Supervised dimension reduction for multivariate time series

M. Matilainen, C. Croux, K. Nordhausen, H. Oja

PII: S2452-3062(17)30034-5
DOI: 10.1016/j.ecosta.2017.04.002
Reference: ECOSTA 58

To appear in: *Econometrics and Statistics*

Received date: 21 March 2016
Revised date: 19 January 2017
Accepted date: 3 April 2017

Please cite this article as: M. Matilainen, C. Croux, K. Nordhausen, H. Oja, Supervised dimension reduction for multivariate time series, *Econometrics and Statistics* (2017), doi: 10.1016/j.ecosta.2017.04.002

Supervised dimension reduction for multivariate time series

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

[a] *Department of Mathematics and Statistics, FI-20014 University of Turku, Finland*
[b] *Faculty of Economics and Business, KU Leuven, Belgium*

Abstract

A regression model where the response as well as the explaining variables are time series is considered. A general model which allows supervised dimension reduction in this context is suggested without considering the form of dependence. The method for this purpose combines ideas from sliced inverse regression (SIR) and blind source separation methods to obtain linear combinations of the explaining time series which are ordered according to their relevance with respect to the response. The method gives also an indication of which lags of the linear combinations are of importance. The method is demonstrated using simulations and a real data example.

Keywords: Blind source separation, joint diagonalization, prediction, sliced inverse regression, SOBI

1. Introduction

In many fields of application many variables are measured regularly over time. Sometimes one variable is of main interest and its relationship to the other variables should be modelled or its future values should be predicted based on the other series. For example in the field of macro-economics usually many possible explaining time series are available which often are also highly correlated and therefore not all might be needed for the modelling or forecasting. When the number of explaining time series is large, modelling is challenging and *supervised dimension reduction* methods can help to reduce the dimension and make therefore visualization and modelling much easier. In supervised dimension reduction the joint distribution of the response and the explaining variables is used in dimension reduction for the explaining variables.

For cross-sectional data there are many supervised dimension reduction methods available (for a recent review see for example Ma and Zhu (2013)) but somehow this has not yet been much considered in the

*Corresponding author
Email address: (M. Matilainen)

time series case. Here the situation is also more difficult. The effect of some explaining time series might manifest itself only in some delayed way. Therefore also past values of the explaining time series have to be considered when reducing the dimension.

In this paper we combine ideas from cross-sectional dimension reduction methods and blind source separation methods for time series to introduce a model that makes such an approach possible. The proposed method suggests linear combinations of the explaining time series that are of most interest when modelling the response series, under weak assumptions. The procedure gives also an idea which of the lags of the linear combinations are relevant.

The closest to our approach are probably Becker and Fried (2003) and Barbarino and Bura (2015) which both use supervised dimension reduction. But we would like to emphasize that our approach is very different from dynamic factor models (see e.g. Forni et al. (2005); Kim and Swanson (2014) and references therein) as there the factors are derived in an unsupervised fashion. Also very recently Fan et al. (2015) have presented a method that uses the idea of SIR and builds a factor model for forecasting. Their approach is cross-sectional unlike ours.

In the paper we use the following notation. We write $\boldsymbol{x} = (\boldsymbol{x}_t)_{t\in\mathbb{Z}}$ for a p-variate time series with index set $\mathbb{Z} = \{0, \pm1, \pm2, \ldots\}$. The term time series, as used here, means both the data as well as the random process that produces the data. For any $p \times p$ matrix $\boldsymbol{A}$ and for any vector $\boldsymbol{b}$, $\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}$ is a time series

$$\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b} = (\boldsymbol{A}\boldsymbol{x}_t + \boldsymbol{b})_{t\in\mathbb{Z}}.$$

As in the case of univariate time series, we say that the multivariate time series $\boldsymbol{x}$ is (strictly) stationary if $(\boldsymbol{x}_{t_1}, \ldots, \boldsymbol{x}_{t_k})' \sim (\boldsymbol{x}_{t_1+s}, \ldots, \boldsymbol{x}_{t_k+s})'$ for all $s, t_1, \ldots, t_k \in \mathbb{Z}$ and weakly stationary if $E(\boldsymbol{x}_t) =: \mu$ (constant) for all $t \in \mathbb{Z}$ and the cross-covariances $\boldsymbol{\Sigma}_s := \mathrm{COV}(\boldsymbol{x}_t, \boldsymbol{x}_{t+s})$, $s, t \in \mathbb{Z}$, exist and depend only on s. For a $p \times q$ matrix $\boldsymbol{A}$, $\|\boldsymbol{A}\| = \sqrt{\sum_{i=1}^{p}\sum_{j=1}^{q} \boldsymbol{A}_{ij}^2}$ denotes the Frobenius norm. For a $p \times p$ matrix $\boldsymbol{A}$, $\mathrm{diag}(\boldsymbol{A})$ is a diagonal matrix with the same diagonal elements as $\boldsymbol{A}$, and $\mathrm{off}(\boldsymbol{A}) = \boldsymbol{A} - \mathrm{diag}(\boldsymbol{A})$. For a p-vector $\boldsymbol{x}$ and a $p \times k$ matrix $\boldsymbol{A}$, $k \le p$, $\boldsymbol{P}_{\boldsymbol{A}} = \boldsymbol{A}(\boldsymbol{A}'\boldsymbol{A})^{-1}\boldsymbol{A}'$ is the projection matrix to the linear subspace $\mathbb{S}_{\boldsymbol{A}}$ of $\mathbb{R}^p$ spanned by the columns of $\boldsymbol{A}$.

The structure of the paper goes as follows. Section 2 deals with sliced inverse regression for iid observations and in Section 3 we generalize it to a time series context. In Section 4 we present an example of how our algorithm works with simulated data. Section 5.1 discusses the prediction problem. In Section 5.2 we present a simulation study and finally in Section 5.3 we show a real data example.

2. Sliced inverse regression (SIR) for iid observations

In sliced inverse regression (SIR) (Li, 1991), the dependence between the p-variate random vector $\boldsymbol{x}$ and a univariate response or target variable y is considered. The goal is to find a $k \times p$ *signal separation matrix* $\boldsymbol{\Gamma}$, $k \ll p$ such that $\boldsymbol{x} \perp\!\!\!\perp y | \boldsymbol{\Gamma x}$, meaning that $\boldsymbol{x}$ and y are independent conditional on $\boldsymbol{\Gamma x}$. The matrix $\boldsymbol{\Gamma}$ is thus used to separate the signal part $\boldsymbol{\Gamma x}$ from the noise part $\boldsymbol{x}|\boldsymbol{\Gamma x}$ in the analysis of dependence between $\boldsymbol{x}$ and y. The idea is to find $\boldsymbol{\Gamma}$ with "minimal" or central dimension reduction subspace $\mathbb{S}_{\boldsymbol{\Gamma}}$. The separation is possible under the following *blind source separation model* for the joint distribution of $\mathbf{x}$ and y.

Assumption 1. Assume that the random vector $\boldsymbol{x} \in \mathbb{R}^p$ is generated by

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

where $\boldsymbol{\mu} \in \mathbb{R}^p$ is a location center, $\boldsymbol{\Omega} \in \mathbb{R}^{p \times p}$ is a full-rank mixing matrix, and $\boldsymbol{z} = \left(\boldsymbol{z}^{(1)\prime}, \boldsymbol{z}^{(2)\prime}\right)'$ with subvectors $\boldsymbol{z}^{(1)} \in \mathbb{R}^k$ and $\boldsymbol{z}^{(2)} \in \mathbb{R}^{p-k}$ satisfies

(A1) $\mathrm{E}(\boldsymbol{z}) = \mathbf{0}$ and $\mathrm{COV}(\boldsymbol{z}) = \boldsymbol{I}_p$, and

(A2) $(y, \boldsymbol{z}^{(1)\prime})' \perp\!\!\!\perp \boldsymbol{z}^{(2)}$

There are however several ambiguities in the model. First note that the vectors $\boldsymbol{z}^{(1)}$ and $\boldsymbol{z}^{(2)}$, and therefore also $\boldsymbol{\Omega}$, are not fully identifiable as the assumptions also hold true for

$$\boldsymbol{z}^{(*1)} = \boldsymbol{U}_1 \boldsymbol{z}^{(1)} \text{ and } \boldsymbol{z}^{(*2)} = \boldsymbol{U}_2 \boldsymbol{z}^{(2)}$$

and

$$\boldsymbol{\Omega}^* = \boldsymbol{\Omega} \begin{pmatrix} \boldsymbol{U}_1' & 0 \\ 0 & \boldsymbol{U}_2' \end{pmatrix}$$

for any orthogonal matrices $\boldsymbol{U}_1 \in \mathbb{R}^{k \times k}$ and $\boldsymbol{U}_2 \in \mathbb{R}^{(p-k)\times(p-k)}$, respectively. This does not cause a problem in the subspace estimation, however. Also, there may be several divisions $\boldsymbol{z} = \left(\boldsymbol{z}^{(1)\prime}, \boldsymbol{z}^{(2)\prime}\right)'$ such that (A1) and (A2) are true. We therefore assume that the division in Assumption 1 is the unique one (up to rotation) having the smallest k.

