Factorized Estimation of High-Dimensional Nonparametric Covariance Models

JIAN ZHANG[1] and JIE LI[1,2]

[1]*School of Mathematics, Statistics and Actuarial Science, University of Kent*
[2]*School of Mathematics and Computer Science, Dali University*

March 17, 2021

Abstract

Estimation of covariate-dependent conditional covariance matrix in a high-dimensional space poses a challenge to contemporary statistical research. The existing kernel estimators may not be locally adaptive due to using a single bandwidth to explore the smoothness of all entries of the target matrix function. In this paper, we propose a novel framework to address this issue, where we factorize the target matrix into factors and estimate these factors in turn by the kernel approach. The resulting estimator is further regularized by thresholding and optimal shrinkage. Under certain mixing and sparsity conditions, we show that the proposed estimator is well-conditioned and uniformly consistent with the underlying matrix even when the sample is dependent. Simulation studies suggest that the proposed estimator significantly outperforms its competitors in terms of integrated root-squared estimation error. We present an application to financial return data.

Keywords : high-dimensional nonparametric covariance models, kernel estimator, shrinkage and consistency, thresholding.

1 Introduction

Nonparametric estimation of covariate-dependent conditional covariance matrix $\Sigma(u)$ in covariance models is fundamental to contemporary scientific research including neuroimaging in neuroscience, disease mapping in health science, daily ozone concentration analysis in environmental science and asset portfolio risk analysis in finance, among others (**ledoitWellconditionedEstimatorLa**
However, most efforts in nonparametric covariance estimation suffer from a curse of dimensionality (**fanNonlinearTimeSeries2003**). For example, in asset portfolio risk analysis, modelling market-dependent co-volatility of p assets by use of historical return data over n consecutive months involves estimating $p(p+1)/2$ nonparametric curves (**famaCapitalAssetPricing2004**). The dataset we are studying in this paper contains historical returns of 75 assets over three time periods, namely before-financial-crisis, in-financial-crisis and after-financial-crisis with n equal to 84, 36 and 95 months respectively. Note that many more assets can be collected for investigation whereas the number of months n in a period is sometimes quite limited (**engleLargeDynamicCovariance2017**). When p is close to or larger than n, the

kernel covariance estimator proposed by **yinNonparametricCovarianceModel2010** can be degenerate or ill-conditioned with a high condition number. Hence, it cannot be reliably inverted to compute the precision matrix which is required in the above risk analysis. In literature, regularization of estimated covariances was often done by banding, thresholding, or truncating the number of the leading eigenvalues (**bickelCovarianceRegularizationThresholding2008yuanRe**
chenDynamicCovarianceModels2016 proposed a method (called DCM) to regularize the kernel covariance model by thresholding covariance entries. These authors pointed out that the resulting covariance estimator can still be ill-conditioned for finite samples, where an ad-hoc and small constant is required to add to its eigenvalues. These authors also established a consistency theory for their estimator when the sample is i.i.d. There are three main issues that arise when we use these existing methods. First, the performance of these methods can be compromised by employing the same smoothing bandwidth for all entries which have varying degrees of smoothness. In particular, under the sparse assumption, the covariance matrix function contains many zero entries which are in favour with an infinite large bandwidth and thus affect estimation of other nonzero entries if we use a single bandwidth for all entries. On the other hand, letting all the entries have their own bandwidths will generate $p(p+1)/2$ tuning constants to choose. The resulting estimator may not be an appropriate covariance matrix estimator as it can be negative definite for finite samples. Secondly, as the ad-hoc eigenvalues adjustment of **chenDynamicCovarianceModels2016** to the estimated matrix is not principle guided, it is desirable to explore an optimal shrinkage procedure. Finally, the existing asymptotic theory holds only for i.i.d. samples although, in most applications, the samples are dependent. For instance, in the above asset portfolio risk analysis, both market returns and asset returns are serially correlated time series.

In this paper, we propose a novel framework to address these issues. It is based on a variance-correlation factorization of $\Sigma(u)$ in the form of $\Sigma(u) = Q_0(u)C_0(u)Q_0(u)^T$, where $Q_0(u)$ is a diagonal matrix function composed by the square roots of the diagonal entries of $\Sigma(u)$ and $C_0(u)$ is the correlation matrix function. We further factorize $C_0(u)$ into the product of invertible band matrix factors of $\Sigma(u)$. In general, we choose band matrices which are less complex than $\Sigma(u)$. In the proposal, we first estimate these band matrices in turn with separate kernel bandwidths, followed by entry-wise thresholding on the resulting estimator of $C_0(u)$. Estimation of these band matrices with different bandwidths is expected to improve the flexibility of the proposal and thus to provide a more accurate estimator for $\Sigma(u)$. Intuitively, performing thresholding on estimated correlations is better than on covariances, since the variation of the estimated correlations is likely to be smaller and more homogenous than that of the estimated covariances. In fact, thresholding correlations has been proved adaptive to the variability of individual entries of covariance matrix (**yuanReproducingKernelHilbert2010**). Finally, a well-conditioned and optimal shrinkage estimator of $\Sigma(u)$ is derived by the principle of minimizing the Frobenius loss. In summary, the proposed framework differs from the DCM in using multiple factorization based bandwidths, thresholding correlations and taking into account a shrinkage effect. The proposal can be viewed as a nonparametric extension of the so-called DCC-GARCH approach (**engleLargeDynamicCovariance2017**), a popular technique for estimating a multivariate time series model.

