Model selection using limiting distributions of second-order blind source separation algorithms

Katrin Illner[a,b], Jari Miettinen[c], Christiane Fuchs[a,b], Sara Taskinen[c], Klaus Nordhausen[d], Hannu Oja[d], Fabian J Theis[a,b,*]

[a] *Institute of Computational Biology, German Research Center for Environmental Health, 85764 Neuherberg, Helmholtz Zentrum München, Germany*
[b] *Chair of Mathematical Modeling of Biological Systems, Center for Mathematics, 85748 Garching, Technische Universität München, Germany*
[c] *Department of Mathematics and Statistics, 40014 University of Jyväskylä, Finland*
[d] *Department of Mathematics and Statistics, 20014 University of Turku, Finland*

Abstract

Signals, recorded over time, are often observed as mixtures of multiple source signals. To extract relevant information from such measurements one needs to determine the mixing coefficients. In case of weakly stationary time series with uncorrelated source signals, this separation can be achieved by jointly diagonalizing sample autocovariances at different lags, and several algorithms address this task. Often the mixing estimates contain close-to-zero entries and one wants to decide whether the corresponding source signals have a relevant impact on the observations or not. To address this question of model selection we consider the recently published second-order blind identification procedures `SOBIdef` and `SOBIsym` which provide limiting distributions of the mixing estimates. For the first time, such distributions enable informed decisions about the presence of second-order stationary source signals in the data. We consider a family of linear hypothesis tests and information criteria to perform model selection as second step after parameter estimation. In simulations we consider different time series models. We validate the model selection performance and demonstrate a good recovery of the true zero pattern of the mixing matrix.

Key words: joint diagonalization, SOBI, asymptotic normality, pattern identification

*Correspondence: fabian.theis@helmholtz-muenchen.de

1. Introduction

Time resolved signals appear in a large variety of contexts, and often one observes a multivariate mixture of different signals rather than separated ones. In blind source separation (BSS) we assume a linear and instantaneous mixing model and aim to estimate the underlying source signals together with the mixing weights. In case of weakly stationary time series with uncorrelated source signals, a mixing matrix can be estimated based on the second-order statistics of the observations. The problem then reduces to jointly diagonalizing sample autocovariances at different lags. Many existing BSS algorithms are based on this idea [1, 2, 3, 4, 5]. A review on joint diagonalization algorithms is given in [6]. Applications range from audio recordings to biomedical signal or image data. For the latter the assumption of uncorrelated source components can be extended to the spatial dimension of the data [7, 8]. In an application to high dimensional functional magnetic resonance imaging (fMRI), for example, patients alternately passed through periods of rest and photic stimulus. In comparison to other BSS methods, joint diagonalization could identify a signal with high coherence to the stimulus [8]. Another widely used measuring technique is electroencephalography (EEG). Here, the brain's electrical activity is recorded and joint diagonalization could successfully separate artifacts like eye movement or blinking from the data [9]. If the EEG signals arise from correlated stimulation of the left and right somatosensory cortices a large number and wide range of time delays is preferable [10].

To draw further conclusions from source separation, one often wants to know whether single source signals are present in a specific observation. More precisely, one wants to decide whether close-to-zero entries of the mixing estimate are actually zero or not. This is commonly done by thresholding which lacks statistical motivation. To provide informative decisions we develop suitable model selection criteria. To that end, we consider the recently published second-order blind identification versions `SOBIdef` [11] and `SOBIsym` [12]. For both algorithms the authors showed that the (un-)mixing estimates are asymptotically normally distributed under mild conditions, and they derived limiting variances of the estimates when the time series length goes to infinity. Based on these distributions we create a framework to perform model selection on the mixing estimates. Here, we use a family of linear

hypothesis tests and different information criterions including the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). To speed up the selection process we also consider an alternative information criterion that does not require the maximum likelihood parameter estimates.

In the first part, we state the second-order source separation problem (Section 2) and shortly review the algorithms `SOBIdef` and `SOBIsym` (Section 3). Their practical estimation performance has not been evaluated yet. To figure it out, we compare both algorithms to the established methods `SOBI` [2] and the non-orthogonal `ACDC` [4]. We find that `SOBIsym` achieves the same estimation results as `SOBI` but with the gain of knowing the limiting distribution of the (un-)mixing estimates (Section 4). In the main part, we then demonstrate how the additional information about the distribution can be used to choose between different candidates for the mixing matrix (Section 5). In simulations we consider a BSS model where the mixing matrix contains zero and close-to-zero entries (Section 6). For both algorithms `SOBIdef` and `SOBIsym` the testing performance could be validated and we show the percentages of correctly reconstructed zero-patterns among different time series models and for the different selection approaches.

Throughout the paper we use bold symbols to denote random variables and solid symbols to denote parameters and realizations of random variables.

2. A second-order blind source separation model

Let $\{\boldsymbol{x}(t)\}_{t\in\mathbb{Z}}$ be a p-variate observable time series that is weakly stationary. This means that the mean and the autocovariance at any lag $\tau \in \mathbb{N}$ do not change with respect to time. After mean-removal we assume a zero-centered process that is generated by the following linear mixing model:

$$\boldsymbol{x}(t) = \Omega \boldsymbol{z}(t) \ , \quad t \in \mathbb{Z} \ . \tag{1}$$

Here, Ω denotes a deterministic full rank $p \times p$ mixing matrix and $\{\boldsymbol{z}(t)\}_{t\in\mathbb{Z}}$ is a p-variate unobservable time series that is weakly stationary as well and has uncorrelated components. More precisely, we assume:

- $(A1)$ $E(\boldsymbol{z}(t)) = 0$,
- $(A2)$ $Cov(\boldsymbol{z}(t), \boldsymbol{z}(t)) = I_p$,
- $(A3)$ $Cov(\boldsymbol{z}(t), \boldsymbol{z}(t+\tau)) = Cov(\boldsymbol{z}(t+\tau), \boldsymbol{z}(t)) = \Lambda_\tau$ is diagonal for all lags $\tau \in \mathbb{N}$, and

(*A*4) for all $i \neq j \in \{1, \ldots, p\}$ there exists a lag $\tau \in \mathbb{N}$ such that $\lambda_{\tau i} \neq \lambda_{\tau j}$ with $\lambda_{\tau i}$ and $\lambda_{\tau j}$ the ith and jth diagonal entry of Λ_τ .

