigraphy.

4.5. Computational Aspects of the Proposed Algorithms

4.5.1. Initialization

The initialization step of all algorithms is crucial for reasonably fast convergence as well as for convergence to biologically meaningful solution since the VB methodology suffers from local minima.

One such possibility is to model several biological sources covering the possible varieties of TACs. The TACs can be selected directly; however, since we adopted a convolution model, we propose initialization of the input function, $\mathbf{b}$, and the convolution kernels, U, from which TACs can be easily computed for the algorithms without the convolution assumption. The initialization for $\mathbf{b}$ is selected as $\exp\left(-\frac{1,\ldots,n}{3}\right)$ to reach exponentially decreasing vector which is close to the reality from our experience. The convolution kernels, vectors $\mathbf{u}_k$, are selected as proposed in Figure 4.4. Here, the selection is motivated by dynamic renal scintigraphy following typical expected convolution kernels of: (1) the blood, (2) the parenchyma, (3) the pelves, (4) the tissue background, (5) the urinary bladder, and the rest of kernels are selected to cover various dynamics of possible sources. The initial matrix X is then computed according to the (3.113) as $X_{\text{init}} = BU$.

Once we have a good guest of the matrix X_{init}, the initial guess of the matrix A_{init} can be computed from the data matrix D using least squares as

$$A_{\text{init}} = DX_{\text{init}}\left(X_{\text{init}}^T X_{\text{init}}\right)^{-1}. \tag{4.2}$$

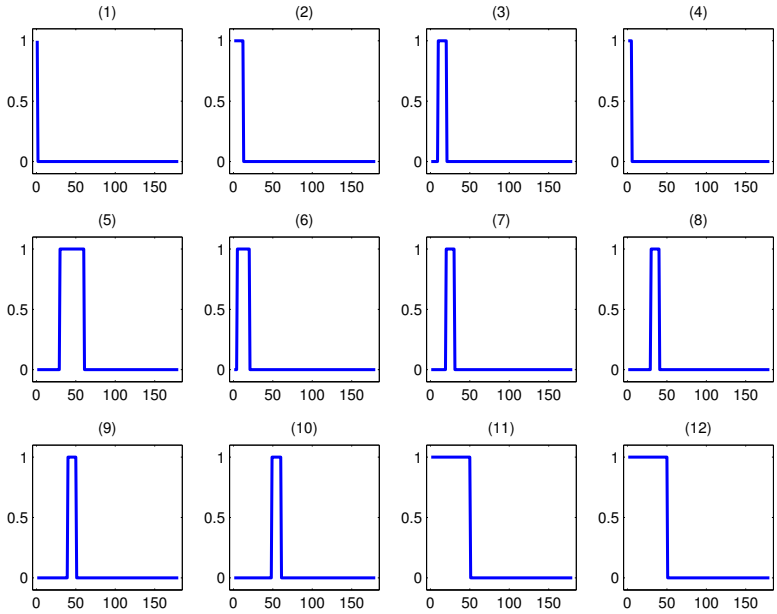

Figure 4.4.: Initialization of the convolution kernels.

Initialization of the variance of the noise is computed according to [97] from eigenvalues of the data matrix $D \in \mathbf{R}^{p \times n}$ as

$$\omega_{\text{init}} = \left| \frac{p}{\min\left(\lambda_{D^T D}\right)} \right|, \tag{4.3}$$

where $\lambda_{D^T D}$ are eigenvalues of the matrix $D^T D$.

The rest of the parameters are selected as ones and the prior parameters (subscripted by 0) are chosen close to zeros such as 10^{-10} approaching uninformative Jeffrey's priors [54].

4.5.2. Estimation of the Number of Sources

The number of sources, r, can be manually preselected for the whole procedure as a static parameter; however, the necessity of choosing the proper number of sources often limits the use of an algorithm. In some algorithms such as BSS+, FAROI, or BCMS, the number of sources was selected using the ARD principle on the whole source images or TACs based on its relevance [76]. In the S-BSS-vecDC algorithm, the ARD principle is used in both, pixel resolution and time-domain resolution; hence, no such principle is used for the whole sources. The ARD prior for the whole sources could be enforced by additional parametrization which would further complicate the algorithm. To avoid this even more complex

modeling, we proposed an alternative: estimation of relevance of each source based on the estimate of the precision parameter ω [108].

Specifically, the VB solution of the scalar version of the model (1.1), $d = ax+e$, yields non-zero signal (i.e. $\hat{a}\hat{x} > 0$) when $d > 2\sqrt{\omega^{-1}}$ [101]. Using this inference bound, the sum of $d_{i,j}^2$ corresponding to a pixel from the kth source, should be n times greater than the noise level:

$$\mathbf{x}_k^T \mathbf{x}_k > 2n\omega^{-1}. \tag{4.4}$$

This observation will be used as a criterion for removal of weak sources within the iterative procedure. Since removal of a source influence all others we disallow further removal for the next 50 iterations of the algorithm and only one source could be removed in one time-point. The algorithm starts from $r_{\max}$ sources and terminates if all sources satisfy (4.4) or the minimum number of sources $r_{\min}$ is reached. The interval $[r_{\min}, r_{\max}]$ can be specified by an expert or heuristically. Here, we propose heuristics based on differences of singular values σ_i of the data matrix D. Specifically, when σ_{i+1}^2 is less than 95% of σ_i^2, then $r_{\max} = i$. We observed that this starting point for r overestimates significantly the true value of r in medical image datasets. The same heuristics is used for selection of $r_{\min}$ but the coefficient is set to 75%. This percentages are valid especially in dynamic scintigraphy and should be carefully considered in other applications of the algorithm.

In general, we recommend to slightly overestimate the number of $r_{\max}$ of the relevant sources since the redundant source would be estimated to be weak or removed by the automatic criteria (4.4). If $r_{\max}$ is chosen lower than the true number of sources, the sources will be always mixed. Condition (4.4) may remove even a valid signal if the number of non-zero elements in $\mathbf{x}_k$ is much lower than n. Removal of sources becomes more aggressive with growing n.

This principle could be used for all derived algorithms.

4.5.3. Numerical Issue with Inversion of Input Function Moments

The equations in Chapter 3 have a potential computational issue since it contains inversions of matrices. We will use the moments for distribution of $\mathbf{u}$ (3.120)–(3.123) from Section 3.3.3.3 as an example:

$$\Sigma_{\mathbf{u}} = \left(\widehat{A^T A} \otimes \widehat{\omega}\widehat{B^T B} + \mathrm{diag}(\mathrm{vec}(\widehat{\Upsilon}))\right)^{-1}, \tag{4.5}$$

$$\mu_{\mathbf{u}} = \Sigma_{\mathbf{u}} \left(\widehat{A^T A} \otimes \widehat{\omega}\widehat{B^T B}\right) \mathrm{vec}\left(\widehat{B^T B}^{-1} \widehat{B}^T D^T \widehat{A} \widehat{A^T A}^{-1}\right), \tag{4.6}$$

where the moment $\widehat{B^T B}^{-1}$ has to be used. This moment suffers from numerical instability when the maximum of the input function, vector $\mathbf{b}$, is not on the first element. We will demonstrate this issue on the following example.

Suppose that $n = 5$; hence, $\mathbf{b} \in \mathbf{R}^{5\times 1}$. Consider two cases, the first with maximum activity in the first element, $\mathbf{b}^1$, and the second in the second element, $\mathbf{b}^2$:

$$\mathbf{b}^{(1)} = \begin{pmatrix} 1 \\ 0.2 \\ 0.1 \\ 0.1 \\ 0.1 \end{pmatrix}, \mathbf{b}^{(2)} = \begin{pmatrix} 0.1 \\ 1 \\ 0.2 \\ 0.1 \\ 0.1 \end{pmatrix}. \tag{4.7}$$

The matrix B is constructed according to (3.77) and resulting terms of interest, B^TB are as follow:

$$\left(B^TB\right)^{(1)} = \begin{pmatrix} 1.07 & 0.24 & 0.13 & 0.12 & 0.1 \\ 0.24 & 1.06 & 0.23 & 0.12 & 0.1 \\ 0.13 & 0.23 & 1.05 & 0.22 & 0.1 \\ 0.12 & 0.12 & 0.22 & 1.06 & 0.2 \\ 0.1 & 0.1 & 0.1 & 0.2 & 1 \end{pmatrix}, \tag{4.8}$$

$$\left(B^TB\right)^{(2)} = \begin{pmatrix} 1.07 & 0.33 & 0.14 & 0.11 & 0.01 \\ 0.33 & 1.06 & 0.32 & 0.12 & 0.01 \\ 0.14 & 0.32 & 1.05 & 0.3 & 0.02 \\ 0.11 & 0.12 & 0.3 & 1.01 & 0.1 \\ 0.01 & 0.01 & 0.02 & 0.1 & 0.01 \end{pmatrix}. \tag{4.9}$$

Now, we can compute the condition number $\kappa_2\left(B^TB\right)$ for spectral norm and symmetric positive matrices for each matrix according to the definition

$$\kappa_2\left(B^TB\right) = \frac{\lambda_{\max}\left(B^TB\right)}{\lambda_{\min}\left(B^TB\right)}, \tag{4.10}$$

where $\lambda_{\max}\left(B^TB\right)$ and $\lambda_{\min}\left(B^TB\right)$ are the maximal and the minimal eigenvalues of the given matrix. Note that condition number κ_2 remain the same for inversion for symmetric positive matrices. The condition numbers in this case are:

$$\kappa_2\left(\left(B^TB\right)^{(1)}\right) = 2.2404, \tag{4.11}$$

$$\kappa_2\left(\left(B^TB\right)^{(2)}\right) = 1.5648 \cdot 10^{10}, \tag{4.12}$$

hence, the computations with the $\mathbf{b}^{(1)}$ are well-conditioned while the operations with $\mathbf{b}^{(2)}$ are ill-conditioned. Therefore, the numerical operations such as inversion with B^TB are very sensitive in the second case.

This can cause a numerical instability in the S-BSS-vecDC algorithm; therefore, we will discuss possibilities how to restore numerical stability.

4.5.3.1. Moore–Penrose Pseudoinverse

We can adopt the Moore–Penrose pseudoinverse [42] instead of the inverse to restore the numerical stability. Specifically, we can discard singular values which are smaller than selected percentage of the mean of all singular values; however, it is very difficult and manual-dependent to selected this percentage. In this work, we discard the singular values that are smaller than 10% of mean of all singular values.

In our example, the condition number $\kappa_2\left(\left(\left(B^TB\right)^{(2)}\right)^{+}\right)$ declines to 2.7638 after use of the pseudoinverse.

4.5.3.2. Covariance Localization

The equations (4.5)–(4.6) can be seen as a solution of a set of linear equations which is typically written as

$$\boldsymbol{\mu}_{\mathbf{u}} = \Sigma_{\mathbf{u}}^{-1}\mathbf{c}, \tag{4.13}$$

where

$$\Sigma_{\mathbf{u}}^{-1} \equiv \widehat{A^TA} \otimes \widehat{\omega}\widehat{B^TB} + \mathrm{diag}(\mathrm{vec}(\widehat{\Upsilon})), \tag{4.14}$$

$$\mathbf{c} \equiv \left(\widehat{A^TA} \otimes \widehat{\omega}\widehat{B^TB}\right) \mathrm{vec}\left(\widehat{B^TB}^{-1}\widehat{B}^TD^T\widehat{A}\widehat{A^TA}^{-1}\right). \tag{4.15}$$

An approach called the covariance localization [36, 47] is used in atmospheric environment research to improve the condition number of a matrix. Here, the matrix L with dominant diagonal elements and decreasing sub-diagonal elements is constructed in a way:

$$L = \begin{pmatrix} 1 & 0.9 & 0.8 & 0.8 & 0.8 \\ 0.9 & 1 & 0.9 & 0.8 & 0.8 \\ 0.8 & 0.9 & 1 & 0.9 & 0.8 \\ 0.8 & 0.8 & 0.9 & 1 & 0.9 \\ 0.8 & 0.8 & 0.8 & 0.9 & 1 \end{pmatrix}, \tag{4.16}$$

where the length of the declining and lower bound of declining are selected parameters. Then, the matrices $\widehat{B^TB}^{-1}$ or $\Sigma_{\mathbf{u}}^{-1}$ can be localized using the matrix L such as

$$\Sigma_{\mathbf{u},\mathrm{loc}}^{-1} = \Sigma_{\mathbf{u}}^{-1} \circ L. \tag{4.17}$$

$$\Sigma_{\mathbf{u},\mathrm{loc}}^{-1} = \Sigma_{\mathbf{u}}^{-1} \circ L, \tag{4.18}$$

$$\widehat{B^TB}_{\mathrm{loc}}^{-1} = \widehat{B^TB}^{-1} \circ L. \tag{4.19}$$

In our example, the condition number $\kappa_2\left(\left(\left(B^TB\right)^{(2)}\right)_{\mathrm{loc}}^{-1}\right)$ declines to $3.5474 \cdot 10^8$ after applying this methodology and this selected localization matrix L.

4.5.3.3. Conjugate Gradients

With notation (4.14)–(4.15), the problem (4.5)–(4.6) can be generally written as a set of linear equations,

$$\Sigma_{\mathbf{u}}\mathbf{c} = \boldsymbol{\mu}_{\mathbf{u}}, \tag{4.20}$$

which is a well studied problem in the literature [69] even in the case of low stability of the matrix $\Sigma_{\mathbf{u}}$. Specifically, we adopt the version of the conjugate gradients (CG) method from MATLAB 2008a. The advantage is that the maximum number of iterations as well as tolerance precision of the method can be specified. Initial guess of $\mu_{\mathbf{u}}$ can be chosen from the previous VB iteration to speedup the computation.

Specifically, the tolerance (residuum, RES_{CG}) of the CG is computed according to the estimated of precision parameter ω as

$$\mathrm{RES}_{CG} = n\omega^{-1}, \tag{4.21}$$

see [101] and Section 4.5.2 for more details. This sensitivity is the key advantage over previous approaches since it allows to reflect the level of noise of the data.

Other numerical methods such as the generalized minimal residual method (GMRES) or biconjugate gradient stabilized method (BiCGSTAB) can be used; however, we do not observe any significant differences in the results from the CG method.

4.5.3.4. Evaluation of Solutions

The example results of the S-BSS-vecDC algorithm versions on synthetic dataset [108] is given in Figure 4.5. The simulated number of sources was 3 while the expected number of sources was set to $r_{max} = 4$ in order to simulate real conditions. The only modification of the used dataset (compared with [108]) is that the delay in the simulated input function is simulated, $b_1 = 0.3$, to study the numerical effect in the moment $\widehat{B^T B}^{-1}$ studied in this Section.

The results of different versions of the S-BSS-vecDC algorithm with different method of evaluation of the unstable inversions as well as the result from the S-BSS-DC algorithm (with conditionally independent convolution kernels, Remark 7) are given in Figure 4.5. None of the algorithms was able to correctly estimate the number of sources. The Figure 4.5 also displays the corresponding mean square errors (MSE) from the simulated data. The MSE is computed according to

$$\mathrm{MSE} = \sum_{k=1}^{r} \frac{1}{n} \sum_{j=1}^{n} \left(\hat{x}_{k,j} - x_{k,j}^{gt}\right)^2, \tag{4.22}$$

where $\hat{x}_{k,j}$ denotes elements of the kth estimated TAC and $x_{k,j}^{gt}$ denotes the kth simulated TAC. Note that by this dataset, the best and comparable results are provided by the S-BSS-vecDC algorithm with the covariance localization and with the conjugate gradients method. However, the chosen solution of

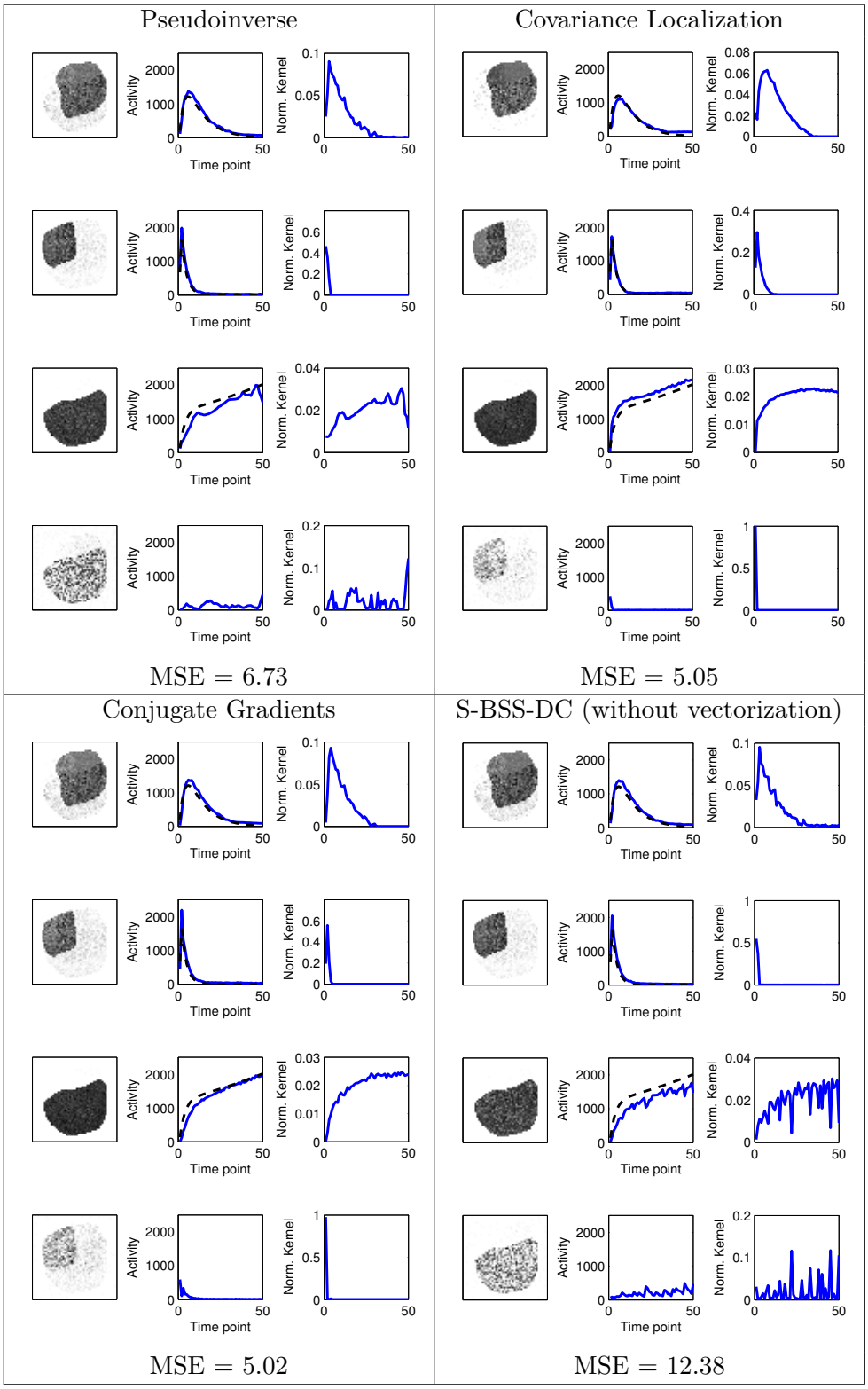

Figure 4.5.: The results of different versions of S-BSS-vecDC algorithms and S-BSS-DC algorithm, full blue lines, on synthetic dataset accompanied with computed MSE from generated data, dashed black lines.

