Nordhausen, Klaus; Virta, Joni

**Ladle Estimator for Time Series Signal Dimension**

*Published in:*
2018 IEEE Statistical Signal Processing Workshop, SSP 2018

*DOI:*
10.1109/SSP.2018.8450695

Published: 29/08/2018

*Document Version*
Peer reviewed version

*Please cite the original version:*
LADLE ESTIMATOR FOR TIME SERIES SIGNAL DIMENSION

Klaus Nordhausen*
Vienna University of Technology
Institute of Statistics &
Mathematical Methods in Economics
Vienna, Austria

Joni Virta
Aalto University
School of Science
Helsinki, Finland

ABSTRACT

We consider a second order source separation model where a set of latent signals is internally mixed with several channels of noise and the goal is to estimate the number of signals. For the purpose we extend the ladle estimator which has been so far considered only for iid methods such as PCA, CCA or FOBI. Using time series bootstrapping methods ladle estimators based on AMUSE and SOBI are presented and a simulation study demonstrates that especially SOBI works well under all considered scenarios if the time series are sufficiently long.

Index Terms— AMUSE, SOBI, bootstrap, order determination

1. BLIND SOURCE SEPARATION MODEL

Blind source separation (BSS) is a well-established multivariate procedure where the goal is to recover latent components while only a linear mixture of them is observed. That is, we assume that the observed multivariate signal \( x_t \in \mathbb{R}^p \) is generated as \( x_t = \mu + \Omega z_t \), where \( \mu \in \mathbb{R}^p, \Omega \in \mathbb{R}^{p \times p} \) are unknown and the latent \( z_t \in \mathbb{R}^p \) is in some sense simpler or more structured than \( x_t \). Clearly the BSS problem is not solvable without further assumptions. The two perhaps most popular BSS approaches are independent component analysis (ICA), where the latent vectors \( z_t \) are assumed to be iid and mutually independent, and second order source separation (SOS), where the latent sources are assumed to be uncorrelated or independent weakly stationary processes. Many different estimators exist for ICA and SOS, such as FOBI, JADE and FastICA or AMUSE and SOBI, respectively, for general references see for example [1, 2].

A general problem in the BSS model is that the order of the components is not fixed and different methods estimate the components in a different order. Hence, contrary to e.g. the principal component analysis (PCA), the ordering of the BSS components usually does not tell anything about the components’ relevance. Moreover, it is often assumed that only a few components are of interest and represent the “signal” while the remaining components can be considered “noise”. Of interest is then to estimate the signal dimension and recover the corresponding components.

In recent works on ICA it has been assumed that the signal is formed solely of non-Gaussian components and that the noise is formed entirely of Gaussian components, see for example [3, 4, 5, 6, 7] for some references. In a similar vein, in this paper our interest lies in estimating the signal/noise subspace dimensions in an SOS context. [8] suggested using the SOS signal-noise model,

\[
\begin{align*}
  \mathbf{S}_t &= \mu + \Omega \mathbf{z}_t = \mu + \Omega_1 \mathbf{z}_{1,t} + \Omega_2 \mathbf{z}_{2,t},
\end{align*}
\]

where \( \Omega = (\Omega_1, \Omega_2) \in \mathbb{R}^{p \times p} \) is non-singular, \( \Omega_1 \in \mathbb{R}^{p \times q}, \Omega_2 \in \mathbb{R}^{p \times (p-q)}, \mu \in \mathbb{R}^p \), and the \( p \)-variate random process \( \mathbf{z}_t = (\mathbf{z}_{1,t}^\top, \mathbf{z}_{2,t}^\top)^\top \) has \( E(\mathbf{z}_t) = 0 \), \( \text{Cov}(\mathbf{z}_t) = I_p \) and satisfies

(A1) \( \mathbf{z}_{1,t} \in \mathbb{R}^q \) and \( \mathbf{z}_{2,t} \in \mathbb{R}^{p-q} \) are mutually independent,

(A2) the components of \( \mathbf{z}_{1,t} \) are independent and exhibit second order serial dependence while the components of \( \mathbf{z}_{2,t} \) are white noise.

Following [9] one could argue that (A2) could rather be

(A3) the components of \( \mathbf{z}_{1,t} \) are independent and exhibit second order serial dependence while the components of \( \mathbf{z}_{2,t} \) are independent but do not exhibit second order serial dependence.

This would allow, for example, the “noise” part to consist of components exhibiting stochastic volatility which is of interest in financial time series analysis. Other types of noise models could also be utilized, see the discussion in Section 7.