In the literature of sliced inverse regression, the assumption (A2) is usually replaced by two conditions

$$\boldsymbol{z}^{(2)} \perp\!\!\!\perp y | \boldsymbol{z}^{(1)} \text{ and } \mathrm{E}(\boldsymbol{z}^{(2)} | \boldsymbol{z}^{(1)}) = \mathbf{0} \text{ (a.s.)},$$

where the latter condition is the so called linearity assumption (indicating a kind of weak independence between $\boldsymbol{z}^{(1)}$ and $\boldsymbol{z}^{(2)}$) (Li, 1991; Cook and Weisberg, 1991). It is straightforward to see that (A2) implies these conditions which in turn imply the key result

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

Remark 1. Let $\mathbb{S}_1, \ldots, \mathbb{S}_H$ be H disjoint intervals (slices) such that $\mathbb{R} = \mathbb{S}_1 + \ldots + \mathbb{S}_H$ and let $y^{sl} := \sum_{h=1}^{H} y_h 1_{y \in \mathbb{S}_h}$ for some choices $y_h \in \mathbb{S}_h$, $h = 1, \ldots, H$. The random variable y^{sl} can then be seen as a discrete approximation of continuous random variable y. Then naturally also

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

and $\mathrm{COV}(\mathrm{E}(\boldsymbol{z}^{(1)}|y^{sl}))$ does not depend on the specific choices of $y_h \in \mathbb{S}_h$, $h = 1, \ldots, H$. The term sliced inverse regression (SIR) then just refers to the use of the inverse regression $\mathrm{E}(\boldsymbol{z}|y^{sl})$ in the analysis of the data. In practice the slices are often chosen so that $\mathbb{P}(y \in \mathbb{S}_h) = \frac{1}{H}$, $h = 1, \ldots, H$, with $H = 10$, for example.

Let $\boldsymbol{\Gamma} \in \mathbb{R}^{k \times p}$ be the matrix of the first k rows of $\boldsymbol{\Omega}^{-1}$. The goal is, based on iid observations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, to find an estimate of $\boldsymbol{\Gamma}$, or rather, the estimate of the k-variate subspace spanned by the rows of $\boldsymbol{\Gamma}$ and given by the projection matrix $\boldsymbol{P}_{\boldsymbol{\Gamma}} = \boldsymbol{\Gamma}'(\boldsymbol{\Gamma}\boldsymbol{\Gamma}')^{-1}\boldsymbol{\Gamma}$. The indeterminacy of $\boldsymbol{\Omega}$ discussed above implies that $\boldsymbol{\Gamma}$ is unique only up to pre-multiplication by a $k \times k$ orthogonal matrix, but $\boldsymbol{P}_{\boldsymbol{\Gamma}}$ does not have any indeterminacy. In practice, however, also k is unknown and has to be estimated from the data. To be more precise in what we are estimating, we define the functional $\boldsymbol{\Gamma} = \boldsymbol{\Gamma}(\mathbf{x}; y)$ using the following steps.

Definition 1. The inverse regression (IR) functional $\boldsymbol{\Gamma}(\boldsymbol{x}; y)$ at the joint distribution of $(y, \boldsymbol{x}')'$ is defined as follows.

1. Standardize $\boldsymbol{x}$ and write $\boldsymbol{x}^{st} := \mathrm{COV}(\boldsymbol{x})^{-1/2}(\boldsymbol{x} - \mathrm{E}(\boldsymbol{x}))$.
2. Find the $k \times p$ matrix $\boldsymbol{W} = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_k)'$, with orthonormal rows $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_k$, that maximizes
$$\left\| \mathrm{diag}\left(\boldsymbol{W}\mathrm{COV}(\mathrm{E}(\boldsymbol{x}^{st}|y))\boldsymbol{W}' \right) \right\|^2 = \sum_{i=1}^{k} \left[\boldsymbol{w}_i{}'\mathrm{COV}(\mathrm{E}(\boldsymbol{x}^{st}|y))\boldsymbol{w}_i \right]^2.$$
3. The value of the functional is then $\boldsymbol{\Gamma}(\boldsymbol{x}; y) = \boldsymbol{W}\mathrm{COV}(\boldsymbol{x})^{-1/2}$.

The functional $\boldsymbol{\Gamma}(\boldsymbol{x}; y^{sl})$ is called the sliced inverse regression (SIR) functional.

First note that, in regular sliced inverse regression, the approximation $\mathrm{COV}(\mathrm{E}(\boldsymbol{x}^{st}|y^{sl}))$ is easier to implement for practical data analysis. Under our model assumption, $\boldsymbol{\Gamma}(\boldsymbol{x}; y)$ is unique if the k eigenvalues of $\mathrm{COV}(\mathrm{E}(\boldsymbol{z}^{(1)}|y))$ are distinct and it provides the matrix of the first k rows of $\boldsymbol{\Omega}^{-1}$ pre-multiplied by a well-specified orthogonal matrix. This can be seen as follows. First, it can be shown that $\boldsymbol{x}^{st} = \boldsymbol{U}\boldsymbol{z}$ for some

$p \times p$ orthogonal matrix $\boldsymbol{U}$, see for example Miettinen et al. (2015). If $\mathrm{COV}(\mathrm{E}(\boldsymbol{z}^{(1)}|y))$ has the eigenvector-eigenvalue decomposition $\boldsymbol{V}_1\boldsymbol{\Lambda}_1\boldsymbol{V}_1'$ then

$$\begin{aligned}\mathrm{COV}(\mathrm{E}(\boldsymbol{x}^{st}|y)) = \boldsymbol{U}\mathrm{COV}(\mathrm{E}(\boldsymbol{z}|y))\mathbf{U}' &= \boldsymbol{U}_1\mathrm{COV}(\mathrm{E}(\boldsymbol{z}^{(1)}|y))\boldsymbol{U}_1' \\ &= \boldsymbol{U}_1\boldsymbol{V}_1\boldsymbol{\Lambda}_1\boldsymbol{V}_1'\boldsymbol{U}_1',\end{aligned}$$

where $\boldsymbol{U}_1$ is a $p \times k$ matrix of the first k columns of $\boldsymbol{U}$. The maximizer of $\left\|\mathrm{diag}\left(\boldsymbol{W}\mathrm{COV}(\mathrm{E}(\boldsymbol{x}^{st}|y))\boldsymbol{W}'\right)\right\|^2$ is therefore $\boldsymbol{W} = (\boldsymbol{U}_1\boldsymbol{V}_1)'$ and, finally,

$$\boldsymbol{\Gamma}(\boldsymbol{x}; y)\boldsymbol{x} = \boldsymbol{V}_1'\boldsymbol{U}_1'\boldsymbol{U}\boldsymbol{z} = \boldsymbol{V}_1'\boldsymbol{z}^{(1)}.$$

The components of $\boldsymbol{\Gamma}(\boldsymbol{x}; y)\boldsymbol{x}$ are standardized and the components of $E(\boldsymbol{\Gamma}(\boldsymbol{x}; y)\boldsymbol{x}|y)$ are uncorrelated and ordered according to their variances $\lambda_1 \geq \ldots \geq \lambda_k$. The larger the variance of $(E(\boldsymbol{\Gamma}(\boldsymbol{x}; y)\boldsymbol{x}|y))_i$, the stronger is the dependence between $(\boldsymbol{\Gamma}(\boldsymbol{x}; y)\boldsymbol{x})_i$ and response y, $i = 1, \ldots, k$. For sliced y^{sl}, the variances simply provide the ANOVA type comparisons for between and within slices variations. Note that, under our assumptions, the $p - k$ smallest eigenvalues of $\mathrm{COV}(\mathrm{E}(\boldsymbol{x}^{st}|y))$ are all zero.

Let $(y_1, \boldsymbol{x}_1')', \ldots, (y_n, \boldsymbol{x}_n')'$ be a random sample of size n from the joint distribution of $(y, \boldsymbol{x}')'$ satisfying Assumption 1. The estimate $\widehat{\boldsymbol{\Gamma}}$ is then obtained if the above procedure is applied to the empirical distribution of $(y_1, \boldsymbol{x}_1')', \ldots, (y_n, \boldsymbol{x}_n')'$: First, use the sample mean vector $\bar{\boldsymbol{x}}$ and sample covariance matrix $\boldsymbol{S}$ to standardize the $\boldsymbol{x}$-observations. Second, choose slices for the empirical distribution of the y-variable, that is, use $(y_1^{sl}, \boldsymbol{x}_1')', \ldots, (y_n^{sl}, \boldsymbol{x}_n')'$ instead of $(y_1, \boldsymbol{x}_1')', \ldots, (y_n, \boldsymbol{x}_n')'$. Third, find eigenvectors and eigenvalues of the empirical version of $\mathrm{COV}(\mathrm{E}(\boldsymbol{x}^{st}|y^{sl}))$ and the first k eigenvectors give you $\widehat{\boldsymbol{W}}$. Fourth, the final estimate is $\widehat{\boldsymbol{\Gamma}} = \widehat{\boldsymbol{W}}\boldsymbol{S}^{-1/2}$. Note that the approach also suggests tests and estimates for the true value of k based on the eigenvalues $\hat{\lambda}_1 \geq \ldots \geq \hat{\lambda}_p$ of the estimate of $\mathrm{COV}(\mathrm{E}(\boldsymbol{x}^{st}|y^{sl}))$. See also Liski et al. (2014) for further discussion on this type of approach for supervised dimension reduction.

3. Sliced Inverse Regression (SIR) for time series

3.1. Blind source separation model for SIR

We consider the following blind source separation model for the joint distribution of a p-variate time series $\boldsymbol{x}$ and the target time series y, that is, for the observed time series

$$(y, \boldsymbol{x}')' = \left((y_t, x_{1t}, \ldots, x_{pt})'\right)_{t\in\mathbb{Z}}.$$

We then have the following.

Assumption 2. Assume that the p-variate time series $\boldsymbol{x}$ is generated by

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

where $\boldsymbol{\mu} \in \mathbb{R}^p$ is a location center, $\boldsymbol{\Omega} \in \mathbb{R}^{p \times p}$ is a full-rank mixing matrix, and $z = \left(z^{(1)\prime}, z^{(2)\prime}\right)'$ is a p-variate time series with k- and $(p-k)$-variate subseries $z^{(1)}$ and $z^{(2)}$. We further assume that

$$(y, z')' = \left(y, z^{(1)\prime}, z^{(2)\prime}\right)' = \left((y_t, z_{1t}, \ldots, z_{pt})'\right)_{t \in \mathbb{Z}}$$

is a stationary $(p+1)$-variate time series, that satisfies

(A1) $\mathrm{E}(z_t) = \mathbf{0}$ and $\mathrm{COV}(z_t) = \boldsymbol{I}_p$, and
(A2) $\left(y, z^{(1)\prime}\right)' \perp\!\!\!\perp z^{(2)}$.