To evaluate the performance of the new proposal, a set of simulation studies are conducted. The results demonstrate that the new proposal substantially outperforms its counterparts in terms of the Frobenius loss and other criteria. The proposed method is illustrated through an application to the analysis of monthly return data for a group of risky assets mentioned above. The analysis reports the following findings: (1) Some asset returns present a striking nonlinear departure from the Capital Asset Pricing Model (CAPM) (**famaCapitalAssetPricing2004**). (2) Both volatility and co-volatility of these

asset returns are market-dependent, see Figure 1 for more details. These two findings provide an empirical support for building a nonparametric CAPM for risk assessment and portfolio selection. We also establish an asymptotic theory for the new proposal: under some mixing and regularity conditions, the proposed estimator is asymptotically consistent with the underlying covariance matrix function even when the samples are dependent. In the procedure, the thresholding step ensures that the resulting estimator converges to the true covariance matrix with a good rate while the shrinkage step makes the resulting estimator not ill-conditioned even in finite samples. To prove the above theory, a dedicated concentration inequality different from **chenDynamicCovarianceModels2016** is employed for dependent samples. In particular, the proof for the convergence rate of the proposed shrinkage is non-trivial. Note that without the extra thresholding step, a standard shrinkage estimator is expected to have convergence rate of $\sqrt{p/(nh)}$, where h is the bandwidth in the kernel estimation (**ledoitWellconditionedEstimatorLargedimensional2004**). After adding the extra thresholding step in the shrinkage procedure, we show that the resulting estimator has a faster convergence rate $\sqrt{\log(p/h)/(nh)}$ than does the standard shrinkage if the underlying covariance matrix is sparse.

[Put Figure 1 here.]

The rest of the article is organized as follows. The proposed factorized estimators are introduced in Section 2. The corresponding algorithms are developed to determine the bandwidths in the related kernel smoothing as well as the levels of thresholding and shrinkage. The uniform consistency and the convergence rate of the proposed estimator are established with dependent samples in Section 3. In Section 4, simulation studies are conducted to evaluate the performance of the proposed method and compare it to the existing method. The proposed procedure is employed to analyse financial returns for a group of assets. We conclude with a discussion in Section 5. The proofs of asymptotic theory and further numerical results are delayed to the Supplementary Materials. Throughout this paper, we let $\lambda_{\min}(\cdot)$ and $\lambda_{\max}(\cdot)$ denote the minimum and maximum eigenvalues of a square matrix. For a vector x, let $\|x\|$ denote its Euclidean norm. For a square matrix $A = (a_{ik})_{p\times p}$, let $\|A\|_F = \sqrt{\mathrm{tr}(AA^T)/p}$, $\|A\| = \lambda_{\max}^{1/2}(AA^T)$, $\|A\|_{\max} = \max_{1\le i,k\le p}|a_{ik}|$ and $\|A\|_\infty = \max_{1\le i\le p}\sum_{k=1}^n |a_{ik}|$ denote its (size-normalized) Frobenius, spectral, max and ∞-norms. Let $\langle A, B\rangle = \mathrm{tr}(AB^T)/p$ be the inner product of square matrices A and B. Let $I(\cdot)$ denote an indicator function. Note that these norms satisfy $\|A\|_F \le \|A\| \le \|A\|_\infty \le \max_{1\le i\le p}\sum_{j=1}^p I(|a_{ik}|>0)\|A\|_{\max}$. Let $\mathrm{diag}(x)$ denote the diagonal matrix with diagonal entries made from the elements of x. Let $c\wedge d$ and $c\vee d$ denote the minimum and maximum of numbers c and d. Let I_p be a p-dimensional identity matrix.

2 Methodology

Let $Y = (Y_1, \ldots, Y_p)^T \in \mathbb{R}^p$ be a $p-$dimensional random vector and $U \in \mathbb{R}$ be the associated index random variable. We model the conditional mean and covariance matrix of Y given $U = u$ as $\boldsymbol{\mu}(u) = E[Y|U=u]$ and $\Sigma(u) = \mathrm{cov}(Y|U=u)$ respectively whose entries are assumed to be the unknown but smooth functions of u. Suppose that $(\mathbf{y}_i, u_i)_{i=1}^n$ with $\mathbf{y}_i = (y_{i1}, \ldots, y_{ip})^T$, are random observations from the population (Y, U), satisfying the equations

$$\mathbf{y}_i = \boldsymbol{\mu}(u_i) + \Sigma^{1/2}(u_i)\varepsilon_i, \quad i = 1, \ldots, n,$$

where $\boldsymbol{\mu}(u_i) = (\mu_1(u_i), \ldots, \mu_p(u_i))^T$ and $(u_i)_{i=1}^n$ is a dependent random sample of U. Also, given $(u_i)_{i=1}^n$, ε_i's are dependent on each other and with zero means and unity covariance matrices (i.e., $E[\varepsilon_i|u_i] = 0_p$, $\mathrm{cov}(\varepsilon_i|u_i) = I_p$ and $E[\varepsilon_i\varepsilon_j^T] \neq 0, i \neq j$).

