Elasso for estimating the signal dimension in ICA

Mengxi Yi
School of Statistics
Beijing Normal University
China
mxyi@bnu.edu.cn
0000-0001-9592-959X

Klaus Nordhausen
Department of Mathematics and Statistics
University of Jyväskylä
Finland
klaus.k.nordhausen@jyu.fi
0000-0002-3758-8501

***Abstract*—Independent component analysis is often considered in a framework where the p observed variables are a mixtures of only $d < p$ latent independent components which are contaminated by white noise. The goal is then to estimate the number of latent components as well as the components itself. Meanwhile several approaches exist to estimate d which are all are based on the eigenvalues of the covariance matrix. However all these approaches were developed and tested in scenarios where p is moderately small. If p is however large the estimation of the eigenvalues suffers. To improve the estimation of d by better estimation of the eigenvalues we employ the recently suggested Elasso which penalizes the eigenvalue structure and groups them together when possible. We show how the Elasso can be used for estimation of d and show in simulations and in an example that it is better than the competing methods when p is large.**

***Index Terms*—noisy ICA, signal dimension, Elasso, order determination**

I. Introduction

One of the fundamental tasks in multivariate data analysis is to extract signals from an observed set of data, which are often assumed to be a linear mixture of the lower-dimensional signals contaminated by noise. A popular way to blindly estimate the signals is the independent component analysis (ICA). Here, the blind means both the signals and the mixing process are unknown. Numerous methods have been proposed to solve the ICA problem, which are wide applicable in many areas, for example, speech and noise filtering, financial time series, telecommunications, medical images, and so on (see [1], [2], [3], and the reference therein, for an overview). Basically, the signals could be recovered by estimating a linear transformation that guarantees independence between signals, which are assumed to be non-Gaussian, as long as the number of the signals are known. However, the signal dimension is usually unknown in practice and needs to be estimated.

The goal of the current work is to estimate the signal dimension in ICA. Throughout the paper, we consider the following noisy ICA model

$$\boldsymbol{x} = \boldsymbol{A}\boldsymbol{s} + \sigma\boldsymbol{\epsilon}, \tag{1}$$

where $\boldsymbol{A}$ is a $p \times d$ *mixing matrix*. $\boldsymbol{s} = (s_1, \cdots, s_d)^\top$ contains the independent signals, with zero means and identity covariance matrix and d is referred to as the signal dimension or signal number. $\boldsymbol{\epsilon}$ is a $p \times 1$ Gaussian white noise vector, in dependent of the signals, with mean vector $\mathbf{0}$ and covariance matrix $\boldsymbol{I}_p$. σ^2 denotes the unknown noise variance parameter. In the above model, only $\boldsymbol{x}$ is observed and $\boldsymbol{A}$ is assumed to be a full rank matrix, with rank $d < p$. The covariance matrix of $\boldsymbol{x}$ is then given by:

$$\boldsymbol{\Sigma} = \boldsymbol{A}\boldsymbol{A}^\top + \sigma^2\boldsymbol{I}_p, \tag{2}$$

which implies the eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$ of $\boldsymbol{\Sigma}$ are respectively equal to

$$(\delta_1 + \sigma^2, \cdots, \delta_d + \sigma^2, \sigma^2, \cdots, \sigma^2), \tag{3}$$

where $\delta_1 \geq \cdots \geq \delta_d > 0$ are the d non-zero eigenvalues of $\boldsymbol{A}\boldsymbol{A}^\top$. The structure of (2) or (3), can also be expressed as the hypothesis:

$$H_{0d} : \lambda_d > \lambda_{d+1} = \cdots = \lambda_p. \tag{4}$$

[4] consider an asymptotic test to consistently estimate the signal dimension d of the ICA model based on the principal component analysis (PCA).

In the literature, the covariance matrix of the form (2)-(4), also known as sub-sphericity models or factor models [5], [6], is well-studied, and the problem of estimating d has been investigated thoroughly, under different settings. For example, in signal processing of IID Gaussian signals, the information theoretic criteria were considered in, e.g., [7] and [8]; in time series factor models, [9] estimate the signal dimension based on the ratios of eigenvalues of autocovariance matrix; and so on. The information theoretic criteria and the asymptotic test are designed to perform well when the sample size is much larger than the dimension, while the ratio-based method is proved to work well for time series data and when the dimension also grows with the sample size. However, it is unknown yet how well those methods work in noisy ICA setting and when the dimension is comparable with the sample size.