Note that (A2) implies that $\left(y_{t_1}, z_{t_1}^{(1)\prime}\right)' \perp\!\!\!\perp z_{t_2}^{(2)}$ or $\left(y_{t_1+s}, z_{t_1}^{(1)\prime}\right)' \perp\!\!\!\perp z_{t_2}^{(2)}$ for all $t_1, t_2, s \in \mathbb{Z}$. Therefore the classical sliced inverse regression (SIR) for marginal distributions of $(y_t, \boldsymbol{x}_t')'$ could also be used to find $z^{(1)}$. The classical SIR however uses only the cross-sectional information and ignores the information coming from the dependencies between the series at different time points. It is important to utilize this information as we often have prediction models that use also information on temporal dependence (See Section 5.1). Note that our model formulation does not separate between independent and dependent explaining series for the modelling of the y series. All the dependence between the $\boldsymbol{x}$ and y series, as a whole, goes through $z^{(1)}$, and the aim is simply to separate between the signal part $z^{(1)}$ and the noise part $z^{(2)}$ of z.

Also in this latent time series model the series $z^{(1)}$ and $z^{(2)}$ are identifiable only up to pre-multiplication by orthogonal matrices. Again we assume that the division $z = \left(z^{(1)\prime}, z^{(2)\prime}\right)'$ is the unique one with the smallest k. As in the iid case, the assumption (A2) implies two conditions

$$z^{(2)} \perp\!\!\!\perp y | z^{(1)} \quad \text{and} \quad \mathrm{E}(z_{t+s}^{(2)} | z_t^{(1)}) = \mathbf{0} \text{ (a.s.) for all } s \in \mathbb{Z},$$

which in turn imply that the cross-covariance matrix

$$\boldsymbol{\Sigma}_s := \mathrm{COV}\left(\mathrm{E}(z_t | y_{t+s})\right) = \begin{pmatrix} \mathrm{COV}\left(\mathrm{E}(z_t^{(1)} | y_{t+s})\right) & 0 \\ 0 & 0 \end{pmatrix} \text{ for all } s \in \mathbb{Z}.$$

Again for H disjoint slices $\mathbb{S}_1, \ldots, \mathbb{S}_H$ such that $\mathbb{R} = \mathbb{S}_1 + \ldots + \mathbb{S}_H$, one can construct a discrete valued time series y^{sl} such that $y_t^{sl} := \sum_{h=1}^H y_h 1_{y_t \in \mathbb{S}_h}$ for some choices $y_h \in \mathbb{S}_h$, $h = 1, \ldots, H$. Then $\mathrm{COV}\left(\mathrm{E}(z_t | y_{t+s}^{sl})\right)$ has the same structure as above and its value does not depend on the specific choices of $y_h \in \mathbb{S}_h$, $h = 1, \ldots, H$.

3.2. The separation matrix functional and estimate

As in the iid case, the goal is, based on an observed series $(y_1, \boldsymbol{x}_1')', \ldots, (y_T, \boldsymbol{x}_T')'$ following Assumption 2, to find an estimate of the first k rows of $\boldsymbol{\Omega}^{-1}$. Then $E(\boldsymbol{x}_t) = \boldsymbol{\mu}$ and $\mathrm{COV}(\boldsymbol{x}_t) = \boldsymbol{\Omega}\boldsymbol{\Omega}' =: \boldsymbol{\Sigma}$. To fix the indeterminacy in estimation, we define the functional $\boldsymbol{\Gamma} = \boldsymbol{\Gamma}(\mathbf{x}; y)$ using the following steps.

Definition 2. The inverse regression (IR) functional $\boldsymbol{\Gamma}(\boldsymbol{x}; y)$ for a stationary time series $(y, \boldsymbol{x}')'$ is obtained as follows.

1. Standardize $\boldsymbol{x}$ and write $\boldsymbol{x}^{st} := \mathrm{COV}(\boldsymbol{x}_t)^{-1/2}(\boldsymbol{x} - \mathrm{E}(\boldsymbol{x}_t))$.
2. Find the $k \times p$ matrix $\boldsymbol{W} = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_k)'$ with orthonormal rows $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_k$ that maximizes
$$\sum_{s\in\mathcal{S}} \left\| \mathrm{diag}\left(\boldsymbol{W}\mathrm{COV}(\mathrm{E}(\boldsymbol{x}_t^{st}|y_{t+s}))\boldsymbol{W}'\right)\right\|^2 = \sum_{i=1}^{k}\sum_{s\in\mathcal{S}} \left[\boldsymbol{w}_i{}'\mathrm{COV}(\mathrm{E}(\boldsymbol{x}_t^{st}|y_{t+s}))\boldsymbol{w}_i\right]^2, \quad (1)$$
for a chosen set of lags $\mathcal{S} \subset \mathbb{Z}_+$.
3. The value of the functional is then $\boldsymbol{\Gamma}(\boldsymbol{x}; y) = \boldsymbol{W}\mathrm{COV}(\boldsymbol{x}_t)^{-1/2}$.

The functional $\boldsymbol{\Gamma}(\boldsymbol{x}; y^{sl})$ is called the sliced inverse regression (SIR) functional. The estimate $\hat{\boldsymbol{\Gamma}}$ of the population value $\boldsymbol{\Gamma}(\boldsymbol{x}; y^{sl})$ is obtained by replacing the covariances, expectations and conditional expectations by their sample counterparts.

Write
$$\lambda_{is} = \left(\boldsymbol{w}_i{}'\mathrm{COV}(\mathrm{E}(\boldsymbol{x}_t^{st}|y_{t+s}))\boldsymbol{w}_i\right)^2, \quad i = 1, \ldots, k;\ s \in \mathcal{S}.$$
Under our model assumption, $\boldsymbol{\Gamma}(\boldsymbol{x}; y)$ is unique if $\lambda_{i\cdot} = \sum_{s\in\mathcal{S}} \lambda_{is}$, $i = 1, \ldots, k$, are distinct. We see it as follows. As in the iid case, $\boldsymbol{x}^{st} = \boldsymbol{U}\boldsymbol{z}$ for some $p \times p$ orthogonal matrix $\boldsymbol{U}$. If $\mathrm{COV}(\mathrm{E}(\boldsymbol{z}_t^{(1)}|y_{t+s})) = \boldsymbol{V}_s\boldsymbol{\Lambda}_s\boldsymbol{V}_s'$, $s \in \mathcal{S}$, (the eigenvector-eigenvalue decompositions) then

$$\begin{aligned}\mathrm{COV}\left(\mathrm{E}(\boldsymbol{x}_t^{st}|y_{t+s})\right) = \boldsymbol{U}\mathrm{COV}\left(\mathrm{E}(\boldsymbol{z}_t|y_{t+s})\right)\mathbf{U}' &= \boldsymbol{U}_1\mathrm{COV}(\mathrm{E}(\boldsymbol{z}_t^{(1)}|y_{t+s}))\boldsymbol{U}_1' \\ &= \boldsymbol{U}_1\boldsymbol{V}_s\boldsymbol{\Lambda}_s\boldsymbol{V}_s'\boldsymbol{U}_1', \quad s \in \mathcal{S},\end{aligned}$$

and $\boldsymbol{U}_1$ is a $p \times k$ matrix of the first k columns of $\boldsymbol{U}$. The maximizer of
$$\sum_{s\in\mathcal{S}} \left\| \mathrm{diag}\left(\boldsymbol{W}\mathrm{COV}(\mathrm{E}(\boldsymbol{x}_t^{st}|y_{t+s}))\boldsymbol{W}'\right)\right\|^2 = \sum_{s\in\mathcal{S}} \left\|\mathrm{diag}\left(\boldsymbol{W}\boldsymbol{U}_1\boldsymbol{V}_s\boldsymbol{\Lambda}_s\boldsymbol{V}_s'\boldsymbol{U}_1'\boldsymbol{W}'\right)\right\|^2$$
is therefore $\boldsymbol{W} = \boldsymbol{V}'\boldsymbol{U}_1'$ for some $k \times k$ orthogonal matrix $\boldsymbol{V}$, and can be computed using joint diagonalization, see Cardoso and Souloumiac (1996). Finally,
$$\boldsymbol{\Gamma}(\boldsymbol{x}; y)\boldsymbol{x} = \boldsymbol{V}'\boldsymbol{U}_1'\boldsymbol{U}\boldsymbol{z} = \boldsymbol{V}'\boldsymbol{z}^{(1)}.$$

The components of $\boldsymbol{\Gamma}(\boldsymbol{x};y)\boldsymbol{x}$ are standardized and ordered so that $\lambda_{1\cdot} \geq \ldots \geq \lambda_{k\cdot}$. Again, the larger the values $\lambda_{i\cdot}$, the stronger the dependence between time series $(\boldsymbol{\Gamma}(\boldsymbol{x};y)\boldsymbol{x})_i$ and y, $i = 1,\ldots,k$. High values of λ_{is} alone indicate a strong dependence between $(\boldsymbol{\Gamma}(\boldsymbol{x};y)\boldsymbol{x})_{it}$ and y_{t+s}, $i = 1,\ldots,k$, $s \in \mathcal{S}$.

Note that our approach finds a $k \times p$ matrix $\boldsymbol{\Gamma}$ such that $\boldsymbol{x} \perp\!\!\!\perp y | \boldsymbol{\Gamma} x$ and uses the values $\lambda_{i\cdot}$, $i = 1,\ldots,p$, for this separation. ($\lambda_{k+1,\cdot} = \ldots = \lambda_{p\cdot} = 0$). However, due to the temporal dependence in the x-series the values λ_{is}, $i = 1,\ldots,k$ and $s \in \mathcal{S}$, may not advice much in the further choice of the number of lags to be used in a future analysis.

