Final Report for EOARD Grant FA9550-17-1-0103

"Big data covariance estimation"

April 20, 2020

Name of Principal Investigators (PI and Co-PIs): Angela Montanari

- e-mail address : angela.montanari@unibo.it
- Institution : Department of Statistical Sciences University of Bologna
- Mailing Address : Via Belle Arti 41 40126 Bologna
- Phone : +39 051 2098241
- Fax : +39 051 232153

Period of Performance: 1/1/2017 – 12/31/2019

Abstract: The research aims at contributing to the development of statistical methods for the analysis of big data. It developed along two lines. On one side it addressed the issue of large covariance estimation based on the assumption that the covariance matrix can be decomposed into the sum of a low rank and a sparse component. In Farnè and Montanari (2020), a new estimator called UNALCE (UNshrunk ALgebraic Covariance Estimator) is proposed. UNALCE is able to recover intermediately spiked latent eigen-structures and intermediately sparse residual components. It is algebraically consistent, i.e. it recovers exactly latent rank and sparsity pattern. Besides optimally estimating the two components, it also gives the optimal covariance estimator in terms of Frobenius loss: this is obtained by the un-shrinkage of the recovered latent eigenvalues.
The second research line was aimed at developing a new model based clustering method suitable for high dimensional data combining model based clustering and random projection ensembles (Anderlucci, Fortunato, Montanari, 2020 under review).
Both research lines have produced a paper and have been presented at many conferences as invited talks. The research has also suggested interesting extensions that are still under development.

List of Publications and Significant Collaborations that resulted from your AOARD supported project: In standard format showing authors, title, journal, issue, pages, and date, for each category list the following:

a) papers published in peer-reviewed journals

Matteo Farnè, Angela Montanari (2020), "A large covariance matrix estimator under intermediate spikiness regimes" Journal of Multivariate Analysis, 176,

b) papers published in peer-reviewed conference proceedings,

Matteo Farnè (2017), 16-18 December 2017, London, UK, 10th International Conference of the ERCIM WG on Computational and Methodological Statistics, Book of Abstracts, http://www.cmstatistics.org/CMStatistics2017/docs/BoA%20CFE-CMStatistics%202017.pdf, p.65, ISBN 978-9963-2227-4-2

Matteo Farnè and Angela Montanari (2018), 3-5 April 2018, Limassol, Cyprus, CRoNoS Workshop and Spring Course on Multivariate Data Analysis and Software (CRoNoS MDA 2018), http://cmstatistics.org/CRONOSMDA2018/docs/CRONOSMDA2018_BoA_ElectronicVersion.pdf, p.5

Matteo Farnè and Matteo Barigozzi (2018), 17-20 September 2018, Joint meeting of the Italian Mathematical Union, the Italian Society of Industrial and Applied Mathematics and the Polish Mathematical Society, http://umi-ptm.im.pwr.edu.pl/wp-content/uploads/2018/09/Talks.pdf, p.290.
Matteo Farnè and Matteo Barigozzi (2018), 14-16 December 2018, London, UK, 11th International Conference of the ERCIM WG on Computational and Methodological Statistics, Book of Abstracts, http://www.cmstatistics.org/CMStatistics2018/docs/BoACFECMStatistics2018.pdf, p.79, ISBN 978-9963-2227-5-9
Matteo Farnè and Matteo Barigozzi (2019), 18-20 March 2019, Bayreuth, Germany, 6th European Conference on Data Analysis, Book of Abstracts, http://www.gfkl.org/ecda2019/wp-content/uploads/sites/7/2019/03/Book_of_Abstracts_FINAL.pdf, p.74

c) conference presentations without papers,

- 11th International Conference of the ERCIM WG on Computational and Methodological Statistics, 16-18 December 2017, London, UK, **Invited talk,** "*Factor model estimation by composite minimization*".
- Theoretical and algorithmic underpinnings of Big Data, Isaac Newton Institute for Mathematical Sciences, 15-19 January 2018, Cambridge, UK, **Poster presentation**, "*Factor model estimation by composite minimization*".
- Multivariate data analysis and software, 3-5 April 2018, Limassol, Cyprus, **Poster presentation**, "*Factor model estimation by composite minimization*".
- University of Wroclaw, 17-20 September 2018, Joint meeting of the Italian Mathematical Union, the Italian Society of Industrial and Applied Mathematics and the Polish Mathematical Society, Wroclaw, Poland, **Invited talk**, "*An algebraic estimator for large spectral matrices*", session entitled "Challenges and Methods of Modern Statistics".
-11th International Conference of the ERCIM WG on Computational and Methodological Statistics, 14-16 December 2018, Pisa, Italy, **Invited talk,** " *An algebraic estimator for large spectral matrices* ".
- 6th European Conference on Data Analysis, 18-20 March 2019, Bayreuth, Germany, **Invited talk**, " *An algebraic estimator for large spectral matrices* ".
StaTalk 2019 @ UniBO March 29th, 2019, Department of Statistical Sciences "Paolo Fortunati", Bologna, "*An algebraic estimator for large spectral density matrices*".
-SIS 2019, Smart Statistics for Smart Applications, 18-21 June, 2019, Milano, Italy, **Invited talk**, "*An algebraic estimator for large spectral density matrices* ".
- 6th RCEA Time Series Econometrics Workshop 22-23 June, 2019, Larnaca, Cyprus, Invited talk, "*An algebraic estimator for large spectral density matrices*"
- "RANDOM PROJECTION ENSEMBLE CLUSTERING" Poster at the Working Group on Mixture Models, Perugia, July 17th – 21st.
L. Anderlucci, F.Fortunato, A. Montanari (2017) "HIGH-DIMENSIONAL CLUSTERING VIA RANDOM PROJECTIONS" Invited talk at the CLADAG conference, Milan, September 13th – 15th.
F. Fortunato, A. Montanari, L. Anderlucci (2017) "RANDOM PROJECTION ENSEMBLE CLUSTERING" Organized Invited Sessions at the ERCIM conference, London, December 16th – 18th.

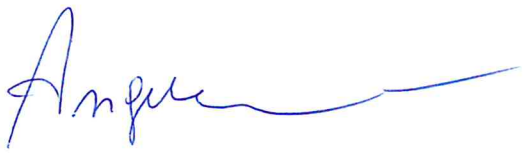
L. Anderlucci, F.Fortunato, A. Montanari (2018) "HIGH-DIMENSIONAL CLUSTERING WITH RANDOM PROJECTIONS" Talk at the Model-based Clustering and Classification (MBC2), Catania, September, 5th – 7th .
L. Anderlucci, F.Fortunato, A. Montanari (2018) "HIGH-DIMENSIONAL M ODEL-BASED CLUSTERING VIA RANDOM PROJECTIONS". Invited tal k at the CLADAG conference, Cassino, September, 11th – 13th.

d) manuscripts submitted but not yet published,

L. Anderlucci, F.Fortunato, A. Montanari (2020) "High-dimensional clustering via Random Projections" Journal of classification (under review)

Attachments: Publications a) and d) listed above

DD882: As a separate document, please complete and sign the inventions disclosure form.



Journal of Multivariate Analysis 176 (2020) 104577

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

A large covariance matrix estimator under intermediate spikiness regimes☆

Matteo Farnè *, Angela Montanari

Department of Statistical Sciences, University of Bologna, Bologna, Italy

ARTICLE INFO

Article history:
Received 9 March 2019
Received in revised form 18 November 2019
Accepted 21 November 2019
Available online 29 November 2019

AMS subject classifications:
62J10
65F35
93E24
65F50
15A18
62J07

Keywords:
Covariance matrix
Nuclear norm
Penalized least squares
Sparsity
Spiked eigenvalues
Un-shrinkage

ABSTRACT

This paper concerns large covariance matrix estimation via composite minimization under the assumption of low rank plus sparse structure. In this approach, the low rank plus sparse decomposition of the covariance matrix is recovered by least squares minimization under nuclear norm plus l_1 norm penalization. The objective is minimized via a singular value thresholding plus soft thresholding algorithm. This paper proposes a new estimator based on an additional least-squares re-optimization step aimed at un-shrinking the eigenvalues of the low rank component estimated in the first step. We prove that such un-shrinkage causes the final estimate to approach the target as closely as possible in spectral and Frobenius norm, while recovering exactly the underlying low rank and sparsity pattern. The error bounds are derived imposing that the latent eigenvalues scale to p^{α} and the maximum number of non-zeros per row in the sparse component scales to p^{δ}, where p is the dimension, $\alpha \in [0, 1]$, $\delta \in [0, 0.5]$, and $\delta < \alpha$. The sample size n is imposed to scale at least to $p^{1.5\delta}$. The resulting estimator is called UNALCE (UNshrunk ALgebraic Covariance Estimator), and it is shown to outperform state-of-the-art estimators, especially for what concerns fitting properties and sparsity pattern detection. The effectiveness of UNALCE is highlighted by a real example regarding ECB (European Central Bank) banking supervisory data.

1. Introduction

Estimation of population covariance matrices from samples of multivariate data is of interest in many high-dimensional inference problems: principal components analysis, classification by discriminant analysis, inferring a graphical model structure, and others. Depending on the goal, the interest is sometimes in inferring the eigenstructure of the covariance matrix (as in principal component analysis) and sometimes in estimating its inverse (as in discriminant analysis or in graphical models). Examples of application areas include gene arrays, functional magnetic resonance imaging, text retrieval, image classification, spectroscopy, climate studies, finance, and macro-economic analysis.

The theory of multivariate analysis for normal variables has been well worked out (see, for example, [2]). However, it soon became apparent that exact expressions were cumbersome, and that multivariate data were rarely Gaussian. The remedy was asymptotic theory for large samples and fixed, relatively small, dimensions. However, in recent years, datasets that do not fit into this framework have become very common, since nowadays the data can be high-dimensional, and sample sizes can be very small relative to dimension.

☆ Funding: This paper is based upon work supported by the Air Force Office of Scientific Research, United States under award number FA9550-17-1-0103.

* Corresponding author.
E-mail address: matteo.farne2@unibo.it (M. Farnè).

https://doi.org/10.1016/j.jmva.2019.104577
0047-259X/

The traditional covariance estimator, the sample covariance matrix, is known to be dramatically ill-conditioned in a large dimensional context, where the process dimension p is larger than or close to the sample size n, even when the population covariance matrix is well-conditioned. Two key properties of the matrix estimation process, i.e., numerical stability and identifiability, assume a particular relevance in large dimensions. Both properties are crucial for the theoretical derivation and the practical use of the estimate. A bad conditioned estimate suffers from collinearity and causes its inverse, the precision matrix, to dramatically amplify any error in the data. A large dimension may make it impossible to identify the unknown covariance structure, thus hampering the interpretation of the results.

Regularization approaches to large covariance matrix estimation are therefore being presented and addressed in the literature, both from theoretical and practical perspectives (see [14] for an exhaustive overview). Some authors propose shrinking the sample covariance matrix toward the identity matrix [24], others suggest applying nonlinear transforms to sample eigenvalues [25] or regularizing them by sample splitting [23], while some others consider tapering the sample covariance matrix, i.e., gradually shrinking the off-diagonal elements toward zero [9,21]. At the same time, a common approach is to encourage sparsity, either by a penalized likelihood approach [20] or by thresholding the sample covariance matrix in different ways: hard-thresholding [6], soft-thresholding [5], generalized thresholding [31], or adaptive thresholding [8]. A consistent bandwidth selection method for all these approaches is described in [30].

A different approach is based on the assumption of a low rank plus sparse structure for the covariance matrix:

$$\mathbf{\Sigma}^* = \mathbf{L}^* + \mathbf{S}^*, \tag{1}$$

where $\mathbf{L}^*$ is low rank with rank $r < p$, $\mathbf{S}^*$ is positive definite and sparse, with at most s non-zero off-diagonal elements, and $\mathbf{\Sigma}^*$ is a positive definite matrix. The generic covariance estimator $\hat{\mathbf{\Sigma}}$ can be written as

$$\hat{\mathbf{\Sigma}} = \mathbf{L}^* + \mathbf{S}^* + \mathbf{W} = \mathbf{\Sigma}^* + \mathbf{W}, \tag{2}$$

where $\mathbf{W}$ is an error term. The error matrix $\mathbf{W}$ may be deterministic or stochastic, as explained in [1]. If the data are Gaussian and $\hat{\mathbf{\Sigma}}$ is the unbiased sample covariance matrix $\mathbf{\Sigma}_n$, then $\mathbf{W}$ is distributed as a re-centred Wishart random matrix.

In [15], a large covariance matrix estimator, called POET (Principal Orthogonal complEment Thresholding), is derived under the assumption in (1). POET combines principal component analysis for the recovery of the low rank component and a thresholding algorithm for the recovery of the sparse component. The underlying model assumptions prescribe an approximate factor model with spiked eigenvalues for the data (growing with p), thus allowing to reasonably use the first r principal components of the sample covariance matrix. Furthermore, at the same time, sparsity in the sense of [5] is imposed on the residual matrix. The rank r of the low rank component is chosen by the information criteria in [4].

Indeed, rank selection represents a relevant issue: when p is large, setting a large rank would cause the estimate $\hat{\mathbf{\Sigma}}$ to be non-positive definite, while setting a small rank would cause a too relevant variance loss. In the discussion of [15], Yu and Samworth point out that the probability to underestimate the latent rank r does not asymptotically vanish if the eigenvalues are not really spiked at rate $O(p)$. In addition, we note that POET systematically overestimates the proportion of variance explained by the factors (given the true rank) because the eigenvalues of $\mathit{\Sigma}_n$ are more spiky than the true ones (as shown in [24]).

POET consistency holds given that a number of assumptions is satisfied. The key assumption is the pervasiveness of latent factors, which implies that the principal component analysis of $\mathbf{\Sigma}_n$ asymptotically identifies the eigenvalues and eigenvectors of $\mathbf{\Sigma}^*$ as p diverges. The results of [15] provide the convergence rates of the relative norm of $\hat{\mathbf{\Sigma}}_{POET} - \mathbf{\Sigma}^*$ (defined as $\|\hat{\mathbf{\Sigma}}_{POET} - \mathbf{\Sigma}^*\|_\Sigma = p^{-1/2}\|\mathbf{\Sigma}^{*-\frac{1}{2}}\hat{\mathbf{\Sigma}}_{POET}\mathbf{\Sigma}^{*-\frac{1}{2}} - \mathbf{I}_p\|_F$), the maximum norm of $\hat{\mathbf{\Sigma}}_{POET} - \mathbf{\Sigma}^*$, and the spectral norm of $\hat{\mathbf{S}}_{POET} - \mathbf{S}^*$. Under stricter conditions, $\hat{\mathbf{S}}_{POET}$ and $\hat{\mathbf{\Sigma}}_{POET}$ are proved to be non-singular with probability approaching 1.

At the same time, a number of non-asymptotic methods have been presented. In [11], the exact recovery of the covariance matrix in the noiseless context is first proved. The result is achieved by minimizing a specific convex non-smooth objective, which is the sum of the nuclear norm of the low rank component and the l_1 norm of the sparse component. In [10], which is an extension of [11], the exact recovery of the inverse covariance matrix in the noisy graphical model setting is provided. The authors prove that, in the worst case, the number of necessary samples in order to ensure consistency is $n = O\left(p^3/r^2\right)$, even if the required condition for the positive definiteness of the estimate is $p \leq 2n$.

An approximate solution to the recovery and identifiability of the covariance matrix in the noisy context is described in [1]. Even there, the condition $p \leq n$ is unavoidable for standard results on large deviations and non-asymptotic random matrix theory. An exact solution to the same problem, based on the results in [10], is then shown in [27]. The resulting estimator is called LOREC (LOw Rank and sparsE Covariance estimator) and is proved to be both algebraically and parametrically consistent in the sense of [10].

In [10], algebraic consistency is defined as follows.

Definition 1. A pair of symmetric matrices $(\mathbf{S}, \mathbf{L})$ with $\mathbf{S}, \mathbf{L} \in \mathbb{R}^{p\times p}$ is an algebraically consistent estimate of the low rank plus sparse decomposition (1) for the covariance matrix $\mathbf{\Sigma}^*$ if the following conditions hold:

(i) The sign pattern of $\mathbf{S}$ is the same as that of $\mathbf{S}^*$: $sign(\mathbf{S}_{ij}) = sign((\mathbf{S}^*)_{i,j})$, $\forall i, j$. Here, we assume that $sign(0) = 0$;
(ii) The rank of $\mathbf{L}$ is the same as the rank of $\mathbf{L}^*$;
(iii) Matrices $\mathbf{L} + \mathbf{S}$, $\mathbf{S}$, and $\mathbf{L}$ are such that $\mathbf{L} + \mathbf{S}$ and $\mathbf{S}$ are positive definite and $\mathbf{L}$ is positive semidefinite.

Parametric consistency holds if the estimates $(\mathbf{S}, \mathbf{L})$ are close to $(\mathbf{S}^*, \mathbf{L}^*)$ in some norm with probability approaching 1. In [10], it is defined as follows.

Definition 2. A pair of symmetric matrices $(\mathbf{S}, \mathbf{L})$ with $\mathbf{S}, \mathbf{L} \in \mathbb{R}^{p\times p}$ is a parametrically consistent estimate of the low rank plus sparse decomposition (1) for the covariance matrix $\mathbf{\Sigma}^*$ if the norm $g_\gamma = \max(\|\mathbf{S} - \mathbf{S}^*\|_\infty/\gamma, \|\mathbf{L} - \mathbf{L}^*\|_2)$, where $\|.\|_\infty$ denotes the maximum norm, converges to 0 with probability approaching 1.

LOREC shows several advantages compared to POET. The most important is that the estimates are both algebraically and parametrically consistent, while POET provides only parametric consistency. Nevertheless, LOREC suffers from some drawbacks, especially concerning the estimated latent eigenvalues. Moreover, the strict condition $p \leq n$ is required, while POET allows for $p\ln(p) \gg n$.

For these reasons, we propose a new estimator, UNALCE (UNshrunk ALgebraic Covariance Estimator), based on the 'unshrinkage' (the technical meaning will be clarified in Section 4) of the estimated eigenvalues of the low rank component, which allows to improve the fitting properties of LOREC systematically. We assume that the non-zero eigenvalues of $\mathbf{L}^*$ are proportional to p^α, $\alpha \in [0, 1]$ (the so called generalized spikiness context). Under the assumption that the maximum number of non-zeros per row in $\mathbf{S}^*$, called 'maximum degree', scales to p^δ (with $\delta \in [0, 0.5]$ and $\delta < \alpha$), we prove that our estimator possesses a non-asymptotic error bound allowing n to be as small as $p^{1.5\delta}$. We derive absolute bounds depending on α for the low rank, the sparse component, and the overall estimate. We also identify the conditions for positive definiteness and invertibility and for rank and sparsity pattern recovery. In this way, we provide a unique framework for covariance estimation via composite minimization under the low rank plus sparse assumption.

The remainder of the paper is organized as follows. In Section 2, we establish the notation, set up the model, briefly recall definitions and key properties of LOREC approach, and outline our novel contributions. In Section 3, we define a new estimator, that we call ALCE (ALgebraic Covariance Estimator), and we state the necessary assumptions for algebraic and parametric consistency. In Section 4, we then define the UNALCE (UNshrunk ALCE) estimator, proving that the unshrinkage of thresholded eigenvalues of the low rank component is the key to improve fitting properties as much as possible given a finite sample, preserving algebraic consistency. In Section 5, we propose a new model selection criterion specifically tailored to our model setting. In Section 6, we provide a real Euro Area banking data example which clarifies the effectiveness of our approach (a thorough simulation study is presented in the supplementary material, Section 2). Finally, in Section 7, we draw conclusions and discuss the most relevant findings. The proofs of all theorems and corollaries are reported in Appendix A.

2. Numerical estimation and spiked eigenvalues

2.1. Notation

Let us define a $p \times p$ symmetric positive-definite matrix $\mathbf{M}$. We denote by $\lambda_i(\mathbf{M})$, $i \in \{1, \ldots, p\}$, the eigenvalues of $\mathbf{M}$ in descending order. Then, we recall the following norm definitions:

(i) Element-wise:

(a) L_0 norm: $\|\mathbf{M}\|_0 = \sum_{i=1}^p \sum_{j=1}^p \mathbb{1}(\mathbf{M}_{ij} \neq 0)$, which is the total number of non-zeros;
(b) L_1 norm: $\|\mathbf{M}\|_1 = \sum_{i=1}^p \sum_{j=1}^p |\mathbf{M}_{ij}|$;
(c) Frobenius norm: $\|\mathbf{M}\|_F = \sqrt{\sum_{i=1}^p \sum_{j=1}^p \mathbf{M}_{ij}^2}$;
(d) Maximum norm: $\|\mathbf{M}\|_\infty = \max_{i\leq p, j\leq p} |\mathbf{M}_{ij}|$.

(ii) Induced by vector:

(a) $\|\mathbf{M}\|_{0,v} = \max_{i\leq p} \sum_{j\leq p} \mathbb{1}(\mathbf{M}_{ij} \neq 0)$, which is the maximum number of non-zeros per column, defined as the maximum 'degree' of $\mathbf{M}$;
(b) $\|\mathbf{M}\|_{1,v} = \max_{i\leq p} \sum_{j\leq p} |\mathbf{M}_{ij}|$;
(c) Spectral norm: $\|\mathbf{M}\|_2 = \lambda_1(\mathbf{M})$.

(iii) Schatten:

(a) Nuclear norm of $\mathbf{M}$, here defined as the sum of the eigenvalues of $\mathbf{M}$: $\|\mathbf{M}\|_* = \sum_{i=1}^p \lambda_i(\mathbf{M})$.

2.2. Model setup

Let us suppose that the population covariance matrix of our data is the sum of a low rank and a sparse component. A p-dimensional random vector $\mathbf{x}$ is said to have a low rank plus sparse structure if its covariance matrix $\mathbf{\Sigma}^*$ satisfies relationship (1):

$$\mathbf{\Sigma}^* = \mathbf{L}^* + \mathbf{S}^*,$$

where:

1. $\mathbf{L}^*$ is a positive semidefinite symmetric $p \times p$ matrix with at most rank $r \ll p$;
2. $\mathbf{S}^*$ is a positive definite $p \times p$ sparse matrix with at most $s \ll p(p-1)/2$ non-zero off-diagonal elements and maximum degree s'.

According to the spectral theorem, we can write $\mathbf{L}^* = \mathbf{U}_L\mathbf{D}\mathbf{U}_L^\top = \mathbf{B}\mathbf{B}^\top$, where $\mathbf{B} = \mathbf{U}_L\mathbf{D}^{1/2}$, $\mathbf{U}_L$ is a $p \times r$ semi-orthogonal matrix, $\mathbf{D}$ is a $r \times r$ diagonal matrix, with $\mathbf{D}_{jj} > 0, j \in \{1, \ldots, r\}$. Let us suppose that the $p \times 1$ random vector $\mathbf{x}$ is generated according to the following model:

$$\mathbf{x} = \mathbf{B}\mathbf{f} + \boldsymbol{\epsilon},$$

where $\mathbf{f}$ is a $r \times 1$ random vector with $\mathrm{E}(\mathbf{f}) = \mathbf{0}_r$, $\mathrm{V}(\mathbf{f}) = \mathbf{I}_r$, and $\boldsymbol{\epsilon}$ is a $p \times 1$ random vector with $\mathrm{E}(\boldsymbol{\epsilon}) = \mathbf{0}_p$, $\mathrm{V}(\boldsymbol{\epsilon}) = \mathbf{S}^*$. The random vector $\mathbf{x}$ is thus assumed to be zero mean, without loss of generality. Given a sample $\mathbf{x}_k$, $k \in \{1, \ldots, n\}$, $\boldsymbol{\Sigma}_n = (n-1)^{-1}\sum_{k=1}^n \mathbf{x}_k\mathbf{x}_k^\top$ is the $p \times p$ sample covariance matrix.

