Truncated Pairwise Likelihood for Sparse High-Dimensional Covariance Estimation

WU Institute of Statistics and Mathematics - Research Seminar Series

Alessandro Casa

Joint work with: D. Ferrari & Z. Huang

Faculty of Economics and Management

Free University of Bozen-Bolzano

alessandro.casa@unibz.it

Wien, 14th May 2025

Framework

- Estimation of covariance matrices is one of the fundamental problems in multivariate analysis (even a wikipedia page!)
  - Principal component analysis
  - Spatio-temporal data
  - Model-based classification and clustering
  - ...
- Given a random sample $\{X_1, \ldots, X_n\}$, with $X_i \sim \mathcal{N}_p(0, \boldsymbol{\Theta})$, for $i = 1, \ldots, n$, the log-likelihood is

$$\ell(\boldsymbol{\Theta}) \propto -\log|\boldsymbol{\Theta}| - \text{tr}(\boldsymbol{\Theta}^{-1} S)$$

It is easy to show that

$$\hat{\boldsymbol{\Theta}}_{\text{MLE}} = \mathbf{S} = \frac{1}{n} \sum_{i=1}^{n} X_i X_i^\top$$

Where the problems start

- Number of free variance/covariance parameters to estimate is

$$|\Theta| = \frac{p(p+1)}{2}$$

  thus growing quadratically with the dimension p

> *"The computational ease with which this abundance of parameters can be estimated should not be allowed to obscure the probable unwisdom of such estimation from limited data"*
>
> Dempster (1972)

Digging into the problem

- In large dimensional scenarios, with p and p/n are large, the sample covariance matrix is known to be an highly unstable estimate of $\boldsymbol{\Theta}$ (Stein's paradox intuition)
- Idea: error of standard MLE increases for larger p and sparser parameter vector. Easier to beat $\hat{\boldsymbol{\Theta}}_{\text{MLE}}$ in case of sparsity

Bet on sparsity principle

Use a procedure that does well in sparse problems, since no procedure does well in dense problems

- This intuition has led to a growing stream of literature

What's out there?

1. Matrix decomposition: among the oldest approaches to provide parsimonious representation of $\boldsymbol{\Theta}$ (e.g. latent factor models)

2. Methods betting on sparsity: obtain sparse estimates of $\boldsymbol{\Theta}$. We might have two different tasks: 1) structure learning; 2) parameter estimation

   - Banding, tapering, thresholding
     Thresholding general idea

$$\hat{\boldsymbol{\Theta}}_{\text{TR},jk} = \begin{cases} S_{jk} & \text{if} \quad j = k \\ \mathcal{T}_\lambda(S_{jk}) & \text{if} \quad j \neq k \end{cases}$$

with S_{jk} the (j,k)-th element of S and $\mathcal{T}_\lambda(\cdot)$ a general thresholding function, depending on tuning parameter λ

Hard thresholding example: $\hat{\boldsymbol{\Theta}}_{\text{TR},jk} = S_{jk}\mathbb{1}(|S_{jk}| > \lambda)$

What's out there?

- Penalized likelihood estimator
  most common approaches to induce sparsity in $\hat{\boldsymbol{\Theta}}$ (or more commonly in the precision matrix $\hat{\boldsymbol{\Theta}}^{-1}$ $\longrightarrow$ graphical lasso)

General idea

maximize $\ell(\boldsymbol{\Theta}) - p_\lambda(\boldsymbol{\Theta})$ where $p_\lambda(\boldsymbol{\Theta})$ encourages shrinkage towards 0 of $\boldsymbol{\Theta}$ off-diagonal elements

- Covariance graphical lasso (Bien and Tibshirani, 2011)

$$\underset{\Sigma}{\operatorname{argmax}} \ \ell(\boldsymbol{\Theta}) - \lambda \|\boldsymbol{\Theta}\|_1$$

where $\|A\|_1 = \sum_{ij} |A_{ij}|$ and λ a tuning parameter

  - **Pros**: connection with covariance graphical models
  - **Cons**: produces biased estimates of the non-zero elements

Composite likelihood - crash course

Composite likelihood - What is it?

