Sliced Average Variance Estimation for Multivariate Time Series

M. Matilainen[a*], C. Croux[b], K. Nordhausen[c] and H. Oja[d]

[a] *Department of Mathematics and Statistics, University of Turku, Turku, Finland and Turku PET Centre, Turku, Finland;*

[b] *EDHEC Business School, Lille, France;*

[c] *Institute of Statistics & Mathematical Methods in Economics, Vienna University of Technology, Austria;*

[d] *Department of Mathematics and Statistics, University of Turku, Turku, Finland*

Supervised dimension reduction for time series is challenging as there may be temporal dependence between the response y and the predictors $\mathbf{x}$. Recently a time series version of sliced inverse regression, TSIR, was suggested, which applies approximate joint diagonalization of several supervised lagged covariance matrices to consider the temporal nature of the data. In this paper we develop this concept further and propose a time series version of sliced average variance estimation, TSAVE. As both TSIR and TSAVE have their own advantages and disadvantages, we consider furthermore a hybrid version of TSIR and TSAVE. Based on examples and simulations we demonstrate and evaluate the differences between the three methods and show also that they are superior to apply their iid counterparts to when also using lagged values of the explaining variables as predictors.

Keywords: blind source separation, supervised dimension reduction, prediction

AMS Subject Classification: 62M10

1. Introduction

Linear supervised dimension reduction has a long tradition for independent and identically distributed (iid) observations with a rich literature reviewed for example in [1]. The idea is to find all linear combinations of a predictor vector $\mathbf{x}$ which are needed to model a response y even when the true functional relationship between the response and explaining variables is not known. In multivariate time series context with temporal dependence, the goal is similarly to model a response time series value y_t at t as a function of the previous history of a stationary multivariate predictor time series $(\mathbf{x}_{t-j})_{j=1,2,\ldots}$. Popular time series dimension reduction methods such as those based on (dynamic) factor models, reviewed for example in [2], are not supervised, and supervised dimension reduction methods are still rare in the literature. Both Xia et al. [3] and Becker and Fried [4] propose the use of standard supervised iid dimension reduction methods simply by explaining a response series value y_t with a vector of lagged predictor time series values in $\mathbf{x}_{t-1}, \ldots, \mathbf{x}_{t-j}$. This may then increase the dimension of the problem dramatically and at the same time reduces the sample size. Barbarino and Bura [5, 6] combine ideas from factor models and standard iid supervised dimension reduction methods.

*Corresponding author. Email: markus.matilainen@utu.fi

Recently, Matilainen et al. [7] proposed a procedure that finds most relevant linear combinations of the predictor series with their most relevant lags when modelling the response series. The approach, called TSIR, is an extension of the sliced inverse regression (SIR), introduced by Li [8], and is based on the approximate joint diagonalization of the covariance matrices of conditional expected values $\mathrm{E}(\mathbf{x}_t|y_{t+j})$, $j = 1, 2, \ldots$ which naturally consider the temporal nature of the data. Considering for example for a p-variate explaining time series $\mathbf{x}_t$, k lags means to jointly approximately diagonalize $k+1$ $p \times p$ matrices while in approaches like [3] and [4] two $p(k+1) \times p(k+1)$ matrices need to be simultaneously diagonalized.

In sliced average variance estimation (SAVE; Cook and Weisberg [9, 10]) for iid observations, one considers the variation of conditional covariance $\mathrm{COV}(\mathbf{x}|y)$ rather than the variation of the conditional expectation $\mathrm{E}(\mathbf{x}|y)$ to detect better the cases of nonlinear dependence. In this paper we suggest the similar use of $\mathrm{COV}(\mathbf{x}_t|y_{t+j})$, $j = 1, 2, \ldots$, in a time series context. This is a time series extension of SAVE, and is called here TSAVE. As TSIR and TSAVE have their own specific drawbacks, a hybrid of TSIR and TSAVE, denoted as TSSH, is also introduced. It can be seen as a weighted combination of TSIR and TSAVE generalizing the hybrid in Zhu et al. [11] to the time series context.

One aim here is also to see whether in a time series context the number of slices and the weight coefficient of the hybrid have similar preferred values as their iid counterparts. This was also not investigated in [7] for TSIR and the number of slices was just assumed to be the same as in the iid case. We will also investigate if these tuning parameters depend much on the underlying stochastic processes.

The structure of the paper is as follows. We first recall SIR and SAVE for iid data. Then we move to the time series context, where first TSIR is reviewed and then in Section 3.3 TSAVE and in Section 3.4 also the hybrid of TSIR and TSAVE are introduced. Section 4 then includes examples and simulation studies. In Section 4.3 we conduct a simulation study to find some guidelines to how many slices we need in practice to estimate the matrices that we approximately jointly diagonalize. Then in Section 4.4 we have another simulation study to find the appropriate weights of TSIR and TSAVE parts for method TSSH and show the hybrid can sometimes be more efficient than these methods separately. Finally in 4.5 we show that also TSAVE is often better than TSIR and that that both methods beat their iid counterparts applied to time series, such as the method in [4].

2. Supervised dimension reduction for iid data

In this section we review iid supervised dimension reduction methods SIR, SAVE and their hybrid version. We formulate the supervised dimension reduction problem as an estimation problem in a blind source separation (BSS) model for the joint distribution of the response variable y and the p-variate vector of observable explaining variables $\mathbf{x}$. The BSS model then assumes that

$$\mathbf{x} = \mathbf{\Omega}\mathbf{z} + \boldsymbol{\mu}, \tag{1}$$

where the full rank $p \times p$ matrix $\mathbf{\Omega}$ is called the mixing matrix and $\boldsymbol{\mu}$ is the location p-vector. The latent p-vector $\mathbf{z}$ can be partitioned as $\mathbf{z} = \left(\mathbf{z^{(1)}}^\top, \mathbf{z^{(2)}}^\top\right)^\top$ with the respective dimensions k and $p-k$, and

(I1) $\mathrm{E}(\mathbf{z}) = \mathbf{0}$ and $\mathrm{COV}(\mathbf{z}) = \mathbf{I}_p$ and

(I2) $(y, \mathbf{z}^{(1)\top})^\top \perp\!\!\!\perp \mathbf{z}^{(2)}$.

Hence in this model $\mathbf{z}^{(1)}$ carries all the information needed to model the response y and $\mathbf{z}^{(2)}$ can be considered as the noise part. Note that assumption (I2) made in this paper for both SIR and SAVE is slightly stronger than those made in the original papers, i.e.

$$\begin{aligned}
&\mathbf{z}^{(2)} \perp\!\!\!\perp y|\mathbf{z}^{(1)},\\
&\mathrm{E}(\mathbf{z}_t^{(2)}|\mathbf{z}_t^{(1)}) = \mathbf{0} \text{ (a.s.) (for SIR) and}\\
&\mathrm{COV}(\mathbf{z}_t^{(2)}|\mathbf{z}_t^{(1)}) = \mathbf{I}_{p-k} \text{ (a.s.) (for SAVE).}
\end{aligned}$$

The assumption (I2) implies all these three assumptions and is needed in [12] to build asymptotic and bootstrap tests for the true subspace dimension k in the SIR methodology.

As there may be several partitions of $\mathbf{z}$ fulfilling (I1) and (I2), we choose the one with the smallest k. The aim is to find a $k \times p$ unmixing matrix $\mathbf{\Gamma}$ such that $\mathbf{\Gamma}\mathbf{x} = \mathbf{z}^{(1)}$ up to a pre-multiplication by a $k \times k$ orthogonal matrix. Note that the latent $\mathbf{z}^{(1)}$ as stated in our model has the same indeterminacy.

A direct consequence of assuming this model is the following.

RESULT 2.1 *Let y denote the response and $\mathbf{z}$ have the properties as stated in (I1) and (I2). Then*

$$\mathrm{COV}[\mathrm{E}(\mathbf{z}|y)] = \begin{pmatrix} \mathrm{COV}[\mathrm{E}(\mathbf{z}^{(1)}|y)] & 0 \\ 0 & 0 \end{pmatrix}$$

and

$$\mathrm{E}[(\mathbf{I}_p - \mathrm{COV}(\mathbf{z}|y))^2] = \begin{pmatrix} \mathrm{E}[(\mathbf{I}_k - \mathrm{COV}(\mathbf{z}^{(1)}|y))^2] & 0 \\ 0 & 0 \end{pmatrix}.$$

As in [13], $\mathbf{x} = \mathbf{\Omega}\mathbf{z} + \boldsymbol{\mu}$ implies that there exists an orthogonal matrix $\mathbf{U}$ such that

$$\mathbf{z} = \mathbf{U}\ \mathrm{COV}(\mathbf{x})^{-1/2}(\mathbf{x} - \mathrm{E}(\mathbf{x})). \tag{2}$$

Then it is shown in [7] that, based upon Result 2.1 (the first equation) and (2), one can define an unmixing matrix in the sliced inverse regression (SIR) [8] using the following steps.

Definition 2.2 The SIR functional $\mathbf{\Gamma}_{SIR}(\mathbf{x}; y)$ is defined as follows.

(1) Consider the standardized variable $\mathbf{x}^{st} := \mathrm{COV}(\mathbf{x})^{-1/2}(\mathbf{x} - \mathrm{E}(\mathbf{x}))$.
(2) Find the $k \times p$ matrix $\mathbf{W}_{SIR} = (\mathbf{w}_1, \ldots, \mathbf{w}_k)^\top$ with orthonormal rows $\mathbf{w}_1, \ldots, \mathbf{w}_k$ which maximizes

$$\sum_{i=1}^{k} \left[\mathbf{w}_i^\top \mathrm{COV}[\mathrm{E}(\mathbf{x}^{st}|y)]\mathbf{w}_i\right]^2.$$

(3) $\mathbf{\Gamma}_{SIR}(\mathbf{x}; y) := \mathbf{W}_{SIR}\mathrm{COV}(\mathbf{x})^{-1/2}$.

Assume now that $\mathbf{V}_1\mathbf{\Lambda}_1\mathbf{V}_1^\top$ is the eigenvector-eigenvalue decomposition of $\mathrm{E}[(\mathbf{I}_k -$

$\mathrm{COV}(\mathbf{z}^{(1)}|y))^2]$. Let $\mathbf{U}_1$ be the $p \times k$ matrix consisting of the first k columns of $\mathbf{U}$. Based upon the second equation of Result 2.1 and (2)

$$\begin{aligned}
\mathrm{E}[(\mathbf{I}_p - \mathrm{COV}(\mathbf{x}^{st}|y))^2] &= \mathrm{E}[(\mathbf{I}_p - \mathbf{U}\mathrm{COV}(\mathbf{z}|y)\mathbf{U}^\top)^2] \\
&= \mathbf{U}\mathrm{E}[(\mathbf{I}_p - \mathrm{COV}(\mathbf{z}|y))^2]\mathbf{U}^\top \\
&= \mathbf{U}_1\mathrm{E}[(\mathbf{I}_k - \mathrm{COV}(\mathbf{z}^{(1)}|y))^2]\mathbf{U}_1^\top \\
&= \boldsymbol{U}_1\mathbf{V}_1\boldsymbol{\Lambda}_1\mathbf{V}_1^\top\mathbf{U}_1^\top.
\end{aligned}$$

Write now $\boldsymbol{W}_{SAVE} = (\boldsymbol{U}_1\boldsymbol{V}_1)^\top$. Then $\mathbf{W}_{SAVE}\mathrm{E}[(\mathbf{I}_p - \mathrm{COV}(\mathbf{x}^{st}|y))^2]\mathbf{W}_{SAVE}^\top = \boldsymbol{\Lambda}_1$ is diagonal. The sliced average variance estimation (SAVE) [9, 10] then has the following steps.

Definition 2.3 The SAVE functional $\boldsymbol{\Gamma}_{SAVE}(\mathbf{x}; y)$ is defined as follows.

(1) Consider the standardized variable $\mathbf{x}^{st} := \mathrm{COV}(\mathbf{x})^{-1/2}(\mathbf{x} - \mathrm{E}(\mathbf{x}))$.
(2) Find the $k \times p$ matrix $\mathbf{W}_{SAVE} = (\mathbf{w}_1, \ldots, \mathbf{w}_k)^\top$ with orthonormal rows $\mathbf{w}_1, \ldots, \mathbf{w}_k$ that maximizes

$$\sum_{i=1}^{k} \left[\mathbf{w}_i^\top \mathrm{E}[(\mathbf{I}_p - \mathrm{COV}(\mathbf{x}^{st}|y))^2]\mathbf{w}_i\right]^2.$$

(3) $\boldsymbol{\Gamma}_{SAVE}(\mathbf{x}; y) = \mathbf{W}_{SAVE}\mathrm{COV}(\mathbf{x})^{-1/2}$.

