Robust model-based clustering for high-dimensional data via covariance matrices regularization

Luca Panzeri, Davide Zaltieri

October 5^{th}, 2023

Advisor:
Dr. Andrea Cappozzo

Introduction

Robust clustering for high-dimensional data represents a significant challenge.

WHY?

Introduction

Robust clustering for high-dimensional data represents a significant challenge.

WHY?

1. Existing robust clustering methods fail when the number of variables is large.
2. Existing clustering approaches for high-dimensional data are not robust.

Introduction

Robust clustering for high-dimensional data represents a significant challenge.

WHY?

1. Existing robust clustering methods fail when the number of variables is large.
2. Existing clustering approaches for high-dimensional data are not robust.

We propose a solution to this challenge.

HOW? $\rightarrow$ by integrating high-dimensional covariance matrix estimators into the efficient TCLUST methodology for robust constrained clustering.

TCLUST formulation

Fritz, Garcìa-Escudero, Mayo-Iscar (2012)

Search for a partition $R_0, R_1, \ldots, R_k$ of indeces $\{1, \ldots, n\}$ with $\# R_0 = \lceil n\alpha \rceil$, centers $\boldsymbol{m}_1, \ldots, \boldsymbol{m}_k$ in $\mathbb{R}^p$, symmetric positive semidefinite $p \times p$ scatter matrices $\boldsymbol{S}_1, \ldots, \boldsymbol{S}_k$ and weights $p_1, \ldots, p_k$, which maximizes

$$\sum_{j=1}^{k} \sum_{i \in \mathbb{R}_j} \log \left(p_j \phi(\boldsymbol{x}_i; \boldsymbol{m}_j, \boldsymbol{S}_j) \right)$$

under the eigenvalue ratio constraint

$$\frac{\max_{j,l} \lambda_l \left(\boldsymbol{S}_j \right)}{\min_{j,l} \lambda_l \left(\boldsymbol{S}_j \right)} \leq c \quad \text{for } j = 1, \ldots, k, \; l = 1, \ldots, p.$$

Limitations

TCLUST fails when dealing with high-dimensional data.

Specifically, it has two main limitations:

Limitations

TCLUST fails when dealing with high-dimensional data.

Specifically, it has two main limitations:

- Initialization issue $\rightarrow$ initial subset size may exceed the total number of observations, increasing the risk of including outliers in the initialization phase.

Limitations

TCLUST fails when dealing with high-dimensional data.

Specifically, it has two main limitations:

- Initialization issue $\rightarrow$ initial subset size may exceed the total number of observations, increasing the risk of including outliers in the initialization phase.
- Singularity issue $\rightarrow$ covariance matrices may become singular, resulting in their impracticability when using the traditional sample covariance matrix estimator that involves matrix inversion.

Limitations

TCLUST fails when dealing with high-dimensional data.

Specifically, it has two main limitations:

- Initialization issue $\rightarrow$ initial subset size may exceed the total number of observations, increasing the risk of including outliers in the initialization phase.
- Singularity issue $\rightarrow$ covariance matrices may become singular, resulting in their impracticability when using the traditional sample covariance matrix estimator that involves matrix inversion.

The use of covariance matrix estimators involving regularization techniques becomes necessary.

Minimum Regularized Covariance Determinant

The Minimum Regularized Covariance Determinant is a robust estimator for covariance matrices.

Its goal is to find $H_{MRCD} \subseteq \{1, \ldots, n\}$ such that:

$$H_{MRCD} = \underset{H \in \mathcal{H}_h}{\operatorname{argmin}} \left(\det(\mathbf{K}(H))^{1/p} \right),$$

where $\mathbf{K}(H) = \rho \mathbf{T} + (1 - \rho) c_\alpha \boldsymbol{S}_U(H)$ is the regularized covariance matrix for a given $H \subseteq \{1, \ldots, n\}$.

Once H_{MRCD} is determined, the MRCD covariance matrix estimate is computed based on this subset.

Linear shrinkage estimator of Ledoit-Wolf

The linear shrinkage estimator of Ledoit-Wolf for $\boldsymbol{\Sigma}$ is found as

$$\boldsymbol{\Sigma}^* = \rho_1 \boldsymbol{I} + \rho_2 \boldsymbol{S}$$

$$\text{that minimizes } E\left[\|\boldsymbol{\Sigma}^* - \boldsymbol{\Sigma}\|^2\right].$$

It computes an asymptotically optimal linear combination of the sample covariance matrix and the identity matrix, effectively shrinking the eigenvalues of the sample covariance matrix towards the identity.