CL function is a type of approximation of a complex likelihood function obtained through the combination of several low-dimensional likelihood objects

- Let $X \in \mathbb{R}^p$ a r.v. with pdf $f(x;\omega)$, for $\omega \in \Omega \subseteq \mathbb{R}^d$, being difficult to specify or to compute, and $\{\mathcal{H}_1, \ldots, \mathcal{H}_K\}$ a set of events with likelihood $\mathcal{L}_k(\omega; x) \propto f(x \in \mathcal{H}_k; \omega)$.
  The composite likelihood is then defined as

$$\mathcal{L}_{CL}(\omega, w) = \prod_{k=1}^{K} \mathcal{L}_k(\omega; x)^{w_k}$$

  with w_k, $k = 1, \ldots, K$, weights to be chosen
- $\mathcal{H}_k$ determines the type of composite likelihood
  - Marginal CL
  - Conditional CL

All that glitters...

- Given a sample $\mathbf{X} = \{X_1, \ldots, X_n\}$, the estimator $\hat{\omega}_{\text{CL}}$ is obtained as the solution of the unbiased estimating equation

$$u(\omega, w; \mathbf{X}) = \sum_{i=1}^{n} \sum_{k=1}^{K} w_k u_k(\omega; X_i) = 0$$

  where $u_k(\omega; X_i) = \nabla_\omega \log \mathcal{L}_k(\omega; x_i)$ and for an appropriate w_k

- Popularity of CL-estimator stems from its properties, resembling those of ML-estimator. In fact

$$\hat{\omega}_{\text{CL}} \overset{a}{\sim} \mathcal{N}(\omega_0, \mathbf{G}(\omega_0, w)^{-1})$$

But there's more

For common exponential families, some specific composite likelihood estimators retain the full efficiency of the ML ones (Mardia, 2009)

...with some open questions

Issues - Lindsay et al. (2011)
Virtually infinite flexibility but less attention on how to properly design composite likelihood

- Focus on the weights w_k as their choice determines both the statistical properties and the computational efficiency of the CL-estimator
- Theory of unbiased estimation equation allows to find *optimal weights* but their computation is unfeasible (Heyde, 2008)
- Often selected heuristically or in application-oriented manner

Sparsifying the composite likelihood

- Idea: start from *all* potential sub-likelihoods and find a data-driven way to select the "useful" ones
- How to: select w_k's by minimizing the penalized score distance

$$\underbrace{\frac{1}{2}\mathbb{E}\left\|u^{ML}(\omega;\mathbf{X}) - u(\omega, w;\mathbf{X})\right\|_2^2}_{\text{Statistical efficiency}} + \underbrace{\lambda \sum_{j<k} |w_{jk}|}_{\text{Sparsity}},$$

where u^{ML} is the ML score, $\|\cdot\|_2$ denotes the L_2-norm and λ a tuning parameter

- When $\lambda > 0$, properties of the L_1-penalty discourage the inclusion of too many sub-scores
- The minimizer $\hat{w}_\lambda(\theta)$ is used for parameter estimation by solving $u(\theta, \hat{w}_\lambda(\theta)) = 0$ and estimator still enjoys good properties

Connecting the dots

Main question

How can we exploit this sparse composite likelihood framework to sparsely estimate covariance matrices?

- Idea: build composite likelihood where each sub-likelihood contains one distinct element of $\boldsymbol{\Theta}$
- How to: consider a **pairwise likelihood** where $K = \frac{p(p+1)}{2}$ and where the score component will be

$$u_k(\boldsymbol{\Theta}; \mathbf{X}) = u_{ij}(\boldsymbol{\Theta}; \mathbf{X}) = \begin{cases} \frac{\partial \log f(X_i; \boldsymbol{\Theta})}{\partial \boldsymbol{\Theta}} & \text{if } i = j \\ \frac{\partial \log f(X_i, X_j; \boldsymbol{\Theta})}{\partial \boldsymbol{\Theta}} & \text{if } i \neq j \end{cases}$$