It is easy to observe that $\mathbf{x}$ follows a low rank plus sparse structure:

$$\mathrm{E}(\mathbf{x}\mathbf{x}^\top) = \mathrm{E}\left\{(\mathbf{B}\mathbf{f} + \boldsymbol{\epsilon})(\mathbf{B}\mathbf{f} + \boldsymbol{\epsilon})^\top\right\} = \mathrm{E}(\mathbf{B}^\top\mathbf{f}^\top\mathbf{f}\mathbf{B}) + \mathrm{E}(\mathbf{B}\mathbf{f}\boldsymbol{\epsilon}^\top) + \mathrm{E}(\boldsymbol{\epsilon}\mathbf{B}^\top\mathbf{f}^\top) + \mathrm{E}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top) = \mathbf{B}\mathbf{B}^\top + \mathbf{S}^* = \mathbf{L}^* + \mathbf{S}^* = \boldsymbol{\Sigma}^*$$

under the usual assumption that $\mathrm{cov}(\mathbf{f}, \boldsymbol{\epsilon}) = \mathrm{E}(\mathbf{f}\boldsymbol{\epsilon}') = \mathbf{0}_{r\times p}$ ($r \times p$ null matrix). Assuming p fixed, it is also useful to recall that $\boldsymbol{\Sigma}_n$ is asymptotically strongly consistent (see [19]). If we assume a normal distribution for $\mathbf{f}$ and $\boldsymbol{\epsilon}$, then $\boldsymbol{\Sigma}_n$ is unbiased for any fixed n (see [2]), and the matrix $\mathbf{W} := \boldsymbol{\Sigma}_n - (\mathbf{L}^* + \mathbf{S}^*)$ is distributed as a re-centred Wishart random matrix. In any case, the normality assumption is not essential for our setting.

2.3. Nuclear norm plus l_1 norm heuristics

Under the assumption in (1), the need arises to develop a method that can consistently estimate the covariance matrix $\boldsymbol{\Sigma}^*$ as well as determine the sparsity pattern of $\mathbf{S}^*$ and the spikiness pattern of the eigenvalues of $\mathbf{L}^*$ simultaneously. Such an estimation problem is stated as

$$\min_{\mathbf{L},\mathbf{S}} \frac{1}{2}\|(\mathbf{L} + \mathbf{S}) - \boldsymbol{\Sigma}_n\|_F^2 + \psi\,\mathrm{rank}(\mathbf{L}) + \rho\|\mathbf{S}\|_{0,off}, \tag{3}$$

where $\|\mathbf{S}\|_{0,off} = \sum_{i=1}^{p-1}\sum_{j=i+1}^{p} \mathbb{1}(\mathbf{S}_{ij}^* \neq 0)$ (because the diagonal of $\mathbf{S}$ is preserved as in [15]). This is a combinatorial problem, which is known to be NP-hard, since both $\mathrm{rank}(\mathbf{L})$ and $\|\mathbf{S}\|_{0,off}$ are non-convex.

The tightest convex relaxation of problem (3), as shown in [17], is

$$\min_{\mathbf{L},\mathbf{S}} \frac{1}{2}\|(\mathbf{L} + \mathbf{S}) - \boldsymbol{\Sigma}_n\|_F^2 + \psi\,\|\mathbf{L}\|_* + \rho\|\mathbf{S}\|_{1,off}, \tag{4}$$

where ψ and ρ are non-negative threshold parameters, and $\|\mathbf{S}\|_{1,off} = \sum_{i=1}^{p-1}\sum_{j=i+1}^{p}|\mathbf{S}_{ij}^*|$. The use of nuclear norm for covariance matrix estimation was introduced in [18]. The feasible set of (4) is the set of all $p \times p$ symmetric positive definite matrices $\mathbf{S}$ and all $p \times p$ symmetric positive semi-definite matrices $\mathbf{L}$.

The objective (4) is minimized according to an alternate thresholding algorithm, composed of a singular value thresholding (SVT, [7]) and a soft thresholding step [12]. In order to speed up convergence, Nesterov's acceleration scheme for composite gradient mapping minimization problems [28] is applied. Details of the algorithm are reported in the supplementary material, Section 2.

From a statistical viewpoint, (4) is penalized least squares heuristics, composed of a smooth least squares term ($0.5\|(\mathbf{L} + \mathbf{S}) - \boldsymbol{\Sigma}_n\|_F^2$) and a non-smooth composite penalty ($\psi\|\mathbf{L}\|_* + \rho\|\mathbf{S}\|_{1,off}$). The choice of (4) allows to lower the condition number of the estimates and the parameter space dimensionality simultaneously. In principle, different losses could be used, like Stein's one [13]. However, the classical Frobenius loss does not require normality and is computationally appealing.

From an algebraic viewpoint, (4) is an algebraic matrix variety recovery problem. In the covariance matrix setting described in Eq. (1), matrices $\mathbf{L}^*$ and $\mathbf{S}^*$ are assumed to come from the following sets of matrices:

$$\mathcal{B}(r) = \{\mathbf{L} \in \mathbb{R}^{p\times p} \mid \mathbf{L} = \mathbf{U}\mathbf{D}\mathbf{U}^\top, \mathbf{U} \in \mathbb{R}^{p\times r}\text{semi} - \text{orthogonal}, \mathbf{D} \in \mathbb{R}^{r\times r}\text{diagonal}\}, \tag{5}$$

$$\mathcal{A}(s) = \{\mathbf{S} \in \mathbb{R}^{p\times p} \mid |\mathrm{support}(\mathbf{S})| \leq s\}. \tag{6}$$

$\mathcal{B}(r)$ is the variety of matrices with at most rank r. $\mathcal{A}(s)$ is the variety of (element-wise) sparse matrices with at most s non-zero elements, where $\mathrm{support}(\mathbf{S})$ is the orthogonal complement of $ker(\mathbf{S})$ and $|\mathrm{support}(\mathbf{S})|$ denotes its dimension.

In [11], a notion of rank-sparsity incoherence is developed. It is expressed as an uncertainty principle between the sparsity pattern of a matrix and its row/column space. This notion has been introduced because $\mathbf{L}^*$ cannot be identified if it is nearly sparse, and $\mathbf{S}^*$ cannot be identified if it is nearly low rank. Denoting by $T(\mathbf{L}^*)$ and $\Omega(\mathbf{S}^*)$ the tangent spaces to $\mathcal{B}(r)$ and $\mathcal{A}(s)$, respectively, the following rank-sparsity incoherence measures between $T(\mathbf{L}^*)$ and $\Omega(\mathbf{S}^*)$ are defined:

$$\xi(T(\mathbf{L}^*)) = \max_{\mathbf{M}\in T(\mathbf{L}^*), \|\mathbf{M}\|_2 \leq 1} \|\mathbf{M}\|_\infty, \tag{7}$$

$$\mu(\Omega(\mathbf{S}^*)) = \max_{\mathbf{M}\in \Omega(\mathbf{S}^*), \|\mathbf{M}\|_\infty \leq 1} \|\mathbf{M}\|_2. \tag{8}$$

Quantities (7) and (8) control the identifiability of $\mathbf{L}^*$ and $\mathbf{S}^*$ in (1). In fact, a necessary and sufficient condition for identifiability is that $T(\mathbf{L}^*)$ and $\Omega(\mathbf{S}^*)$ have a transverse intersection, i.e., they intersect only at the origin. In [11], it is proved that $T(\mathbf{L}^*)$ and $\Omega(\mathbf{S}^*)$ are transverse if and only if (7) and (8) are small. Therefore, the product $\mu(\Omega(\mathbf{S}^*))\xi(T(\mathbf{L}^*))$ is a rank-sparsity incoherence measure and bounding it controls both for identification and recovery.

The described approach was first used for deriving the LOREC estimator in [27]. Therein, the reference matrix class imposed to $\mathbf{\Sigma}^*$ is

$$\mathbf{\Sigma}^*(\epsilon_0) = \left\{\mathbf{\Sigma}^* \in \mathbb{R}^{p\times p} : 0 < \epsilon_0 \leq \lambda_i(\mathbf{\Sigma}^*) \leq \epsilon_0^{-1},\ i \in \{1, \ldots, p\}\right\}$$

which is the class of positive definite matrices having uniformly bounded eigenvalues. In the context so far described, Luo (cf. [27]) proves that $\mathbf{L}$ and $\mathbf{S}$ can be identified and recovered with bounded error, and the rank of $\mathbf{L}$ as well as the sparsity pattern of $\mathbf{S}$ are exactly recovered.

The proof reproduces a similar proof in [5], but neglects a fundamental assumption on which that paper relies, i.e., that $\max_{i\leq p}\sum_{j\leq p}|\mathbf{\Sigma}^*_{ij}|^q = o(p)$ for some $q \in [0, 1[$. As stated in the supplementary material (Section 1), this can happen only if $\mathbf{L}^*$ is sparse, which contradicts rank-sparsity incoherence, thus making the model in [27] not identifiable.

2.4. Contribution of the paper

In this article, we propose an estimation algorithm for $\mathbf{\Sigma}^*$ under the assumption in (1) based on a nuclear norm plus l_1 norm penalization, as in [27]; however, contrary to [27], we derive the properties of the estimator under the sparsity assumption on $\mathbf{S}^*$ (see Assumption 4) and not on $\mathbf{\Sigma}^*$. This allows to avoid the non-identifiability trap and to enormously enlarge the set of recoverable pairs of matrices. We explicitly control the magnitude of $\xi(T(\mathbf{L}^*))$ and $\mu(\Omega(\mathbf{S}^*))$ with respect to p. More importantly, we allow for the generalized spikiness of the eigenvalues of $\mathbf{L}^*$ (cf. Yu and Samworth, [15], p. 656), thus modelling a large variety of real situations. In addition, we overcome the strict assumption $p \leq n$ by linking n to the degree of sparsity of $\mathbf{S}^*$. We call the resulting estimator ALCE (ALgebraic Covariance Estimator). In the end, since the singular value thresholding procedure has a significantly strong impact on sample eigenvalues when p is large and the latent eigenvalues are spiked, we apply an un-shrinkage step to the estimates of the latent eigenvalues. We name the resulting estimator UNALCE (UNshrunk ALCE). We prove that UNALCE is both algebraically and parametrically consistent. Within the class of algebraically consistent estimates, it minimizes the overall loss in Frobenius norm, given the finite sample and the threshold pair (ψ, ρ) in (4).

3. The ALgebraic Covariance Estimator (ALCE)

3.1. Component estimates and consistency

Let us suppose that the eigenvalues of $\mathbf{\Sigma}^*$ are intermediately spiked. This amounts to assume the generalized spikiness of latent eigenvalues in the sense of Yu and Samworth ([15], p. 656):

Assumption 1. All the eigenvalues of the $r \times r$ matrix $p^{-\alpha}\mathbf{B}^\top\mathbf{B}$ are bounded away from 0 for all p and $\alpha \in [0, 1]$.

If p is finite, Assumption 1 is equivalent to state that

$$\lambda_{1,\ldots,r}(\mathbf{\Sigma}^*) > \delta_\alpha p^\alpha, \tag{9}$$

$$\lambda_{r+1,\ldots,p}(\mathbf{\Sigma}^*) < \delta_\alpha p^\alpha, \tag{10}$$

for some $\delta_\alpha > 0$. We aim to study the properties of the covariance estimates obtained by heuristics (4) under the generalized spikiness assumption in a non-asymptotic context.

In order to reach this goal, we need to impose the following assumptions in our finite sample context.

Assumption 2. There exist $k_L, k_S > 0$, $\delta \in [0, 0.5]$, such that $\xi(T(\mathbf{L}^*)) = \sqrt{r/(k_L^2 p^{2\delta})}$, $\mu(\Omega(\mathbf{S}^*)) = k_S p^\delta$, $k_S/k_L \leq 1/54$ with $\delta < \alpha$.

Assumption 3. There exist $r_1, r_2 > 0$ and $b_1, b_2 > 0$ such that, for any $t > 0$, $k \leq n$, $i \leq r$, $j \leq p$:

$$\Pr(|\mathbf{f}_{ik}| > s) \leq \exp(-b_1/t), \qquad \Pr(|\boldsymbol{\epsilon}_{jk}| > s) \leq \exp(-b_2/t).$$

Assumption 4. There exist $c_1, c_2, c_3, \delta_2, \delta_2' > 0$, $\delta' \in [0, \delta+0.5]$ such that $\lambda(\mathbf{S}^*)_{min} > c_1$, $\min_{i,j\leq p}\mathrm{var}(\boldsymbol{\epsilon}_{ik}\boldsymbol{\epsilon}_{jk}) > c_2$ for any $k \leq n$, $i, j \leq p$, $s^*_{ii} \leq c_3$ for any $i \leq p$, $s' = \max_{i\leq p}\sum_{j\leq p}\mathbb{1}(\mathbf{S}^*_{ij} = 0) \leq \delta_2 p^\delta$ with $\delta_2 \geq k_S$ and $\|\mathbf{S}\|_{1,v} = \max_{i\leq p}\sum_{j\leq p}|\mathbf{S}^*_{ij}| \leq \delta_2' p^{\delta'}$.

Assumption 5. There exist $\delta_3, \delta_4 > 0$ such that $r = \delta_3 \ln p$ and $n \geq \delta_4 p^{1.5\delta}$.

Under those assumptions, we prove Theorem 1 which provides a non-asymptotic consistency result, particularly useful when p is not that large and $\alpha < 1$. In fact, in that case, principal components are far from convergence, and therefore, POET approach becomes suboptimal.

Theorem 1. *Let $T = T(\mathbf{L}^*)$ and $\Omega = \Omega(\mathbf{S}^*)$ be the tangent spaces to (5) and (6), respectively. Suppose that Assumptions 1–5 hold. Define $\psi = (1/\xi(T))(p^\alpha/\sqrt{n})$ and $\rho = \gamma\psi$, where $\xi(T)$ has been defined in (7), $\alpha \in [0, 1]$, $\gamma \in [9\xi(T), 1/(6\mu(\Omega))]$, and $\mu(\Omega)$ has been defined in (8). In addition, suppose that the minimum eigenvalue of $\mathbf{L}^*$ ($\lambda_r(\mathbf{L}^*)$) is greater than $C_2\psi/\xi^2(T)$. Then, with probability $1 - O(1/\min(p, n)^2)$, the pair $(\hat{\mathbf{L}}, \hat{\mathbf{S}})$ minimizing (4) recovers the rank of $\mathbf{L}^*$ ($\operatorname{rank}(\hat{\mathbf{L}}) = \operatorname{rank}(\mathbf{L}^*)$). Moreover, the matrix losses for each component are bounded as follows:*

$$\|\hat{\mathbf{L}} - \mathbf{L}^*\|_2 \leq C\psi, \qquad \|\hat{\mathbf{S}} - \mathbf{S}^*\|_\infty \leq C\rho.$$

We call ALCE (ALgebraic Covariance Estimator) the estimator of $\mathbf{\Sigma}^*$ in (2) obtained by estimating $\mathbf{L}^*$ by $\hat{\mathbf{L}}$ and $\mathbf{S}^*$ by $\hat{\mathbf{S}}$:

$$\hat{\mathbf{\Sigma}}_{ALCE} = \hat{\mathbf{L}}_{ALCE} + \hat{\mathbf{S}}_{ALCE}. \tag{11}$$

Theorem 1 states that, under all the prescribed assumptions, the losses of the pair $(\hat{\mathbf{L}}_{ALCE}, \hat{\mathbf{S}}_{ALCE})$ obtained by minimizing (4) with respect to the true $(\mathbf{L}^*, \mathbf{S}^*)$ are bounded, and the rank of $\mathbf{L}^*$ is exactly recovered, provided that the minimum latent eigenvalue is large enough, as well as the underlying matrix varieties T and Ω are transverse enough. Exploiting the consistency norm of [10], i.e.,

$$g_\gamma = \max\left(\frac{\|\hat{\mathbf{S}}_{ALCE} - \mathbf{S}^*\|_\infty}{\gamma}, \|\hat{\mathbf{L}}_{ALCE} - \mathbf{L}^*\|_2\right),$$

it follows from Theorem 1 that

$$g_\gamma(\hat{\mathbf{S}}_{ALCE} - \mathbf{S}^*, \hat{\mathbf{L}}_{ALCE} - \mathbf{L}^*) \leq C\frac{1}{\xi(T)}\frac{p^\alpha}{\sqrt{n}}$$

with probability $1 - O(1/\min(p, n)^2)$.

In the proof of Theorem 1, Assumption 2 is needed in order to ensure consistency and rank recovery. In fact, an identifiability condition for problem (4), as shown in Theorem 1, is $\xi(T(\mathbf{L}^*))\mu(\Omega(\mathbf{S}^*)) \leq 1/54$. According to [11], $\sqrt{r/p} \leq \xi(T(\mathbf{L}^*)) \leq 1$ and $\min_{i\leq p}\sum_{j\leq p}\mathbb{1}(\mathbf{S}^*_{ij} \neq 0) \leq \mu(\Omega(\mathbf{S}^*)) \leq \max_{i\leq p}\sum_{j\leq p}\mathbb{1}(\mathbf{S}^*_{ij} \neq 0)$. It follows that $\xi(T(\mathbf{L}^*)) = 1$ with $\delta = 0$ in the worst case scenario and $\xi(T(\mathbf{L}^*)) = \sqrt{r/p}$ with $\delta = 0.5$ in the best case scenario, under the condition $k_S/k_L \leq 1/54$. Such assumption is essential to ensure the parametric consistency of the estimated pair in terms of matrix norms and the recovery of the underlying algebraic matrix varieties under model (2) (cf. [10]). The assumption $\delta < \alpha$ is required in order to ensure that conditions (9) and (10) hold under the condition $\lambda_r(\mathbf{L}^*) > C_2\psi/\xi^2(T)$ of Theorem 1 and to rule out the degenerate case $\delta = \alpha$.

Assumption 3 is necessary to ensure that large deviation theory can be applied to $\mathbf{f}_{ik}$, $\boldsymbol{\epsilon}_{jk}$, and $\mathbf{f}_{ik}\boldsymbol{\epsilon}_{jk}$ for all $i \leq r, j \leq p$, and $k \leq n$ (cf. [15]). Assumption 4 is necessary in order to apply the results of [5] on the sparse component which prescribe that $\mathbf{S}^*$ must be well conditioned with uniformly bounded diagonal elements. We stress that the maximum degree s' must be bounded to ensure parametric and algebraic consistency, because Assumption 2 ensures $\mu(\Omega(\mathbf{S}^*)) = k_S p^\delta$ with $\delta \leq 0.5$. This condition is different from the corresponding one in [15], which prescribes $\max_{i\leq p}\sum_{j\leq p}|\mathbf{S}^*_{ij}|^q < c_4, q \in [0, 1], c_4 > 0$.

In general, we can allow for $\|\mathbf{S}^*\|_2 \leq \|\mathbf{S}^*\|_{1,v} \leq \delta_2' p^{\delta'}$, $\|\mathbf{S}^*\|_1 \leq p\|\mathbf{S}^*\|_{1,v} \leq \delta_2' p^{1+\delta'}$, and $\|\mathbf{S}^*\|_0 = p + s \leq ps' \leq \delta_2 p^{1+\delta}$. In addition, we can also write $\|\mathbf{S}^*\|_2 \leq \|\mathbf{S}^*\|_{0,v} = \delta_2 p^\delta$ (as shown in [5]). The assumption $\delta' \leq \delta + 0.5$ is needed to respect the inequality $\|\mathbf{S}^*\|_{1,v} \leq \sqrt{p}\|\mathbf{S}^*\|_2$. We stress that the assumption $\delta < \alpha$ is enough to ensure $\|\mathbf{S}^*\|_2 = o(p)$ as p diverges, thus also guaranteeing POET consistency for each α given the true rank (see Yu and Samworth, [15], p. 656).

Assumption 5 prescribes that the latent rank is infinitesimal with respect to p and the sample size n is possibly smaller than p, but not smaller than $\delta_4 p^{1.5\delta}$. It leads to overcoming the restrictive condition $p \leq n$, since $\delta \leq 0.5$. The need for it arises in order to ensure coherence with Assumptions 1 and 4.

In Corollary 1, we prove the asymptotic consistency of ALCE estimates.

Corollary 1. *Suppose that Assumptions 1–5 hold. If the limit $\lim_{v\to\infty}\min_v(p_v^{2\alpha+2\delta}, n_v) = \infty$ with the path-wise restriction $\lim_{v\to\infty} p_v^{2\alpha+2\delta}/n_v = 0$ holds, then $\lim_{v\to\infty}\psi_v = 0$ for $\psi_v = p_v^\alpha/(\xi(T)\sqrt{n_v})$.*

Corollary 1 shows how p and n may cause the probabilistic error to annihilate in the limit. For the terminology about limit sequences, see [3]. Moreover, $\psi_v/p_v^{\alpha+\delta} \to 0$ as $\lim_{v\to\infty}\min_v(p_v^{2\alpha+2\delta}, n_v) = \infty$, thus establishing the asymptotic consistency in relative terms, resembling the 'blessing of dimensionality' described in [15].

In order to prove the recovery of the residual sparsity pattern, we add to the previous ones the following assumption.

Assumption 6. $2\delta \leq \alpha \leq 2\delta + \delta'$ and $0 < (C_3\delta)/(k_L\delta_4) < \delta'$.

We can then prove Theorem 2.

Theorem 2. *Suppose that all the assumptions of Theorem 1 hold. If the minimum absolute value of the non-zero off-diagonal entries of $\mathbf{S}^*$, $S_{min,off}$, is greater than $(C_3\psi)/\mu(\Omega)$ and Assumption 6 holds, then the matrix $\hat{\mathbf{S}}$ minimizing (4) exactly recovers the sparsity pattern of $\mathbf{S}^*$ with probability $1 - O(1/\min(p, n)^2)$ ($sign(\hat{\mathbf{S}}) = sign(\mathbf{S}^*)$).*

Theorem 2 states that the sparsity pattern of $\mathbf{S}^*$ is also recovered if the minimum absolute non-zero off-diagonal entry of $\mathbf{S}^*$ is large enough and Assumption 6 holds. Consequently, we can state that the condition on the minimum latent eigenvalue and the assumption $\delta < \alpha$ are more important than the condition on the minimum absolute non-zero off-diagonal entry. In fact, the former is strictly necessary both for proving parametric consistency and rank recovery. The latter is necessary only for proving sparsity pattern recovery, as an additional result, given that the former hold. The only consequence of its violation is that some non-zero elements of $\mathbf{S}^*$ are not recovered.

Assumption 6 is necessary for the following reason. Since the product between the minimum absolute non-zero off-diagonal entry of $\mathbf{S}^*$, $S_{min,off}$, and the maximum degree of $\mathbf{S}^*$, s', cannot overcome the L_1 norm of $\mathbf{S}^*$, $\max_{i\leq p}\sum_{j\leq p}|\mathbf{S}^*_{ij}|$, it follows from the condition $S_{min,off} > (C_3\psi)/\mu(\Omega)$ of Theorem 2 and Assumption 4 that

$$0 < \frac{C_3\psi}{\mu(\Omega)}s' < S_{min,off}s' < \max_{i\leq p}\sum_{j\leq p}|\mathbf{S}^*_{ij}| \leq \delta_2' p^{\delta'}. \tag{12}$$

Inequality (12), under Assumptions 2, 4, and 5, boils down to $(C_3\delta_2 p^{\alpha-2\delta})/(k_L\delta_4) < \delta_2' p^{\delta'}$ and $0 < (C_3\delta_2 p^{\alpha-2\delta})/(k_L\delta_4)$, which hold if and only if Assumption 6 is satisfied.

We stress that the conditions $\lambda_r(\mathbf{L}^*) > (C_2\psi)/\xi^2(T)$ and $S_{min,off} > (C_3\psi)/\mu(\Omega)$ under Assumptions 2 and 5 become $\lambda_r(\mathbf{L}^*) > C_2 p^{\alpha}$ and $S_{min,off} > C_3 p^{\alpha-2\delta}$, respectively. The latter in turn leads to (12), which holds under Assumption 6. Therefore, the resultant model setting is fully consistent with Assumptions 1 and 4.

A representative selection of the latent eigenvalue and sparsity patterns admitted under the described conditions is reported in the supplementary material, Section 2. We emphasize that, e.g. the algebraic consistency no longer forces the latent eigenvalues to scale to p, provided that the maximum degree of the residual component is scaled accordingly. In general, it is necessary that the minimum latent eigenvalue and absolute non-zero residual entry should be large enough to ensure algebraic consistency, but they can both depend on p^{α}, with α potentially smaller than 1. In particular, if we increase α, both $\lambda_r(\mathbf{L}^*)$ and $S_{min,off}$ must be larger to ensure identifiability. The same happens if p increases. On the contrary, if r increases, then $\mathbf{L}^*$ can have less spiked eigenvalues, while if δ increases, then $S_{min,off}$ is allowed to be smaller.

3.2. Error bounds for $\hat{\mathbf{S}}_{ALCE}$ and $\hat{\mathbf{\Sigma}}_{ALCE}$ in spectral and frobenius norm

Within the same framework, we can complete our analysis with the bounds for $\hat{\mathbf{S}}_{ALCE}$.