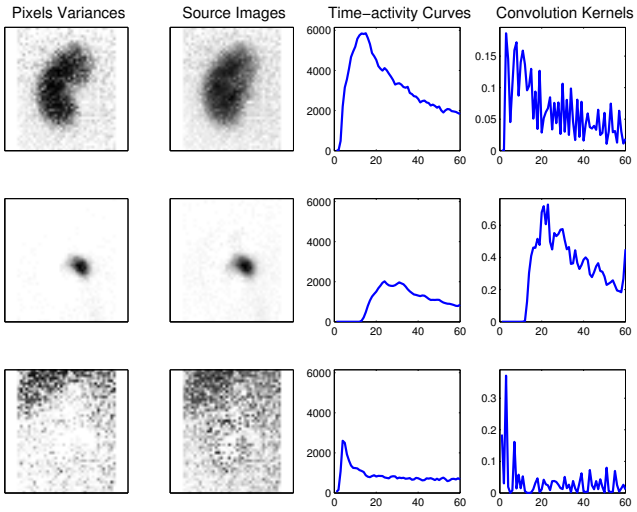

Figure 4.6.: Example results from the S-BSS-vecDC algorithm with conjugate gradients modification on selected sequence from dynamic renal scintigraphy.

numerical issue with the moment $\widehat{B^T B}^{-1}$ will be conjugate gradients method since its performance is based on estimates of the precision parameter ω which is more universal and flexible, especially on real datasets.

As an example, the same data as in Figure 4.3 were used to demonstrate the performance of the S-BSS-vecDC algorithm with conjugate gradients modification. The results are in Figure 4.6 where significantly smoother convolution kernels are obtained.

Remark 8. In addition, the performance of the simplified S-BSS-DC method is studied, Figure 4.5, bottom right. The result of the algorithm is not satisfactory since it does not take into account the correlation of the sources and tends to split the strongest source into different sources.

4.6. Other Algorithms for Blind Source Separation

4.6.1. Successive Nonnegative Projection Algorithm

Successive non-negative projection algorithm (SNPA) [40] for robust non-negative blind source separation provides near-separable non-negative matrix factorization. The non-negative matrix $D \in \mathbf{R}^{p \times n}$ is near-separable if exists index set $\mathcal{K}$ of the size r and non-negative matrix X of the size $n \times r$ such as

$$D \approx D\left(:, \mathcal{K}\right) X^T + E, \tag{4.23}$$

where $D(:, \mathcal{K})$ denotes matrix composed from $\mathcal{K}$ columns of the matrix D and E is the noise matrix of the same size as the matrix D. Robustness of the algorithm [40] is proven for any sufficiently small noise. The preselected number of sources r is expected in the algorithm.

We conjecture that the assumption of clear source images in the sequence is the main limitation of this algorithm since this condition is rarely met in dynamic medical imaging analysis.

4.6.2. Non-negative Matrix Factorization

Non-negative matrix factorization (NMF) [68, 52, 17] solves the following problem (in notation of [68]): given the data in the form of a non-negative matrix $D \in \mathbf{R}^{p \times n}$, find non-negative matrices A and X approximating the data V as

$$D \approx AX^T, \tag{4.24}$$

so that factors $A \in \mathbf{R}^{p \times r}$ and $X \in \mathbf{R}^{n \times r}$ are also non-negative.

The algorithm for NMF [68] is based on iterative calculation of matrices A and X using the multiplying the current values by a factor depending on quality of approximation (4.24). This quality depends on a chosen cost function. We consider the version of the NMF algorithm [17] with the Euclidean distance between the matrices D and AX^T defined as

$$||D - AX^T||_2 = \sum_{i,j} \left(D_{ij} - (AX^T)_{ij} \right)^2, \tag{4.25}$$

which is naturally bounded by zero and the convergence to a local optimum is guaranteed [68].

The number of sources r has to be preselected.

4.6.3. Convex Analysis of Mixtures - Compartment Modeling Algorithm

Interpretation of the superposed signal as a mixture model is used in the Convex analysis of mixtures - compartment modeling (CAM-CM) method [24, 23]. Here, the signal at the ith pixel and at time t, $d(i,t)$, of the tracer concentration is expressed as a non-negative linear combination of compartment-specific TACs $x_j(t)$ weighted by the relative tissue type proportions $A_j(i)$ at this pixel. Hence, pixel activity is described as

$$d(i,t) = x_1(t)A_1(i) + \cdots + x_J(t)A_J(i), \tag{4.26}$$

where J is the number of compartments [49]. A convex set can be defined using these parameters as

$$\mathcal{X} = \left\{ \sum_{j=1}^{J} \mathbf{x}_j A_j(i) \Big| A_j(i) \geq 0, \sum_{j=1}^{J} A_j(i) = 1, i = 1, \ldots, N \right\}, \tag{4.27}$$

where $\mathbf{x}_j$ is vector of $A_j(t)$ over time. It is shown that the corner points of the convex hull $\mathcal{H}(\mathcal{X})$ of pixel time-series correspond to pure-volume pixels for each tissue compartments [118]. Then, pharmacokinetic parameters are estimated using compartment modeling for pure-volume pixel time series. The TACs results from convolution between a common input function (tracer concentration in plasma) and exponential tissue-specific kernels. Although the study on estimation of number of sources is presented [24], the number of sources J has to be preselected.

4.7. Experiment with Synthetic Phantom Study

The performance of the described methods will be tested on a synthetic phantom sequence generated according to the convolution model (3.79); hence, $D = AU^TB^T + E$. The spatial resolution of the sequence is 50×50 and the number of simulated time points is 50. We simulate 3 sources, see Figure 4.7, top row, where source images and source TACs are given. The homogeneous Gaussian noise for this particular sequence is generated with standard deviation 0.3 of the signal strength. The number of sources is set as $r = 3$ for all tested algorithms.

The results of all mentioned methods are given in Figure 4.7, rows 2–7. The estimated source images are displayed in the first three columns and the estimated TACs are displayed in the second three columns as solid blue lines while the ground truth simulated sources are dashed red lines. Note that all estimated TACs are normalized between 0 and 1 in order to compare them with the ground truth data.

Algorithms BSS+ (the second row), FAROI (the third row), S-BSS-vecDC (the fifth row), and NMF (the seventh row) were able to estimate correct source images and correct TACs. Results of the BCMS algorithm (the forth row) suffer from application-specific model assumption of piece-wise linear convolution kernels with activity from the beginning which can be clearly seen by all estimated TACs from the BCMS algorithm. Results of the SNPA algorithm (the sixth row) suffer from the assumption of at least one clear image of each source in the sequence which is not valid here. Results of the CAM-CM algorithm (the eighth row) suffer from noise presented in estimated TACs while the source images are estimated correctly.

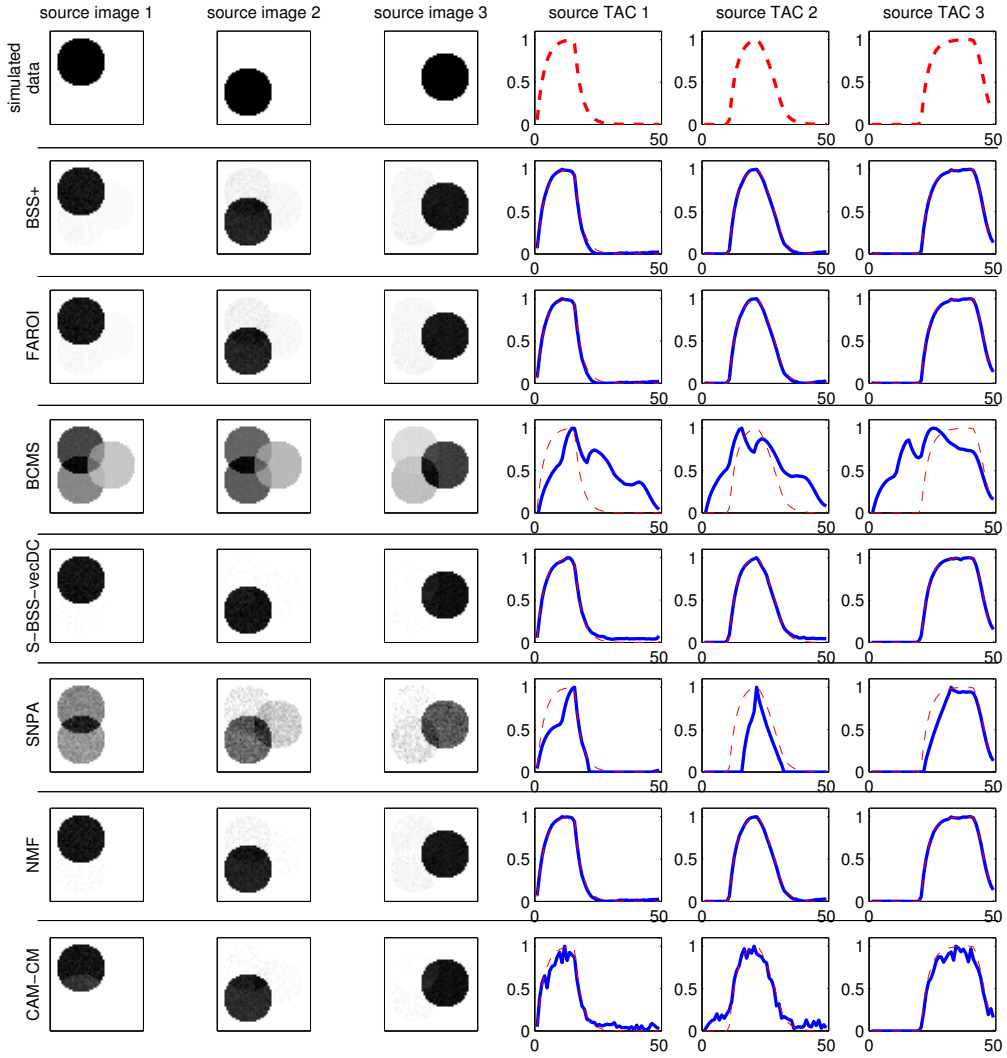

Figure 4.7.: Results of all described algorithms on a synthetic phantom study, rows 2–7, and the ground truth data, row 1. Source images are in the first three columns and related TACs are in the second three columns. The ground truth TACs are displayed using dashed red lines while the estimated TACs are displayed using solid blue lines.

5. Experiments with Dynamic Renal Scintigraphy

If prior opinions can differ from one researcher to the next, what happens to scientific objectivity in data analysis?

(Leonard Jimmie Savage)

Several methods for blind source separation of dynamic medical data were derived or described in Chapter 4, namely: BSS+ [76], FAROI [100], BCMS [113], S-BSS-vecDC [108], NMF [17], SNPA [40], and CAM-CM [24] algorithms. In this chapter, we will study performance of the algorithms on clinical data from dynamic renal scintigraphy.

Images in the real sequence may not fit into the assumptions of the model, due to movement of the patient or elastic deformation of the tissues. These phenomena can be mitigated by by preprocessing using image registration and motion correction [129, 55]. For the purpose of this work, we assume that the data were already preprocessed and the model of superposition, (1.1) is valid.

5.1. Dynamic Renal Scintigraphy

Dynamic renal scintigraphy [29, 10, 19] is a nuclear medicine method based on application of a radiopharmaceuticals into the body. The spatial distribution of the radiopharmaceuticals can be measured at given time-frames and a sequence of images is obtained, see Figure 5.1 as an example. The task is to separate the original sources (tissues) while they can overlap with other sources and the whole sequence is typically degraded by strong noise. The noise is Poisson distributed in renal scintigraphy; hence, the assumption of homogeneous noise (3.4) can be too restrictive. We perform scaling of the scintigraphy data D_{orig} using the correspondence analysis [10]:

$$d_{ij} = \frac{d_{ij,orig}}{\sqrt{\sum_{i=1}^{p} d_{ij,orig} \sum_{j=1}^{n} d_{ij,orig}}}. \tag{5.1}$$

This scaling transform the noise to be asymptotically independent normal distributed with identical variance modeled using ω^{-1} in (3.3). When this operation is performed, inverse scaling needs to be applied to the estimates of source images and source TACs for their presentation in the original scale. The scaling

(5.1) is only asymptotically optimal for Poisson noise, and may introduce bias for low count scenarios.

The anatomy of a physiological kidney is important for understanding the results. A healthy kidney is composed of parenchyma, a spongy tissue covering the whole kidney, and pelvis, a small structure serving to drain the urine from the kidney. Biologically, the parenchyma is activated directly from the blood while the pelvis is activated from the parenchyma with delay approximately $100 - 180$ seconds [32]. This delay is known as the uptake time.

The task of medical examination is to compute medical coefficients such as the differential renal function (DRF) which serve to diagnostics. For example, DRF [16, 87] is a percentage of function of the left kidney and the right kidney. The DRF is estimated from the sum of activity in the left (L) and in the right (R) parenchyma during the uptake time. Then, $DRF_L = \frac{L}{L+R} \times 100$ % and DRF_R can be computed analogically, both weighted by their time activity curves. Historically, the activity is taken only from the uptake time. Generally, these coefficients serve as quantification of a specific aspect of the sequence and are mainly used in investigation of diseases such as urinary obstruction, renal artery stenosis, renovascular hypertension [95], pelvi-uretric junction [85], renal transplantation etc.

In the following experiments, we will study the quality of estimation of TACs of sources as well as DRF estimation and theirs validation in comparison to the estimates obtained by an expert physician.

5.2. Qualitative Experiments with Selected Sequences

Here, we will show and discuss results from the algorithms on selected sequences. The sequences are chosen from database [1] where one typical study and two problematic ones are selected in order to demonstrate several aspect of the studied algorithms.

In all cases, we will not analyze the whole sequence but only selected regions of interest as shown in Figure 5.5, middle. This choice is motivated by clinical practice and suppresses the influence of sources such as heart, lungs, and urinary bladder. Each sequence consists of 180 images taken after each 10 seconds. The size of the regions of interest of the selected kidneys are 37×47 pixels.

5.2.1. Typical Study

The first sequence serves as a demonstration of typically observed source images and TACs as described in Section 5.1 as well as demonstration of results from all mentioned algorithms. The source sequence is shown in Figure 5.1 for illustration and the results are shown in Figure 5.2. The estimated source images are in the first four columns and the estimated TACs are in the second four columns while each row is related to the associated algorithm. The first column is recognized as the parenchyma, the second column as the pelvis, the

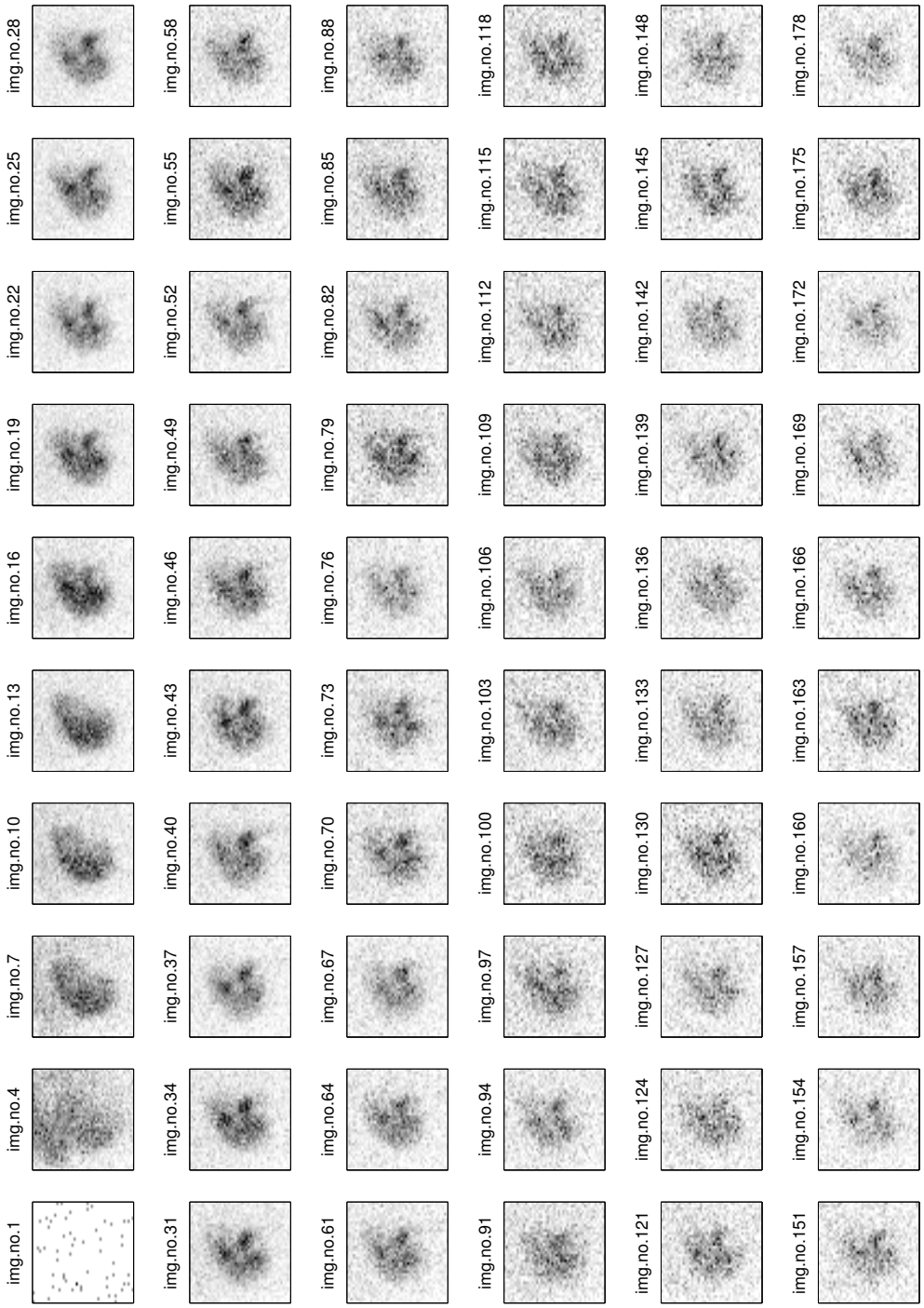

Figure 5.1.: Example sequence, every 3th image is shown.