For a random sample $(y_1, \mathbf{x}_1), \ldots, (y_n, \mathbf{x}_n)$, the population values of the two functionals can be consistently estimated if y is discrete with a finite number of values. In practice, the continuous y is then often replaced by its discretized version utilizing H disjoint slices $S_1, ..., S_H$, $S_1 + \ldots + S_H = \mathbb{R}$. One can for example define a discrete variable $y^{sl} \in \{1, ..., H\}$ by the condition $y^{sl} = h \Leftrightarrow y \in S_h$, $h = 1, ..., H$. Note that the condition (I2) and the model still holds true for $(y^{sl}, \mathbf{x})$, but the dimension k and the functionals may change.

Remark 2.4 Both SIR and SAVE can be seen as an approach which jointly diagonalizes two matrices [14], the regular covariance matrix and the supervised matrices $\mathrm{COV}[\mathrm{E}(\mathbf{x}|y)]$ or $\mathrm{E}[(\mathbf{I}_p - \mathrm{COV}(\mathbf{x})^{-1/2}\mathrm{COV}(\mathbf{x}|y)\mathrm{COV}(\mathbf{x})^{-1/2})^2]$, respectively. Hence, both methods can be solved using a generalized eigenvector-eigenvalue decomposition, where under model (1) there are $\hat{k} \leq k$ non-zero eigenvalues, and the functional is given by the corresponding generalized eigenvectors. Hence the functionals are unique if all non-zero eigenvalues are distinct.

Remark 2.5 The estimation of the dimension of the subspace k has some issues. For example the number of slices H can change the estimated value for the subspace, see e.g. [15]. However, the block diagonal structures in Result 2.1 still exist for different values of H, but the block sizes may be different. Also the method used may not find the whole subspace. In case of SIR for example when y is a quadratic function of a component, SIR fails to capture it (see e.g. [9]). In general SAVE is able to capture a larger portion of the subspace than SIR, as Cook and Critchley have shown in [16]. Due to these issues, it is possible that $\hat{k} < k$.

In the practical data analysis with unknown k, the estimated eigenvalues and the variation of the eigenvectors have been used to estimate k, see for example especially

in the context of SIR, see [12, 17, 18] and the references therein. The magnitude of the eigenvalues indicates the relevance of the corresponding source to model the response.

As [16] show, SAVE is in general considered more comprehensive when estimating the subspace of interest and SIR can be seen in certain situations as a special case. This increased flexibility of SAVE is however considered costly and it is usually said that SAVE needs more data than SIR [16]. This is also reflected when considering the numbers of slices used.

For SIR the slices are often chosen so that $\mathbb{P}(y \in \mathbb{S}_h) = 1/H$, $h = 1, \ldots, H$, with $H = 10$. In simulations in [8] it was shown that SIR is not very sensitive to the choice of H. The rank of $\mathrm{COV}(\mathrm{E}(\mathbf{x}|y^{sl}))$ and the maximum number of non-zero eigenvalues is $H - 1$, which however gives the restriction $H > \hat{k} + 1$.

SAVE, unlike SIR, is more sensitive to the choice of H as it uses higher moments and therefore needs more observations per slice than SIR, see for example [10, 16, 19]. Zhu et al. have conducted some simulations for SAVE [11]. With a data length of $n = 480$, SAVE with $H = 6$ still produces proper results in all of their settings, but with $H = 24$ not anymore. However, we should note that these results are based only on some specific simulation settings.

Asymptotic properties of SAVE estimator $\mathbf{W}_{SAVE}$ have been investigated e.g. in [20] in order to find a way an estimate of the subspace dimension k. Li and Zhu [19] have examined the consistency of the SAVE estimator. SAVE can achieve consistency and in the case where the response is discrete and takes finite values, SAVE can also achieve $\sqrt{n}$ consistency. However, generally SAVE cannot achieve this, unlike SIR. For asymptotic properties of the SIR estimator, including $\sqrt{n}$ consistency and asymptotic normality of the estimator, see [21].

In [16] it is argued that SAVE with sufficient data is superior to SIR but in practice it would be better to use both and complement them to uncover the structures of interest. A hybrid method based on SIR and SAVE, using a convex combination $(1-a)\mathrm{COV}(\mathrm{E}(\mathbf{x}^{st}|y^{sl})) + a\mathrm{E}[(\mathbf{I}_p - \mathrm{COV}(\mathbf{x}^{st}|y^{sl}))^2]$, with $a \in [0, 1]$, has been proposed. It is discussed first briefly in [22] and then more closely with the discussion of the choice of the coefficient a in [11]. In Section 3.4 we introduce a time series version of this hybrid method.

Also [23] have combined the strengths of SIR and SAVE by suggesting the $SAVE|SIR$ method. As SIR is efficient in finding the linear relationships, it can be used to find a partial dimension reduction subspace and SAVE is then used to find the rest.

Note also that SIRα, mentioned already in the rejoinder of Li's SIR paper [8] and developed further in [24] and [25], is the first kind of hybrid method of the first and second moments in supervised dimension reduction. However, this is not a combination of SIR and SAVE.

As recently [7] extended SIR to the time series framework it is therefore natural also to extend SAVE, which will be done in the following sections.

3. Linear supervised dimension reduction for time series

3.1. *The blind source separation model for linear supervised dimension reduction for time series*

Assume that $y = (y_t)_{t\in\mathbb{Z}}$ and $\mathbf{x} = (\mathbf{x}_t)_{t\in\mathbb{Z}}$ are (weakly and jointly) stationary univariate and p-variate time series, respectively. In this paper the term time series is used for both the observed realizations and the stochastic process producing them.

In the time series prediction problem, it is usually assumed that the response series y at time t is an unspecified function of $\mathbf{x}_t, \mathbf{x}_{t-1}, ...$ and $\epsilon_t, \epsilon_{t-1}, ...$ where $\epsilon = (\epsilon_t)_{t\in\mathbb{Z}}$ is an unspecified stationary noise process independent from $\mathbf{x}$, i.e.,

$$y_t = f(\mathbf{x}_t, \mathbf{x}_{t-1}, \ldots; \epsilon_t, \epsilon_{t-1}, \ldots).$$

As in the iid case we assume the blind source separation (BSS) model which states that only $k \ll p$ linear combinations of $\mathbf{x}$ are needed in the prediction model. In the following, if $\mathbf{A}$ is a $k \times p$ matrix and $\mathbf{b}$ a k-vector, $\mathbf{Ax} + \mathbf{b}$ is a k-variate time series with the value $\mathbf{Ax}_t + \mathbf{b}$ at t. In the time series BSS model, we assume that

$$\mathbf{x} = \mathbf{\Omega z} + \boldsymbol{\mu},$$

where $\mathbf{\Omega}$ is a full rank $p \times p$ mixing matrix and $\boldsymbol{\mu}$ is a location vector. Furthermore, just like in the iid case, the stationary p-variate source time series $\mathbf{z}$ can be partitioned as $\mathbf{z} = \left(\mathbf{z^{(1)}}^\top, \mathbf{z^{(2)}}^\top\right)^\top$ with the dimensions k and $p-k$ of the subseries, respectively. Dimension k is the smallest one to fulfil the conditions

(T1) $\mathrm{E}(\mathbf{z}_t) = \mathbf{0}$ and $\mathrm{COV}(\mathbf{z}_t) = \mathbf{I}_p$ and
(T2) $(y, \mathbf{z^{(1)}}^\top)^\top \perp\!\!\!\perp \mathbf{z}^{(2)}$.

As in Section 2, from (T2) it follows therefore that

$$\begin{aligned}
&\mathbf{z}^{(2)} \perp\!\!\!\perp y | \mathbf{z}^{(1)}, \\
&\mathrm{E}(\mathbf{z}^{(2)}_{t+s} | \mathbf{z}^{(1)}_t) = \mathbf{0} \text{ (a.s.) for all } s \in \mathbb{Z} \text{ and} \\
&\mathrm{COV}(\mathbf{z}^{(2)}_{t+s} | \mathbf{z}^{(1)}_t) = \mathbf{I}_{p-k} \text{ (a.s.) for all } s \in \mathbb{Z}.
\end{aligned}$$

All the information needed to model y is therefore contained in the process $\mathbf{z}^{(1)}$ and one can write

$$y_t = f(\mathbf{x}_t, \mathbf{x}_{t-1}, \ldots; \epsilon_t, \epsilon_{t-1}, \ldots) = f_0(\mathbf{z}^{(1)}_t, \mathbf{z}^{(1)}_{t-1}, \ldots; \epsilon_t, \epsilon_{t-1}, \ldots) \tag{3}$$

with another unspecified function f_0, possibly depending on $\mathbf{\Omega}$ and $\boldsymbol{\mu}$.

Also in this time series case the model is ill-defined in the sense that both $\mathbf{z}^{(1)}$ and $\mathbf{z}^{(2)}$ can be multiplied by respective orthogonal matrices and they still fulfil (T1) and (T2). The goal is therefore the estimation of the unmixing matrix $\mathbf{\Gamma}$ such that $\mathbf{\Gamma x} = \mathbf{z}^{(1)}$ up to orthogonal transformations. The function f_0 should then be parametrized to allow all linear combinations of the elements of $\mathbf{z}^{(1)}_t$, for example. One should also identify which lagged values $\mathbf{z}^{(1)}_t, \mathbf{z}^{(1)}_{t-1}, ...$ contribute in the model.

In this section, the TSIR method from [7] is first reviewed and then the methods TSAVE and TSSH, a combination of TSIR and TSAVE are introduced. Finally we recall the choosing of the number of important sources and lags mentioned in [7].

3.2. *SIR for time series*

In [7] the sliced inverse regression for time series uses the matrices

$$G_{0,j}(\mathbf{z}, y) = \mathrm{COV}(\mathrm{E}(\mathbf{z}_t | y_{t+j})), \quad j \in \mathbf{Z}_+$$

with the following important property.

RESULT 3.1 *Under (T1) and (T2)*

$$\mathrm{COV}[\mathrm{E}(\mathbf{z}_t|y_{t+j})] = \begin{pmatrix} \mathrm{COV}[\mathrm{E}(\mathbf{z}_t^{(1)}|y_{t+j})] & 0 \\ 0 & 0 \end{pmatrix},$$

for all lags $j \in \mathbf{Z}_+$.

Based on Result 3.1, [7] then finds the time series sliced inverse regression (TSIR) estimate of the unmixing matrix with the following three steps.

Definition 3.2 The TSIR functional $\mathbf{\Gamma}_{TSIR}(\mathbf{x}; y)$ is defined as follows.

1. Find $\mathbf{x}^{st} := \mathrm{COV}(\mathbf{x})^{-1/2}(\mathbf{x} - \mathrm{E}(\mathbf{x}))$.
2. Find the $k \times p$ matrix $\mathbf{W}_{TSIR} = (\mathbf{w}_1, \ldots, \mathbf{w}_k)^\top$ with orthonormal rows $\mathbf{w}_1, \ldots, \mathbf{w}_k$ which maximizes

$$\sum_{j \in S} \sum_{i=1}^{k} \left[\mathbf{w}_i^\top G_{0,j}(\mathbf{x}^{st}, y) \mathbf{w}_i \right]^2, \tag{4}$$

   for a chosen set of lags $S = \{S_1, \ldots, S_s\}$ with $S_j \geq 1$.
3. $\mathbf{\Gamma}_{TSIR}(\mathbf{x}; y) = \mathbf{W}_{TSIR}\mathrm{COV}(\mathbf{x})^{-1/2}$.

The matrix $\mathbf{W}_{TSIR}$ is a $k \times p$ matrix. If the approach for time series used in [3] and [4] were applied here, this matrix would be of size $k \times (|S| + 1)p$. This would make the method less stable when the number of time series and the number of lags used increases.

3.3. *SAVE for time series*

To make a time series version of SAVE a natural extension for the matrix of interest is

$$G_{1,j}(\mathbf{z}, y) = \mathrm{E}((\mathbf{I}_p - \mathrm{COV}(\mathbf{z}_t|y_{t+j}))^2), \quad j \in \mathbf{Z}_+$$

that depends a joint distribution of y and $\mathbf{x}$. We then have the following.

RESULT 3.3 *Under (T1) and (T2)*

$$\mathrm{E}((\mathbf{I}_p - \mathrm{COV}(\mathbf{z}_t|y_{t+j}))^2) = \begin{pmatrix} E((\mathbf{I}_p - \mathrm{COV}(\mathbf{z}_t^{(1)}|y_{t+j}))^2) & 0 \\ 0 & 0 \end{pmatrix},$$

for all lags $j \in \mathbf{Z}_+$.

The following unmixing matrix estimate is then called the time series sliced average variance estimator (TSAVE) functional:

Definition 3.4 The TSAVE functional $\mathbf{\Gamma}_{TSAVE}(\mathbf{x}; y)$ is defined as follows.

1. Find $\mathbf{x}^{st} := \mathrm{COV}(\mathbf{x})^{-1/2}(\mathbf{x} - \mathrm{E}(\mathbf{x}))$.
2. Find the $k \times p$ matrix $\mathbf{W}_{TSAVE} = (\mathbf{w}_1, \ldots, \mathbf{w}_k)^\top$ with orthonormal rows $\mathbf{w}_1, \ldots, \mathbf{w}_k$

that maximizes

$$\sum_{j\in S}\sum_{i=1}^{k}\left[\mathbf{w}_i^\top G_{1,j}(\mathbf{x}^{st},y)\mathbf{w}_i\right]^2, \tag{5}$$

for a chosen set of lags $S=\{S_1,\ldots,S_s\}$ with $S_j\geq 1$.