With the scaling to unit variance in (*A*2) and the assumption (*A*4) the mixing becomes unique up to a sign-changing permutation: If $\boldsymbol{x}(t) = \Omega_1 \boldsymbol{z}_1(t) = \Omega_2 \boldsymbol{z}_2(t)$, then $\Omega_2 = \Omega_1 B$ and $\boldsymbol{z}_2(t) = B^{-1}\boldsymbol{z}_1(t)$, where B contains exactly one non-zero entry per row and column and these entries equal ± 1. This restriction on B follows from the spectral theorem [13].

In second-order source separation we consider the second-order statistics of the observable process, and with these we estimate the mixing matrix Ω as well as the unobservable process $\{\boldsymbol{z}(t)\}_{t\in\mathbb{Z}}$. The autocovariance of $\{\boldsymbol{x}(t)\}_{t\in\mathbb{Z}}$ at lag $\tau \in \mathbb{N}$ is of the form:

$$Cov(\boldsymbol{x}(t), \boldsymbol{x}(t+\tau)) = \Omega \Lambda_\tau \Omega' \,,$$

where $\Lambda_0 = I_p$ at lag zero. Let now $x(1), \ldots, x(T)$ be observations at subsequent time points. The sample autocovariance at lag τ is then given as

$$S_\tau = \frac{1}{T-\tau} \sum_{t=1}^{T-\tau} x(t) x(t+\tau)' \,.$$

To determine an unmixing estimate we jointly diagonalize sample autocovariances at distinct lags $\tau_1, \ldots, \tau_K$. We assume that $\{\tau_1, \ldots, \tau_K\} \subseteq \mathbb{N}$ is such that (*A*4) also holds for $\{\tau_1, \ldots, \tau_K\}$ instead of $\mathbb{N}$. For better readability, we denote the corresponding autocovariances as $S_1, \ldots, S_K$ even if the lags are different from $1, \ldots, K$. An unmixing estimate is then a $p \times p$ matrix $\Gamma = (\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_p)'$ that minimizes the off-diagonal elements of $\Gamma S_k \Gamma'$ for all $k = 1, \ldots, K$ in the sense that

$$f^*(\Gamma) = \sum_{k=1}^{K} \left\| \mathrm{off}(\Gamma S_k \Gamma') \right\|_{\mathrm{F}}^2$$

is minimized under the constraint $\Gamma S_0 \Gamma' = I_p$. Here, $\mathrm{off}(M) = M - \mathrm{diag}(M)$ with $\mathrm{diag}(M)$ a diagonal matrix consisting of the diagonal entries of M, and $\|\,.\,\|_{\mathrm{F}}$ denotes the Frobenius norm of a matrix. The above minimization is equivalent to the maximization of

$$f(\Gamma) = \sum_{k=1}^{K} \left\| \mathrm{diag}(\Gamma S_k \Gamma') \right\|_{\mathrm{F}}^2 = \sum_{j=1}^{p} \sum_{k=1}^{K} (\boldsymbol{\gamma}_j' S_k \boldsymbol{\gamma}_j)^2 \tag{2}$$

under the same constraint. From the spectral theorem it follows that an optimal solution Γ is indeed an estimate of the unmixing matrix.

An important class of joint diagonalization algorithms is restricted to the estimation of orthogonal mixing matrices. In this case we first pre-whiten the data using $\tilde{\boldsymbol{x}}(t) = V\boldsymbol{x}(t) = (V\Omega)\boldsymbol{z}(t) = \tilde{\Omega}\boldsymbol{z}(t)$ with $V = S_0^{-1/2}$ for $t = 1, \ldots, T$. The whitened process has unit variance and from the constraint it follows that $\tilde{\Gamma}$ and $\tilde{\Omega}$ are orthogonal. From the diagonal unmixing estimate for the whitened process we then get an unmixing estimate for the original process by multiplication with V from the left.

3. The algorithms SOBIdef and SOBIsym

Recently, Miettinen et al. proposed the new second-order blind identification algorithms `SOBIdef` [11] and `SOBIsym` [12]. In the deflation based approach (`SOBIdef`) the single rows of an unmixing matrix $\Gamma = (\gamma_1, \ldots, \gamma_p)'$ are estimated one after the other, such that at each step j only $f(\gamma_j) = \sum_{k=1}^{K} (\gamma_j' S_k \gamma_j)^2$ is maximized. In the symmetric approach (`SOBIsym`), in contrast, all rows of Γ are estimated at once, such that the complete sum in (2) is maximized. Using Lagrange multiplier techniques, Miettinen et al. formulated in both cases estimating equations for an optimal solution Γ and derived iterative algorithms from these equations. Both algorithms contain an update step, where single rows of the current estimate are replaced by $H(\gamma_j) = \sum_{k=1}^{K} (\gamma_j' S_k \gamma_j) S_k \gamma_j$ and an orthogonalization step based on the Gram-Schmidt process or singular value decomposition, respectively.

`SOBIdef`. For $j = 1, \ldots, p-1$ initialize γ_j (discussed below), and then alternate until convergence:

$$\begin{aligned} &\text{step 1.} \quad \gamma_j \leftarrow H(\gamma_j) \\ &\text{step 2.} \quad \gamma_j \leftarrow (I_p - \textstyle\sum_{r=1}^{j-1} \gamma_r \gamma_r')\gamma_j \\ &\qquad\qquad \gamma_j \leftarrow \gamma_j / \left\|\gamma_j\right\| \end{aligned}$$

`SOBIsym`. Initialize Γ randomly, and then alternate until convergence:

$$\begin{aligned} &\text{step 1.} \quad \Gamma \leftarrow (H(\gamma_1), \ldots, H(\gamma_p)) \\ &\text{step 2.} \quad \Gamma \leftarrow \mathrm{svd}^p(\Gamma) \end{aligned}$$

Here, $\text{svd}^p(\Gamma)$ denotes orthogonalization of Γ using singular value decomposition. If $\Gamma = U\Sigma V'$ with U, V orthogonal and Σ diagonal, then step 2 results in $\Gamma = UV'$. This is the closest orthogonal matrix to Γ in terms of the Frobenius norm.

The performance of `SOBIdef` depends on the extraction order of the rows, or equivalently, on permutations of the initial vectors $\gamma_1, \ldots, \gamma_p$. If we initialize `SOBIdef` with all $p!$ permutations of these vectors, we get up to $p!$ different estimates. To directly determine the estimate with the highest value for the maximization function (2), we introduce a randomization at each step j. Among a set of $100\,(p-j+1)$ random vectors orthogonal to $\gamma_1, \ldots, \gamma_{j-1}$ we choose the vector γ_j with the highest value $f(\gamma_j) = \sum_{k=1}^{K} (\gamma_j' S_k \gamma_j)^2$ and use it as initialization. With this, both algorithms are independent of the initial guess. Further, the resulting estimate Γ is affine equivariant, i.e. if Γ and $\tilde{\Gamma}$ are the estimates derived from $\boldsymbol{x}(t)$ and $\tilde{\boldsymbol{x}}(t) = B\boldsymbol{x}(t)$ for $t = 1, \ldots, T$ and any invertible matrix B, then $\Gamma = \tilde{\Gamma} B$.