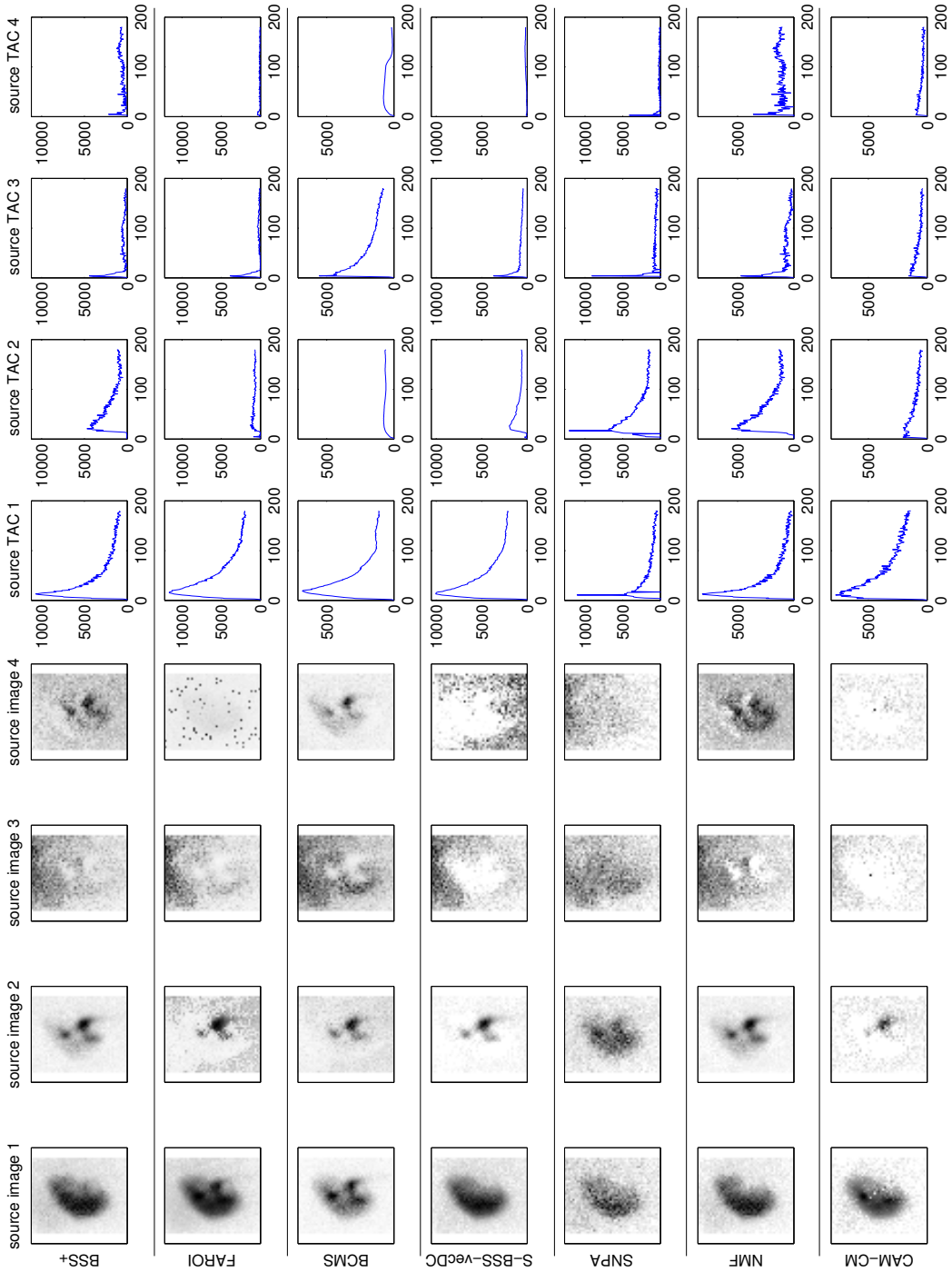

Figure 5.2.: Example separation of the sequence from Figure 5.1 using algorithms: BSS+, FAROI, BCMS, S-BSS-vecDC, SNPA, NMF, and CAM-CM.

third column as the tissue background, and the fourth column is either noise or another tissue background.

The main differences between the results of different algorithms can be observed in the first and the second sources, i.e. the parenchyma and pelvis. The BSS+, S-BSS-vecDC, NMF, and CAM-CM algorithms provided the expected source image and TAC of the parenchyma while the results of FAROI and BCMS algorithms suffers from mixed images of parenchyma and pelvis. Moreover, the expected zero activity at the beginning of the pelvis TAC, see Figure 5.2, the sixth column, is not respected in the case of BCMS due to the unrealistic piecewise linear model of the convolution kernels. The SNPA algorithm suffer from its assumption of presence of a clear source image in the sequence resulting in peaks and zeros in each TACs.

Overall, the results of separations using the BSS+, S-BSS-vecDC, and NMF algorithms can be taken as an examples of the desired separation of the scintigraphy sequence.

5.2.2. Study with Low Signal and Non-physiological Shapes

The second sequence is selected to show a more demanding case and ambiguity in its analysis. The results are shown in Figure 5.3 in the same layout as in Section 5.2.1.

First, note that the activity in this sequence is significantly lower than in the sequence in Figure 5.1. Second, the shapes of parenchyma and pelvis are non-typical and it is very difficult to distinguish between them for a non-trained person. One such clue could be the TACs of the estimated source; however, only the S-BSS-vecDC and NMF algorithms were able to separate the parenchyma and pelvis correctly which is demonstrated by the typical zero activity region at the beginning of the pelvis TAC (source TAC 2).

5.2.3. Study with Non-typical Separation Results

The third sequence is selected to show another possible ambiguity in separation and interpretation of the results. The results of separation are shown on Figure 5.4 for all tested algorithms.

The main issue with this sequence is that majority of the algorithms, all except CAM-CM, found the parenchyma divided into two structures: outer part, the first source, and the inner part, the third source. We conjecture that the outer part could be probably the blood activity in the kidney due to early signal activation in the related TAC, while the inner part could be nephrons and renal pyramids with slightly delayed activity than the blood. However, the automatic classification of this results is difficult and ambiguous since it does not fit in the expected form. These two structures could be considered as the substructures of the parenchyma and we could add them together to obtain whole parenchyma

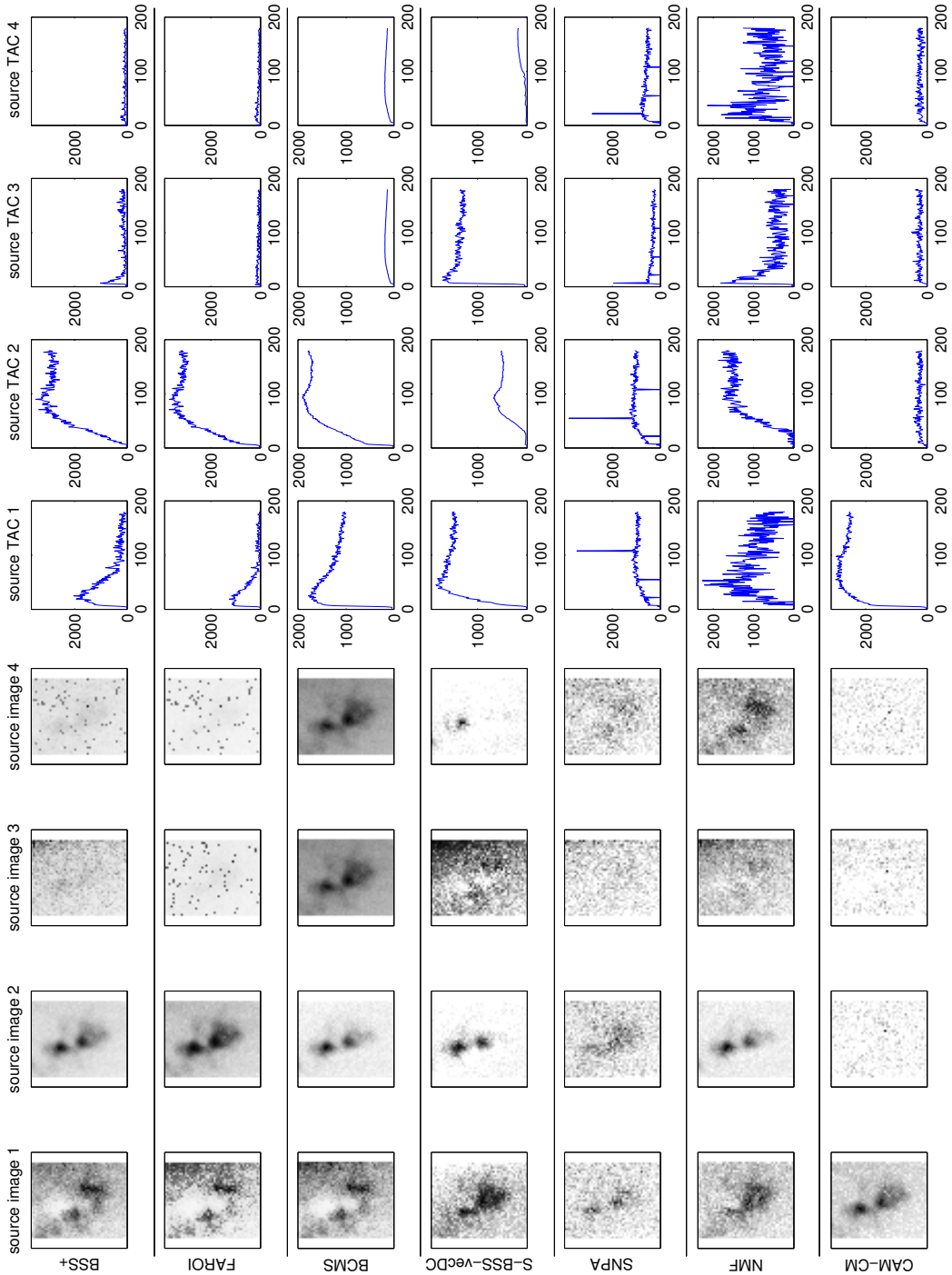

Figure 5.3.: Example of problematic separation of the selected sequence using algorithms: BSS+, FAROI, BCMS, S-BSS-vecDC, SNPA, NMF, and CAM-CM.

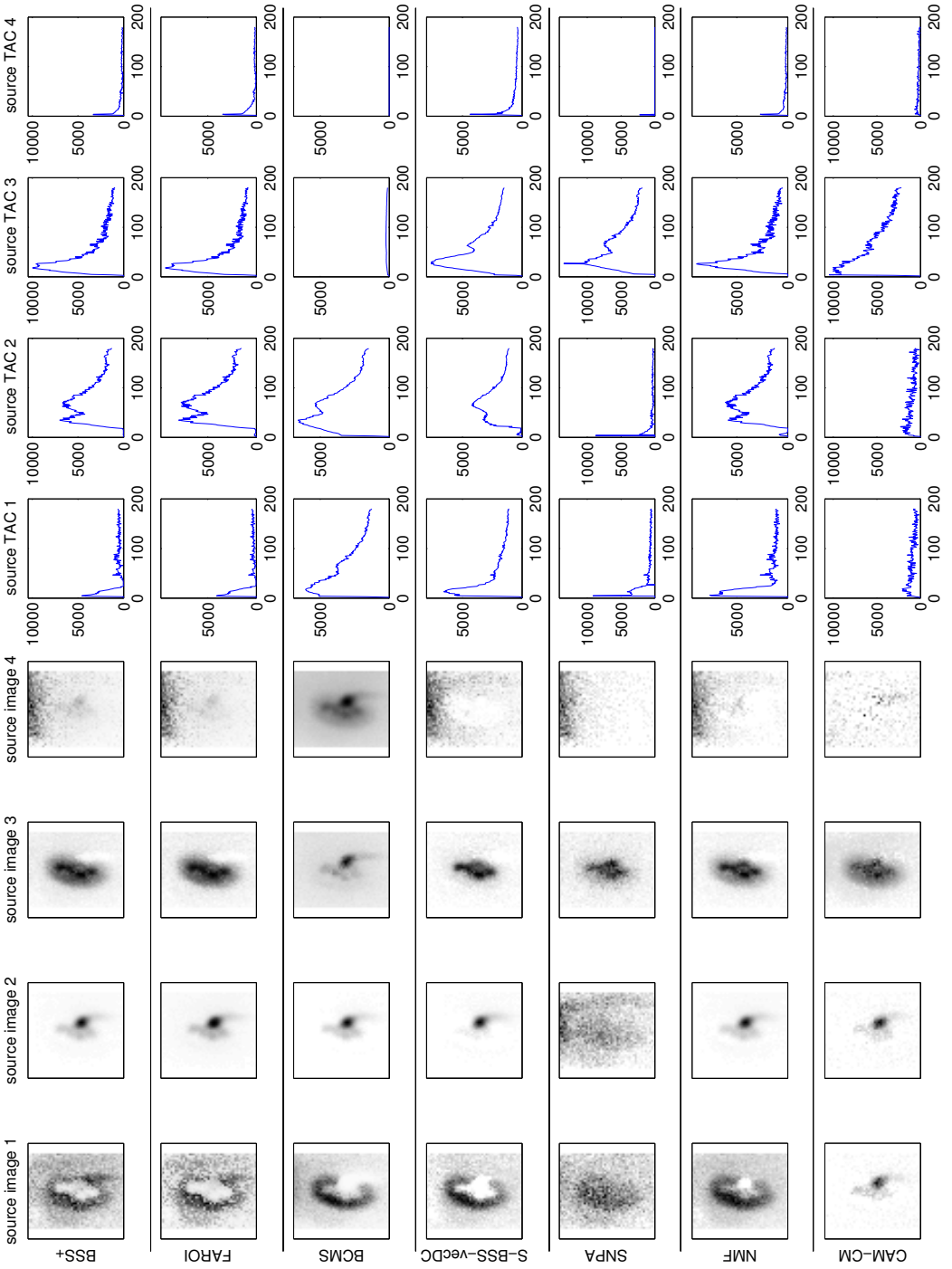

Figure 5.4.: Example of problematic separation of the selected sequence using algorithms: BSS+, FAROI, BCMS, S-BSS-vecDC, SNPA, NMF, and CAM-CM.

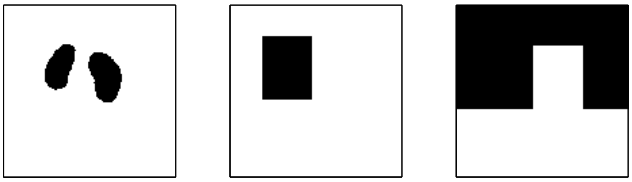

Figure 5.5.: Regions of interest separating left and right kidney. Left: anatomically motivated ROIs of the left and the right kidney. Middle: heuristic rectangular ROI of the left kidney. Right: ROI of the left kidney with excluded rectangular region of the right kidney.

or we could treat these structures separately. Further analysis of this type of results is still an open question.

5.2.4. Classification of Estimated Sources

For quantitative evaluation and comparison of the results presented, e.g., in Sections 5.2.1 – 5.2.3, we need to classify the results. This means to recognize the following structures: the heart or blood tissue, parenchyma, pelves, and urinary bladder. However, this is very difficult task because of huge variance in the shapes of their images and in their TACs as demonstrated in Figure 5.3 where non-typical shapes of parenchyma and pelvis are observed. In addition, location of a kidney can also vary a lot.

For automated classification and further evaluation, we proposed the following manual-based classification procedure for the heart and parenchyma, here refereed to as tissues:

1. we have manually drawn a region with the tissue and stored a mask with ones and zeros, see example in Figure 5.5, left,

2. for all separated sources of a given sequence, we compute the correlation between the mask and the source images,

3. the source image with the highest correlation is assumed to represent the corresponding tissue.

We are aware that this procedure in not flawless; however, it is the best way for automated analysis of large datasets to the best of our knowledge and we will use this procedure in the following quantitative evaluations.

5.3. Quantitative Evaluation of Large Datasets

In Section 5.2, we provided example results of the studied algorithms on selected sequences in order to demonstrate their performance and some issues of further

analysis of these results. In this section, we provide statistical evaluation of the results on large amount of data, on two datasets with 18 and 99 patients.

5.3.1. Dataset18

The Dataset18 consists of 18 sequences from dynamic renal scintigraphy. These data were chosen from a large set of anonymous data considered to be included into the database [1] by its author. The criterion of choice was clear visibility of dynamic structures. Each sequence has spatial resolution 128×128 pixels and the number of images vary from 100 to 180 taken with the sampling period of 10 seconds. In those data, we do not know the ground true solution in contrast to the synthetic phantom. Hence, we will validate the methods by comparing the automatically obtained results with those obtained by an experienced physician. While different physicians may reach different conclusion [18], we still consider this measure to be valid since the expert considers also anatomical knowledge and clinical experience.

This data are accompanied with the TACs of the left and right parenchyma and the heart extracted by an experienced physician. The ROIs of these tissues were defined manually on a computer screen with the best anatomical and physiological knowledge while recommended procedures were used for suppressing tissue and vascular backgrounds. Mutual overlap of all tissues is resolved iteratively following the best practice procedure [43]. These manually extracted TACs are, of course, not the ground truth; however, they are the best way how to evaluate the results of automated algorithms in clinical data in our opinion.