From $\|\hat{\mathbf{S}}_{ALCE} - \mathbf{S}^*\|_2 \leq s'\|\hat{\mathbf{S}}_{ALCE} - \mathbf{S}^*\|_\infty$, we obtain

$$\|\hat{\mathbf{S}}_{ALCE} - \mathbf{S}^*\|_2 \leq Cs'\xi(T)\psi = \phi_{\mathbf{S}} \tag{13}$$

From $\|\hat{\mathbf{S}}_{ALCE} - \mathbf{S}^*\|_F \leq \sqrt{ps'}\|\hat{\mathbf{S}}_{ALCE} - \mathbf{S}^*\|_\infty$, we obtain

$$\|\hat{\mathbf{S}}_{ALCE} - \mathbf{S}^*\|_F \leq C\sqrt{ps'}\xi(T)\psi. \tag{14}$$

$\hat{\mathbf{S}}_{ALCE}$ is positive definite if and only if $\lambda_p(\mathbf{S}^*) > \phi_{\mathbf{S}}$. Bounds (13) and (14) hold with probability $1 - O(1/\min(p,n)^2)$. For the inverse of $\hat{\mathbf{S}}_{ALCE}$, $\hat{\mathbf{S}}^{-1}_{ALCE}$, the same bounds hold with probability $1 - O(1/\min(p,n)^2)$:

$$\|\hat{\mathbf{S}}^{-1}_{ALCE} - \mathbf{S}^{*-1}\|_2 \leq Cs'\xi(T)\psi = \phi_{\mathbf{S}}, \qquad \|\hat{\mathbf{S}}^{-1}_{ALCE} - \mathbf{S}^{*-1}\|_F \leq C\sqrt{ps'}\xi(T)\psi.$$

if and only if $\lambda_p(\mathbf{S}^*) \geq 2\phi_{\mathbf{S}}$.

From Theorem 1, we can derive with probability $1 - O(1/\min(p,n)^2)$ the following bounds for $\hat{\mathbf{\Sigma}}_{ALCE}$:

$$\|\hat{\mathbf{\Sigma}}_{ALCE} - \mathbf{\Sigma}^*\|_2 \leq C(s'\xi(T) + 1)\psi = \phi, \qquad \|\hat{\mathbf{\Sigma}}_{ALCE} - \mathbf{\Sigma}^*\|_F \leq C(\sqrt{ps'}\xi(T) + \sqrt{r})\psi.$$

$\hat{\mathbf{\Sigma}}_{ALCE}$ is positive definite if and only if $\lambda_p(\mathbf{\Sigma}^*) > \phi$. The same bounds hold for the inverse covariance estimate $\hat{\mathbf{\Sigma}}^{-1}_{ALCE}$ with probability $1 - O(1/\min(p,n)^2)$:

$$\|\hat{\mathbf{\Sigma}}^{-1}_{ALCE} - \mathbf{\Sigma}^{*-1}\|_2 \leq C(s'\xi(T) + 1)\psi = \phi, \qquad \|\hat{\mathbf{\Sigma}}^{-1}_{ALCE} - \mathbf{\Sigma}^{*-1}\|_F \leq C(\sqrt{ps'}\xi(T) + \sqrt{r})\psi$$

given that $\lambda_p(\mathbf{\Sigma}^*) \geq 2\phi$.

Overall, ALCE estimator allows to recover a relaxed spiked eigen-structure, thus overcoming the condition $p \leq n$, even using the sample covariance matrix as estimation input (the ratio p/n directly impacts the error bound). Our bounds are in absolute norms and reflect the underlying degree of spikiness α. Our theory relies on the probabilistic convergence of the sample covariance matrix under the assumption that the data follow an approximate factor model with a sparse residual. If all the assumptions of Theorems 1 and 2 and Corollary 1 hold with $\lambda_p(\mathbf{S}^*) > \phi_S$ and $\lambda_p(\mathbf{\Sigma}^*) > \phi$, then both algebraic and parametric consistency are ensured in the sense of Definitions 1 and 2, respectively.

Compared to LOREC, ALCE minimizes the same heuristics but is consistent for a much wider range of real situations, including high-dimensional settings ($p > n$). However, they both share a problem related to input eigenvalues: as p increases and the latent eigenvalues are spiked, the nuclear norm heuristics may lead to eigenvalue over-shrinkage, as shown in the following Section. For this reason, we further improve ALCE by un-shrinking the estimates of latent eigenvalues.

4. The UNALCE estimator: A re-optimized ALCE solution

4.1. Motivation

As previously mentioned, when p is large and the latent eigenvalues are spiked, the singular value thresholding procedure may lead to eigenvalue over-shrinkage, because in that case, the top r eigenvalues of $\mathbf{\Sigma}_n$ estimate increasingly better the latent eigenvalues as p increases. Therefore, shrinking the top r sample eigenvalues leads to too small estimates of the latent eigenvalues, and this also inevitably affects the residual and overall estimate.

Let us define $\mathbf{\Delta}_L = \hat{\mathbf{L}}_{ALCE} - \mathbf{L}^*$, $\mathbf{\Delta}_S = \hat{\mathbf{S}}_{ALCE} - \mathbf{S}^*$, $\mathbf{\Delta}_\Sigma = \hat{\mathbf{\Sigma}}_{ALCE} - \mathbf{\Sigma}^*$. Another key aspect of Theorem 1 is that the two losses in $\mathbf{L}^*$ and $\mathbf{S}^*$ are bounded separately. This fact results in a negative effect on the overall performance of $\hat{\mathbf{\Sigma}}_{ALCE}$, represented by the loss $\|\mathbf{\Delta}_\Sigma\|_2$, since $\|\mathbf{\Delta}_\Sigma\|_2$ is simply derived as a function of $\|\mathbf{\Delta}_L\|_2$ and $\|\mathbf{\Delta}_S\|_2$ according to the triangle inequality $\|\mathbf{\Delta}_\Sigma\|_2 \leq \|\mathbf{\Delta}_L\|_2 + \|\mathbf{\Delta}_S\|_2$. Therefore, the need arises to also correct for this drawback, re-shaping $\hat{\mathbf{\Sigma}}_{ALCE}$, as the ALCE solution is somehow sub-optimal for the whole covariance matrix.

We approach these issues by a finite-sample analysis, which could be referred to as a re-optimized least squares method. We refer to the usual objective function (4) with $\|\mathbf{S}\|_1 = \|\mathbf{S}\|_{1,off} = \sum_{i=1}^{p-1}\sum_{j=i+1}^{p}|\mathbf{S}_{ij}|$, which is the l_1 norm of $\mathbf{S}$ excluding the diagonal entries, consistently with POET approach. We define $\mathbf{Y}_{pre}$ and $\mathbf{Z}_{pre}$ as the last updates in the gradient step of the minimization algorithm of (4) (see Section 2 in the supplementary material). $\mathbf{Y}_{pre}$ and $\mathbf{Z}_{pre}$ are the two matrices we condition upon in order to derive our finite-sample re-optimized estimates.

We note some analogies between our approach and the restricted maximum likelihood (REML) method as explained in [22,29]. More precisely, the minimization of (4) acts as the ML estimator of fixed effects, while our re-optimized least squares step acts as the estimator of variance components.

Let us define the recovered rank $\hat{r} = \text{rank}(\hat{\mathbf{L}}_{ALCE})$ and the recovered number of residual non-zeros $\hat{s} = |\text{support}(\hat{\mathbf{S}}_{ALCE})|$. In the second step, we exploit the consistency properties of the varieties $\hat{\mathcal{B}}(\hat{r})$ and $\hat{\mathcal{A}}(\hat{s})$ recovered in the first step, defined as

$$\hat{\mathcal{B}}(\hat{r}) = \{\mathbf{L} \in \mathbb{R}^{p\times p} \mid \mathbf{L} = \hat{\mathbf{U}}_{ALCE}\mathbf{D}\hat{\mathbf{U}}_{ALCE}^\top, \mathbf{D} \in \mathbb{R}^{\hat{r}\times\hat{r}}\text{diagonal}\}, \tag{15}$$

$$\hat{\mathcal{A}}(\hat{s}) = \{\mathbf{S} \in \mathbb{R}^{p\times p} \mid |\text{support}(\mathbf{S})| \leq \hat{s}\}. \tag{16}$$

In particular, based on Theorems 1 and 2, we rely on the parametric guarantees offered by $\hat{\mathcal{B}}(\hat{r})$ and $\hat{\mathcal{A}}(\hat{s})$, and we condition upon the recovered latent rank and residual off-diagonal sparsity pattern. In this way, conditioning on the first step, we can focus on re-optimizing our pair of estimates to improve the overall fitting as much as possible, restricting our search into $\hat{\mathcal{B}}(\hat{r})$ and $\hat{\mathcal{A}}(\hat{s})$.

4.2. Optimality

The recovered varieties $\hat{\mathcal{B}}(\hat{r})$ and $\hat{\mathcal{A}}(\hat{s})$ ensure the algebraic consistency of $(\hat{\mathbf{S}}_{ALCE}, \hat{\mathbf{L}}_{ALCE})$ under all the assumptions of Theorems 1 and 2. One might look for the solution (say $(\hat{\mathbf{L}}_{New}, \hat{\mathbf{S}}_{New})$) of the problem

$$\min_{\mathbf{L}\in\hat{\mathcal{B}}(\hat{r}),\mathbf{S}\in\hat{\mathcal{A}}(\hat{s})} \text{STL}(\mathbf{L},\mathbf{S}) = \|\mathbf{\Sigma}_n - (\mathbf{L}+\mathbf{S})\|_F^2, \tag{17}$$

where $\text{STL}(\mathbf{L},\mathbf{S})$ stands for *Sample Total Loss*. The sample covariance matrix follows the model $\mathbf{\Sigma}_n = \mathbf{L}^* + \mathbf{S}^* + \mathbf{W}$, given a sample of p-dimensional data vectors $\mathbf{x}_k$, $k \in \{1,\dots,n\}$. Our problem essentially is as follows: which pair $\mathbf{L} \in \hat{\mathcal{B}}(\hat{r})$, $\mathbf{S} \in \hat{\mathcal{A}}(\hat{s})$ satisfying algebraic consistency shows the best approximation properties of $\mathbf{\Sigma}_n$?

We prove the following result.

Theorem 3. *Define the spectral decomposition of* $\hat{\mathbf{L}}_{ALCE}$ *as* $\hat{\mathbf{U}}_{ALCE}\hat{\mathbf{D}}_{ALCE}\hat{\mathbf{U}}_{ALCE}^\top$ *and* $\hat{\mathbf{L}}_{New} = \hat{\mathbf{U}}_{ALCE}(\hat{\mathbf{D}}_{ALCE} + \breve{\psi} I_r)\hat{\mathbf{U}}_{ALCE}^\top$, *where* $\breve{\psi} > 0$ *is any prescribed threshold parameter. Define* $\hat{\mathbf{S}}_{New}$ *such that its off-diagonal elements are the same as* $\hat{\mathbf{S}}_{ALCE}$, *and* $\hat{\mathbf{\Sigma}}_{New}$ *such that its diagonal elements are the same as* $\hat{\mathbf{\Sigma}}_{ALCE}$, *respectively. In addition, set* $\text{diag}(\hat{\mathbf{S}}_{New}) = \text{diag}(\hat{\mathbf{\Sigma}}_{ALCE}) - \text{diag}(\hat{\mathbf{L}}_{New})$. *Then, supposing that all the assumptions of Theorems 1 and 2 hold, the minimum*

$$\min_{\mathbf{L}\in\hat{\mathcal{B}}(\hat{r}),\mathbf{S}\in\hat{\mathcal{A}}(\hat{s})} \|\mathbf{\Sigma}_n - (\mathbf{L}+\mathbf{S})\|_F^2$$

conditioning on $\mathbf{Y}_{pre}$ *and* $\mathbf{Z}_{pre}$ *is achieved with probability* $1 - O(1/\min(p,n)^2)$ *if and only if* $\mathbf{L} = \hat{\mathbf{L}}_{New}$ *and* $\mathbf{S} = \hat{\mathbf{S}}_{New}$.

Theorem 3 essentially states that the sample total loss (17) is minimized if we un-shrink the eigenvalues of $\hat{\mathbf{L}}_{ALCE}$ (re-adding the threshold $\breve{\psi}$). We call the resulting overall estimator $\hat{\mathbf{\Sigma}}_{New} = \hat{\mathbf{L}}_{New} + \hat{\mathbf{S}}_{New}$ UNALCE (UNshrunk ALgebraic Covariance Estimator). We stress the importance of conditioning on $\mathbf{Y}_{pre}$ and $\mathbf{Z}_{pre}$. Since $\mathbf{Y}_{pre}$ and $\mathbf{Z}_{pre}$ are the matrices minimizing $0.5\|\mathbf{\Sigma}_n - (\mathbf{L}+\mathbf{S})\|_F^2$ and $\breve{\psi}\|\mathbf{L}\|_* + \breve{\rho}\|\mathbf{S}\|_1$ jointly considered (see (4)), our finite-sample re-optimization step aims to re-compute $\min\|\mathbf{\Sigma}_n - (\mathbf{L}+\mathbf{S})\|_F^2$, once the effect of the composite penalty $\breve{\psi}\|\mathbf{L}\|_* + \breve{\rho}\|\mathbf{S}\|_1$ has been removed.

As shown in Appendix A, problem (17) can be decomposed into two problems: one involving $\mathbf{L}$ and the other involving $\mathbf{S}$ (see (A.8)). The problem in $\mathbf{L}$ is solved by the covariance matrix formed by the top $\hat{r}$ principal components of $\mathbf{Y}_{pre}$, which belongs by construction to $\hat{\mathcal{B}}(\hat{r})$ and is equal to $\hat{\mathbf{U}}_{ALCE}(\hat{\mathbf{D}}_{ALCE} + \breve{\psi}\mathbf{I}_r)\hat{\mathbf{U}}_{ALCE}^\top = \hat{\mathbf{L}}_{UNALCE}$. The problem in $\mathbf{S}$ collapses to the problem

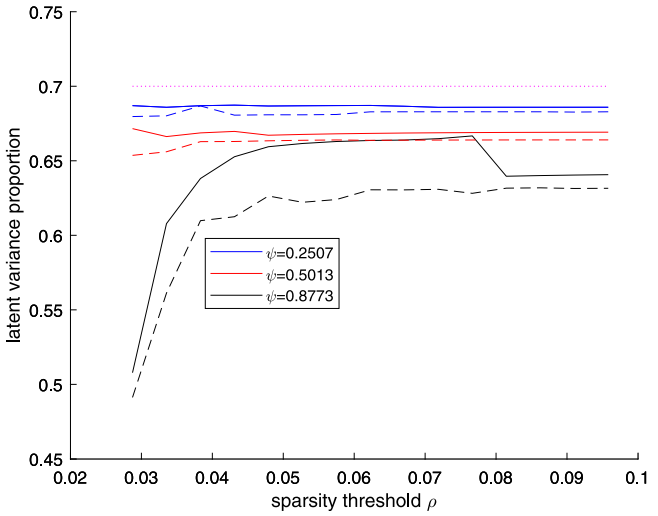


Fig. 1. This figure shows the proportion of latent variance θ estimated by UNALCE (solid line) and ALCE (dashed line) in correspondence to three selected values of latent eigenvalue thresholds ψ across twenty values of sparsity thresholds ρ. The reference value $\theta = 0.7$ is represented as a dotted line. The used sample is drawn from Setting 1 (see Section 2 in the supplementary material).

in $\mathbf{L}$ under the prescribed assumptions on the off-diagonal elements of $\hat{\mathbf{S}}_{UNALCE}$ (which ensures $\hat{\mathbf{S}}_{UNALCE} \in \hat{\mathcal{A}}(\hat{\mathbf{s}})$) and on the diagonal elements of $\hat{\mathbf{\Sigma}}_{UNALCE}$. The new estimate of the diagonal of $\mathbf{S}^*$ is simply the difference between the diagonal of the original $\hat{\mathbf{\Sigma}}_{ALCE}$ and that of the newly computed $\hat{\mathbf{L}}_{UNALCE}$. Note that our re-optimization step depends entirely on $\mathbf{\Sigma}_n$, as $\mathbf{Y}_{pre}$ and $\mathbf{Z}_{pre}$ are functions of $\mathbf{\Sigma}_n$.

Fig. 1 reports the proportion of latent variance $\hat{\theta} = (\sum_{i=1}^{p} \hat{\mathbf{L}}_{ii})/(\sum_{i=1}^{p} \hat{\mathbf{\Sigma}}_{ii})$ estimated by UNALCE and ALCE for three selected latent eigenvalue thresholds ψ over twenty sparsity thresholds ρ. We note that $\hat{\theta}$ gets systematically closer to the true $\theta = 0.7$ for $\hat{\mathbf{\Sigma}}_{UNALCE}$ with respect to $\hat{\mathbf{\Sigma}}_{ALCE}$ for all threshold pairs, and the performance difference is proportional to ψ. The sample used is drawn from our Setting 1 (see the supplementary material, Section 2 for more details).

4.3. Consequences

Four consequences of Theorem 3 are reported in Corollary 2.

Corollary 2. *Under the assumptions of Theorem 3, the differences between the total losses from the target in the spectral norm of $\hat{\mathbf{L}}_{ALCE}$ and $\hat{\mathbf{L}}_{UNALCE}$ and of $\hat{\mathbf{S}}_{ALCE}$ and $\hat{\mathbf{S}}_{UNALCE}$ are strictly positive and bounded with probability $1 - O(1/\min(p, n)^2)$ by $\breve{\psi}$. The differences between the total losses from the target in the Frobenius norm of $\hat{\mathbf{L}}_{ALCE}$ and $\hat{\mathbf{L}}_{UNALCE}$ and of $\hat{\mathbf{S}}_{ALCE}$ and $\hat{\mathbf{S}}_{UNALCE}$ are strictly positive and bounded with probability $1 - O(1/\min(p, n)^2)$ by $\sqrt{r}\breve{\psi}$.*

Two further relevant consequences of Theorem 3 are reported in Corollary 3.

Corollary 3. *Under the assumptions of Theorem 3, the difference between the sample total losses in the spectral norm of $\hat{\mathbf{\Sigma}}_{ALCE}$ and $\hat{\mathbf{\Sigma}}_{UNALCE}$ is strictly positive and bounded with probability $1 - O(1/\min(p, n)^2)$ by $\breve{\psi}$. The difference between the sample total losses in the Frobenius norm of $\hat{\mathbf{\Sigma}}_{ALCE}$ and $\hat{\mathbf{\Sigma}}_{UNALCE}$ is strictly positive and bounded with probability $1 - O(1/\min(p, n)^2)$ by $\sqrt{r}\breve{\psi}$.*

The following result compares the losses of $\hat{\mathbf{\Sigma}}_{UNALCE}$ and $\hat{\mathbf{\Sigma}}_{ALCE}$ from the target $\mathbf{\Sigma}^*$.

Theorem 4. *Under the assumptions of Theorem 3, the difference between the total losses from the target $\mathbf{\Sigma}^*$ in the spectral norm of $\hat{\mathbf{\Sigma}}_{ALCE}$ and $\hat{\mathbf{\Sigma}}_{UNALCE}$ is strictly positive and bounded with probability $1 - O(1/\min(p, n)^2)$ by $\breve{\psi}$. The difference between the total losses from the target in the Frobenius norm of $\hat{\mathbf{\Sigma}}_{ALCE}$ and $\hat{\mathbf{\Sigma}}_{UNALCE}$ is strictly positive and bounded with probability $1 - O(1/\min(p, n)^2)$ by $\sqrt{r}\breve{\psi}$.*

The rationale of the reported claims is as follows. We accept to pay the price of a non-optimal solution in terms of nuclear norm (we allow to increase $\|\hat{\mathbf{L}}\|_*$ by $r\breve{\psi}$), but we have a better fitting performance for the whole covariance matrix, decrementing the squared Frobenius loss of $\hat{\mathbf{\Sigma}}$ by a quantity bounded by $r\breve{\psi}^2$. The l_1 norm of $\mathbf{S}$ excluding the diagonal, $\|\hat{\mathbf{S}}\|_{1,off}$, is unvaried, while the norm $\|\hat{\mathbf{S}}\|_1$ (including the diagonal) is decreased by a quantity bounded by $\sqrt{r}\breve{\psi}$.

In Figs. 2 and 3, we report the differences between the sample total losses and the total losses of ALCE and UNALCE computed over the same sample of Fig. 1 for three selected latent eigenvalue thresholds ψ over twenty sparsity thresholds ρ.

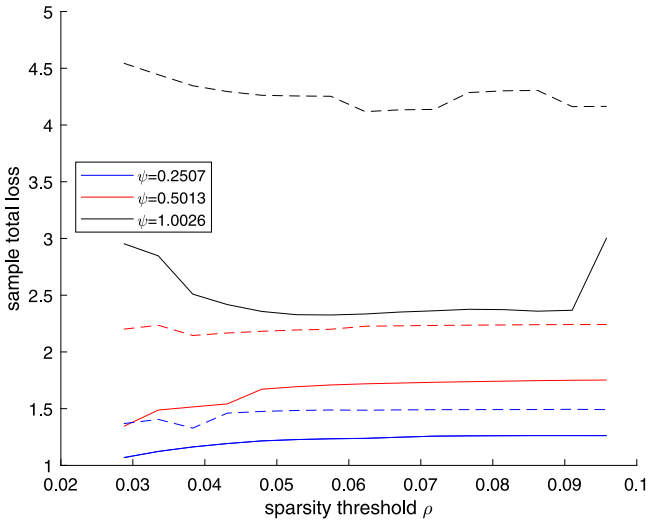


Fig. 2. This figure shows the sample total losses of $\hat{\mathbf{\Sigma}}_{UNALCE}$ (solid line) and $\hat{\mathbf{\Sigma}}_{ALCE}$ (dashed line) in one sample drawn from Setting 1 (see Section 2 in the supplementary material) in correspondence to three selected values of latent eigenvalue thresholds ψ across twenty values of sparsity thresholds ρ.

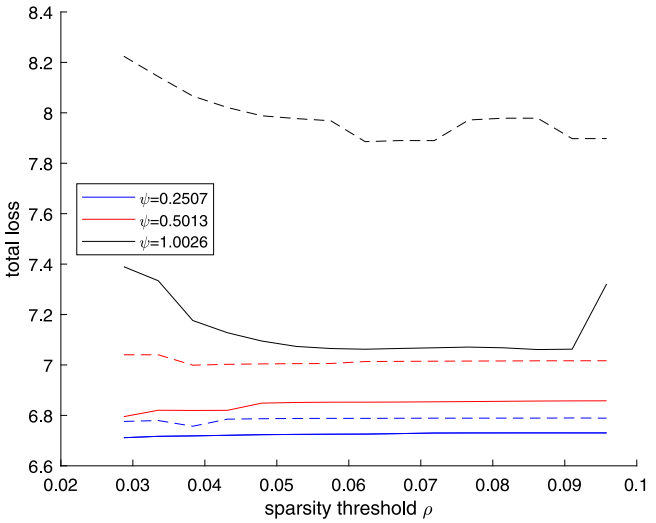


Fig. 3. This figure shows the total losses of $\hat{\mathbf{\Sigma}}_{UNALCE}$ (solid line) and $\hat{\mathbf{\Sigma}}_{ALCE}$ (dashed line) drawn from Setting 1 (see Section 2 in the supplementary material) in correspondence to three selected values of latent eigenvalue thresholds ψ across twenty values of sparsity thresholds ρ.

We note that the gain is relevant for UNALCE over all threshold pairs, is proportional to ψ, and never overcomes its theoretical maximum $\sqrt{r}\psi$ (in Figs. 2 and 3 $r = 4$). We stress that the gain is positive for each prescribed threshold pair $(\breve{\psi}, \breve{\rho})$, satisfying the conditions of Theorem 1, while the overall performance also depends on the threshold selection criterion (see Section 5).