Remark 2. The form (1) is a double sum over all linear combinations and over all lags and therefore allows a deeper analysis of the dependence than an alternative objective function

$$\left\| \operatorname{diag}\left(\boldsymbol{W} \sum_{s\in\mathcal{S}} \operatorname{COV}(\operatorname{E}(\boldsymbol{x}_t^{st}|y_{t+s}))\boldsymbol{W}' \right) \right\|^2 = \sum_{i=1}^{k} \left(\sum_{s\in\mathcal{S}} \boldsymbol{w}_i' \operatorname{COV}(\operatorname{E}(\boldsymbol{x}_t^{st}|y_{t+s}))\boldsymbol{w}_i \right)^2,$$

suggested by an anonymous referee.

Remark 3. We briefly discuss, as requested by the reviewers, the computation and future work needed for the limiting distribution of the estimate $\hat{\boldsymbol{\Gamma}}$, first with known dimension k and number of lags s'. Let $\boldsymbol{S} = \operatorname{COV}(\boldsymbol{x}_t)$ and $\boldsymbol{S}_j = \operatorname{COV}(\boldsymbol{x}_t|y_{t+j})$, $j = 1,\ldots,s'$. Next write $\boldsymbol{R}_j = \boldsymbol{S}^{-1/2}\boldsymbol{S}_j\boldsymbol{S}^{-1/2}$, $j = 1,\ldots,s'$. Then $\boldsymbol{\Gamma}(\boldsymbol{x};y) = \boldsymbol{W}\operatorname{COV}(\boldsymbol{x}_t)^{-1/2}$ where $\boldsymbol{W} = (\mathbf{w}_1,...,\mathbf{w}_k)'$ is the $k \times p$ matrix with orthonormal rows maximizing $\sum_{i=1}^{k}\sum_{j=1}^{s'} \left(\boldsymbol{w}_i'\boldsymbol{R}_j\boldsymbol{w}_i\right)^2$. Using the Lagrange multiplier technique one finds the estimating equations

$$\boldsymbol{W}\boldsymbol{T}' = \boldsymbol{T}\boldsymbol{W}' \text{ and } \boldsymbol{W}\boldsymbol{W}' = \boldsymbol{I}_k$$

where $\boldsymbol{T} = \boldsymbol{T}(\boldsymbol{W}) = (\boldsymbol{t}(\boldsymbol{w}_1),\ldots,\boldsymbol{t}(\boldsymbol{w}_k))'$ with $\boldsymbol{t}(\boldsymbol{w}_i) = \sum_{j=1}^{s'}(\boldsymbol{w}_i'\boldsymbol{R}_j\boldsymbol{w}_i)\boldsymbol{R}_j\boldsymbol{w}_i$, $i = 1,\ldots,k$. The estimating equations suggest a fixed-point algorithm with steps $\boldsymbol{W} \leftarrow (\boldsymbol{T}\boldsymbol{T}')^{-1/2}\boldsymbol{T}$ (providing the same solution as joint diagonalization in Cardoso and Souloumiac (1996), see Illner et al. (2015)) and can be used to find the limiting distribution of $\hat{\boldsymbol{W}}$ and $\hat{\boldsymbol{\Gamma}}$ as follows.

Due to affine equivariance of the estimate is is not a restriction to consider the case with population values $\mathbf{S} = \boldsymbol{I}_p$, $\boldsymbol{W} = (\boldsymbol{I}_k, \mathbf{0})$ and $\lambda_{1\cdot} > \ldots > \lambda_{k\cdot}$. Next we need to assume that the joint limiting distribution of $\sqrt{n}(\hat{\boldsymbol{S}} - \boldsymbol{I}_p)$ and $\sqrt{n}(\hat{\boldsymbol{S}}_j - \boldsymbol{S}_j)$, $j = 1,\ldots,s'$ is known. To find this joint distribution one may need strong model assumptions. Denote $\hat{\boldsymbol{t}}(\boldsymbol{w}_i) = \sum_{j=1}^{s'}(\boldsymbol{w}_i'\hat{\boldsymbol{R}}_j\boldsymbol{w}_i)\hat{\boldsymbol{R}}_j\boldsymbol{w}_i$, $i = 1,\ldots,k$ and $\hat{\boldsymbol{T}}(\boldsymbol{W}) = (\hat{\boldsymbol{t}}(\boldsymbol{w}_1),\ldots,\hat{\boldsymbol{t}}(\boldsymbol{w}_k))'$. The joint limiting distribution of $\hat{\boldsymbol{W}}$ and $\hat{\boldsymbol{T}} = \hat{\boldsymbol{T}}(\hat{\boldsymbol{W}})$ then satisfies

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

and

$$\sqrt{n}(\hat{\boldsymbol{W}} - \boldsymbol{W})\boldsymbol{W}' = -\boldsymbol{W}\sqrt{n}(\hat{\boldsymbol{W}} - \boldsymbol{W})' + o_P(1)$$

and these can further be used to find the joint limiting distribution of $\sqrt{n}(\hat{\boldsymbol{W}} - \boldsymbol{W})$ and $\sqrt{n}(\hat{\boldsymbol{S}} - \boldsymbol{I}_p)$. Finally $\sqrt{n}(\hat{\boldsymbol{\Gamma}} - \boldsymbol{W}) = \sqrt{n}(\hat{\boldsymbol{W}} - \boldsymbol{W}) - \frac{1}{2}\boldsymbol{W}\sqrt{n}(\hat{\boldsymbol{S}} - \boldsymbol{I}_p) + o_P(1)$. See Miettinen et al. (2016) for similar derivations in the Second Order Blind Identification (SOBI) for time series.

For testing the null hypothesis $H_0 : k = k_0$, that is, $\lambda_{1\cdot} \geq \ldots \geq \lambda_{k_0\cdot} > \lambda_{k_0+1\cdot} = \lambda_{p\cdot} = 0$, one can first find a $p \times p$ matrix $\hat{\boldsymbol{W}}$ with $\hat{\lambda}_{1\cdot} > \ldots > \hat{\lambda}_{p\cdot}$ and then use $L = \sum_{i=k_0+1}^{p} \hat{\lambda}_{i\cdot}$ as a test statistic. In the regular SIR for the iid case with n observations, the limiting distribution of nL is a chi square distribution with $(p - k_0)(H - K_0 - 1)$ degrees of freedom. See Nordhausen et al. (2016) and references therein. To find the limiting distribution for L in the time series context is still an open problem. Also, it is not clear how to make inference on s'.

Remark 4. The inverse regression functional $\boldsymbol{\Gamma}(\boldsymbol{x}; y)$ is thus a functional for supervised dimension reduction in the time series context. For unsupervised dimension reduction, assume that $\boldsymbol{x} = \boldsymbol{\Omega}\boldsymbol{z} + \boldsymbol{\mu}$ where $\boldsymbol{z}$ is second order stationary with $\mathrm{E}(\boldsymbol{z}_t) = \boldsymbol{0}$, $\mathrm{COV}(\boldsymbol{z}_t) = \boldsymbol{I}_p$ and diagonal $\mathrm{COV}(\mathrm{E}(\boldsymbol{x}_t^{st}(\boldsymbol{x}_{t+s}^{st})'))$ for all $s = 0, \pm 1, \ldots$. The Second Order Blind Identification (SOBI) functional (Belouchrani et al., 1997) for time series is then $\boldsymbol{\Gamma}(\boldsymbol{x}) = \boldsymbol{W}\mathrm{COV}(\boldsymbol{x}_t)^{-1/2}$ where $\boldsymbol{W}$ is an orthogonal matrix that maximizes

$$\sum_{s \in \mathcal{S}} \left\| \mathrm{diag}\left(\boldsymbol{W}\mathrm{COV}(\mathrm{E}(\boldsymbol{x}_t^{st}(\boldsymbol{x}_{t+s}^{st})'))\boldsymbol{W}'\right)\right\|^2 .$$

For different versions of SOBI, their computation and statistical properties see also Miettinen et al. (2012, 2014, 2016); Taskinen et al. (2016). For other independent component time series models, see for example Shi et al. (2009); Nordhausen (2014); Matilainen et al. (2015, 2016) and references therein.

3.3. The choice of the dimension and the number of lags

Consider next the case with lags in $\mathcal{S} = \{1, \ldots, s\}$ that is natural when y_t-values are predicted with lagged $\boldsymbol{x}$ values $\boldsymbol{x}_{t-1}, \ldots, \boldsymbol{x}_{t-s}$. Let then $\boldsymbol{\Lambda} = (\lambda_{ij})$ be the matrix of

$$\lambda_{ij} = c \cdot \left(\boldsymbol{w}_i'\mathrm{COV}(\mathrm{E}(\boldsymbol{x}_t^{st}|y_{t+j}))\boldsymbol{w}_i\right)^2, \quad i = 1, \ldots, k;\ j = 1, \ldots, s,$$

where $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_k$ are the k rows of $\boldsymbol{W}$ in the functional $\boldsymbol{\Gamma}(\boldsymbol{x}; y) = \boldsymbol{W}\mathrm{COV}(\boldsymbol{x}_t)^{-1/2}$ and c is chosen so that $\sum_{i=1}^{k}\sum_{j=1}^{s}\lambda_{ij} = 1$. Write also $\lambda_{i\cdot} = \sum_{j=1}^{s}\lambda_{ij}$, $i = 1, \ldots, k$, and $\lambda_{\cdot j} = \sum_{i=1}^{k}\lambda_{ij}$, $j = 1, \ldots, s$, for the row and column sums of $\boldsymbol{\Lambda}$, respectively. Assume also that the latent series $\boldsymbol{w}_1'\boldsymbol{x}^{st}, \ldots, \boldsymbol{w}_k'\boldsymbol{x}^{st}$ are ordered so that $\lambda_{1\cdot} \geq \ldots \geq \lambda_{k\cdot}$. The $k \times p$ matrix $\boldsymbol{\Lambda}$ thus provides measures of dependence between

$$y_t \ \text{ and } \ (\boldsymbol{w}_i'\boldsymbol{x}^{st})_{t-j}, \quad i = 1, \ldots, k;\ j = 1, \ldots, s.$$

In the simulations and in the real data example $k = p$ is used. There are several strategies for further choosing between the $k' \cdot s'$ variables $(\boldsymbol{w}_i{}'\boldsymbol{x}^{st})_{t-j}$ in a way that most of the dependence measured by the λ_{ij}'s, say $100\times\pi$ percent, is still remaining.