3.1. Algorithm performance

To give an idea about the estimation performance of `SOBIdef` and `SOBIsym` we compare both algorithms to the following well-established methods.

`SOBI` [2] is the original second-order blind identification algorithm and it is based on Jacobi rotations. Starting with an orthogonal initial guess for the unmixing matrix, the algorithm determines for each pair of rows in turn an optimal Jacobi rotation to maximize (2). The current unmixing estimate is then rotated in the plane spanned by the two rows. For `SOBI` as well as for `SOBIdef` and `SOBIsym` the final mixing estimate is orthogonal by construction. Thus, these algorithms require a pre-whitening of the data.

`ACDC` [4] is a non-orthogonal algorithm. Iteratively, the algorithm optimizes in the AC-step a single row of the unmixing estimate and updates in the DC-step the estimate of the diagonal source autocovariances. Although `ACDC` does not require pre-whitened data, we observed a better convergence on such data. In this case, we need to include $S_0 = I_p$ to the set of autocovariances that we jointly diagonalize to assure an orthogonal mixing estimate.

As performance measure we use the minimum distance index [14]. This index compares the product of unmixing estimate and true mixing to the identity matrix and is defined as

$$\text{MDI}(\hat{\Gamma}, \Omega) = \frac{1}{\sqrt{p-1}} \inf_{C \in \mathcal{C}} \|C\hat{\Gamma}\Omega - I_p\| \,, \tag{3}$$

where $\mathcal{C}$ is the set of $p \times p$ matrices with one non-zero entry per row and column. Thus the index is independent of sign and permutation of the rows estimates. All MDI values are in $[0, 1]$, and we say that the mixing estimate $\hat{\Omega}$ is close at the true mixing matrix Ω if the value is low.

Finally, we consider four different time series models to generate data:

(*i*) *AR(4)-model*: three AR(4)-processes with coefficient vectors $(0.2, -0.5, 0.5, -0.4)$, $(0.3, 0.1, -0.7, 0.2)$, $(-0.2, 0.3, 0.1, 0.1)$ and normal innovations

(*ii*) *ARMA-model*: three ARMA-processes with AR-coefficient vectors $(-0.4, 0.2, -0.3)$, $(0.2, 0.5, -0.1)$, $(0.5, -0.1, 0.1)$ and MA-coefficient vectors $(0.1 - 0.3, 0.2, 0.2, -0.1)$, $(0.7, 0.4, -0.3, 0.1, -0.2)$, $(-0.5, -0.4, -0.2, 0.5, 0.1)$ and normal innovations

(*iii*) *Mixed model*: one AR(3)-, one AR(1)-, one MA(10)-process with coefficient vectors $(0.5, 0.1, 0.3)$, (0.7), $(0.4, 0.2, -0.1, -0.4, 0.3, 0.2, 0.6, 0.1, -0.3, -0.1)$ and normal innovations

(*iv*) *Close-coefficient model*: three MA(3)-processes with coefficient vectors $(-0.25, 0.1, 0.5)$, $(-0.3, 0.1, 0.35)$, $(-0.2, 0.07, 0.4)$ and normal innovations

From all models we generate times series of length T and scale each component to unit variance. We mix observations from these source signals using a random mixing matrix Ω with entries from $\mathcal{U}[-1, 1]$. For joint diagonalization we consider sample autocovariances at lags $\tau = 1, \ldots, K$. Note that all algorithms are applied to the whitened data, and we need to transform the estimate to the original coordinate system afterwards. Further, all algorithms contain loops to update the single vectors or matrices iteratively. In the simulations we repeat these loops till the change in terms of the Frobenius norm (of the updated vector or matrix) is less than 10^{-6}, or a maximum number of 1 000 iterations is achieved. If there is no convergence after this maximum number of iterations we consider the run as non-convergent. All non-convergent runs are excluded from the performance results.

In Figure 1 we generated data from models (*i*)-(*iv*) with a sample size of $T = 10\,000$ and used sample autocovariances at lags $\tau = 1, \ldots, 10$ for joint diagonalization. All algorithms are initialized with the identity matrix – except `SOBIdef` which has an internal randomization for correct row selection.

In the abundant data situations (i)-(iii) all algorithms achieve comparably good performances, where `SOBIdef` is slightly slower in terms of runtime. In the more challenging data situation (iv) the estimates of `SOBIdef` show a decrease in performance and the runtime of `ACDC` increases. Note that `SOBIsym` and `SOBI` lead to exactly the same estimates after convergence, and this is true for any data situation.

4. Limiting distributions for SOBIdef and SOBIsym

The crucial strength of `SOBIdef` and `SOBIsym` is that we do not only get estimates for the mixing matrix but also know the asymptotic distribution of these estimates. Under general multivariate time series assumptions, the (un-)mixing matrix estimates derived from `SOBIdef` or `SOBIsym` converge in probability to a true (un-)mixing matrix with limiting multivariate normal distribution, and the limiting variances of the estimates can be calculated. In case of real data, where we do not know the underlying time series model, we can still estimate the finite-sample variances from the estimated source signals. In the following we shortly outline necessary assumptions and the findings from our earlier papers [11, 12]. In Section 5 we then use the finite-sample variances to decide whether single entries of the mixing matrix are zero.

In the BSS model with assumptions $(A1)$-$(A4)$ from Section 2 the separation into mixing matrix and unobservable process is unique only up to sign-changing permutations. For the remaining part we need to clear this unidentifiability since we want to compare mixing estimate and true mixing matrix on the level of single entries. Therefore, we replace $(A4)$ by the stronger assumption

$(A4)^*$ $\sum_k \lambda_{k1}^2 > \ldots > \sum_k \lambda_{kp}^2$, where $\lambda_{k1}, \ldots, \lambda_{kp}$ are the diagonal entries of the autocovariances of $\{\boldsymbol{z}(t)\}_{t\in\mathbb{Z}}$ at lags $k \in \mathbb{N}$, and $\omega_j' 1_p \geq 0$ for all columns ω_j of Ω, where 1_p denotes a p-dimensional vector with ones.