A consequence of Corollaries 2 and 3 and Theorem 4 is that we can reduce the numerical instability of our estimates as much as possible in terms of the expected variance of estimated eigenvalues. In fact, defining $\mu_{\mathbf{L}} = \mathrm{E}(tr(\mathbf{L})/p)$, $\mu_{\mathbf{S}} = \mathrm{E}(tr(\mathbf{S})/p)$, and $\mu_{\mathbf{\Sigma}} = \mathrm{E}(tr(\mathbf{\Sigma})/p)$ and recalling the following equalities according to [24]

$$\frac{1}{p}\mathrm{E}\left\{\sum_{i=1}^{p}(\hat{\lambda}_{L,i} - \mu_L)^2 \mid \mathbf{\Sigma}_n\right\} = \frac{1}{p}\sum_{i=1}^{p}(\lambda_{L,i} - \mu_L)^2 + \mathrm{E}(\|\hat{\mathbf{L}} - \mathbf{L}^*\|^2|\mathbf{\Sigma}_n),$$

$$\frac{1}{p}\mathrm{E}\left\{\sum_{i=1}^{p}(\hat{\lambda}_{S,i} - \mu_S)^2 \mid \mathbf{\Sigma}_n\right\} = \frac{1}{p}\sum_{i=1}^{p}(\lambda_{S,i} - \mu_S)^2 + \mathrm{E}(\|\hat{\mathbf{S}} - \mathbf{S}^*\|^2|\mathbf{\Sigma}_n),$$

$$\frac{1}{p}\mathrm{E}\left\{\sum_{i=1}^{p}(\hat{\lambda}_{\Sigma,i}-\mu_{\Sigma})^2 \mid \mathbf{\Sigma}_n\right\} = \frac{1}{p}\sum_{i=1}^{p}(\lambda_{\Sigma,i}-\mu_L)^2 + \mathrm{E}(\|\hat{\mathbf{\Sigma}}-\mathbf{\Sigma}^*\|^2|\mathbf{\Sigma}_n),$$

we note that, under the assumptions of Theorem 3 the UNALCE estimated eigenvalues are maximally concentrated with probability $1-O(1/\min(p,n)^2)$, because $\hat{\mathbf{L}}_{UNALCE} = \min_{\mathbf{L}\in\hat{\mathcal{B}}(\hat{r})}(\|\mathbf{L}-\mathbf{L}^*\|^2|\mathbf{\Sigma}_n)$, $\hat{\mathbf{S}}_{UNALCE} = \min_{\mathbf{S}\in\hat{\mathcal{A}}(\hat{s})}(\|\mathbf{S}-\mathbf{S}^*\|^2|\mathbf{\Sigma}_n)$, $\hat{\mathbf{\Sigma}}_{UNALCE} = \min_{\mathbf{\Sigma}=\mathbf{L}+\mathbf{S},\mathbf{L}\in\hat{\mathcal{B}}(\hat{r}),\mathbf{S}\in\hat{\mathcal{A}}(\hat{s})}(\|\mathbf{\Sigma}-\mathbf{\Sigma}^*\|^2|\mathbf{\Sigma}_n)$, given the finite sample and a threshold pair $(\breve{\psi},\breve{\rho})$ satisfying the conditions of Theorem 1.

The following corollary extends our framework to the performance of the inverse covariance estimate $\hat{\mathbf{\Sigma}}^{-1}_{UNALCE}$.

Corollary 4. *Under the assumptions of Theorem 3, the difference between the total losses from the target in the spectral norm of $\hat{\mathbf{\Sigma}}^{-1}_{ALCE}$ and $\hat{\mathbf{\Sigma}}^{-1}_{UNALCE}$ is strictly positive and bounded with probability $1-O(1/\min(p,n)^2)$ by $\breve{\psi}$. The difference between the total losses from the target in the Frobenius norm of $\hat{\mathbf{\Sigma}}^{-1}_{ALCE}$ and $\hat{\mathbf{\Sigma}}^{-1}_{UNALCE}$ is strictly positive and bounded with probability $1-O(1/\min(p,n)^2)$ by $\sqrt{r}\breve{\psi}$.*

Finally, we study how the necessary conditions to ensure the positive definiteness of UNALCE estimates evolve with respect to the ALCE ones. The following corollary holds.

Corollary 5. *$\hat{\mathbf{L}}_{UNALCE}$ is positive semi-definite if $\lambda_r(\mathbf{L}^*) > C_2p^{\alpha}-\breve{\psi}$. $\hat{\mathbf{S}}_{UNALCE}$ is positive definite if $\lambda_p(\mathbf{S}^*) > \phi_S + r\breve{\psi}/p$. $\hat{\mathbf{\Sigma}}_{UNALCE}$ is positive definite if $\lambda_p(\mathbf{\Sigma}^*) > \phi + r\breve{\psi}/p$.*

Theorems 1, 2, and 3 and Corollaries 1 and 5 ensure the algebraic and parametric consistency of the pair $(\hat{\mathbf{L}}_{UNALCE}, \hat{\mathbf{S}}_{UNALCE})$ in the sense of Definitions 1 and 2.

5. A new model selection criterion: *MC*

In empirical applications, the selection of thresholds ψ and ρ in Eq. (4) requires a model selection criterion consistent with the described estimation method and the consistency norm g_γ (recall that $g_\gamma = \max(\|\mathbf{S}-\mathbf{S}^*\|_\infty/\gamma, \|\mathbf{L}-\mathbf{L}^*\|_2)$). Our aim is to detect the optimal threshold pair (ψ,ρ) with respect to the spikiness/sparsity trade-off. In order to exploit g_γ with model selection purposes, we need to make the two terms comparable, i.e., the need of rescaling both arguments of g_γ arises.

First, we note that if all the estimated latent eigenvalues are equal, then we have $\|\hat{\mathbf{L}}\|_* = \hat{r}\|\hat{\mathbf{L}}\|_2$. As the condition number of $\hat{\mathbf{L}}$ increases, we have $\hat{r}\|\hat{\mathbf{L}}\|_2 > \|\hat{\mathbf{L}}\|_*$. Consequently, the quantity $\hat{r}\|\hat{\mathbf{L}}\|_2$ acts as a penalization term against the presence of too small eigenvalues. Analogously, if $\hat{\mathbf{S}}$ is diagonal, it holds $\|\hat{\mathbf{S}}\|_\infty = \|\hat{\mathbf{S}}\|_{1,v}$. As the number of non-zeros increases, it holds $\|\hat{\mathbf{S}}\|_{1,v} > \|\hat{\mathbf{S}}\|_\infty$. Therefore, the quantity $\|\hat{\mathbf{S}}\|_{1,v}$ acts as a penalization term against the presence of too many non-zeros.

In order to compare the magnitude of the two quantities, we divide the former by the trace of $\hat{\mathbf{L}}$, estimated by $\hat{\theta}\mathrm{trace}(\mathbf{\Sigma}_n)$, and the latter by the trace of $\hat{\mathbf{S}}$, estimated by $(1-\hat{\theta})\mathrm{trace}(\mathbf{\Sigma}_n)$. Our maximum criterion *MC* can be therefore defined as follows:

$$MC(\psi,\rho) = \max\left\{\frac{\hat{r}\|\hat{\mathbf{L}}\|_2}{\hat{\theta}}, \frac{\|\hat{\mathbf{S}}\|_{1,v}}{\gamma(1-\hat{\theta})}\right\}, \tag{18}$$

where $\gamma = \rho/\psi$ is the ratio between the sparsity and the latent eigenvalue threshold.

MC criterion is by definition mainly intended to catch the proportion of variance explained by the factors. For this reason, it tends to choose quite sparse solutions with a small number of non-zeros and a small proportion of absolute residual covariance, unless the non-zero entries of $\hat{\mathbf{S}}$ are prominent, as Theorem 2 prescribes. The *MC* method performs considerably better than the usual cross-validation using H-fold Frobenius loss (cf. [27]). In fact, minimizing a loss based on a sample approximation such as the Frobenius one causes the parameter $\hat{\theta}$ to be significantly shrunk. The threshold setting which shows a minimum for *MC* criterion (given that the estimate $\hat{\mathbf{\Sigma}}$ is positive definite) is the best in terms of composite penalty, considering the latent low rank and sparse structure simultaneously.

Since we apply *MC* criterion to choose thresholds both for UNALCE and ALCE, we observe that the overall performance of the two methods is very similar, even if a little margin in favour of UNALCE is always present (see Section 2.2 in the supplementary material for more details).

6. A Euro Area banking data example

This section provides a real example on the performance of POET and UNALCE based on a selection of Euro Area banking data. We acknowledge the assistance of the European Central Bank, where one of the authors spent a semester as a PhD trainee, in providing access to high-level banking data. Here, we use the covariance matrix computed on a selection of balance sheet indicators relative to the last quarter of 2014 for some of the most relevant Euro Area banks. The overall number of banks (our sample size) is $n = 365$. These indicators are the ones needed for supervisory reporting, and they include capital and financial variables.

Table 1
Supervisory data: this table reports the main results of $\hat{\mathbf{\Sigma}}_{UNALCE}$ and $\hat{\mathbf{\Sigma}}_{POET}$ estimated on a selection of 382 supervisory indicators referred to 365 Euro Area banks with reference date Q4,2014. In particular, $\hat{r}$ is the latent rank, $\hat{s}$ is the number of recovered off-diagonal non-zeros in $\hat{\mathbf{S}}$, $\hat{\pi}_{\hat{s}}$ is the percentage of recovered non-zeros over the number of off-diagonal elements in $\hat{\mathbf{S}}$, $\hat{\theta} = (100 \sum_{i=1}^{p} \hat{\mathbf{L}}_{ii})/(\sum_{i=1}^{p} \hat{\mathbf{\Sigma}}_{ii})$ is the percentage of latent variance, $\hat{\rho}_{\hat{\mathbf{S}}} = (100 \sum_{i=1}^{p} \sum_{j=i+1}^{p} |\hat{\mathbf{S}}_{ij}|)/(\sum_{i=1}^{p} \sum_{j=i+1}^{p} |\hat{\mathbf{\Sigma}}_{ij}|)$ is the percentage of absolute residual covariance, $\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}_n\|_F$ is the sample total loss, $cond(\hat{\mathbf{\Sigma}}) = \lambda_{max}(\hat{\mathbf{\Sigma}})/\lambda_{min}(\hat{\mathbf{\Sigma}})$ is the condition number of the overall estimate, $cond(\hat{\mathbf{S}}) = \lambda_{max}(\hat{\mathbf{S}})/\lambda_{min}(\hat{\mathbf{S}})$ is the condition number of the sparse estimate, and $cond(\hat{\mathbf{L}}) = \lambda_{max}(\hat{\mathbf{L}})/\lambda_{min}(\hat{\mathbf{L}})$ is the condition number of the low rank estimate.

Supervisory data	UNALCE	POET
$\hat{r}$	6	6
$\hat{s}$	328	404
$\hat{\pi}_{\hat{s}}$	0.45	0.56
$\hat{\theta}$	32.47	61.23
$\hat{\rho}_{\hat{\mathbf{S}}}$	16.87	1.61
$\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}_n\|_F$	0.0337	0.0645
$cond(\hat{\mathbf{\Sigma}})$	$6.35e+15$	$6.68e+15$
$cond(\hat{\mathbf{S}})$	$2.78e+15$	$1.11e+15$
$cond(\hat{\mathbf{L}})$	3.1335	2.5625

The chosen raw variables were rescaled to the total asset of each bank. Then, a screening based on the importance of each variable, intended as the absolute amount of correlation with all the other variables, was performed in order to remove identities. The resulting very sparse data matrix contains $p = 382$ variables; here, we are in the typical $p > n$ case, where the sample covariance matrix is completely ineffective. We plot sample eigenvalues in the left panel of Fig. 4.

UNALCE estimation method selects a solution with a latent rank equal to $\hat{r} = 6$. The number of surviving non-zeros in the sparse component is $\hat{s} = 328$, which corresponds to a percentage $\hat{\pi}_{\hat{s}} = 0.45\%$ of 72 771 off-diagonal elements. Conditioning properties are inevitably very bad. In order to obtain a POET estimate, we exploit the algebraic consistency of $\hat{\mathbf{\Sigma}}_{UNALCE}$, setting the rank to $\hat{r} = 6$, and we perform cross-validation for threshold selection. The number of non-zeros estimated by POET is $\hat{s} = 404$ ($\hat{\pi}_{\hat{s}} = 0.56\%$). The results of both methods are reported in Table 1.

Apparently, one could argue that POET estimate is better; the estimated percentage of latent variance $\hat{\theta}$ is 61.23%, and the percentage of absolute residual covariance $\hat{\rho}_{\hat{\mathbf{S}}}$ is 1.61%. On the contrary, UNALCE method outputs $\hat{\theta} = 32.47\%$ and $\hat{\rho}_{\hat{\mathbf{S}}} = 16.87\%$. A relevant question thus arises: how much is the true percentage of variance explained by the factors? In fact, such a large percentage of latent variance, which depends on the use of the first six principal components, causes the absolute residual covariance percentage to be very low. Therefore, POET procedure gives *a priori* preference to the low rank part. This pattern does not change even if we choose a lower value for the rank.

On the contrary, the UNALCE estimate, which depends on a double-step iterative thresholding procedure, requires a larger magnitude of the non-zero elements in the sparse component. In fact, the percentage of lost covariance during the procedure is here 29.39%. Consequently, via rank/sparsity detection, UNALCE shows better approximation properties compared to POET; its Sample Total Loss is sensibly lower than that of POET (0.337 VS 0.645).

For UNALCE, the covariance structure appears so complex that a relevant percentage of absolute residual covariance is present. This allows us to explore the importance of variables, i.e., to explore which variables have the largest systemic power (the most relevant communality) or the largest idiosyncrasy (the most relevant residual variance).

In the right panel of Fig. 4, we plot in descending order the degree of each variable with respect to the estimated residual component $\hat{\mathbf{S}}_{UNALCE}$. The degree of the variable i with respect to a p-dimensional covariance matrix $\mathbf{M}$ is defined as

$$deg_{\mathbf{M},i} = \sum_{j=1}^{p} \mathbb{1}(\mathbf{M}_{ij} \neq 0). \tag{19}$$

We observe that only 62 out of 382 variables have at least one non-zero residual covariance with other variables.

In Table 2, we report the top six variables by estimated degree. These variables are related to the largest number of other variables. They are mainly credit-based indicators: financial assets through profit and loss, impaired assets, allowances to credit institutions and non-financial corporations, and cash.

In Table 3, we report the top five variables by estimated communality, defined for variable i as

$$comm_i = \frac{\hat{\mathbf{L}}_{UNALCE,ii}}{\hat{\mathbf{\Sigma}}_{UNALCE,ii}}, \qquad i \in \{1, \ldots, 382\}. \tag{20}$$

The results are very meaningful; the most systemic variables are debt securities, loans and advances to households, specific allowances for financial assets, and advances which are not loans to central banks. All these are fundamental indicators for banking supervision because they represent key indicators for the assessment of bank performance.

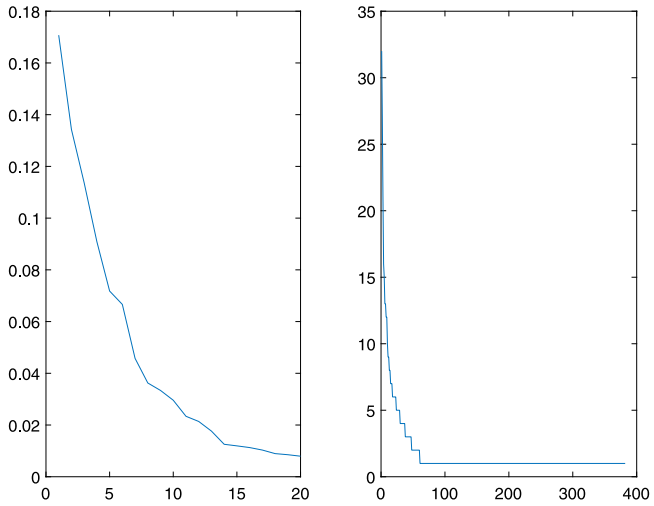


Fig. 4. Supervisory data: the left panel of this figure shows the twenty largest eigenvalues of the sample covariance matrix computed on a selection of 382 supervisory indicators referred to 365 Euro Area banks with reference date Q4,2014. The right panel of this figure plots in descending order the estimated degree of each supervisory indicator with respect to $\hat{\mathbf{S}}_{UNALCE}$, defined for variable i as $deg_{\hat{\mathbf{S}}_{UNALCE},i} = \sum_{j=1}^{p} \mathbb{1}(\hat{\mathbf{S}}_{UNALCE,ij} \neq 0)$, $i \in \{1, \ldots, 382\}$.

Table 2
Supervisory data: this table reports the top six variables by estimated degree with respect to $\hat{\mathbf{S}}_{UNALCE}$, defined for variable i as $deg_{\hat{\mathbf{S}}_{UNALCE},i} = \sum_{j=1}^{p} \mathbb{1}(\hat{\mathbf{S}}_{UNALCE,ij} \neq 0)$, $i \in \{1, \ldots, 382\}$. This measure counts how many variables are related to variable i that their estimated residual covariance is not null. Therefore, the reported variables are the most connected with all the others.

Supervisory indicator	Estimated degree
Financial assets designated at fair value through profit or loss	34
Central banks, Impaired assets [gross carrying amount]	25
Credit institutions, Collective allowances for incurred but not reported losses	20
Other financial corporations, Collective allowances for incurred but not reported losses	19
Cash, Cash balances at central banks and other demand deposits	16
Other financial corporations, Specific allowances for financial assets, collectively estimated	16

Table 3
Supervisory data: this table reports the top six variables by estimated communality via UNALCE, defined for variable i as $comm_i = \hat{\mathbf{L}}_{UNALCE,ii}/\hat{\mathbf{\Sigma}}_{UNALCE,ii}$. Therefore, the reported variables have a strong explanation power for banking supervision.

Supervisory indicator	Estimated communality
Debt securities	0.8414
Households, carrying amount	0.8210
Non-financial corporations, specific allowances for financial assets	0.8110
Loans and advances, specific allowances for financial assets, collectively estimated	0.7592
Advances that are not loans, central banks	0.7439

In Table 4, we report the top five variables by estimated idiosyncrasy, defined for variable i as

$$idio_i = \frac{\hat{\mathbf{S}}_{UNALCE,ii}}{\hat{\mathbf{\Sigma}}_{UNALCE,ii}}, \qquad i \in \{1, \ldots, 382\}. \tag{21}$$

We note that those indicators have a marginal power in the explanation of the common covariance structure and are much less relevant for supervisory analysis than the previous five.

In conclusion, our UNALCE procedure offers a more realistic view of the underlying covariance structure of a set of variables, allowing a larger part of covariance to be explained by the residual sparse component compared to POET.

7. Conclusions

In this work, we propose an estimator for large covariance matrices which are assumed to be the sum of a low rank and a sparse component. Estimation is performed by solving a regularization problem where the objective function is composed of a smooth Frobenius loss and a non-smooth composite penalty, which is the sum of the nuclear norm of the low rank component and the l_1 norm of the sparse component. Our estimator is called UNALCE (UNshrunk ALgebraic

Table 4
Supervisory data: this table reports the top six variables by estimated idiosyncrasy via UNALCE, defined for variable i as $idio_i = \hat{\mathbf{S}}_{UNALCE,ii}/\hat{\mathbf{\Sigma}}_{UNALCE,ii}$. Therefore, the reported variables have a marginal explanation power for banking supervision.

Supervisory indicator	Estimated idiosyncrasy
Credit card debt, central banks	0.9995
Other collateralized loans, other financial corporations	0.9986
Equity instruments, central banks, carrying amount	0.9971
Equity instruments, other financial corporations, carrying amount	0.9970
General governments, carrying amount of unimpaired assets	0.9970

Covariance Estimator). UNALCE provides consistent recovery of the low rank and the sparse component, as well as of the overall covariance matrix, under a generalized assumption of spikiness of latent eigenvalues and sparsity of the residual component. Thanks to the addition of an un-shrinkage step of the estimated latent eigenvalues, we can also improve numerical properties and minimize the overall loss given the finite sample and the threshold pair, while preserving algebraic consistency. In addition, we can overcome the restrictive condition $p \leq n$.

Moreover, in this paper, we also compare UNALCE and POET (Principal Orthogonal complEment Thresholding, see [15]), an asymptotic estimator which performs principal component analysis in order to recover the low rank component and uses a thresholding algorithm to recover the sparse component. Both estimators provide the usual parametric consistency, while UNALCE also provides the algebraic consistency of the estimate, i.e., the rank and position of residual non-zeros are simultaneously recovered by the solution algorithm. This automatic recovery is a crucial advantage compared to POET; the latent rank, in fact, is automatically selected and the sparsity pattern of the residual component is recovered considerably better.

In particular, we prove that UNALCE can effectively recover the covariance matrix even in the presence of spiked eigenvalues with rate $O(p)$, exactly as POET estimator does, allowing n to be as small as $O(p^{1.5\delta})$, where the maximum number of non-zeros per row in the sparse component is proportional to $O(p^{\delta})$. Moreover, we prove that the recovery is actually effective even if the eigenvalues show an intermediate degree of spikiness p^{α}, $\alpha \in [0, 1]$. The resulting loss is bounded accordingly to p^{α}, and all latent eigenvalues are recovered under the assumption $\delta < \alpha$. In this way, we obtain a generalized estimator of large covariance matrices by low rank plus sparse decomposition.

A real example of a set of Euro Area banking data shows that our tool is particularly useful for mapping the covariance structure among variables even in a large dimensional context. The variables with the largest systemic power, i.e., the ones mostly affecting the common covariance structure, can be identified, as well as the variables having the largest idiosyncratic power, that is, the ones characterized by the largest residual variance. In addition, the variables showing the largest idiosyncratic covariances can be identified. Particular forms of the residual covariance pattern can thus be detected, if present.

Our research may provide a basis for possible future developments in many directions. In the time series context, this procedure can be potentially extended to covariance matrix estimation under dynamic factor models. Another fruitful extension of our procedure is related to the spectral matrix estimation context. Finally, this tool can be potentially used in the Big data context, where both the dimension and the sample size are very large. This poses new computational and theoretical challenges, the solution of which is crucial in order to further extend the power of statistical modelling and its effectiveness in detecting patterns and underlying drivers of real phenomena.

CRediT authorship contribution statement

Matteo Farnè: Investigation, Methodology, Software, Data curation, Writing — original draft. **Angela Montanari:** Conceptualization, Supervision, Writing — review & editing, Resources, Funding acquisition.

Appendix A. Proofs

Proof of Theorem 1

First, we note that the deterministic analysis needed to ensure the identifiability of the matrix varieties $\mathcal{B}(r)$ and $\mathcal{A}(s)$ is directly inherited by [26]. In fact, Propositions 12, 13, and 14 in [26] may be directly applied to our setting, provided that the assumption $\xi(T(\mathbf{L}^*))\mu(\Omega(\mathbf{S}^*)) \leq 1/54$ and the conditions $\lambda_r(\mathbf{L}^*) > C_2\psi/\xi^2(T)$ and $\rho = \gamma\psi$ hold with $\gamma \in [9\xi(T), 1/(6\mu(\Omega))]$. In that case, it descends from the mentioned Propositions that $g_\gamma(\hat{\mathbf{S}} - \mathbf{S}^*, \hat{\mathbf{L}} - \mathbf{L}^*)$ is bounded, $\hat{\mathbf{L}} \in T(\mathbf{L}^*)$, $\hat{\mathbf{S}} \in \Omega(\mathbf{S}^*)$, and $\text{rank}(\hat{\mathbf{L}}) = \text{rank}(\mathbf{L}^*)$.

We stress that the remaining assumptions of Theorem 1 are not needed for this purpose. We also remark that parametric and rank consistency are not affected even if Assumption 6 and the condition $S_{min,off} > (C_3\psi)/\mu(\Omega)$ do not hold. The only consequence of that is that some residual non-zeros are not recovered (cf. [10], Corollary D.4 and D.6, and Proposition D.5 for more details).

Hence, we now focus on probabilistic analysis. Recalling that $\boldsymbol{\Sigma}_n = (n-1)^{-1}\sum_{k=1}^n \mathbf{x}_k\mathbf{x}_k^\top$ and $\mathbf{x}_k = \mathbf{B}\mathbf{f}_k + \boldsymbol{\epsilon}_k$, where $\mathbf{f}_k$ and $\boldsymbol{\epsilon}_k$, $k \in \{1,\dots,n\}$, are respectively the vectors of factor scores and residuals for each observation, we can decompose the error matrix $\mathbf{E}_n = \boldsymbol{\Sigma}_n - \boldsymbol{\Sigma}^*$ in four components as follows (cf. [15]):

$$\mathbf{E}_n = \boldsymbol{\Sigma}_n - \boldsymbol{\Sigma}^* = \hat{\mathbf{D}}_1 + \hat{\mathbf{D}}_2 + \hat{\mathbf{D}}_3 + \hat{\mathbf{D}}_4,$$

where:

$$\hat{\mathbf{D}}_1 = \left(n^{-1}\mathbf{B}\sum_{k=1}^n \mathbf{f}_k\mathbf{f}_k^\top - I_r\right)\mathbf{B}^\top,\ \hat{\mathbf{D}}_2 = n^{-1}\sum_{k=1}^n\left(\boldsymbol{\epsilon}_k\boldsymbol{\epsilon}_k^\top - \mathbf{S}^*\right),\ \hat{\mathbf{D}}_3 = n^{-1}\mathbf{B}\sum_{k=1}^n \mathbf{f}_k\boldsymbol{\epsilon}_k^\top,\ \hat{\mathbf{D}}_4 = \hat{\mathbf{D}}_3^\top.$$

Following [15], we note that

$$\|\hat{\mathbf{D}}_1\|_2 \le \left\|\frac{1}{n}\sum_{k=1}^n \mathbf{f}_{ik}\mathbf{f}_{jk} - \mathrm{E}(\mathbf{f}_{ik}\mathbf{f}_{jk})\right\|_2 \|\mathbf{B}\mathbf{B}^\top\|_2 \le r\left\|\frac{1}{n}\sum_{k=1}^n \mathbf{f}_{ik}\mathbf{f}_{jk} - \mathrm{E}(\mathbf{f}_{ik}\mathbf{f}_{jk})\right\|_\infty p^\alpha$$

where the second inequality depends on standard matrix norm properties and Assumption 1.

