High-dimensional covariance estimation in YUIMA package

Yuta Koike

Mathematics and Informatics Center, University of Tokyo

June 27, 2019

The 3nd YUIMA Conference
@ Brixen-Bressanone

Background

- $Y_t = (Y_t^1, \dots, Y_t^d)^\top$ ($t \in [0,1]$): d-dimensional semimartingale
- Aim Estimating the quadratic covariation matrix

$$\Sigma_Y := [Y, Y]_1 = ([Y^i, Y^j]_1)_{1 \le i,j \le d}$$

  from discrete observation data of Y
  - The observation data may be noisy and/or non-synchronously observed
- Σ_Y can be considered as a kind of "(conditional) covariance matrix" and thus plays an important role in financial risk management

Background

- In high-frequency financial econometrics, this subject has been extensively studied in the past two decades, and a number of statistical methods have been proposed
    - Survey: K. & Yoshida (2019) "Covariance estimation and quasi likelihood analysis", to appear in Routledge handbook
- The R package `yuima` offers the function `cce` to implement some of those methods with a simple command
    - Cumulative Covariance Estimator
    - Currently, totally 12 methods (plus various options) have been implemented
    - See also the function `lmm` which implements the *local method of moments estimator* from Bibinger et al. (2014), which is theoretically the best possible (i.e. asymptotically efficient) in some situations

Background

- The aim of this talk is to discuss how we can take account of the **high-dimensionality**, i.e. the case with (extremely) large d
- Ignoring computational cost, we can use the function `cce` in *any* dimension, but ...
  1. the higher the dimension, the less accurate the estimates
  2. the estimated covariance matrices might be singular

Background

- The non-singularity of estimated covariance matrices is particularly important in financial applications
  - In the recent years, the "smart beta", which is a class of the alternative indices to traditional ones (such as S&P500), has attracted financial institutions
  - Some smart beta indices are constructed via solving optimization problems using covariance matrices as an input
    - ★ Ex. The *minimum volatility index* determines its weight vector $\boldsymbol{w} = (w_1, \dots, w_d)^\top \in \mathbb{R}^d$ by solving the following optimization problem:

$$\min_{\boldsymbol{w}} \boldsymbol{w}^\top \Sigma_Y \boldsymbol{w} \qquad \text{subject to } \sum_{j=1}^{d} w_j = 1$$

      (in practice, we often impose additional constraints such as short selling constraint $w_j \geq 0$ $(j = 1, \dots, d)$)
  - Minimum volatility type indices have already been sold by some index venders (such as MSCI)

Factor structure

- In financial applications, it is important to take account of the **factor structure** of financial data, which also serves as resolving issues of the high-dimensionality
- The factor structure of financial data is suggested by **both** theory and empirical results
  - Theory: CAPM, Arbitrage pricing theory, ...
  - Empirical: Fama-French 3-factor model, ...

Factor structure

- Specifically, suppose that we have a known factor process $X = (X^1, \dots, X^r)^\top$ and consider the following continuous-time factor model:

$$Y = \beta X + Z.$$

  - β: factor loading (non-random $d \times r$ matrix)
  - $Z = (Z^1, \dots, Z^d)^\top$: residual process
  - We suppose that both X and Z are semimartingales and satisfy $[Z^j, X^k] \equiv 0$ for $j = 1, \dots, d$ and $k = 1, \dots, r$
- Even if we do not know the factor process, we can (at least formally) construct a pseudo factor process by PCA
  - In some situations, this procedure has been formally validated; see Aït-Sahalia and Xiu (2017); Dai et al. (2019); Fan and Kim (2018); Pelger (2019)

Factor structure

- We are interested in estimating Σ_Y based on observation data for X and Y with taking account of the factor structure
  - We can compute traditional estimators $\hat{\Sigma}_{Y,n}$ for $\Sigma_Y := [Y, Y]_1$, $\hat{\Sigma}_{X,n}$ for $\Sigma_X := [X, X]_1$ and $\hat{\Sigma}_{YX,n}$ for $\Sigma_{YX} := [Y, X]_1$ by e.g. cce
- By assumption Σ_Y is written as follows:

$$\Sigma_Y = \beta \Sigma_X \beta^\top + \Sigma_Z. \tag{1}$$

- Provided that Σ_X is a.s. invertible, we can write β as $\beta = \Sigma_{YX} \Sigma_X^{-1}$
- Hence we can naturally estimate β by $\hat{\beta}_n := \hat{\Sigma}_{YX,n} \hat{\Sigma}_{X,n}^{-1}$, provided that $\hat{\Sigma}_{X,n}$ is invertible
  - The invertibility of $\hat{\Sigma}_{X,n}$ is usually not problematic as long as the number of factors r is sufficiently small compared to the sample size

