Estimating the number of sources in a noisy convolutive mixture using BIC

Rasmus Kongsgaard Olsson and Lars Kai Hansen

Technical University of Denmark, Informatics and Mathematical Modelling, B321,
DK-2800 Lyngby, Denmark
email: rko@isp.imm.dtu.dk, lkh@imm.dtu.dk

Abstract. The number of source signals in a noisy convolutive mixture is determined based on the exact log-likelihoods of the candidate models. In (Olsson and Hansen, 2004), a novel probabilistic blind source separator was introduced that is based solely on the time-varying second-order statistics of the sources. The algorithm, known as 'KaBSS', employs a Gaussian linear model for the mixture, i.e. AR models for the sources, linear mixing filters and a white Gaussian noise model. Using an EM algorithm, which invokes the Kalman smoother in the E-step, all model parameters are estimated and the exact posterior probability of the sources conditioned on the observations is obtained. The log-likelihood of the parameters is computed exactly in the process, which allows for model evidence comparison assisted by the BIC approximation. This is used to determine the activity pattern of two speakers in a convolutive mixture of speech signals.

1 Introduction

We are pursuing a research program in which we aim to understand the properties of mixtures of independent source signals within a generative statistical framework. We consider *convolutive* mixtures, i.e.,

$$\mathbf{x}_t = \sum_{k=0}^{L-1} \mathbf{A}_k \mathbf{s}_{t-k} + \mathbf{n}_t, \tag{1}$$

where the elements of the source signal vector, $\mathbf{s}_t$, i.e., the d_s statistically independent source signals, are convolved with the corresponding elements of the filter matrix, $\mathbf{A}_k$. The multichannel sensor signal, $\mathbf{x}_t$, are furthermore degraded by additive Gaussian white noise.

It is well-known that separation of the source signals based on second order statistics is infeasible in general. Consider the second order statistic

$$\langle \mathbf{x}_t \mathbf{x}_{t'}^\top \rangle = \sum_{k,k'=0}^{L-1} \mathbf{A}_k \langle \mathbf{s}_{t-k} \mathbf{s}_{t'-k'}^\top \rangle \mathbf{A}_{k'}^\top + \mathbf{R},$$

where $\mathbf{R}$ is the (diagonal) noise covariance matrix. If the sources are white noise stationary, the source covariance matrix can be assumed proportional to the unit matrix without loss of generality, and we see that the statistic is symmetric to a common rotation of all mixing matrices $\mathbf{A}_k \rightarrow \mathbf{A}_k\mathbf{U}$. This rotational invariance means that the statistic is not informative enough to identify the mixing matrix, hence, the source time series.

However, if we consider stationary sources with *known*, non-trivial, autocorrelations $\langle \mathbf{s}_t \mathbf{s}_{t'}^\top \rangle = \mathbf{C}(t-t')$, and we are given access to measurements involving multiple values of $\mathbf{C}(t-t')$, the rotational degrees of freedom are constrained and we will be able to recover the mixing matrices up to a choice of sign and scale of each source time series. Extending this argument by the observation that the mixing model (1) is invariant to filtering of a given column of the convolutive filter provided that the inverse filter is applied to corresponding source signal, we see that it is infeasible to identify the mixing matrices if these arbitrary inverse filters can be chosen to that they 'whiten' the sources.

For non-stationary sources, on the other hand, the autocorrelation functions vary through time and it is not possible to choose a single common whitening filter for each source. This means that the mixing matrices may be identifiable from multiple estimates of the second order correlation statistic (2) for non-stationary sources. Parra and Spence [1] provide analysis in terms of the number of free parameters vs. the number of linear conditions.

Also in [1], the constraining effect of source non-stationarity was exploited by simultaneously diagonalizing multiple estimates of the source power spectrum. In [2] we formulated a generative probabilistic model of this process and proved that it could estimate sources and mixing matrices in noisy mixtures. A state-space model -a Kalman filter- was specialized and augmented by a stacking procedure to model a noisy convolutive mixture of non-stationary colored noise sources, and a forward-backward EM approach was used to estimate the source statistics, mixing coefficients and the diagonal noise covariance matrix. The EM algorithm furthermore provides an exact calculation of the likelihood as it is possible to average over all possible source configurations. Other approaches based on EM schemes for source inference are [3], [4] and [5]. In [6], a non-linear state-space model is proposed.