CovGlasso

The CovGlasso estimator for $\mathbf{\Sigma}$ is found by minimizing

$$\log(\det(\mathbf{\Sigma})) + \mathrm{tr}(\mathbf{\Sigma}^{-1}\boldsymbol{S}) + \lambda\|\boldsymbol{P} * \mathbf{\Sigma}\|_1,$$

which is minus the penalized log-likelihood of a p-variate Gaussian distribution with zero mean and covariance matrix $\mathbf{\Sigma}$.

It estimates a sparse covariance matrix, specifically using a fast coordinate descent algorithm to solve the covariance graphical lasso.

Modifying TCLUST algorithm

The initial proposal is to improve the TCLUST methodology by integrating the Minimum Regularized Covariance Determinant estimator.

We work on the TCLUST code, specifically arranging the **initialization** and **parameter estimation**:

Modifying TCLUST algorithm

The initial proposal is to improve the TCLUST methodology by integrating the Minimum Regularized Covariance Determinant estimator.

We work on the TCLUST code, specifically arranging the **initialization** and **parameter estimation**:

- by initially assigning observations to clusters randomly, designating a portion α of them as outliers.

Modifying TCLUST algorithm

The initial proposal is to improve the TCLUST methodology by integrating the Minimum Regularized Covariance Determinant estimator.

We work on the TCLUST code, specifically arranging the **initialization** and **parameter estimation**:

- by initially assigning observations to clusters randomly, designating a portion α of them as outliers.
- by computing cluster means and covariance matrices using the MRCD estimator.

Modifying TCLUST algorithm

The initial proposal is to improve the TCLUST methodology by integrating the Minimum Regularized Covariance Determinant estimator.

We work on the TCLUST code, specifically arranging the **initialization** and **parameter estimation**:

- by initially assigning observations to clusters randomly, designating a portion α of them as outliers.
- by computing cluster means and covariance matrices using the MRCD estimator.
- by incorporating a threshold for the minimum cluster size.

Modifying TCLUST algorithm

The initial proposal is to improve the TCLUST methodology by integrating the Minimum Regularized Covariance Determinant estimator.

We work on the TCLUST code, specifically arranging the **initialization** and **parameter estimation**:

- by initially assigning observations to clusters randomly, designating a portion α of them as outliers.
- by computing cluster means and covariance matrices using the MRCD estimator.
- by incorporating a threshold for the minimum cluster size.

In the main function, at the end of a nested loop of multiple initializations and iterations, we select the best initialization, which is the one resulting in the highest objective function value.

Heuristic or likelihood-based?

Our algorithm is using the same objective function as the original TCLUST methodology → it does not consider the contribution of the regularization applied to the covariance matrix.

To incorporate this regularization into the objective function, we need to explore the possibility of reformulating the MRCD problem in terms of likelihood. This would allow us to rewrite the objective function of our MRCD in TCLUST as a summation of k penalized log-likelihoods, each representing the objective function of MRCD for an individual cluster, thus transforming MRCD in TCLUST into a likelihood-based methodology.

However, our analysis concludes that the MRCD estimation problem cannot be reformulated in terms of likelihood. As a result, MRCD in TCLUST remains heuristic.

New objective function formulation

We develop another algorithm founded on the TCLUST framework, this time incorporating the Gaussian-based CovGlasso estimator with the aim of creating a likelihood-based methodology.

We can formulate the objective function of CovGlasso in TCLUST as the summation of k minus penalized log-likelihoods, each representing the objective function of the CovGlasso methodology for an individual cluster:

$$\sum_{j=1}^{k} \left(\log \left(\det \left(\widehat{\boldsymbol{\Sigma}}_j \right) \right) + \operatorname{tr} \left(\widehat{\boldsymbol{\Sigma}}_j^{-1} \boldsymbol{S}_j \right) + \lambda \left\| \boldsymbol{P} * \widehat{\boldsymbol{\Sigma}}_j \right\|_1 \right).$$

This objective function needs to be collectively minimized, as it is the sum of k CovGlasso objective functions, each requiring minimization.

Algorithm updates

- A new type of initial cluster assignments, involving the application of TCLUST on a low-dimensional subset of the original variables, is introduced.

Algorithm updates

- A new type of initial cluster assignments, involving the application of TCLUST on a low-dimensional subset of the original variables, is introduced.
- Parameter estimation within clusters is enhanced by incorporating the CovGlasso estimator.