With this, the separation in the BSS model becomes unique. To assure $(A4)^*$ also for the `SOBIdef` and `SOBIsym` estimates we need to add some post-processing. Let $S_1, \ldots, S_K$ be sample autocovariances at lags $\tau_1, \ldots, \tau_K$ that we consider for joint diagonalization. Similarly to Section 2., we assume

that $(A4)^*$ still holds if we consider the finite set of lags $\{\tau_1, \ldots, \tau_K\}$ instead of $\mathbb{N}$. For the `SOBIsym` unmixing estimates we then sort the rows such that $\sum_k (\hat{\gamma}_j S_k \hat{\gamma}_j')^2$ is decreasing for $j = 1, \ldots, p$. For `SOBIdef` this step is included in the initialization process. In addition we may need to multiply single columns of the finite mixing estimates by -1 to assure positive column sums.

We now specify our assumptions about the unobservable process $\{\boldsymbol{z}(t)\}_{t\in\mathbb{Z}}$ and assume a p-variate MA(∞)-process, i. e.

$$\boldsymbol{z}(t) = \sum_{j=-\infty}^{\infty} \Psi_j \, \boldsymbol{\varepsilon}(t-j) \ , \quad \text{for } t \in \mathbb{Z} \ .$$

According to $(A2)$-$(A3)$, the matrices Ψ_j for $j \in \mathbb{Z}$ are diagonal and satisfy $\sum_{j=-\infty}^{\infty} \Psi_j^2 = I_p$, and we assume $\boldsymbol{\varepsilon}(t) \sim \mathcal{N}(0, I_p)$. Using Wold's decomposition [15] every second-order stationary process with normal components can be transformed into such an MA(∞)-process.

From now on we consider the `SOBIdef` and `SOBIsym` (un-)mixing estimates as p^2-variate random variables rather than concrete estimates and we use bold symbols $\hat{\boldsymbol{\Omega}} = (\hat{\boldsymbol{\omega}}_{ij})$ and $\hat{\boldsymbol{\Gamma}} = (\hat{\boldsymbol{\gamma}}_{ij})'$ for visual distinction. As before, let Ω denote the true mixing matrix and $\Gamma = \Omega^{-1}$ its inverse. In [11, 12] Miettinen et al. showed that $\sqrt{T}\mathrm{vec}(\hat{\boldsymbol{\Omega}} - \Omega)$ and $\sqrt{T}\mathrm{vec}(\hat{\boldsymbol{\Gamma}} - \Gamma)$ are asymptotically normally distributed with a mean vector zero. The covariance matrices $ASV(\hat{\boldsymbol{\Omega}})$ and $ASV(\hat{\boldsymbol{\Gamma}})$ depend on the autocovariances of the unobservable process and explicit formulas are available.

Given a finite sample $x(1), \ldots, x(T)$, the deflation-based or symmetric mixing estimate $\hat{\boldsymbol{\Omega}}$ is then approximately normally distributed as

$$\mathcal{N}(\mathrm{vec}(\Omega) \, , \frac{1}{T} ASV(\hat{\boldsymbol{\Omega}})) \ .$$

Since in general the true mixing matrix and the source model are unknown, we can approximate the distribution using the mixing estimate $\hat{\Gamma}$ and the estimated source signals $\hat{\Gamma}x(1), \ldots, \hat{\Gamma}x(T)$. To determine the variance we use sample autocovariances of the source estimates and consider lags from a finite subset of $\mathbb{N}$. In addition, infinite sums are approximated by finite sums. We denote the resulting finite-sample variance by $\widehat{ASV}(\hat{\boldsymbol{\Omega}})$. Functions to compute the asymptotic and the finite-sample variance can be found in the R-package 'BSSasymp' [16].

5. Identification of the mixing pattern

In applications of mixing models we sometimes face the question whether single source signals are present in a specific observation or not. In the several speakers problem, for example, we want to decide whether a single speakers' sound is recorded by a specific microphone. This question is related to the question whether single entries in the mixing matrix are zero: If the j-th source signal $\{s_j(t)\}_{t\in\mathbb{Z}}$ is not present in the i-th observation $\{x_i(t)\}_{t\in\mathbb{Z}}$ then the entry $\Omega(i,j)$ of the mixing matrix is zero. On the other hand, blind source separation algorithms typically estimate a dense matrix $\hat{\Omega}$ where no entry is exactly equal to zero. Simple thresholding implies the cruical choice of an appropriate cut-off and does not appear convincing. Another idea is to add a penalty term to the joint diagonalization problem (2). In the supplement we show that numerical optimization of such a penalized version fails in practice and we discuss the reasons. As an alternative, we perform informed pattern identification using the limiting distributions of the `SOBIdef` and `SOBIsym` estimates. With this, we can soundly decide whether close-to-zero-entries are actually zero. In Section 6 we present simulation results.

5.1. Pattern identification via hypothesis tests

First, we investigate hypothesis tests on linear combinations of the mixing entries. Let therefore $x(1),\dots,x(T)$ be observations of a BSS model with mixing matrix Ω. We consider a family of linear null hypotheses H_0: $A\operatorname{vec}(\Omega) = b$ and alternatives H_1: $A\operatorname{vec}(\Omega) \neq b$, where A is a $k \times p^2$ matrix and b is a k-vector. If the rows of A contain only one non-zero entry and this entry equals 1 we can test whether single entries of $\hat{\Omega}$ are different from zero. Under the above null hypothesis and with $\hat{\Omega}$ the `SOBIdef` or `SOBIsym` estimate we have

$$\sqrt{T}(A\operatorname{vec}(\hat{\Omega}) - b) \;\to_d\; \mathcal{N}_k(0, A(ASV(\hat{\mathbf{\Omega}}))A')$$

in distribution. This can be used in a test construction (still under H_0) as

$$M := (A\operatorname{vec}(\hat{\Omega}) - b)' \left(A(\frac{1}{T}\widehat{ASV}(\hat{\mathbf{\Omega}}))A'\right)^{-1}(A\operatorname{vec}(\hat{\Omega}) - b) \;\to_d\; \chi^2_k\,.$$

Here χ^2_k denotes the chi-squared distribution with k degrees of freedom which equals the number of linear equations in $A\operatorname{vec}(\Omega) = b$. If M is larger than the upper αth quantile of χ^2_k, we reject H_0 with (asymptotic) probability of false

alarm equal to α. Similar test statistics have been introduced by Ollia et al. [17] for the independent component model (ICA) and the fastICA estimate.