- Keep the k' first directions and s' first lags, that is,

$$\left(\boldsymbol{w}_i{}'\boldsymbol{x}^{st}\right)_{t-j}, \quad i = 1, \ldots, k'; j = 1, \ldots, s',$$

  with $k' \leq k$ and $s' \leq s$. The strategies then are, for example, as follows.

  1. Keep all lags: Choose $s' = s$ and find smallest k' such that $\sum_{i=1}^{k'} \lambda_{i\cdot} \geq \pi$.
  2. Keep all directions: Choose $k' = k$ and find smallest s' such that $\sum_{j=1}^{s'} \lambda_{\cdot j} \geq \pi$.
  3. Find k' and s' with the smallest product $k's'$ such that $\sum_{i=1}^{k'} \sum_{j=1}^{s'} \lambda_{ij} \geq \pi$.

- Find the smallest number r of elements $(i_1, j_1), \ldots, (i_r, j_r)$ of $\boldsymbol{\Lambda}$ such that $\sum_{k=1}^{r} \lambda_{i_k j_k} \geq \pi$. The chosen variables are then

$$\left(\boldsymbol{w}'_{i_1}\boldsymbol{x}^{st}\right)_{t-j_1}, \ldots, \left(\boldsymbol{w}'_{i_r}\boldsymbol{x}^{st}\right)_{t-j_r}.$$

These strategies are illustrated with the $\hat{\lambda}_{ij}$'s in Table 1 obtained from model B (see Section 4 below) where the first two components are $AR(1)$ models with $\phi = 0.2$. However, for illustration purposes here the time series length is only $T = 1000$ and the innovation distribution is $N(0, 0.1)$. The two dashed lines represent the 80% thresholds when all lags and all directions are kept, respectively. The rectangle with solid lines gives k' and s' for the third strategy. The last strategy gives $r = 11$ with the $\hat{\lambda}_{ij}$'s separated by grey background.

4. Sliced inverse regression in three simulated multivariate time series

The four $\boldsymbol{z}$ series are simulated as follows.

$$\begin{aligned}
\text{First component} &: AR(1) \text{ with } \phi = 0.2 \text{ (or } 0.8).\\
\text{Second component} &: AR(1) \text{ with } \phi = 0.2 \text{ (or } 0.8).\\
\text{Third component} &: ARMA(1,1) \text{ with } \phi = 0.3 \text{ and } \theta = 0.4.\\
\text{Fourth component} &: MA(1) \text{ with } \theta = -0.4, \text{ respectively.}
\end{aligned}$$

The first and the second component have the same AR-coefficient $\phi = 0.2$ or $= 0.8$, the cases of low or high dependencies, respectively. In our three models A, B and C, the y series depends on z_1 and z_2 series in the following way.

	$\boldsymbol{w}_1'\boldsymbol{x}^{st}$	$\boldsymbol{w}_2'\boldsymbol{x}^{st}$	$\boldsymbol{w}_3'\boldsymbol{x}^{st}$	$\boldsymbol{w}_4'\boldsymbol{x}^{st}$	Sum
$t-1$	0.006	0.218	0.007	0.006	0.237
$t-2$	0.009	0.012	0.005	0.004	0.030
$t-3$	0.004	0.005	0.009	0.006	0.024
$t-4$	0.024	0.005	0.002	0.006	0.037
$t-5$	0.460	0.005	0.010	0.003	0.478
$t-6$	0.017	0.010	0.011	0.009	0.047
$t-7$	0.007	0.011	0.001	0.016	0.035
$t-8$	0.006	0.009	0.004	0.003	0.022
$t-9$	0.002	0.009	0.003	0.004	0.018
$t-10$	0.005	0.007	0.011	0.001	0.024
$t-11$	0.004	0.009	0.005	0.010	0.028
$t-12$	0.003	0.007	0.006	0.003	0.019
Sum	0.547	0.307	0.074	0.071	1

Table 1: Estimated dependencies $\hat{\lambda}_{ij}$ between y_t and $(\boldsymbol{w}_i'\boldsymbol{x}^{st})_{t-j}$ with an illustration of different choices of $(\boldsymbol{w}_i'\boldsymbol{x}^{st})_{t-j}$.

A: $y_t = 2z_{1,t-1} + 3z_{2,t-1} + \epsilon_t$ with iid $N(0,1)$-distributed innovations ϵ_t.

B: $y_t = 2z_{1,t-1} + 3z_{2,t-5} + \epsilon_t$ with iid $N(0,1)$-distributed innovations ϵ_t.

C: $y_t = z_{1,t-1}/(0.5 + (z_{2,t-1} + 1.5)^2) + \epsilon_t$ with iid $N(0,1)$-distributed innovations ϵ_t.

As our procedure is affine equivariant, it is not a restriction to consider only the case $\boldsymbol{\Omega} = \boldsymbol{I}_4$ so that $\boldsymbol{x}_t = \boldsymbol{z}_t$ for all $t \in \mathbb{Z}$. We use the standardized $\boldsymbol{z}$ series and the length of the time series was $T = 10000$. The 'true' linear combinations and lags then are

A: $((1,1,0,0)'\boldsymbol{x})_{t-1}$

B: $((1,0,0,0)'\boldsymbol{x})_{t-1}$ and $((0,1,0,0)'\boldsymbol{x})_{t-5}$

C: $((1,0,0,0)'\boldsymbol{x})_{t-1}$ and $((0,1,0,0)'\boldsymbol{x})_{t-1}$

Note that in the last case the dependence is non-linear. For the choice of the number of directions and the number of lags we use the threshold value $\pi = 0.8$.

Results for Model A. Table 2 gives the matrix $\hat{\mathbf{\Lambda}}$. Depending on the method used, we can take values inside the rectangle with solid lines (i.e. k' first directions and s' first lags) or pick the values marked in gray (r largest λ_{ij}'s). Here both lead to the same choices. Dashed lines, which represent which lags to take when all the directions are used and vice versa, lead us to take to either one direction with all lags or all directions with one lag in each case.

1. Case $\phi = 0.2$: $k' = s' = 1$ and $r = 1$, as $\lambda_{11} = \lambda_{i_1,j_1} = 0.926 \geq 0.8$.
2. Case $\phi = 0.8$: $k' = 1$ and $s' = 4$, and $r = 4$, as $\sum_{j=1}^{4} \lambda_{1j} = \sum_{r=1}^{4} \lambda_{i_r,j_r} = 0.813 \geq 0.8$.

The case of weak dependence leads to the first direction with only the first lag to be chosen, as expected. In the strong dependence case also the subsequent lags of the first direction are important, as values depend largely on the previous ones. The chosen lags and directions would then be used for modelling or prediction.

To illustrate this situation with the low dependence case, first the original time series are plotted together with the first selected direction and then y_t is plotted along with the chosen direction. The direction is chosen using the rectangle method with $\pi = 0.8$. In Figure 1 the first 100 values are plotted. It is rather easy to see that the previous values of the chosen direction correspond to the current values of (standardized) y_t, as the changes in those values follow a very similar track. Also from the left panel of the figure it can be seen that the first two original variables (darker gray lines) contribute to the chosen direction. Hence time series SIR is also helpful for data visualization, by reducing the dimension of multivariate time series (here from 4 to one) in an informative way.

Results for Model B. We conclude from Table 3

1. Case $\phi = 0.2$: $k' = 2$ and $s' = 5$, as $\sum_{i=1}^{2} \sum_{j=1}^{5} \lambda_{ij} = 0.954 \geq 0.8$. Also $r = 2$, as $\sum_{r=1}^{2} \lambda_{i_r,j_r} = 0.915 \geq 0.8$, where pairs (i_r, j_r) are $(5, 1)$ and $(1, 2)$.
2. Case $\phi = 0.8$: $k' = 1$ and $s' = 7$, as $\sum_{j=1}^{7} \lambda_{1j} = 0.803 \geq 0.8$. Also $r = 7$, as $\sum_{r=1}^{7} \lambda_{i_r,j_r} = 0.803 \geq 0.8$.

In the weak dependence case the last method chooses two (i_r, j_r) pairs, which are included also in the solid rectangle. In the strong dependence case both selection methods lead again to the same choices; the values in the first column are already high enough to explain at least 80% of the dependence.

The methods should lead to choices where two directions and lags one and five are important. This is the case when there is weak dependence in the source time series. On the other hand, for the strong dependency case, the high value in lag five of the first column combined with strong dependence for adjacent values leads to a first direction overshadowing the second one (Table 3, right panel).