Recently, to improve the poor performance of the sample covariance when sample size is small relative to the dimension of the data, [10] propose a class of nonsmooth penalty functions, called *Elasso*, that can group the empirical eigenvalues together. This method can be used to determine the number

MY's work was supported partly by the National Natural Science Foundation of China (12101119) and by the Fundamental Research Funds for the Central Universities (310422113) and KN was partly supported by the Austrian Science Fund (P31881-N32).

of signals when the sample size is not large enough. The goal of this work is to present the benefit of its use for the signal estimation, in a noisy ICA model of large variable dimension, which has not yet been considered, and compare this procedure with the above mentioned asymptotic test, ratio-based and information criteria method in an extensive simulation study and an real data example.

II. Methodologies

Given a set of n independent observations $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n$ of $\boldsymbol{x}$, the sample covariance matrix is computed by $\boldsymbol{S} = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^\top$. Let $l_1 > \cdots > l_p > 0$ be the eigenvalues of $\boldsymbol{S}$, and denote its eigen-decomposition as $\boldsymbol{S} = \boldsymbol{P}\boldsymbol{L}\boldsymbol{P}^T$, where $\boldsymbol{L} = \mathrm{diag}\{l_1, \cdots, l_p\}$, and $\boldsymbol{P}$ is an orthornormal matrix contains the eigenvectors of $\boldsymbol{S}$. Consider a set of null hypotheses:

$$H_{0k} : d = k, \quad k = 0, \cdots, p-1.$$

Under H_{0k}, the eigenvalues of $\boldsymbol{\Sigma}$ are $\lambda_i = \delta_i + \sigma^2 (i = 1, 2, \cdots, k)$ and $\lambda_{k+j} = \sigma^2 (j = 1, 2, \cdots, p-k)$.

When H_{0k} is true and the noise are normally distributed,

$$T_k \equiv \frac{n(\sum_{j=k+1}^{p} l_j^2 - (p-k)^{-1}(\sum_{j=k+1}^{p} l_j)^2)}{2((p-k)^{-1}\sum_{j=k+1}^{p} l_j)^2} \\ \to \chi^2_{(p-k-1)(p-k+2)/2}.$$

And the signal dimension can be consistently estimated by

$$\mathrm{Asym}(\hat{d}) = \underset{k=0,1\cdots,p-2}{\arg\min} \{T_k < c_{k,n}\},$$

where $c_{k,n} \to \infty$ and $c_{k,n} = o(n)$. In practice, it is a nontrivial task to select the sequences $c_{k,n}$ that guarantees the performance of $\hat{d}$ in finite samples. Therefore usually a divide-and-conquer strategy is employed to estimate the dimension; see for example [11]. Due to this limitation, the asymptotic test performs poorly when p is large.

When the signals and noise are all normally distributed, [8] proposed the IID likelihood-based information theoretic criteria. The AIC criterion is given by

$$\mathrm{AIC}(\hat{d}) = \underset{k=0,1,\cdots,p-1}{\arg\min} \{-2\log L_k + 2k(2p-k)\}$$

where $L_k = \left(\frac{\prod_{i=k+1}^{p} l_i^{1/(p-k)}}{\frac{1}{p-k}\sum_{i=k+1}^{p} l_i}\right)^{(p-k)n}$ is the Gaussian likelihood ratio under the assumption of k signals. Similarly, the MDL criterion is defined as

$$\mathrm{MDL}(\hat{d}) = \underset{k=0,1,\cdots,p-1}{\arg\min} \{-\log L_k + \frac{1}{2}k(2p-k)\log n\}.$$

It was argued in [8] that MDL yields a consistent estimate of the number of signals, while the AIC tends to overestimate. The Gaussian assumption was relaxed in [7] and they showed that the consistency of MDL also holds when $\boldsymbol{x}$ in (1) has an elliptical distribution. However, it is unclear if they still have a good performance in the noisy ICA setting.