In this presentation we elaborate on the generative model and its applications. In particular, we use the exact likelihood calculation to make inference about the dimensionality of the model, i.e. the number of sources. Choosing the incorrect model order can lead to either a too simple, biased model or a too complex model. We use the so-called Bayes Information Criterion (BIC) [7] to approximate the Bayes factor for competing hypotheses.

The model is stated in section 2, while the learning in the particular model described in section 3. Model order selection using BIC is treated in section 4. Experiments for speech mixtures are shown in section 5.

2 The model

As indicated above, the sources must be assumed non-stationary in order to uniquely retrieve the parameters and sources, since the estimation is based on second-order statistics. In line with [1], this is obtained by *segmenting* the signals into frames, in which the wide-sense stationarity of the sources is assumed. A separate source model is assumed for each segment. The channel filters and observation noise covariance are assumed stationary across segments in the entire observed signal.

The colored noise sources are modelled by AR(p) random processes. In segment n, source i is represented by:

$$s_{i,t}^n = f_{i,1}^n s_{i,t-1}^n + f_{i,2}^n s_{i,t-2}^n + \ldots + f_{i,p}^n s_{i,t-p}^n + v_{i,t}^n \tag{2}$$

where $n \in \{1, 2, .., N\}$ and $i \in \{1, 2, .., d_s\}$. The innovation noise, $v_{i,t}$, is white Gaussian. In order to make use of well-established estimation theory, the above recursion is fitted into the framework of Gaussian linear models, for which a review is found in e.g. [8]. The Kalman filter model is an instance of this model that particularly treats continuous Gaussian linear models used widely in e.g. control and speech enhancement applications. The general Kalman filter with no control inputs is defined:

$$\begin{aligned} \mathbf{s}_t &= \mathbf{F}\mathbf{s}_{t-1} + \mathbf{v}_t \\ \mathbf{x}_t &= \mathbf{A}\mathbf{s}_t + \mathbf{n}_t \end{aligned} \tag{3}$$

where $\mathbf{v}_t$ and $\mathbf{n}_t$ are white Gaussian noise signals that drive the processes.

In order to incorporate the colored noise sources, equation (2), into the Kalman filter model, the well-known principle of *stacking* must be applied, see e.g [9]. At any time, the stacked source vector, $\bar{\mathbf{s}}_t^n$, contains the last p samples of all d_s sources:

$$\bar{\mathbf{s}}_t^n = \left[(\mathbf{s}_{1,t}^n)^\top \; (\mathbf{s}_{2,t}^n)^\top \; \ldots \; (\mathbf{s}_{d_s,t}^n)^\top \right]^\top$$

The component vectors, $\mathbf{s}_{i,t}^n$, contain the p most recent samples of the individual sources:

$$\mathbf{s}_{i,t}^n = \left[s_{i,t}^n \; s_{i,t-1}^n \; \ldots \; s_{i,t-p+1}^n \right]^\top$$

In order to maintain the statistical independency of the sources, a constrained format must be imposed on the parameters:

$$\bar{\mathbf{F}}^n = \begin{bmatrix} \bar{\mathbf{F}}_1^n & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \bar{\mathbf{F}}_2^n & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \bar{\mathbf{F}}_{d_s}^n \end{bmatrix}, \bar{\mathbf{F}}_i^n = \begin{bmatrix} f_{i,1}^n & f_{i,2}^n & \cdots & f_{i,p-1}^n & f_{i,p}^n \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix}$$

$$\bar{\mathbf{Q}}^n = \begin{bmatrix} \bar{\mathbf{Q}}_1^n & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \bar{\mathbf{Q}}_2^n & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \bar{\mathbf{Q}}_{d_s}^n \end{bmatrix}, \; (\bar{\mathbf{Q}}_i^n)_{jj'} = \begin{cases} q_i^n & j = j' = 1 \\ 0 & j \neq 1 \bigvee j' \neq 1 \end{cases}$$