Let $K(u)$ be a kernel density function, $K_h(u) = h^{-1}K(u/h)$ (the scaled kernel function with a bandwidth $h > 0$) and $w_{ih}(u) = K_h(u_i - u)/\sum_{k=1}^n K_h(u_k - u)$ (the weighting function). **yinNonparametricCovarianceModel2010** considered the following kernel estimators for $\boldsymbol{\mu}(\cdot)$ and $\Sigma(\cdot)$:

$$\begin{aligned} \hat{\boldsymbol{\mu}}(u) &= \sum_{i=1}^n w_{ih_a}(u)\mathbf{y}_i, \\ \hat{\Sigma}(u) &= \sum_{i=1}^n w_{ih}(u)(\mathbf{y}_i - \hat{\boldsymbol{\mu}}(u_i))(\mathbf{y}_i - \hat{\boldsymbol{\mu}}(u_i))^T \hat{=} (\hat{\sigma}_{kj}(u))_{1\le k,j\le p}, \end{aligned} \tag{1}$$

where h_a and h are bandwidths for mean and covariance matrix functions respectively.

2.1 The variance-correlation based approach

To improve the above covariance estimator, we consider a variance-correlation factorization in the form

$$\Sigma(u) = Q_0(u)C_0(u)Q_0(u), \tag{2}$$

where $Q_0(u) = \text{diag}(\Sigma(u))^{1/2}$ and $C_0(u) = Q_0(u)^{-1}\Sigma(u)Q_0(u)^{-1}$. The proposed variance-correlation based Q_0-procedure can be implemented in the following three steps.

Step 1: Estimate $Q_0(u)$ and $C_0(u)$. We first estimate the diagonal entries, $\hat{Q}_0(u) = \text{diag}(\hat{\sigma}_{kk}(u) : 1 \le k \le p)$ with a $Q_0(u)$-specified bandwidth $h = h_0$. Then, we standardize $\mathbf{y}_i, 1 \le i \le n$ by using $\hat{\boldsymbol{\mu}}(u_i)$ and $\hat{Q}_0(u_i)$:

$$\tilde{\mathbf{y}}_i = \hat{Q}_0^{-1}(u_i)(\mathbf{y}_i - \hat{\boldsymbol{\mu}}(u_i)), \quad 1 \le i \le n$$

and estimate $C_0(u)$ by

$$\hat{C}_0(u) = \sum_{i=1}^n w_{ih_0}(u)\tilde{\mathbf{y}}_i\tilde{\mathbf{y}}_i^T \tag{3}$$

with bandwidth h_0.

Step 2: Threshold $\hat{C}_0(u)$. Note that the dimension p is frequently larger than the local sample size nh. This results in a degenerate estimator $\hat{C}_0(u)$. Following **bickelCovarianceRegularizationThresh** we regularize the above correlation matrix estimator by thresholding its entries as follows:

$$\hat{C}_0^{(t)}(u) = \left(\hat{c}_{jk}(u)I\left(|\hat{c}_{jk}(u)| > t_0(u)\sqrt{\log(p/h)/(nh)}\right)\right)_{1\le j,k\le p},$$

where $\hat{c}_{jk}(u)$ is the (j,k)-th entry of $\hat{C}_0(u)$ and $I(\cdot)$ is an indicator function and $t_0(u)$ is a positive function of u. The above rate of thresholding is suggested by Theorem 2 in Section 3 below. Here, p/h is related to the dimension of an approximate parametric model to the original model: $[a, b]$ is partitioned into $(b-a)/h$ intervals in which the p-dimensional nonparametric model are approximated by a $p(b-a)/h$-dimensional step model. Note that unlike the covariance matrix, the correlation matrix is scale-invariant and with homogenous diagonals. Therefore, thresholding correlation matrix is expected to make less errors than does thresholding covariance matrix, in particular, when individual variances $\sigma_{kk}(u), 1 \le k \le p$ greatly differ from each other (**yuanReproducingKernelHilbert2010**). Using the above estimators, we construct a plug-in estimator of $\Sigma(u)$ in form

