NOVELIST estimator for large covariance matrix

THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

Na Huang† and Piotr Fryzlewicz‡

Department of Statistics, London School of Economics, UK

E-mail: n.huang1@lse.ac.uk†, p.fryzlewicz@lse.ac.uk‡

ABSTRACT

We propose a NOVEL Integration of the Sample and Thresholded covariance estimator (NOVELIST) to estimate large covariance matrix. It is shrinkage of the sample covariance towards a general thresholding target, especially soft or hard thresholding estimators. The benefits of NOVELIST include simplicity, ease of implementation, and the fact that its application avoids eigenanalysis. We obtain an explicit convergence rate in the operator norm over a large class of covariance matrices when dimension p and sample size n satisfy $\log p/n \to 0$. Further we show the rate is a trade-off between sparsity, shrinkage intensity, thresholding level, dimension and sample size under different covariance structures. The simulation results will be presented and comparison with other competing methods will also be given.

INTRODUCTION

- Importance and Challenging

- Estimating covariance matrix Σ or precision matrix (inverse of covariance matrix) Σ^{-1} has always been an important part of multivariate analysis.
  - Financial risk management, Linear discriminant analysis (LDA), Principal component analysis (PCA)
  - Graphic networks, large portfolio selection, and so on
- In high-dimensional cases, where the dimension p is of the same order as the sample size n, or even much larger than n, estimating covariance or precision matrix is much more challenging.
  - When $p > n$, noninvertability of sample covariance matrix
  - Even p is of the same order as n, parameters needed to be estimated is $p(p-1)/2$ can be much larger than n.

- Existing Methods

- Impose some structure on the estimators
  - Bandable structure: banding or tapering (Bickel and Levina, 2008a, Furrer and Bengtsson, 2007)
  - Sparsity structure: thresholding (El Karoui, 2008; Bickel and Levina, 2008b), lasso (Friedman *et al.*, 2007), neighbourhood selection (Meinshausen and Buhlmann, 2006)
  - Factor model: POET (Fan, Liao and Mincheva, 2013)
- Shrinkage estimation
  - Shrinkage on sampe eigenvalues (linear or nonlinear) (Ledoit and Wolf, 2004 & 2012)
  - Shrinkage of sample covariance matrix towards other knowledge-added estimators (James-Stein estimator) (Ledoit and Wolf, 2003; Schaefer and Strimmer, 2005)

METHODS

- NOVELIST estimator

- Let $\boldsymbol{X} = (X_1, X_2, \cdots, X_p)^T$ be a p-dimensional random variables, distributed according to a distribution F, with $E\boldsymbol{X} = \mathbf{0}$, and $E\boldsymbol{X}\boldsymbol{X}^T = \Sigma = \{\sigma_{ij}\}$, then NOVELIST estimator is defined as $\hat{\Sigma}^N$

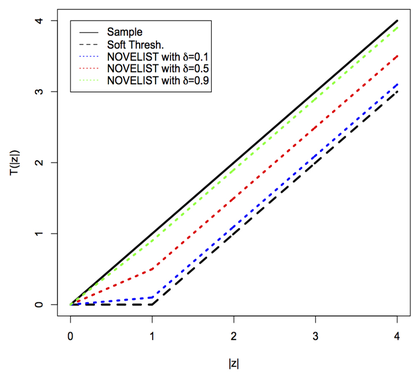
$$\hat{\Sigma}^N(\delta, \lambda) = \underbrace{(1-\delta)\hat{\Sigma}_{sam}}_{\text{pattern(nonsparse part)}} + \underbrace{\delta T(\hat{\Sigma}_{sam}, \lambda)}_{\text{noise(sparse part)}} \tag{1}$$

where
- $\hat{\Sigma}_{sam}$ is $p \times p$ sample covariance matrix of $\boldsymbol{X}$, $\hat{\Sigma}_{sam} = \frac{1}{n}\boldsymbol{X}^T\boldsymbol{X}$, n is sample size.
- δ is a weight or shrinkage intensity, which is usually within the range $[0, 1]$,
- λ is thresholding level, $\lambda \in (0, 1)$.
- $T(\hat{\Sigma}_{sam}, \lambda)$ is thresholding estimator, which obtains soft, hard or other thresholding estimator of $\hat{\Sigma}_{sam}$ with corresponding thresholding level specified by λ.