To determine the zero pattern of the mixing matrix we independently test $H_0 : \omega_{ij} = 0$ vs. $H_1 : \omega_{ij} \neq 0$ for all mixing entries. If H_0 is not rejected we assume that the corresponding entry is zero. This first approach for pattern identification is rather simplistic since the dependence structure of the mixing entries is not taken into account. In Section 6 we refer to it as *h-test*.

5.2. Pattern identification via information criteria

We now move on to information criteria to select between different zero-patterns of the mixing matrix. Let Ω^h denote a $p \times p$ matrix with zero entries at positions given in the set h. In the following we define a model for these *reduced* mixing matrices.

According to Section 4, the mixing estimate $\hat{\mathbf{\Omega}}$ is asymptotically normally distributed with the mean being the true mixing matrix and variance given by the limiting variance. Based on observations $x(1), \ldots, x(T)$ one can estimate the variance using the finite sample variance. A model for the (unknown) reduced mixing matrix Ω^h for any zero-pattern h is then given by

$$\mathcal{N}(\mathrm{vec}(\Omega^h)\,, \frac{1}{T}\widehat{ASV}(\hat{\mathbf{\Omega}}))\ ,$$

where $\widehat{ASV}(\hat{\mathbf{\Omega}})$ is the finite sample variance calculated from $x(1), \ldots, x(T)$. The number of model parameters equals the number of non-zero entries in Ω^h, and for $h = \emptyset$ we get the full model with p^2 parameters. The observations are given by the mixing estimate $\hat{\Omega}$ and with this the likelihood function is defined as $\ell(\Omega^h) = \ln f(\hat{\Omega}; \Omega^h)$. Finally, let $\hat{\Omega}^h = \mathrm{argmax}\ \ell(\Omega^h)$ denote the maximum likelihood estimate of the reduced mixing matrix.

To determine the most appropriate zero-pattern of the mixing matrix, we study information criteria of the form

$$IC(h) = -2\,\ell(\hat{\Omega}^h) + kc\ , \tag{4}$$

where h is any zero-pattern, $k = |h|$ denotes the number of model parameters, and c is some constant. For $c = 2$ the above equation yields the Akaike information criterion (AIC) and for $c = \ln(T)$ the equation yields the Bayesian information criterion (BIC) where T is the length of the observed time series. With this, we identify the lowest value $IC(h)$ among all zero-patterns and

the resulting h is the estimated zero-pattern for the mixing matrix. In the result part we refer to this approach as *AIC*, *BIC*, or *IC*.

In the above approach one needs to maximize the likelihood function for all zero-patterns h to determine the reduced estimate $\hat{\Omega}^h$. To save computational time we invent a more heuristic variant:

Since $\hat{\Omega}$ itself is a mixing estimate, the non-zero entries of $\hat{\Omega}^h$ will typically be close at the corresponding entries of $\hat{\Omega}$. Thus, we might directly set entries of $\hat{\Omega}$ to zero and leave all other entries unchanged. For a zero-pattern h we set $\hat{\Omega}^h(i,j) = \hat{\Omega}(i,j)$ for $(i,j) \notin h$ and zero otherwise. Note that this approach yields different estimates than before whenever $\widehat{ASV}(\hat{\mathbf{\Omega}})$ contains non-zero off-diagonal entries. Using this modified estimate $\hat{\Omega}^h$ we again determine the zero-pattern with the lowest IC value. We refer to this approach as *AICmod*, *BICmod*, or *ICmod*.

6. Simulations

In the following we first validate the test statistics from Section 5.1 and investigate the impact of noise. We then compare the three pattern identification methods IC, ICmod and h-test and consider mixing matrices with different numbers of zero entries.

We consider the AR(4)-model (i) from Section 3.1 and generate 3-dimensional data with mixing matrix $\Omega = I_3$. For $j = 1, 2, 3$ we then test the hypothesis $H_0^{(j)} : \omega_j = \mathrm{e}_j$ vs. $H_1^{(j)} : \omega_j \neq \mathrm{e}_j$, where e_j is the canonical unit vector with 1 at the jth component. In addition we consider the complete mixing matrix and test $H_0^{(all)} : \mathrm{vec}(\Omega) = \mathrm{vec}(I_3)$ vs. $H_1^{(all)} : \mathrm{vec}(\Omega) \neq \mathrm{vec}(I_3)$. For all these tests we can easily define a matrix A with entries in $\{0, 1\}$ such that $A\,\mathrm{vec}(\Omega) = \mathrm{e}_j$ $(j = 1, 2, 3)$ or $A\,\mathrm{vec}(\Omega) = \mathrm{vec}(I_3)$, respectively. In the first type of test the degrees of freedom of χ^2_k equal p in the latter p^2. Table 1 shows the percentage of (falsely) rejected null hypotheses at significance level 0.05 over 5 000 runs for a sample length of $T = 500, 1\,000, 10\,000$. We find a better identification for the first column of the estimate, but for $T = 10\,000$ all tests come close to the expected value of 5%.

We further address the question of how large entries of the mixing matrix must be such that they can be identified as non-zero. We therefore replace the previous mixing matrix, and we assume now that the first column of Ω is of the form $\omega_1 = (1, \varepsilon, 0)'$. All other entries are chosen randomly from the uniform distribution $\pm\mathcal{U}\,[0.1, 1.0]$ (U1) or $\pm\mathcal{U}\,[0.5, 1.0]$ (U2). Here, $\pm\mathcal{U}\,[a, b]$

for $0 < a < b$ denotes a uniform distribution with support $[-b, -a] \cup [a, b]$. We test $H_0^{(1)} : \omega_1 = \mathrm{e}_1$ vs. $H_1^{(1)} : \omega_1 \neq \mathrm{e}_1$ for increasing $\varepsilon = 0, 0.01, \ldots, 0.05$ and with 1 000 runs in each case. The percentage of correctly rejected null hypotheses increases with the value of ε and already at a value of $\varepsilon = 0.02$ we observe a rejection rate of 80% (Figure 2).

To perform pattern identification we generate data using the AR(4)-model (i) and fix the sample size at $T = 10\,000$. For the mixing matrix we consider the following four zero-patterns:

$$\Omega_1 = \begin{pmatrix} * & 0 & * \\ * & * & * \\ * & * & * \end{pmatrix}, \ \Omega_2 = \begin{pmatrix} * & 0 & * \\ * & * & 0 \\ * & * & * \end{pmatrix}, \ \Omega_3 = \begin{pmatrix} * & 0 & 0 \\ * & * & * \\ * & * & * \end{pmatrix}, \ \Omega_4 = \begin{pmatrix} * & 0 & 0 \\ 0 & * & * \\ 0 & * & * \end{pmatrix},$$

where (*) denotes the non-zero entries. In case 1, for example, the second source signal has no impact on the first observation, and in case 3 the first observation depends only on the first source signal. Let h_i denote the set of zero entries in each case, i.e. $h_1 = \{(1,2)\}$, $h_2 = \{(1,2),(2,3)\}$, $h_3 = \{(1,2),(1,3)\}$ and $h_4 = \{(1,2),(1,3),(2,1),(3,1)\}$. The non-zero entries of the mixing matrix are chosen randomly from the uniform distribution $\pm\mathcal{U}[0.1, 1.0]$.