Main proposal

Rationale

info about Θ_{ij} is contained only in the sub-score $u_{ij}(\boldsymbol{\Theta}; \mathbf{X})$

$\Downarrow$

When ith and jth variable have negligible correlation, the score $u_{ij}(\boldsymbol{\Theta}; \mathbf{X})$ does not contribute meaningful information and we should have $w_{ij} = 0$

Idea

Model selection problem of determining non-zero covariance entries is recast in terms of selecting informative sub-scores among $p(p-1)/2$ possible contributions

Practical estimation

- For a given $\lambda \geq 0$, we estimate the weight vector $\hat{w} = \{\hat{w}_{jk}\}_{j \leq k}$ by minimizing the empirical objective function

$$\hat{d}_\lambda(w) = \frac{1}{2} w^\top \hat{\mathbf{J}} w + w^\top \operatorname{diag}(\hat{\mathbf{J}}) + \frac{\lambda}{n} \sum_{j<k} \frac{|w_{jk}|}{S_{jk}^2}$$

  where $\hat{\mathbf{J}}$ is the sample estimate of the score covariance matrix
- Convex minimization problem, solved via coordinate descent algorithm. Solutions provide insights on the selection process
- Given $\hat{w}$, the final estimator is found by solving $u(\boldsymbol{\Theta}, \hat{w}; \mathbf{X}) = 0$. We have the thresholding mechanism

$$\hat{\boldsymbol{\Theta}}_{\text{TPL},jk} = \begin{cases} S_{jk} & \text{if } \hat{w}_{jk} \neq 0 \\ 0 & \text{if } \hat{w}_{jk} = 0 \end{cases}$$

Theoretical guarantees

- Let $\mathcal{E} = \{jk : j < k, \Theta_{jk} \neq 0\}$, $\hat{\mathcal{E}}$ its estimated counterpart. Under some regularity conditions we have

  Selection consistency for large p
  $\mathbb{P}(\hat{\mathcal{E}} = \mathcal{E}) \to 1$ as $n \to \infty$, with p potentially exponentially increasing with n

Convergence to the oracle maximum likelihood estimator
Under the assumption used to prove selection consistency

$$\lim_{n \to \infty} \mathbb{P}(\hat{\mathbf{\Theta}}_{\text{TPL}} = \hat{\mathbf{\Theta}}_{\text{OML}}) = 1$$

where $\hat{\mathbf{\Theta}}_{\text{OML}}$ is the ML estimator which assume the knowledge of the location of non-zero entries

Additional insights

- Estimation of the weight vector $\hat{w}$ shed light on the sub-score selection mechanism
- In particular

$$\hat{\mathcal{E}} = \left\{ jk : |\text{diag}\{\hat{\mathbf{J}}\}_{jk} - \hat{\mathbf{J}}_{jk,.}\hat{w}| \geq \frac{\lambda}{nS_{jk}^2} \mathbb{1}(j \neq k) \right\}$$

Rationale

The jkth sub-score is included when it contributes relatively large information in the overall pairwise likelihood.

Contribution is measured by the difference between marginal Fisher information and information already present in previously selected score

Relationships with adaptive thresholding

- Adaptive thresholding introduces the following estimator

$$\hat{\Theta}_{\text{ATR},jk} = S_{jk}\mathbb{1}\left(\frac{|S_{jk}|}{SE_{jk}} > \lambda\right)$$

  - Pro: extremely simple in high-dim problems
  - Con: it is marginal

- Our estimator can be expressed as

$$\hat{\Theta}_{\text{TPL},jk} = S_{jk}\mathbb{1}\left(\frac{|S_{jk}|}{SE^{\text{adj}}_{jk}} > \lambda\right)$$

with SE^{adj}_{jk} the adjusted std error, computed removing the portion of variance already explained by already selected scores

Some thoughts on λ selection

- Usually different heuristic can be followed
  Example: stop adding sub-scores (i.e., decreasing λ) as soon as they do not increase substantially the explained variability