Factor structure

- Then, from (1), Σ_Z is estimated by

$$\hat{\Sigma}_{Z,n} := \hat{\Sigma}_{Y,n} - \hat{\beta}_n \hat{\Sigma}_{X,n} \hat{\beta}_n^\top$$

- Due to the high-dimensionality, $\hat{\Sigma}_{Z,n}$ might be a poor estimator for Σ_Z
  - In particular, $\hat{\Sigma}_{Z,n}$ might NOT be positive definite even when Σ_Z is
  - In contrast, one can show that $\hat{\beta}_n \hat{\Sigma}_{X,n} \hat{\beta}_n^\top$ is a "good" estimator for $\beta \Sigma_X \beta^\top$ even in high-dimensional situations under appropriate assumptions
- To overcome this issue, we need to "regularize" $\hat{\Sigma}_{Z,n}$ in an appropriate way
  - In the context of HF econometrics, this approach was first studied in Fan et al. (2016)

Factor structure

- Given a regularized version $\tilde{\Sigma}_{Z,n}$ of $\hat{\Sigma}_{Z,n}$, we can estimate Σ_Y by

$$\tilde{\Sigma}_{Y,n} := \hat{\beta}_n \hat{\Sigma}_{X,n} \hat{\beta}_n^\top + \tilde{\Sigma}_{Z,n}$$

- If $\tilde{\Sigma}_{Z,n}$ is positive definite, $\tilde{\Sigma}_{Y,n}$ is also positive definite (as long as $\hat{\Sigma}_{X,n}$ is positive semi-definite)
- There are a number of approaches on how to regularize a covariance matrix estimator
- Some of them directly regularize estimated covariance matrices and do not use the particular structure of a model (at least formally), which are appropriate to our purpose

Implementation in YUIMA: The function `cce.factor`

- In summary, there are basically three ingredients in the estimation procedure described above, and each ingredient contain several options according to situations
  1. Covariance estimation: Non-synchronous and/or noisy and/or jumps
  2. Factor modeling: No/known/unknown
  3. Regularization: How to regularize the residual covariance matrix
- The function `cce.factor`, which will be implemented in future versions of the package `yuima`, systematically combines these three ingredients and provides several options for each one

Description of cce.factor

```
cce.factor(yuima, method = "HY", factor = NULL, PCA = FALSE
    , nfactor = "interactive", regularize = "tapering",
    taper, group = 1:(dim(yuima) - length(factor)),lambda =
    "bic", weight = TRUE, nlambda = 10, ratio, N, thr.type =
     "soft", thr = NULL, tau = NULL, par.alasso = 1, par.
    scad = 3.7, frequency = 300, utime, ...)
```

- method indicates the method used in cce
- factor indicates which components of yuima are factors
- PCA Use PCA to construct factors?
- regularize indicates the regularization method applied to the residual covariance matrix; four methods are currently available (tapering, glasso, eigen.cleaning and thresholding)
- Other arguments are options for each method

Description of `cce.factor`

- Brief description of each regularization method
  - `tapering`: Taking the entry-wise product of $\hat{\Sigma}_{Z,n}$ and some pre-determined $d \times d$ matrix $\mathcal{T}_d$: $\tilde{\Sigma}_{Z,n} := \hat{\Sigma}_{Z,n} \circ \mathcal{T}_d$ ($\circ$ denotes the entry-wise product)
  - `glasso`: ℓ_1-penalized Gaussian MLE for the inverse of Σ_Z
  - `eigen.cleaning`: shrinking eigenvalues of $\hat{\Sigma}_{Z,n}$; here the procedure described in Hautsch et al. (2012) is implemented
  - `thresholding`: The entries below a pre-determined threshold are set to 0 (hard thresholding) or shrunk toward 0 (soft thresholding)