The `SOBIdef` and `SOBIsym` mixing estimates and their distributions are based on sample autocovariances at lags $\tau = 1, \ldots, 10$. From these estimates we determine the most appropriate zero-patterns following the three approaches in Section 5. For evaluation we compare the zero entries of the true mixing matrix to those of the estimated pattern. Figure 3 shows the percentage of correctly determined patterns (filled areas) as well as the percentage of partly determined patterns (shaded areas), where not all or more zero-entries were detected. We considered 500 samples from time series model (i) with random mixing matrices. We found a crucial increase in performance if we used AIC/BIC with parameter maximization. In this case BIC determined nearly all zero-patterns correctly. Corresponding figures for the other time series models are added in the supplement.

We further investigated the impact of the information criterion constant c in (4) as well as the sample size T. Figure 4 shows the percentage of correctly determined zero-patterns for the `SOBIsym` estimate. The data was generated from time series models (i) and (iii) with a sample length of $T = 500, 1\,000, 10\,000$. We increased $c = 1, \ldots, 100$ where the BIC is given for $c = 6.2$, 6.9, and 9.2 depending on T. We find that in nearly all settings IC clearly outperforms the modified IC. In comparison to the h-test the

information criterion is only slightly better. The highest rates of correct zero detection are achieved for $c = \ln(T)$ (BIC). Furthermore, the performance depends on the underlying time series model; for the AR(4)-model (i) we find higher recovery rates compared to the mixed-model (iii). Results for the remaining time series models (ii) and (iv) can be found in the supplement.

7. Conclusions and outlook

In this paper we considered a second-order BSS model and discussed how one can select source signals that have a real impact on a specific component of the time series data. Until now, this was done setting arbitrary cut-off values below which entries of the mixing matrix are considered negligible. In contrast, we propose more sound methods. To that end, we focused on the recently published algorithms `SOBIdef` and `SOBIsym`, for which the distributions of the mixing estimates are known. In a comparison study we showed that `SOBIsym` provides a reasonable alternative to established algorithms in terms of performance and runtime. From the distribution of the estimates we derived methods to decide whether small entries of the mixing estimate are actually zero or not. In simulations with different time series models the Bayesian information criterion leads to the best reconstruction of the true zero-pattern. Our findings give insight whenever one observes a linear and instantaneous mixture of stationary time series signals and is interested in the source signals that are actually present in the data.

For further research it can be interesting to relax the model assumptions. Second-order source separation has for example been translated to the case of non-stationary signals [18]. Moreover, non-instantaneous (or convolutive) mixtures have been considered [19]. For the latter, the mixing model can be traced back to an instantaneous mixing. To adopt our model selection approaches to models with such relaxed assumptions one needs to reformulate the limiting variances of the mixing estimates.

Acknowledgements

This work was financially supported by the German Federal Ministry of Education and Research (BMBF) within the GerontoSys project "Stromal Aging" (Grant no. FKZ 0315576C), and the European Union within the ERC Grant "LatentCauses". Moreover, the work of Jari Miettinen, Sara Taskinen, Klaus Nordhausen and Hannu Oja was supported by the academy of Finland (Grant nos. 256291 and 268703).

References

[1] Tong, L., Soon, V.C., Huang, Y.F., Liu, R. (1990). AMUSE: a new blind identification algorithm. In: Proceedings of IEEE International Symposium on Circuits and Systems 1990, pp.1784–1787.

[2] Belouchrani, A., Abed-Meraim, K., Cardoso, J.-F., Moulines, E. (1997). A blind source separation technique using second-order statistics. IEEE Trans. on Signal Process. 45, pp.434–444.

[3] Ziehe, A., Müller, K.-R. (1998). TDSEP - an efficient algorithm for blind separation using time structure. In Proceedings of International Conference on Artificial Neural Networks (ICANN'98), pp.675–680.

[4] Yeredor, A. (2002). Non-orthogonal joint diagonalization in the least-squares sense with application in blind source separation. IEEE Trans. on Signal Process. 50, pp.1545–1553.

[5] Ziehe, A., Laskov, P., Müller, K.R., Nolte, G. (2003). A linear least-squares algorithm for joint diagonalization. In: Proceedings of 4th International Symposium on Independent Component Analysis and Blind Signal Separation (ICA2003), pp.469–474.

[6] Theis, F.J., Inouye, Y. (2006). On the use of joint diagonalization in blind signal processing. In: Proceedings of IEEE International Symposium on Circuits and Systems 2006.

[7] Theis, F.J., Müller, N.S., Plant, C., Böhm, C. (2010). Robust second-order source separation identifies experimental responses in biomedical imaging. In: Latent Variable Analysis and Signal Separation, pp.466–473.

[8] Theis, F.J., Gruber, P., Keck, I.R., Lang, E.W. (2008). A robust model for spatiotemporal dependencies. Neurocomputing, 71(10), pp.2209–2216.

[9] Joyce, C.A., Gorodnitsky, I.F., Kutas, M. (2004). Automatic removal of eye movement and blink artifacts from EEG data using blind component separation. Psychophysiology, 41(2), pp.313–325.

[10] Tang, A.C., Liu, J.Y., Sutherland, M.T. (2005). Recovery of correlated neuronal sources from EEG: the good and bad ways of using SOBI. Neuroimage, 28(2), pp.507–519.

[11] Miettinen, J., Nordhausen, K., Oja, H., Taskinen, S. (2014). Deflation-based separation of uncorrelated stationary time series. J. of Multivar. Anal. 123, pp.214–227.

[12] Miettinen, J., Illner, K., Nordhausen, K., Oja, H., Taskinen, S., Theis, F.J. (2015). Separation of uncorrelated stationary time series using autocovariance matrices. Submitted for publication.

[13] Fischer, G. (2005). Lineare Algebra. Springer-Verlag, Germany.

[14] Ilmonen, P., Nordhausen, K., Oja, H., Ollila, E. (2010). A new performance index for ICA: Properties, computation and asymptotic analysis. In: Latent Variable Analysis and Signal Separation, pp.229–236.