- We *(try to)* exploit the connection with adaptive thresholding whose condition resembles the rejection region of a test
- Idea: propose a data-driven criterion to select λ based on sequential testing of null hypotheses

$$\begin{aligned} H_0 &: \Theta_{jk} = 0 \,|\, \hat{w}_{jk} \neq 0 \\ H_1 &: \Theta_{jk} \neq 0 \,|\, \hat{w}_{jk} \neq 0 \end{aligned}$$

based on a level α (type I error control)

Synthetic data exploration

- We considered $n = \{40, 100, 250\}$ and $p = \{20, 50, 150\}$ for different levels of sparsity $\tau = \{0.5, 0.9\}$ and two structures
  - Block diagonal
  - Sparse at random

- Insights from simulations
  - Method works better when $\tau = 0.9$ (as expected)
  - Sparse at random is more challenging
    - structurally, a lot of non-zero entries are closed to zero
    - with $\tau = 0.5$ we have visible over-sparsification producing many false negatives but leading to better MSE with respect to $\hat{\mathbf{\Theta}}_{\text{OML}}$

Some results - Urban land cover data

- $n = 675$ aerial images for 9 types of different urban land cover
- $p = 147$ numerical features

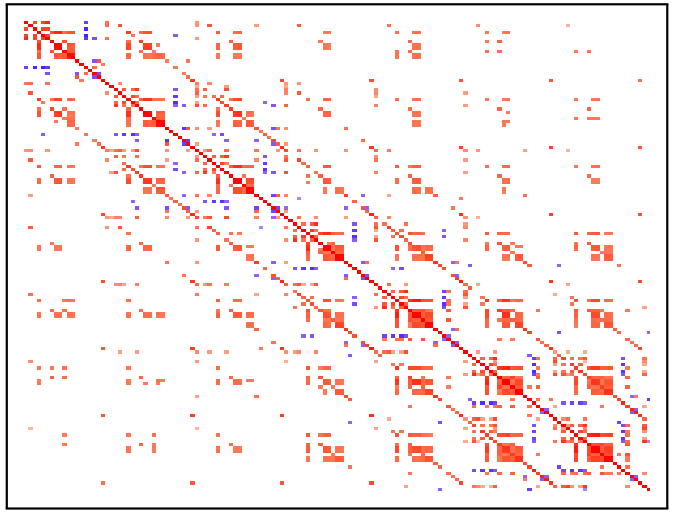
$$\Downarrow$$

21 variables repeated at 7 different scales

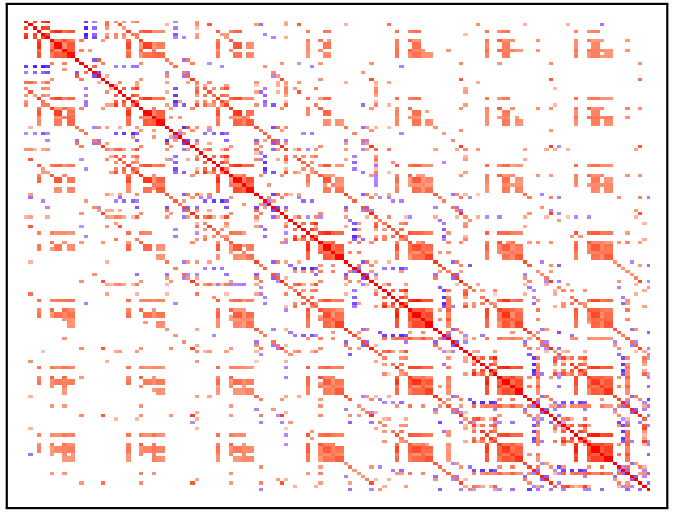
- $n < p$ for all the classes `shadow`: $n_{\text{shadow}} = 45$
- Changes in α produces the expected effect in λ selection and the amount of sparsity