Algorithm updates

- A new type of initial cluster assignments, involving the application of TCLUST on a low-dimensional subset of the original variables, is introduced.
- Parameter estimation within clusters is enhanced by incorporating the CovGlasso estimator.
- The new objective function is implemented.

Algorithm updates

- A new type of initial cluster assignments, involving the application of TCLUST on a low-dimensional subset of the original variables, is introduced.
- Parameter estimation within clusters is enhanced by incorporating the CovGlasso estimator.
- The new objective function is implemented.
- In the main function, we modify the final condition for selecting the best initialization to be the one that yields the lowest objective function value.

From doubly-robust to individually-robust approach

The MRCD in TCLUST methodology presents a doubly-robust extension, effectively addressing outliers in two ways:

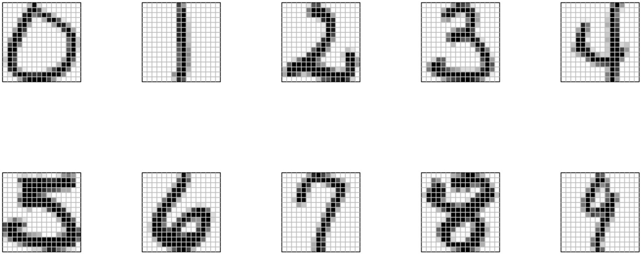
1. within TCLUST, when the trimming procedure is applied.
2. within MRCD, taking advantage of its robust-based estimation when computing the regularized covariance matrices.

The Ledoit-Wolf estimator, which is a particular case of MRCD, lacks robustness to outliers. Nevertheless, when integrated into the TCLUST algorithm, it becomes a sensible choice, as TCLUST already ensures robustness $\rightarrow$ we replace MRCD with the Ledoit-Wolf estimator, resulting in the new Ledoit-Wolf in TCLUST, which is still a heuristic methodology.

Dataset presentation

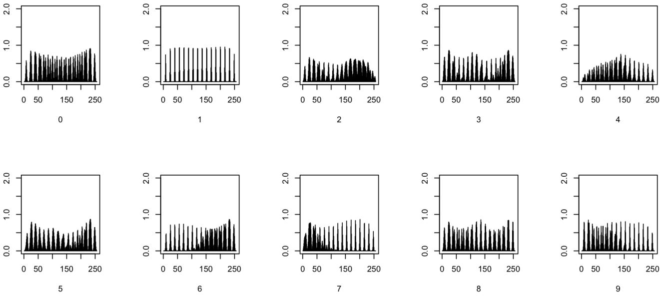
We focus on the task of recognizing handwritten digits sourced from the USPS dataset, which is available through the UCI Machine Learning Repository.

This dataset contains images of handwritten digits ranging from 0 to 9, each partitioned into a 16×16 grid and resulting in 256 pixels as the feature set.



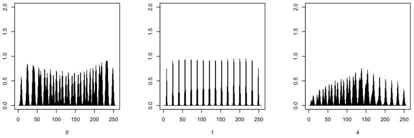
Analysis of multivariate means

Our initial goal is to identify two separate groups of digits by plotting their multivariate means. One group comprises quite distinguishable digits, while the other consists of digits that share the highest similarities among themselves. This approach enables us to evaluate the performance of our algorithms across datasets with different levels of complexity.

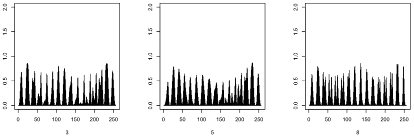


Analysis of multivariate means

Digits 0, 1 and 4 exhibit clearly distinct behaviors.



Digits 3, 5 and 8 display remarkably similar multivariate means.



Preprocessing

- USPS014 creation: we create the dataset USPS014 by randomly selecting 50 data points from each subset of digits 0, 1 and 4, while also including 5 randomly chosen outliers from the subset of all the other digits.
- USPS358 creation: we repeat the same process for digits 3, 5 and 8, resulting in the creation of the dataset USPS358.

Preprocessing

- USPS014 creation: we create the dataset USPS014 by randomly selecting 50 data points from each subset of digits 0, 1 and 4, while also including 5 randomly chosen outliers from the subset of all the other digits.
- USPS358 creation: we repeat the same process for digits 3, 5 and 8, resulting in the creation of the dataset USPS358.
- Variable selection: we perform variable selection to reduce the dimensionality of our datasets by eliminating irrelevant features. We set a variance threshold of 0.5 and discard variables with variance below it, retaining approximately **130 features** out of the total 256.