5.3.1.1. Experiment on Data with Manual Reference Curves

Since the TACs of the parenchyma and heart from the physician are available, we will compare them with the estimated TACs using methods described in Chapter 4. The task of our analysis is to recover the TAC of the parenchyma structure of each kidney while it has to be separated from the pelvis, vascular and tissue background. As a first step for automatic analysis, we need to separate the left and the right kidney regions and solve each region as an independent source separation problem. Heuristics has been designed for automatic selection of the rectangular region of the kidneys, Figure 5.5 middle. The data used for automatic analysis of the kidney of interest are composed from the full image with excluded rectangular region of the second kidney and urinary bladder, Figure 5.5 right. The advantage of this choice is that it preserves enough information about vascular and tissue background.

We use algorithms: BSS+, FAROI, BCMS, S-BSS-vecDC, NMF, and SNPA. We do not use the CAM-CM algorithm for these data since it has computational issues on such a large data. The algorithms will be applied to both kidneys, using ROI from Figure 5.5 right. Hence, 38 sequences will be analyzed. The estimates related to the medically relevant sources, i.e. the parenchyma (outer part of the

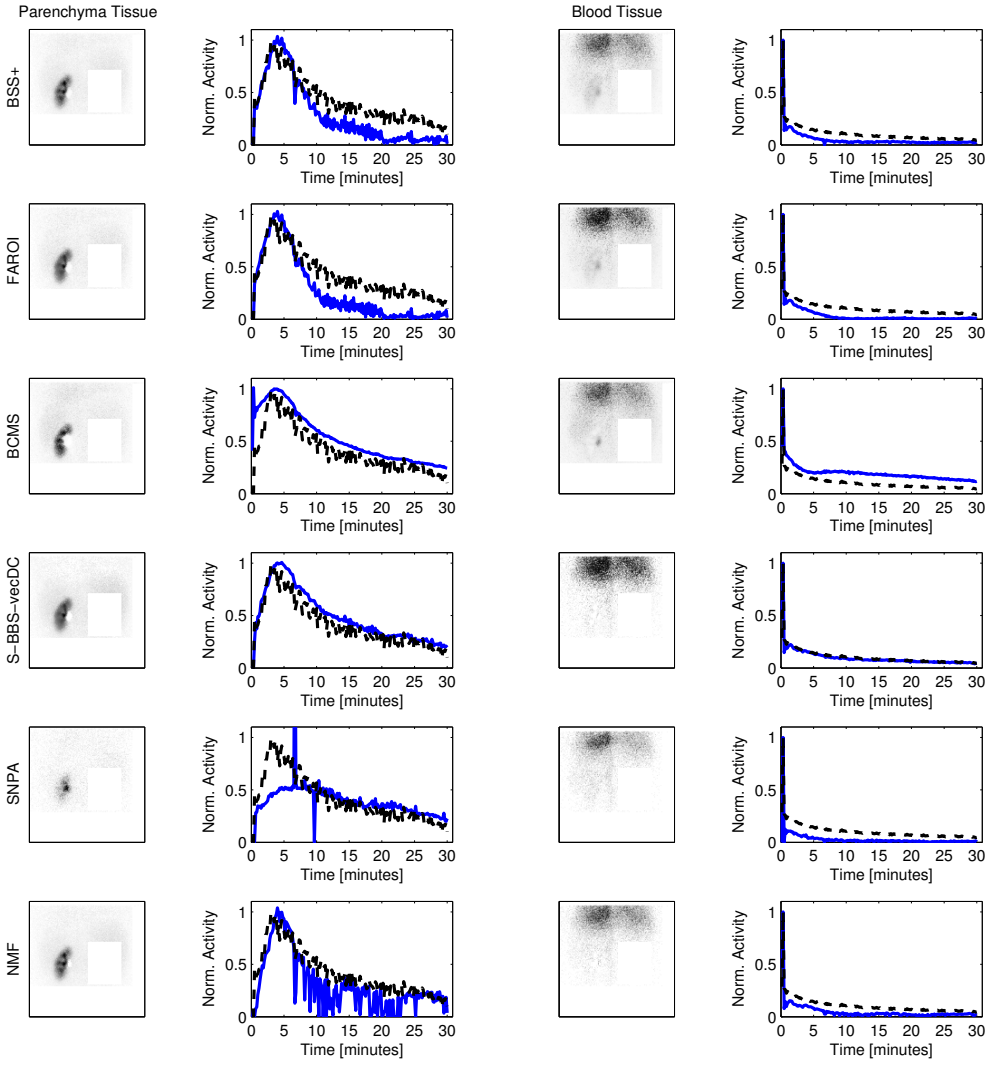

Figure 5.6.: Estimates of the sources of interest, parenchyma and blood tissues, are displayed for all tested algorithms. The solid blue line denotes the estimated TACs and the dashed black line denotes the TACs from the physician.

Table 5.1.: Statistical comparison of proximity of the estimated parenchyma TACs to those obtained by the expert physician for 38 data sets in terms of MSE and MAE.

	38 parenchyma curves			
algorithm	$\mu^{\text{par}}_{\text{MSE}} \pm \sigma^{\text{par}}_{\text{MSE}}$	$p^{\text{S-BSS-vecDC}}_{\text{MSE}}$	$\mu^{\text{par}}_{\text{MAE}} \pm \sigma^{\text{par}}_{\text{MAE}}$	$p^{\text{S-BSS-vecDC}}_{\text{MAE}}$
BSS+	0.0349±0.0271	0.0005	0.1225±0.0552	0.0005
FAROI	0.0426±0.0323	<0.0001	0.1382±0.0620	<0.0001
BCMS	0.0764±0.0930	0.0006	0.1799±0.1316	0.0002
S-BSS-vecDC	**0.0184**±0.0161	-	**0.0880**±0.0429	-
SNPA	0.0286±0.0207	0.0129	0.0998±0.0478	0.1975 (h=0)
NMF	0.0548±0.0421	<0.0001	0.1675±0.0685	<0.0001
CAM-CM	N/A	N/A	N/A	N/A

Table 5.2.: Statistical comparison of proximity of the estimated blood TACs to those obtained by the expert physician for 38 data sets in terms of MSE and MAE.

	38 heart curves			
algorithm	$\mu^{\text{blood}}_{\text{MSE}} \pm \sigma^{\text{blood}}_{\text{MSE}}$	$p^{\text{S-BSS-vecDC}}_{\text{MSE}}$	$\mu^{\text{blood}}_{\text{MAE}} \pm \sigma^{\text{blood}}_{\text{MAE}}$	$p^{\text{S-BSS-vecDC}}_{\text{MAE}}$
BSS+	0.0120±0.0146	0.0082	0.0765±0.0394	0.0025
FAROI	0.0144±0.0147	<0.0001	0.0870±0.0381	<0.0001
BCMS	0.0169±0.0396	0.2653 (h=0)	0.0731±0.0794	0.4911 (h=0)
S-BSS-vecDC	**0.0090**±0.0126	-	**0.0626**±0.0336	-
SNPA	0.0156±0.0163	<0.0001	0.0878±0.0374	<0.0001
NMF	0.0146±0.0102	0.0061	0.0907±0.0332	<0.0001
CAM-CM	N/A	N/A	N/A	N/A

kidney) and to the heart tissue, will be studied since these were provided by the physician. The physician provided only the TACs of these tissues, without the tissue images. Therefore, we compare only the shapes of the TAC and not their scale. Comparison of the scales will be done in Section 5.3.2.1. In this experiment, we scale each estimated TAC of match the peak of the expert chosen TAC and compute statistics of their deviations. The maximum number of tissue is set to 5, i.e. $r_{max} = 5$, in order to ensure the same conditions for each method. The parenchyma and the heart tissues are automatically selected using correlation of the estimated tissue images with the manually obtained ROIs of parenchyma and heart, Section 5.2.4.

Results of all tested algorithms for a particular sequence are displayed in Figure 5.6. The parenchyma images are in the first column, the parenchyma TACs are in the second column, the blood tissue images are in the third column, and the blood tissue TACs are in the fourth column. The solid blue line denotes the estimated TACs and the dashed black line denotes the TACs from the physician.

Statistical comparison of all 38 kidneys is given in Table 5.1 where mean MSE, denoted as μ_{MSE}, with standard deviation, denoted as σ_{MSE}, are computed according to

$$\mu_{\mathrm{MSE}}^{\mathrm{par}} = \frac{1}{38}\sum_{l=1}^{38}\left(\frac{1}{n}\sum_{j=1}^{n}\left(\hat{x}_{j,par}^{(l)} - x_{j,par}^{(l),gt}\right)^2\right), \tag{5.2}$$

$$\sigma_{\mathrm{MSE}}^{\mathrm{par}} = \sqrt{\frac{1}{38-1}\sum_{l=1}^{38}\left(\mu_{\mathrm{MSE},l}^{\mathrm{par}} - \mu_{\mathrm{MSE}}^{\mathrm{par}}\right)^2} \tag{5.3}$$

where $\hat{\mathbf{x}}_{par}^{(l)}$ denotes the estimated parenchyma TAC for the lth sequence and $\mathbf{x}_{par}^{(l),gt}$ denotes the corresponding estimate from the physician. The $p_{\mathrm{MSE}}^{\text{S-BSS-vecDC}}$ denotes the p-value of the statistical paired sample two-tailed t-test of MSEs from the S-BSS-vecDC method and from other methods, demonstrating that the improvement of the S-BSS-vecDC is statistically significant. Statistics for the blood tissue and the corresponding mean absolute error (MAE) are evaluated analogically and summarized in Table 5.2.

We conclude that the S-BSS-vecDC algorithm outperforms all other algorithms in proximity of shapes of the estimated TACs to those obtained by the physician for both tissues. The improvement was found statistically significant in most cases. The only statistically insignificant cases are:

1. MAE of SNPA algorithm in the case of parenchyma curves,
2. Both, MSE and MAE, of the BCMS algorithm in the case of heart curves (blood tissue).

However, the S-BSS-vecDC algorithm is found to be systematically better than any other tested algorithm.

5.3.2. Dataset99

The Dataset99 consists of 99 sequences from dynamic renal scintigraphy dataset publicly available on [1]. Originally, this dataset consists of 107 sequences; however, we excluded 8 sequences since they do not contain two kidneys. Each sequence has spatial resolution 128×128 pixels and the number of images is 180 taken with sampling period 10 seconds. For detailed clinical description see [1, 94].

For the purposes of evaluation, each sequence is accompanied by a manually computed differential renal function (DRF) [43] provided by an experienced physician using standard clinical procedures. Basically, the DRF is a percentual performance of each kidney, hence, it is a relative number. Once again, it is a subjective value since the resulting DRF may differ from physician to physician depending on the used technique and analyzing procedures. However, it allows us to quantitatively compare performance of all methods based on proximity of the estimated DRF to the values from the physician.

Table 5.3.: The number of results which differ from physician's results less than 3%, less than 5%, less than 10%, and more or equal than 10% computed for all 99 patients.

algorithm	<3%	<5%	<10%	≧10%
BSS+	38	57	78	21
FAROI	43	58	83	16
BCMS	42	59	82	17
S-BSS-vecDC	**46**	**68**	**86**	**13**
SNPA	25	41	72	27
NMF	40	52	81	18
CAM-CM	30	48	63	36

5.3.2.1. Differential Renal Function Estimation

As the first step for automatic analysis, we need to separate the left and the right kidney regions and solve each region as an independent source separation problem. Heuristics has been designed for automatic selection of the rectangular region of the kidneys, see Figure 5.5 middle.

The estimated TACs of parenchyma are used to obtain the DRF coefficient:

$$\mathrm{DRF}_L = \frac{P_l}{P_l + P_r} \times 100\%, \tag{5.4}$$

where DRF_L is DRF coefficient of the left kidney, P_l is the total activity of the left parenchyma, and P_r is the total activity of the right parenchyma. Total activity of the kth source is the sum $P_k = \sum_{j=1}^{n} \sum_{i=1}^{p} a_{i,k} x_{j,k}$.

The ROI in Figure 5.5, middle, ensures no influence of other tissue except those surrounding the kidney. The number of tissues is assumed to be 2, $r = 2$, since we expect separation of only the parenchyma and the background. All mentioned algorithms are capable to separate these sequences; hence, algorithms BSS+, FAROI, BCMS, S-BSS-vecDC, NMF, SNPA, and CAM-CM are used in this experiment.

For each sequence we compute the DRF using all competing methods and compare it with the reference value from the expert denoted as DRF_L^{gt}. The results are summarized in Table 5.3 using the number of tested sequences for which difference $|\mathrm{DRF}_L - \mathrm{DRF}_L^{gt}|$ is lower than 3%, lower than 5%, lower than 10%, and more or equal to 10%. Once again, the S-BSS-vecDC algorithm provides the closest results to those from the physician. Note that even for this method, the estimated DRF on 13 sequences differ from that of the physicians by more than 10%. This difference can be caused by: (i) great ambiguity of the estimation due to low signal to noise ratio which happens for seriously harmed kidney, (ii) background tissues may have TAC very similar to that of the parenchyma; in that case the blind source separation methods do not separate the background

and the activity of the parenchyma is overestimated, (iii) the estimate of the DRF from the expert may be biased.

5.3.2.2. Input Function Estimation

The proposed experiment with the Database99 is an extended version of that published in [113]. Two proposed algorithms, the BCMS and S-BSS-vecDC, are using convolution models and are capable to estimate the input function (IF) as their output. However, we do not have any ground truth data of the input function (IF), i.e. measurement of the tracer activity after its application, that could be used for comparison with the estimates of the algorithms. Hence, we proposed an indirect verification of the consistency of the IF estimates.

First, we compute the IFs for both left and right ROIs of kidneys separately. Note that we take only the first 50 images from each sequence. Second, we compare the estimated IFs. If the assumption of the common IF valid, the estimated IFs should be close to each other. Since we compare only shapes and not amplitudes of the estimated IFs, each IF will be normalized as follows

$$\mathbf{b}^l = \frac{\mathbf{b}^l}{\max\left(\left[\frac{\mathbf{b}^l}{\sum_{j=1}^{n} b_j^l}, \frac{\mathbf{b}^r}{\sum_{j=1}^{n} b_j^r}\right]\right)}. \tag{5.5}$$

This choice of normalization ensures independence on scale (with maximum 1) and comparability of all IF as well as between different algorithms. For each sequence and each algorithm, we compute an area of differences (AOD) defined as

$$\mathrm{AOD} = \sum_{j}^{n} \left| b_j^l - b_j^r \right|. \tag{5.6}$$

As an example, the estimated IFs for both algorithms are displayed in Figure 5.7. Here, the plot with IFs from the BCMS are in the top and the plot with IFs from the S-BSS-vecDC algorithm are in the bottom, both accompanied with the computed AOD for illustration. The solid blue lines are the estimated IFs from the ROI with the left kidney and the dashed black lines are the estimated IFs from the ROI with the right kidney.

The results from both algorithms are displayed in Figure 5.8 via cumulative histograms of the computed AODs. The Figure 5.8 could suggest that the S-BSS-vecDC algorithm outperforms the BCMS algorithm in proximity of left and right estimated IFs. However, we conjecture that estimates of the IF from the S-BSS-vecDC algorithm has limited potential in physiological interpretation of its IF estimates, see Figure 5.7, bottom, as an example. Most of its IF estimates are formed from shifted pulse with relatively low activity elsewhere since the model of convolution kernels in the S-BSS-vecDC algorithm is very unrestricted and the resulting TACs do not need common IF. On the other hand, the BCMS algorithm has a very restricted piece-wise model of the convolution kernels;

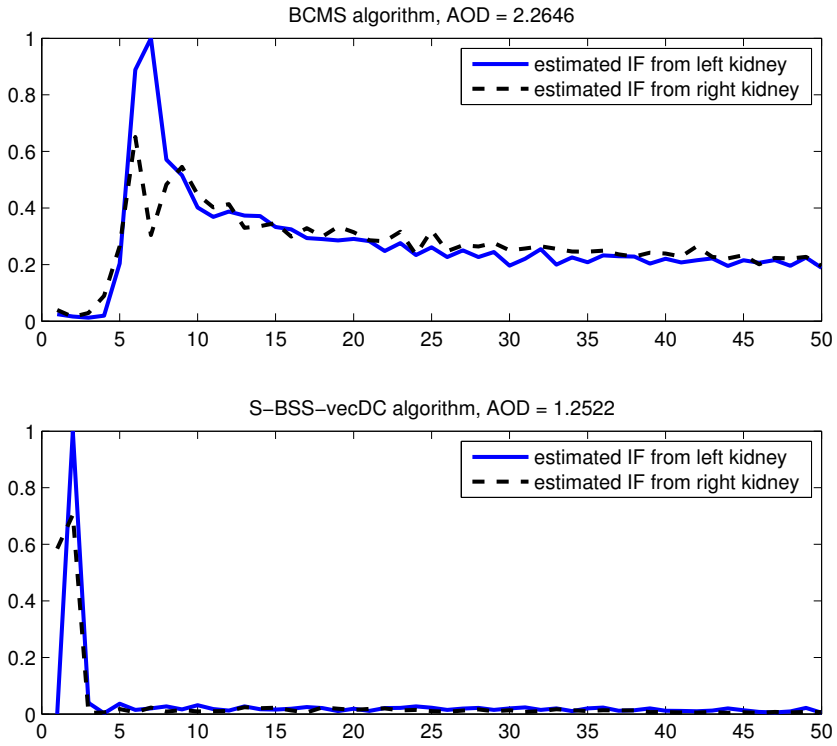

Figure 5.7.: Examples of estimated input functions from the BCMS algorithm, top, and the S-BSS-vecDC algorithm, bottom. The computed AOD is displayed in the title.