Under Assumption 3, we can apply Lemma 4 in [15], which claims

$$\max_{i,j\le r}\left|\frac{1}{n}\sum_{k=1}^n \mathbf{f}_{ik}\mathbf{f}_{jk} - \mathrm{E}(\mathbf{f}_{ik}\mathbf{f}_{jk})\right| \le C\frac{1}{\sqrt{n}}, \tag{A.1}$$

with probability $1 - O(1/n^2)$. Consequently, we obtain

$$\|\hat{\mathbf{D}}_1\|_2 \le Cr\sqrt{\frac{1}{n}}p^\alpha \le Cp^\alpha\sqrt{\frac{1}{n}} \tag{A.2}$$

because Assumption 5 prescribes that $r = \delta_3 \ln p$ and $\ln p = o(n)$.

Consider now the uniformity class of sparse matrices in [5]:

$$\left\{\mathbf{S}^* : \mathbf{S}^*_{ii} \le c_3, \qquad \max_{i\le p}\sum_{j\le p}\mathbb{1}(\mathbf{S}^*_{ij} \ne 0) \le c_0(p), \qquad \forall i\right\}.$$

Therein, Assumption 4 holds with $\delta = 1$ (assuming $q = 0$). Therefore, it is possible to write

$$\lambda_{max}(\mathbf{S}^*) \le \max_{i\le p}\sum_{j\le p}\mathbb{1}(\mathbf{S}^*_{ij} \ne 0) \le c_3c_0(p),$$

since the quantity $c_0(p)$ is constant with respect to p. Consequently, Lemma A.3 on p. 220 in [6] can be applied, which leads to the claim

$$\max_{i,j\le p}\left|\frac{1}{n}\sum_{k=1}^n \boldsymbol{\epsilon}_{ik}\boldsymbol{\epsilon}_{jk} - \mathrm{E}(\boldsymbol{\epsilon}_{ik}\boldsymbol{\epsilon}_{jk})\right| \le C\sqrt{\frac{\ln p}{n}}, \tag{A.3}$$

that holds with probability $1 - O(1/p^2)$.

Under Assumptions 2 and 4, however, the quantity $c_0(p)$ must be replaced by $c_0(p^\delta)$, $\delta \le 0.5$. Consequently, with respect to p, the rate in (A.3) is now too strong. Therefore, applying the recalled Lemma A.3 in [6], the following claim holds with probability $1 - O(1/p^2)$:

$$\|\hat{\mathbf{D}}_2\|_\infty = \max_{i,j\le p}\left|\frac{1}{n}\sum_{k=1}^n \boldsymbol{\epsilon}_{ik}\boldsymbol{\epsilon}_{jk} - \mathrm{E}(\boldsymbol{\epsilon}_{ik}\boldsymbol{\epsilon}_{jk})\right| \le Cp^{\delta-1}\sqrt{\frac{\ln p}{n}}. \tag{A.4}$$

Consequently, by (A.4), we can derive

$$\|\hat{\mathbf{D}}_2\|_2 \le Cp\|\hat{\mathbf{D}}_2\|_\infty = Cp^\delta\sqrt{\frac{\ln p}{n}} = Cp^\delta\sqrt{\frac{1}{n}}, \tag{A.5}$$

because $\ln p \ll n$ by Assumption 5.

To conclude, we study the random term $\max_{i\le r, j\le p}\left|n^{-1}\sum_{k=1}^n \mathbf{f}_{ik}\boldsymbol{\epsilon}_{jk}\right|$. We know from Lemma 3 in [15] that this term has exponential-type tails, due to Assumption 3. Thus, we only need to study how its standard deviation evolves in our context. We consider the following Cauchy–Schwarz inequality:

$$\max_{i\le r, j\le p}\left|\frac{1}{n}\sum_{k=1}^n \mathbf{f}_{ik}\boldsymbol{\epsilon}_{jk}\right| \le C\max_i\sqrt{\hat{V}(\mathbf{f}_i)}\max_j\sqrt{\hat{V}(\boldsymbol{\epsilon}_j)}.$$

From (A.1), we know that $\max_i\sqrt{\hat{V}(\mathbf{f}_i)} \leq C/\sqrt[4]{n}$ with probability $1-O(1/n^2)$. From (A.4), we know that $\max_j\sqrt{\hat{V}(\boldsymbol{\epsilon}_j)} \leq Cp^{(\delta-1)/2}\sqrt[4]{(\ln p)/n}$ with probability $1-O(1/p^2)$. It follows that with probability $1-O(1/\min(p,n)^2)$, it holds

$$\left\| n^{-1}\sum_{k=1}^{n}\mathbf{f}_k\boldsymbol{\epsilon}_k{}^\top \right\|_2 \leq \sqrt{pr}\left\| n^{-1}\sum_{k=1}^{n}\mathbf{f}_k\boldsymbol{\epsilon}_k{}^\top \right\|_\infty = C\sqrt{pr}\sqrt[4]{\frac{1}{n}}p^{(\delta-1)/2}\sqrt[4]{\frac{\ln p}{n}}.$$

Exploiting Assumption 5, the bound then becomes $Cp^{\delta/2}\sqrt{n^{-1}}$, since $r=\delta_3\ln p$ and $\ln p \ll n$.

Consequently, we obtain with probability $1-O(1/p^2)$ the following claim

$$\|\hat{\mathbf{D}}_3\|_2 \leq \left\| \frac{1}{n}\sum_{k=1}^{n}\mathbf{f}_k\boldsymbol{\epsilon}_k^\top \right\| \times \|\mathbf{B}\| \leq C\left(p^{\frac{\delta}{2}}\sqrt{\frac{1}{n}}\right)\left(p^{\frac{\alpha}{2}}\right) = Cp^{\frac{\alpha}{2}+\frac{\delta}{2}}\sqrt{\frac{1}{n}}, \tag{A.6}$$

because $\|\mathbf{B}\| = O(p^{\alpha/2})$ by Assumption 1.

Putting (A.2), (A.5), and (A.6) together, the following bound is proved with probability $1-O(1/\min(p,n)^2)$

$$\|\boldsymbol{\Sigma}_n - \boldsymbol{\Sigma}^*\|_2 \leq C\frac{p^\alpha}{\sqrt{n}}, \tag{A.7}$$

because $\delta<\alpha$ from Assumption 2. In fact, if $\delta>\alpha$, the condition of Theorem 1 $\lambda_r(\mathbf{L}^*) > (C_2\psi)/\xi^2(T)$ would result in $\lambda_r(\mathbf{L}^*) > C_2p^\delta$, thus violating Assumption 1 under Assumption 5.

In other words, the bound (A.7) means $\|\mathbf{E}_n\|_2 \to 0 \Leftrightarrow p^\alpha/\sqrt{n}\to 0$. Exploiting the basic property $\|.\|_\infty \leq \|.\|_2$ and the minimum for γ in the range of Theorem 1, we can also write $\|\mathbf{E}_n\|_\infty \to 0 \Leftrightarrow \xi(T)p^\alpha/\sqrt{n}\to 0$.

In order to prove Theorem 1, we observe from [26] that the only probabilistic component of the error norm $g_\gamma(\hat{\mathbf{S}}-\mathbf{S}^*, \hat{\mathbf{L}}-\mathbf{L}^*)$ is $g_\gamma(\mathbf{E}_n) = \max(\|\mathbf{E}_n\|_\infty/\gamma, \|\mathbf{E}_n\|_2)$, which in turn depends on $\|\mathbf{E}_n\|_2$ and $\|\mathbf{E}_n\|_\infty$. Therefore, setting $\psi = (1/\xi(T))(p^\alpha/\sqrt{n})$, it follows that the claims

$$g_\gamma(\hat{\mathbf{S}}-\mathbf{S}^*, \hat{\mathbf{L}}-\mathbf{L}^*) \leq C\frac{1}{\xi(T)}\frac{p^\alpha}{\sqrt{n}}, \operatorname{rank}(\hat{\mathbf{L}}) = \operatorname{rank}(\mathbf{L}^*)$$

hold with probability $1-O(1/\min(p,n)^2)$ under all the assumptions of Theorem 1. Parametric and rank consistency are thus guaranteed.

Proof of Corollary 1

We observe that, under Assumption 2, the bound $\psi = (1/\xi(T))(p^\alpha/\sqrt{n})$ tends to 0 if and only if $p^{2\alpha+2\delta}/n = o(1)$ as $\lim_{\nu\to\infty}\min_\nu(p_\nu^{2\alpha+2\delta}, n_\nu) = \infty$. As expected, the absolute bound vanishes only in the small dimensional case ($n \gg p^{\alpha+\delta}\log(p)$).

Proof of Theorem 2

If, in addition to all the assumptions and conditions of Theorem 1, Assumption 6 and the condition $S_{min,off} > (C_3\psi)/\mu(\Omega)$ hold, then we can fully apply Corollary D.4, D.6, Proposition D.5, and Lemma D.7 in [10] and conclude that the recovered sparsity pattern is also consistent: $sign(\hat{\mathbf{S}}_{ALCE}) = sign(\mathbf{S}^*)$.

Proof of Theorem 3

Conditioning on $\mathbf{Y}_{pre}$, $\mathbf{Z}_{pre}$, and $\boldsymbol{\Sigma}_{pre} = \mathbf{Y}_{pre}+\mathbf{Z}_{pre}$, we aim to solve

$$\min_{\mathbf{L}\in\hat{\mathcal{B}}(\hat{r}),\mathbf{S}\in\hat{\mathcal{A}}(\hat{s}),\boldsymbol{\Sigma}=\mathbf{L}+\mathbf{S}}\|\boldsymbol{\Sigma}-\boldsymbol{\Sigma}_n\|_F^2 = \|\boldsymbol{\Sigma}-\boldsymbol{\Sigma}_{pre}+\boldsymbol{\Sigma}_{pre}-\boldsymbol{\Sigma}_n\|_F^2.$$

By Cauchy–Schwarz inequality, it can be shown that

$$\|\boldsymbol{\Sigma}-\boldsymbol{\Sigma}_{pre}+\boldsymbol{\Sigma}_{pre}-\boldsymbol{\Sigma}_n\|_F^2 \leq \|\boldsymbol{\Sigma}-\boldsymbol{\Sigma}_{pre}\|_F^2 + \|\boldsymbol{\Sigma}_{pre}-\boldsymbol{\Sigma}_n\|_F^2.$$

$\boldsymbol{\Sigma}_{pre}$ solves the problem

$$\min_{\mathbf{L}\in\hat{\mathcal{B}}(\hat{r}),S\in\hat{\mathcal{A}}(\hat{s}),\boldsymbol{\Sigma}=\mathbf{L}+\mathbf{S}}\|\boldsymbol{\Sigma}_{pre}-\boldsymbol{\Sigma}_n\|_F^2$$

conditioning on the fact that $\breve{\psi}\|\mathbf{L}\|_* + \breve{\rho}\|\mathbf{S}\|_1$ is minimum over the same set.

Then, we can write

$$\|\boldsymbol{\Sigma}-\boldsymbol{\Sigma}_{pre}\|_F^2 = \|\mathbf{L}+\mathbf{S}-\mathbf{Y}_{pre}-\mathbf{Z}_{pre}\|_F^2.$$

By Cauchy–Schwarz inequality, it can be shown that

$$\|\mathbf{L}+\mathbf{S}-\mathbf{Y}_{pre}+\mathbf{Z}_{pre}\|_F^2 \leq \|\mathbf{L}-\mathbf{Y}_{pre}\|_F^2 + \|\mathbf{S}-\mathbf{Z}_{pre}\|_F^2. \tag{A.8}$$

Hence,

$$\min_{\mathbf{L}\in\hat{\mathcal{B}}(\hat{r}),\mathbf{S}\in\hat{\mathcal{A}}(\hat{s}),\boldsymbol{\Sigma}=\mathbf{L}+\mathbf{S}}\|\mathbf{L}+\mathbf{S}-\mathbf{Y}_{pre}+\mathbf{Z}_{pre}\|_F^2 \leq \min_{\mathbf{L}\in\hat{\mathcal{B}}(\hat{r})}\|\mathbf{L}-\mathbf{Y}_{pre}\|_F^2+\min_{\mathbf{S}\in\hat{\mathcal{A}}(\hat{s})}\|\mathbf{S}-\mathbf{Z}_{pre}\|_F^2.$$

The problem in $\mathbf{L}$ is solved by taking out the first $\hat{r}$ principal components of $\mathbf{Y}_{pre}$. By construction, the solution is $\hat{\mathbf{U}}_{ALCE}(\hat{\mathbf{D}}_{ALCE}+\breve{\psi}\mathbf{I}_r)\hat{\mathbf{U}}_{ALCE}^{\top}=\hat{\mathbf{L}}_{UNALCE}$. The problem in $\mathbf{S}$, assuming that the diagonal of $\hat{\boldsymbol{\Sigma}}_{ALCE}$ is given and the off-diagonal elements of $\hat{\mathbf{S}}$ are invariant, leads to:

$$\begin{aligned}\min_{\mathbf{S}\in\hat{\mathcal{A}}(\hat{s})}\|\mathbf{S}-\mathbf{Z}_{pre}\|_F^2=\min_{\mathbf{L}\in\hat{\mathcal{B}}(\hat{r})}\|(\hat{\boldsymbol{\Sigma}}-\mathbf{L})-(\boldsymbol{\Sigma}_{pre}-\mathbf{Y}_{pre})\|_F^2 &= \min_{\mathbf{L}\in\hat{\mathcal{B}}(\hat{r})}\|(\hat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}_{pre})-(\mathbf{L}-\mathbf{Y}_{pre})\|_F^2 \leq \\ &\leq \|\hat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}_{pre}\|_F^2+\|\mathbf{L}-\mathbf{Y}_{pre}\|_F^2=\mathbf{B}'+\mathbf{B}''.\end{aligned}$$

The following question now arises: which diagonal elements of $\mathbf{L}$ ensure the minimum of $\mathbf{B}'+\mathbf{B}''$? Term $\mathbf{B}'$ is fixed with respect to $\mathbf{L}$ because we are assuming the invariance of diagonal elements in $\hat{\boldsymbol{\Sigma}}$ ($\text{diag}(\hat{\boldsymbol{\Sigma}}_{UNALCE})=\text{diag}(\hat{\boldsymbol{\Sigma}}_{ALCE})$). The minimization of term $\mathbf{B}''$, given that $\text{rank}(\mathbf{L})=\hat{r}$, falls back into the previous case, i.e., $\mathbf{B}''$ is minimum if and only if $\hat{\mathbf{L}}=\hat{\mathbf{L}}_{UNALCE}=\hat{\mathbf{U}}_{UNALCE}(\hat{\mathbf{D}}_{UNALCE}+\breve{\psi}I_r)\hat{\mathbf{U}}'_{UNALCE}$.

Optimality holds over the Cartesian product of the set of all symmetric positive semi-definite matrices with a rank smaller or equal to r, $\hat{\mathcal{B}}(\hat{r})$, and the set of all symmetric sparse matrices with the same sparsity pattern as $\hat{\mathbf{S}}_{ALCE}$ such that $\text{diag}(\mathbf{S})=\text{diag}(\hat{\boldsymbol{\Sigma}}_{ALCE}-\mathbf{L})$, $\mathbf{L}\in\hat{\mathcal{B}}(\hat{r})$ (we call this set $\hat{\mathcal{A}}_{diag}(\hat{s})$).

Consequently, we can write:

$$\hat{\mathbf{S}}_{UNALCE,ii}=\hat{\boldsymbol{\Sigma}}_{ALCE,ii}-\hat{\mathbf{L}}_{UNALCE,ii},\ \hat{\mathbf{S}}_{UNALCE,ij}=\hat{\mathbf{S}}_{ALCE,ij},\ i\neq j.$$

Proof of Corollary 2

We know that $\|\hat{\mathbf{L}}_{UNALCE}-\hat{\mathbf{L}}_{ALCE}\|_2=\breve{\psi}$. We can prove that $\hat{\mathbf{L}}_{UNALCE}=\min_{\mathbf{L}\in\hat{\mathcal{B}}(\hat{r})}\|\mathbf{L}-\mathbf{L}^*\|_F^2$, conditioning on the event $\min_{\mathbf{L}\in\hat{\mathcal{B}}(\hat{r}),\mathbf{S}\in\hat{\mathcal{A}}(\hat{s}),\boldsymbol{\Sigma}=\mathbf{L}+\mathbf{S}}\|\boldsymbol{\Sigma}-\boldsymbol{\Sigma}_n\|_F^2$ under prescribed assumptions (see Theorem 3). In fact, we can write

$$\min_{\mathbf{L}\in\hat{\mathcal{B}}(\hat{r})}\|\mathbf{L}-\mathbf{L}^*\|_F^2\leq\min_{\mathbf{L}\in\hat{\mathcal{B}}(\hat{r})}\|\mathbf{L}-\mathbf{Y}_{pre}\|_F^2+\|\mathbf{Y}_{pre}-\mathbf{L}^*\|_F^2,$$

because $\mathbf{Y}_{pre}$ is uniquely determined by the conditioning event. The same inequality holds in the spectral norm.

Since

$$\|\hat{\mathbf{L}}_{ALCE}-\mathbf{L}^*\|_2\leq\|\hat{\mathbf{L}}_{UNALCE}-\hat{\mathbf{L}}_{ALCE}\|_2+\|\hat{\mathbf{L}}_{UNALCE}-\mathbf{L}^*\|_2,$$

it follows

$$0<\|\hat{\mathbf{L}}_{ALCE}-\mathbf{L}^*\|_2-\|\hat{\mathbf{L}}_{UNALCE}-\mathbf{L}^*\|_2\leq\breve{\psi}$$

given the conditioning event. Consequently, since $\|\hat{\mathbf{L}}_{UNALCE}-\hat{\mathbf{L}}_{ALCE}\|_F=tr(\hat{\mathbf{L}}_{UNALCE}-\hat{\mathbf{L}}_{ALCE})^2=r\breve{\psi}^2$, we obtain

$$0<\|\hat{\mathbf{L}}_{ALCE}-\mathbf{L}^*\|_F-\|\hat{\mathbf{L}}_{UNALCE}-\mathbf{L}^*\|_F\leq\sqrt{r}\breve{\psi}.$$

The analogous triangular inequality for the sparse component is

$$\|\hat{\mathbf{S}}_{ALCE}-\mathbf{S}^*\|_F^2\leq\|\hat{\mathbf{S}}_{UNALCE}-\hat{\mathbf{S}}_{ALCE}\|_F^2+\|\hat{\mathbf{S}}_{UNALCE}-\mathbf{S}^*\|_F^2.$$

In order to quantify $\|\hat{\mathbf{S}}_{UNALCE}-\hat{\mathbf{S}}_{ALCE}\|_F^2$, we need to study the behaviour of the term $\sum_{i=1}^{p}(\hat{\mathbf{L}}_{UNALCE,ii}-\hat{\mathbf{L}}_{ALCE,ii})^2$, which is less than or equal to $r\breve{\psi}^2$, because it is less than or equal to $tr(\hat{\mathbf{L}}_{UNALCE}-\hat{\mathbf{L}}_{ALCE})^2$.

Consequently, we have $\|\hat{\mathbf{S}}_{UNALCE}-\hat{\mathbf{S}}_{ALCE}\|_F\leq\sqrt{r}\breve{\psi}$. Analogously to $\hat{\mathbf{L}}_{UNALCE}$, we can prove that

$$\hat{\mathbf{S}}_{UNALCE}=\min_{\mathbf{S}\in\hat{\mathcal{A}}(\hat{s})}\|\mathbf{S}-\mathbf{S}^*\|_F^2,$$

conditioning on the event

$$\min_{\mathbf{L}\in\hat{\mathcal{B}}(\hat{r}),\mathbf{S}\in\hat{\mathcal{A}}(\hat{s}),\boldsymbol{\Sigma}=\mathbf{L}+\mathbf{S}}\|\boldsymbol{\Sigma}-\boldsymbol{\Sigma}_n\|_F^2$$

under prescribed assumptions (see Theorem 3). In fact, we can write

$$\min_{\mathbf{S}\in\hat{\mathcal{A}}_{diag}(\hat{s})}\|\mathbf{S}-\mathbf{S}^*\|_F^2\leq\min_{\mathbf{S}\in\hat{\mathcal{A}}_{diag}(\hat{s})}\|\mathbf{S}-\mathbf{Z}_{pre}\|_F^2+\|\mathbf{Z}_{pre}-\mathbf{S}^*\|_F^2,$$

because $\mathbf{Z}_{pre}$ is uniquely determined by the conditioning event.

Therefore, we can write

$$0<\|\hat{\mathbf{S}}_{ALCE}-\mathbf{S}^*\|_F-\|\hat{\mathbf{S}}_{UNALCE}-\mathbf{S}^*\|_F\leq\sqrt{r}\breve{\psi}.$$

The claim on $\|\hat{\mathbf{S}}_{UNALCE}-\mathbf{S}^*\|_2$ is less immediate. We recall that $\|\hat{\mathbf{L}}_{UNALCE}-\hat{\mathbf{L}}_{ALCE}\|_2=\|\hat{\mathbf{U}}\breve{\psi}\mathbf{I}_r\hat{\mathbf{U}}^{\top}\|_2=\breve{\psi}$. $\hat{\mathbf{U}}\breve{\psi}\mathbf{I}_r\hat{\mathbf{U}}^{\top}$ can be divided into the contribution coming from diagonal elements and the rest: $\|\text{diag}(\hat{\mathbf{L}}_{UNALCE}-\hat{\mathbf{L}}_{ALCE})+\text{off}-\text{diag}(\hat{\mathbf{L}}_{UNALCE}-\hat{\mathbf{L}}_{ALCE})\|_2$. Both contributions are part of $\hat{\mathbf{U}}\breve{\psi}\mathbf{I}_r\hat{\mathbf{U}}^{\top}$.

Given the matrix of eigenvectors $\hat{\mathbf{U}}$, we can write $\text{diag}(\hat{\mathbf{L}}_{UNALCE} - \hat{\mathbf{L}}_{ALCE}) = \sum_{i=1}^{p} \|\hat{\mathbf{u}}_i^\top\|^2 \mathbf{K}_{ii}$, where $\mathbf{K}_{ii}$ is a null matrix, except for the ith diagonal element equal to $\breve{\psi}$, and $\hat{\mathbf{u}}_i^\top$ is the ith row of $\hat{\mathbf{U}}$. Similarly, we can write $\text{off} - \text{diag}(\hat{\mathbf{L}}_{UNALCE} - \hat{\mathbf{L}}_{ALCE}) = \sum_{i=1}^{p} \sum_{j \neq i} \hat{\mathbf{u}}_i^\top \hat{\mathbf{u}}_j \mathbf{K}_{ij}$ where $\mathbf{K}_{ij}$ is a null matrix, except for the element ij equal to $\breve{\psi}$. Note that the rows of $\hat{\mathbf{U}}$, differently from the columns, are not orthogonal.