Note that both the asymptotic test and the information criteria depend on the fact that l_k starts to become stable at $k = d+1$. This fact could also be described by the behavior of the ratio of the empirical eigenvalues. So, motivated by [12] and [9], we propose the following ratio-based estimator

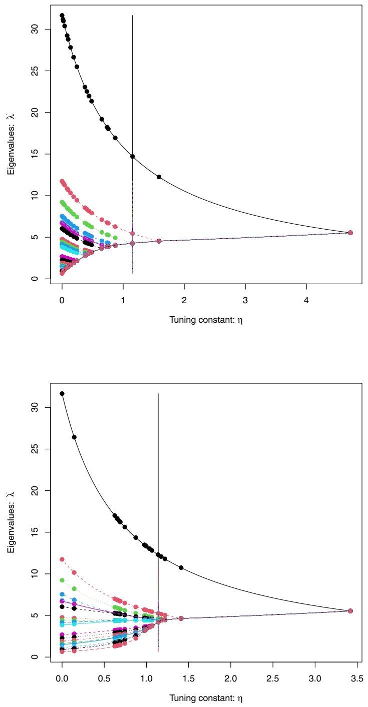

Fig. 1. Elasso solution path for three signals when the sample size is 50 and the dimension is 20. The upper figure is plotted using the SS weights and the lower figure is plotted using the MP weights. The vertical bars give the cut-off point based on cross-validation.

$$S(\hat{d}) = \underset{i=1,\cdots,M}{\arg\min} \frac{l_i - l_{i+1}}{l_i + l_{i+1}},$$

where $d < M < p$ is a predefined constant. Note that we define the optimal as the minimum of the ratio, instead of the maximum as done in most literature of factor models, because, in noisy ICA setting, the true optimal ratio drop to 0 when $i = d+1$. In the factor model context, as discussed by [13] and the reference therein, it is often argued that it would be more stable to use the eigenvalues of the correlation matrix instead of the eigenvalues of the covariance matrix. In the noisy ICA model however, it is not clear that the correlation matrix would have the eigenvalue structure as postulated in hypothesis (4) and therefore we advise against this approach. As the original ratio-based estimator defined in [9] was shown to work better, when the dimension of time series increases as the sample size, we expect $S(\hat{d})$ would perform better than the asymptotic test and the information criterior for large p.

Testing a sequence of H_{0k} is also a type of model selection problem, where covariance matrix with grouped eigenvalues is regarded as a model. Recently, Tyler and Yi [10] studied penalized sample covariance matrix estimators based on a class of non-smooth penalties, that can automatically partition the