3. $\mathbf{\Gamma}_{TSAVE}(\mathbf{x};y)=\mathbf{W}_{TSAVE}\mathrm{COV}(\mathbf{x})^{-1/2}$.

TSIR and TSAVE can be seen as procedures which jointly diagonalize $|S|+1$ matrices (in terms of the Frobenius norm), that is, the covariance matrix of $\mathbf{x}_t$ and $|S|$ matrices depending on the joint distributions of $\mathbf{x}$ and y with lags in S. Under the blind source separation model assumed in this paper all the matrices of interest can be jointly diagonalized. However, for finite data this can be done only approximately and it has to be solved using algorithms using some objective criterion, as for example stated in the algorithmic outline above. While many other criteria are possible and many algorithms exist in the literature, for practical purposes we use the approach based on Jacobi rotations [26] in search for the matrices $\mathbf{W}_{TSIR}$ and $\mathbf{W}_{TSAVE}$ that maximize (4) and (5), respectively. This algorithm was recommended in [27], and for more details about joint diagonalization in BSS see for example [13, 28–33] and the references therein. In the following we will not distinguish between joint diagonalization and joint approximate diagonalization.

Given a solution $\mathbf{W}=(\mathbf{w}_1,...,\mathbf{w}_k)'$, the maximum value of the criterion function is $\sum_{j\in S}\sum_{i=1}^{k}\lambda_{ij}$, where

$$\lambda_{ij}=\left(\mathbf{w}_i^\top \mathrm{E}[(\mathbf{I}_p-\mathrm{COV}(\mathbf{x}_t^{st}|y_{t+j}))^2]\mathbf{w}_i\right)^2,$$

in a sense that it measures the contribution of the sth lag of the ith linear combination to this maximum value, $i=1,\ldots,k$; $j\in S$. $\mathbf{\Gamma}_{TSAVE}(\mathbf{x};y)$ is unique if $\lambda_{i\cdot}=\sum_{j\in\mathcal{S}}\lambda_{ij}$, $i=1,\ldots,k$ are distinct. The components $\mathbf{\Gamma}_{TSAVE}(\mathbf{x};y)\mathbf{x}$ are standardized and can be naturally ordered so that $\lambda_{1\cdot}\geq\ldots\geq\lambda_{k\cdot}$. The large value $\lambda_{i\cdot}$ indicates a strong dependence between the time series $(\mathbf{\Gamma}(\mathbf{x};y)\mathbf{x})_i$ and y. The higher the value of λ_{ij}, the stronger is the dependence between $(\mathbf{\Gamma}(\mathbf{x};y)\mathbf{x})_{it}$ and y_{t+j}. Identifying the relevant sources and lags is however difficult due to the possible serial correlations in $\mathbf{x}$ which means that λ_{ij} might vanish only slowly to zero with s for irrelevant lags.

As for SAVE, the unmixing matrix estimate is obtained for the sliced version $\mathbf{\Gamma}_{TSAVE}(\mathbf{x};y^{sl})$, where y^{sl} is a discrete time series such that $y_t^{sl}=h\Leftrightarrow y_t\in S_h$, $h=1,...,H$. As with SAVE it can be also here concluded that TSAVE will be more sensitive to the number of slices H as also TSAVE, just like SAVE, estimates a higher order moment and therefore needs more information (see Section 4.3).

Consider the following important property of TSAVE.

RESULT 3.5 *Let $\mathbf{x}^*=\mathbf{A}\mathbf{x}+\mathbf{b}$, where $\mathbf{A}$ is a full rank $p\times p$ matrix and $\mathbf{b}$ a p-vector. TSAVE is affine equivariant in the sense that, for all transformed time series $\mathbf{x}^*$, $\mathbf{\Gamma}_{TSAVE}(\mathbf{x}^*;y)\mathbf{x}^*=\mathbf{\Gamma}_{TSAVE}(\mathbf{x};y)\mathbf{x}$ up to the signs and the location of the component series.*

Result 3.5 also means that $\mathbf{\Gamma}_{TSAVE}(\mathbf{x}^*;y)=\mathbf{J}\mathbf{\Gamma}_{TSAVE}(\mathbf{x};y)\mathbf{A}^{-1}$, where $\mathbf{J}$ is a $p\times p$ diagonal matrix with diagonal elements ±1, up to the location. The proof is straightforward and hence is omitted from here.

To derive asymptotic properties of TSAVE the challenge consists of deriving the joint limiting distributions of $\sqrt{T}(\widehat{\mathrm{COV}}(\mathbf{x}) - \mathrm{COV}(\mathbf{x}))$ and $\sqrt{T}(\hat{G}_{1,j}(\mathbf{x}, y) - G_{1,j}(\mathbf{x}, y))$ for all lags $j \in S$ for which probably stronger assumptions on the process $\mathbf{x}_t$ need to be made. Therefore this is beyond the scope of this paper and we just outline how the asymptotics could be derived given the joint distribution of these matrices and that the signal dimension k would be known.

Write $\mathbf{G}_j := G_{1,j}(\mathbf{x}, y)$, for all $j = 1, \ldots, S$. The maximization (5) can also be written as

$$\sum_{j \in S} \sum_{i=1}^{k} \left[\mathbf{w}_i^\top \mathbf{G}_j^* \mathbf{w}_i \right]^2, \tag{6}$$

where $\mathbf{G}_j^* = \mathrm{COV}(\mathbf{x})^{-1/2} \mathbf{G}_j \mathrm{COV}(\mathbf{x})^{-1/2}$. Denote then $\mathbf{\Gamma} := \mathbf{\Gamma}_{TSAVE}$ and $\mathbf{M} = M(\mathbf{W}) = ((m(\mathbf{w}_i))_{i=1,\ldots,k}^\top$, where $m(\mathbf{w}_i) = \sum_{j=1}^{\hat{s}} \left[\mathbf{w}_i^\top \mathbf{G}_j^* \mathbf{w}_i \right] \mathbf{G}_j^* \mathbf{w}_i$, for $i = 1, \ldots, k$. Now we can use the Lagrange multiplier technique, which yields $\mathbf{WM} = \mathbf{MW}^\top$ and $\mathbf{WW}^\top = \mathbf{I}_k$. These equations lead to a fixed-point algorithm (see e.g. [26]) with a step $\mathbf{W} \leftarrow (\mathbf{MM}^\top)^{-1/2}\mathbf{M}$.

Using this we can search for the limiting distributions of $\hat{\mathbf{\Gamma}}$ and $\hat{\mathbf{W}}$, with known dimension k and lags in S. As the estimate is also affine equivariant, we can wlog consider here the case, where $\mathrm{COV}(\mathbf{x}) = \mathbf{I}_p$ and $\mathbf{W} = (\mathbf{I}_k, \mathbf{0})$.

Let T be the length of the time series. After deriving the joint limiting distribution of $\sqrt{T}(\widehat{\mathrm{COV}}(\mathbf{x}) - \mathbf{I}_p)$ and $\sqrt{T}(\hat{\mathbf{G}}_j - \mathbf{G}_j)$ and then noticing that the joint limiting distribution of $\hat{\mathbf{W}}$ and $\hat{\mathbf{M}}$ satisfies the conditions

$$\sqrt{T}(\hat{\mathbf{W}} - \mathbf{W})\mathbf{M}' - \mathbf{M}\sqrt{T}(\hat{\mathbf{W}} - \mathbf{W})' = \sqrt{T}(\hat{\mathbf{M}} - \mathbf{M})\mathbf{W}' - \mathbf{W}\sqrt{T}(\hat{\mathbf{M}} - \mathbf{M})' + o_P(1)$$

and

$$\sqrt{T}(\hat{\mathbf{W}} - \mathbf{W})\mathbf{W}' = -\mathbf{W}\sqrt{T}(\hat{\mathbf{W}} - \mathbf{W})' + o_P(1),$$

one can further derive $\sqrt{T}(\hat{\mathbf{W}} - \mathbf{W})$ and $\sqrt{T}(\widehat{\mathrm{COV}}(\mathbf{x}) - \mathbf{I}_p)$ and then finally get $\sqrt{T}(\hat{\mathbf{\Gamma}} - \mathbf{W}) = \sqrt{T}(\hat{\mathbf{W}} - \mathbf{W}) - \frac{1}{2}\mathbf{W}\sqrt{T}(\widehat{\mathrm{COV}}(\mathbf{x}) - \mathbf{I}_p) + o_P(1)$. For similar derivations based on the Langrangian multiplier technique, see for example [32]. The derivations for the regular SAVE asymptotics in the iid case already appeared to be quite complicated [19]. Here the derivations are much more demanding as numerous matrices are jointly diagonalized with serially dependent observations.

3.4. *A hybrid of TSIR and TSAVE*

As both SIR and SAVE have their advantages and drawbacks, a hybrid of SIR and SAVE using a convex combination of the two supervised matrices was proposed in [11]. As the time series versions of SIR and SAVE show similar behaviour to the original SIR and TSAVE, respectively, we similarly propose here a convex combination of TSIR and TSAVE methods. We call this time series SIR and SAVE hybrid method TSSH. In TSSH we are searching for a $k \times p$ matrix $\mathbf{W}_{TSSH} = (\mathbf{w}_1, \ldots, \mathbf{w}_k)^\top$ with orthonormal

rows $\mathbf{w}_1, \ldots, \mathbf{w}_k$ that maximizes

$$\sum_{j \in S} \sum_{i=1}^{k} \left(\mathbf{w}_i^\top \left((1-a) * G_{1,j}(\mathbf{x}^{st}, y) + a * G_{2,j}(\mathbf{x}^{st}, y) \right) \mathbf{w}_i \right)^2,$$

where $a \in [0,1]$ and S a set of chosen lags as before. Then $a = 0$ gives TSIR and $a = 1$ gives TSAVE.

Note that the results similar to Results 3.1 and 3.3 and the Result 3.5 apply to TSSH as well. Also the search for the number of latent sources and lags goes as for TSIR and TSAVE.

In addition to the issues that TSIR and TSAVE have, a proper value for the coefficient a needs to be found. This is discussed in Section 4.4.

Extending the $SIR|SAVE$ method by [23] for time series is challenging, as we need to search not only for sources but also lags corresponding to each of the sources.

3.5. *Identification of k and the lags of interest*

In practice the number of sources, i.e. the value of k, is not known and needs to be estimated as well. Also the important lags regarding the sources need to be found. We can choose these by using the quantities λ_{ij}. However, at this stage of the development of TSAVE, formal testing is not possible yet and we suggest to use the same strategies as suggested for TSIR in [7].

For that purpose consider the matrix $\mathbf{L} = l_{ij}$ where

$$l_{ij} = \frac{\lambda_{ij}}{\sum_{i=1}^{k} \sum_{j=1}^{s} \lambda_{ij}}, \quad i = 1, \ldots, k; \; j = 1, \ldots, s,$$

contains the scaled pseudo eigenvalues and the scaling is chosen such that the elements of $\mathbf{L}$ add up to 1. Note that we have assume here that we use the first s lags, as currently we do not have information that would suggest to use some other set of lags. Row and column sums of $\mathbf{L}$ will be again denoted as $l_{i\cdot} = \sum_{j=1}^{s} l_{ij}$, as before, and $l_{\cdot j} = \sum_{i=1}^{k} l_{ij}$.

Assume then that $\Gamma_{TSAVE}(\mathbf{x}; y)$ is defined such that the latent sources are ordered according to their magnitudes of $\lambda_{1\cdot} \geq \ldots \geq \lambda_{k\cdot}$ and $k = p$. Then [7] suggested different strategies to find the appropriate amount of sources and the lags corresponding to those sources, by trying to explain similar as in principal component analysis (PCA) $100 \cdot P\%$ of the dependence between the latent sources and the response series.

The suggested strategies can be summarized as:

ALL LAGS: keep all s lags and find the smallest value $\hat{k}$ such that $\sum_{i=1}^{\hat{k}} l_{i\cdot} \geq P$.

ALL SOURCES: keep all k sources and find the smallest $\hat{s}$ in such way that $\sum_{j=1}^{\hat{s}} l_{\cdot j} \geq P$.

RECTANGLE: find $\hat{k}$ and $\hat{s}$ with the smallest product $\hat{k}\hat{s}$ in such way that $\sum_{i=1}^{\hat{k}} \sum_{j=1}^{\hat{s}} l_{ij} \geq P$.

BIGGEST VALUES: find the smallest number $\hat{r}$ of elements $(i_1, j_1), \ldots, (i_{\hat{r}}, j_{\hat{r}})$ of $\mathbf{L}$ in such way that $\sum_{k=1}^{\hat{r}} l_{i_k j_k} \geq P$.