- NOVELIST estimator is a weighted combination of two parts. In high dimensional setting, the first part is a non-sparse, low-rank matrix, and the other part is a sparse matrix that is to ensure invertibility of the estimator.

- Illustration of NOVELIST

- NOVELIST operators with soft thresholding target $T(z) = (z - \text{Sign}(z) \cdot \lambda)1(|z| > \lambda)$ is shown as



- Motivation: link to ridge regression

- For linear regression

$$Y = \boldsymbol{X}\beta + \varepsilon, \tag{2}$$

possibly with $p > n$. Consider a criterion

$$||Y - \boldsymbol{X}(1-\delta)\beta||_2^2 + \delta(1-\delta)\beta^T f(\boldsymbol{X}^T\boldsymbol{X})\beta \tag{3}$$

where $f(\boldsymbol{X}^T\boldsymbol{X})$ is any modification of the matrix $\boldsymbol{X}^T\boldsymbol{X}$, and δ is a constant. To minimising (3), we get

$$\hat{\beta} = [(1-\delta)\boldsymbol{X}^T\boldsymbol{X} + \delta f(\boldsymbol{X}^T\boldsymbol{X})]^{-1}\boldsymbol{X}^T Y \doteq E^{-1}\boldsymbol{X}^T Y. \tag{4}$$

- If $f(\boldsymbol{X}^T\boldsymbol{X}) = \mathbb{I}$, (3) is actually a rescaled and weighted ridge regression and the part E reduces to shrinkage estimator with diagonal target.
- If $f(\boldsymbol{X}^T\boldsymbol{X}) = T(\boldsymbol{X}^T\boldsymbol{X}, \lambda)$, the part E is in the form of NOVELIST estimator.
- What to penalise for NOVELIST estimator? It places penalty on the *products* $\beta_i\beta_j$ of those coefficients of β for which the sample covariance between X_i and X_j exceeds a certain threshold λ. In other words, if the covariance (correlation) is high, we penalise the product of the corresponding β's, i.e. hope that not both are simultaneously large. While the others β's are not penalised.

CONSISTENCY AND PROPERTIES

- Consistency

- A uniformity class of covariance matrices introduced by Bickel and Levina (2008b)

$$\mathcal{U}_\tau(q, c_0(p), M) = \left\{\Sigma : \sigma_{ii} \le M, \sum_{j=1}^{p} |\sigma_{ij}|^q \le c_0(p), \text{for all } i\right\}, \quad \text{for } 0 \le q < 1. \tag{5}$$

- To ensure invertability, $\mathcal{U}_\tau(q, c_0(p), M)$ is narrowed down to

$$\mathcal{U}_\tau(q, c_0(p), M, \varepsilon_0) = \{\Sigma : \Sigma \in \mathcal{U}_\tau(q, c_0(p), M) \text{ and } \lambda_{min}(\Sigma) \ge \varepsilon_0 > 0\}, \tag{6}$$

- **Proposition 1** *If $T(z, \lambda)$ and F satisfies $\int_0^\infty exp(\lambda t)dG_j(t) < \infty$, for $0 < |\lambda| < \lambda_0$, for some $\lambda_0 > 0$, where G_j is the cdf of X_{1j}^2. Then, uniformly on $\mathcal{U}_\tau(q, c_0(p), M)$, for sufficiently large M', if $\lambda = M'\sqrt{\frac{\log p}{n}} = o(1)$,*

$$||\hat{\Sigma}^N - \Sigma|| = \underbrace{O_p((1-\delta)p\sqrt{\frac{\log p}{n}})}_{(A)} + \underbrace{O_p(\delta c_0(p)(\frac{\log p}{n})^{(1-q)/2})}_{(B)}. \tag{7}$$

Moreover, the same result holds for $||(\hat{\Sigma}^N)^{-1} - \Sigma^{-1}||$ uniformly on $\mathcal{U}_\tau(q, c_0(p), M, \varepsilon_0)$.