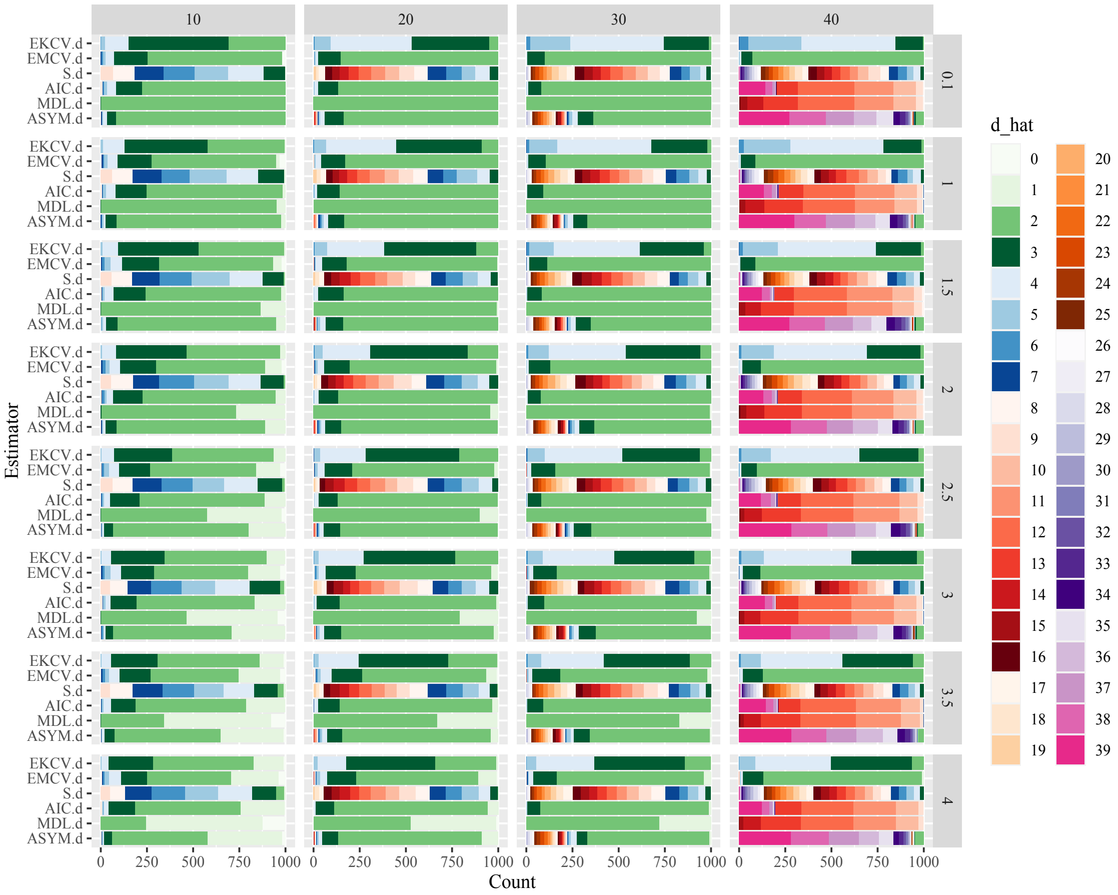

Fig. 2. Estimated signal number when the sample size is $n = 50$ based on 1,000 repetitions.The columns correspond to different dimension of the observable data and the rows to different variance levels.

empirical eigenvalues into distinct groups; thus is referred as the Elasso-estimators. The Elasso estimator is defined as:

$$\hat{\boldsymbol{\Sigma}}_\eta = \arg\min_{\boldsymbol{\Sigma}>\boldsymbol{0}} Tr(\boldsymbol{\Sigma}^{-1}\boldsymbol{S}) + \log|\boldsymbol{\Sigma}| + \eta\sum_{i=1}^{p} a_i \log\lambda_i,$$

where $a_1 \geq \cdots \geq a_p$ are the weights satisfying $\sum_{i=1}^{p} a_i = 0$. It was shown that the solution $\hat{\boldsymbol{\Sigma}}_\eta$ has the form

$$\hat{\boldsymbol{\Sigma}}_\eta = \boldsymbol{P}\boldsymbol{\Lambda}_\eta\boldsymbol{P}^\top, \quad \boldsymbol{\Lambda}_\eta = \mathrm{diag}\{\hat{\lambda}_1, \cdots, \hat{\lambda}_p\},$$

where $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_p$ is the solution of

$$\min_{\lambda_1 \geq \cdots \geq \lambda_q > 0} \sum_{j=1}^{p} \{l_j/\lambda_j + (1+\eta a_j)\log(\lambda_j)\}.$$

Different values of the tuning parameter η yield different multiplicities of eigenvalues that solve the above optimization problem. The best tuning could be selected by regular K-fold cross-validation or model cross-validation; see [10] for details. If the best tuning suggests the roots have r different groups, with the $p-k$ smallest eigenvalues in one group, thus having weight (multiplicity) $w_r = p-k$, and all the larger eigenvalues are distinct, with weight $w_i = 1, (i = 1, \cdots, r-1)$, respectively, then the eigenvalues of $\hat{\boldsymbol{\Sigma}}_\eta$ has the structure

$$\hat{\lambda}_1 > \cdots > \hat{\lambda}_k > \hat{\lambda}_{k+1} = \cdots = \hat{\lambda}_p.$$

The signal dimension could thus be estimated by

$$\mathrm{Elasso}(\hat{d}) = \arg\min_{i=1,\cdots,r}\{w_i > 1\} - 1.$$

The choice of the weights $a_j, j = 1, \cdots, p$ depends on the application of interest. For example, if we want to group the smallest eigenvalues together, we could choose the SS weight, defined as

$$a_1 = 1, a_2 = 1, \cdots, a_{p-1} = 1, a_p = -(p-1). \tag{5}$$

If we want to get general multiplicities of eigenvalues, [10] suggests using the Marchenko-Pastur (MP) weight,

$$a_{mp,i} = \xi_j - \bar{\xi}, \quad \xi_j = F^{-1}\{(p-j+0.5)/p; p/n\},$$

where $F(x;\nu)$ is the Marchenko-Pastur distribution function with parameter ν. Many useful properties, like the model consistency when both n and p grow to infinity, for the MP weight, have been shown in the paper. If using the MP weight, we could first find the group index that contains the most