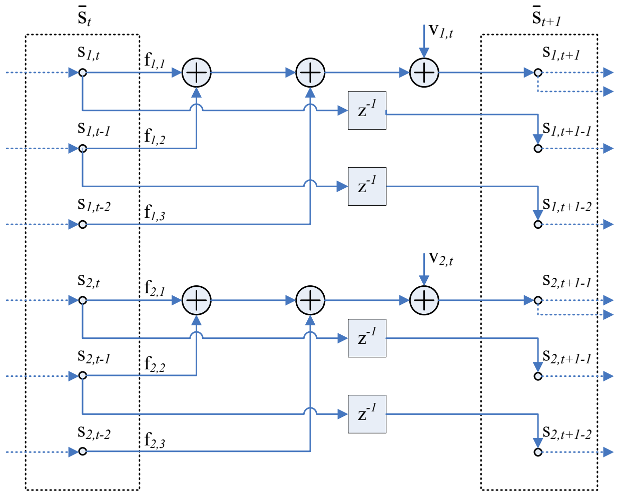

Fig. 1. The multiplication of $\bar{\mathbf{F}}$ on $\bar{\mathbf{s}}_t$ and the addition of innovation noise, $\mathbf{v}_t$, shown for an example involving two AR(3) sources. The special contrained format of $\bar{\mathbf{F}}$ simultaneously ensures the storage of past samples.

The matrix $\mathbf{A}$ of (3) is left unconstrained but its dimensions must be expanded to $d_x \times (p \times d_s)$ to reflect the stacking of the sources. Conveniently, its elements can be interpreted as the impulse responses of the channel filters of (1):

$$\bar{\mathbf{A}} = \begin{bmatrix} \mathbf{a}_{11}^\top & \mathbf{a}_{12}^\top & .. & \mathbf{a}_{1d_s}^\top \\ \mathbf{a}_{21}^\top & \mathbf{a}_{22}^\top & .. & \mathbf{a}_{2d_s}^\top \\ \mathbf{a}_{d_x1}^\top & \mathbf{a}_{d_x2}^\top & .. & \mathbf{a}_{d_xd_s}^\top \end{bmatrix}$$

where $\mathbf{a}_{ij} = [a_{ij,1}, a_{ij,2}, .., a_{ij,L}]^\top$ is the filter between source i and sensor j. Having defined the stacked sources and the constrained parameter matrices, the total model is:

$$\begin{aligned} \bar{\mathbf{s}}_t^n &= \bar{\mathbf{F}}^n \bar{\mathbf{s}}_{t-1}^n + \bar{\mathbf{v}}_t^n \\ \mathbf{x}_t^n &= \bar{\mathbf{A}} \bar{\mathbf{s}}_t^n + \mathbf{n}_t^n \end{aligned}$$

where $\bar{\mathbf{v}}_t^n \sim (\mathbf{0}, \bar{\mathbf{Q}}^n)$ and $\mathbf{n}_t^n \sim (\mathbf{0}, \bar{\mathbf{F}}^n)$. Figures 1 and 2 illustrate the updating of the stacked source vector, $\bar{\mathbf{s}}_t$ and the effect of multiplication by $\bar{\mathbf{A}}$, respectively.

3 Learning

Having described the convolutive mixing problem in the general framework of linear Gaussian models, more specifically the Kalman filter model, optimal inference of the sources is obtained by the Kalman smoother. However,

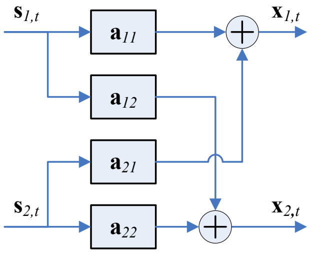

Fig. 2. The effect of the matrix multiplication $\bar{\mathbf{A}}$ on $\bar{\mathbf{s}}_t$ is shown in the system diagram. The source signals are filtered (convolved) with the impulse responses of the channel filters. Observation noise and the segment index, n, are omitted for brevity.

since the problem at hand is effectively *blind*, the parameters are estimated. Along the lines of, e.g. [8], an EM algorithm will be used for this purpose, i.e. $\mathcal{L}(\theta) \geq \mathcal{F}(\theta, \hat{p}) \equiv \mathcal{J}(\theta, \hat{p}) - \mathcal{R}(\hat{p})$, where $\mathcal{J}(\theta, \hat{p}) \equiv \int d\mathbf{S}\hat{p}(\mathbf{S}) \log p(\mathbf{X}, \mathbf{S}|\theta)$ and $\mathcal{R}(\hat{p}) \equiv \int d\mathbf{S}\hat{p}(\mathbf{S}) \log \hat{p}(\mathbf{S})$ were introduced. In accordance with standard EM theory, $\mathcal{J}(\theta, \hat{p})$ is optimized wrt. θ in the M-step. The E-step infers the model posterior, $\hat{p} = p(\mathbf{S}|\mathbf{X}, \theta)$. The combined E and M steps are guaranteed not to decrease $\mathcal{L}(\theta)$.