$$\hat{\Sigma}^{(t)}(u) = \hat{Q}_0(u)\hat{C}_0^{(t)}(u)\hat{Q}_0(u).$$

Step 3: Shrink $\hat{\Sigma}^{(t)}(u)$. In Section 3 below, under sparsity and regularity conditions, we show that under certain regularity conditions the above thresholded covariance estimator is consistent with the underlying covariance matrix function as n and p tend to infinity. However, for a finite sample, the proposed estimator may still be ill-conditioned. To ameliorate it, we propose to shrink $\hat{\Sigma}^{(t)}(u)$ to the identity matrix I_p, where the amount of shrinkage is optimized in terms of the data-driven Frobenius loss. There are other covariance shrinkage methods in the literature, but most of them were developed for estimating covariance models without covariates. See **jolliffePrincipalComponentAnalysis2002baiSpectralAnalysisLarge2010** and references therein. To find the optimal amount of shrinkage, we first consider a population version, namely a linear combination of I_p and $\hat{\Sigma}^{(t)}(u)$, $\Sigma^*(u) = \rho a I_p + (1-\rho)\hat{\Sigma}^{(t)}(u)$, whose expected Frobenius loss $E\|\Sigma^*(u) - \Sigma(u)\|_F^2$ attains the minimum with respect to $0 \le \rho \le 1$ and $a \in \mathbb{R}$. The resulting solutions depend on $\Sigma(u)$ as well as variability of $\hat{\Sigma}^{(t)}(u)$. Replacing these unknown quantities by their estimators, we obtain the following plug-in estimator of $\Sigma(u)$ with a data-driven optimal amount of shrinkage:

$$\hat{\Sigma}^{(st)}(u) = \frac{\hat{\beta}_p^2(u)}{\hat{\alpha}_p^2(u) + \hat{\beta}_p^2(u)} p^{-1}\mathrm{tr}(\hat{\Sigma}^{(t)}(u)) I_p + \frac{\hat{\alpha}_p^2(u)}{\hat{\alpha}_p^2(u) + \hat{\beta}_p^2(u)} \hat{\Sigma}^{(t)}(u), \tag{4}$$

where

$$\begin{aligned}
\hat{\alpha}_p^2(u) &= \left\| \hat{\Sigma}^{(t)}(u) - \langle \hat{\Sigma}^{(t)}(u), I_p \rangle I_p \right\|_F^2. \\
\hat{\beta}_p^2(u) &= \frac{1}{p} \sum_{j=1}^{p} \sum_{k=1}^{p} \sum_{i=1}^{n} w_{ih}^2(u)((y_{ij} - \hat{\mu}_j(u_i))(y_{ik} - \hat{\mu}_k(u_i)) - \hat{\sigma}_{jk}(u))^2 \\
&\quad \times I(|\hat{\sigma}_{jk}(u)| > t_0(u)\sqrt{\log(p/h)/(nh)}).
\end{aligned}$$

Note that $\hat{\alpha}_p^2(u)$ is a plug-in bias when we use $p^{-1}\mathrm{tr}(\hat{\Sigma}^{(t)}(u))I_p$ to estimate $\Sigma(u)$ while $\hat{\beta}_p^2(u)$ gauges the variability of $\hat{\Sigma}^{(t)}(u)$ as an estimator of $\Sigma(u)$. So estimator (4) is intended to strike a balance between variability and bias of covariance estimators. Our idea is general, which can be directly used to improve other nonparametric covariance matrix estimators including the DCM. See the Appendix A, the Supplementary Materials for the detailed derivation.

2.2 Effects of unknown zero-entries on bandwidth selection

In Step 1 above, we explore the smoothness of $C_0(u)$ with a single bandwidth. This may introduce a large bias to estimating non-zero entries when $C_0(u)$ contains many unknown zero-entries as illustrated by the following toy example.

Let $(n, p) = (250, 100)$. For simplicity, we assumed that both $\boldsymbol{\mu}(u)$ and Q_0 were known and estimated $C_0(u)$ by using a cross-validated (CV) kernel. To generate an i.i.d. sample $(\mathbf{y}_i, u_i)_{1\le i\le n}$, we first randomly drew $(u_i)_{1\le i\le n}$ from the uniform distribution over $[-0.95, 0.95]$. Then, given u_i, we defined $\Sigma(u_i)$ through its square root matrix $R(u_i) = (r_{kj})_{p\times p}$. For a pre-selected $\theta \in [0,1]$, we randomly selected $p^* = \lfloor p_0\theta \rfloor$ entries from the strictly lower triangle part of $R(u_i)$ and assigned zeros to them. To keep the symmetry of $R(u_i)$, we reflected these zero-entries to the upper triangle part of $R(u_i)$. For the remaining entries, for example, (k,j)th entry, we set $r_{kj}(u_i) = \exp(100u_i \sin(kj))\sin(\pi u_i)$. Finally, we obtained $C_0(u_i) = Q_0(u_i)^{-1}R(u_i)^2Q_0(u_i)^{-1}$. We calculated the sparsity index S_{C_0} defined as the proportion of zero-entries in $C_0(u_i)$. Given u_i and $C_0(u_i)$, we drew $\mathbf{y}_i$ from the multivariate normal $N(0, C_0(u_i))$.