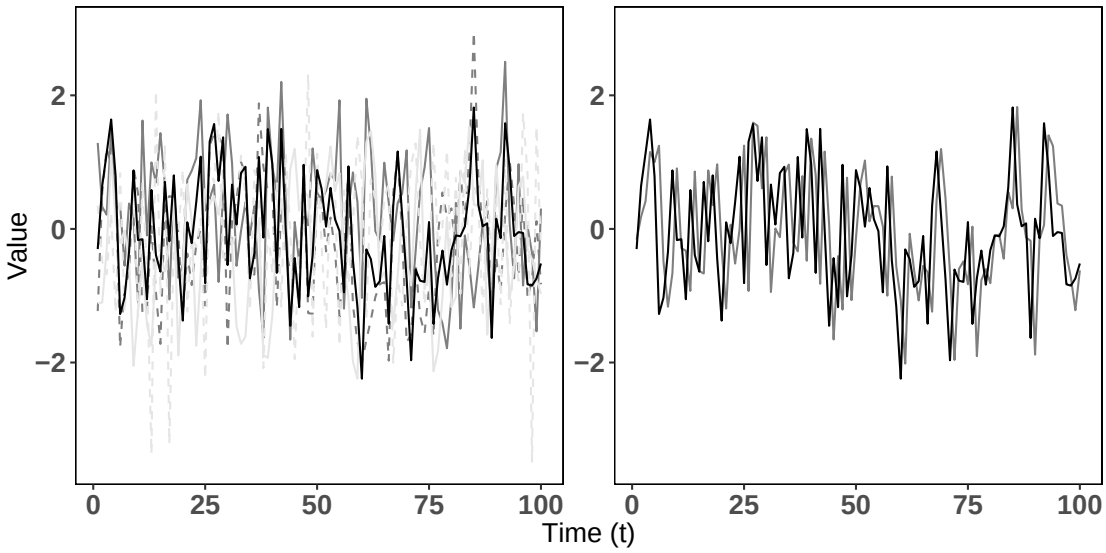

Figure 1: Model A with $\phi = 0.2$. Left panel: The original variables (gray lines) with the chosen direction (black line). Right panel: the first 100 values of standardized y_t (gray line) with the chosen direction (black line).

Results for Model C. Table 4 provides the estimates for λ_{ij}'s. In this model different strategies in case of strong dependence lead to slightly different results.

1. Case $\phi = 0.2$: $k' = 2$ and $s' = 1$, as $\sum_{i=1}^{2} \lambda_{i1} = 0.82 \geq 0.8$. Also $r = 2$, as $\sum_{r=1}^{2} \lambda_{i_r,j_r} = 0.82 \geq 0.8$, where pairs (i_r, j_r) are $(1, 1)$ and $(1, 2)$.
2. Case $\phi = 0.8$: $k' = 2$ and $s' = 5$, as $\sum_{i=1}^{2} \sum_{j=1}^{5} \lambda_{1j} = 0.855 \geq 0.8$. Also $r = 8$, as $\sum_{r=1}^{8} \lambda_{i_r,j_r} = 0.820 \geq 0.8$, where pairs (i_r, j_r) are $(1, 1), \ldots, (6, 1), (1, 2)$ and $(2, 2)$.

In the weak dependence case both these methods lead to same choices, i.e. λ_{11} and λ_{12} are chosen. In the strong dependence case here these methods differ a bit more from each other than in other examples.

In both settings two directions are found. In the case of weak dependence only the first lag is enough. In the strong dependence case more lags are needed, and depending on the choice of method, either an equal or unequal number of lags are chosen for the two directions.

	$\boldsymbol{w}_1'\boldsymbol{x}^{st}$	$\boldsymbol{w}_2'\boldsymbol{x}^{st}$	$\boldsymbol{w}_3'\boldsymbol{x}^{st}$	$\boldsymbol{w}_4'\boldsymbol{x}^{st}$	Sum
$t-1$	0.926	0.000	0.001	0.001	0.929
$t-2$	0.027	0.001	0.001	0.001	0.030
$t-3$	0.002	0.001	0.001	0.001	0.005
$t-4$	0.001	0.001	0.001	0.000	0.003
$t-5$	0.001	0.002	0.001	0.001	0.004
$t-6$	0.001	0.001	0.002	0.001	0.004
$t-7$	0.001	0.002	0.001	0.001	0.005
$t-8$	0.001	0.001	0.001	0.001	0.004
$t-9$	0.001	0.001	0.001	0.001	0.003
$t-10$	0.001	0.000	0.001	0.001	0.004
$t-11$	0.001	0.001	0.001	0.001	0.005
$t-12$	0.001	0.002	0.001	0.000	0.005
Sum	0.965	0.013	0.011	0.011	1

	$\boldsymbol{w}_1'\boldsymbol{x}^{st}$	$\boldsymbol{w}_2'\boldsymbol{x}^{st}$	$\boldsymbol{w}_3'\boldsymbol{x}^{st}$	$\boldsymbol{w}_4'\boldsymbol{x}^{st}$	Sum
$t-1$	0.351	0.001	0.000	0.000	0.352
$t-2$	0.224	0.001	0.000	0.000	0.225
$t-3$	0.144	0.000	0.000	0.000	0.145
$t-4$	0.093	0.000	0.000	0.000	0.094
$t-5$	0.062	0.000	0.000	0.000	0.063
$t-6$	0.040	0.001	0.001	0.000	0.041
$t-7$	0.027	0.001	0.000	0.000	0.028
$t-8$	0.017	0.001	0.000	0.000	0.018
$t-9$	0.010	0.001	0.001	0.000	0.012
$t-10$	0.007	0.001	0.000	0.000	0.008
$t-11$	0.005	0.001	0.001	0.000	0.007
$t-12$	0.004	0.002	0.000	0.000	0.006
Sum	0.983	0.010	0.003	0.003	1

Table 2: Estimated dependencies $\hat{\lambda}_{ij}$ between y_t and $(\boldsymbol{w}_i'\boldsymbol{x}^{st})_{t-j}$ for model A with $\phi = 0.2$ (left panel) and $\phi = 0.8$ (right panel) for the whole example data.

	$\boldsymbol{w}_1'\boldsymbol{x}^{st}$	$\boldsymbol{w}_2'\boldsymbol{x}^{st}$	$\boldsymbol{w}_3'\boldsymbol{x}^{st}$	$\boldsymbol{w}_4'\boldsymbol{x}^{st}$	Sum
$t-1$	0.001	0.276	0.001	0.001	0.279
$t-2$	0.001	0.011	0.000	0.002	0.013
$t-3$	0.002	0.001	0.001	0.000	0.004
$t-4$	0.023	0.001	0.000	0.001	0.025
$t-5$	0.639	0.000	0.000	0.000	0.640
$t-6$	0.021	0.000	0.001	0.000	0.023
$t-7$	0.002	0.000	0.000	0.001	0.003
$t-8$	0.001	0.001	0.001	0.000	0.003
$t-9$	0.001	0.001	0.001	0.001	0.004
$t-10$	0.000	0.000	0.000	0.001	0.002
$t-11$	0.000	0.001	0.001	0.000	0.002
$t-12$	0.001	0.001	0.001	0.001	0.003
Sum	0.691	0.292	0.009	0.008	1

	$\boldsymbol{w}_1'\boldsymbol{x}^{st}$	$\boldsymbol{w}_2'\boldsymbol{x}^{st}$	$\boldsymbol{w}_3'\boldsymbol{x}^{st}$	$\boldsymbol{w}_4'\boldsymbol{x}^{st}$	Sum
$t-1$	0.073	0.041	0.000	0.000	0.114
$t-2$	0.084	0.017	0.000	0.000	0.101
$t-3$	0.104	0.005	0.000	0.000	0.109
$t-4$	0.138	0.000	0.000	0.000	0.139
$t-5$	0.198	0.003	0.001	0.000	0.202
$t-6$	0.126	0.001	0.001	0.000	0.128
$t-7$	0.079	0.001	0.001	0.000	0.081
$t-8$	0.050	0.001	0.000	0.000	0.051
$t-9$	0.031	0.000	0.000	0.000	0.032
$t-10$	0.019	0.001	0.000	0.000	0.021
$t-11$	0.012	0.000	0.000	0.000	0.013
$t-12$	0.008	0.001	0.000	0.000	0.009
Sum	0.923	0.071	0.004	0.002	1

Table 3: Estimated dependencies $\hat{\lambda}_{ij}$ between y_t and $(\boldsymbol{w}_i'\boldsymbol{x}^{st})_{t-j}$ for model B with $\phi = 0.2$ (left panel) and $\phi = 0.8$ (right panel) for the whole example data.

	$\boldsymbol{w}_1'\boldsymbol{x}^{st}$	$\boldsymbol{w}_2'\boldsymbol{x}^{st}$	$\boldsymbol{w}_3'\boldsymbol{x}^{st}$	$\boldsymbol{w}_4'\boldsymbol{x}^{st}$	Sum
$t-1$	0.686	0.139	0.003	0.002	0.830
$t-2$	0.024	0.009	0.004	0.002	0.039
$t-3$	0.006	0.004	0.002	0.003	0.015
$t-4$	0.002	0.002	0.005	0.003	0.013
$t-5$	0.002	0.004	0.001	0.006	0.014
$t-6$	0.003	0.003	0.010	0.002	0.019
$t-7$	0.003	0.003	0.005	0.003	0.014
$t-8$	0.002	0.004	0.002	0.000	0.008
$t-9$	0.002	0.001	0.003	0.002	0.008
$t-10$	0.006	0.006	0.002	0.001	0.015
$t-11$	0.004	0.002	0.005	0.001	0.012
$t-12$	0.003	0.004	0.003	0.003	0.014
Sum	0.744	0.182	0.044	0.029	1

	$\boldsymbol{w}_1'\boldsymbol{x}^{st}$	$\boldsymbol{w}_2'\boldsymbol{x}^{st}$	$\boldsymbol{w}_3'\boldsymbol{x}^{st}$	$\boldsymbol{w}_4'\boldsymbol{x}^{st}$	Sum
$t-1$	0.280	0.065	0.001	0.001	0.347
$t-2$	0.178	0.042	0.001	0.001	0.222
$t-3$	0.114	0.027	0.001	0.001	0.143
$t-4$	0.067	0.020	0.001	0.001	0.089
$t-5$	0.045	0.017	0.001	0.001	0.064
$t-6$	0.028	0.010	0.001	0.001	0.040
$t-7$	0.017	0.007	0.002	0.001	0.027
$t-8$	0.011	0.007	0.001	0.001	0.020
$t-9$	0.008	0.005	0.002	0.001	0.016
$t-10$	0.007	0.003	0.001	0.001	0.012
$t-11$	0.006	0.001	0.001	0.002	0.010
$t-12$	0.006	0.001	0.002	0.001	0.010
Sum	0.768	0.205	0.015	0.012	1

Table 4: Estimated dependencies $\hat{\lambda}_{ij}$ between y_t and $(\boldsymbol{w}_i{}'\boldsymbol{x}^{st})_{t-j}$ for model C with $\phi = 0.2$ (left panel) and $\phi = 0.8$ (right panel) for the whole example data.