Popular approaches to solve the SOS-model (1) are AMUSE [10, 11] and SOBI [12, 13, 14], which both first standardize the series using the regular covariance matrix \( \Sigma \), i.e. \( x_t := \Sigma^{-1/2}(x_t - \mu) \). Denote next by \( \Sigma_\tau = E(x_t^\top x_{t+\tau}^\top) \) the autocovariance matrix of the standardized series at lag \( \tau > 0 \). The idea in AMUSE is to find an orthogonal matrix \( U_\tau \) which diagonalizes \( \Sigma_\tau \) for a selected \( \tau \) while SOBI
uses a set of $K$ distinct lags $\mathcal{T} = \{\tau_1, \ldots, \tau_K\}$ and finds an orthogonal matrix $U_\mathcal{T}$ as

$$U_\mathcal{T}^\top = \arg\max_{V^\top V = I, \tau \in \mathcal{T}} \sum_{\tau \in \mathcal{T}} \left\| \text{diag} \left( V^\top \Sigma, V \right) \right\|_F^2,$$

(2)

where $\| \cdot \|_F$ denotes the Frobenius norm. The AMUSE and SOBI unmixing matrices are then $\Gamma_\tau = \Sigma^{-1/2} U_\tau$ and $\Gamma_\tau = \Sigma^{-1/2} U_\mathcal{T}$ respectively. The question is then how to decide which of the components are signal and which are noise? In this work we will approach the question via the ladle estimator [7] which combines information from eigenvectors and eigenvalues to estimate the dimensions. In our joint diagonalization framework, the counterparts for the concepts are clearly the matrix $U_\mathcal{T}$ and the diagonal elements of the matrices $\text{diag}(U_\tau \Sigma, U_\tau)^2$, $\tau \in \mathcal{T}$. To combine the information in different lags for SOBI, we sum over the lag set $\mathcal{T}$ to obtain the diagonal matrix $\Lambda = \sum_{\tau \in \mathcal{T}} \text{diag}(U_\tau \Sigma, U_\tau)^2$, the diagonal elements of which we further order in decreasing order (for AMUSE we just use $\Lambda = \text{diag}(U_\mathcal{T} \Sigma, U_\mathcal{T})^2$). To now identify the signal space, note that under model (1) $\Lambda$ has always $d \leq q$ non-zero elements and exactly $d = q$ non-zero elements if all signal components exhibit serial dependence for some lag in $\mathcal{T}$. This insight, along with the bootstrap variability [15] of the vectors in $U_\mathcal{T}$, will be used in Section 5 to define an SOS version of the ladle estimator. But before that, we first define its individual components, starting with describing suitable bootstrap strategies for dependent data in Section 2.

### 2. TIME SERIES BOOTSTRAP

The standard bootstrap [16] is not usable with time series data as drawing observations randomly from the observed $T$ time points ignores the time dependency of the series and fails to reproduce anything resembling the original data. A solution is provided by fixed block bootstrap [17] where we choose a block length $B$ and instead of resampling individual time points, resample blocks of size $B$. Assuming for simplicity that the time series length is a multiple of the block length, $T = hB$, in the fixed block bootstrap we thus randomly sample $h$ indices with replacement from $\{1, \ldots, T - B + 1\}$ and the length-$B$ blocks starting with these indices are concatenated to obtain a bootstrap replication of the time series.

However, the fixed block bootstrap, while already a powerful resampling method, does not in general preserve the possible stationarity of the original time series. This drawback is averted in stationary bootstrap [18] which differs from the moving block bootstrap in that it lets the block sizes vary individually. That is, letting $p \in (0, 1)$ be a fixed parameter, for each sampled index we generate the corresponding block length from the geometric distribution with the success probability $p$. This procedure is repeated for enough indices to obtain a time series replicate with length at least $T$, after which the leftover time points can be cut off, leaving us with a stationary replicate of the original time series, see [18]. As our SOS model is assumed to be stationary, the stationary bootstrap can thus be seen as better suited to our current problem than the fixed block bootstrap, see Section 6 for their comparison. In the next section the bootstrap methods will be used to define the first part of the ladle estimator. See also [19] for a comprehensive treatise of bootstrapping for dependent data.