While the first two strategies assume some prior knowledge about the number of lags or sources respectively, the last two methods seem suitable for general use. As in the iid

case, the 'real' amount of sources k may not be found due to the slicing (value of H) and/or the method used, and hence it is possible that $\hat{k} < k$.

Natural values for P are then for example 0.5 or 0.8. How the different strategies perform will also be considered in the example and simulation section.

4. Examples and simulations

In this section the differences between TSIR and TSAVE are first visualized. Then the simulation settings and the prediction models are presented. In Section 4.3 we search for the best values for the number of slices H in TSIR and TSAVE and in Section 4.4 the appropriate values for the coefficient a for TSSH. In both cases we aim to give some guidelines how to choose them in practice. Finally we compare TSAVE with TSIR, SIR and SAVE (applied to time series case) in Section 4.5.

Note that TSIR, TSAVE and TSSH, together with the different selection strategies described in the previous section, are implemented in the R package tsBSS [34] and are used together with the R package JADE [33] in this section.

4.1. *Visualization of the differences of TSIR and TSAVE*

It is already well established that the regular SIR works efficiently with linear relationships, but not when the relationship in $y = g(\mathbf{z}^{(1)}) + \epsilon$ is specified by a symmetric function g [8, 9]. On the other hand, the regular SAVE works with a symmetric function g.

To consider the differences between the time series versions of both methods, consider the examples where the response y at time t be

M1: $y_t = x_{t-1} + x_{t-3} + \epsilon_t$
M2: $y_t = 1 + x_{t-1}^2 + x_{t-3}^2 + \epsilon_t$

where $\epsilon_t \sim N(0, 0.2)$ and x_t follows an AR(1) model with $\phi = 0.1$. To illustrate how and where TSIR and TSAVE work, we plot the values of y_t against the values of x_{t-j}, where $j = 1, 2, 3$ or 10 for all the models (see Figures 1 and 2).

The larger black dots in the figures denote the sample values of $\mathrm{E}(x_t|y_{t+j})$ in each slice. Also the variance of these values is added to each figure as text. A non-zero variance value indicates that TSIR is able to find a relationship between y_t and x_{t-j}.

The width of the dark gray bars added around the black dots corresponds to the sample value of $(1-\mathrm{Var}(x_t|y_{t+j}))^2$ in each slice. When TSAVE cannot find a relationship between y_t and x_{t-j}, all these values are so close to zero that the bars are hardly visible. Also the means of these values are added to each figure as text. A non-zero mean indicates that TSAVE is able to find a relationship between y_t and x_{t-j}.

From Figure 1 and 2 it can be seen that TSIR cannot find any relationship between y_t and x_{t-j}, when $j = 2$ or 10, as y_t did not depend on x_t with those lags. However, the results are different with $j = 1$ and 3. As seen on the left side panels of Figure 1, y_t and x_{t-j} have a strong linear relationship. The variance of the slice means is clearly non-zero and indicates that TSIR finds the relationship. Also a non-zero value for the mean of the conditional variances indicates that also TSAVE works with linear relationships.

On the left side panels of Figure 2, y_t and x_{t-j} have a strong quadratic relationship. It can easily be seen that TSAVE finds the quadratic relationship. However, as the mean values for each slice are close to zero, TSIR fails to find the quadratic relationship, similar to the regular SIR in iid regression.

Based on the figures and the values in them, it can be concluded that TSAVE finds

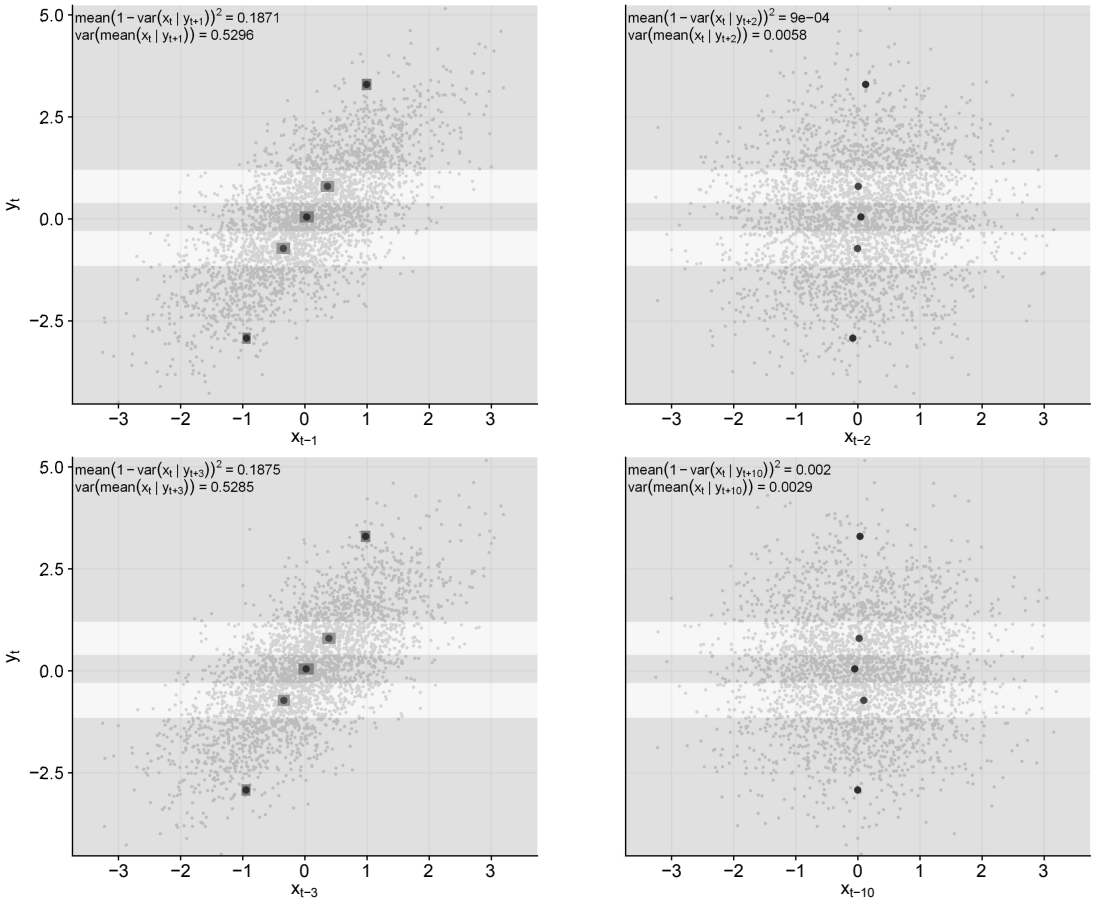

Figure 1. Model *M1*: Scatterplot of y_t and x_{t-j}, $j = 1, 2, 3, 10$ with slices of y_t as the shaded areas

both the linear and the quadratic relationship between y_t and x_t with lags 1 and 3 in the models *M1* and *M2*.

Next we illustrate how the appropriate lags and the amount of the latent sources are chosen in TSAVE and TSIR using the strategies mentioned in Section 3.5.

Consider a 4-variate time series $\mathbf{x} = (x_1, x_2, x_3, x_4)^\top$, where x_1 and x_2 are AR(1) processes with $\phi = 0.2$, x_3 is ARMA(1,1) with $\phi = 0.3$ and $\theta = -0.4$, and x_4 is MA(1) model with $\theta = -0.4$. As both methods are affine equivariant, we have chosen $\mathbf{\Omega} = \mathbf{I}_4$ as the mixing matrix and therefore $\mathbf{x} = \mathbf{z}$ for all $t \in \mathbf{Z}$. The time series $\mathbf{z}$ are standardized and the length of the time series is $T = 10000$.

In choosing the number of sources and the lags, we use $P = 0.8$ as the threshold value, and $H = 5$ as the number of slices. Tables are constructed as the average values of the elements l_{ij} over 100 repetitions, and we have used lags $1, \ldots, 12$ for both methods. Assume now that the response y at a time t depends on the predictors as follows.

$$y_t = z_{1,t-1}^2 + 3z_{2,t-5} + \epsilon_t,$$

where $\epsilon_t \sim N(0, 1)$. Table 1 includes the $\mathbf{L}$ matrices for both TSAVE and TSIR based on this model. It can be seen that TSAVE finds two latent sources and five lags. Also already the values of the elements l_{12} and l_{51} are clearly bigger than others and together they already explain more than 80 % of the dependence between the response y and the predictors. On the other hand TSIR finds only one source (and five lags), which seems

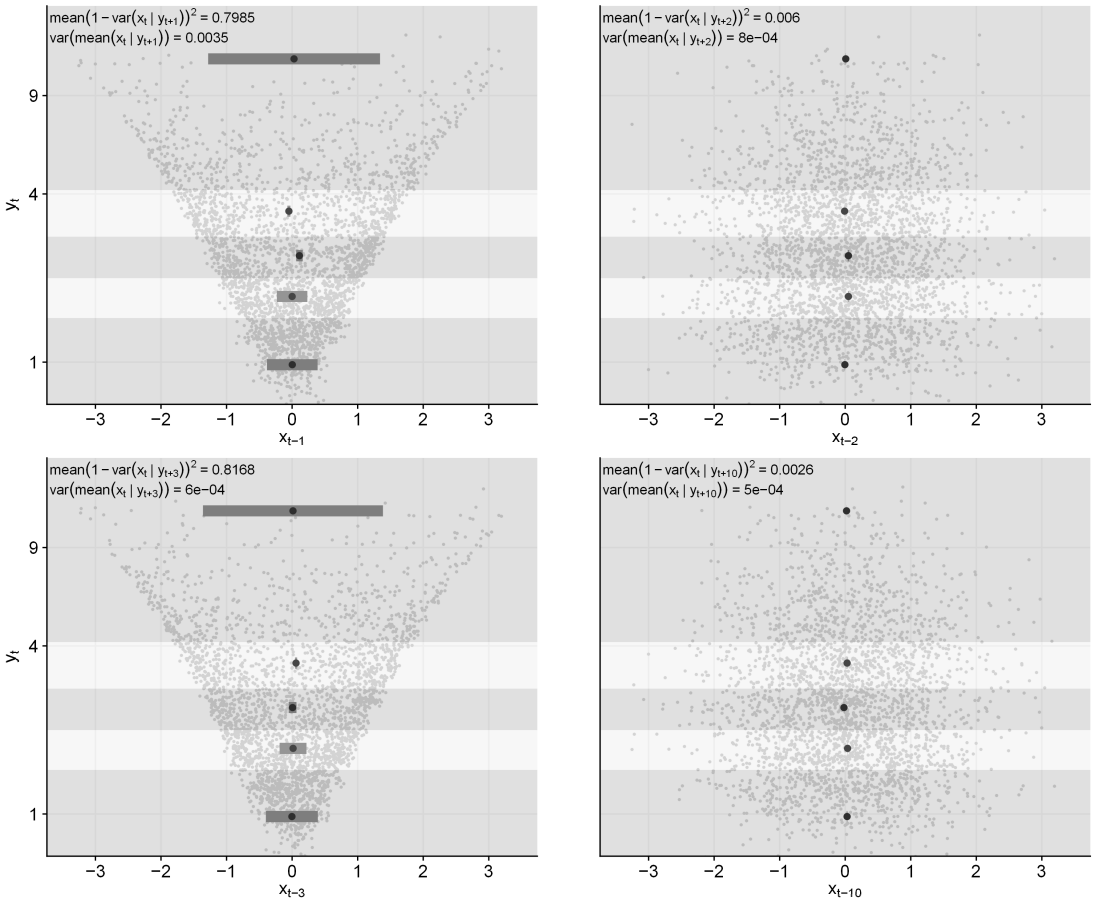

Figure 2. Model *M2*: Scatterplot of y_t and x_{t-j}, $j = 1, 2, 3, 10$ with slices of y_t as the shaded areas. The y-axis is in a logarithmic scale.

	$\mathbf{w}_1'\mathbf{x}^{st}$	$\mathbf{w}_2'\mathbf{x}^{st}$	$\mathbf{w}_3'\mathbf{x}^{st}$	$\mathbf{w}_4'\mathbf{x}^{st}$	Sum
$t-1$	0.002	0.317	0.002	0.002	0.323
$t-2$	0.002	0.003	0.002	0.002	0.010
$t-3$	0.002	0.002	0.003	0.002	0.009
$t-4$	0.003	0.002	0.003	0.002	0.010
$t-5$	0.576	0.002	0.002	0.002	0.582
$t-6$	0.003	0.002	0.003	0.002	0.010
$t-7$	0.002	0.002	0.003	0.002	0.009
$t-8$	0.002	0.002	0.003	0.002	0.009
$t-9$	0.002	0.002	0.002	0.002	0.009
$t-10$	0.002	0.002	0.003	0.002	0.009
$t-11$	0.002	0.002	0.003	0.002	0.009
$t-12$	0.002	0.002	0.002	0.002	0.009
Sum	0.603	0.342	0.029	0.025	1.000

	$\mathbf{w}_1'\mathbf{x}^{st}$	$\mathbf{w}_2'\mathbf{x}^{st}$	$\mathbf{w}_3'\mathbf{x}^{st}$	$\mathbf{w}_4'\mathbf{x}^{st}$	Sum
$t-1$	0.001	0.001	0.001	0.001	0.004
$t-2$	0.001	0.001	0.001	0.001	0.004
$t-3$	0.002	0.001	0.001	0.001	0.006
$t-4$	0.035	0.001	0.001	0.001	0.038
$t-5$	0.878	0.001	0.001	0.001	0.881
$t-6$	0.036	0.001	0.001	0.001	0.039
$t-7$	0.002	0.001	0.001	0.001	0.006
$t-8$	0.001	0.001	0.001	0.001	0.004
$t-9$	0.001	0.001	0.001	0.001	0.004
$t-10$	0.001	0.001	0.001	0.001	0.004
$t-11$	0.001	0.001	0.001	0.001	0.004
$t-12$	0.001	0.001	0.001	0.001	0.004
Sum	0.961	0.015	0.013	0.011	1.000

Table 1. The matrix $\mathbf{L}$ with row sums and column sums: TSAVE (left panel) and TSIR (right panel)

to explain by far the most of the dependence, but fails to find the other source with the quadratic relationship.