[15] Brockwell, P.J., Davis, R.A. (1991). Time Series: Theory and Methods. Second edition, Springer-Verlag, New York.

[16] Miettinen, J., Nordhausen, K., Oja, H., Taskinen, S. (2013). BSSasymp: Asymptotic Covariance Matrices of Some BSS Mixing and Unmixing Matrix Estimates. R-package version 1.0-0. http://cran.r-project.org/web/packages/BSSasymp.

[17] Ollila, E., Hyon-Jung, K. (2011). On testing hypotheses of mixing vectors in the ICA model using FastICA. IEEE International Symposium on Biomedical Imaging: From Nano to Macro, pp.325–328.

[18] Choi, S., Cichocki, A., Beloucharni, A. (2002). Second order nonstationary source separation. J. VLSI Signal Process. Syst. Signal Image Video Technol. 32(1-2), pp.93–104.

[19] Ye, Z., Chang, C., Wang, C., Zhao, J., Chan, F.H. (2003). Blind separation of convolutive mixtures based on second order and third order statistics. IEEE International Conference on Acoustics, Speech, and Signal Processing, 5, pp.305–308.

Table 1: Hypothesis tests on mixing entries. We generate data from the AR(4)-model (i) with a time series length of $T = 10\,000$. The true mixing matrix is chosen as identity matrix and for each column ω_j $(j=1,2,3)$ we test $H_0^{(j)}: \omega_j = e_j$ vs. $H_1^{(j)}: \omega_j \neq e_j$. In addition, we consider the complete mixing matrix and test $H_0^{(all)}: \mathrm{vec}(\Omega) = \mathrm{vec}(I_3)$ vs. $H_1^{(all)}: \mathrm{vec}(\Omega) \neq \mathrm{vec}(I_3)$. The table shows the percentage of (falsely) rejected null hypotheses at significance level 0.05 over 5 000 samples.

	`SOBIdef`				`SOBIsym`			
T	$H_0^{(1)}$	$H_0^{(2)}$	$H_0^{(3)}$	$H_0^{(all)}$	$H_0^{(1)}$	$H_0^{(2)}$	$H_0^{(3)}$	$H_0^{(all)}$
500	5.50	6.16	7.98	8.28	5.84	6.36	7.42	8.38
1 000	5.62	5.46	6.70	6.50	5.96	5.16	6.92	6.58
10 000	4.32	5.26	5.26	4.62	4.46	5.50	5.04	4.66

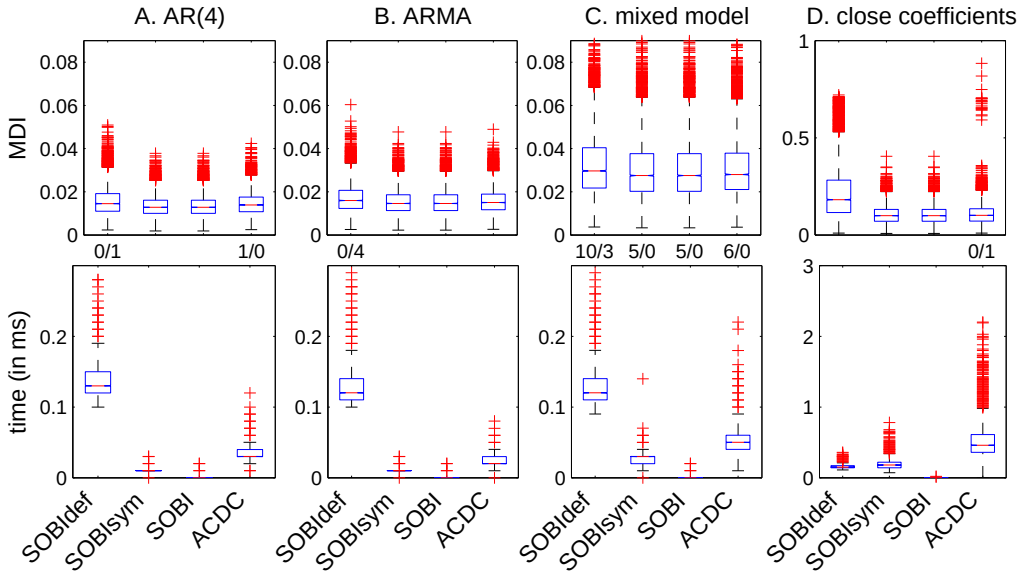

Figure 1: **Algorithm performance.** We compare `SOBIdef` and `SOBIsym` to the well-established methods `SOBI` and `ACDC`. Shown is the median MDI (upper plots) and the median runtime (lower plots) of the mixing estimates over 10 000 repetitions. The generate the data we considered in A. the AR(4)-model (*i*), in B. the ARMA-model (*ii*) , in C. the mixed model (*iii*), and in D. the close-coefficient model (*iv*). The sample size is fixed at $T = 10\,000$ and for joint diagonalization we used autocovariances at lags $\tau = 1, \ldots, 10$. All algorithms are initialized with the identity matrix. The numbers between the upper and lower plots indicate the counts of larger MDI/time values above the axis scaling.

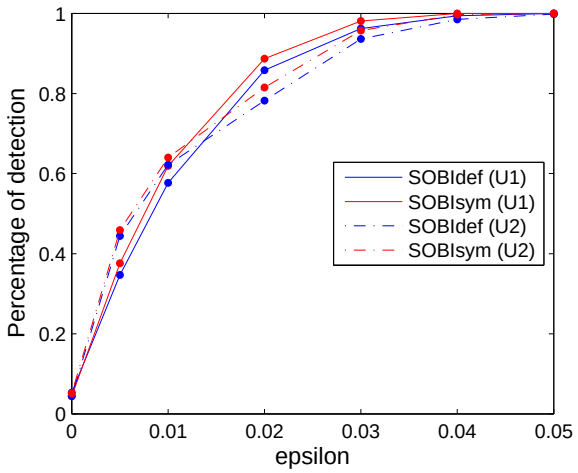

Figure 2: **Hypothesis tests on noisy mixing entries.** The first column of the mixing matrix is chosen as $\omega_1 = (1, \varepsilon, 0)$ for increasing ε, all other entries are randomly sampled from $U1 = \pm\mathcal{U}[0.1, 1.0]$ and $U2 = \pm\mathcal{U}[0.5, 1.0]$. The figure shows the percentage of rejected null hypothesis $H_0^{(1)} : \omega_1 = e_1$ at significance level 0.05. In case of $\varepsilon = 0$, this is a wrong decision, otherwise a correct one. We used 1 000 repetitions for each ε.