Preprocessing

- USPS014 creation: we create the dataset USPS014 by randomly selecting 50 data points from each subset of digits 0, 1 and 4, while also including 5 randomly chosen outliers from the subset of all the other digits.
- USPS358 creation: we repeat the same process for digits 3, 5 and 8, resulting in the creation of the dataset USPS358.
- Variable selection: we perform variable selection to reduce the dimensionality of our datasets by eliminating irrelevant features. We set a variance threshold of 0.5 and discard variables with variance below it, retaining approximately **130 features** out of the total 256.
- Evaluation metrics choice: we choose **overall accuracy** and **Adjusted Rand Index** as similarity measures betweeen the estimated labels and the true labels of the digits to evaluate the performance of our algorithms.

Results on USPS014

Ledoit-Wolf in TCLUST:

- **Overall accuracy** = 90.3%
- **ARI** = 0.729
- 5/5 outliers correctly detected

group	0	1	4	out
0	45	3	2	0
1	0	50	0	0
4	1	9	40	0
out	0	0	0	5

CovGlasso in TCLUST:

- **Overall accuracy** = 96.8%
- **ARI** = 0.905
- 5/5 outliers correctly detected

group	0	1	4	out
0	48	0	2	0
1	0	49	1	0
4	0	2	48	0
out	0	0	0	5

Results on USPS358

Ledoit-Wolf in TCLUST:

- **Overall accuracy** = 60.0%
- **ARI** = 0.172
- 5/5 outliers correctly detected

group	3	5	8	out
3	29	9	12	0
5	16	29	5	0
8	9	11	30	0
out	0	0	0	5

CovGlasso in TCLUST:

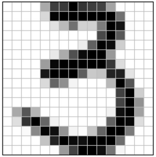
- **Overall accuracy** = 69.7%
- **ARI** = 0.385
- 5/5 outliers correctly detected

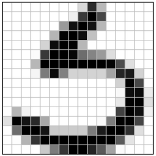
group	3	5	8	out
3	48	2	0	0
5	29	19	2	0
8	4	10	36	0
out	0	0	0	5

Results on USPS358

CovGlasso in TCLUST struggles to correctly identify the true 5's, often assigning them to the estimated cluster of 3's.

This issue arises from the high similarity between the two types of digits:

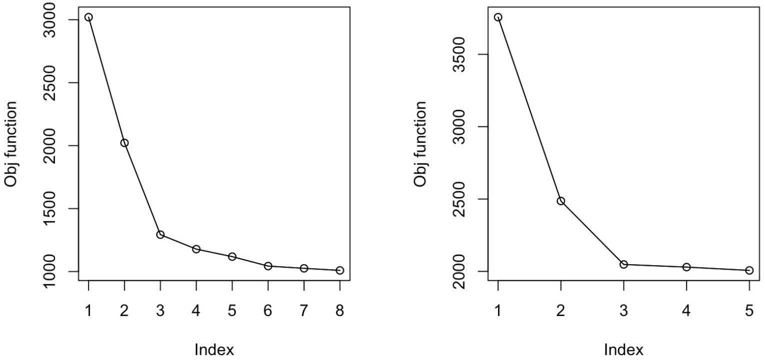




Nevertheless, CovGlasso in TCLUST produces satisfactory results on this complex dataset.

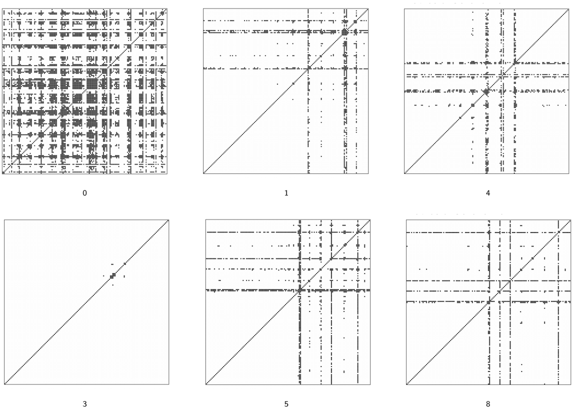
Validation of CovGlasso in TCLUST

Decreasing trend in the objective function:



Validation of CovGlasso in TCLUST

Sparsity patterns in the covariance matrices:



Conclusions

- Both Ledoit-Wolf in TCLUST and CovGlasso in TCLUST demonstrate robustness and effectiveness when clustering high-dimensional and contaminated data.

Conclusions

- Both Ledoit-Wolf in TCLUST and CovGlasso in TCLUST demonstrate robustness and effectiveness when clustering high-dimensional and contaminated data.
- CovGlasso in TCLUST demonstrate robustness and effectiveness when clustering high-dimensional, contaminated and limitedly separated data, outperforming Ledoit-Wolf in TCLUST.