4.2. *Models and prediction*

The results presented here are based on the following ARMA models, where the four $\mathbf{z}$ component series are as follows.

$$\begin{aligned} \text{Components 1 and 2:} &\quad AR(1) \text{ with } \phi = 0.2 \text{ (or } 0.8). \\ \text{Component 3:} &\quad ARMA(1,1) \text{ with } \phi = 0.3 \text{ and } \theta = 0.4. \\ \text{Component 4:} &\quad MA(1) \text{ with } \theta = -0.4, \text{ respectively.} \end{aligned}$$

Note that for the first two components two different ϕ values are used to compare how the level of the autocorrelation affects the results. The response series y depends then on the first two components z_1 and z_2, in the following different ways:

$$\begin{aligned} \text{Model } A\text{: } y_t &= 2z_{1,t-1} + 3z_{2,t-1} + \epsilon_t \\ \text{Model } B\text{: } y_t &= z_{1,t-1}^2 + 3z_{2,t-5} + \epsilon_t \\ \text{Model } C\text{: } y_t &= (2z_{1,t-1} + 3z_{2,t-1})^2 + \epsilon_t \\ \text{Model } D\text{: } y_t &= z_{1,t-1}^2 + 3z_{2,t-5}^2 + \epsilon_t \\ \text{Model } E\text{: } y_t &= 2z_{1,t-1}^3 + 3z_{2,t-5}^2 + \epsilon_t \end{aligned}$$

All the models have iid $N(0,1)$-distributed innovations ϵ_t. As before $\mathbf{\Omega} = \mathbf{I}_4$ is used as the mixing matrix.

Note that we have also performed additional simulations which evaluated what happens if there are almost non-stationary components, stochastic volatility components or components with heavy-tailed innovations. The exact settings are detailed in the supplementary material together with the corresponding results. While there are maybe minor differences, we believe that the guidelines derived on the settings specified above suffice in practice.

For prediction we use the prediction model (3). As an approximation of the function f we use both simple linear regression and also regression with quadratic B-splines (for model E cubic B-splines, as it includes a factor of the form z^3). The size of the testing set is 100, i.e. we predict the last 100 values of the data. To predict the value for $T-100+i$, $i = 1, 2, \ldots, 100$, we use the observations $i, \ldots, T-100+(i-1)$ as a training set ('rolling window approach').

We estimate the accuracy of the prediction by calculating the root mean square error (RMSE) based on the one-step-ahead prediction errors $\hat{\epsilon}_t$ of the last 100 observations (testing set). The lags used to create the matrix $\mathbf{L}$ are $1, \ldots, 12$ and the number of repetitions is 500.

4.3. *On the number of slices for TSIR and TSAVE*

In [7] simulations for TSIR are conducted only with value $H = 10$, which is a common value used with SIR. Here we aim to go a bit deeper and provide some guidelines for choosing the value H for TSAVE as well as for TSIR.

To find the optimal number of slices H, we conduct an experiment with different strategies and with two different threshold values $P = 0.5$ and 0.8 to find the important lags and the number of sources. We use time series lengths $T = 500, 1000, 2000$ and 3000.

First we predict the values using all the strategies with linear and spline predictions. With $H = 2, 5, 10, 20$ and 40 and models A – D, we calculate the RMSE values compared

to value $H = 10$, which is commonly used in e.g. SIR and has been used in TSIR in [7]. The results are for the components with low and high autocorrelation as well as for time series lengths $T = 500$, 1000, 2000 and 3000.

The choices $H = 20$ and $H = 40$ do not work that well in any of the settings in TSAVE. We conclude from this that there are then simply too few observations per slice. Thus we show here results only concerning values $H = 2, 5$ and 10. The choices $H = 2$ and $H = 5$ are compared to $H = 10$. If the choice is better than $H = 10$, then the relative RMSE values will generally be lower than one.

In model A the linear predictions are the most efficient and spline predictions may add a little bit of additional noise. In models B – D, however, the linear predictions are not very good at determining the best value of H. Only with $T = 500$ the choice $H = 2$ seems to be a bit better than others, while in longer time series any possible difference is barely visible. This is expected, since the relationship between the predictors and the response is not linear. Thus the spline predictions is preferred for determining the optimal value for H.

In both low and high dependence settings the threshold value $P = 0.5$ seems to be working well for models A and C, where only one source is expected to be found, and choosing $P = 0.8$ has only a very small effect on results. For models B and D it seems that $P = 0.5$ might be enough in short time series ($T = 500$), but in longer time series, especially when we have sources with high autocorrelation, using $P = 0.8$ is crucial for the second source to be found. Thus $P = 0.8$ is seems to be a safe choice in general.

When comparing the different strategies to select the number of sources and lags, the biggest values strategy seems to produce almost always the best results, and in the remaining few cases it is very close to the best. Figures 3–6 show then based on that strategy the relative RMSE for models A – D and the different sample sizes. And these figures clearly indicate that using only 2 or 5 slices for TSAVE is clearly better than 10 slices. Only with increasing sample size the differences vanish which means that then all slices contain enough observations. The same can also be observed using other strategies for the selection (not shown here) although then in rare cases can be 5 slices better than 2 slices.

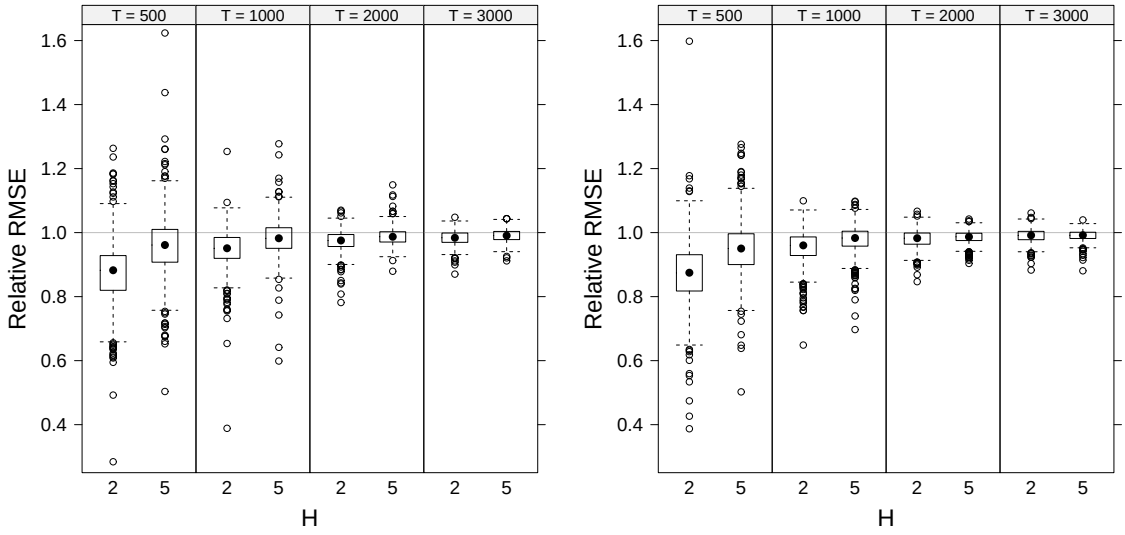

Figure 3. TSAVE. Model A with the biggest values strategy. Relative RMSE values compared to $H = 10$ with $\phi = 0.2$ (left panel) and $\phi = 0.8$ (right panel).

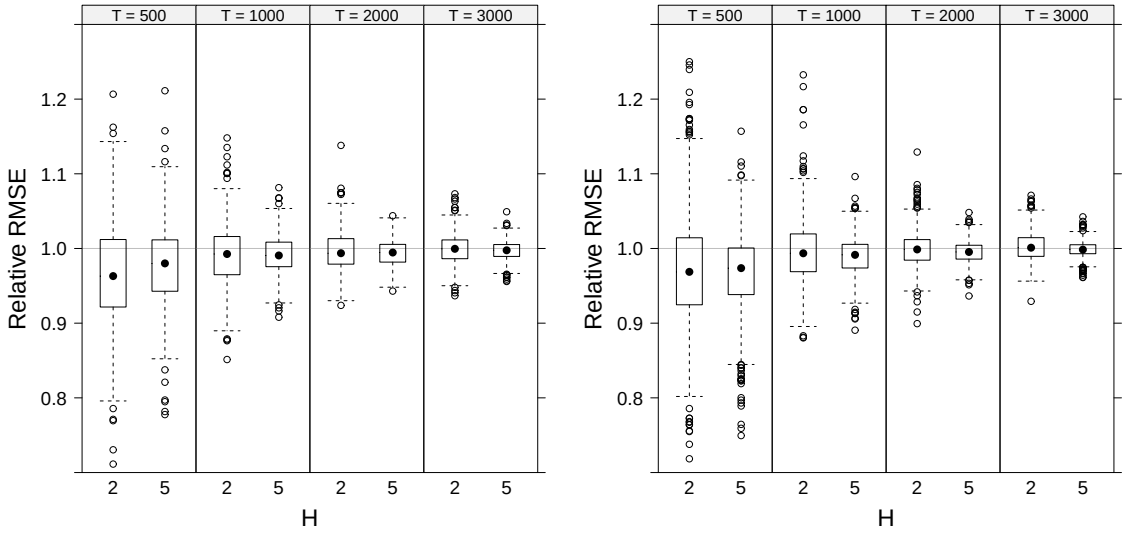

Figure 4. TSAVE. Model B with the biggest values strategy. Relative RMSE values compared to $H = 10$ with $\phi = 0.2$ (left panel) and $\phi = 0.8$ (right panel).

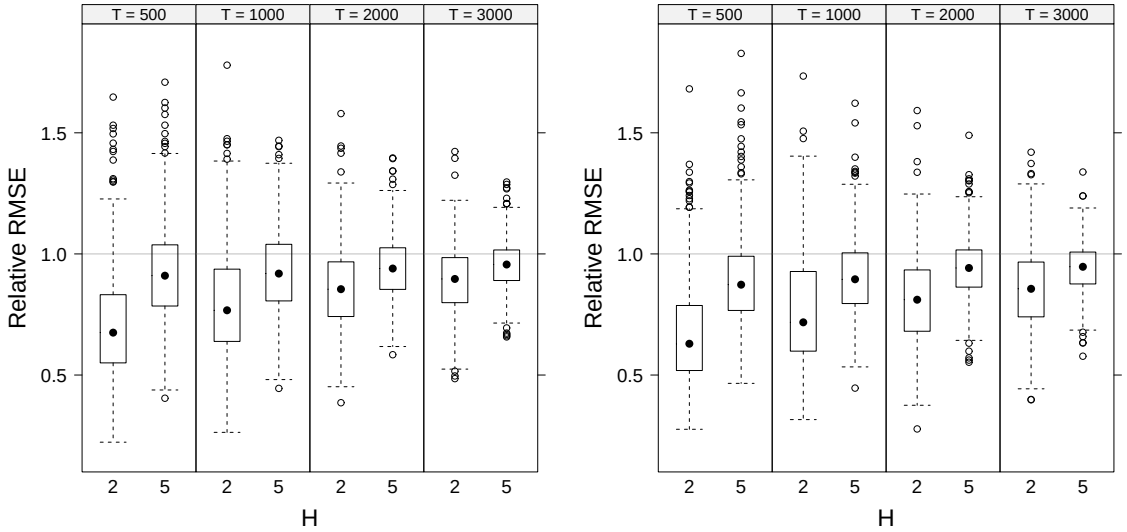

Figure 5. TSAVE. Model C with the biggest values strategy. Relative RMSE values compared to $H = 10$ with $\phi = 0.2$ (left panel) and $\phi = 0.8$ (right panel).

For TSIR the choice $H = 10$ seems to be the safest when linear predictions are used. However, when using spline predictions, $H = 2$ and $H = 5$ may be better choices with shortest time series, i.e. with $T = 500$. Figure 7 includes results from spline predictions using the biggest values strategy. Note that with the models $B - D$, TSIR does not work well and thus the evaluation of the value of H for TSIR is not considered in those models.