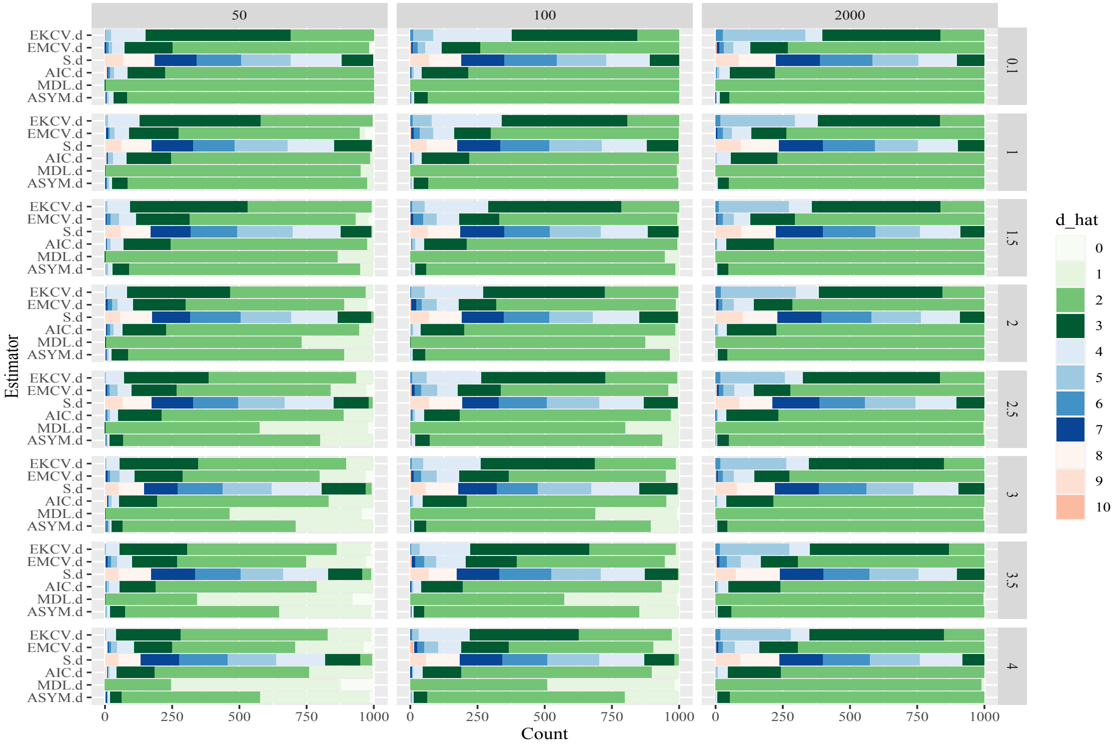

Fig. 3. Estimated signal number when the variable dimension is $p = 10$ based on 1,000 repetitions. The columns correspond to different sample size of the observable data and the rows to different noise variance levels.

eigenvalues and the signal number could then be estimated by the sum of the multiplicities before that index group.

We illustrate the above arguments with a toy example. Figure 1 plots the Elasso solution path using the data, with $n = 50$, generated from a three-signal 20-variate noisy ICA model described in detail in Section III. The upper figure path uses the SS weights and the lower path uses the MP weights, with the vertical lines indicating the corresponding optimal partition selected by the regular 5-fold cross-validation and the model 5-fold cross-validation [10], respectively. It can be seen from the figures that, the SS weights group first the smallest eigenvalues and the best tuning partitions the roots into three distinct groups with respective multiplicities $w_1 = 1, w_2 = 1$, and $w_3 = 18$, indicating that the model has two different signals; the MP weights may group the smallest, middle-large and/or large eigenvalues simultaneously and the best tuning partitions the roots into five distinct groups with respective multiplicities $w_1 = 1, w_2 = 1, w_3 = 1, w_4 = 8, w_5 = 9$, indicating that the model has eleven signals.