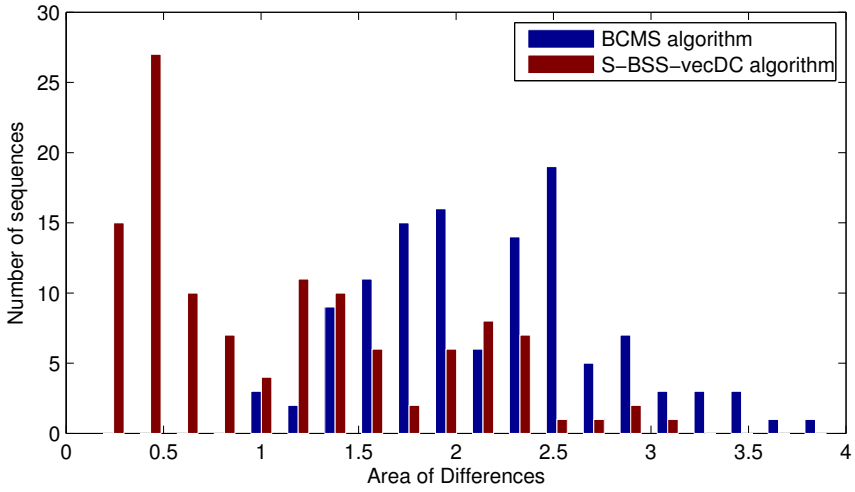

Figure 5.8.: The histogram with computed area of differences of the left and the right input function for the BCMS and S-BSS-vecDC algorithms.

hence, the shape of the TAC is much more important. Therefore, we consider the resulting estimates to be closer to the realistic input function.

5.3.2.3. Subjective Evaluation of Separation Quality

Although we do not have the ground truth source images and the related TACs, we can still try to evaluate the quality of separation in the same way as discussed in Section 5.2. We can set several conditions and subjectively evaluate if these conditions are met. We choose three main conditions indicating that the sequences are properly separated in the case of selected ROI (denoted here as ROI_0) of kidney, Figure 5.5, middle:

1. There have to be at least three main separated sources: parenchyma, pelvis, and tissue background.
2. The parenchyma image should be cleared from tissue and vascular backgrounds.
3. The pelvis tissue has a typical initial delay, its activity starts approximately near the peak of the parenchyma activity.

If all the conditions are met, the result of separation is marked as correct; if at least one is not met, the result of separation is considered as incorrect.

Another possible indication of correctness is non-sensitivity of the result to small variations in the selected ROI. For this experiment, we design 5 ROIs for each kidney. When the ROI_0 is given, Figure 5.5, middle, $ROI_1, \ldots, ROI_4$ arise from the ROI_0 as follows: the ROI_s arises from the ROI_0 using addition of s

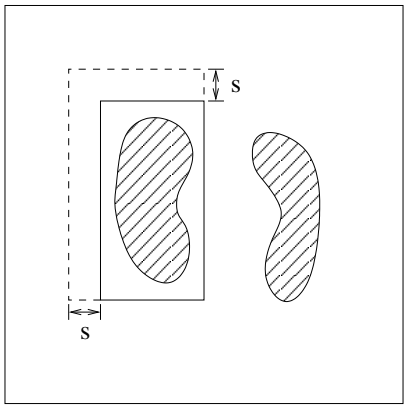

Figure 5.9.: Example of the ROI_0 of the left kidney, solid line, and its extention to ROI_s, dashed line.

Table 5.4.: The number of correctly separated sequences, N_{total}, and the number of sequences where correct separation were observed for all ROIs, N_{allROI}, are displayed for the BSS+, FAROI, and S-BSS-vecDC algorithms.

quality of separation		
algorithm	N_{total}(from 990)	N_{allROI}(from 198)
BSS+	426	68
FAROI	493	81
S-BSS-vecDC	**633**	**99**

pixels to the outer side and s pixels upwards, see Figure 5.9, dashed line. As a result, we obtain 198 sequences (99 sequences with 2 kidneys each) with 5 ROIs to be analyzed; hence, 990 sequences in total. This experiment was conducted only for the BSS+, FAROI, and S-BSS-vecDC algorithms.

The subjective evaluation of the quality of separation allow us to study the performance of algorithms in two ways: (i) the total number of correctly separated sequences from total number of 990 sequences, denoted as N_{total}, and (ii) the number of sequences where correct separation is observed for all ROIs, $ROI_0, \ldots, ROI_4$, from total 198 sequences, denoted as N_{allROI}.

The results are summarized in Table 5.4 using computed N_{total} and N_{allROI} for each algorithm. It is clearly seen that the S-BSS-vecDC algorithm outperform simpler models in the number of correct separation. Moreover, Figure 5.10 shows the histograms of sequences where $\{0, 1, 2, 3, 4, 5\}$ correct separations were observed for each algorithm. This show a tendency of the S-BSS-vecDC algorithm to correctly separate tissues more often than other algorithms.

We conclude that these results validate the assumptions of the S-BSS-vecDC algorithm (i.e. sparsity of source images and the convolution model of TACs)

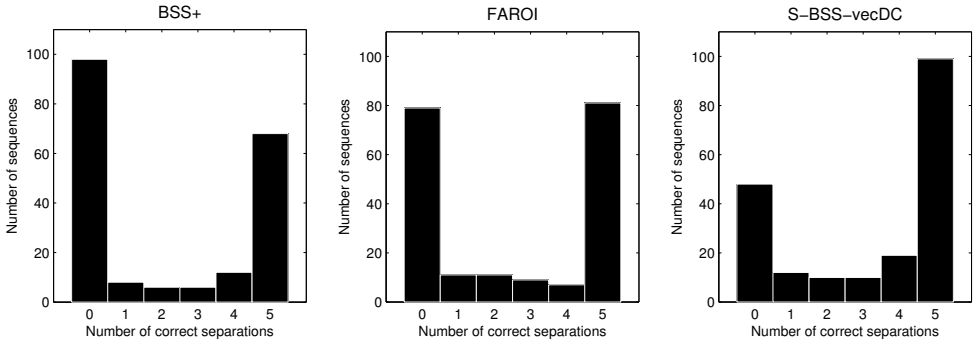

Figure 5.10.: Histograms with the number of sequences with the total number of correct separation 0 to 5 are displayed for the BSS+, FAROI, and S-BSS-vecDC algorithms.

on real data.

6. Experiments with Other Imaging Modalities

Far better an approximate answer to the right question, ...
than an exact answer to the wrong question.

(John Tukey)

The experiments conducted in Chapter 5 were possible due to large publicly available dataset and thanks to the expert provided solutions that were accompanied with the dataset. In other imaging modalities, there are no such dataset available as far as we know. However, we provide experiments on two studies, dynamic positron emission tomography and functional magnetic resonance imaging respectively. Here, the superposition or partial volume effect of underlaying structures is also observed and the proposed blind source separation algorithms can be directly applied on these data.

In addition, we conduct an experiment with hyper-spectral images in order to demonstrate the variability of the proposed models on data from other area but with the same data model.

6.1. Dynamic Positron Emission Tomography

Positron emission tomography (PET) [45] is another example where medical image sequences arise and can be described using superposition model (1.1).

In practice, the analysis of dynamic PET is often based on input function (IF), i.e. blood curve, knowledge [81]. This can be achieved using arterial blood sampling [44] which is very invasive and sensitive to errors. A number of methods has been proposed to lower invasiveness of measurement [26] or derive the IF directly from the dynamic PET images [127]. In some cases, blood structure can be directly observed on the images. In this cases, a region of interest (ROI) can be manually placed on vascular structures and its related time-activity curve (TAC) can be obtained [38]. Manual selection of the ROIs may suffer from subjectivity. This issues has been addressed using automatic clustering methods [70].

We will demonstrate the separation ability of the compared algorithms on a real brain data from dynamic PET [70]. In this study, ^{18}F-altanserin was applied to a patient and scanned with an 18-ring GE-Advance scanner (General Electric Medical System, Milwaukee, WI, USA) which is able to record 3D scans. Each

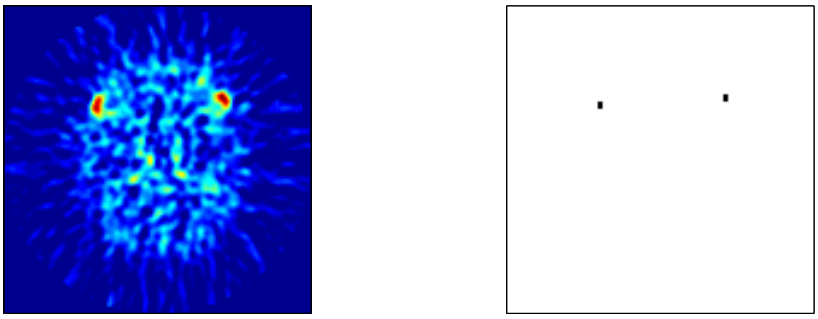

Figure 6.1.: The source image for slice number 9 is displayed on the left image. The manually selected ROIs of arterial veins are displayed on the right image.

scan consists of 35 image slices with an interslice distance 4.25 mm. The data were reconstructed into a sequence of $128 \times 128 \times 35$ voxel matrices, $2 \times 2 \times 4.25$ mm each voxel, using software provided by the manufacturer. The sequence consists of 40 voxel matrices, $n = 40$.

The aim of this experiment [107] is to compare the methodology for the BSS on data with easy-to-find TAC of the blood. We analyze data from dynamic PET of brain where structures with arterial blood are obvious and thus this data can be used as the simplest possible benchmark of separation methods. At first, we will study the separation on one selected slice. At second, we will study the separation performance for the whole volume.

6.1.1. Analysis of One Slice

For simplification and demonstration, we selected the 10th slice to be analyzed. For this slice, we placed the ROI of the arterial veins on the tested sequence, see Figure 6.1. We run all tested methods with preselected number of sources as $r = 4$ except CAM-CM algorithm which has computational issue with this data. From the results given in Figure 6.2, we selected the blood sources and compared them with the arterial TAC obtained manually. Note that the results from the BCMS algorithm are not displayed in Figure 6.2 since the assumptions of this algorithm do not hold for this data and the results are not satisfactory; nevertheless, we add the blood source to the comparison for illustration.

The comparison of the manually obtained blood TAC with TACs obtained using automatic method is given in Figure 6.3. It can be observed that the S-BSS-vecDC provides far closer blood TAC than other competing methods.

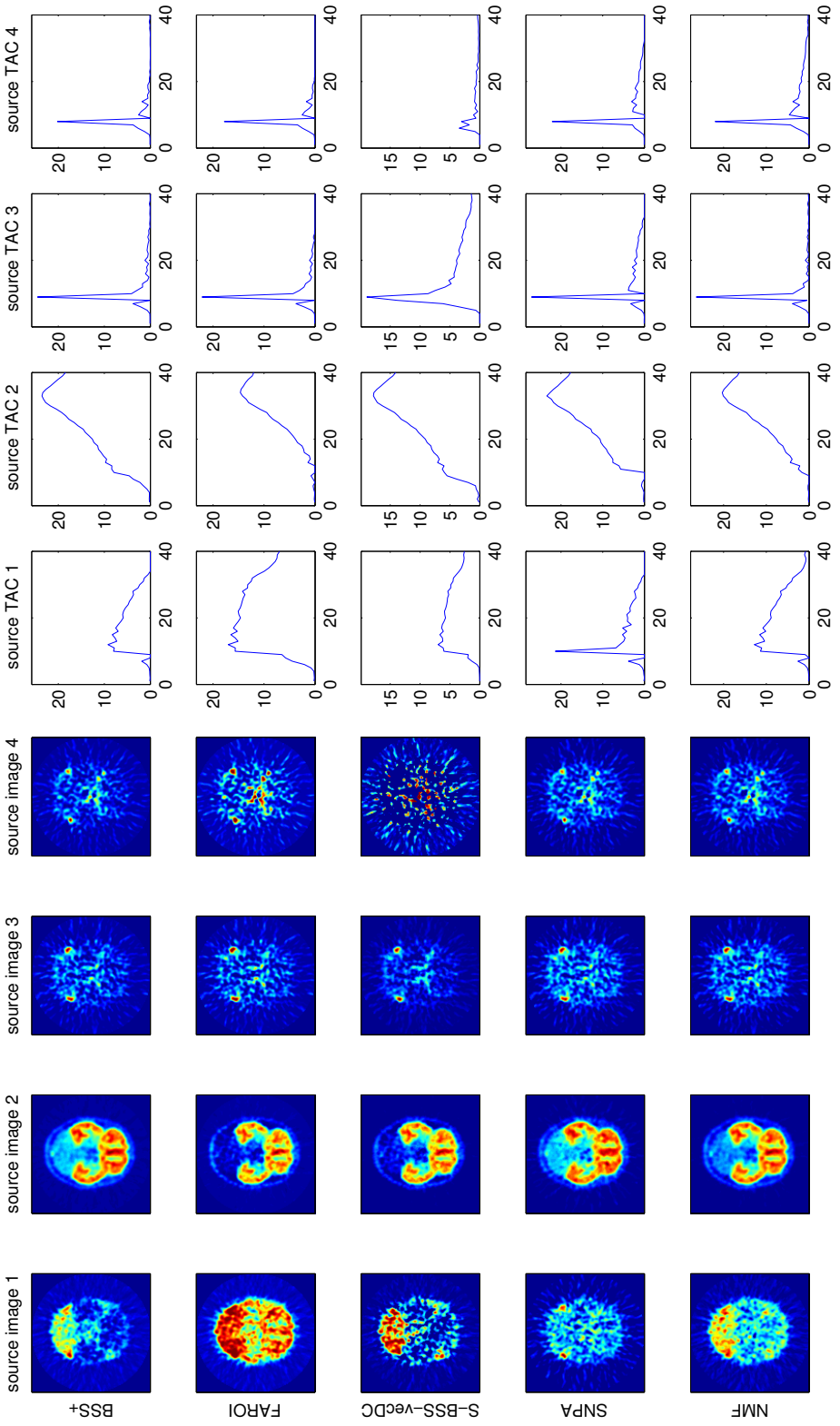

Figure 6.2.: The results of the BSS methods on the 10th slice of the PET dataset.

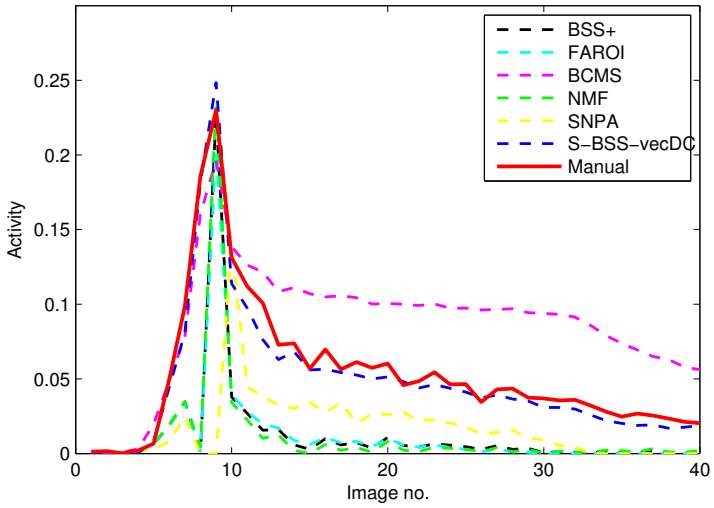

Figure 6.3.: TACs of blood tissues from all comparing methods on slice 10 are shown.

6.1.2. Analysis of Whole Volume

The goal of this experiment is to estimate input function from the whole measured volume. Again, we will not use the CAM-CM algorithm in this experiment since it has computational issue with this data. All remaining methods are now being compared.

Firstly, we created a manual ROI in several slices in the same sense as in Figure 6.1. We obtained a manually derived TAC of blood using this approach. This TAC will be used for comparison with the blood estimates from the compared BSS algorithms. Secondly, we ran the BSS algorithms on the whole volume and the blood curves were selected from each result. The estimates of the blood source are displayed in Figure 6.4. The curves are compared with manually obtained blood TAC, the red line.

All estimates of the blood tissue is similar to those obtained using only one slice, Figure 6.3, with the S-BSS-vecDC algorithm being the closest. However, all blind source separation algorithms provide slightly lower estimates of the blood TAC than the manual method which could be caused by additional background tissue activity in manually selected blood ROI.

6.2. Functional Magnetic Resonance Imaging

Functional magnetic resonance imaging (fMRI) data are the third type of data for testing of algorithms from Chapter 4. Here, the signal is based on blood-level-oxygenation contrast (BOLD) where fMRI detects local increases in blood

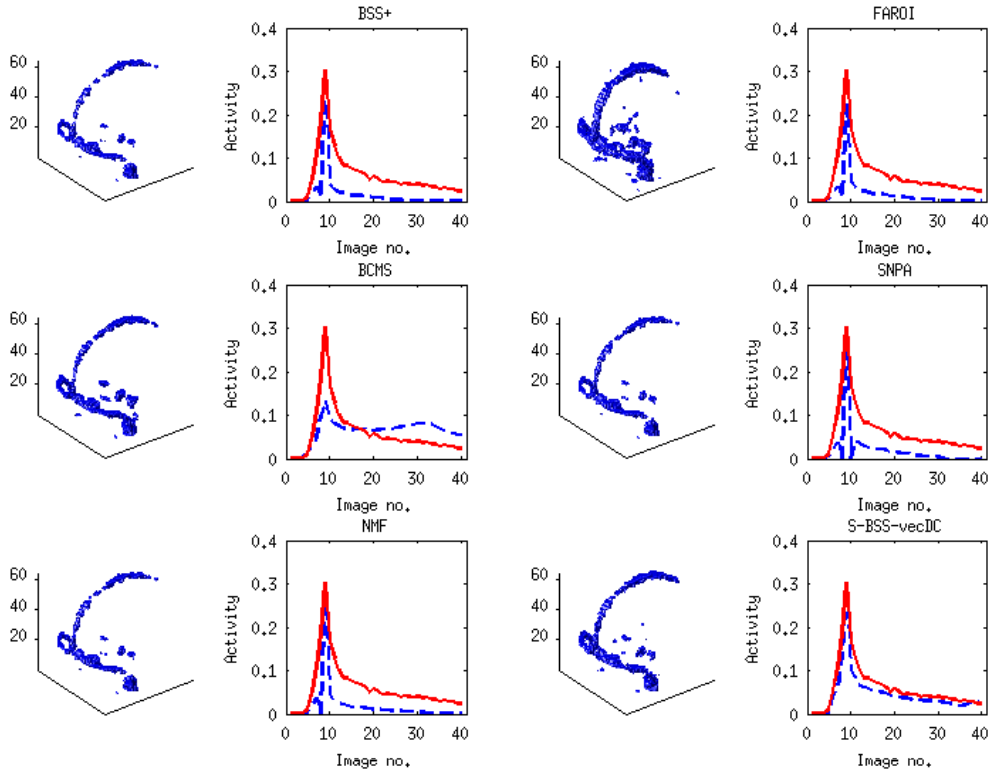

Figure 6.4.: Resulted blood source from all tested algorithms from analysis of the whole volume of the PET data.

oxygenation that is probably consequence of neurotransmitter action [72].