We could also look for the best H with the $\mathbf{L}$ matrix. For the models A and C we can check, if only one source with at least lag 1 is found, as only one is expected. For the models B and D we can first check that if one source with lag 1 and one source with lag 5 are found. If that is true, then we can check that if only the two sources are found.

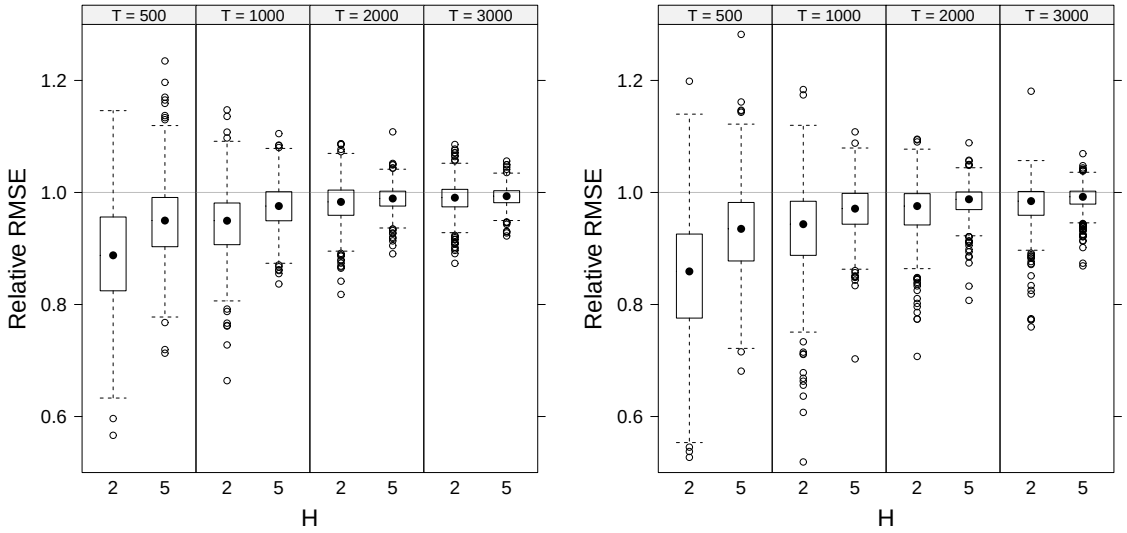

Figure 6. TSAVE. Model D with the biggest values strategy. Relative RMSE values compared to $H = 10$ with $\phi = 0.2$ (left panel) and $\phi = 0.8$ (right panel).

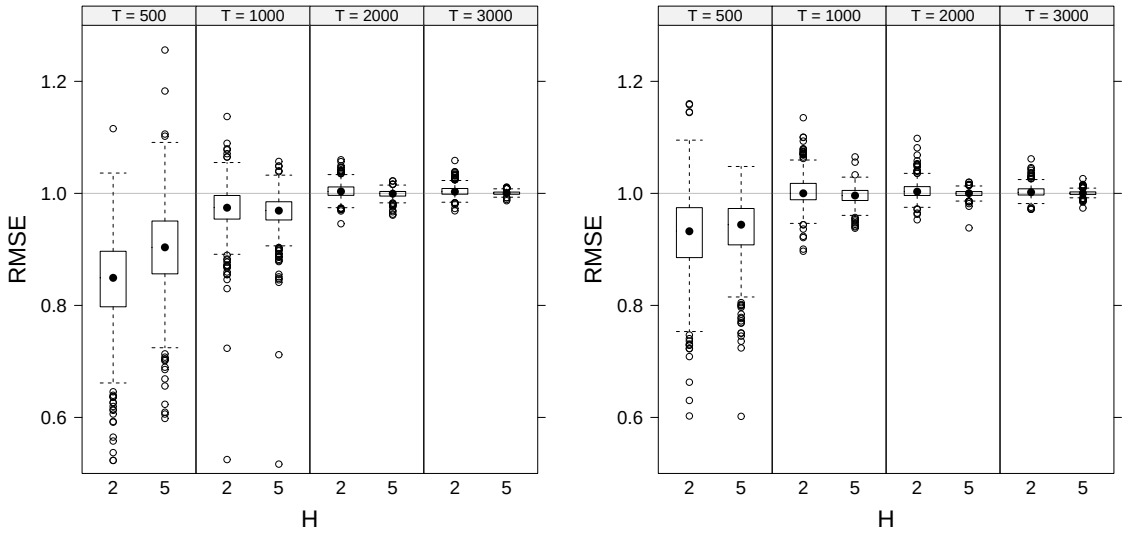

Figure 7. TSIR. Model A with the biggest values strategy. Relative RMSE values compared to $H = 10$ with low (left panel) and high (right panel) dependency of the sources.

With the biggest values strategy with $T = 3000$ for model A, generally $H = 2, 5$ and 10 are good choices when $P = 0.8$, and $H = 2$ is the best when $P = 0.5$. In all the other models also $H = 2$ seems to be the most efficient choice when $P = 0.8$. For model C the choices $H = 2$ and 5 are generally safe, while with $P = 0.8$ also $H = 10$ and 20 seem to be good enough. As an example, Figure 8 has the results for model C with $T = 3000$. In model D the threshold value $P = 0.5$ seems to be generally too low for efficiently finding the right amount of sources, while with $P = 0.8$ choices $H = 2$ and $H = 5$ seem to produce the expected results. For TSIR any value $H \leq 10$ seems to be safe for model A. For a short time series ($T = 500$), $H = 2$ seems mostly the safest choice for TSAVE and $H = 2, 5$ and 10 for TSIR.

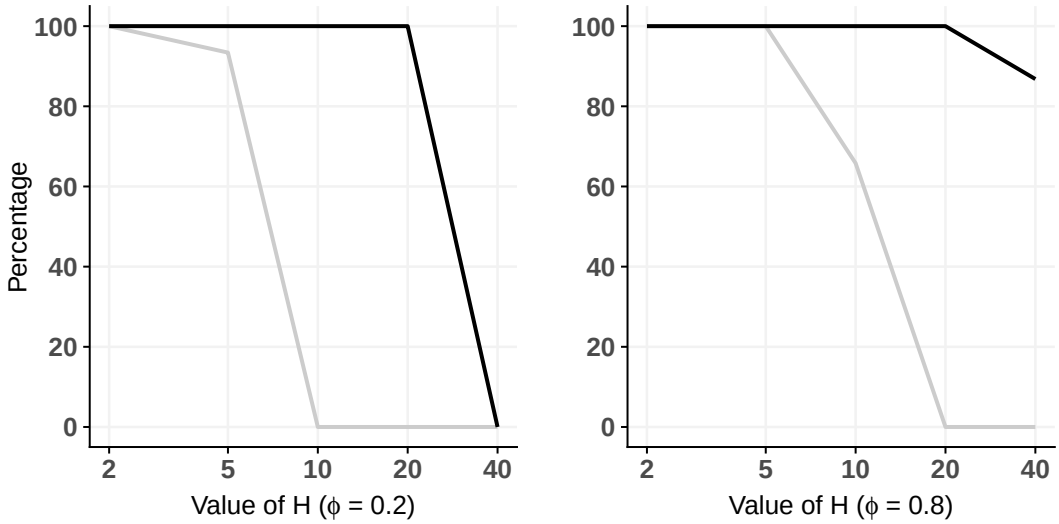

Figure 8. TSAVE. Model C with $T = 3000$ using the biggest values strategy. Percentage of cases that finds the correct lags and correct amount of sources. Black line: $P = 0.5$, gray line: $P = 0.8$.

To conclude this section, we can say that for TSIR $H = 10$ is generally a reliable choice, however, with short time series a lower H might be beneficial, depending on the prediction method used. For TSAVE the number of observations per slice is more important and depends also on the data generation process. Based on our simulations we recommend to have at least 100 observations per slice and for the time series lengths considered here our preferred choice is $H = 2$, but also $H = 5$ seems good. For a shorter time series $H = 2$ might be the only choice, but the longer the time series the smaller differences there are between the values of H and then also a larger H would be reliable.

Furthermore, the simulations suggest that the value of P has a big influence on the number of selected components and $P = 0.5$ is more restrictive than $P = 0.8$, which is however also very intuitive. The biggest values strategy to choose the amount of sources and the lags corresponding to them is also recommended.

4.4. *On the TSSH method and the choice of coefficient a*

In model A, the relationship between the response and the predictors is linear. Already [7] show that TSIR works in such models efficiently. On the other hand, the models *B*-*E* have symmetric parts. As seen in Section 4.1, TSIR in unable to find the relationship in such case (see Figure 2). This is also seen later in the simulation results of Section 4.5. Also from Figure 2 it can be seen that TSAVE still works with the linear relationships, but not as efficiently as TSIR.

This preference for different structures of the different methods was the main motivation for the introduction of the hybrid in Section 4.3. Now we consider the optimal value of a for model E. This model is similar to the model 4 in [11], for which the hybrid of iid SIR and SAVE shows clearly better performance than SIR or SAVE separately. Therefore we could expect here that the TSIR part uncovers the asymmetric part of the dependence $2z_{1,t-1}^3$ efficiently and TSAVE the symmetric part $3z_{2,t-5}^2$, and hopefully both together work even better. The question how much weight should be given to which method, i.e. the proper value for a. For the iid hybrid method [11] recommend as general rule of thumb to use $a = 0.5$.

Following our general guidelines from the previous section, in the following presentation the results are based on using $H = 10$ slices for TSIR part and $H = 2$ slices for TSAVE part. To select the components, P is set to 0.8 and cubic B-splines are used for the prediction. The RMSE values are then computed for the values $a = 0, 0.1, 0.2, \ldots, 1$, where $a = 0$ refers to the pure TSIR method and $a = 1$ to the pure TSAVE method. We show here the RMSE for the time series lengths $T = 500$ and 3000 in Figure 9 and Figure 10. It seems that there are not so big differences as long as not all or almost all of the weight is given to TSIR or all the weight given to TSAVE. The central values of a seem to be a little bit better.

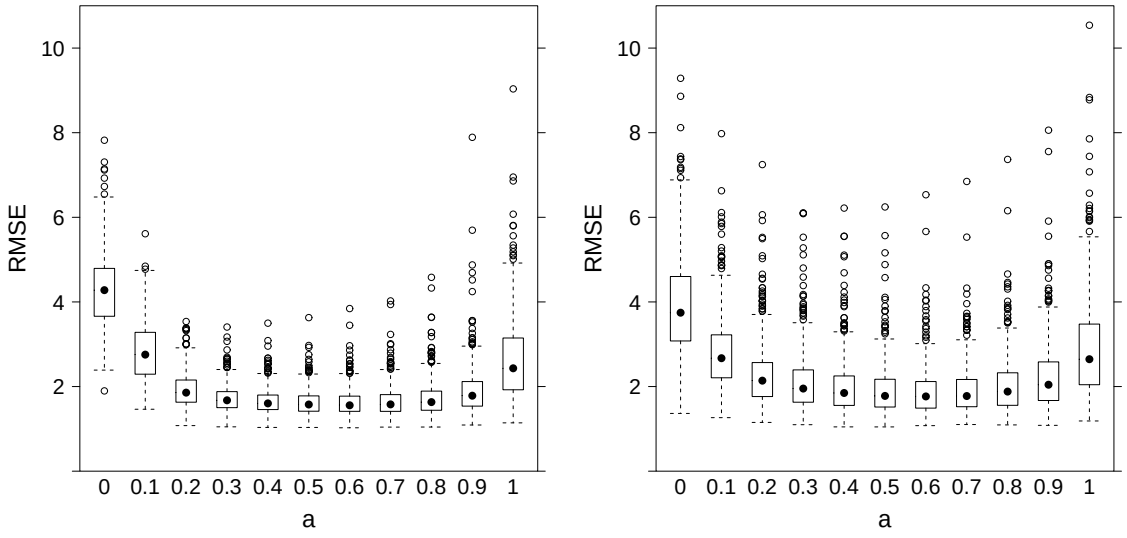

Figure 9. TSSH. Model E with the biggest values strategy: $T = 500$ and $H = 2$ for TSAVE part and $H = 10$ for TSIR part. RMSE values with $\phi = 0.2$ (left panel) and $\phi = 0.8$ (right panel).