Since all summands are orthogonal to each other ($\mathbf{A} \perp \mathbf{B} \Leftrightarrow tr(\mathbf{A}\mathbf{B}^\top) = 0$), the triangular inequalities relative to $\|\text{diag}(\hat{\mathbf{L}}_{UNALCE} - \hat{\mathbf{L}}_{ALCE})\|_2$, $\|\text{off} - \text{diag}(\hat{\mathbf{L}}_{UNALCE} - \hat{\mathbf{L}}_{ALCE})\|_2$ and $\|\hat{\mathbf{L}}_{UNALCE} - \hat{\mathbf{L}}_{ALCE}\|_2$ become equalities. Therefore, we can write:

$$\|\text{diag}(\hat{\mathbf{L}}_{UNALCE} - \hat{\mathbf{L}}_{ALCE})\|_2 = \sum_{i=1}^{p} \|\hat{\mathbf{u}}_i^\top\|^2 \times \|\mathbf{K}_{ii}\| = \sum_{i=1}^{p} \|\hat{\mathbf{u}}_i^\top\|^2 \breve{\psi};$$

$$\|\text{off} - \text{diag}(\hat{\mathbf{L}}_{UNALCE} - \hat{\mathbf{L}}_{ALCE})\|_2 = \sum_{i=1}^{p} \sum_{j \neq i} \hat{\mathbf{u}}_i^\top \hat{\mathbf{u}}_j \|\mathbf{K}_{ij}\| = \sum_{i=1}^{p} \sum_{j \neq i} \hat{\mathbf{u}}_i^\top \hat{\mathbf{u}}_j \breve{\psi};$$

$$\|\hat{\mathbf{L}}_{UNALCE} - \hat{\mathbf{L}}_{ALCE}\|_2 = \sum_{i=1}^{p} \|\hat{\mathbf{u}}_i^\top\|^2 \times \|\mathbf{K}_{ii}\| + \sum_{i=1}^{p} \sum_{j \neq i} \hat{\mathbf{u}}_i^\top \hat{\mathbf{u}}_j \|\mathbf{K}_{ij}\| = \breve{\psi}.$$

From this consideration, it follows that

$$\|\text{diag}(\hat{\mathbf{L}}_{UNALCE} - \hat{\mathbf{L}}_{ALCE})\|_2 \leq \|\hat{\mathbf{L}}_{UNALCE} - \hat{\mathbf{L}}_{ALCE}\|_2 = \breve{\psi}.$$

Since, by definition, $\|\text{diag}(\hat{\mathbf{S}}_{UNALCE} - \hat{\mathbf{S}}_{ALCE})\|_2 = \|\text{diag}(\hat{\mathbf{L}}_{UNALCE} - \hat{\mathbf{L}}_{ALCE})\|_2$ (because $\text{diag}(\hat{\mathbf{S}}_{UNALCE} - \hat{\mathbf{S}}_{ALCE}) = -\text{diag}(\hat{\mathbf{L}}_{UNALCE} - \hat{\mathbf{L}}_{ALCE})$), and recalling that $\hat{\mathbf{S}}_{UNALCE}$ has the best approximation property (for Theorem 3) given the conditioning event, we can conclude

$$0 < \|\hat{\mathbf{S}}_{ALCE} - \mathbf{S}^*\|_2 - \|\hat{\mathbf{S}}_{UNALCE} - \mathbf{S}^*\|_2 \leq \breve{\psi}.$$

Proof of Corollary 3

The relevant triangular inequality for the overall estimate is

$$\|\boldsymbol{\Sigma}_n - \hat{\boldsymbol{\Sigma}}_{ALCE}\|_2 \leq \|\|\hat{\boldsymbol{\Sigma}}_{UNALCE} - \hat{\boldsymbol{\Sigma}}_{ALCE}\|_2 + \|\boldsymbol{\Sigma}_n - \hat{\boldsymbol{\Sigma}}_{UNALCE}\|_2.$$

By definition, $\|\hat{\boldsymbol{\Sigma}}_{UNALCE} - \hat{\boldsymbol{\Sigma}}_{ALCE}\|_2 = \|\text{off} - \text{diag}(\hat{\mathbf{L}}_{UNALCE} - \hat{\mathbf{L}}_{ALCE})\|_2$. For the same considerations explained before,

$$\|\text{off} - \text{diag}(\hat{\mathbf{L}}_{UNALCE} - \hat{\mathbf{L}}_{ALCE})\|_2 \leq \|\hat{\mathbf{L}}_{UNALCE} - \hat{\boldsymbol{\Sigma}}_{ALCE}\|_2 = \breve{\psi}.$$

Consequently, recalling that $\hat{\boldsymbol{\Sigma}}_{UNALCE} = \min_{\boldsymbol{\Sigma} = \mathbf{L} + \mathbf{S}, \mathbf{L} \in \hat{\mathcal{B}}(\hat{r}), \mathbf{S} \in \hat{\mathcal{A}}(\hat{s})} \|\boldsymbol{\Sigma} - \boldsymbol{\Sigma}_n\|_F^2$ under the described assumptions, it follows

$$0 < \|\boldsymbol{\Sigma}_n - \hat{\boldsymbol{\Sigma}}_{ALCE}\|_2 - \|\boldsymbol{\Sigma}_n - \hat{\boldsymbol{\Sigma}}_{UNALCE}\|_2 \leq \breve{\psi}. \tag{A.9}$$

Since $\|\hat{\mathbf{L}}_{UNALCE} - \hat{\mathbf{L}}_{ALCE}\|_F^2 = tr(\hat{\mathbf{L}}_{UNALCE} - \hat{\mathbf{L}}_{ALCE})^2 = r\breve{\psi}^2$, we have

$$0 < \|\text{off} - \text{diag}(\hat{\mathbf{L}}_{UNALCE} - \hat{\mathbf{L}}_{ALCE})\|_F \leq \sqrt{r}\breve{\psi}.$$

We can then claim

$$0 < \|\boldsymbol{\Sigma}_n - \hat{\boldsymbol{\Sigma}}_{ALCE}\|_F - \|\boldsymbol{\Sigma}_n - \hat{\boldsymbol{\Sigma}}_{UNALCE}\|_F \leq \sqrt{r}\breve{\psi}.$$

Therefore, the real gain in terms of the approximation of $\boldsymbol{\Sigma}_n$ with respect to ALCE measured in the squared Frobenius norm is strictly positive and bounded from $r\breve{\psi}^2$.

Proof of Theorem 4

Conditioning on $\boldsymbol{\Sigma}_n$, we can easily write

$$\|\hat{\boldsymbol{\Sigma}}_{UNALCE} - \boldsymbol{\Sigma}^*\|_2 = \|\hat{\boldsymbol{\Sigma}}_{UNALCE} - \boldsymbol{\Sigma}_n + \boldsymbol{\Sigma}_n - \boldsymbol{\Sigma}^*\|_2 \leq \|\hat{\boldsymbol{\Sigma}}_{UNALCE} - \boldsymbol{\Sigma}_n\|_2 + \|\boldsymbol{\Sigma}_n - \boldsymbol{\Sigma}^*\|_2. \tag{A.10}$$

The term $\|\boldsymbol{\Sigma}_n - \boldsymbol{\Sigma}^*\|_2$ only depends on the estimation input $\boldsymbol{\Sigma}_n$.

Therefore, by (A.9) and (A.10), it is straightforward that

$$0 < \|\hat{\boldsymbol{\Sigma}}_{ALCE} - \boldsymbol{\Sigma}^*\|_2 - \|\hat{\boldsymbol{\Sigma}}_{UNALCE} - \boldsymbol{\Sigma}^*\|_2 \leq \breve{\psi}.$$

Analogously, it is easy to prove that

$$0 < \|\hat{\boldsymbol{\Sigma}}_{ALCE} - \boldsymbol{\Sigma}^*\|_F - \|\hat{\boldsymbol{\Sigma}}_{UNALCE} - \boldsymbol{\Sigma}^*\|_F \leq \sqrt{r}\breve{\psi}. \tag{A.11}$$

Proof of Corollary 4

Let us recall the following expression:

$$\|(\hat{\mathbf{L}}+\hat{\mathbf{S}})^{-1}-(\boldsymbol{\Sigma}^{*-1})\|_F = \|(\hat{\mathbf{L}}+\hat{\mathbf{S}})^{-1}[\hat{\mathbf{L}}+\hat{\mathbf{S}}-\boldsymbol{\Sigma}^*](\boldsymbol{\Sigma}^{*-1})\|_F \leq \|(\hat{\mathbf{L}}+\hat{\mathbf{S}})^{-1}\|_2 \cdot \|\hat{\mathbf{L}}+\hat{\mathbf{S}}-\boldsymbol{\Sigma}^*\|_F \cdot \|\boldsymbol{\Sigma}^{*-1}\|_2.$$

From (A.11), we can conclude that

$$0 < \|(\hat{\mathbf{L}}_{ALCE}+\hat{\mathbf{S}}_{ALCE})^{-1}-\boldsymbol{\Sigma}^{*-1}\|_F - \|(\hat{\mathbf{L}}_{UNALCE}+\hat{\mathbf{S}}_{UNALCE})^{-1}-\boldsymbol{\Sigma}^{*-1}\|_F \leq \sqrt{r}\breve{\psi}.$$

Analogously, since it holds

$$\|(\hat{\mathbf{L}}+\hat{\mathbf{S}})^{-1}-(\boldsymbol{\Sigma}^{*-1})\|_2 = \|(\hat{\mathbf{L}}+\hat{\mathbf{S}})^{-1}[\hat{\mathbf{L}}+\hat{\mathbf{S}}-\boldsymbol{\Sigma}^*](\boldsymbol{\Sigma}^{*-1})\|_2 \leq \|(\hat{\mathbf{L}}+\hat{\mathbf{S}})^{-1}\|_2 \cdot \|\hat{\mathbf{L}}+\hat{\mathbf{S}}-\boldsymbol{\Sigma}^*\|_2 \cdot \|\boldsymbol{\Sigma}^{*-1}\|_2,$$

it is straightforward that

$$0 < \|(\hat{\mathbf{L}}_{ALCE}+\hat{\mathbf{S}}_{ALCE})^{-1}-\boldsymbol{\Sigma}^{*-1}\|_2 - \|(\hat{\mathbf{L}}_{UNALCE}+\hat{\mathbf{S}}_{UNALCE})^{-1}-\boldsymbol{\Sigma}^{*-1}\|_2 \leq \breve{\psi}.$$

Proof of Corollary 5

The three claims of the corollary are proved in sequence.

1. We start to note that $\hat{\mathbf{L}}_{UNALCE}$, $\hat{\mathbf{L}}_{ALCE}$, and $\hat{\mathbf{U}}_{ALCE}\breve{\psi}\mathbf{I}_r\hat{\mathbf{U}}_{ALCE}^\top$ are r-ranked. Let the respective spectral decompositions be:

   (a) $\hat{\mathbf{B}}_{UNALCE}\hat{\mathbf{B}}_{UNALCE}^\top$ with $\hat{\mathbf{B}}_{UNALCE} = \hat{\mathbf{U}}_{ALCE}\sqrt{\hat{\mathbf{D}}_{UNALCE}}$;
   (b) $\hat{\mathbf{B}}_{ALCE}\hat{\mathbf{B}}_{ALCE}^\top$ with $\hat{\mathbf{B}}_{ALCE} = \hat{\mathbf{U}}_{ALCE}\sqrt{\hat{\mathbf{D}}_{ALCE}}$;
   (c) $(\hat{\mathbf{U}}_{ALCE}\sqrt{\breve{\psi}})(\hat{\mathbf{U}}_{ALCE}\sqrt{\breve{\psi}})^\top$.

   Consequently, we note that

   $$\lambda_r(\hat{\mathbf{L}}_{UNALCE}) = \lambda_r(\hat{\mathbf{L}}_{ALCE}+\hat{\mathbf{U}}_{ALCE}\breve{\psi}\mathbf{I}_r\hat{\mathbf{U}}_{ALCE}^\top) =$$
   $$\lambda_r(\hat{\mathbf{U}}_{ALCE}\hat{\mathbf{D}}_{ALCE}\hat{\mathbf{U}}_{ALCE}+\hat{\mathbf{U}}_{ALCE}\breve{\psi}\mathbf{I}_r\hat{\mathbf{U}}_{ALCE}^\top) = \lambda_r(\hat{\mathbf{L}}_{ALCE})+\breve{\psi},$$

   which proves the claim on $\hat{\mathbf{L}}_{UNALCE}$.
2. By Lidskii dual inequality (see [32]), we note that

   $$\lambda_p(\hat{\mathbf{S}}_{UNALCE}) = \lambda_p(\hat{\mathbf{S}}_{ALCE}-\mathrm{diag}(\hat{\mathbf{U}}_{ALCE}\breve{\psi}\mathbf{I}_r\hat{\mathbf{U}}_{ALCE}^\top)) \geq \lambda_p(\hat{\mathbf{S}}_{ALCE})+\lambda_p(-\mathrm{diag}(\hat{\mathbf{U}}_{ALCE}\breve{\psi}\mathbf{I}_r\hat{\mathbf{U}}_{ALCE}^\top)).$$

   The matrix $-\mathrm{diag}(\hat{\mathbf{U}}_{ALCE}\breve{\psi}\mathbf{I}_r\hat{\mathbf{U}}_{ALCE}^\top)$ is a p-dimensional squared matrix having as ith element the quantity $-\|\mathbf{u}_i^\top\|^2\breve{\psi}$, where $\mathbf{u}_i^\top$, $i \in \{1,\ldots,p\}$, is the ith row of the matrix $\hat{\mathbf{U}}_{ALCE}$. Since $\mathrm{tr}(-\mathrm{diag}(\hat{\mathbf{U}}_{ALCE}\breve{\psi}\mathbf{I}_r\hat{\mathbf{U}}_{ALCE}^\top)) = \mathrm{tr}(-\hat{\mathbf{U}}_{ALCE}\breve{\psi}\mathbf{I}_r\hat{\mathbf{U}}_{ALCE}^\top) = -r\breve{\psi}$, it follows that $\lambda_p(\mathrm{diag}(\hat{\mathbf{U}}_{ALCE}\breve{\psi}\mathbf{I}_r\hat{\mathbf{U}}_{ALCE}^\top)) \leq r\breve{\psi}/p$, i.e.,

   $$-\frac{r}{p}\breve{\psi} \leq \lambda_p(-\mathrm{diag}(\hat{\mathbf{U}}_{ALCE}\breve{\psi}\mathbf{I}_r\hat{\mathbf{U}}_{ALCE}^\top)) \leq 0.$$

   Therefore, we obtain

   $$\lambda_p(\hat{\mathbf{S}}_{UNALCE}) \geq \lambda_p(\hat{\mathbf{S}}_{ALCE})-\frac{r}{p}\breve{\psi},$$

   which proves the claim on $\hat{\mathbf{S}}_{UNALCE}$.
3. By Lidskii dual inequality, we note that

   $$\lambda_p(\hat{\boldsymbol{\Sigma}}_{UNALCE}) = \lambda_p(\hat{\boldsymbol{\Sigma}}_{ALCE}+\hat{\mathbf{U}}_{ALCE}\breve{\psi}\mathbf{I}_r\hat{\mathbf{U}}_{ALCE}^\top-\mathrm{diag}(\hat{\mathbf{U}}_{ALCE}\breve{\psi}\mathbf{I}_r\hat{\mathbf{U}}_{ALCE}^\top))$$
   $$\geq \lambda_p(\hat{\boldsymbol{\Sigma}}_{ALCE})+\lambda_p(\hat{\mathbf{U}}_{ALCE}\breve{\psi}\mathbf{I}_r\hat{\mathbf{U}}_{ALCE}^\top)-\lambda_p(\mathrm{diag}(\hat{\mathbf{U}}_{ALCE}\breve{\psi}\mathbf{I}_r\hat{\mathbf{U}}_{ALCE}^\top)).$$

   Recalling the argument above and noting that $\lambda_p(\hat{\mathbf{U}}_{ALCE}\breve{\psi}\mathbf{I}_r\hat{\mathbf{U}}_{ALCE}^\top) = 0$ because $\mathrm{rank}(\hat{\mathbf{U}}_{ALCE}\breve{\psi}\mathbf{I}_r\hat{\mathbf{U}}_{ALCE}^\top) = \hat{r}$, it follows

   $$\lambda_p(\hat{\boldsymbol{\Sigma}}_{UNALCE}) \geq \lambda_p(\hat{\boldsymbol{\Sigma}}_{ALCE})+0-\breve{\psi} = \lambda_p(\hat{\boldsymbol{\Sigma}}_{ALCE})-\frac{r}{p}\breve{\psi},$$

   which proves the claim on $\hat{\boldsymbol{\Sigma}}_{UNALCE}$.

Appendix B. Supplementary data

This paper is complemented by a supplement containing a discussion of LOREC assumptions and a simulation study. In addition, the MATLAB functions `UNALCE.m` and `POET.m`, performing UNALCE and POET procedures, respectively, can be downloaded at [16]. Both functions contain the detailed explanation of input and output arguments. Finally, the

MATLAB dataset `supervisory_data.mat`, which contains the covariance matrix, `C`, and the relative labels of supervisory indicators, `Labgood`, can also be downloaded at the same link, which we refer to for the details.

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jmva.2019.104577.

References

[1] A. Agarwal, S. Negahban, M.J. Wainwright, Noisy matrix decomposition via convex relaxation: Optimal rates in high dimensions, Ann. Statist. 40 (2) (2012) 1171–1197.
[2] T. Anderson, Multivariate Statistical Analysis, Wiley and Sons, New York, NY, 1984.
[3] J. Bai, Inferential theory for factor models of large dimensions, Econometrica 71 (1) (2003) 135–171.
[4] J. Bai, S. Ng, Determining the number of factors in approximate factor models, Econometrica 70 (1) (2002) 191–221.
[5] P.J. Bickel, E. Levina, Covariance regularization by thresholding, Ann. Statist. 36 (6) (2008) 2577–2604.
[6] P.J. Bickel, E. Levina, Regularized estimation of large covariance matrices, Ann. Statist. 36 (1) (2008) 199–227.
[7] J.-F. Cai, E.J. Candès, Z. Shen, A singular value thresholding algorithm for matrix completion, SIAM J. Optim. 20 (4) (2010) 1956–1982.
[8] T. Cai, W. Liu, Adaptive thresholding for sparse covariance matrix estimation, J. Amer. Statist. Assoc. 106 (494) (2011) 672–684.
[9] T. Cai, C.-H. Zhang, H.H. Zhou, Optimal rates of convergence for covariance matrix estimation, Ann. Statist. 38 (4) (2010) 2118–2144.
[10] V. Chandrasekaran, P.A. Parrilo, A.S. Willsky, Latent variable graphical model selection via convex optimization, Ann. Statist. 40 (4) (2012) 1935–1967.
[11] V. Chandrasekaran, S. Sanghavi, P.A. Parrilo, A.S. Willsky, Rank-sparsity incoherence for matrix decomposition, SIAM J. Optim. 21 (2) (2011) 572–596.
[12] I. Daubechies, M. Defrise, C. De Mol, An iterative thresholding algorithm for linear inverse problems with a sparsity constraint, Commun. Pure Appl. Math. 57 (11) (2004) 1413–1457.
[13] D.K. Dey, C. Srinivasan, Estimation of a covariance matrix under Stein's loss, Ann. Statist. 13 (4) (1985) 1581–1591.
[14] J. Fan, Y. Liao, H. Liu, An overview of the estimation of large covariance and precision matrices, 2016.
[15] J. Fan, Y. Liao, M. Mincheva, Large covariance estimation by thresholding principal orthogonal complements, J. R. Stat. Soc. Ser. B Stat. Methodol. 75 (4) (2013) 603–680.
[16] M. Farné, A. Montanari, A large covariance matrix estimator under intermediate spikiness regimes, 2018, http://dx.doi.org/10.17632/nh97vfvhkt, https://data.mendeley.com/datasets/nh97vfvhkt.
[17] M. Fazel, Matrix Rank Minimization with Applications (Ph.D. thesis), Stanford University, 2002.
[18] M. Fazel, H. Hindi, S.P. Boyd, A rank minimization heuristic with application to minimum order system approximation, in: American Control Conference, 2001. Proceedings of the 2001, Vol. 6, IEEE 2001, pp. 4734–4739.
[19] T.S. Ferguson, A Course in Large Sample Theory, Routledge, 2017.
[20] J. Friedman, T. Hastie, R. Tibshirani, Sparse inverse covariance estimation with the graphical lasso, Biostatistics 9 (3) (2008) 432–441.
[21] R. Furrer, T. Bengtsson, Estimation of high-dimensional prior and posterior covariance matrices in Kalman filter variants, J. Multivariate Anal. 98 (2) (2007) 227–255.
[22] D.A. Harville, Maximum likelihood approaches to variance component estimation and to related problems, J. Amer. Statist. Assoc. 72 (358) (1977) 320–338.
[23] C. Lam, Nonparametric eigenvalue-regularized precision or covariance matrix estimator, Ann. Statist. 44 (3) (2016) 928–953.
[24] O. Ledoit, M. Wolf, A well-conditioned estimator for large-dimensional covariance matrices, J. Multivariate Anal. 88 (2) (2004) 365–411.
[25] O. Ledoit, M. Wolf, Spectrum estimation: A unified framework for covariance matrix estimation and PCA in large dimensions, J. Multivariate Anal. 139 (2015) 360–384.
[26] X. Luo, High dimensional low rank and sparse covariance matrix estimation via convex minimization, 2011, Arxiv preprint arxiv:1111.1133v1.
[27] X. Luo, Recovering model structures from large low rank and sparse covariance matrix estimation, 2011, Arxiv preprint arxiv:1111.1133v2.
[28] Y. Nesterov, Gradient methods for minimizing composite functions, Math. Program. 140 (1) (2013) 125–161.
[29] H. Patterson, R. Thompson, Recovery of inter-block information when block sizes are unequal, Biometrika 58 (3) (1971) 545–554.
[30] Y. Qiu, S.X. Chen, Bandwidth selection for high-dimensional covariance matrix estimation, J. Amer. Statist. Assoc. 110 (511) (2015) 1160–1174.
[31] A.J. Rothman, E. Levina, J. Zhu, Generalized thresholding of large covariance matrices, J. Amer. Statist. Assoc. 104 (485) (2009) 177–186.
[32] T. Tao, Topics in Random Matrix Theory, Vol. 132, American Mathematical Society, 2012.

Noname manuscript No.
(will be inserted by the editor)

High-dimensional clustering via Random Projections

Laura Anderlucci · Francesca Fortunato · Angela Montanari

Abstract In this work, we address the unsupervised classification issue by exploiting the general idea of RP ensemble. Specifically, we propose to generate a set of low dimensional independent random projections and to perform model-based clustering on each of them. The top B^* projections, i.e. the projections which show the best grouping structure are then retained. The final partition is obtained by aggregating the clusters found in the projections via consensus. The performances of the method are assessed on both real and simulated datasets. The obtained results suggest that the proposal represents a promising tool for high-dimensional clustering.

Keywords High-dimensional clustering; random projections; model-based clustering

1 Introduction

Data clustering plays a key role in modern statistics as it represents one of the most effective tools to understand the underlying structure of a given data set. The aim of clustering is essentially to categorize data into 'clusters' (or groups) such that observations belonging to the same cluster are more similar to each others than those in different groups. This problem has been studied extensively and the state-of-the-art is exposed in surveys that have appeared regularly over the years; see, for example, McLachlan et al. (2019), Bouveyron and Brunet-Saumard (2014), Maugis et al. (2009b), Xu and Tian (2015).

Clustering in low-dimensional spaces requires limited resources; the complexity of the problem indeed increases with the number of observed features, p. When dealing with high-dimensional data, the use of traditional unsupervised classification algorithms faces several limitations; in particular, the presence of noisy or irrelevant information can mislead these methods due to the '*curse of dimensionality*', as coined by Bellman (1957). In order to overcome this problem, often dimension reduction procedures are applied before carrying out any clustering.

Generally, the term 'dimension reduction' refers to two different approaches; namely, it includes both *feature selection* methods that embed the high-dimensional points into a lower subspace by selecting some 'relevant' variables, and *feature extraction* algorithms

Department of Statistical Sciences - University of Bologna
via delle Belle Arti 41, 40126 Bologna (Italy)

which find an embedding by constructing new artificial features that are, for example, linear combinations of the original ones. Variable selection strategies have been frequently used to handle high-dimensional clustering issues, but feature extraction procedures could be generally more efficient. Feature selection techniques indeed may discard some potentially important variables, e.g. variables that are not predictive if individually considered, but that could provide significant benefits when taken in conjunction with other features.

Traditionally, variable combination methods involve the projection of high-dimensional data onto a lower subspace with the intent of capturing as much of the data variability as possible (e.g. Principal Component Analysis). Albeit this approach has been successfully used in many applications, its aim does not always coincide with that of a clustering task. In fact, the useful information about the group structure is not necessarily contained in the subspaces with the largest variance, as exposed by Chang (1983). A recent approach for dimension reduction that has been gaining increasing attention is based on Random Projections (RPs) and consists in mapping at random the original high-dimensional data onto a lower subspace by using a random matrix with orthogonal columns of unit length. Specifically, the key point of RP is that, regardless of the original data dimension, the final solution still preserves the global information almost perfectly. Such a result is guaranteed by the Johnson and Lindenstrauss (1984) Lemma, which states that any n-point set in p dimensions ($X = [\mathbf{x}_1, \ldots, \mathbf{x}_i, \ldots, \mathbf{x}_n]^T$, $\mathbf{x}_i \in \mathbb{R}^p \ \forall i = 1, \ldots, n$) can be linearly projected onto $d = O(\log(n)/\varepsilon^2)$ coordinates (with $d \ll p$) , by using a random matrix $A \in \mathbb{R}^{p \times d}$ with orthonormal columns, while preserving pairwise distances within a factor $1 \pm \varepsilon$. More precisely, with high probability over the randomness of A:

$$(1-\varepsilon)||\mathbf{x}_i - \mathbf{x}_j||_2 \leq ||A^\top \mathbf{x}_i - A^\top \mathbf{x}_j||_2 \leq (1+\varepsilon)||\mathbf{x}_i - \mathbf{x}_j||_2, \tag{1}$$

where $||\cdot||_2$ indicates the L_2 norm.