In this experiment, the data are provided by the Hvidovre Universitets Hospital, Denmark, available online [84]. A flickering checkerboard were shown in front of eyes of a human subject. Each flash of the checkerboard lasted for 50 frames of the sequence while one frame in the sequence is equal to 0.333 second. The data consist of 280 two-dimensional images of the brain covering the visual cortex with spatial resolution 78×57 pixels. The task of this experiment is to find the visual cortex and to study the activity of it and its relation to the activation of the checkerboard. The sampling period is fast; hence, a delay between the measured signal and the activity of the checkerboard is expected. The initialization of input function in case of all algorithms with convolution is slightly different from those in dynamic scintigraphy and dynamic PET. Here, we do not have any premise for input function shape; hence, we initialize it as a constant curve as

$$\mathbf{b}_{init} = \mathbf{1}_{280,1}. \tag{6.1}$$

The results are shown in Figure 6.5. The CAM-CM algorithm was not able to handle such a large dataset; hence, only results from algorithms BSS+, FAROI, BCMS, S-BSS-vecDC, SNPA, and NMF are provided in rows. Source images are in the first three columns and related TACs are in the second three columns. Here, the amplitude of the signal is unnecessary; hence, the resulting TACs are normalized. The blue lines are estimated normalized TACs while the red doted lines symbolically represent flashes of the checkerboard.

The results in Figure 6.5 suggest that the BSS+, FAROI, and S-BSS-vecDC algorithms are capable to clearly estimate the activity in visual cortex in accordance with the flashing of the checkerboard and with assumption of the delay in the measured signal. The BCMS and NMF algorithms estimate recognizable visual cortex; however, the related estimated TACs are very noisy degrading the possibility of evaluation. The SNPA algorithm was not able to estimate any significant signal in these data.

6.3. Analysis of Hyper-spectral Images

Hyper-spectral image arise when more than three channels of images is captured, e.g. in hyper-spectral digital imagery collection experiment (HYDICE) [90], where set of images of the same scenery is taken with different wavelengths. The task is to classify or recognize pixels with similar bands [88]. The problem can be considered as a dimensionality reduction and spectral classification. General algorithms such as independent component analysis (ICA) have been studied; however, suitability of these methods for hyperspectral imaging is limited since e.g. assumption of independence of sources in ICA is not valid for hyperspectral images [80]. Another issue of BSS methods is assuming that all pixels of a potential source contribute the to source activity equally [7] while the sparsity

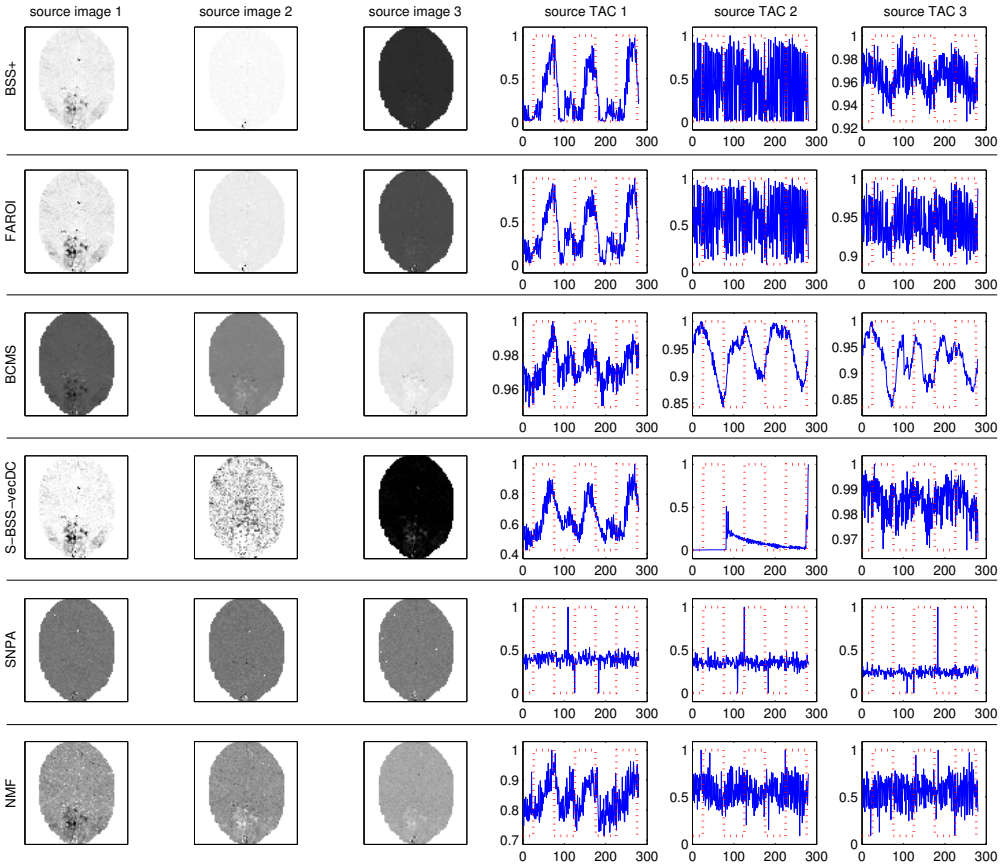

Figure 6.5.: Results of all tested algorithm on fMRI brain data with visual cortex activation. Source images are in the first three columns and related TACs are in the second three columns. The blue lines are estimated normalized activity while the red doted lines symbolically represent flashes of checkerboard.

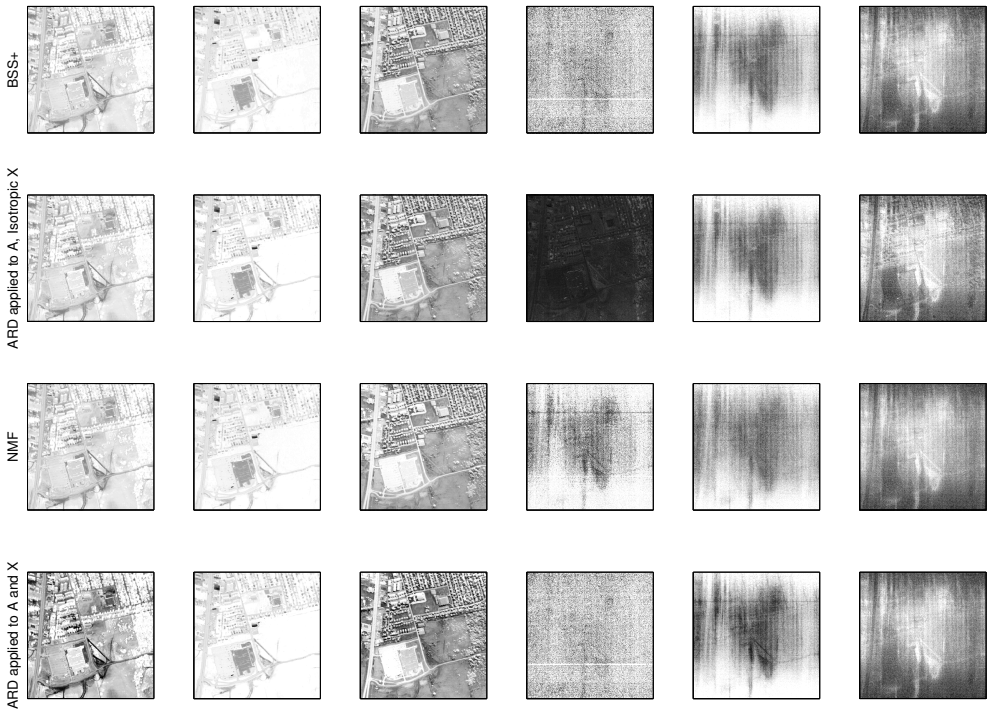

Figure 6.6.: Source images from hyperspectral image data estimated using compared methods (in rows).

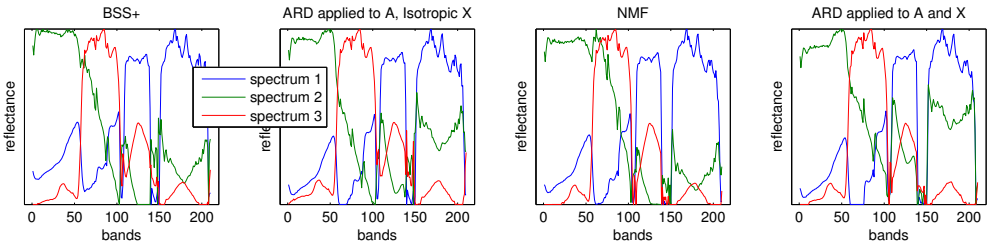

Figure 6.7.: 3 most significant estimated spectra from hyperspectral image data estimated using compared methods.

of each source image and source spectrum seems to be a natural assumption in hyperspectral image analysis [111].

The measured data are stored columnwise in the matrix D where each spectrum corresponds to one column. The decomposition (1.1) provides the matrix A with decomposed images in gray scale, i.e. intensities, and the matrix X with mean band activities for related images (reflectance or spectra). The example of hyper-spectral image separation are given in Figure 6.6 on Urban dataset[1] accompanied with estimated spectra in Figure 6.7. Here, the hyper-spectral image of the size 307×307 pixels with 210 bands is analyzed using 5 algorithms: (i) the BSS+ algorithm, Section 4.1, (ii) sparse model of images using ARD while the model of spectra remain isotropic, i.e. Sections 3.2.3 and 3.3.1, (iii) the NMF algorithm, Section 4.6.2, (iv) sparse model using ARD of both, images and spectral weights, i.e. Sections 3.2.3 and 3.3.2, (v) the SNPA algorithm, Section 4.6.1, however, SNPA performance is very poor with noisy data and the noisy bands need to be removed manually to obtain comparable results; hence, we removed these results from comparison. Convolution models are not tested since no convolution on the spectral bands is expected.

It can be seen that all algorithms distinguish between vegetation (the third column), metal object such as roofs (the second image), and roads or dirts (the first column). The remaining images represent the noise presented in original sequence and well separated here from meaningful results.

We conjecture that the best results are provided by the algorithm with sparse ARD priors on both, images and spectra. In estimated images, Figure 6.6, it reaches better contrast in comparison with other methods. In estimated spectra, Figure 6.7, the sparsity on weights are beneficial since the suppression of noisy observations, e.g. bands 104–109 or 139–151, are far better using methods with sparse priors. All methods are used directly without any parameters tunning.

[1] http://www.agc.army.mil/

7. Conclusion

I beseech you, in the bowels of Christ, think it possible you may be mistaken.

(Oliver Cromwell)

The aim of this thesis was to study and to extend methods of blind source separation for analysis of dynamic medical image sequences where the images arise as superposition of underlaying sources weighted by their time-activity curves. We have studied the general stochastic superposition model (Section 3.1) and existing prior models of its parameters. We proposed novel prior models for parameters representing both, the source images and the source time-activity curves. The variational Bayes method was used to derive posterior distributions for selected combinations of the prior models (Chapter 4).

Performance of the derived algorithms is heavily influenced by their initialization or numerical stability of internal steps. These issues were studied in detail to show that the derived algorithms are easy to use without any "tunning" parameters with the exception of the maximum number of sources.

We have tested all derived algorithms together with the state of the art algorithms (Section 4.6) on real data from dynamic medical imaging. Specifically, data from dynamic renal scintigraphy, dynamic brain positron emission tomography, and dynamic brain magnetic resonance imaging were used. We have shown that the proposed S-BSS-vecDC algorithm which is based on the use of sparse priors provides the best estimates from all tested algorithms.

MATLAB implementations of all derived algorithms are freely available for download.

7.1. Key Contributions of the Thesis

The key contributions of this thesis are summarized in the following list:

Chapter 2: Methods for approximation of Bayesian posterior distributions were reviewed with emphasis on the Variational Bayes approximation. In addition, we have shown that the automatic relevance determination (ARD) principle within the VB method impose sparse solution more aggressively than the classical VB solution in two cases, on unrestricted support and on positive support.

Chapter 3: After reviewing isotropic priors of the source images and time-activity curves (TACs) and priors for isotropic Gaussian noise within the blind source separation model, we proposed:

- sparse prior models of source images using a mixture model and using the ARD model, and sparse prior of TACs using the ARD model,
- deconvolution model of TACs with two prior models of convolution kernels: the piece-wise linear model and sparse model using the ARD principle.

Chapter 4: We have proposed several algorithms based on prior models from Chapter 3. We have used the advantage of the VB approximation where the posterior estimates of one parameter are computed using only posterior moments of the other parameters. This allows us

- to arbitrary combine the prior models of source images and TACs to obtain a specific algorithm.

In addition,

- we have studied various aspects common to all derived algorithms such as initialization, number of sources estimation, or numerical stability.

The behavior of the derived algorithms together with the state of the arts algorithms is studied on synthetic phantom study where the ground truth can be compared with their estimates.

Chapter 5: We applied the proposed and the state of the art algorithms on data from dynamic renal scintigraphy. Three selected sequences are studied to demonstrate typical issues of the analysis of medical image sequences. Key results are quantitative results on large dataset. Namely:

- comparison of estimated TACs of the heart and the parenchyma with TACs derived manually by physician is given for 19 sequences. The advantage of the proposed S-BSS-vecDC algorithm over other algorithms is proven with statistical significance,
- differential renal functions (DRF) are computed for 99 sequences. The results of the proposed S-BSS-vecDC algorithm are closest to those obtained by an expert physician.

In addition, we proposed evaluation of the estimated input function from convolution models and experiment with subjective evaluation of results.

Chapter 6: We have tested all algorithms on other problems where mixture of sources arise:

- dynamic positron emission tomography (PET) brain data and functional magnetic resonance imaging (fMRI) brain slices with visual cortex have been analyzed,
- hyperspectral image data have been analyzed,

with promising results.

7.2. Future Research

We summarize some of possible directions of future research and open questions.

7.2.1. Integration in Variational Bayes Approximation

The VB approximation was chosen as a compromise between accurate modeling and computational feasibility; however, it is not statistically optimal [96]. The VB approximation is based on factorization over latent variable and model parameters. One possibility of improvement in the VB approximation is to integrate the model parameters exactly, leaving only the latent variables for the VB approximation procedure. This technique is described for conjugate-exponential family models and called Latent-Space VB (LSVB) [103].

7.2.2. Prior Model Improvements

There are various possibilities in modeling improvements, we will mention three of them which we have in various states of completion.

7.2.2.1. Full Prior Model of Convolution Kernels

We conjecture that the ARD priors of the convolution kernels are so far the best choice in many cases in dynamic medical image data analysis. For this model, the matrix of TACs, X, is modeled as (3.113), $X = BU$, where the matrix B is composed from input function (3.77) and the matrix U stores convolution kernels in columns, or in a single vector $\mathbf{u} = \mathrm{vec}(U)$. In the ARD case, we model only the main diagonal of the covariance matrix, see (3.114)–(3.115). However, it is possible to model the full covariance matrix using Wishart distribution

$$f(\mathbf{u}|\Upsilon) = t\mathcal{N}_{\mathbf{u}}\left(\mathbf{0}_{nr,1}, \Upsilon^{-1}\right), \tag{7.1}$$

$$f(\Upsilon) = \mathcal{W}_{\Upsilon}\left(\alpha_0 I_{nr}, \beta_0\right), \tag{7.2}$$

where $\mathcal{W}(.,.)$ denotes the Wishart distribution of the covariance matrix Υ with prior parameters α_0, β_0, see Appendix A.5. We pioneered this approach in [110]

and we shown that this approach should has advantages in modeling of smoothness of convolution kernels, especially in connection with matrix localization [47].

Similar approach can be used directly for the matrix X [109].

7.2.2.2. Model of Input Function

The only influence on input function (IF) $\mathbf{b}$, defined in Section 3.3, is in initialization. Within the iterations, the estimate of IF can have any possible shape.

Depending on application, we could use another model for the IF. For example, a parametric model for IF such as exponential with possible delay can be used in scintigraphy, or bi-exponential [48] or sum of three gamma function [56] can be used in in modalities with fast sampling period.

7.2.2.3. Prior Image Knowledge Incorporation

In all considered models, we assumed zero mean value of the prior models of each pixel of each source image, i.e. all elements of the matrix A. However, in specific areas such as in dynamic renal scintigraphy, typical or average images of specific source could be observed. The advantage of probabilistic models is that we can use this average images for initialization or as a prior mean value. As an example, we extracted average images of left and right parenchyma, heart, and urinary bladder from results of the BSS+ algorithm on database of 99 scintigraphic sequences taken from [1]. We run the BSS+ algorithm for each sequence and manually label the source of interest. The labeled sources are averaged and the pixels with less than 35% of maximum activity of respected image is cropped to 0. The resulting source images are given in Figure 7.1.

Using this information as an initialization is straightforward. Incorporation of this information into the model cloud be as follows:

$$f\left(\mathbf{a}_k\right) = t\mathcal{N}_{\mathbf{a}_k}\left(\mathbf{p}_{\mathbf{a}_k}, I_p\right), \tag{7.3}$$

where $\mathbf{p}_{\mathbf{a}_k}$ is the kth prior source image stored in vector columnwise.

The same approach could be used for TACs as well.

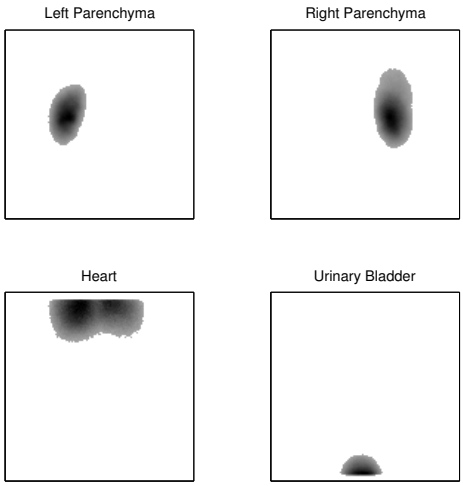

Figure 7.1.: Average images of left and right parenchyma, heart, and urinary bladder.