5. Sliced inverse regression in prediction

5.1. Prediction

To predict values of the response y_t we need to extract the meaningful linear combinations $\boldsymbol{d} = \boldsymbol{\Gamma}\boldsymbol{x}$ and the number of lags s as explained in Section 3.3. We use the prediction model

$$y_t = f(\boldsymbol{d}_{t-1}, \ldots, \boldsymbol{d}_{t-s}) + \epsilon_t, \quad t \in \mathbb{Z}, s \in \mathbb{Z}_+,$$

where the components of $\boldsymbol{d}$ are the chosen directions and $\boldsymbol{\epsilon}$ is a white noise process. Let $\hat{f}_t$ be the estimate of the prediction function only using past observations $\boldsymbol{x}_{t-1}, \boldsymbol{x}_{t-2}, \ldots$ and $y_{t-1}, y_{t-2}, \ldots$, then the one-step-ahead predictor of y_t is $\hat{f}_t(\boldsymbol{d}_{t-1}, \ldots, \boldsymbol{d}_{t-s})$. If $\hat{\epsilon}_t = y_t - \hat{f}_t(\boldsymbol{d}_{t-1}, \ldots, \boldsymbol{d}_{t-s})$ is calculated for $t = a+1, a+2, \ldots, a+b$ then the prediction power of different choices of $\boldsymbol{\Gamma}$ and s can be compared using the Root Mean Square Error (RMSE)

$$RMSE = \sqrt{\frac{1}{b}\sum_{t=a+1}^{a+b} (\hat{\epsilon}_t)^2}.$$

In our simulations and in the real data example we use a simple linear regression to approximate f. We also set a equal to 75% of the length of the time series, being $a + b$. Hence we make one-step-ahead predictions of the last 25% of the series.

5.2. Simulated time series

In the following simulation study we consider the performance of our method in the context of one-step-ahead prediction as described above. We continue to consider the models A, B and C of Section 4 with cases $\phi = 0.2$ and $\phi = 0.8$ and fix the time series length to $T = 5000$. The number of directions and relevant lags is chosen in our method using the first three approaches described in Section 3.3 with $\pi = 0.8$, a value common for example in PCA as threshold for the proportion of variance to be explained. Although sometimes also lower values might be appropriate, especially when the dimension is large. To see the impact of the value of π we also consider $\pi = 0.5$ for the 'rectangle' method where the number of directions k' and lags s' is chosen such that $k's'$ is minimal for $\sum_{i=1}^{k'}\sum_{j=1}^{s'} \lambda_{ij} \geq \pi$. In all simulations we consider the lags s of interest to be $1, \ldots, 12$, as these values are common in blind source separation methods. As a baseline we choose an Oracle estimator which knows the value of k' and the number of lags relevant but still needs to estimate the parameters of the regression model for the prediction.

As mentioned earlier, there are not really comparable other methods around. The closest seems the approach suggested in Becker and Fried (2003) where the iid SIR is applied to a modified matrix of explaining

variables $\mathbf{x}^*$. In our context this would mean that $\mathbf{x}_t^* = (\mathbf{x}_t', \mathbf{x}_{t-1}', \ldots, \mathbf{x}_{t-s}')'$, i.e. if it is assumed that at most s lags are relevant, then all time series are added as new variables shifted by each lag $1, \ldots, s$. Hence even for moderate p and s the dimension of $\mathbf{x}^*$ will be huge while at the same time the number of available time points is reduced by s. In the context of Becker and Fried (2003) they argue that $s = 2$ is sufficient. Also no direct rules about how to choose the number of directions here are available. In our simulation study we choose for their method k' as the minimal value for which $\sum_{i=1}^{k'} \lambda_i / \sum_{i=1}^{s(p+1)} \lambda_i \geq \pi = 0.8$, where λ_i, $i = 1, \ldots, s(p+1)$ are the ordered eigenvalues of the empirical supervised covariance matrix $\mathrm{COV}(E(\mathbf{x}^{*,st}|y^{sl})$.

Figure 2 presents the relative RMSE values of 500 repetitions compared to the Oracle estimator, using 80% as the threshold value unless otherwise stated. In models A and C the rectangle method with choice $\pi = 0.5$ gives at least as good as or sometimes even better results than choice $\pi = 0.8$, but all methods work well. In model B some of our methods work very well (small variation like in other models) and better than Becker and Fried (2003). However, some methods do not work at all because the second direction is not found. This is especially true with the choice $\pi = 0.5$. Keeping all directions gives the best results here.

In general, except in model B, the loss in estimating the number of directions and lags is minimal compared to the oracle estimator. In addition $\pi = 0.8$ seems to be a more reliable choice. It can be concluded that the method keeping all the directions and then choosing the number of lags seems to be the best from a prediction point of view, as other methods do not always find the proper amount of directions. It is the best or one of the best in all three settings under both low and high serial dependence and in all of the simulations it works better than Becker and Fried (2003). However, it should be noted that keeping all directions in high dimensional time series may be infeasible.

Also if the underlying relationship is linear or non-linear seems not to matter much in this context. We had used also a more flexible spline regression model to approximate the f, but do not present the results here as the gain was minimal. This might be however more relevant when predicting more than one step ahead.

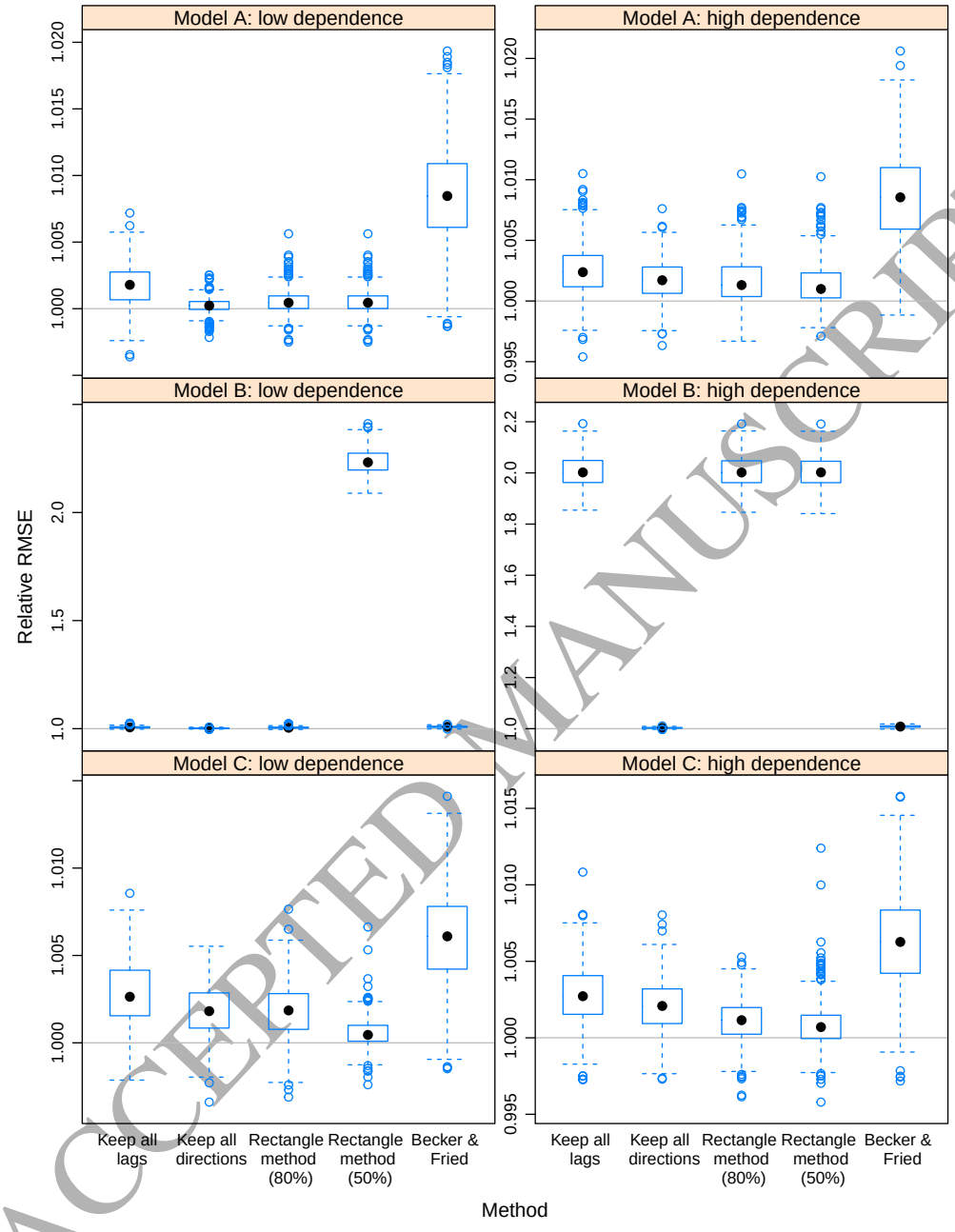

Figure 2: Relative RMSE values of 500 repetitions of all models with low ($\phi = 0.2$) and high ($\phi = 0.8$) dependence compared to the Oracle estimator.