3.1 E-step

The forward-backward recursions which comprise the Kalman smoother is employed in the E-step to infer the source posterior, $p(\mathbf{S}|\mathbf{X}, \theta)$, i.e. the joint posterior of the sources conditioned on all observations. The relevant second-order statistics of this distribution in segment n is the posterior mean, $\hat{\bar{\mathbf{s}}}_t^n \equiv \langle \bar{\mathbf{s}}_t^n \rangle$, and autocorrelation, $\mathbf{M}_{i,t}^n \equiv \langle \mathbf{s}_{i,t}^n (\mathbf{s}_{i,t}^n)^\top \rangle \equiv [\, \mathbf{m}_{i,1,t}^n \; \mathbf{m}_{i,2,t}^n \; .. \; \mathbf{m}_{i,L,t}^n \,]^\top$, along with the time-lagged covariance, $\mathbf{M}_{i,t}^{1,n} \equiv \langle \mathbf{s}_{i,t}^n (\mathbf{s}_{i,t-1}^n)^\top \rangle \equiv [\, \mathbf{m}_{i,1,t}^{1,n} \; \mathbf{m}_{i,2,t}^{1,n} \; .. \; \mathbf{m}_{i,L,t}^{1,n} \,]^\top$. In particular, $m_{i,t}^n$ is the first element of $\mathbf{m}_{i,1,t}^n$. All averages are performed over $p(\mathbf{S}|\mathbf{X}, \theta)$. The forward recursion also yields the likelihood $\mathcal{L}(\theta)$.

3.2 M-step

The estimators are derived by straightforward optimization of $\mathcal{J}(\theta, \hat{p})$ wrt. the parameters. It is used that the data model, $p(\mathbf{X}, \mathbf{S}|\theta)$, factorizes. See, e.g., [8] for background, or [2] for details. The estimators for source i in segment n are:

$$
\begin{aligned}
\mu_{i,\mathbf{new}}^n &= \hat{\mathbf{s}}_{i,1}^n \\
\boldsymbol{\Sigma}_{i,\mathbf{new}}^n &= \mathbf{M}_{i,1}^n - \mu_{i,\mathbf{new}}^n (\mu_{i,\mathbf{new}}^n)^\top \\
(\mathbf{f}_{i,\mathbf{new}}^n)^\top &= \Big[\sum_{t=2}^{\tau} (\mathbf{m}_{i,t}^{1,n})^\top \Big] \Big[\sum_{t=1}^{\tau} \mathbf{M}_{i,t-1}^n \Big]^{-1} \\
q_{i,\mathbf{new}}^n &= \frac{1}{\tau - 1} \Big[\sum_{t=2}^{\tau} m_{i,t}^n - (\mathbf{f}_{i,\mathbf{new}}^n)^\top \mathbf{m}_{i,t}^{1,n} \Big]
\end{aligned}
$$

The stacked estimators, $\bar{\mu}^n_{\mathbf{new}}$, $\bar{\mathbf{\Sigma}}^n_{\mathbf{new}}$, $\bar{\mathbf{F}}^n_{\mathbf{new}}$ and $\bar{\mathbf{Q}}^n_{\mathbf{new}}$ are reconstructed from the above as defined in section 2. The constraints on the parameters cause the above estimators to differ from those of the general Kalman model, which is not the case for $\bar{\mathbf{A}}_{\mathbf{new}}$ and $\mathbf{R}_{\mathbf{new}}$:

$$\begin{aligned}
\bar{\mathbf{A}}_{\mathbf{new}} &= \Big[\sum_{n=1}^{N}\sum_{t=1}^{\tau}\mathbf{x}^n_t(\hat{\bar{\mathbf{s}}}^n_t)^\top\Big]\Big[\sum_{n=1}^{N}\sum_{t=1}^{\tau}\bar{\mathbf{M}}^n_t\Big]^{-1}\\
\mathbf{R}_{\mathbf{new}} &= \frac{1}{N\tau}\sum_{n=1}^{N}\sum_{t=1}^{\tau}\mathrm{diag}[\mathbf{x}^n_t(\mathbf{x}^n_t)^\top - \bar{\mathbf{A}}_{\mathbf{new}}\hat{\bar{\mathbf{s}}}^n_t(\mathbf{x}^n_t)^\top]
\end{aligned}$$