For $\theta = 0.855, 0.91, 0.95, 0.97, 0.98, 0.985, 0.99, 0.995$, we used the above sampling procedure to obtain a sample for each case with the sparsity index S_{C_0} taking values around

$0.10, 0.40, 0.74,\ 0.88, 0.93, 0.95, 0.97, 0.98$ respectively. For each sample, we calculated the cross-validated kernel estimates for $C_0(u_i), 1 \le i \le n$ and obtained cross-validation values for bandwidth grid points, which resulted in a CV curve. Note that these curves may have different minimum values. To make them comparable, we divided these curves by their minimum values respectively. The plots, displayed in Figure 2, showed that when the sparsity index increased from 0.10 to 0.98, the curves became flat and the CV-selected bandwidth tended to infinity. This implied that the CV-based bandwidth selection was gradually dominated by zero-entries where there were many unknown zero-entries. This can be explained by the fact that the arithmetic average of $y_{ik}y_{ij}^T, i = 1, \ldots, n$ (equivalent to choosing an infinite large bandwidth) is an optimal unbiased estimate for the (k, j)th zero-entry, but not for non-zero entries.

[Put Figure 2 here.]

To assess the zero-entry effect with multiple samples, we performed the above sampling procedure 90 times, obtaining 90 independent datasets. Applying the single-bandwidth kernel procedure to each sample, for each u_i, we calculated the Frobenius loss between the estimated and the true values for all zero off-diagonal entries and for all non-zero off-diagonal entries respectively. We calculated the average Frobenius losses over these u_i's. Box-plots of these average Frobenius losses were made for each $\theta \in \{0.855, 0.91, 0.95, 0.97, 0.98, 0.985, 0.99, 0.995\}$ as shown in Figure 2. The result indicated that the average Frobenius losses did increase for non-zero entries and decreasing for zero-entries when the sparsity index was increasing.

[Put Figure 3 here.]

To tackle the problem, we factorize $C_0(u)$ further. For example, we can construct a new estimator $\hat{C}_0(u) = \hat{Q}_1(u)\hat{C}_1(u)\hat{Q}_1^T$, where $\hat{Q}_1$ and $\hat{C}_1$ are defined in the next subsection. We applied this new procedure to each of the 90 datasets for the sparsity index $S_{C_0} = 0.97$. The average Frobenius losses were presented as box-plots in Figure 4. The result indicated that adding another factorization reduced the Frobenius loss for estimating non-zero entries by 35% while the Frobenius losses for estimating zero-entries under two methods were kept almost the same.

[Put Figure 4 here.]

2.3 A multiple factorization approach

To reduce the above zero-entry effect, we further factorize $C_0(u)$ into $C_0(u) = Q(u)C_m(u)Q(u)^T$ with $Q(u) = Q_1(u)Q_2(u)\cdots Q_m(u)$ and $C_m(u) = Q(u)^{-1}\Sigma(u)Q(u)^{-T}$ by using pre-selected invertible band matrix factors Q_k's. We hope $C_m(u)$ contains fewer zero-entries than does $C_0(u)$. This can reduce the bias effect of these zero-entries. See the Appendix A, the Supplementary Material for some details. For the pre-fixed m, we estimate $Q_v(u), \quad 0 \le v \le m$, and $C_m(u)$ in turn with separate kernel bandwidths, followed by entry-wise thresholding on the resulting plug-in estimator of $C_0(u)$. We replaced Step 1 in the variance-correlation based procedure by the above multiple factorization approach with multiple bandwidths. Here, with multiple bandwidths we aim to improve the flexibility of the proposed procedure in addressing varying smoothness across entries of $C_0(u)$.

There are a few matrix factorization algorithms for estimating a covariance matrix in the literature, for example, the Cholesky algorithm (**rothmanNewApproachCholeskybased2010**). In this paper, we first estimate the diagonal entries, $\hat{Q}_0(u) = \text{diag}(\hat{\sigma}_{kk}(u) : 1 \le k \le p)$ with a $Q_0(u)$-specified bandwidth $h = h_0$. Then, we standardize $\mathbf{y}_i, 1 \le i \le n$ by using $\hat{\boldsymbol{\mu}}(u_i)$ and $\hat{Q}_0(u_i)$:

$$\tilde{\mathbf{y}}_i = \hat{Q}_0^{-1}(u_i)(\mathbf{y}_i - \hat{\boldsymbol{\mu}}(u_i)), \quad 1 \le i \le n.$$