As the Elasso estimator $\hat{\Sigma}_\eta$ was proposed to solve the insufficient sample size problem, we anticipate that the Elasso-based method would perform well for small n. And according to the underlying eigenvalue structure of the noisy ICA model, we would prefer to use the SS weight.

III. Simulation Study

In this section, we conduct a simulation study to compare the Elasso method with four other methods. We assume the data follow the noisy ICA model (1) and consider a similar set-up inspired by [4], that has $d = 3$ independent signals

$$s_1 \sim \text{logsitic}(0, 1), \quad s_2 \sim t_5, \quad s_3 \sim \text{unif}[0, 1].$$

All signals have been standardized to have mean zero and variance one. The noise $\boldsymbol{\epsilon}$ is taken from a $N_p(\mathbf{0}, \boldsymbol{I}_p)$ distribution and all the $p \times d$ elements of the mixing matrix $\boldsymbol{A}$ are generated independently from $N(0, 1)$. Eight different noise levels σ^2 are used for comparison:

$$\sigma^2 = (0.1, 1, 1.5, 2, 2.5, 3, 3.5, 4).$$

We consider two scenarios:

- Fixed n: $n = 50, p = 10, 20, 30, 40$;
- Fixed p: $p = 10, n = 50, 100, 2000$.

The following estimators of the signal dimension are to be compared: (1) S.d, ratio method based on $\boldsymbol{S}$, with $M = p-1$; (2) AIC.d, minimization of the AIC criterion; (3) MDL.d, minimization of the MDL criterion; (4) ASYM.d, asymptotic test method applied using a divide and conquer strategy where each test uses $\alpha = 0.05$; (5) EKCV.d, the Elasso method using the SS weights (5), with the tuning η determined by 5-fold cross-validation; (6) EMCV.d, the Elasso method obtained by 5-fold model cross-validation with the Marchenko-Pastur weights.

The simulations were repeated over 1000 runs, and the estimated signal dimension of different estimators, for the two scenarios, are plotted in Figure 2 and Figure 3, respectively.

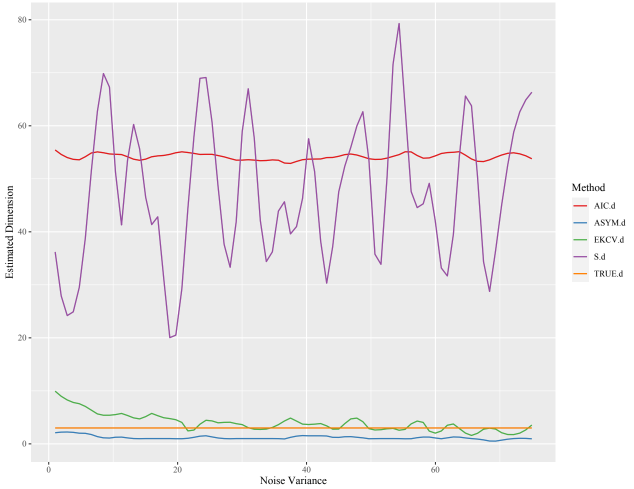

Fig. 4. Estimated signal dimension for the sound example for various estimators when the noise variance changes.

In general, the Elasso method EKCV performs better than any other method, especially for the increasing dimension case. When the dimension p is close to the sample size n, only the two Elasso-based estimators can still provide a reasonable estimator: the EMCV has a large probability to underestimate the signal number and the EKCV tends to overestimate; see the last column of Figure 2. For the increasing sample size case, Figure 3, the benefits of the Elasso methods are not prominent when $n = 50$ and become better when n gets larger, while the performance of the other estimators vary little for different sample size and noise level. Note also that the noise level affect the performance of Elasso estimators slightly, in some cases even in a positive way.