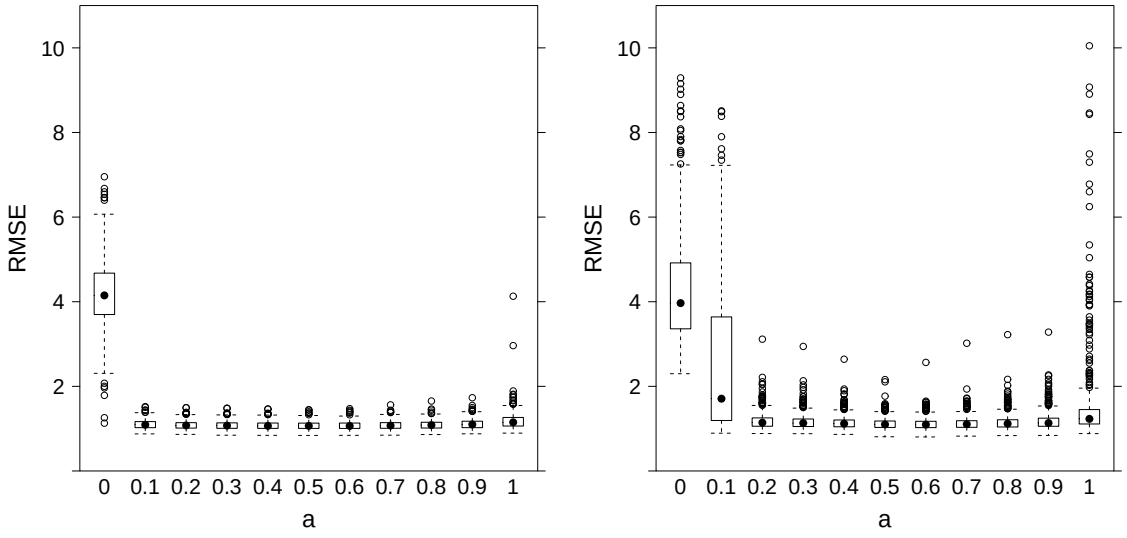

Figure 10. TSSH. Model E with the biggest values strategy: $T = 3000$ and $H = 2$ for TSAVE part and $H = 10$ for TSIR part. RMSE values with $\phi = 0.2$ (left panel) and $\phi = 0.8$ (right panel).

To investigate this a bit further, we also compare the choices of $P = 0.5$ and $P = 0.8$ using the $\mathbf{L}$ matrix. Figure 11 gives then for $T = 3000$ the percentages of right number of sources and appropriate lags chosen, based on the biggest value strategy, for $a = 0, 0.01, 0.02, \ldots, 1$

From Figure 11 we can conclude that values for a around 0.5 and 0.6 are reasonable choices. Considering results using other selection strategies not shown here we can in general recommend the value $a = 0.5$, which coincides with the recommendation of the regular SIR and SAVE hybrid.

The main feature we observed for the hybrid is that with values closer to $a = 0.5$, the value of P is less crucial and one often comes to the same conclusion. Quite different from what we have seen when using only TSIR or only TSAVE.

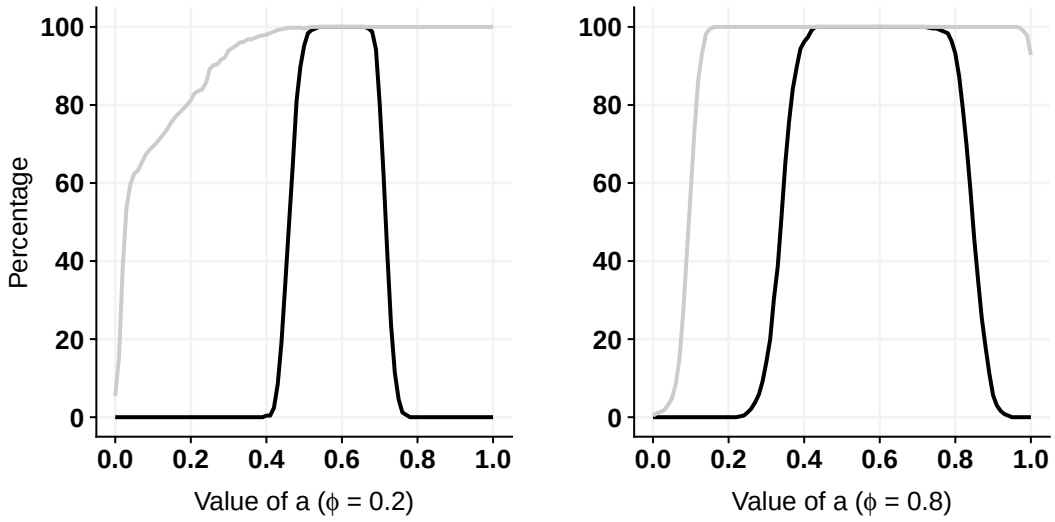

Figure 11. TSSH. Model E with the biggest values strategy: $T = 3000$ and $H = 2$ for TSAVE part and $H = 10$ for TSIR part. Percentage of cases that finds the correct lags and correct amount of sources. Black line: $P = 0.5$, gray line: $P = 0.8$.

4.5. *Comparison to other approaches*

To compare different methods, we simulate with time series length $T = 3000$ using $H = 2$ for TSAVE and $H = 10$ for TSIR, $P = 0.8$ and the biggest values strategy, as recommended in Section 4.3. For models A – D the relative RMSE values are compared to Oracle estimator, where the functional form of the relationship between the response and the predictors is known, but the coefficients are estimated.

Figures 12 – 15 have the relative RMSE values based on different methods compared to the Oracle estimator. The methods here are TSAVE and TSIR as well as the original SAVE and SIR (Becker & Fried SIR [4]) with the lagged values of $\mathbf{x}$ as predictors, i.e. with $\mathbf{x}_t^* = (\mathbf{x}_{t-1}, \ldots, \mathbf{x}_{t-s})$. Here we have used $s = 12$.

For original SIR we have used $H = 10$ [8] and for original SAVE $H = 5$, as $H = 2$ may be too low for SAVE (see for example [19]). To choose the number of sources, we use the ordered empirical eigenvalues λ_i, $i = 1, \ldots, s \cdot p$, of the supervised matrices $\mathrm{COV}[\mathrm{E}(\mathbf{x}^{*,st}|y^{sl})]$ in the original SIR and $\mathrm{E}[(\mathbf{I}_p - \mathrm{COV}(\mathbf{x}^{*,st}|y^{sl}))^2]$ in the original SAVE. The chosen number of sources is the minimal $\hat{k}$ for which $\sum_{i=1}^{\hat{k}} \lambda_i / \sum_{i=1}^{s \cdot p} \lambda_i \geq P = 0.8$.

From Figure 12 we can see that TSIR and TSAVE both work very well. From Figures

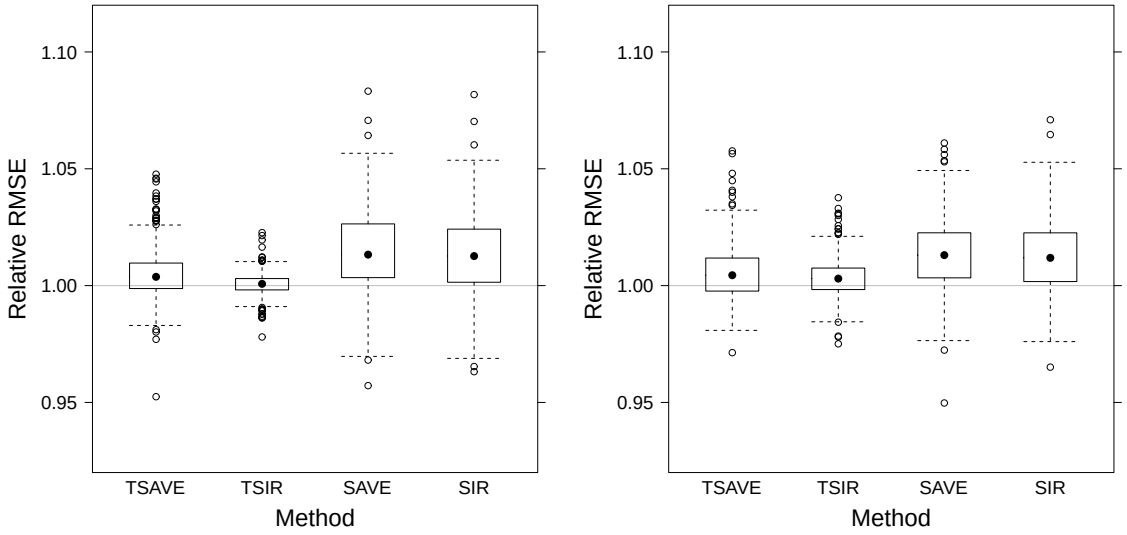

Figure 12. Model A with biggest values strategy. Relative RMSE values compared to Oracle estimator with $\phi = 0.2$ (left panel) and $\phi = 0.8$ (right panel).

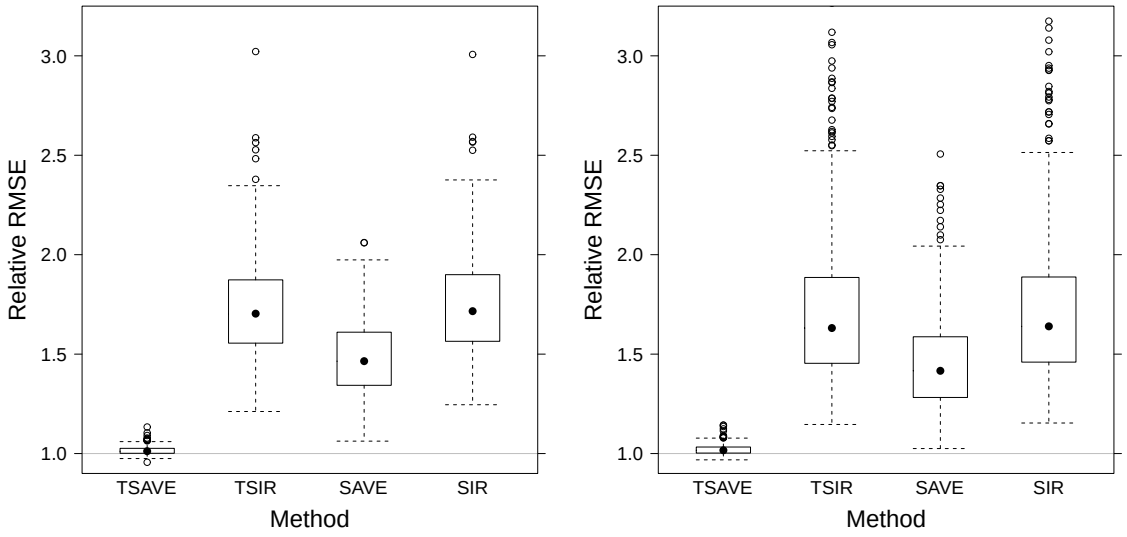

Figure 13. Model B with biggest values strategy. Relative RMSE values compared to Oracle estimator with $\phi = 0.2$ (left panel) and $\phi = 0.8$ (right panel).

13 – 15 we see that TSAVE clearly works the best in the models B, C and D, while also the original SAVE with lagged variables as predictors works better than TSIR and the original SIR. The iid versions using the lagged variables as predictors do not work that well except in the linear case (Model A). If we used $H = 5$ instead of $H = 2$, the results would be very similar.

To evaluate the effect of the dimension we included as a final setting also a setup, where we have $p = 10$ components and the 'true' number of series that the response depends on is $k = 3$. The simulations were conducted using several different time series lengths ($T = 500, 1000, 2000, 3000$ and 5000), but here we show only the results based on $T = 3000$ and other results can be found in the supplementary material.

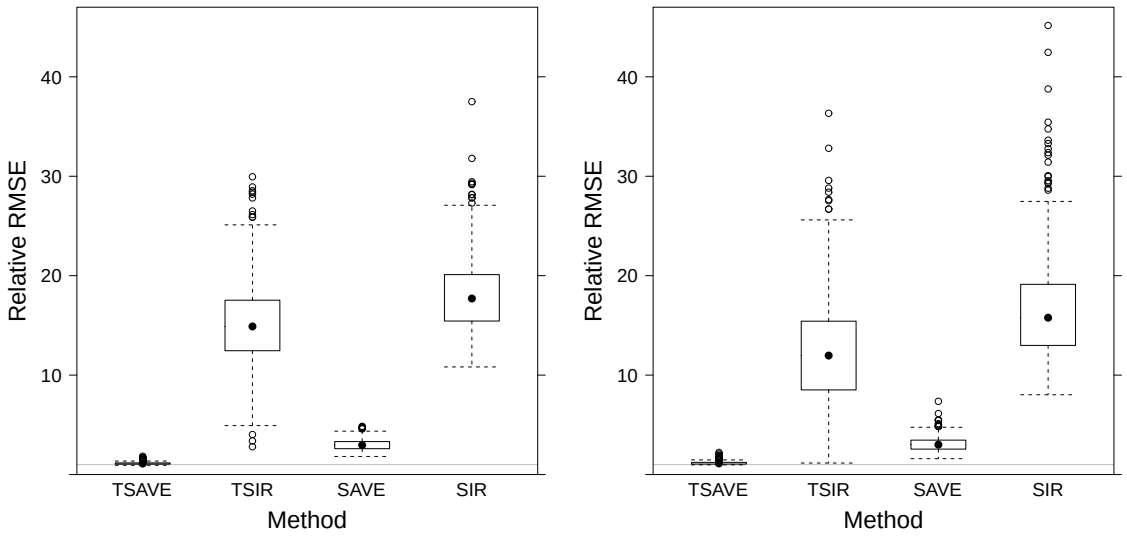

Figure 14. Model C with biggest values strategy. Relative RMSE values compared to Oracle estimator with $\phi = 0.2$ (left panel) and $\phi = 0.8$ (right panel).