Let $\tilde{\mathbf{y}}^{(m)}$ denote the mth row of the standardized data matrix $\tilde{\mathbf{y}} = (\tilde{\mathbf{y}}_1, \ldots, \tilde{\mathbf{y}}_n)$. To identify band matrix factors for $C_0(u)$, we re-couple the coordinates of Y by maximizing the marginal

correlations of consecutive coordinates as follows: Let $m_1 = 1$ and $S_1 = \{2, \ldots, p\}$. For $k = 2, \ldots, p$, define

$$m_k = \arg \max_{m \in S_{k-1}} \text{Corr}(\tilde{\mathbf{y}}^{(m)}, \tilde{\mathbf{y}}^{(m_{k-1})})$$

and $S_k = S_{k-1} \setminus \{m_k\}$, where $\text{Corr}(\cdot, \cdot)$ denotes the operator for calculating the sample correlation between two random vectors. Let $Y^* = (Y_{m_1}, \ldots, Y_{m_p})$ be the re-coupled Y. Then the large entries in $C_0(u)$ are likely re-arranged to be close to the diagonal band. To simplify the notation, in the following we assume that the underlying coordinates have already been coupled, i.e., satisfying $Y = Y^*$ with $m_k = k$, $1 \le k \le p$. In many applications, the above assumption may hold when coordinates in Y have a natural ordering.

We opt for the following band matrices as factors whose inverses can be explicitly calculated:

$$\hat{Q}_1(u) = (\hat{q}^{(1)}_{kj})_{1 \le k,j \le p}, \quad \hat{q}^{(1)}_{kj} = \begin{cases} 1, & 1 \le k = j \le p. \\ \sum_{i=1}^n w_{ih_1}(u) \tilde{y}_{ki} \tilde{y}_{(k+1)i}, & j = k+1, 1 \le k \le p-1. \\ 0, & \text{Otherwise} \end{cases}$$

$$\hat{Q}_2(u) = (\hat{q}^{(2)}_{kj})_{1 \le k,j \le p}, \quad \hat{q}^{(2)}_{kj} = \begin{cases} 1, & 1 \le k = j \le p. \\ \sum_{i=1}^n w_{ih_2}(u) \tilde{y}_{ki} \tilde{y}_{(k+2)i}, & j = k+2, 1 \le k \le p-2. \\ 0, & \text{Otherwise} \end{cases}$$

$$\vdots$$

$$\hat{Q}_m(u) = (\hat{q}^{(m)}_{kj})_{1 \le k,j \le p}, \quad \hat{q}^{(m)}_{kj} = \begin{cases} 1, & 1 \le k = j \le p. \\ \sum_{i=1}^n w_{ih_m}(u) \tilde{y}_{ki} \tilde{y}_{(k+m)i}, & j = k+m, 1 \le k \le p-m. \\ 0, & \text{Otherwise} \end{cases}$$

with bandwidths $h_1, h_2, \ldots, h_m$. Using the above band matrices, we make the following transformation

$$\breve{\mathbf{y}}_i = \hat{Q}_m^{-1} \cdots \hat{Q}_1^{-1} \tilde{\mathbf{y}}_i, \quad 1 \le i \le n.$$

After the transformation, the new random vector $\breve{\mathbf{y}}_i$ may contain much fewer zero-entries in its covariance matrix $\text{cov}(\breve{\mathbf{y}}_i | u_i)$ than does the original $\tilde{\mathbf{y}}_i$. For $m \ge 1$, we estimate $C_m(u)$ by

$$\hat{C}_m(u) = \sum_{i=1}^n w_{ih_r}(u) \breve{\mathbf{y}}_i \breve{\mathbf{y}}_i^T \tag{5}$$

with bandwidth h_r. With these estimated factors, we reconstruct the following estimator for $C_0(u)$:

$$\hat{C}_0(u) = \begin{cases} \sum_{i=1}^n w_{ih_0}(u) \breve{\mathbf{y}}_i \breve{\mathbf{y}}_i^T, & m = 0 \\ \hat{Q}_1(u) \hat{Q}_2(u) \cdots \hat{Q}_m(u) \hat{C}_m(u) (\hat{Q}_1(u) \hat{Q}_2(u) \cdots \hat{Q}_m(u))^T, & m \ge 1. \end{cases} \tag{6}$$

With $\hat{C}_0$ and $\hat{Q}_0$, we constructed a new plug-in estimator $\hat{\Sigma}(u) = \hat{Q}_0(u) \hat{C}_0(u) \hat{Q}_0(u)$.

Finally, replacing Step 1 in the variance-correlation based procedure by the multiple factorization step above, we end up with a general procedure for estimating covariance matrix.

3 Theory

In this section, we develop an asymptotic theory for the proposed estimators which covers both i.i.d. and non i.i.d. cases and thus is more general than **chenDynamicCovarianceModels2016**. Under certain regularity conditions, the proposed estimators are shown to be consistent with

the underlying matrix function if we let the related bandwidths be different from each other but have the same convergence rate to zero.