Description of `cce.factor`

- Theoretical validity of each regularization method in the HF context
- Factors are known
  - `tapering`/`thresholding`: Fan et al. (2016) for the synchronous & non-noisy case and Dai et al. (2019) for the non-synchronous & noisy case
  - `glasso`: Brownlees et al. (2018) (see also K. (2019))
  - `eigen.cleaning`: No theoretical validity
- Factors are unknown
  - `tapering`/`thresholding`: Aït-Sahalia and Xiu (2017) for the synchronous & non-noisy case and Dai et al. (2019) for the non-synchronous & noisy case
  - `glasso`: No result is available
  - `eigen.cleaning`: No theoretical validity

What we can do by `cce.factor`

cce method		**factor**	**regularization**
HY	QMLE	No	tapering
PHY	SIML	Known	glasso
MRC	THY	Unknown	eigen.cleaning
TSCV	PTHY		thresholding
GME	SRC		
RK	SBPC		

12 × 3 × 4 = **144**

Some simulation results

- Model for the factor process X: We set $r = 3$ and

$$dX_t^j = \mu_j dt + \sqrt{v_t^j} dW_t^j,$$
$$dv_t^j = \kappa_j(\theta_j - v_t^j)dt + \eta_j \sqrt{v_t^j} \left(\rho_j dW_t^j + \sqrt{1-\rho_j^2} d\widetilde{W}_t^j \right), \quad j = 1, 2, 3,$$

  where $W^1, W^2, W^3, \widetilde{W}^1, \widetilde{W}^2, \widetilde{W}^3$ are independent standard Wiener processes
- We set $\kappa = (3, 4, 5), \theta = (0.09, 0.04, 0.06), \eta = (0.3, 0.4, 0.3), \rho = (-0.6, -0.4, -0.25)$ and $\mu = (0.05, 0.03, 0.02)$
- The entries of the loading β are independently drawn as

$$\beta^{i1} \overset{i.i.d.}{\sim} \mathcal{U}[0.25, 2.25], \quad \beta^{i2}, \beta^{i3} \overset{i.i.d.}{\sim} \mathcal{U}[-0.5, 0.5]$$

  ($\mathcal{U}[a, b]$: the uniform distribution on $[a, b]$)

Some simulation results

- Model of the residual process Z: d-dimensional Wiener process with covariance matrix Q
- We consider the following two designs for Q

  Design 1 Q is a block diagonal matrix with 10 blocks of size $(d/10) \times (d/10)$. Each block has diagonal entries independently generated from $\mathcal{U}[0.2, 0.5]$ and a constant correlation of 0.25.

  Design 2 We simulate a Chung-Lu random graph $\mathcal{G}$ and set $Q := (\mathsf{E}_d + \boldsymbol{D} - \boldsymbol{A})$, where $\boldsymbol{D}$ and $\boldsymbol{A}$ are respectively the degree and adjacent matrices of the random graph $\mathcal{G}$. We use the same parameters for the Chung-Lu random graph as in the simulation study of Barigozzi et al. (2018).

Some simulation results

- $d = 500$
- We observe the process Y at the equi-spaced sampling times $t_i = i/n$ ($i = 0, 1, \dots, n$) on the interval $[0, 1]$ and the realized covariance matrices are used as the estimators $\hat{\Sigma}_{Y,n}, \hat{\Sigma}_{X,n}$ and $\hat{\Sigma}_{YX,n}$
- We vary n as $n \in \{78, 130, 195, 390, 780\}$
- Based on 10,000 Monte Carlo iterations for each scenario
- Regularization methods

NO	No regularization
glasso	Graphical Lasso
wglasso	Weighted graphical Lasso (graphical Lasso based on the correlation matrix)
tapering	Tapering with $\mathcal{T}_d = (1_{\{\Sigma_Z^{ij} \neq 0\}})_{1 \leq i,j \leq d}$ (only for Design 1)
eigen	Eigen cleaning method proposed in Hautsch et al. (2012)