- δ and rate of convergence

- To find the optimal final rate of convergence, we make (A) equal to (B), which yields,

$$\tilde{\delta} = \frac{p(\frac{\log p}{n})^{q/2}}{c_0(p) + p(\frac{\log p}{n})^{q/2}} \tag{8}$$

(a) If $c_0(p) = o_p(p(\frac{\log p}{n})^{q/2})$, we have $\tilde{\delta} \to 1$, as $n \to \infty$
(b) If $p(\frac{\log p}{n})^{q/2} = o_p(c_0(p))$, we have $\tilde{\delta} \to 0$, as $n \to \infty$
(c) If $p(\frac{\log p}{n})^{q/2} \asymp c_0(p)$, we have $\tilde{\delta} \in (0, 1)$, as $n \to \infty$

The corresponding final rate of convergence will be (a) $c_0(p)(\frac{\log p}{n})^{(1-q)/2}$, (b) $p\sqrt{\frac{\log p}{n}}$, and (c) a rate between the two above.

- **Scenario 1** $q = 0$. *(Example: β-sparse covariance matrix)*

  Corollary 1 *Under Scenario 1 and conditions of Proposition 1, $\tilde{\delta}$ is a function of p only. And the following holds.*
  1. $\tilde{\delta} \in (0, 1)$ *for fixed* p.
  2. $\tilde{\delta} \to 1$, *as* $p \to \infty$, *as long as* $c_0(p) = o_p(p)$.

- **Scenario 2** $q \ne 0$, $c_0(p) < \infty$ *as* $p \to \infty$. *(Examples: MA and AR covariance matrix)*

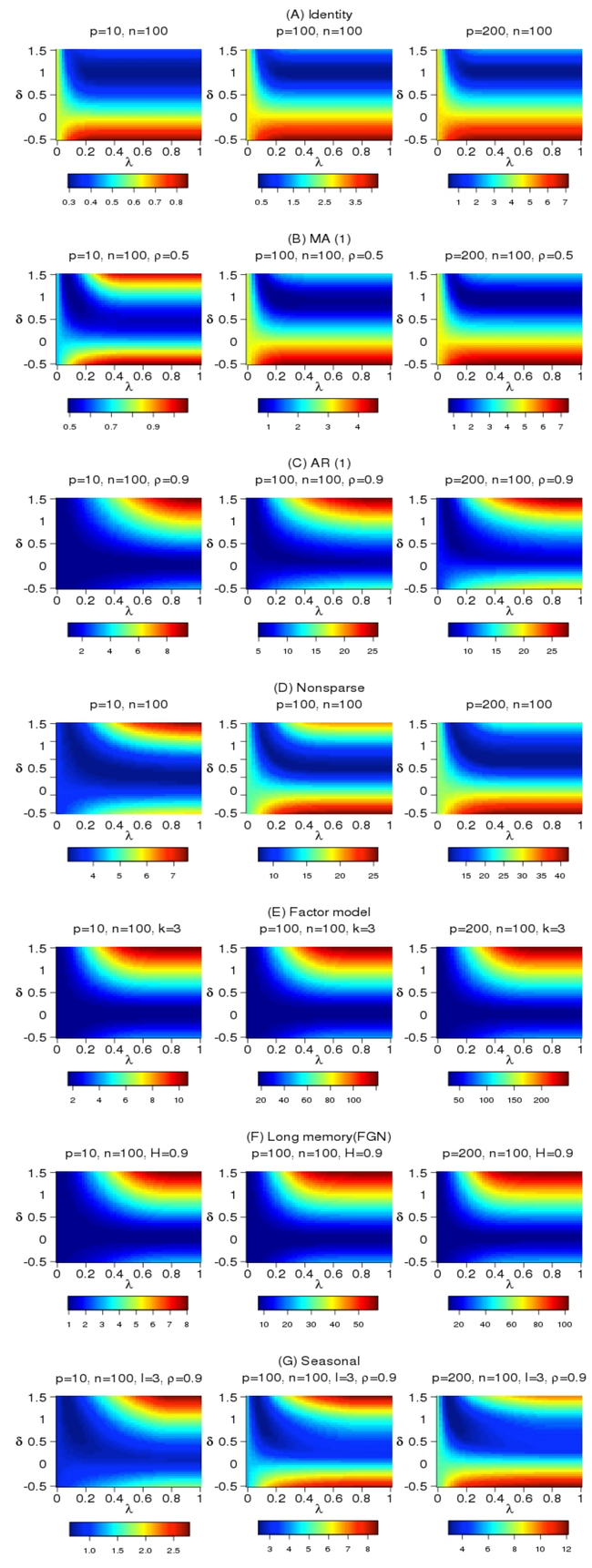
  Corollary 2 *Under Scenario 2 and conditions of Proposition 1, $\tilde{\delta}$ is a function of p and n. Assume $\log p = C_1 n^\alpha$, $0 < \alpha < 1$, as $n \to \infty$, the following holds.*
  1. $\tilde{\delta} \to 0$, *if* $p = o_p(n^{\frac{(1-\alpha)q}{2}})$.
  2. $\tilde{\delta} \to 1$, *if* $n = o_p(p^{\frac{2}{(1-\alpha)q}})$.
  3. $\tilde{\delta} \in (0, 1)$, *if* $p \asymp n^{\frac{(1-\alpha)q}{2}}$.