Let $\mathcal{F}_{k_0}$ and $\mathcal{F}_{k_0+k}^{\infty}$ be the σ-algebras generated by $\{(\mathbf{y}_i, u_i) : 1 \le i \le k_0\}$ and $\{(\mathbf{y}_i, u_i) : k_0 + 1 \le k < \infty\}$. Define

$$\alpha(k) = \max_{k_0 \ge 1} \sup_{A \in \mathcal{F}_{k_0}, B \in \mathcal{F}_{k_0+k}^{\infty}} |P(A)P(B) - P(A \cap B)|.$$

We assume the following regularity conditions:

(C1) The symmetric kernel function $K(\cdot)$ on $\mathbb{R}$ with derivative $K'(\cdot)$ satisfies

$$\begin{aligned} K_0 = \sup_z K(z) &< +\infty, \quad K_1 = \sup_z |K'(z)| < +\infty, \quad \int K(z)dz = 1, \\ \int zK(z)dz &= 0, \quad \int z^2 K(z)dz < +\infty, \quad \int |z|^3 k(z)dz < \infty. \end{aligned}$$

(C2) The density function of U, $g(u)$, has the second order continuous derivative $g''(\cdot)$ over a compact support $[a, b]$ and $\inf_{u \in [a,b]} g(u) > 0$. For any $i \ne i_1$, the joint density of u_i and u_{i_1}, $\max_{i \ne i_1} \sup_{z, z_1 \in [a,b]} g_{ii_1}(z, z_1)$ is bounded.

(C3) There exist positive constants τ_2 and $\kappa_2 < 1$ such that for $k \ge 1$, $\alpha(k) \le \exp(-\tau_2 k^{\kappa_2})$.

(C4) There exist constants $0 < \kappa_1 \le 1, \tau_1 > 0$ such that

$$\max_{1 \le j \le p} P(|y_{ij}| > v) \le \exp(1 - \tau_1 v^{\kappa_1}).$$

(C5) The second derivatives of $\mu_j(u) = E[y_{1j}|U = u]$, $1 \le j \le p$ are uniformly bounded in the sense that $\max_{1 \le j \le p} \sup_{u \in [a,b]} |\mu_j''(u)| < \infty$.

(C6) The conditional variance functions $\sigma_j^2(u) = E[(y_{ij} - \mu_j(u_i))^2 | u_i = u]$ are bounded below from zero uniformly for $1 \le j \le p$ and $u \in [a, b]$. Their first order derivatives are also uniformly bounded. The conditional expectations $E[(y_{ij} - \mu_j(u_i))(y_{(i+t)j} - \mu_j(u_{i+t}))|u_i = z, u_{i+t} = z_1]$ with $z, z_1 \in [a, b]$, $1 \le i < \infty$, $1 \le t \le \infty$, $1 \le j \le p$, are uniformly bounded in i, t, z and z_1.

The above conditions are routinely used in the literature of nonlinear time series analysis (**fanNonlinearTimeSeries2003lamFactorModelingHighdimensional2012azhangLinearlyConstrained**). It follows from (C5) that $b_2 \hat{=} \max_{1 \le j \le p} \sup_{u \in [a,b]} |\mu_j(u)| < \infty$. (C3) and (C4) assume that the response observations have an exponentially fast mixing rate and sub-exponential tails. Note that these conditions are imposed to facilitate the proofs and thus may not be the weakest possible for establishing the theory below.

Let $\hat{g}_{h_a}(u) = \frac{1}{n} \sum_{i=1}^{n} K_{h_a}(u_i - u)$ be a kernel density estimator of $g(u)$. It follows from Proposition 0.1 in the Supplementary Materials that $\hat{g}_{h_a}(u)$ is uniformly consistent with $g(u)$.

Letting $1/\gamma_1 = 1/\kappa_1 + 1/\kappa_2$, we state a uniform consistency result for estimator $\hat{\mu}_j(u)$ in the following theorem.

Theorem 1. *Under Conditions (C1)∼(C6), if as $n, p \to \infty$ and $h_a \to 0$,*

$$\begin{aligned} &(\log(p))^{2/\gamma_1 - 1}/n = O(1), \quad \frac{\log(h_a^{-4} np)}{(nh_a \log(p/h_a))^{\gamma_1/2}} = O(1), \\ &\frac{(\log(nh_a \log(p/h_a)))^{\gamma_1} \log(1/h_a)}{(nh_a \log(p/h_a))^{\gamma_1(1-\gamma_1)/2}} = O(1), \end{aligned}$$

then

$$\max_{1\le j\le p}\sup_{u\in[a,b]}|\hat{\mu}_j(u)-\mu_j(u)| = O_p\left(\sqrt{\frac{\log(p/h_a)}{nh_a}}\right)+O(h_a^2).$$

Note that $0<\gamma_1<1/2$ as $\kappa_1\le 1$ and $\kappa_2<1$. The above bandwidth condition imposed on h_a holds and $\sqrt{\log(p/h_a)/(nh_a)}=o(1)$ if $h_a=c_0n^{-1/5}$ and $(\log(p))^d/n=o(1)$ for a constant c_0 and $d=\max\{1/(2\gamma_1),2/\gamma_1-1\}$.

Let $1/\gamma_2=2/\kappa_1+1/\kappa_2$. In the next theorem, we show that the entries of the proposed covariance matrix estimator are consistent with the underlying ones uniformly in u and indices $1\le j,k\le p$. We say $h_a,h_v,h_r,h\to 0$ with the same convergence rate if $h/h_a+h_a/h=O(1)$, $h/h_r+h_r/h=O(1)$, $h/h_v+h_v/h=O(1)$, $0\le v\le m$.