5.3. Real data example

To demonstrate our method also for real data, we consider the well-known Stock and Watson (2002) monthly economic time series data set. The version of the data we use contains 125 macroeconomic variables for the whole period from January 1959 to December 2003 ($T = 540$). Following Carriero et al. (2011), which give also more detailed information about the data set, we follow their recommendation and use a subset of, partly transformed, $p = 52$ time series as chosen there. As in Stock and Watson (2002), monthly growth rate of the industrial production index (total index, code IPS10) is used as the response series. 527 observations are then left after data preparation.

We use our approach and Becker and Fried (2003) as described in the simulation section for one-step-ahead prediction of the last 25% of the data. However, now only lags $s = 1, \ldots, 6$ are utilized, as Becker and Fried (2003) cannot cope with a larger number of lags. Our methods did however not really improve when using the larger lag set $s = 1, \ldots, 12$. The threshold values $\pi = 0.5$ and 0.8 have been chosen as before.

For $\pi = 0.8$ when keeping all lags 23 directions are chosen and when keeping all directions 6 lags are chosen. For the rectangle method $k' = 23$ and $s' = 6$. For $\pi = 0.5$ when keeping all lags 10 directions are chosen and when keeping all directions 3 lags are chosen. For the rectangle method $k' = 10$ and $s' = 6$. As seen from the results, real data are much more complex than simulated data, and therefore here so many directions are needed and also $\pi = 0.5$ seems to be enough.

Table 5 presents the results using Becker and Fried (2003) with $\pi = 0.5$ as the baseline. While in the simulations the differences between Becker and Fried (2003) and our methods seemed not so big, it is quite different here. All our methods are clearly better than Becker and Fried (2003) and rectangle method along with keeping all lags and then choosing the number of directions seem to be the best approaches. Results also show that choosing all directions and then the number of lags is not so good in practice, likely due to too many directions causing excess noise. It can also be noted that the results with $\pi = 0.5$ give more satisfactory results.

Method	RMSE ($\pi = 0.5$)	RMSE ($\pi = 0.8$)
Keep all s lags	0.469	0.579
Keep all k directions	0.596	0.807
Rectangle method	0.469	0.579
Becker & Fried	1.000*	0.957

Table 5: Relative RMSE values for all methods compared to Becker & Fried with $\pi = 0.5$.

6. Discussion

Especially in macroeconomics forecasting one time series using multiple other time series is a common problem. Many approaches were suggested for this problem. In dynamic factor models the latent factors are usually estimated in an unsupervised fashion and then a linear regression model is fitted. Another approach is variable selection methods like for example Gelper and Croux (2008) where a special version of least angle regression (LARS) for time series is suggested in the linear regression context.

Barbarino and Bura (2015) combine supervised and unsupervised ideas by applying SIR to selected principal components of the explaining time series jointly with the lagged values of the response. Becker and Fried (2003) also naturally can use the lagged response values in the augmented explaining series $\mathbf{x}_t^*$. Similarly, in our approach one could first regress the response on its past values and apply our methodology to the resulting residuals. This will be still explored in future work as it makes then an assumption on linearity. The method we suggested in this paper on the other hand makes no assumption on the form of the relationship between and y_t and $\mathbf{x}_t$ and under weak assumptions suggests a number of linear combinations which are relevant and can even indicate important lags. As all these different approaches here make quite different assumptions, a general comparison is difficult.

In the case of independent and identically distributed observations asymptotic results, including consistency, have been developed in the literature. Asymptotics for subspace estimation is considered in Zhu and Ng (1995) and Li and Zhu (2007) for example. If the dimension of the subspace is unknown as well, asymptotics tools for testing and estimation of the subspace dimension are available. Li (1991) proposes a chi-squared test for finding the dimension when $\boldsymbol{x}_t$ is normal. Bura and Cook (2001) shows that the normality assumption is not necessary and just the conditional covariance structure of the predictors needs to have some restrictions. See also Nordhausen et al. (2016). Zhu et al. (2010) suggests a BIC-type criterion to find the dimension. If the observations are multivariate time series, the problem is much more challenging with unknown number of lags and unknown dimensions for each lag. In this paper we just provide heuristic tools to choose the lags, dimensions and subspaces in the time series context.

We consider this paper as an opening for a formal supervised dimension reduction framework for time series and plan to consider in future extensions of other iid supervised dimension reduction methods like SAVE (Cook, 2000) and so on to the time series context. Also different ways on how to incorporate past values of the response series will be considered.

Acknowledgments

We thank the Associate Editor and the referees for careful reading of the paper and helpful comments.

References

Barbarino, A., Bura, E., 2015. Forecasting with sufficient dimension reductions. Finance and Economics Discussion Series 2015-074 doi:http://dx.doi.org/10.17016/FEDS.2015.074.

Becker, C., Fried, R., 2003. Sliced inverse regression for high-dimensional time series, in: Schwaiger, M., Opitz, O. (Eds.), Exploratory Data Analysis in Empirical Research. Springer Berlin Heidelberg. Studies in Classification, Data Analysis, and Knowledge Organization, pp. 3–11.

Belouchrani, A., Abed Meraim, K., Cardoso, J.F., Moulines, E., 1997. A blind source separation technique based on second order statistics. IEEE Transactions on Signal Processing 45, 434–444.

Bura, E., Cook, R., 2001. Extending sliced inverse regression: the weighted chi-squared test. Journal of the American Statistical Association 96, 996–1003.

Cardoso, J.F., Souloumiac, A., 1996. Jacobi angles for simultaneous diagonalization. SIAM J. Mat. Anal. Appl 17, 161–164.

Carriero, A., Kapetanios, G., Marcellino, M., 2011. Forecasting large datasets with bayesian reduced rank multivariate models. Journal of Applied Econometrics 26, 735–761.

Cook, R., 2000. Save: A method for dimension reduction and graphics in regression. Communications in Statistics - Theory and Methods 29, 2109–2121.

Cook, R., Weisberg, S., 1991. Sliced inverse regression for dimension reduction: Comment. Journal of the American Statistical Association 86, 328–332.

Fan, J., Xue, L., Yao, J., 2015. Sufficient forecasting using factor models. URL: http://arxiv.org/abs/1505.07414v2.

Forni, M., Hallin, M., Lippi, M., Reichlin, L., 2005. The generalized dynamic factor model: one-sided estimation and forecasting. Journal of the American Statistical Association 100, 830–840.

Gelper, S., Croux, C., 2008. Least angle regression for time series forecasting with many predictors. Technical Report. Katholieke Universiteit Leuven.

Illner, K., Miettinen, J., Fuchs, C., Taskinen, S., Nordhausen, K., Oja, H., Theis, F.J., 2015. Model selection using limiting distributions of second-order blind source separation algorithms. Signal Processing 113, 95–103.

Kim, H.H., Swanson, N.R., 2014. Forecasting financial and macroeconomic variables using data reduction methods: New empirical evidence. Journal of Econometrics 178, 352–367.

Li, K.C., 1991. Sliced inverse regression for dimension reduction. Journal of the American Statistical Association 86, 316–327.

Li, Y., Zhu, L.X., 2007. Asymptotics for sliced average variance estimation. Annals of Statistics 35, 41–69.

Liski, E., Nordhausen, K., Oja, H., 2014. Supervised invariant coordinate selection. Statistics: A Journal of Theoretical and Applied Statistics 4, 711–731.

Ma, Y., Zhu, L., 2013. A review on dimension reduction. International Statistics Review 81, 134–150.

Matilainen, M., Miettinen, J., Nordhausen, K., Oja, H., Taskinen, S., 2016. ICA and stochastic volatility models, in: Aivazian, S., Filzmoser, P., Kharin, Y. (Eds.), Proceedings of the XI International Conference on Computer Data Analysis and Modeling, Publishing center of BSU, Minsk. pp. 30–37.

Matilainen, M., Nordhausen, K., Oja, H., 2015. New independent component analysis tools for time series. Statistics & Probability Letters 105, 80–87.

Miettinen, J., Illner, K., Nordhausen, K., Oja, H., Taskinen, S., Theis, F., 2016. Separation of uncorrelated stationary time series using autocovariance matrices. Journal of Time Series Analysis 37, 337–354.

Miettinen, J., Nordhausen, K., Oja, H., Taskinen, S., 2012. Statistical properties of a blind source separation estimator for stationary time series. Statistics & Probability Letters 82, 1865–1873.

Miettinen, J., Nordhausen, K., Oja, H., Taskinen, S., 2014. Deflation-based separation of uncorrelated stationary time series. Journal of Multivariate Analysis 123, 214–227.

Miettinen, J., Taskinen, S., Nordhausen, K., Oja, H., 2015. Fourth moments and independent component analysis. Statistical Science 30, 372–390.

Nordhausen, K., 2014. On robustifying some second order blind source separation methods for nonstationary time series. Statistical Papers 55, 141–156.

Nordhausen, K., Oja, H., Tyler, D., 2016. Asymptotic and bootstrap tests for subspace dimension. arXiv:1611.04908 .

Shi, Z., Jiang, Z., Zhou, F., 2009. Blind source separation with nonlinear autocorrelation and non-Gaussianity. J. Comput. Appl. Math. 223, 908–915.

Stock, J.H., Watson, M.W., 2002. Macroeconomic forecasting using diffusion indexes. Journal of Business & Economic Statistics 20, 147–162.

Taskinen, S., Miettinen, J., Nordhausen, K., 2016. A more efficient second order blind identification method for separation of uncorrelated stationary time series. Statistics & Probability Letters 116, 21–26.

Zhu, L.P., Wang, T., Zhu, L., Ferré, L., 2010. Sufficient dimension reduction through discretization-expectation estimation. Biometrika 97, 295–304.

Zhu, L.X., Ng, K.W., 1995. Asymptotics of sliced inverse regression. Statistica Sinica 5, 727–736.