- **Scenario 3** $q \ne 0$, $c_0(p) = \infty$ *as* $p \to \infty$ *(Example: long memory covariance matrix)*

  Corollary 3 *Under Scenario 3 and conditions of Proposition 1, if the true covariance entries satisfy $\sigma_{i,j} = |i-j|^{-\gamma}$, $0 \le \gamma \le 1$, then $\tilde{\delta}$ is a function of p and n. Assume $\log p = C_1 n^\alpha$, $0 < \alpha < 1$ as $n \to \infty$, the following holds.*
  1. $\tilde{\delta} \to 0$, *if* $p = o_p(n^{\frac{1-\alpha}{2\gamma}})$.
  2. $\tilde{\delta} \to 1$, *if* $n = o_p(p^{\frac{2\gamma}{1-\alpha}})$.
  3. $\tilde{\delta} \in (0, 1)$, *if* $p \asymp n^{\frac{1-\alpha}{2\gamma}}$.

SIMULATION STUDY

- Figure 1: Image plots of operator norm error for NOVELIST estimators with different δ and λ under 7 covariance structures



- Table 1: Operator norm error to Σ for different estimators based on optimal results (50 replications)

	$p=10$ $\hat{\Sigma}$	$n=100$ T_s	B	$\hat{\Sigma}^N_{opt}$
(A) Identity	0.912	0.304	0.304	0.304
(B) MA(1)	0.983	0.567	0.391	0.456
(C) AR(1)	1.487	1.287	1.191	1.251
(D) Nonsparse	4.539	3.303	3.493	2.960
(E) Factor	1.716	1.370	1.808	0.982
(F) FGN	1.377	1.251	1.353	0.907
(G) Seasonal	0.796	0.632	0.695	0.608
	$p=100$ $\hat{\Sigma}$	$n=100$ T_s	B	$\hat{\Sigma}^N_{opt}$
(A) Identity	2.867	0.423	0.423	0.423
(B) MA(1)	2.976	0.711	0.640	0.712
(C) AR(1)	4.053	2.393	1.598	2.363
(D) Nonsparse	16.942	8.373	9.017	7.922
(E) Factor	22.998	20.733	21.378	18.870
(F) FGN	7.581	7.581	7.508	6.888
(G) Seasonal	2.168	2.092	2.715	2.008
	$p=200$ $\hat{\Sigma}$	$n=100$ T_s	B	$\hat{\Sigma}^N_{opt}$
(A) Identity	4.709	0.600	0.600	0.600
(B) MA(1)	4.901	0.836	0.745	0.827
(C) AR(1)	6.600	2.920	1.782	2.864
(D) Nonsparse	26.883	11.835	12.437	10.983
(E) Factor	29.998	27.733	28.378	25.950
(F) FGN	15.732	15.732	15.732	14.840
(G) Seasonal	6.897	3.153	3.014	3.081

Where $\hat{\Sigma}$ is sample covariance, T_s is soft thresholding, B is banding, $\hat{\Sigma}^N_{opt}$ is optimal NOVELIST.