Bhattacharya et al. (2009) proved that also the Hellinger distance between any two distributions P and Q, defined as

$$H(P,Q) = \frac{1}{2}\left(||\sqrt{P} - \sqrt{Q}||_2\right)^2,$$

admits a low disorsion JL-type embedding. In the model-based clustering context, where data are considered as coming from a distribution that is a mixture of two or more components, this theorem directly implies that the distance between the density of any pair of components is preserved with arbitrarily small distortion. In other words, it states that if two component densities are sufficiently far apart in the high-dimensional space, then they would be expected approximately the same also in the reduced d-dimensional space.

This interesting result motivated us to employ random projections within a model-based clustering framework. Specifically, inspired by the original idea of Cannings and Samworth (2017) for supervised classification, we propose to generate a set of B low dimensional independent random projections and to apply a Gaussian Mixture Model (GMM) on each of them. Our Random Projection Ensemble Clustering (RPE Clu) algorithm then obtains the final partition by combining via consensus the clustering results from the top B^* projections, i.e. the projections which show the best grouping structure according to a given criterion.

The paper is organized as follows. Section 2 recalls the model-based clustering framework. In the same section, some popular dimension reduction procedures for high-dimensional clustering are briefly presented. In Section 3, the Random Projection Ensemble Clustering algorithm (RPE Clu) is introduced and defined in detail. Section 4 is devoted to practical

considerations about the computational complexity of the algorithm, the choice of the number of random projections and the dimension of the projected space. Section 5 presents a simulation study where the proposed methodology is compared with some benchmark clustering techniques. In Section 6, RPE Clu is applied to two sets of high-dimensional real data. A final discussion on the obtained results concludes the paper.

2 High-dimensional model-based clustering

In model-based clustering (see McLachlan and Peel (2004) for a detailed review), data are assumed to derive from a common source with G different sub-populations. In particular, each sub-population is modelled separately (typically by members of the same parametric density family) and the overall population is but a mixture of them. The resulting model is a finite mixture and it is described by the following probability density function (pdf):

$$f(\mathbf{x}) = \sum_{k=1}^{G} \pi_k f_k(\mathbf{x}|\theta_k).$$

Here, f_k and θ_k are the density and the parameters of the k-th component of the mixture, respectively, whereas π_k is the prior probability that an observation belongs to the k-th component ($\pi_k \geq 0$, $\sum_{k=1}^{G} \pi_k = 1$). For clustering purposes, units are allocated to the component whose posterior probability is maximum.

A common choice for $f_k(\cdot)$ is the multivariate normal distribution, $\phi_k(\cdot)$, parameterized by its mean μ_k and its covariance matrix Σ_k:

$$\phi_k(\mathbf{x}|\mu_k, \Sigma_k) = (2\pi)^{-(p/2)}|\Sigma_k|^{-(1/2)} \exp\left\{-\frac{1}{2}(\mathbf{x}-\mu_k)'\Sigma_k^{-1}(\mathbf{x}-\mu_k)\right\}.$$

Following this approach, the entire data set is modeled by a Gaussian Mixture model:

$$f(\mathbf{x}) = \sum_{k=1}^{G} \pi_k \phi_k(\mathbf{x}|\mu_k, \Sigma_k).$$

In presence of high-dimensional data the GMM tends to perform poorly, due to the large number of parameters to estimate with relatively few observations. In fact, the number of parameters increases quadratically with p and thus the maximum-likelihood estimation problem becomes ill-posed very quickly. The earliest approaches which appeared in the literature to overcome this limit and attain parsimony propose alternative parameterizations of the component densities. For instance, Banfield and Raftery (1993) and Celeux and Govaert (1995) introduce a parsimonious parameterizations of the covariance matrix in terms of its eigenvalue decomposition so as to control the volume, shape and orientation of the Gaussian ellipsoids. Biernacki and Lourme (2014) define different parsimonious models based on a variance-correlation decomposition of the covariance matrices.

When performing variable selection for clustering, the aim is essentially to identify those features that bring relevant information about the underlying group structure. In the model-based context, the definition of 'relevance' should be expressed in terms of probabilistic dependence (or independence) with respect to $\mathbf{z}$, i.e. the variable which describes the latent class membership. Specifically, the distribution of relevant variables directly depends on $\mathbf{z}$ as these features contain the key clustering information. Conversely, both redundant and uninformative variables do not provide any additional or useful information and, thus, they

can be assumed to be conditionally independent given the relevant variables or completely independent of the group structure, respectively. Following this approach, several authors have recast the variable selection problem for clustering in a model selection one. Namely, relevant variables are sought through a stepwise procedure that, at each step, compares models that differ in the role assigned to the variables in explaining the clustering structure.

Pioneers of this framework were Raftery and Dean (2006), who introduced a procedure in which the decision for inclusion or exclusion of a generic (set of) variable(s) $\mathbf{x}^P$ into the current set of clustering ones $\mathbf{x}^C$ is taken by comparing two competing models in terms of their Bayesian Information Criterion (BIC). In particular, Model I assumes that $\mathbf{x}^P$ carries relevant information about the cluster membership, whereas Model II states that $\mathbf{x}^P$ does not depend on $\mathbf{z}$. The BIC associated to these models are:

$$\begin{aligned} \text{BIC}_I &= \text{BIC}_{\text{clust}}(\mathbf{x}^C, \mathbf{x}^P) \\ \text{BIC}_{II} &= \text{BIC}_{\text{no clust}}(\mathbf{x}^C) + \text{BIC}_{\text{reg}}(\mathbf{x}^P|\mathbf{x}^C) \end{aligned} \tag{2}$$

Here, $\text{BIC}_{\text{clust}}(\mathbf{x}^C, \mathbf{x}^P)$ is the BIC of the GMM in which $\mathbf{x}^P$ adds useful information, $\text{BIC}_{\text{no clust}}(\mathbf{x}^C)$ is the BIC of the GMM on the current set of clustering variables only and $\text{BIC}_{\text{reg}}(\mathbf{x}^P|\mathbf{x}^C)$ is the BIC of the regression of $\mathbf{x}^P$ on $\mathbf{x}^C$. If $\text{BIC}_I - \text{BIC}_{II} > 0$, then $\mathbf{x}^P$ is added to the set of clustering variables $\mathbf{x}^C$.

This method has been further improved by Maugis et al. (2009a) and Maugis et al. (2009b) under the assumption that the irrelevant variables can be independent of some relevant ones.

Recently, two further extensions of the above modeling appeared in the literature: Scrucca (2016) suggests to overcome the sub-optimality of a stepwise model search by employing genetic algorithms; Galimberti et al. (2018) take into account the possibility that different variable vectors provide information about different clustering structures.

Although effective in many applications, in the unsupervised classification context the variable selection problem is ill-posed: clusters indeed strongly depend on the selected features and the features are selected according to the clusters (see Ruiz et al. (2009)). For this reason, feature extraction procedures would rather be preferred.

Dasgupta (2013) demonstrated that RPs can be successfully used to handle high-dimensional clustering issues with a model-based approach. Firstly, he showed that a mixture of G Gaussians can be embedded onto just $O(\log G)$ random coordinates without destroying the original group structure. Second, he proved that even if the original Gaussians exhibit eccentric elliptical contours, their projected counterparts are always more spherical. These two benefits are of major importance and they definitely facilitate the learning of a Gaussian Mixture Model. In particular, dimension reduction saves a lot of time and computational costs on one hand; on the other, clusters of low eccentricity reduce the EM algorithmic challenges ensuring that intermediate covariance matrices are not singular or close to singular.

3 Random projection ensemble clustering

As discussed in the previous section, high-dimensional data pose many challenges to model-based clustering. Methods in this class indeed become rapidly over-parameterized since the number of parameters to estimate increases quadratically with the number of observed features p.

Random projections have shown to provide promising results for the analysis of high-dimensional data. Their main inconvenience is that they are highly unstable: namely, different random projections of the original data may provide completely different classification results. That is the reason why most of the successful proposals on RPs resorts to ensembles. For example, Fern and Brodley (2003) propose to aggregate the clustering results of a GMM on different random projections of the data into a similarity matrix containing the probability "estimates" that any two data points belong to the same cluster; then, they suggest to perform an agglomerative clustering procedure on such a matrix to produce the final groups.

In this paper, we also exploit the general idea of RP ensemble for high-dimensional clustering. In particular, our novel proposal consists of applying a Gaussian Mixture Model to *carefully chosen* random projections of the original data, but differently from Fern and Broadley, we use the GMM properties for both projection selection and consensus aggregation.

3.1 On the choice of random projections

Differently from other transformation techniques (such as, for example, principal components or projection pursuit), the random projection method does not exploit any 'interestingness' criterion to identify the 'optimal' projection. High-dimensional data are just embedded into a lower dimensional subspace by using a random projection matrix A with orthogonal and unit length columns. As a consequence of that, results from distinct configurations of the same data can be even dramatically different: some projections indeed can highlight a clear group structure in the lowered data, whilst some others can derail any hope of learning by confusing all the groups together.

In this section, we propose a method for choosing 'good' random projections, that is, a criterion for identifying those projections that give a partition of the data close to the underlying group structure.

Hennig (2019) provides a detailed review of the validation indexes proposed in the literature to evaluate the quality of a clustering procedure. Although effective, many of these indexes rely on a measure of distance/dissimilarity and, therefore, they may seem inconsistent with a model-based framework. Furthermore, since in the unsupervised context no *apriori* information about the structure being looked for is available, we believe it makes sense to consider the RP selection as a part of the clustering algorithm, i.e. as the choice of the model that best fits the data according to a specific criterion (e.g. the BIC).

The BIC of mixture models fitted to different random projections cannot in principle be compared, because they are referred to different variables generated by the different random projections. On the contrary, the BIC of different models defined in the original variable space can be compared. We search for the solution that maximizes the log-likelihood of the GMM fitted on the original data, penalized by the number of free parameters.

In practice, in order to avoid the drawbacks associated with the high-dimensional spaces, a feasible solution consists in considering the following variable partition

$$Y^* = [Y, \bar{Y}] = [XA|X\bar{A}],$$

where $X \in \mathbb{R}^{n\times p}$ is the original high-dimensional data matrix, $A \in \mathbb{R}^{p\times d}$ is the random projection matrix and $\bar{A} \in \mathbb{R}^{p\times(p-d)}$ is its orthogonal complement. The basic idea is to perform model-based clustering on the reduced data $Y = XA$, assuming that the underlying

group structure may be well approximated by the one in the d dimensions of the block matrix Y^*, i.e.:

$$f(Y|\mathbf{z}) = \sum_{k=1}^{G} \pi_k \phi_k(Y|\mu_{Y_k}, \Sigma_{Y_k}, \mathbf{z}). \tag{3}$$

This assumption does not imply that $\bar{Y}$ is not useful for clustering, but only that it contains some information on the group membership $\mathbf{z}$ that is very similar to that already available in Y. Therefore, in terms of distributional representation, it seems reasonable to think of $\bar{Y}$ as conditionally independent of $\mathbf{z}$ given Y; it could be necessary for the clustering, but only if Y is not present (Fop and Murphy 2018). This amounts to assume that:

$$f(\bar{Y}|Y) = \phi(\bar{Y}|\mu_{\bar{Y}|Y}, \Sigma_{\bar{Y}|Y}, Y) \tag{4}$$

where

$$\begin{aligned} \mu_{\bar{Y}|Y} &= \mu_{\bar{Y}} + \Sigma_{\bar{Y}Y}\Sigma_Y^{-1}(Y - \mu_Y) \\ \Sigma_{\bar{Y}|Y} &= \Sigma_{\bar{Y}} - \Sigma_{\bar{Y}Y}\Sigma_Y^{-1}\Sigma_{Y\bar{Y}}. \end{aligned} \tag{5}$$

Equation (5) describes the Schur complemement of the block Σ_Y in the $p \times p$ block-matrix

$$\Sigma_{Y^*} = \begin{bmatrix} \Sigma_Y & \Sigma_{Y\bar{Y}} \\ \Sigma_{\bar{Y}Y} & \Sigma_{\bar{Y}} \end{bmatrix}.$$

The distribution of Y^* is the product of the marginal density of Y, $f(Y|\mathbf{z})$, and the conditional density of $\bar{Y}|Y$, $f(\bar{Y}|Y)$:

$$f(Y^*|\mathbf{z}) = \left[\sum_{k=1}^{G} \pi_k \phi_k(Y|\mu_{Y_k}, \Sigma_{Y_k}, \mathbf{z})\right] \phi(\bar{Y}|\mu_{\bar{Y}|Y}, \Sigma_{\bar{Y}|Y}, Y) = \sum_{k=1}^{G} \pi_k \phi_k(Y^*|\mu_{Y_k^*}, \Sigma_{Y_k^*}, \mathbf{z}). \tag{6}$$

Equation (6) can be easily rewritten in terms of log-likelihood as:

$$\sum_{i=1}^{n} \log[f(y_i^*|\mathbf{z})] = \sum_{i=1}^{n} \log[f(y_i)|\mathbf{z_i}] + \sum_{i=1}^{n} \log[f(\bar{y}_i|y_i)]. \tag{7}$$

The BIC corresponding to Equation (7) is:

$$\text{BIC} = \text{BIC}_{\text{GMM}}(Y) + \text{BIC}_{\text{reg}}(\bar{Y}|Y), \tag{8}$$

where $\text{BIC}_{\text{GMM}}(Y) = 2\log[f(Y)] - q_Y \log(n)$ is the BIC associated to the Gaussian mixture fitted on the d-dimensional data and $\text{BIC}_{\text{reg}}(\bar{Y}|Y) = 2\log[f(\bar{Y}|Y)] - q_{\bar{Y}} \log(n)$ is the BIC for the linear regression of the $(p-d)$ last columns of Y^* on the first d ones. The number of free parameters of the GMM on Y and those of the linear regression are described by q_Y and $q_{\bar{Y}}$, respectively. In order to allow for great flexibility, $\Sigma_{\bar{Y}|Y}$ is assumed to have a general form and, thus,

$$q_{\bar{Y}} = (p-d)(d+1) + \frac{(p-d)[(p-d)+1)]}{2}.$$

When the number of observed features p is particularly large with respect to d, a restricted form for $\Sigma_{\bar{Y}|Y}$ is suggested. Namely, $\Sigma_{\bar{Y}|Y} = \text{diag}(\sigma_1^2, \ldots, \sigma_{p-d}^2)$. In this case, the number of free parameters for the regression model reduces to $q_{\bar{Y}} = (p-d)(d+1) + (p-d)$.

As depicted in Figure 1, the criterion we propose seems capable to correctly rank the random projections according to the goodness of the partition they induce. Specifically, the largest values for the Adjusted Rand Index (ARI), i.e. a measure of the similarity between the classification yielded by the GMM on the reduced data and the true class membership, correspond to models with the largest BIC.

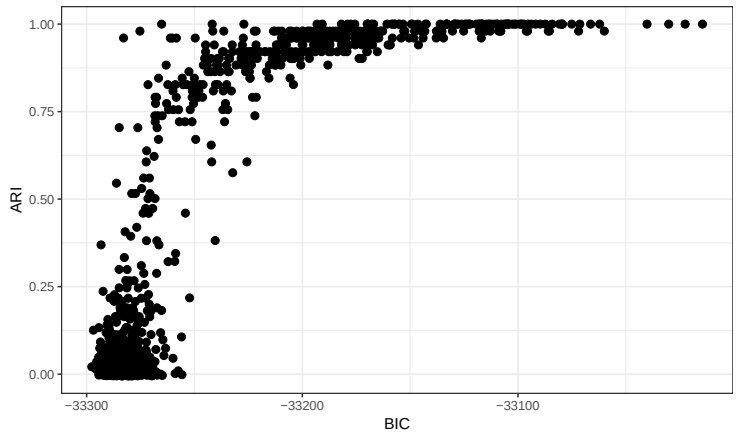


Fig. 1: ARI of the classification yielded by the GMM on $B = 1000$ different 8-dimensional projections of a simulated dataset with $p = 100$, $G = 2$, $n_1 = n_2 = 100$ and the true class membership, ordered by increasing values of the BIC.

3.2 On the result aggregation

A possible solution to the inherent instability associated with random projections involves the use of cluster ensembles that combine multiple individual partitions into a single consensus one. This process was pioneered by Strehl and Ghosh (2002) who proved that ensembles can provide robust and stable solutions across different problem domains. A detailed review of the state-of-the-art cluster ensemble methods can be found in Boongoen and Iam-On (2018), where both theoretical aspects and empirical applications are widely discussed.

Consensus clustering algorithms generally derive the ultimate data partition by minimizing an objective function that measures how dissimilar each hard or soft consensus candidate $C \in \mathbf{C}$ is from the ensemble members. Among the various consensus functions that have been developed in the literature, the most popular ones are of the form

$$L(C) = \sum_b w_b d(C_b, C)^l, \tag{9}$$

where w_b is the weight given to element C_b of the ensemble, $d(\cdot)$ is a suitable dissimilarity measure (e.g. Euclidean, Manhattan, ...) and $l \geq 1$.

In this work, we suggest to derive the final unit allocation by using the greedy algorithm introduced by Hornik (2005). This iterative procedure determines, at each step b, the locally optimal permutation matrix Π_b for relabeling by minimizing the Euclidean distance between the previously determined consensus candidate C^*_{b-1} (note that at the initial step $C_0 \equiv C_1$) and all the possible permutations of the membership vector C_b. Then, it derives the new consensus partition by taking the weighted average of C^*_{b-1} and $C_b\Pi_b$. In so doing, this sequential method helps to tackle the issue of simultaneous combination of all partitions, otherwise computationally unfeasible.

3.3 Random projection ensemble clustering algorithm

In this paper, a new model-based clustering method for high-dimensional data based on random projections, is introduced. The final partition is obtained through the following steps:

(a) Generate B independent d-dimensional random projections A_b, $b = 1,\dots,B$, according to a specific measure, e.g. the Haar measure;
(b) Compute the BIC as described in Equation 8 for the partition C_b induced by the GMM fitted on the projected data $Y = XA$ and by the linear regression of $\bar{Y}$ on Y;
(c) Among the B possible solutions, select the B^* projections that exhibit the highest values for the BIC: $A = [A_1;A_2;\dots;A_{B^*}]$;
(d) Aggregate the cluster membership vector of the best B^* projections via consensus.

4 Practical considerations

4.1 Computational complexity

The algorithm we propose derives the final partition by aggregating the results of Gaussian Mixture Model clustering performed on an 'optimal' subset of random projections.

The first step of this procedure involves the computation of B random projection matrices. The cost of this operation varies according to the method used: namely, generating a single RP from the Haar measure requires $O(pd^2)$ operations, whilst choosing each entry of this matrix uniformly and independently from $[-1,1]$ takes time only $O(dp)$ (see Achlioptas 2003).

Once the projections have been generated, the original high-dimensional data should be embedded onto the lowered spaces; each projection requires $O(npd)$ operations.

Then, for $b = 1,\dots,B$, a GMM is performed on the reduced set $Y = XA_b$ with a total cost of $O(Gd^3) \cong O(d^3)$. Simultaneously, a multiple linear regression of $\bar{Y} = X\bar{A}_b$ on Y is computed. The cost of this step is $O((p-d)^3)$. Finally, the BIC values computed as in Equation (8) are sorted and observations are clustered by using the best B^* projections (i.e. those yielding the highest values for the BIC). These steps involve $O(B^*)$[1] and $O(B^*nd)$ resources, respectively.

4.2 Choice of B and B^*

The random projection ensemble clustering performances strongly depend on the possibility to identify those random projections that induce a very clear group structure in the reduced space.

The choice of B^*, i.e. the number of 'base' models to retain in the final ensemble, is more insidious. Several studies have shown that ensembles of classifiers are generally more effective when they are constructed from members whose errors are dissimilar; see, for example, Kittler et al. (1998). In fact, aggregating the base results of models that agree on how a dataset should be partitioned does not provide any improvement. The random projection method itself represents a valid technique to introduce artificial instability (and thus *diversity*) to an ensemble as it allows to generate clustering results from different perturbed configurations of the original data. However, as Fern and Brodley (2003) point

[1] See the R Documentation for the `sort` function with default settings.

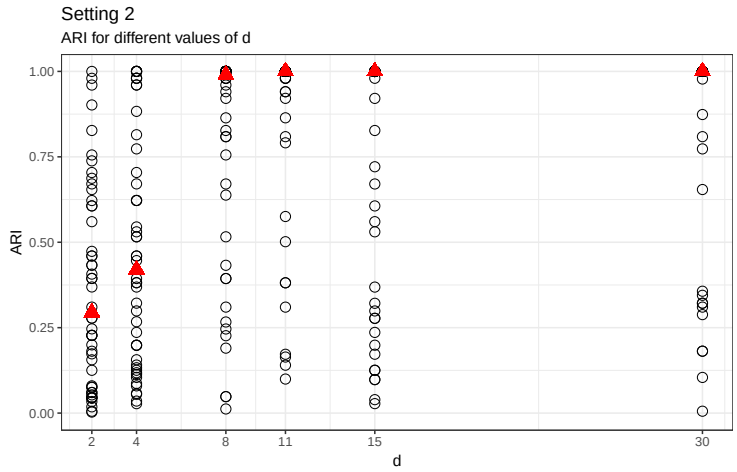


Fig. 2: Setting 2. ARI of the random projection ensemble clustering algorithm for different dimensions of the projected space, d. The numbers of generated and selected projections are set equal to $B = 1000$ and $B^* = 100$, respectively. Empty points indicate single performances, whereas the full red triangle refers to the median ARI estimated over 50 replicates of the experiment.

out, taking into account too many projections may degrade the final result, especially when the original features are highly correlated; furthermore, it surely increases the computational cost of the procedure. On the other side, considering a very small ensemble can be risky, too. In fact, since in clustering no *apriori* knowledge of the true data structure is available, identifying the best predictors is not a trivial task and, therefore, any criterion (including the BIC we propose) could be confused. In order to avoid the selection of too similar or inaccurate base classifiers, a compromise solution for B^* is highly suggested.

On the basis of the numerical evidences we suggest $B = 1000$ and $B^* = 100$ as generally good choices.

4.3 Choice of d

Dasgupta (2013) proved that data from an arbitrary mixture of G Gaussian distributions can be randomly embedded into a subspace of just $O(\log G)$ dimensions, while preserving the group structure almost perfectly. Furthermore, if $d < \log G$, the worsening of the mapping performance is gradual. This result is particularly appealing as it proves that the projected dimension is independent of the original dimensions of the data, that is, d does not depend upon n nor p. A couple of numerical experiments conducted on our datasets corroborate Dasgupta's result: in fact, Figures 2 and 3 clearly show that a choice of $d = O(10 \log G)$ works pretty well; higher values of d do dot noticeably improve the final performance.

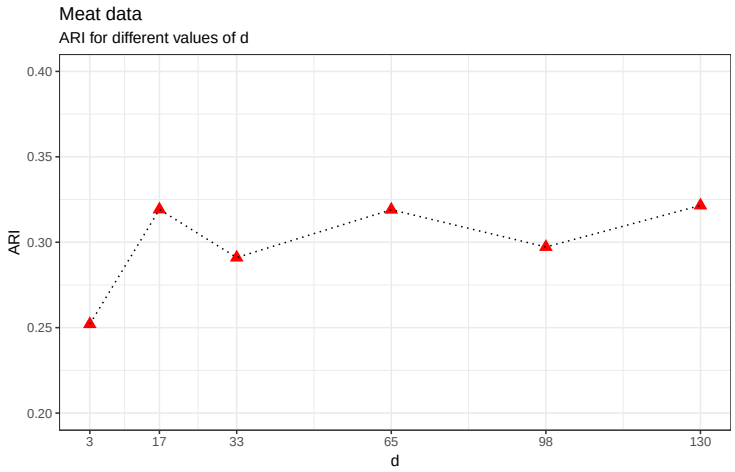


Fig. 3: Meat data. ARI of the random projection ensemble clustering algorithm for different dimensions of the projected space, d. The numbers of generated and selected projections are set equal to $B = 1000$ and $B^* = 100$, respectively.