### 3. JOINT DIAGONALIZER BOOTSTRAP VARIABILITY

The use of bootstrap variability of eigenvectors to estimate the rank of a symmetric matrix was proposed in [15] and is based on the essential uniqueness of eigenvectors corresponding to eigenvalues of multiplicity one. As a motivating example, consider the estimation of the rank of the covariance matrix $\Sigma = \text{diag}(3, 1, 0, 0)$ from an infinite supply of random samples of size $T$ coming from a distribution with the particular covariance structure. Inspecting the eigendecomposition of a single estimate, $\Sigma = \hat{V} \Lambda \hat{V}^\top$, the first eigenvector having multiplicity one guarantees that the first eigenvector is for large $T$ always (up to its sign) close to the first canonical basis vector $e_1 \in \mathbb{R}^4$. Similarly, the second eigenvector is for all of our infinite replicates expected to be close to $e_2$. But the analogous no longer holds true for the third eigenvector: as the corresponding eigenvalue has multiplicity greater than one we only know that the eigenvectors eventually lie in $\text{span}(e_3, e_4)$, and are thus expected to exhibit greater variation than the previous two. This explosion in the variation of the eigenspace can (in our simple example) be used to determine the exact point where the eigenvalues turn to zero. As no such thing as an infinite supply of samples is available for us in practice, the replicates need to be drawn as bootstrap samples from the original data.

The bootstrap variability of eigenvectors has been previously used only in the context of an eigendecomposition [15, 7] and the next description, tailored to our current situation, generalizes it to apply to joint diagonalization. Let $x_1, \ldots, x_T \in \mathbb{R}^p$ be a sample from the SOS model (1) and fix an SOS-estimator and a time series bootstrapping method along with their parameters. Denote then by $B_k, B_{k,j}^*$, $j = 1, \ldots, n$, the first $k$ columns of the joint diagonalizer (2) computed respectively from the original sample and the $j$th bootstrap sample. The average variation of the space spanned by the first $k$ vectors of the joint diagonalizer can now be quantified with the function,

$$f_0(k) = \frac{1}{n} \sum_{j=1}^n \left\{ 1 - \left| \det \left( B_k^\top B_{k,j}^* \right) \right| \right\},$$

for $k = 1, \ldots, p - 1$ and $f_0(0) = 0$. The absolute determinant takes values between 0 and 1 and measures the similarity of the column spaces of $B_k$ and $B_{k,j}^*$, making $f_0$ a sort of “variance” of the column spaces of $B_{k,j}^*$ around the column space.
of $\hat{B}_k$. As in [7] we further normalize the variance function as $f(k) = f_0(k)/(1 + \sum_{l=0}^{p-1} f_0(l))$ and similar analysis as done therein can be used to show that if all signal series in $z_{1,t}$ exhibit autocorrelation for at least one lag included in the joint diagonalization, then the function $f$ has a positive jump between the values $f(q)$ and $f(q + 1)$ where $q$ is the true dimension of $z_{1,t}$. This observation already gives us a good starting point for determining $q$ but, depending on the dependency structures of the signal series, the function $f$ could have jumps also between other values, and as in [7] we supplement the information contained in the joint diagonalization “eigenvectors” by that contained in the joint diagonalization “eigenvalues”, described in detail in the next section.

4. JOINT DIAGONALIZER SCREE PLOT

The way we use the eigenvalues is in the form of the classical scree plot used commonly in principal component analysis [20]. Let $\lambda_0^2, \lambda_1^2, \ldots, \lambda_{p-1}^2$ be the (re-indexed) diagonal elements of $\Lambda$ ordered in decreasing order, see Section 1. Assuming again that all signal series exhibit suitable autocorrelations, the squared sums experience a negative jump between $\lambda_{q-1}^2$ and $\lambda_q^2$, where the transition from signal to noise lies. Mirroring our development of $f$ in the previous section, we define the normalized function $\phi$ from $\{0, \ldots, p - 1\}$ to $\mathbb{R}$ as $\phi(k) = \lambda_k^2/(1 + \sum_{l=0}^{p-1} \lambda_l^2)$, for which the jump occurs at the same spot. The normalization is required to make also $\phi$ scale-free, so that in the next section we can combine the two functions, $f$ and $\phi$, to obtain the ladle, a powerful plot-based estimator for the signal dimension $q$.