Correlation

α = 0.05



α = 0.2



Some results - Wine data

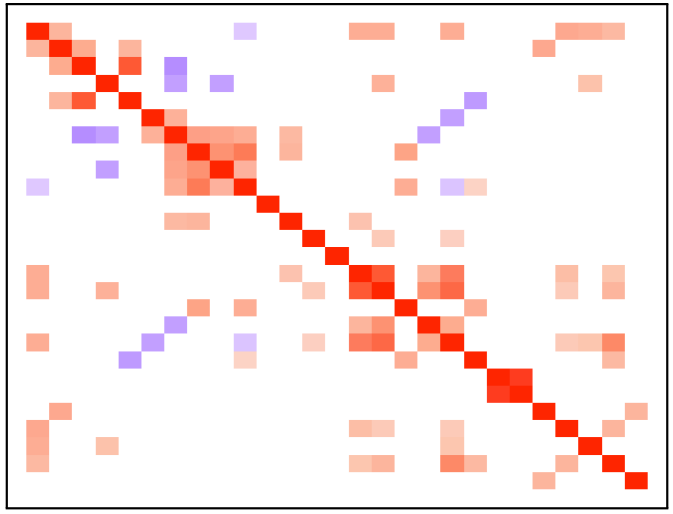
- $n = 178$ wine samples
  $G = 3$ different types
  $p = 27$ chemical measurements
- Sparse estimates might lead to easier interpretation
- Data often considered in the clustering literature

  $\Downarrow$

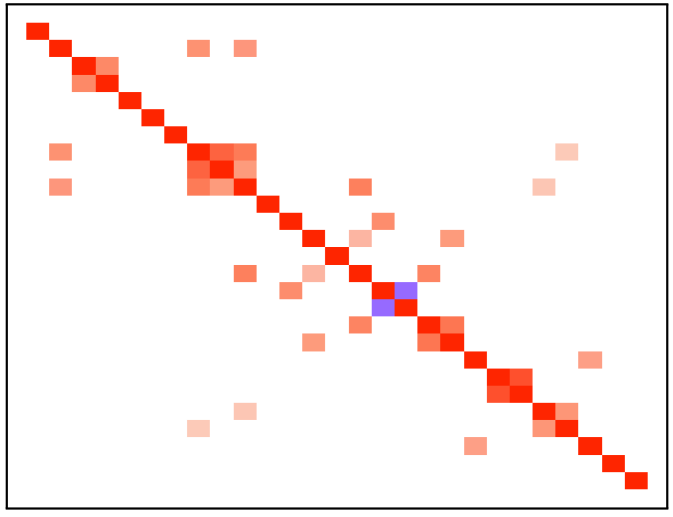
  Possible to embed the estimation strategy into classification tools?

Correlation

Barolo



Barbera



Wrapping up

We introduce a method which is able to produce reliable sparse estimate of covariance matrices even when $p > n$

- Different from standard CL penalized strategies
  $\Longrightarrow$ sparsity is a byproduct of clever sub-scores selection
- Penalty on the sub-likelihoods rather than on the parameters
  $\Longrightarrow$ not introducing bias in the final estimates

The method enjoys asymptotically model selection consistency and shares ML estimator properties

Issues and future directions

- Possible to extend it to precision matrix estimation?
- Possible to embed it into predictive tools?
- Closer look to efficient computational solutions

Some references

Casa, A., Ferrari, D. & Huang, Z. (2024+).
High-dimensional covariance estimation by pairwise likelihood truncation
arXiv:2407.07717

Other relevant references

- Bien, J. & Tibshirani, R. J. (2011). Sparse estimation of a covariance matrix. *Biometrika*, 98(4), 807-820.
- Huang, Z. & Ferrari, D. (2024). Fast construction of optimal composite likelihood. *Statistica Sinica*, 24, 47-66
- Lindsay, B. G., Yi, G. Y. & Sun, J. (2011). Issues and strategies in the selection of composite likelihood. *Statistica Sinica*, 71-105.
- Pourahmadi, M. (2013). *High-dimensional covariance estimation: with high-dimensional data*. John Wiley & Sons.
- Varin, C., Reid, N. & Firth, D. (2011). An overview of composite likelihood methods. *Statistica Sinica*, 5-42.