4 Estimating the number of sources using BIC

In the following is described a scheme for determining d_s based on the likelihood of the parameters. A similar approach was taken in previous work, see [10]. Model control in a strictly Bayesian sense amounts to selecting the most probable hypothesis, based on the posterior probability of the model conditioned on the data:

$$p(d_s|\mathbf{X}) = \frac{p(\mathbf{X}|d_s)p(d_s)}{\sum_{d_s} p(\mathbf{X}, d_s)} \tag{4}$$

In cases where all models, a priori, are to be considered equally likely, (4) reduces to $p(d_s|\mathbf{X}) \propto p(\mathbf{X}|d_s)$. The Bayes factor, $p(\mathbf{X}|d_s)$, is defined:

$$p(\mathbf{X}|d_s) = \int d\theta p(\mathbf{X}|\theta, d_s)p(\theta|d_s) \tag{5}$$

Bayes information criterion (BIC), see [7], is an approximation of (5) to be applied in cases where the marginalization of θ is intractable:

$$p(\mathbf{X}|d_s) \approx p(\mathbf{X}|\theta_{ML}, d_s)\tau^{-\frac{|\theta|}{2}} \tag{6}$$

The underlying assumptions are that (5) can be evaluated by Laplace integration, i.e. $\log p(\mathbf{X}|\theta, d_s)$ is well approximated by a quadratic function for large amounts of data ($\tau \rightarrow \infty$), and that the parameter prior $p(\theta|d_s)$ can be assumed constant under the integral.

5 Experiments

In order to demonstrate the applicability of the model control setup, a convolutive mixture of speech signals was generated and added with observation noise. The four models/hypotheses that we investigate in each time frame are that only one of two speakers are active, **1** and **2**, respectively, that both of them are active, **1+2**, or that none of them are active, **0**.

Recordings of male speech[1], which were also used in [11], were filtered through the $(2 \times 2 = 4)$ known channel filters:

$$\bar{\mathbf{A}} = \begin{bmatrix} 1.00\ 0.35\ -0.20 & 0.00\ 0.00, & 0.00\ 0.00\ -0.50\ -0.30\ 0.20 \\ 0.00\ 0.00\ \ \ 0.70 & -0.20\ 0.15, & 1.30\ 0.60\ \ \ 0.30\ \ \ 0.00\ 0.00 \end{bmatrix}$$

Observation noise was added to simulate SNR=15dB in the two sensor signals. KaBSS was then invoked in order to separate the signals and estimate $\bar{\mathbf{A}}$ and $\mathbf{R}$, as shown in [2]. The signals were segmented into frames of $\tau = 160$ samples. The obtained estimates of $\bar{\mathbf{A}}$ and $\mathbf{R}$ were treated as known true parameters in the following. In each segment and for each model-configuration, KaBSS was separately reinvoked to estimate the source model parameters, $\bar{\mathbf{F}}^n$, $\bar{\mathbf{Q}}^n$, and obtain the log-likelihood, $\mathcal{L}(\theta)$, of the various models. The four resulting $\mathcal{L}(\theta)$'s were then processed in the BIC model control scheme described in section 4. The number of samples in (6) were set to τ although the sensor signals are not i.i.d. This approximation is, however, acceptable due to the noisy character of speech. Figure 3 displays the source signals, the mixtures and the most likely hypothesis in each time frame. Convincingly, the MAP speech activity detector selects the correct model.

6 Conclusion

An EM algorithm, 'KaBSS', which builds on probabilistic inference in a generative linear convolutive mixture model with Gaussian sources was introduced in [2]. This contribution expands the model and its utility by showing that the exact computation of the log-likelihood, which is readily available as an output of the forward-backward recursion, can be exploited in a BIC-based model selection scheme. The result is an exploratory tool capable of determining the correct number of sources in a convolutive mixture. In particular, it was shown that the activity pattern of two speech sources in a convolutive mixture can be well estimated. Potential applications include the ability to select the correct model in speech enhancement and communication algorithms, hopefully resulting in more robust estimation.