- Table 2: Operator norm error to Σ for data-driven estimators (50 replications)

	$p=10$ $\hat{\Sigma}$	$n=100$ S	P	$\hat{\Sigma}^N_{cv}$
(A) Identity	0.912	0.029	1.815	0.304
(B) MA(1)	0.983	0.465	1.554	0.460
(C) AR(1)	1.487	1.061	1.823	1.454
(D) Nonsparse	4.539	3.296	6.870	3.470
(E) Factor	1.716	1.431	0.394	1.251
(F) FGN	1.377	1.155	1.594	1.109
(G) Seasonal	0.796	0.675	1.089	0.690
	$p=100$ $\hat{\Sigma}$	$n=100$ S	P	$\hat{\Sigma}^N_{cv}$
(A) Identity	2.867	0.037	3.768	0.423
(B) MA(1)	2.976	0.646	3.781	0.726
(C) AR(1)	4.053	2.958	4.536	2.368
(D) Nonsparse	16.942	8.542	20.943	8.502
(E) Factor	22.998	25.478	17.951	19.820
(F) FGN	7.581	7.571	8.228	8.282
(G) Seasonal	2.168	2.325	2.599	2.284
	$p=200$ $\hat{\Sigma}$	$n=100$ S	P	$\hat{\Sigma}^N_{cv}$
(A) Identity	4.709	0.215	5.607	0.602
(B) MA(1)	4.901	0.754	5.733	0.902
(C) AR(1)	6.600	3.870	7.119	2.901
(D) Nonsparse	26.883	12.128	30.708	11.01
(E) Factor	29.998	32.478	14.382	26.371
(F) FGN	15.732	18.457	17.192	15.502
(G) Seasonal	6.897	4.192	7.497	3.479

Where S is shrinkage estimator with diagonal target, P is POET estimator, $\hat{\Sigma}^N_{cv}$ is cross validation NOVELIST.

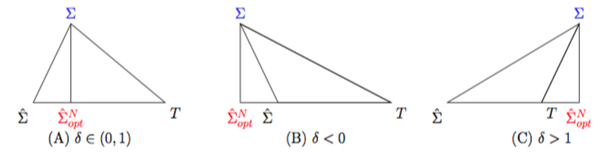
DISCUSSION AND SUMMARY

- Advantage and disadvantage

- Simplicity: easy of implementation, avoid eigenanalysis.
- Flexibility: work well in a wide range of underlying covariance matrix and in high dimensional setting, which is benefit from balancing of shrinkage intensity and thresholding level.
- Performance: better performances than other competing methods under several models.
- Computationally expensive: two parameter cross validation.

- δ outside $(0, 1)$

- Notice that for certain cases, the empirical optimal choice δ^* is beyond the theoretical boundary $(0, 1)$. The diagram explanation is as below. If the thresholding target is actually misspecification, the shrinkage intensity δ could be < 0 in order to make NOVELIST estimator far away from the target, as shown in the middle graph.



References

[1] Bickel, P. J. and Levina, E. (2008a). "Regularized estimation of large covariance matrices", *The Annals of Statistics*, **36**, 199-227.

[2] Bickel, P. J. and Levina, E. (2008b). "Covariance Regularization by Thresholding", *The Annals of Statistics*, **36**, 2577-2604.

[3] Drothman, A. J., Levina, E, and Zhu, J. (2009). "Generalized Thresholding of Large Covariance Matrices", *Journal of American Statistical Association*, **104**, 177-186.

[4] Fan, J. Q., Liao, Y. and Mincheva, M. (2013). "Large Covariance Estimation by Thresholding Principal Orthogonal Complements", *Journal of Royal Statistical Society. Series B*, **75**(4), 603-680.

[5] Ledoit, O. and Wolf, M. (2003). "Improved estimation of the covariance matrix of stock returns with an application to portfolio selection", *Journal of Empirical Finance*, **10**,603-621.

[6] Ledoit, O. and Wolf, M.(2004) "A well-conditioned estimator for large-dimensional covariance matrices". *Journal Multivariate Analysis*, **88**, 365-411.

[7] Ledoit, O. and Wolf, M.(2012) "Nonlinear Shrinkage and Estimation of Large-dimensional Covariance Matrices". *The Annals of Statistics*, **40**(2), 1024-1060.

[8] Schäfer, J. and Strimmer, K. (2005) "A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomic", *Statistical Applications in Genetics and Molecular Biology*, **4**(1), Article no. 32..