Some simulation results

Table 1: Estimation accuracy of different methods in Design 1

measure	n	NO	glasso	wglasso	tapering	eigen
$\vert\vert\vert\hat{\Sigma}_Y^{-1} - \Sigma_Y^{-1}\vert\vert\vert_2$	78	6.576	3.419	3.420	138.442	23.269
	130	6.508	3.193	3.193	28.384	20.187
	195	6.480	3.094	3.097	14.307	18.508
	390	203.038	2.133	2.100	6.446	16.545
	780	93.354	1.782	1.693	3.562	15.335
$\vert\vert\vert\hat{\Sigma}_Y - \Sigma_Y\vert\vert\vert_2$	78	21.771	21.829	21.829	21.477	21.782
	130	16.919	17.184	17.184	16.693	16.914
	195	13.844	14.287	14.289	13.656	13.840
	390	9.762	9.991	9.959	9.628	9.759
	780	6.869	7.031	6.978	6.772	6.867

$|||\cdot|||_2$ denotes the spectral norm. The Moore-Penrose generalized inverse is used when $\hat{\Sigma}_Y$ is singular.

Some simulation results

Table 2: Estimation accuracy of different methods in Design 2

measure	n	NO	glasso	wglasso	eigen
	78	17.805	7.857	7.843	16.847
	130	17.798	7.954	7.866	16.835
$\|\|\|\hat{\Sigma}_Y^{-1} - \Sigma_Y^{-1}\|\|\|_2$	195	17.752	8.006	7.742	16.832
	390	87.239	8.059	7.416	16.823
	780	55.619	8.065	6.072	16.809
	78	27.907	27.707	27.708	27.729
	130	21.552	21.397	21.399	21.413
$\|\|\|\hat{\Sigma}_Y - \Sigma_Y\|\|\|_2$	195	17.569	17.447	17.449	17.462
	390	12.368	12.284	12.284	12.298
	780	8.722	8.665	8.664	8.678

$|||\cdot|||_2$ denotes the spectral norm. The Moore-Penrose generalized inverse is used when $\hat{\Sigma}_Y$ is singular.

Conclusions and future work

- We overview the recent studies on high-dimensional covariance estimation in high-frequency data
- We introduce the function `cce.factor` to systematically implement the methods proposed by those studies in the framework of YUIMA
- Future work
    1. Simulator for continuous factor models
        - ⋆ The diffusion case is straightforward. It becomes somewhat complicated when we introduce different types of jumps/hurst parameters to the factor and residual processes
    2. Implementing formal methods to select the number of factors
    3. Implementing statistical testing procedures
    4. Implementing additional regularization methods
    5. (Machine learning approach to select the “best” method)

References I

Y. Aït-Sahalia and D. Xiu. Using principal component analysis to estimate a high dimensional factor model with high-frequency data. *J. Econometrics*, 201:384–399, 2017.

M. Barigozzi, C. Brownlees, and G. Lugosi. Power-law partial correlation network models. *Electron. J. Stat.*, 12:2905–2929, 2018.

M. Bibinger, N. Hautsch, P. Malec, and M. Reiß. Estimating the quadratic covariation matrix from noisy observations: local method of moments and efficiency. *Ann. Statist.*, 42(4):80–114, 2014.

C. Brownlees, E. Nualart, and Y. Sun. Realized networks. *J. Appl. Econometrics*, 33(7): 986–1006, 2018.

C. Dai, K. Lu, and D. Xiu. Knowing factors or factor loadings, or neither? Evaluating estimators of large covariance matrices with noisy and asynchronous data. *J. Econometrics*, 208:43–79, 2019.

J. Fan and D. Kim. Robust high-dimensional volatility matrix estimation for high-frequency factor model. *J. Amer. Statist. Assoc.*, 113(523):1268–1283, 2018.

J. Fan, A. Furger, and D. Xiu. Incorporating global industrial classification standard into portfolio allocation: A simple factor-based large covariance matrix estimator with high-frequency data. *J. Bus. Econom. Statist.*, 34(4):489–503, 2016.

References II

N. Hautsch, L. M. Kyj, and R. C. Oomen. A blocking and regularization approach to high-dimensional realized covariance estimation. *J. Appl. Econometrics*, 27:625–645, 2012.

Y. Koike. De-biased graphical lasso for high-frequency data. Working paper. arXiv:1905.01494, 2019.

M. Pelger. Large-dimensional factor modeling based on high-frequency observations. *J. Econometrics*, 208:23–42, 2019.