References

1. Parra, L., Spence C., Convolutive blind separation of non-stationary sources. IEEE Transactions, Speech and Audio Processing (5), 320-7, 2000.
2. Olsson, R. K., Hansen L. K., Probabilistic blind deconvolution of non-stationary source. Proc. EUSIPCO, 2004, *submitted*.
3. Moulines E., Cardoso J. F., Gassiat E., Maximum likelihood for blind separation and deconvolution of noisy signals using mixture models, Proc. ICASSP (5), 3617-20, 1997.
4. Attias H., New EM algorithms for source separation and deconvolution with a microphone array. Proc. ICASSP (5), 297-300, 2003.

[1] Available at http://www.ipds.uni-kiel.de/pub_exx/bp1999_1/Proto.html.

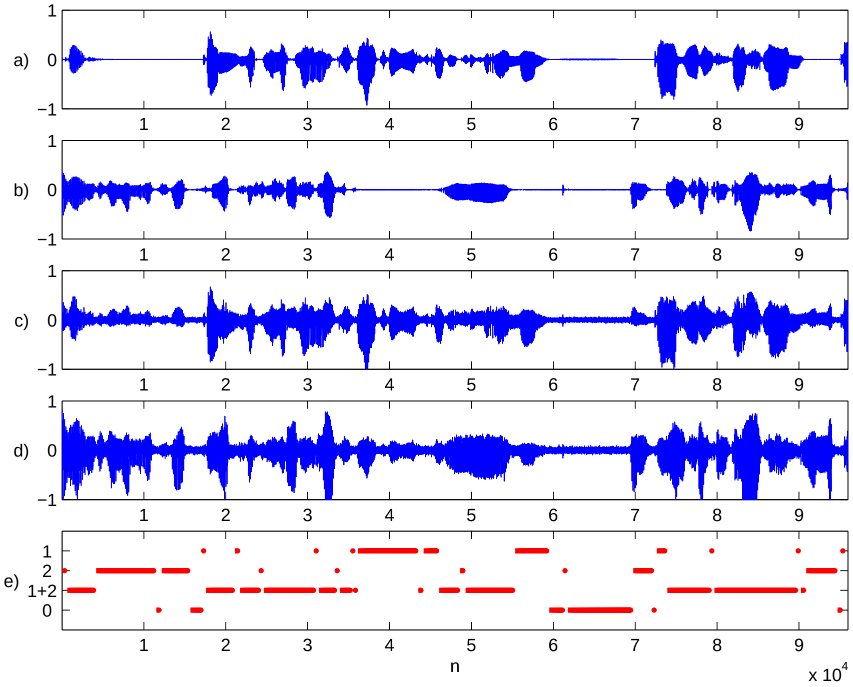

Fig. 3. From top to bottom, **a & b**) the original speech signals, **c & d**) the noisy mixtures and **e**) the most likely model in each segment. The four models are,
1: first speaker exclusively active, **2**: second speaker exclusively active, **1+2**: both speakers simultaneously active and **0**: no speaker activity. A segment of 6 seconds of speech, sampled at $F_s = 16\text{kHz}$, is shown.

5. Todorovic-Zarkula S., Todorovic B., Stankovic M., Moraga C., Blind separation and deconvolution of nonstationary signals using extended Kalman filter. South-Eastern European workshop on comp. intelligence and IT, 2003.
6. Valpola H., Karhunen J, An unsupervised ensemble learning method for nonlinear dynamic state-space models. Neural Computation 14 (11), MIT Press, 2647-2692, 2002.
7. Schwartz G., Estimating the dimension of a model. Annals of Statistics (6), 461-464, 1978.
8. Roweis S., Ghahramani Z., Spence C., A unifying review of linear Gaussian models. Neural Computation (11), 305-345, 1999.
9. Doblinger G., An adaptive Kalman filter for the enhancement of noisy AR signals. IEEE Int. Symp. on Circuits and Systems (5), 305-308, 1998.
10. Højen-Sørensen P. A. d. F. R., Winther O., Hansen L. K., Analysis of functional neuroimages using ICA with adaptive binary sources. Neurocomputing (49), 213-225, 2002.
11. Peters B., Prototypische Intonationsmuster in deutscher Lese- und Spontansprache. AIPUK (34), 1-177, 1999.