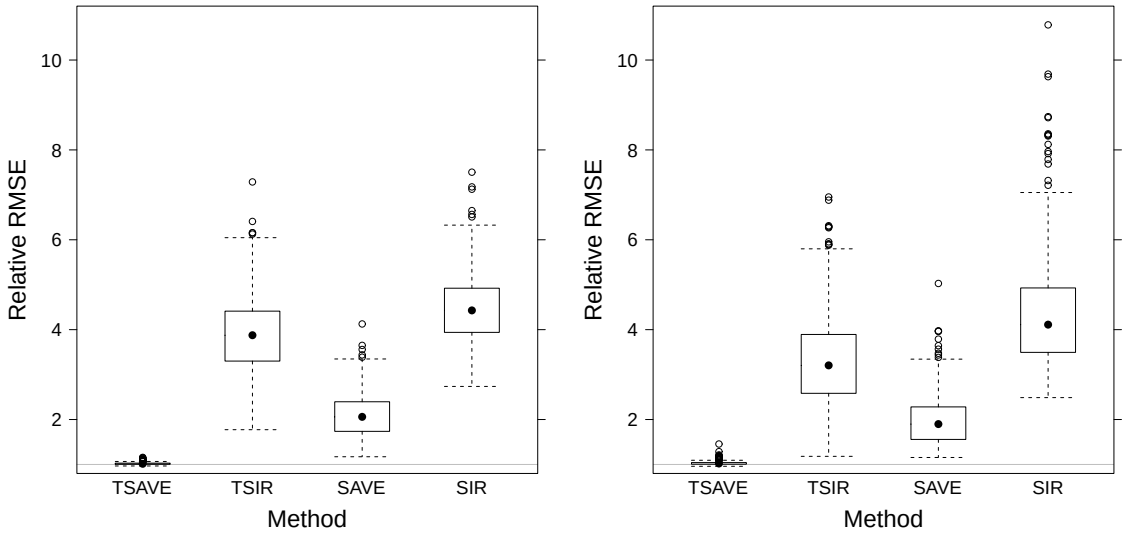

Figure 15. Model D with biggest values strategy. Relative RMSE values compared to Oracle estimator with $\phi = 0.2$ (left panel) and $\phi = 0.8$ (right panel).

The innovations are standard normal unless otherwise stated. Components included here are

z_1: AR(1) with $\phi = -0.2$
z_2: AR(1) with $\phi = 0.8$ with heavy-tailed t_4 innovations
z_3: GARCH(1,1) with $\alpha = 0.05$ and $\beta = 0.93$
z_4: AR(1) with $\phi = 0.6$ with light-tailed $U(-1, 1)$ innovations
z_5: AR(1) with $\phi = 0.98$
z_6: ARCH(2) with $\alpha_1 = 0.3$ and $\alpha_2 = 0.4$
z_7: GARCH(1,1) with $\alpha = 0.1$ and $\beta = 0.8$

z_8: ARMA(1,1) with $\phi = 0.3$ and $\theta = -0.6$
z_9: iid $N(0,1)$
z_{10}: iid t_4

All are standardized to meet the requirements for $\mathbf{z}$. The response is created as

$$y_t = z_{1,t-1} + z_{2,t-2} + 0.5z_{3,t-4} + \epsilon_t,$$

where $\epsilon_t \sim N(0,1)$. In order to examine what happens if y_t is not well approximated in the respective spline class, we have used predictions based both on cubic ('optimal') and quadratic ('non-optimal') splines for all the methods used. The RMSE values with $T = 3000$ in Figure 16 reveal that TSAVE with optimal prediction models produces clearly the best results. Also TSAVE with non-optimal prediction model gives very slightly better results than TSIR with optimal prediction model and much better than TSIR with non-optimal predictions. All the time series versions behave better than the vectorized iid versions.

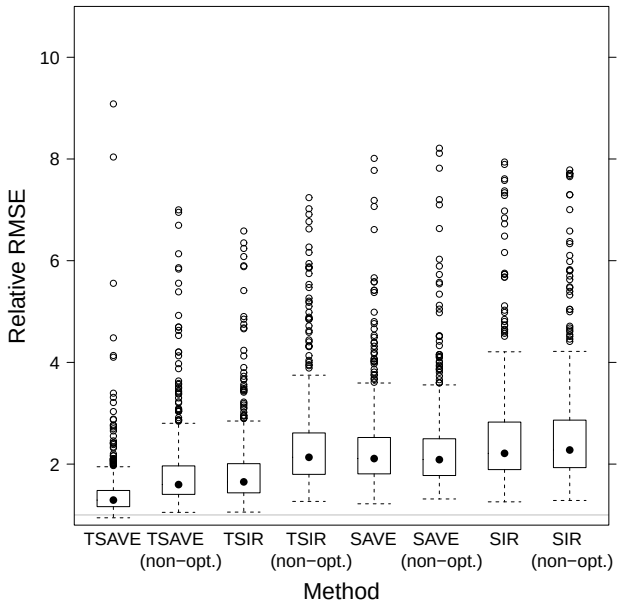

Figure 16. Big setting with biggest values strategy. Relative RMSE values compared to Oracle estimator.

For $T = 2000$ and $T = 5000$ the above is also clearly true. For $T = 1000$ both choices are about equally good. With $T = 500$ it seems that the 'non-optimal' prediction is better. It might be safer not to use 'too accurate' prediction models when time series lengths are short and the number of dimensions large. Also in shorter time series the vectorized SIR starts to produce better and better results compared to vectorized SAVE. Figures are included in the supplementary material.

We have also compared the computation times of SIR, SAVE, TSIR and TSAVE methods considering different lags numbers, dimensions and time series lengths. According to the conducted simulations, the time series versions seem to outperform the iid versions especially when p is small, time series length T increases and the number of lags increase. Detailed results of the simulations are available in the supplementary material.

5. Final comments

SAVE and hybrids that include SIR and SAVE parts are well established for the iid case. After [7] introduced TSIR as a time series extension for SIR, we suggested in this paper TSAVE and TSSH as time series extensions of SAVE and the hybrid of SIR and SAVE, respectively. We demonstrated that these methods are superior for supervised dimension reduction in a time series context than applying their iid counterparts to the explaining variables and their lagged values. We furthermore explored further the strategies for components and lag selection in the time series case suggested in [7] and could now give some general recommendations how to apply TSIR, TSAVE and TSSH in practice. Not so surprisingly many of the recommendations of iid methods apply also in the time series context. Also, while in [7] only $H = 10$ was used as the number of slices, here we conducted a simulation study to give some guidelines for choosing H for TSIR and TSAVE.

The popularity of the iid versions of SIR, SAVE and hybrids of them also led to the introduction of modified versions of SAVE, like CSAVE [19] and ESAVE [11] and hybrids between these and SIR, or other versions of hybrids with SIR like $SIRII_a$ (see the rejoinder of [8]). Future work will be to investigate how these modified versions can be moved to a time series framework and if they are improvements compared to TSIR, TSAVE and TSSH.

Acknowledgements

We wish thank the Associate Editor and the reviewers for careful reading of the paper and their valuable comments.

Funding

The work of M. Matilainen and H. Oja was supported by the Academy of Finland under Grant 268703; and the work of K. Nordhausen was supported by the CRoNoS COST Action IC1408.

References

[1] Ma Y, Zhu L. A review on dimension reduction. International Statistics Review. 2013;81:134–150.

[2] Ensor KB. Time series factor models. Wiley Interdisciplinary Reviews: Computational Statistics. 2013;5(2):97–104.

[3] Xia Y, Tong H, Li WK, Zhu LX. An adaptive estimation of dimension reduction space. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2002;64(3):363–410.

[4] Becker C, Fried R. Sliced inverse regression for high-dimensional time series. In: Schwaiger M, Opitz O, editors. Exploratory data analysis in empirical research. Springer Berlin Heidelberg; 2003. p. 3–11.

[5] Barbarino A, Bura E. Forecasting with sufficient dimension reductions. Washington: Board of Governors of the Federal Reserve System; 2015. Finance and economics discussion series 2015-074.

[6] Barbarino A, Bura E. A unified framework for dimension reduction in forecasting. Washington: Board of Governors of the Federal Reserve System; 2017. Feds working paper no. 2017–004.
[7] Matilainen M, Croux C, Nordhausen K, Oja H. Supervised dimension reduction for multivariate time series. Econometrics and Statistics. 2017;4:57–69.
[8] Li KC. Sliced inverse regression for dimension reduction. Journal of the American Statistical Association. 1991;86(414):316–327.
[9] Cook R, Weisberg S. Sliced inverse regression for dimension reduction: Comment. Journal of the American Statistical Association. 1991;86:328–332.
[10] Cook R. SAVE: A method for dimension reduction and graphics in regression. Communications in Statistics – Theory and Methods. 2000;29:2109–2121.
[11] Zhu LX, Ohtaki M, Li Y. On hybrid methods of inverse regression-based algorithms. Computational Statistics & Data Analysis. 2007;51(5):2621–2635.
[12] Nordhausen K, Oja H, Tyler D. Asymptotic and bootstrap tests for subspace dimension. 2017; submitted; Available from: https://arxiv.org/abs/1611.04908v2.
[13] Miettinen J, Taskinen S, Nordhausen K, Oja H. Fourth moments and independent component analysis. Statistical Science. 2015;30:372–390.
[14] Liski E, Nordhausen K, Oja H. Supervised invariant coordinate selection. Statistics: A Journal of Theoretical and Applied Statistics. 2014;4:711–731.
[15] Bura E, Cook R. Estimating the structural dimension of regressions via parametric inverse regression. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2001; 63:393–410.
[16] Cook RD, Critchley F. Identifying regression outliers and mixtures graphically. Journal of the American Statistical Association. 2000;95(451):781–794.
[17] Portier F. An empirical process view of inverse regression. Scandinavian Journal of Statistics. 2016;43(3):827–844.
[18] Luo W, Li B. Combining eigenvalues and variation of eigenvectors for order determination. Biometrika. 2016;103(4):875–887.
[19] Li Y, Zhu LX. Asymptotics for sliced average variance estimation. The Annals of Statistics. 2007;35(1):41–69.
[20] Cook RD, Ni L. Sufficient dimension reduction via inverse regression. Journal of the American Statistical Association. 2005;100(470):410–428.
[21] Zhu LX, Ng KW. Asymptotics of sliced inverse regression. Statistica Sinica. 1995;5:727–736.
[22] Ye Z, Weiss RE. Using the bootstrap to select one of a new class of dimension reduction methods. Journal of the American Statistical Association. 2003;98(464):968–979.
[23] Shaker AJ, Prendergast LA. Iterative application of dimension reduction methods. Electronic Journal of Statistics. 2011;5:1471–1494.
[24] Saracco J. Pooled slicing methods versus slicing methods. Communications in Statistics – Simulation and Computation. 2001;30(3):489–511.
[25] Gannoun A, Saracco J. An asymptotic theory for SIR_α method. Statistica Sinica. 2003; 13(2):297–310.
[26] Cardoso JF, Souloumiac A. Jacobi angles for simultaneous diagonalization. SIAM Journal on Matrix Analysis and Applications. 1996;17:161–164.
[27] Illner K, Miettinen J, Fuchs C, Taskinen S, Nordhausen K, Oja H, Theis FJ. Model selection using limiting distributions of second-order blind source separation algorithms. Signal Processing. 2015;113:95–103.
[28] Theis FJ, Inouye Y. On the use of joint diagonalization in blind signal processing. In: IEEE International Symposium on Circuits and Systems. IEEE; 2006. p. 3589–3593.
[29] Chabriel G, Kleinsteuber M, Moreau E, Shen H, Tichavsky P, Yeredor A. Joint matrices decompositions and blind source separation: A survey of methods, identification, and applications. IEEE Signal Processing Magazine. 2014;31(3):34–43.
[30] Miettinen J, Nordhausen K, Oja H, Taskinen S. Deflation-based separation of uncorrelated stationary time series. Journal of Multivariate Analysis. 2014;123:214–227.
[31] Matilainen M, Nordhausen K, Oja H. New independent component analysis tools for time

series. Statistics & Probability Letters. 2015;105:80–87.
[32] Miettinen J, Illner K, Nordhausen K, Oja H, Taskinen S, Theis F. Separation of uncorrelated stationary time series using autocovariance matrices. Journal of Time Series Analysis. 2016; 37(3):337–354.
[33] Miettinen J, Nordhausen K, Taskinen S. Blind source separation based on joint diagonalization in R: The packages JADE and BSSasymp. Journal of Statistical Software. 2017; 76(2):1–31.
[34] Matilainen M, Croux C, Miettinen J, Nordhausen K, Oja H, Taskinen S. tsBSS: Tools for blind source separation and supervised dimension reduction for time series. 2018; R package version 0.5.2; Available from: https://CRAN.R-project.org/package=tsBSS.