5 Simulation study

The performance of the RPE Clu algorithm is evaluated in a variety of scenarios through an extensive simulation study. In particular, $G = \{2,4\}$ different Gaussian clusters of size 100 are generated in $p = \{100,500,1000\}$ dimensions by using the `clusteval` `R` package (Ramey 2012). Each population has a p-dimensional multivariate normal distribution, with mean vector

$$\mu_k = \frac{1}{2}\sum_{j=1}^{p/G} e_{(p/G)(k-1)+j},$$

where e_k is the k-th basis vector; therefore, the first p/G dimensions of μ_1 are set to 1 and all the remaining to 0, the second p/G dimensions of μ_2 are set to 1 and all the remaining to 0, and so on. The k-th population covariance matrix is

$$\Sigma_k = (1-\tau_k)\mathbb{1}_p + \tau_k I_p,$$

where $\mathbb{1}_p$ and I_p denote the $p \times p$ matrix of ones and identity matrix, respectively. Here, $-(p-1)^{-1} < \tau_k < 1$ governs the intra-class correlation; throughout the study, we evaluate different levels of correlation between variables, i.e. we take $\tau_k = \{0.1,0.3,0.4,0.6\}$, so as to explore how the clustering algorithm behaves in different situations. Furthermore, we consider scenarios characterized by both *homoscedastic* (settings 1–12) and *heteroscedastic* (settings 13–16) components. Scenarios with *heteroscedastic rotated* components are also investigated (settings 17–20). In this case, as depicted in the illustrative example of Figure 4, the first fifty odd variables of half of the groups are rotated with respect to the axis $x = 0$. A brief description of the simulation settings considered for the analysis is given in Table 1; more details are given in the Supplementary Material.

In addition, we studied the behaviour of our proposal in contexts where original data deviate from Normality. In particular, settings 21–23 consider the exponential, the logarithm

Table 1: Summary description of the simulation settings. When only one value for τ is given, it means that *homoscedastic* Gaussian components are considered. The $*$ indicates *rotated* components.

Setting	p	G	τ
1	100	2	0.1
2	500	2	0.1
3	1000	2	0.1
4	100	4	0.1
5	500	4	0.1
6	1000	4	0.1
7	100	2	0.4
8	500	2	0.4
9	1000	2	0.4
10	100	4	0.4
11	500	4	0.4
12	1000	4	0.4
13	100	2	0.1-0.6
14	100	2	0.1-0.3
15	500	2	0.1-0.6
16	500	2	0.1-0.3
17	100	2	0.1-0.6*
18	100	2	0.1-0.3*
19	500	2	0.1-0.6*
20	500	2	0.1-0.3*

and the square-root transformation of p-variate Gaussian distributions, respectively (p=100, n_g=100, g=1,...,G); the number of groups is set to two and only 50% of the variables are relevant for clustering. Scenarios 24-26 extend the study to the case of four groups.

To validate the proposal, we apply other clustering algorithms on the same settings: the 'standard' Gaussian Mixture Model (McLachlan and Peel 2000) (via `Mclust` function of the `mclust` package), the K-means algorithm (Lloyd 1982) (via `kmeans` function), Ward's agglomerative hierarchical clustering (Ward 1963) (via `hclust` function) and the Partition Around Medoids (pam) (Kaufman and Rousseeuw 2009) (via `pam` function of the `cluster` package). Two recent procedures that have shown good performances in the context of high-dimensional unsupervised classification are also included: namely, the Spectral clustering approach (Ng et al. 2002) (`specc` function of the `kernlab` package) and the Affinity Propagation algorithm (Frey and Dueck 2007) (`apclusterK` function of the `apcluster` package). A further comparison is with the variable selection methodology for Gaussian model-based clustering (Cl VarSel) presented in section 2. This procedure is implemented by using the `clustvarsel` function included in the namesake `R` package (Scrucca and Raftery 2018).

The number of groups G is always taken as known. The default settings of each algorithms are considered, except for the K-means which run with 5 starts. As previously discussed, the RPE Clu algorithm is performed with $B = 1000$, $B^* = 100$ and $d = \{8, 15\}$ with $G = \{2, 4\}$, respectively.

Figure 5 contains the aggregated results for the considered scenarios: (a)-(b) *homoscedastic* Gaussian components with highly correlated features, with two and four groups respectively; (c)-(d) *homoscedastic* Gaussian components with mildly related features, with two and four groups respectively; (e) *heteroscedastic* Gaussian components and *heteroscedastic rotated* Gaussian components; (f)-(g) *non-Gaussian* components, with two and four groups respectively. The boxplots show the distribution of the ARI over 100 simulations of each

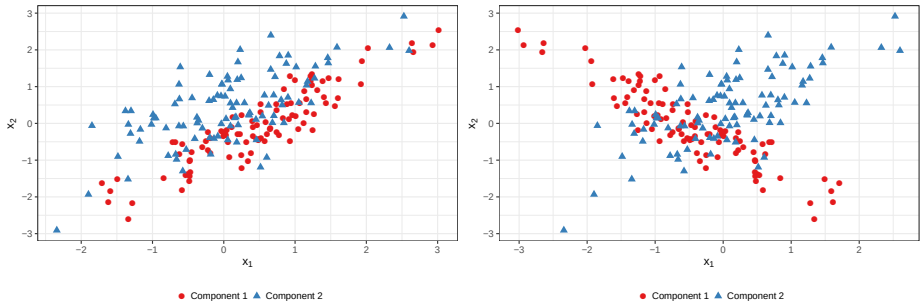


Fig. 4: An example of bivariate dataset with *heteroscedastic rotated* components: variable x_1 of Component 1 (red points) is rotated with respect to the axis $x = 0$.

setting; the horizontal line helps the comparison with the other approaches, by highlighting the median ARI for the random projection ensemble clustering algorithm. Individual Adjusted Rand Indexes of each setting are reported in the Supplementary Material.

Results coming from this numerical study clearly show the general effectiveness of the algorithm we introduce. In fact, for all the situations considered in the boxplots of Figure 5, the RPE Clu produces better solutions than those from the other state-of-the-art methods, including the two procedures that usually work well in high-dimensional contexts (i.e. spectral and affinity propagation clustering algorithms). Not surprisingly, this aspect is particularly evident in those scenarios where the original features are strongly related as some approaches tend to discard this kind of information. With reference to the Mixtures of Gaussians, for example, when p is very large compared to n, `mclust` is able to estimate only those models that have a small number of parameters, i.e. models with spherical, diagonal, or homoscedastic covariance matrix. Furthermore, the K-means algorithm can be viewed as a procedure which attempts to model the data as a mixture of Gaussian distributions with diagonal covariance matrices and thus it does not account for the variable correlation. Scenarios with mildly related features, i.e. 7-9 and 10-12, appear to be very hard tasks: basically all the considered methods perform poorly in terms of recovering the 'true' grouping structure.

As expected, K-means algorithm, hierarchical agglomerative clustering with Ward's method and pam often fail because the distance measures they rely on become increasingly meaningless in high-dimensions; however, with non-Gaussian data they exhibit an acceptable performance.

A special mention should be made for the variable selection procedure (ClVarSel) that seems capable to correctly identify relevant clustering information in most of the settings. Nevertheless, it underperforms the RPE Clu, especially in the case of homoscedastic components with highly correlated features or in case of non-Gaussian data. This outcome corroborates our initial idea that feature extraction techniques are generally more effective than feature selection ones.

Globally, the capability of the RPE Clu in recovering the cluster membership does not change too much with p nor with the number of groups. In addition, it is quite robust to deviations from Gaussianity: plots (f) and (g) show that RPE Clu outperforms the other methods almost always.

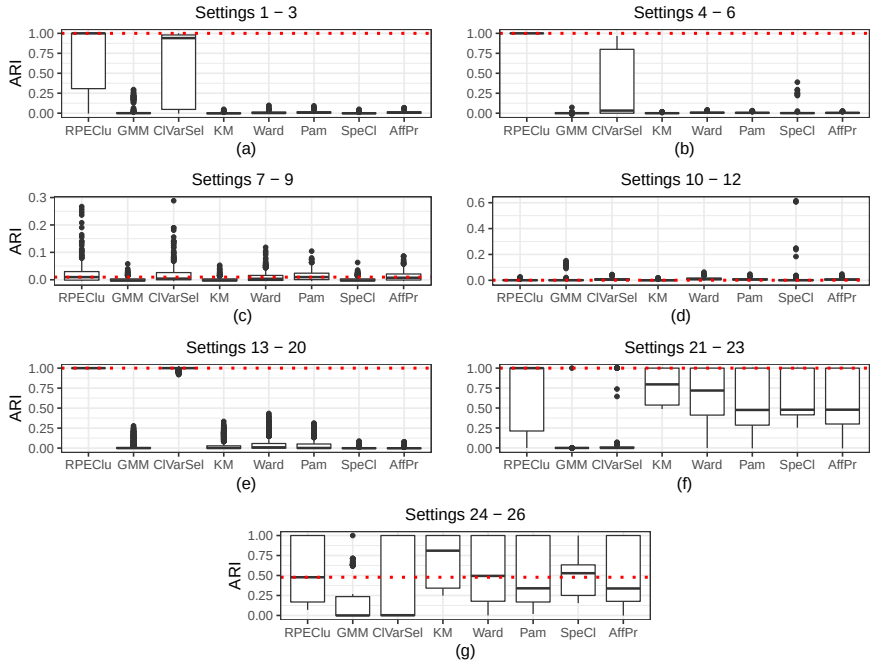


Fig. 5: Performance of different clustering algorithms. The labels along the horizontal axis refer to the different methods: *RPEClu*, Random Projections Ensemble Clustering; *GMM*, Gaussian Mixture Model; *ClVarSel*, Gaussian Mixture Model with Variable Selection; *KM*, k-means clustering; *Ward*, hierarchical clustering with Ward's method; *KM*, k-means clustering; *Pam*, Partition Around Medoids algorithm; *SpeCl*, spectral clustering; *AffPr*, affinity propagation. The seven panels show the distribution of the Adjusted Rand Index for (a) homoscedastic Gaussian clusters ($G = 2$) with highly correlated features, (b) homoscedastic Gaussian clusters ($G = 4$) with highly correlated features, (c) homoscedastic Gaussian clusters ($G = 2$) with mildly correlated features, (d) homoscedastic Gaussian clusters ($G = 4$) with mildly correlated features (e) heteroscedastic Gaussian clusters ($G = 2$), (f) non-Gaussian clusters ($G = 2$) with 50% of relevant features and (g) non-Gaussian clusters ($G = 4$) with 50% of relevant features.

6 Real data examples

For illustration, we evaluate the performances of the clustering algorithms described in the previous section on two different real data experiments. Namely, we use the set of near infrared spectroscopic meat data originally described in the study of Downey et al. (2000) and the Lymphoma Gene Expression dataset used by Chung and Keles (2010).

6.1 Meat Data

This dataset contains $n = 231$ samples of homogenized raw meat coming from $G = 5$ different animal species. The distribution of the samples is described in Table 2. The spectra

Table 2: Distribution of the meat samples

Species	Samples
Beef	32
Chicken	55
Lamb	34
Pork	55
Turkey	55

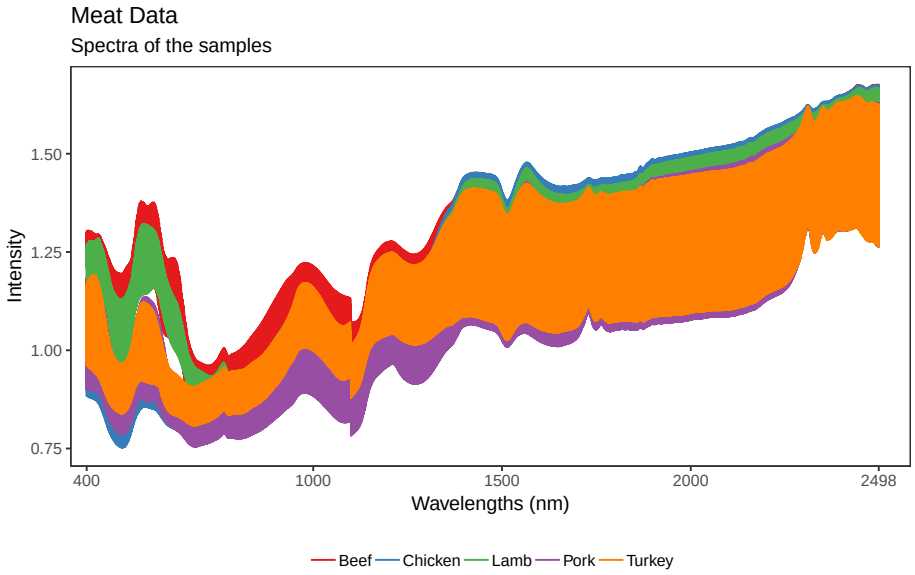


Fig. 6: Meat data. Spectra of the samples, grouped by type of meat.

are recorded over the wavelength range 400 – 2498 nm, with measurements taken every 2 nm. The total number of variables is thus $p = 1050$. Figure 6 shows the spectrum of each sample, grouped by type of meat.

The objective of the analysis is to partition the set of 231 samples so as to reflect the corresponding type of meat by employing the information coming from their spectra. The number of groups $G = 5$ is taken as known; we set $B = 1000$, $B^* = 100$ and $d = 17$. Table 3 contains the Adjusted Rand Index yielded by each method.

Globally, none of the employed methods is able to perfectly recover the original cluster membership of the meat data. Nevertheless, the RPE Clu algorithm provides an Adjusted Rand Index that is considerably superior to all the other solutions. The GMM performs poorly; this is probably due to the fact that, as p is very large, `Mclust` could only estimate mixtures of Gaussians with spherical or diagonal covariance matrices, while data require a model that accounts for the high correlation between the features. The Clust VarSel methodology could not run because variables are too correlated.

Table 3: ARI for the Meat Data.

Method	ARI
RPEClu	0.32
GMM	0.14
k-means	0.18
h-ward	0.23
pam	0.18
Clust VarSel	NA
Specc	0.25
AClust	0.18

Table 4: ARI for the Gene Expression Data.

Method	ARI
RPEClu	1.00
GMM	0.95
k-means	0.95
h-ward	0.95
pam	0.84
Clust VarSel	0.49
Specc	0.95
AClust	0.85

6.2 Gene Expression Data

The lymphoma dataset (taken from the `R` package `spls`) contains the expression levels of $p = 4026$ genes for $n = 62$ patients. The study reports that 42 subjects have diffuse large B-cell lymphoma (DLBCL), 9 follicular lymphoma (FL), and 11 chronic lymphocytic leukemia (CLL). All gene expression profiles were base 10 log-transformed and, in order to prevent single arrays from dominating the analysis, standardized to zero mean and unit variance, as described in Dettling and Bühlmann (2002) and Dettling (2004).

The objective of the analysis is to group patients according to the corresponding lymphoma diagnosis, by using the information on their gene expression levels. RPE Clu procedure run with $B = 1000$, $B^* = 100$ and $d = 12$; the number of groups is taken as known and set equal to 3 for all the methods. Clustering results in terms of ARI are reported in Table 4. As can be seen, the performance of the random projection ensemble clustering algorithm is capable to perfectly detect the grouping structure identified by the diagnosis. Mixture of Gaussians, *K*-means and hierarchical agglomerative clustering with Ward's method provide exactly the same (good) result, up to a label switching. This is due to the fact that, when $p \gg n$, `Mclust` only works on the restricted set of parsimonious models (e.g. spherical or diagonal models) and, therefore, its optimal solution often slightly improves the one yielded by the hierarchical algorithm.

7 Discussion

In this work we propose a novel procedure for model-based clustering of high-dimensional data. This procedure is based on Random Projections and it has been firstly inspired by the original idea of Cannings and Samworth (2017) in the context of supervised classification.

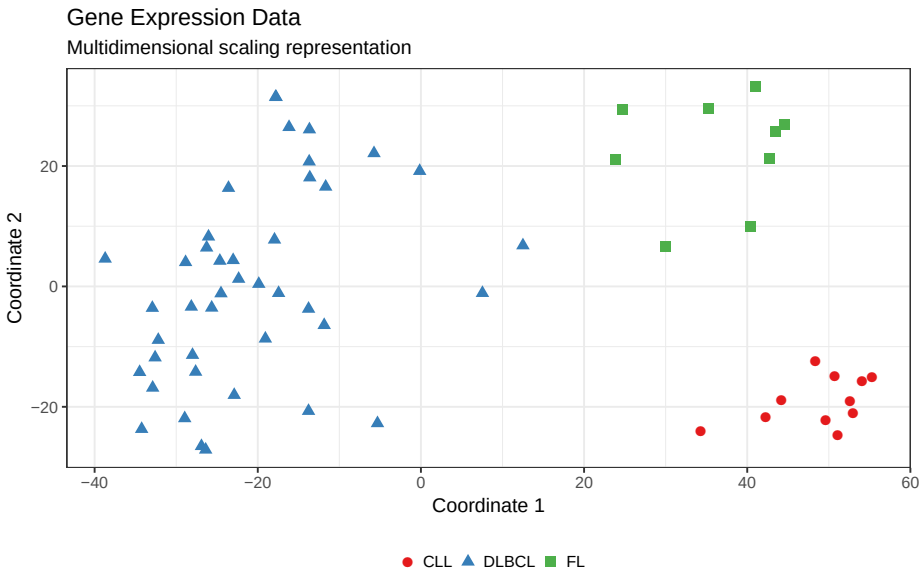


Fig. 7: Gene Expression Data. Multidimensional scaling representation of the samples, grouped by type of diagnosis.

More in detail, we suggest to apply a Gaussian Mixture Model to random projections of the high-dimensional data and to select a subset of solutions accordingly to the Bayesian Information Criterion, computed here as discussed in Raftery and Dean (2006); the multiple 'base' results are then aggregated via consensus to obtain the final partition.

Such proposal has been initially motivated by some benefits associated to RPs for learning Mixture of Gaussians. Dasgupta (2013) proved that a mixture of G Gaussians can be embedded onto just $O(\log G)$ random coordinates without destroying the original group structure too much; furthermore, he demonstrated that even when the original mixing components exhibit elliptical contours, their projected counterparts are always more spherical.

Method performances, evaluated in terms of ARI with respect to the true class membership on both synthetic and real datasets, seem to confirm our motivating ideas. Overall results indeed show that RPs represent a key ingredient that decisively facilitates the learning of high-dimensional mixtures of Gaussians. Moreover, the advantage of their use in conjunction with GMM becomes even more evident as the correlation between the original variables increases. In fact, when dealing with high-dimensional sets, `Mclust` search is restricted to models with few parameters only (i.e. `EII`, `VEI`, `VII`, `VVI`, `EEI` and `EVI`) whereas data would require more complex parameterizations.

The RPE Clu algorithm is a very general tool for model-based clustering of high-dimensional data. We explore in detail its behavior within the Gaussian Mixture model framework only; however, many other distributions can in principle be used. Moreover, further options for combining the clustering results can be tested.

The number of clusters G is fixed here; estimating its value is left to future work.

Acknowledgements. This paper is based upon work supported by the Air Force Office of Scientific Research under award number FA9550-17-1-010.

References

Achlioptas, D. (2003). Database-friendly random projections: Johnson-lindenstrauss with binary coins. *Journal of computer and System Sciences 66*(4), 671–687.

Banfield, J. D. and A. E. Raftery (1993). Model-based gaussian and non-gaussian clustering. *Biometrics 49*, 803–821.

Bellman, R. (1957). *Dynamic programming*. Princeton University Press.

Bhattacharya, A., P. Kar, and M. Pal (2009). On low distortion embeddings of statistical distance measures into low dimensional spaces. In *International Conference on Database and Expert Systems Applications*, pp. 164–172. Springer.

Biernacki, C. and A. Lourme (2014). Stable and visualizable gaussian parsimonious clustering models. *Statistics and Computing 24*(6), 953–969.

Boongoen, T. and N. Iam-On (2018). Cluster ensembles: A survey of approaches with recent extensions and applications. *Computer Science Review 28*, 1–25.

Bouveyron, C. and C. Brunet-Saumard (2014). Model-based clustering of high-dimensional data: A review. *Computational Statistics & Data Analysis 71*, 52–78.

Cannings, T. I. and R. J. Samworth (2017). Random-projection ensemble classification. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 79*(4), 959–1035.

Celeux, G. and G. Govaert (1995). Gaussian parsimonious clustering models. *Pattern recognition 28*(5), 781–793.

Chang, W.-C. (1983). On using principal components before separating a mixture of two multivariate normal distributions. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 32*(3), 267–275.

Chung, D. and S. Keles (2010). Sparse partial least squares classification for high dimensional data. *Statistical applications in genetics and molecular biology 9*(1).

Dasgupta, S. (2013). Experiments with random projection. *preprint arXiv:1301.3849*.

Dettling, M. (2004). Bagboosting for tumor classification with gene expression data. *Bioinformatics 20*(18), 3583–3593.

Dettling, M. and P. Bühlmann (2002). Supervised clustering of genes. *Genome biology 3*(12), research0069–1.

Downey, G., J. McElhinney, and T. Fearn (2000). Species identification in selected raw homogenized meats by reflectance spectroscopy in the mid-infrared, near-infrared, and visible ranges. *Applied Spectroscopy 54*(6), 894–899.

Fern, X. Z. and C. E. Brodley (2003). Random projection for high dimensional data clustering: A cluster ensemble approach. In *Proceedings of the 20th international conference on machine learning (ICML-03)*, pp. 186–193.

Fop, M. and T. B. Murphy (2018). Variable selection methods for model-based clustering. *Statistics Surveys 12*, 18–65.

Frey, B. J. and D. Dueck (2007). Clustering by passing messages between data points. *science 315*(5814), 972–976.

Galimberti, G., A. Manisi, and G. Soffritti (2018). Modelling the role of variables in model-based cluster analysis. *Statistics and Computing 28*(1), 145–169.

Hennig, C. (2019). Cluster validation by measurement of clustering characteristics relevant to the user. *Data Analysis and Applications 1: Clustering and Regression, Modeling-estimating, Forecasting and Data Mining 2*, 1–24.

Hornik, K. (2005). A clue for cluster ensembles. *Journal of Statistical Software 14*(12), 1–25.

Johnson, W. B. and J. Lindenstrauss (1984). Extensions of lipschitz mappings into a hilbert space. *Contemporary mathematics 26*(189-206), 1.

Kaufman, L. and P. J. Rousseeuw (2009). *Finding groups in data: an introduction to cluster analysis*, Volume 344. John Wiley & Sons.

Kittler, J., M. Hatef, R. P. Duin, and J. Matas (1998). On combining classifiers. *IEEE transactions on pattern analysis and machine intelligence 20*(3), 226–239.

Lloyd, S. (1982). Least squares quantization in pcm. *IEEE transactions on information theory 28*(2), 129–137.

Maugis, C., G. Celeux, and M.-L. Martin-Magniette (2009a). Variable selection for clustering with gaussian mixture models. *Biometrics 65*(3), 701–709.

Maugis, C., G. Celeux, and M.-L. Martin-Magniette (2009b). Variable selection in model-based clustering: A general variable role modeling. *Computational Statistics & Data Analysis 53*(11), 3872–3882.

McLachlan, G. and D. Peel (2004). *Finite Mixture Models*. John Wiley & Sons.

McLachlan, G. J., S. X. Lee, and S. I. Rathnayake (2019). Finite mixture models. *Annual review of statistics and its application 6*, 355–378.

McLachlan, G. J. and D. Peel (2000). *Finite Mixture Models*. Wiley.

Ng, A. Y., M. I. Jordan, and Y. Weiss (2002). On spectral clustering: Analysis and an algorithm. In *Advances in neural information processing systems*, pp. 849–856.

Raftery, A. E. and N. Dean (2006). Variable selection for model-based clustering. *Journal of the American Statistical Association 101*(473), 168–178.

Ramey, J. A. (2012). *clusteval: Evaluation of Clustering Algorithms*. R package version 0.1.

Ruiz, F. E., P. S. Pérez, and B. I. Bonev (2009). *Information theory in computer vision and pattern recognition*. Springer Science & Business Media.

Scrucca, L. (2016). Genetic algorithms for subset selection in model-based clustering. In *Unsupervised Learning Algorithms*, pp. 55–70. Springer.

Scrucca, L. and A. E. Raftery (2018). clustvarsel: A package implementing variable selection for gaussian model-based clustering in R. *Journal of Statistical Software 84*(1), 1–28.

Strehl, A. and J. Ghosh (2002). Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *Journal of machine learning research 3*(Dec), 583–617.

Ward, J. H. J. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American statistical association 58*(301), 236–244.

Xu, D. and Y. Tian (2015). A comprehensive survey of clustering algorithms. *Annals of Data Science 2*(2), 165–193.