5. LADLE ESTIMATOR FOR TIME SERIES

The ladle [7] is now obtained simply by summing the normalized eigenvector and eigenvalue functions $f, \phi$ to obtain the ladle function $g(k) = f(k) + \phi(k)$. Heuristically, as $\phi$ goes down between $q - 1$ and $q$ and $f$ goes up between $q$ and $q + 1$, we expect the function $g$ to have a (local) minimum at the true dimension $q$. The ladle estimate $\hat{q}$ is then defined as,

$$\hat{q} = \arg\min_{q=0,\ldots,p-1} \{g(q)\},$$

the value which minimizes $g$, and under suitable assumptions [7] show that the minimum is actually a consistent estimator of $q$ (in the case of the standard eigendecomposition-based ladle). Investigating whether these assumptions and the claim itself hold also for our case of dependent time series data, time series bootstrapping and the joint diagonalization version of the ladle constitutes a future work. In this preliminary study we will however focus on validating the proposed procedure using the simulation studies conducted in the next section. Sample plots of all three functions, $f, \phi$ and $g$, will also be presented therein.

6. SIMULATION STUDIES

We conducted an extensive simulation study in R [21], using the packages boot [22, 23] and JADE [24], to evaluate how well the ladle estimates the signal dimension. Throughout the simulations we considered model (1) with $p = 10$ and $q = 5$ where the five signal components are ARMA(0.6; (0, -0.4)), ARMA((0.4, 0.1, 0.3); (0.2, 0.4)), AR(0.4), MA(0.5, 0.3, 0.1, 0.2) and AR(0.1, 0.1, 0.3), all having Gaussian innovations. In Setting 1 the five noise components are all Gaussian white noise whereas in Setting 2 the five noise components are all GARCH(1,1) processes. Due to affine equivariance we used in both settings the identity mixing $\Omega = I_{10}$. As process lengths we considered $T = 200, 500, 1000, 2000$ where the ladle was computed for AMUSE with lag $\tau = 1$ and for SOBI with $T = \{1, \ldots, 12\}$. We used 200 bootstrap samples and considered fixed block bootstrapping (F) with block lengths 10, 20 and 40 and stationary block bootstrapping (S) with the expected block lengths 10, 20 and 40. All the simulation results are based on 2000 repetitions.

Before discussing the results of the simulations we show in Figure 1 the ladle plot based on SOBI for one case of Setting 1 with $T = 1000$ and using the stationary block bootstrap with the expected block length 20. The plot contains the main criterion $g$ which has a clear minimum at the true signal subspace dimension $q = 5$ and includes also the two components $f$ and $\phi$ which sum up to $g$. For all simulations we then recorded the minimizer of the function $g$ and Figures 2 and 3 give the counts of how often each method estimated each dimensionality. In all three settings it is evident that 200 or 500 observations are not sufficient. However starting with $T=1000$ SOBI works well and is clearly better than AMUSE. While the form of the time series bootstrapping seems of minor relevance here, the (expected) block length has an impact and 40 seems to be preferable. We tried also the ladle estima-
tor based on FOBI [7] in all settings but it failed completely and is therefore omitted from the figures.

7. DISCUSSION

We introduced in this paper the ladle plot to determine the signal space dimension in SOS models using AMUSE or SOBI. The plot is easy to interpret and as the simulations show it reliably gives the correct dimension estimate especially for SOBI when the time series are sufficiently long. How the time series bootstrapping was done seems of minor importance given that the block length is large enough. Future work will investigate the properties of the estimates further and might consider also parametric bootstrap methods.

Apart from model (1), another viable model to apply the ladle to would be

\[
x_t = \mu + \Omega z_t + \epsilon_t,
\]

where \( \mu \) is again a \( p \)-vector, \( \Omega \in \mathbb{R}^{p \times q} \), \( z_t \) is a \( q \)-variate process and \( \epsilon_t \) is \( p \)-variate white noise vector that is independent of \( z_t \). The natural assumption to make on the latent \( z_t \) having \( E(z_t) = 0 \) and \( \text{Cov}(z_t) = I_q \) would then be

\[ (B1) \] the components of \( z_t \) are independent and exhibit second order serial dependence.

The main difference between models (1) and (3) is that in (1) the noise is internal and the signals can be estimated undistorted while in (3) the noise is external and the signals can only be estimated with distortion. Although AMUSE and SOBI are unable to extract the signal components from model (3), it can be shown that the ladle can still be used to correctly estimate the dimension \( q \) of \( z_t \). Some preliminary simulations (not included here) also verify that this is indeed the case. Furthermore, under the alternative model the dimension information given by the ladle could then be combined with, for example, the joint diagonalization of two autocovariance matrices to estimate both the dimension and the signals in (3). This will also be explored further in future work.

![Fig. 2. The distributions of the ladle estimates in Setting 1.](image)

![Fig. 3. The distributions of the ladle estimates in Setting 2.](image)
8. REFERENCES