IV. Real Data Example

The results of the simulation suggest that if one uses the Elasso method with cross-validation, the resulting estimator can yield significant improvements in estimating the signal number of a noisy ICA setting. In practice, when the variable dimension is comparable with the sample size, we recommend choosing the SS weights (5) and select the tuning parameter by regular cross-validation. To further study the behavior of EKCV.d for large dimension, we consider three sound samples, publicy available in the R package JADE [14]. We mix the three sound signals with a 100×3 matrix $\boldsymbol{A}$, whose elements are randomly generated from a uniform distribution on $[0, 1]$, and add then 100-variate Gaussian noise to the mixed signals and let the noise variance vary. For simplicity, we use only the first $1,000$ observations and study how the noise level affects the estimation of the signal number, considering only AIC.d, S.d, ASYM.d and EKCV.d. The estimated signal numbers are plotted against the noise variance in Figure 4. From Figure 4 it is clear that in this example AIC and the gaps of the sample covariance matrix eigenvalues are not useful at all and hugely overestimate the number of signals. Also the successively applied hypothesis test is not performing well, it seems to find in most cases only one signal. The Elasso approach however performs well. In most cases it estimates 3 or little bit higher signal dimension, therefore reducing the dimension dramatically without loosing a signal.

V. Conclusion

Most ICA applications require a reduction of the signal dimension, assuming a known signal dimension; this does not hold in general. Recent advances propose several methods to estimate the signal dimension for noisy ICA. These methods are applicable when the sample size n is considerably larger than the variable dimension p, which, however, does not hold when n/p is not so favorable. Here we estimate the signal dimension in noisy ICA using a newly proposed Elasso method, which groups eigenvalues of the covariance together. Meanwhile we compare ratio statistic of the eigenvalues of the covariance as often used in the context of dynamic factor models in time series. Our simulations and example show that Elasso works well when n compared to p is less ideal, thus a valuable new tool in the blind source separation workflow. Further, Elasso does not presume the ICA independence assumption, but works with any model of the covariance structure (2), thus has broader applications. Also, when $n < p$, a regularized covariance estimator could replace S for the Elasso method, this is however a topic for future research.

References

[1] P. Comon and C. Jutten, *Handbook of Blind Source Separation: Independent component analysis and applications*. Academic press, 2010.

[2] A. Hyvärinen, “Independent component analysis: recent advances,” *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 371, no. 1984, p. 20110534, 2013.

[3] K. Nordhausen and H. Oja, “Independent component analysis: A statistical perspective,” *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 10, no. 5, p. e1440, 2018.

[4] J. Virta and K. Nordhausen, “Estimating the number of signals using principal component analysis,” *Stat*, vol. 8, no. 1, p. e231, 2019.

[5] T. W. Anderson, *An introduction to multivariate statistical analysis*. New York: Wiley, 2003.

[6] K. Nordhausen, H. Oja, and D. E. Tyler, “Asymptotic and bootstrap tests for subspace dimension,” *Journal of Multivariate Analysis*, vol. 188, p. 104830, 2022.

[7] L. Zhao, P. R. Krishnaiah, and Z. Bai, “On detection of the number of signals in presence of white noise,” *Journal of multivariate analysis*, vol. 20, no. 1, pp. 1–25, 1986.

[8] M. Wax and T. Kailath, “Determining the number of signals by information theoretic criteria,” in *ICASSP’84. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 9. IEEE, 1984, pp. 232–235.

[9] C. Lam and Q. Yao, “Factor modeling for high-dimensional time series: inference for the number of factors,” *The Annals of Statistics*, pp. 694–726, 2012.

[10] D. E. Tyler and M. Yi, “Lassoing eigenvalues,” *Biometrika*, vol. 107, no. 2, pp. 397–414, 2020.

[11] C. Muehlmann, F. Bachoc, K. Nordhausen, and M. Yi, “Test of the latent dimension of a spatial blind source separation model,” *To appear in Statistica Sinica*, 2022.

[12] A. Onatski, “Testing hypotheses about the number of factors in large factor models,” *Econometrica*, vol. 77, no. 5, pp. 1447–1479, 2009.

[13] J. Fan, J. Guo, and S. Zheng, “Estimating number of factors by adjusted eigenvalues thresholding,” *Journal of the American Statistical Association*, vol. 117, no. 538, pp. 852–861, 2022.

[14] J. Miettinen, K. Nordhausen, and S. Taskinen, “Blind source separation based on joint diagonalization in `R`: The packages `JADE` and `BSSasymp`,” *Journal of Statistical Software*, vol. 76, no. 2, pp. 1–31, 2017.