A. Required Probability Distribution

A.1. Normal Distribution

A.1.1. Multivariate Normal Distribution

Let the vector $\mathbf{x} \in \mathbf{R}^{n\times 1}$. Then, the multivariate normal distribution of the vector $\mathbf{x}$ with mean vector $\mu_{\mathbf{x}} \in \mathbf{R}^{n\times 1}$ and symmetric positive definite covariance matrix $\Sigma_{\mathbf{x}} \in \mathbf{R}^{n\times n}$ is

$$\mathcal{N}_{\mathbf{x}}(\boldsymbol{\mu}_{\mathbf{x}}, \Sigma_{\mathbf{x}}) = \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma_{\mathbf{x}}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}\left(\mathbf{x}-\mu_{\mathbf{x}}\right)^T \Sigma_{\mathbf{x}}^{-1}\left(\mathbf{x}-\mu_{\mathbf{x}}\right)\right). \tag{A.1}$$

The moments of the multivariate normal distribution (A.1) are

$$\widehat{\mathbf{x}} = \boldsymbol{\mu}_{\mathbf{x}}, \tag{A.2}$$

$$\widehat{\mathbf{x}\mathbf{x}^T} = \Sigma_{\mathbf{x}} + \boldsymbol{\mu}_{\mathbf{x}}\boldsymbol{\mu}_{\mathbf{x}}^T, \tag{A.3}$$

$$\widehat{\mathbf{x}^T\mathbf{x}} = \mathrm{tr}\left(\Sigma_{\mathbf{x}}\right) + \boldsymbol{\mu}_{\mathbf{x}}^T\boldsymbol{\mu}_{\mathbf{x}}. \tag{A.4}$$

For real matrix $C \in \mathbf{R}^{n\times n}$ holds

$$C\mathbf{x} \sim \mathcal{N}(C\mu_{\mathbf{x}}, C\Sigma_{\mathbf{x}}C^T). \tag{A.5}$$

A.1.2. Matrix Normal Distribution

Let the matrix $X \in \mathbf{R}^{n\times p}$. The matrix normal distribution of the matrix X is defined as

$$\begin{aligned}\mathcal{N}_X(\mu_X, \Sigma_n \otimes \Phi_p) = &\frac{1}{(2\pi)^{\frac{np}{2}}|\Sigma_n|^{\frac{p}{2}}|\Phi_p|^{\frac{n}{2}}} \times \\ &\times \exp\left(-\frac{1}{2}\mathrm{tr}\left[\Sigma_n^{-1}(X-\mu_X)(\Phi_p^{-1})^T(X-\mu_X)^T\right]\right),\end{aligned} \tag{A.6}$$

where matrices $\Sigma_n \in \mathbf{R}^{n\times n}$ and $\Phi_p \in \mathbf{R}^{p\times p}$ are symmetric positive definite matrices. The moments of normal matrix distribution (A.6) are

$$\widehat{X} = \mu_X, \tag{A.7}$$

$$\widehat{XX^T} = \mathrm{tr}(\Phi_p)\Sigma_n + \mu_X\mu_X^T, \tag{A.8}$$

$$\widehat{X^TX} = \mathrm{tr}(\Sigma_n)\Phi_p + \mu_X^T\mu_X. \tag{A.9}$$

For real matrices $C \in \mathbf{R}^{n\times n}$ and $D \in \mathbf{R}^{p\times p}$ holds

$$CXD \sim \mathcal{N}\left(C\mu_X D, C\Sigma_n C^T \otimes D^T \Phi_p D\right), \tag{A.10}$$

$$\widehat{X^T D X} = \mathrm{tr}(\Sigma_n D)\Phi_p + \mu_X^T D \mu_X, \tag{A.11}$$

$$\widehat{XCX^T} = \mathrm{tr}(C\Phi_p)\Sigma_n + \mu_X C \mu_X^T. \tag{A.12}$$

A.1.3. Truncated Scalar Normal Distribution

Truncated normal distribution, denoted as $t\mathcal{N}$, of a scalar variable x on interval $[a; b]$ is defined as

$$t\mathcal{N}(\mu, \sigma, [a, b]) = \frac{\sqrt{2}\exp\left(-\frac{1}{2\sigma}(x-\mu)^2\right)}{\sqrt{\pi\sigma}(erf(\beta) - erf(\alpha))}\chi_{[a,b]}(x), \tag{A.13}$$

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

The moments of truncated normal distribution are

$$\widehat{x} = \mu - \sqrt{\sigma}\frac{\sqrt{2}[\exp(-\beta^2) - \exp(-\alpha^2)]}{\sqrt{\pi}(\mathrm{erf}(\beta) - \mathrm{erf}(\alpha))}, \tag{A.14}$$

$$\widehat{x^2} = \sigma + \mu\widehat{x} - \sqrt{\sigma}\frac{\sqrt{2}[b\exp(-\beta^2) - a\exp(-\alpha^2)]}{\sqrt{\pi}(\mathrm{erf}(\beta) - \mathrm{erf}(\alpha))}. \tag{A.15}$$

A.1.4. Truncation in Matrix Normal Distribution

Assume the matrix $X \in \mathbf{R}^{n\times r}$ and matrix normal distribution $f(X) = \mathcal{N}_X(\mu_X, \Sigma_n \otimes \Phi_r)$ defined using equation (A.6). The truncation to given support $\mathcal{N}_X(\mu_X, \Sigma_n \otimes \Phi_r, [a, b])$ is computed according to the scalar truncated normal distribution as

$$f(X) = f(\mathbf{x}) \approx \prod_{l=1}^{np} t\mathcal{N}_{\mathbf{x}_l}\left(\mu_{\mathbf{x}_l}, \sigma_{\mathbf{x}_l}, [a, b]\right), \tag{A.16}$$

where $\mathbf{x} = \mathrm{vec}(X)$, $\boldsymbol{\mu}_{\mathbf{x}} = \mathrm{vec}(\mu_X)$ $\boldsymbol{\sigma}_{\mathbf{x}} = \mathrm{diag}(\Sigma_n \otimes \Phi_r)^{-1}$. The moments of the matrix normal distribution with truncated support are computed as

$$\widehat{\mathbf{x}} = \boldsymbol{\mu}_{\mathbf{x}} - \frac{1}{\sqrt{\boldsymbol{\sigma}_{\mathbf{x}}}} \circ \frac{\sqrt{2}[\exp(-\beta^2) - \exp(-\alpha^2)]}{\sqrt{\pi}(\mathrm{erf}(\beta) - \mathrm{erf}(\alpha))}, \tag{A.17}$$

$$\widehat{\mathbf{x} \circ \mathbf{x}} = \boldsymbol{\sigma}_{\mathbf{x}} + \boldsymbol{\mu}_{\mathbf{x}} \circ \widehat{\mathbf{x}} - \frac{1}{\sqrt{\boldsymbol{\sigma}_{\mathbf{x}}}} \circ \frac{\sqrt{2}[b\exp(-\beta^2) - a\exp(-\alpha^2)]}{\sqrt{\pi}(\mathrm{erf}(\beta) - \mathrm{erf}(\alpha))}, \tag{A.18}$$

where α, β, and function erf() are defined in Section A.1.3. The matrix-shape moments are

$$\widehat{X} = \left[\widehat{\mathbf{x}}_{1:n}, \ldots, \widehat{\mathbf{x}}_{(nr-n):(nr)}\right], \tag{A.19}$$

$$\widehat{X^T X} = \text{diag}\left(\left[\widehat{\mathbf{x} \circ \mathbf{x}}_{1:n}, \ldots, \widehat{\mathbf{x} \circ \mathbf{x}}_{(nr-n):(nr)}\right]^T \mathbf{1}_{n,1}\right) + (\mathbf{1}_{r,r} - I_r)\left(\widehat{X}^T \widehat{X}\right), \tag{A.20}$$

where subscript $i:j$ denotes selection of elements from i to j from respective vector.

The whole procedure is denoted using functions

$$\widehat{X} = \text{M}_1^{\text{tN}}\left(\mu_X, \Sigma_X, a, b\right), \tag{A.21}$$

$$\widehat{X^T X} = \text{M}_2^{\text{tN}}\left(\widehat{X}, \mu_X, \Sigma_X, a, b\right), \tag{A.22}$$

where $\Sigma_X = \Sigma_n \otimes \Phi_r$.

A.2. Gamma Distribution

The gamma distribution of a random scalar variable x is defined as

$$\mathcal{G}_x(\alpha, \beta) = \frac{1}{\Gamma(\alpha)} \frac{1}{\beta^{-\alpha}} x^{\alpha-1} e^{-x\beta} \tag{A.23}$$

for $x, \alpha, \beta > 0$ and $\Gamma(x) = \int_0^\infty t^{x-1} \exp(-t) \mathrm{d}t$ for $x > 0$.

Moments of the gamma distribution are given as

$$\widehat{x} = \frac{\alpha}{\beta}, \tag{A.24}$$

$$\widehat{x^2} = \frac{\alpha}{\beta^2}. \tag{A.25}$$

A.3. Truncated Exponential Distribution

The truncated exponential distribution of a random scalar variable x on the interval $(a, b]$ is defined as

$$\text{tExp}_x\left(\lambda, (a, b]\right) = \frac{1}{\exp(\lambda b) - \exp(\lambda a)} \exp(-\lambda x) \chi_{(a,b]}(x), \tag{A.26}$$

where $\chi_{(a,b]}(x)$ is characteristic function of the interval defined in Section A.1.3.

The moment of the truncated exponential distribution is

$$\widehat{x} = \frac{\exp(\lambda b)(1 - \lambda b) - \exp(\lambda a)(1 - \lambda a)}{\lambda(\exp(\lambda a) - \exp(\lambda b))}. \tag{A.27}$$

A.4. Uniform Distribution

The uniform distribution of a random scalar variable x on the interval (a, b) is defined as

$$\mathcal{U}_x\left((a,b)\right) = \frac{1}{b-a}\chi_{(a,b)}(x), \tag{A.28}$$

where $\chi_{(a,b]}(x)$ is characteristic function of the interval defined in Section A.1.3.

The moment of the truncated exponential distribution is

$$\widehat{x} = \frac{a+b}{2}. \tag{A.29}$$

A.5. Wishart Distribution

Wishart distribution $\mathcal{W}$ of the positive-definite matrix $X \in \mathbf{R}^{p\times p}$ is defined as

$$\mathcal{W}_p(\Sigma,\nu) = |X|^{\frac{\nu-p-1}{2}} 2^{-\frac{\nu p}{2}} |\Sigma|^{-\frac{\nu}{2}} \Gamma_p^{-1}\left(\frac{\nu}{2}\right) \exp\left(-\frac{1}{2}\mathrm{tr}\left(\Sigma^{-1}X\right)\right), \tag{A.30}$$

where $\Gamma_p\left(\frac{\nu}{2}\right)$ is the gamma function. The required moment is:

$$\widehat{X} = \nu\Sigma. \tag{A.31}$$

Bibliography

[1] Database of dynamic renal scintigraphy. http://www.dynamicrenalstudy.org. Accessed: 2014-13-12.

[2] J.Y. Ahn, D.S. Lee, J.S. Lee, S.K. Kim, G.J. Cheon, J.S. Yeo, S.A. Shin, J.K. Chung, and M.C. Lee. Quantification of regional myocardial blood flow using dynamic $H_2{}^{15}O$ pet and factor analysis. *Journal of Nuclear Medicine*, 42(5):782, 2001.

[3] M.F. Alf, M.T. Wyss, A. Buck, B. Weber, R. Schibli, and S.D. Krämer. Quantification of brain glucose metabolism by ^{18}F-FDG PET with real-time arterial and image-derived input function in mice. *Journal of Nuclear Medicine*, 54(1):132–138, 2013.

[4] T.W. Anderson. *An introduction to multivariate statistical analysis*, volume 3. Wiley New York, 2003.

[5] M.C.U. Araújo, T.C.B. Saldanha, R.K.H. Galvão, T. Yoneyama, H.C. Chame, and V. Visani. The successive projections algorithm for variable selection in spectroscopic multicomponent analysis. *Chemometrics and Intelligent Laboratory Systems*, 57(2):65–73, 2001.

[6] H. Attias and C.E. Schreiner. Blind source separation and deconvolution: the dynamic component analysis algorithm. *Neural Computation*, 10(6):1373–1424, 1998.

[7] N. Bali and A. Mohammad-Djafari. Bayesian approach with hidden markov modeling and mean field approximation for hyperspectral data analysis. *Image Processing, IEEE Transactions on*, 17(2):217–225, 2008.

[8] M.J. Beal. *Variational Algorithms for Approximate Bayesian Inference*. PhD thesis, University College London, 2003.

[9] M.J. Beal and Z. Ghahramani. The variational bayesian em algorithm for incomplete data: with application to scoring graphical model structures. *Bayesian statistics*, 7:453–464, 2003.

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

[11] H. Bergmann, E. Dworak, B. König, A. Mostbeck, and M. Šámal. Improved automatic separation of renal parenchyma and pelvis in dynamic renal scintigraphy using fuzzy regions of interest. *European Journal of Nuclear Medicine and Molecular Imaging*, 26(8):837–843, 1999.

[12] J.M. Bernardo. Expected information as expected utility. *The Annals of Statistics*, pages 686–690, 1979.

[13] J.M. Bernardo and A.F.M. Smith. *Bayesian theory*, volume 405. John Wiley & Sons, 2009.

[14] C.M. Bishop. Variational principal components. *IET Conference Proceedings*, pages 509–514(5), January 1999.

[15] C.M. Bishop and M.E. Tipping. Variational relevance vector machines. In *Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence*, pages 46–53, 2000.

[16] M.D. Blaufox, M. Aurell, B. Bubeck, E. Fommei, A. Piepsz, C. Russell, A. Taylor, H.S. Thomsen, D. Volterrani, et al. Report of the radionuclides in nephrourology committee on renal clearance. *Journal of nuclear medicine: official publication, Society of Nuclear Medicine*, 37(11):1883, 1996.

[17] B. Bodvarsson, L.K. Hansen, C. Svarer, and G. Knudsen. Nmf on positron emission tomography. In *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, volume 1, pages I–309. IEEE, 2007.

[18] A. Brink, M. Šámal, and M.D. Mann. The reproducibility of measurements of differential renal function in paediatric 99mtc-mag3 renography. *Nuclear medicine communications*, 33(8):824–831, 2012.

[19] I. Buvat, H. Benali, and R.D. Paola. Statistical distribution of factors and factor images in factor analysis of medical image sequences. *Physics in medicine and biology*, 43:1695, 1998.

[20] M. Caglar, G.K. Gedik, and E. Karabulut. Differential renal function estimation by dynamic renal scintigraphy: influence of background definition and radiopharmaceutical. *Nuclear medicine communications*, 29(11):1002, 2008.

[21] F. Calamante, M. Mørup, and L.K. Hansen. Defining a local arterial input function for perfusion mri using independent component analysis. *Magnetic resonance in Medicine*, 52(4):789–797, 2004.

[22] J.-F. Cardoso. Blind signal separation: statistical principles. *Proceedings of the IEEE*, 86(10):2009–2025, 1998.

[23] L. Chen, T.-H. Chan, P.L. Choyke, E.M.C. Hillman, Z.M. Bhujwalla, G. Wang, S.S. Wang, Z. Szabo, Y. Wang, et al. CAM-CM: a signal deconvolution tool for in vivo dynamic contrast-enhanced imaging of complex tissues. *Bioinformatics*, 27(18):2607–2609, 2011.

[24] L. Chen, P.L. Choyke, T.-H. Chan, C.-Y. Chi, G. Wang, and Y. Wang. Tissue-specific compartmental analysis for dynamic contrast-enhanced MR imaging of complex tumors. *Medical Imaging, IEEE Transactions on*, 30(12):2044–2058, 2011.

[25] S. Chen, C.A. Bouman, and M.J. Lowe. Clustered components analysis for functional mri. *Medical Imaging, IEEE Transactions on*, 23(1):85–98, 2004.

[26] E. Croteau, E. Poulin, S. Tremblay, V. Dumulon-Perreault, O. Sarrhini, M. Lepage, and R. Lecomte. Arterial input function sampling without surgery in rats for positron emission tomography molecular imaging. *Nuclear medicine communications*, 35(6):666–676, 2014.

[27] M. Davies. Identifiability issues in noisy ica. *IEEE Signal processing letters*, 11(5):470–473, 2004.

[28] A.P. Dempster, N.M. Laird, D.B. Rubin, et al. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39(1):1–38, 1977.

[29] R. Di Paola, J.P. Bazin, F. Aubry, A. Aurengo, F. Cavailloles, J.Y. Herry, and E. Kahn. Handling of dynamic sequences in nuclear medicine. *Nuclear Science, IEEE Transactions on*, 29(4):1310–1321, 1982.

[30] B.L. Diffey, F.M. Hall, and J.R. Corfield. The 99mTc-DTPA dynamic renal scan with deconvolution analysis. *Journal of Nuclear Medicine*, 17(5):352, 1976.

[31] Ch.H.Q. Ding, X. He, and H.D. Simon. On the equivalence of nonnegative matrix factorization and spectral clustering. In *SDM*, volume 5, pages 606–610. SIAM, 2005.

[32] E. Durand, M.D. Blaufox, K.E. Britton, O. Carlsen, P. Cosgriff, E. Fine, J. Fleming, C. Nimmon, A. Piepsz, A. Prigent, et al. International Scientific Committee of Radionuclides in Nephrourology (ISCORN) consensus on renal transit time measurements. In *Seminars in nuclear medicine*, volume 38, pages 82–102. Elsevier, 2008.

[33] J.S. Fleming and P.M. Kemp. A comparison of deconvolution and the Patlak-Rutland plot in renography analysis. *Journal of Nuclear Medicine*, 40(9):1503, 1999.

[34] M. Funaro, E. Oja, and H. Valpola. Independent component analysis for artefact separation in astrophysical images. *Neural networks*, 16(3):469–478, 2003.

[35] E.V. Garcia, R. Folks, S. Pak, and A. Taylor. Totally automatic definition of renal regions-of-interest from Tc-99m MAG3 renograms: Validation in patients with normal kidneys and in patients with suspected renal obstruction. *Nuclear medicine communications*, 31(5):366, 2010.