Conclusions

- Both Ledoit-Wolf in TCLUST and CovGlasso in TCLUST demonstrate robustness and effectiveness when clustering high-dimensional and contaminated data.
- CovGlasso in TCLUST demonstrate robustness and effectiveness when clustering high-dimensional, contaminated and limitedly separated data, outperforming Ledoit-Wolf in TCLUST.

CovGlasso in TCLUST results in our final methodology for robustly and effectively addressing clustering challenges in complex, contaminated and high-dimensional data.

Bibliography

- J. Bien and R. J. Tibshirani. Sparse estimation of a covariance matrix. Biometrika, pages 807–820, 2011.
- K. Boudt, P. J. Rousseeuw, S. Vanduffel, and T. Verdonck. The minimum regularized covariance determinant estimator. Statistics and Computing, 30:113–128, 2020.
- C. Bouveyron, G. Celeux, B. Murphy, and A. Raftery. Model-based Clustering and Classification for Data Science, with Applications in R - Chapter 8, volume 4 of 50. Cambridge University Press, 6 2019. ISBN 9781108644181.
- A. Casa, A. Cappozzo, and M. Fop. Group-wise shrinkage estimation in penalized model-based clustering. Journal of Classification, 39:648–674, 10 2022.
- H. Fritz, L. A. Garcìa-Escudero, and A. Mayo-Iscar. A fast algorithm for robust constrained clustering. Computational Statistics and Data Analysis, pages 124–136, 11 2012.
- H. Fritz, L. A. Garcìa-Escudero, and A. Mayo-Iscar. tclust: An R package for a trimming approach to cluster analysis. 12(47):1–26, 5 2012.
- L. A. Garcìa-Escudero, A. Gordaliza, C. Matràn, and A. Mayo-Iscar. A general trimming approach to robust cluster analysis. The Annals of Statistics, 36(3):1324–1345, 6 2008.
- O. Ledoit and M. Wolf. A well-conditioned estimator for large-dimensional covariance matrices. Journal of Multivariate Analysis, 88(2):365–411, 2 2004.
- J. Raymaekers and P. J. Rousseeuw. The cellwise minimum covariance determinant estimator. 7 2022.
- V. Todorov and P. Filzmoser. An object-oriented framework for robust multivariate analysis. Journal of Statistical Software, 32(3):1–47, 10 2009.
- P. Tseng. Convergence of a block coordinate descent method for nondifferentiable minimization. Journal of optimization theory and applications, 109(3):475–494, 6 2001.
- H. Wang. Coordinate descent algorithm for covariance graphical lasso. Statistics and Computing, 24:521–529, 2013.

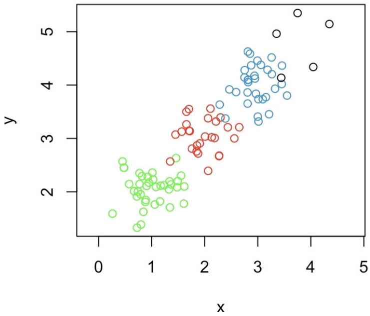
THANK YOU FOR YOUR ATTENTION!

Luca Panzeri
Davide Zaltieri

Data generation and visualization

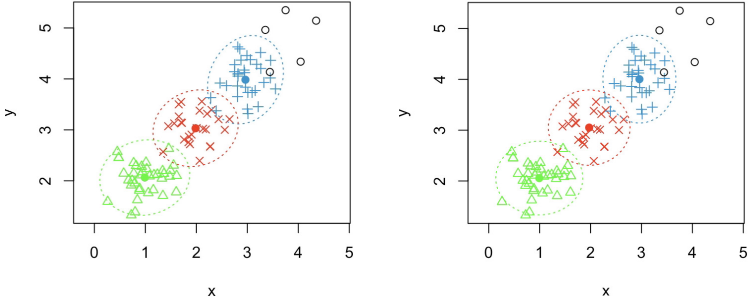
We simulate a high-dimensional three-component mixture distribution, with each component modeled as a Gaussian, and additionally introduce outliers by using a separate high-dimensional Gaussian distribution.

Simulated data in the first two dimensions of the feature space:



Results of algorithms application

- Both algorithms correctly detect all outliers $\rightarrow$ **robustness**.
- Both algorithms accurately identify all data points $\rightarrow$ **effectiveness**.



Example: sparsity plot of a 10x10 covariance matrix