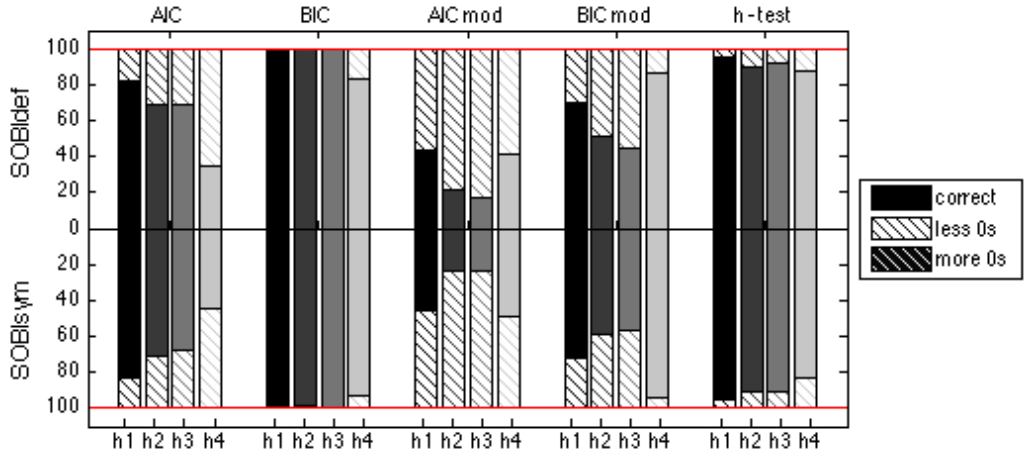

Figure 3: **Pattern identification to determine zero entries of the mixing matrix.** The data is generated using the AR(4)-model (i) with a time series length of $T = 10\,000$. The true mixing matrix contains zeros at positions $h_1 = \{(1,2)\}$, $h_2 = \{(1,2),(1,3)\}$, $h_3 = \{(1,2),(2,3)\}$ and $h_4 = \{(1,2),(1,3),(2,1),(3,1)\}$. We reconstruct these zero-patterns from the `SOBIdef` and `SOBIsym` mixing estimates using the selection methods AIC, BIC, AICmod, BICmod and h-test from Section 5. The percentages of correctly determined, under- and overdetermined patterns over 500 repetitions are shown as filled, shaded and dark shaded areas, respectively.

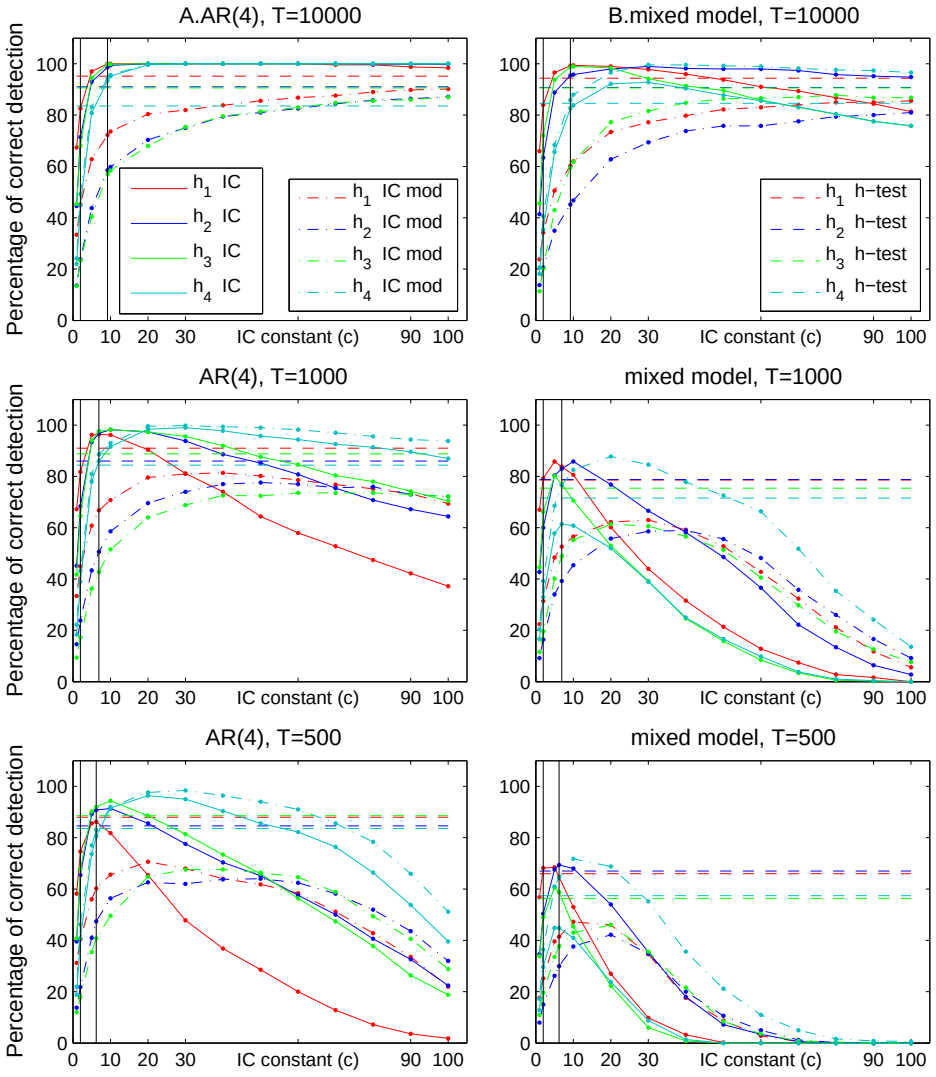

Figure 4: **Pattern identification for increasing constant** c**.** The data is generated using A. the AR(4)-model (i) and B. the mixed model (iii) with a time series length of $T = 10\,000$, $T = 1\,000$ and $T = 500$ in the single rows. The true mixing matrix contains zeros at positions $h_1 = \{(1,2)\}$, $h_2 = \{(1,2),(1,3)\}$, $h_3 = \{(1,2),(2,3)\}$ and $h_4 = \{(1,2),(1,3),(2,1),(3,1)\}$. We reconstruct these zero-patterns from the `SOBIsym` mixing estimates using the information criterion for increasing constant c (with and without maximization) and the h-test. The figure shows the percentage of correctly determined patterns over 500 repetitions. The black vertical lines indicate the constant values $c = 2$ (AIC) and $c = \ln(T)$ (BIC).