[36] G. Gaspari and S.E. Cohn. Construction of correlation functions in two and three dimensions. *Quarterly Journal of the Royal Meteorological Society*, 125(554):723–757, 1999.

[37] P. Gebouský, M. Kárný, and A. Quinn. Lymphoscintigraphy of upper limbs: A bayesian framework. *Bayesian Statistics*, 7:543–552, 2003.

[38] G. Germano, B.C. Chen, et al. Use of the abdominal aorta for arterial input function determination in hepatic and renal pet studies. *Journal of nuclear medicine*, 33(4):613, 1992.

[39] Z. Ghahramani and M.J. Beal. Variational Inference for Bayesian Mixtures of Factor Analysers. *Advances in Neural Information Processing Systems 12: Proceedings of the 1999 Conference*, 2000.

[40] N. Gillis. Successive nonnegative projection algorithm for robust nonnegative blind source separation. *SIAM Journal on Imaging Sciences*, 7(2):1420–1450, 2014.

[41] N. Gillis and S.A Vavasis. Fast and robust recursive algorithmsfor separable nonnegative matrix factorization. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 36(4):698–714, April 2014.

[42] G. Golub and W. Kahan. Calculating the singular values and pseudo-inverse of a matrix. *Journal of the Society for Industrial & Applied Mathematics, Series B: Numerical Analysis*, 2(2):205–224, 1965.

[43] I. Gordon, A. Piepsz, and R. Sixt. Guidelines for standard and diuretic renogram in children. *European journal of nuclear medicine and molecular imaging*, 38(6):1175–1188, 2011.

[44] H.N.J.M. Greuter, R. Boellaard, et al. Measurement of 18f-fdg concentrations in blood samples: comparison of direct calibration and standard solution methods. *Journal of nuclear medicine technology*, 31(4):206–209, 2003.

[45] R.N. Gunn, S.R. Gunn, and V.J. Cunningham. Positron emission tomography compartmental models. *Journal of Cerebral Blood Flow & Metabolism*, 21(6):635–652, 2001.

[46] H.R. Ham. Is renography suitable for deconvolution analysis? *Journal of Nuclear Medicine*, 37(2):403–404, 1996.

[47] T.M. Hamill, J.S. Whitaker, and Ch. Snyder. Distance-dependent filtering of background error covariance estimates in an ensemble kalman filter. *Monthly Weather Review*, 129(11):2776–2790, 2001.

[48] M. Heilmann, Ch. Walczak, J. Vautier, J.-L. Dimicoli, C.D. Thomas, M. Lupu, J. Mispelter, and A. Volk. Simultaneous dynamic T 1 and T 2* measurement for AIF assessment combined with DCE MRI in a mouse tumor model. *Magnetic Resonance Materials in Physics, Biology and Medicine*, 20(4):193–203, 2007.

[49] E.M.C. Hillman, A. Devor, M.B. Bouchard, A.K. Dunn, G.W. Krauss, J. Skoch, B.J. Bacskai, A.M. Dale, and D.A. Boas. Depth-resolved optical imaging and microscopy of vascular compartment dynamics during somatosensory stimulation. *Neuroimage*, 35(1):89–104, 2007.

[50] P.D. Hoff. *A first course in bayesian statistical methods*. Springer Verlag, 2009.

[51] H. Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417, 1933.

[52] P.O. Hoyer. Non-negative matrix factorization with sparseness constraints. *The Journal of Machine Learning Research*, 5:1457–1469, 2004.

[53] H. Iida, Ch.G. Rhodes, R. de Silva, L.I. Araujo, P.M. Bloomfield, A.A. Lammertsma, and T. Jones. Use of the left ventricular time-activity curve as a noninvasive input function in dynamic oxygen-15-water positron emission tomography. *Journal of Nuclear Medicine*, 33:1669–1677, 1992.

[54] H. Jeffreys. *The theory of probability*. Oxford University Press, 1998.

[55] M. Jenkinson, P. Bannister, M. Brady, and S. Smith. Improved optimization for the robust and accurate linear registration and motion correction of brain images. *Neuroimage*, 17(2):825–841, 2002.

[56] R. Jiřík, K. Souček, M. Mézl, M. Bartoš, E. Dražanová, F. Dráfi, L. Grossová, J. Kratochvíla, O. Macíček, K. Nylund, et al. Blind deconvolution in dynamic contrast-enhanced mri and ultrasound. In *Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE*, pages 4276–4279. IEEE, 2014.

[57] I. Jolliffe. *Principal component analysis*. Wiley Online Library, 2002.

[58] C.A. Joyce, I.F. Gorodnitsky, and M. Kutas. Automatic removal of eye movement and blink artifacts from EEG data using blind component separation. *Psychophysiology*, 41(2):313–325, 2004.

[59] Ch. Jutten and J. Herault. Blind separation of sources, part I: An adaptive algorithm based on neuromimetic architecture. *Signal processing*, 24(1):1–10, 1991.

[60] R.E. Kass and A.E. Raftery. Bayes Factors. *Journal of the American Statistical Association*, 90(430), 1995.

[61] H. Kim and H. Park. Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis. *Bioinformatics*, 23(12):1495–1502, 2007.

[62] R. Klein, R.S. Beanlands, A. Adler, and R. deKemp. Model-based factor analysis of dynamic sequences of cardiac positron emission tomography. In *Nuclear Science Symposium Conference Record, 2008. NSS'08. IEEE*, pages 5198–5202. IEEE, 2008.

[63] S. Kullback and R.A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22(1):79–86, 1951.

[64] A. Kuruc, W.J.H. Caldicott, and S. Treves. Improved Deconvolution Technique for the Calculation of Renal Retention Functions. *COMP. AND BIOMED. RES.*, 15(1):46–56, 1982.

[65] B. Lanz, C. Poitry-Yamate, and R. Gruetter. Image-derived input function from the vena cava for 18F-FDG PET studies in rats and mice. *Journal of Nuclear Medicine*, 55(8):1380–1388, 2014.

[66] R.S. Lawson. Application of mathematical methods in dynamic nuclear medicine studies. *Physics in medicine and biology*, 44:R57–R98, 1999.

[67] D.D. Lee and H.S. Seung. Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, pages 556–562, 2000.

[68] D.D. Lee and H.S. Seung. Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, pages 556–562, 2001.

[69] J. Liesen and Z. Strakoš. *Krylov subspace methods: principles and analysis*. Oxford University Press, 2012.

[70] M. Liptrot, K.H. Adams, L. Martiny, L.H. Pinborg, M.N. Lonsdale, N.V. Olsen, S. Holm, C. Svarer, and G.M. Knudsen. Cluster analysis in kinetic modelling of the brain: a noninvasive alternative to arterial sampling. *NeuroImage*, 21(2):483–493, 2004.

[71] D.J.C. MacKay. *Information theory, inference, and learning algorithms*. Cambridge University Press, 2003.

[72] P.M. Matthews and P. Jezzard. Functional magnetic resonance imaging. *Journal of Neurology, Neurosurgery & Psychiatry*, 75(1):6–12, 2004.

[73] N. Michoux, J.P. Vallee, A. Pechere-Bertschi, X. Montet, L. Buehler, and B.E. Van Beers. Analysis of contrast-enhanced MR images to assess renal function. *Magnetic Resonance Materials in Physics, Biology and Medicine*, 19(4):167–179, 2006.

[74] G.W. Middleton, W.H. Thomson, I.H. Davies, and A. Morgan. A multiple regression analysis for accurate background subtraction in 99Tcm-DTPA renography. *Nuclear Medicine Communications*, 10(5):315, 1989.

[75] T.P. Minka. Expectation propagation for approximate bayesian inference. In *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*, pages 362–369. Morgan Kaufmann Publishers Inc., 2001.

[76] J.W. Miskin. *Ensemble learning for independent component analysis*. PhD thesis, University of Cambridge, 2000.

[77] K. Mouridsen, S. Christensen, L. Gyldensted, and L. Østergaard. Automatic selection of arterial input function using cluster analysis. *Magnetic resonance in medicine*, 55(3):524–531, 2006.

[78] J.E.M. Mourik, M. Lubberink, U.M.H. Klumpers, E.F. Comans, A.A. Lammertsma, and R. Boellaard. Partial volume corrected image derived input functions for dynamic PET brain studies: Methodology and validation for [11 C] flumazenil. *Neuroimage*, 39(3):1041–1050, 2008.

[79] S. Moussaoui, H. Hauksdottir, F. Schmidt, Ch. Jutten, J. Chanussot, D. Brie, S. Douté, and J.A. Benediktsson. On the decomposition of mars hyperspectral data by ica and bayesian positive source separation. *Neurocomputing*, 71(10):2194–2208, 2008.

[80] J.M.P. Nascimento and J.M. Bioucas Dias. Does independent component analysis play a role in unmixing hyperspectral data? *Geoscience and Remote Sensing, IEEE Transactions on*, 43(1):175–187, 2005.

[81] C.S. Patlak, R.G. Blasberg, J.D. Fenstermacher, et al. Graphical evaluation of blood-to-brain transfer constants from multiple-time uptake data. *J Cereb Blood Flow Metab*, 3(1):1–7, 1983.

[82] K. Pearson. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901.

[83] M.S. Pedersen, J. Larsen, U. Kjems, and L.C. Parra. A survey of convolutive blind source separation methods. *Multichannel Speech Processing Handbook*, pages 1065–1084, 2007.

[84] K.S. Petersen, L.K. Hansen, T. Kolenda, E. Rostrup, and S. Strother. On the independent components of functional neuroimages. In *Proc. Int. Conf. on Independent Component Analysis and Blind Signal Separation*, pages 615–620, 2000.

[85] A. Piepsz, M. Tondeur, C. Nogarède, F. Collier, K. Ismaili, M. Hall, A. Dobbeleir, and H. Ham. Can severely impaired cortical transit predict which children with pelvi-ureteric junction stenosis detected antenatally might benefit from pyeloplasty? *Nuclear Medicine Communications*, 32(3):199, 2011.

[86] R.A. Poldrack. Region of interest analysis for fmri. *Social cognitive and affective neuroscience*, 2(1):67, 2007.

[87] A. Prigent and P. Cosgriff. Consensus report on quality control of quantitative measurements of renal function obtained from the renogram: International consensus committee from the scientific committee of radionuclides in nephrourology. In *Seminars in nuclear medicine*, volume 29, pages 146–159. Elsevier, 1999.

[88] H. Ren and C.-I. Chang. A generalized orthogonal subspace projection approach to unsupervised multispectral image classification. *Geoscience and Remote Sensing, IEEE Transactions on*, 38(6):2515–2528, 2000.

[89] R.A. Reyment. *Applied factor analysis in the natural sciences*. Cambridge University Press, 1997.

[90] L.J. Rickard, R.W. Basedow, E.F. Zalewski, P.R. Silverglate, and M. Landers. Hydice: An airborne system for hyperspectral imaging. In *Optical Engineering and Photonics in Aerospace Sensing*, pages 173–179. International Society for Optics and Photonics, 1993.

[91] M. Šámal, M. Kárný, H. Surová, E. Maříková, and Z. Dienstbier. Rotation to simple structure in factor analysis of dynamic radionuclide studies. *Physics in medicine and biology*, 32:371, 1987.

[92] M. Šámal and C. Nimmon. Automatic definition of renal regions of interest. *Nuclear medicine communications*, 32(5):419–420, 2011.

[93] M. Šámal, C.C. Nimmon, K.E. Britton, and H. Bergmann. Relative renal uptake and transit time measurements using functional factor images and fuzzy regions of interest. *European Journal of Nuclear Medicine and Molecular Imaging*, 25(1):48–54, 1997.

[94] 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*, volume 39, pages S170–S171. Springer, 2012.

[95] A. Schlotmann, J.H. Clorius, and S.N. Clorius. Diuretic renography in hydronephrosis: renal tissue tracer transit predicts functional course and thereby need for surgery. *European journal of nuclear medicine and molecular imaging*, 36(10):1665–1673, 2009.

[96] V. Šmídl. *The Variational Bayes Approach in Signal Processing*. PhD thesis, University of Dublin, Trinity College, 2004.

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

[98] V. Šmídl and A. Quinn. On bayesian principal component analysis. *Computational statistics & data analysis*, 51(9):4101–4123, 2007.

[99] V. Šmídl and M. Šámal. Robust detection of linear part of Patlak-Rutland plots. ÚTIA AV ČR, v.v.i, Research Report 2243, 2008.

[100] V. Šmídl and O. Tichý. Automatic Regions of Interest in Factor Analysis for Dynamic Medical Imaging. In *2012 IEEE International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2012.

[101] V. Šmídl and O. Tichý. Sparsity in Bayesian Blind Source Separation and Deconvolution. In Hendrik Blockeel et al., editor, *Machine Learning and Knowledge Discovery in Databases (ECML/PKDD 2013)*, volume 8189 of *Lecture Notes in Computer Science*, pages 548–563. Springer Berlin Heidelberg, 2013.

[102] V. Šmídl, O. Tichý, and M. Šámal. 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.

[103] J. Sung, Z. Ghahramani, and S.-Y. Bang. Latent-space variational Bayes. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 30(12):2236–2242, 2008.

[104] D. Sutton. Deconvolution of the renogram. In C.D. Greaves, editor, *Mathematical techniques in nuclear medicine*, pages 106–129. Institute of Physics and Engineering in Medicine, York, 2011.

[105] W.M. Sy, S. Seo, P.C. Sze, M.A. Kimmel, C.J. Homs, J.G. McBride, and K.F. Smith. A patient with three kidneys: a correlative imaging case report. *Clinical nuclear medicine*, 24(4):264–266, 1999.

[106] O. Tichý. Integral Models for Dynamic Renal Scintigraphy, 2010. Thesis, FNSPE CTU.

[107] O. Tichý and V. Šmídl. Kinetic modeling of the dynamic PET brain data using blind source separation methods. In *7th International Conference on BioMedical Engineering and Informatics*. LTU, 2014.

[108] O. Tichý and V. Šmídl. Bayesian blind separation and deconvolution of dynamic image sequences using sparsity priors. *Medical Imaging, IEEE Transaction on*, 34(1):258–266, January 2015.

[109] O. Tichý and V. Šmídl. Bayesian blind source separation with unknown prior covariance. In *The 12th International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA 2015)*, Liberec, Czech Republic, August 2015. (submitted).

[110] O. Tichý and V. Šmídl. Non-parametric bayesian models of response function in dynamic image sequences. *Pre-print submitted to Computer Vision and Image Understanding*, 2015. (arXiv:1503.05684 [stat.ML]).

[111] O. Tichý and V. Šmídl. Variational blind source separation toolbox and its application to hyperspectral image data. In *European Signal Processing Conference 2015 (EUSIPCO 2015)*, Nice, France, August 2015. (submitted).

[112] O. Tichý, V. Šmídl, and M. Šámal. Model-based Extraction of Input and Organ Functions in Dynamic Medical Imaging. In *ECCOMAS Conference on Computational Vision and Medical Image Processing (VipImage 2013)*. Taylor and Francis, 2013.

[113] O. Tichý, V. Šmídl, and M. Šámal. Model-based extraction of input and organ functions in dynamic scintigraphic imaging. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, 2014. (in print, doi:10.1080/21681163.2014.916229).

[114] M.E. Tipping. Sparse Bayesian learning and the relevance vector machine. *The journal of machine learning research*, 1:211–244, 2001.

[115] M.E. Tipping and C.M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622, 1999.

[116] Y. Tomaru, T. Inoue, N. Oriuchi, K. Takahashi, and K. Endo. Semi-automated renal region of interest selection method using the double-threshold technique: inter-operator variability in quantitating 99m Tc-MAG3 renal uptake. *European Journal of Nuclear Medicine and Molecular Imaging*, 25(1):55–59, 1997.

[117] D. Vriens, L.-F. de Geus-Oei, W.J.G. Oyen, and E.P. Visser. A curve-fitting approach to estimate the arterial plasma input function for the

assessment of glucose metabolic rate and response to treatment. *Journal of Nuclear Medicine*, 50(12):1933–1939, 2009.

[118] F.Y. Wang, Ch.-Y. Chi, T.-H. Chan, and Y. Wang. Nonnegative least-correlated component analysis for separation of dependent sources by volume maximization. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(5):875–888, 2010.

[119] J.M. Winn and Ch.M. Bishop. Variational message passing. In *Journal of Machine Learning Research*, pages 661–694, 2005.

[120] D.P. Wipf and S.S. Nagarajan. A new view of automatic relevance determination. In *Advances in neural information processing systems*, pages 1625–1632, 2008.

[121] D.P. Wipf, B.D. Rao, and S. Nagarajan. Latent variable bayesian models for promoting sparsity. *Information Theory, IEEE Transactions on*, 57(9):6236–6255, 2011.

[122] D.P. Wipf and H. Zhang. Revisiting bayesian blind deconvolution. *Journal of Machine Learning Research*, 2014.

[123] K.P. Wong, D. Feng, S.R. Meikle, and M.J. Fulham. Segmentation of dynamic pet images using cluster analysis. *Nuclear Science, IEEE Transactions on*, 49(1):200–207, 2002.

[124] M.W. Woolrich. Bayesian inference in fmri. *NeuroImage*, 62(2):801–810, 2012.

[125] W. Xu, X. Liu, and Y. Gong. Document clustering based on non-negative matrix factorization. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 267–273. ACM, 2003.

[126] J.J. Zaknun, H. Rajabi, A. Piepsz, I. Roca, and M. Dondi. The international atomic energy agency software package for the analysis of scintigraphic renal dynamic studies: a tool for the clinician, teacher, and researcher. In *Seminars in Nuclear Medicine*, volume 41, pages 73–80. Elsevier, 2011.

[127] P. Zanotti-Fregonara, R. Maroy, M.A. Peyronneau, R. Trebossen, and M. Bottlaender. Minimally invasive input function for 2-18 f-fluoro-a-85380 brain pet studies. *European journal of nuclear medicine and molecular imaging*, pages 1–9, 2012.

[128] S. Zhou, K. Chen, E.M. Reiman, D. Li, and B. Shan. A method of generating image derived input function in quantitative 18F-FDG PET study based on the monotonicity of the input and output function curve. *Nuclear medicine communications*, 33(4), 2012.

[129] B. Zitová and J. Flusser. Image registration methods: a survey. *Image and vision computing*, 21(11):977–1000, 2003.