Theorem 2. *Under Conditions (C1)∼(C6), if as $n,p\to\infty$, $h_a,h_v,h_r,h\to 0$ with the same rate, for $w=1,2$, $\log(p)^{2/\gamma_w-1}/n=O(1)$ and*

$$\frac{\log(nph^{-4})}{(nh\log(p/h))^{\gamma_w/2}}=O(1),\quad \frac{(\log(nh\log(p/h)))^{\gamma_w}\log(1/h)}{(nh\log(p/h))^{\gamma_w(1-\gamma_w)/2}}=O(1),$$

then

$$\begin{aligned}\max_{1\le j,k\le p}\sup_{u\in[a,b]}|\hat{\sigma}_{jk}(u)-\sigma_{jk}(u)| &= O_p\left(\sqrt{\frac{\log(p/h)}{nh}}+h^2\right),\\ \max_{1\le j,k\le p}\sup_{u\in[a,b]}|\hat{c}_{jk}(u)-c_{jk}(u)| &= O_p\left(\sqrt{\frac{\log(p/h)}{nh}}+h^2\right).\end{aligned}$$

Note that $0<\gamma_2<1/3$ as $\kappa_1\le 1$ and $\kappa_2<1$. The bandwidth condition imposed on h_a,h_v,h_r,h holds and $\sqrt{\log(p/h)/(nh)}=o(1)$ if $h_a=c_0n^{-1/5}$ (which is the optimal bandwidth for the univariate nonparametric regression estimator with c_0 a constant) and $(\log(p))^d/n=o(1)$ for $d=\max\{1/(2\gamma_1),2/\gamma_1-1,1/(2\gamma_2),2/\gamma_2-1\}$.

Let $\alpha_p(u)=\|\Sigma(u)-\langle\Sigma(u),I_p\rangle I_p\|_F$ and $\tau_{np}=\sqrt{\log(p/h)/(nh)}$. Let $\hat{t}_0(u)$ be an estimator of the thresholding function $t_0(u)$ used in $\hat{\Sigma}^{(t)}(u)$ and $\hat{\Sigma}^{(st)}(u)$. Let $m_p(u)=\max_{1\le k\le p}\sum_{j=1}^p I(\sigma_{kj}(u)>0)$ be a sparsity index of $\Sigma(u)$. The smaller $m_p(u)$, the sparser $\Sigma(u)$ is. To state the next theorem, we introduce the following conditions on separability between $\Sigma(u)$ and I_p, sparsity and bounds of $\Sigma(u)$ respectively.

(C7) $\tau_{np}/(\log(p/h)\inf_{u\in[a,b]}\alpha_p^2(u))=O(1)$, $\sup_{u\in[a,b]}m_p(u)\tau_{np}/\alpha_p(u)=o(1)$.

(C8) There exists a positive constant s_1 such that $\sup_{u\in[a,b]}\|\Sigma(u)\|\le s_1$.

(C9) There exists a positive constant s_{0p} such that as $p\to\infty$,

$$s_{0p}/\sqrt{\sup_{u\in[a,b]}m_p(u)\tau_{np}}\to\infty,\quad \inf_{u\in[a,b]}\|\Sigma(u)\|\ge s_{0p}.$$

(C10) $\sup_{u\in[a,b]}|\hat{t}_0(u)-t_0(u)|=o(1)$ and there exist positive constants $t_a<t_b$ such that for $t_a<\inf_{u\in[a,b]}t_0(u)\le\sup_{u\in[a,b]}t_0(u)<t_b$.

Note that Condition (C7) implies that $\Sigma(u)$ is not close to cI_p in a distance less than the product of the sparsity index and the rate $\tau_{np}/\log(p/h)$, where c is any arbitrary constant. Conditions (C8) and (C9) are about the uniform boundedness of $\|\Sigma(u)\|$ from above and away from zero in an order of τ_{np} timed by the sparsity index. Finally, we can see from Theorem 3 below that although (C10) requires the tuning constant $\hat{t}_0(u)$ has a finite limit as n tends to infinity, the order of the convergence rate of the corresponding estimator $\hat{\Sigma}^{(st)}(u)$ is independent of such a limit.

Under these conditions, we state a uniform consistent result for $\hat{\Sigma}^{(st)}(u)$ as follows.

Theorem 3. *Under Conditions (C1)∼(C8), if as $n, p \to \infty$, $h_a, h_v, h_r, h \to 0$ with the same rate, and for $w = 1, 2$ $\log(p)^{2/\gamma_w - 1}/n = O(1)$, $nh^5/\log(p/h) = O(1)$ and*

$$\frac{\log(nph^{-4})}{(nh\log(p/h))^{\gamma_w/2}} = O(1), \quad \frac{(\log(nh\log(p/h)))^{\gamma_w}\log(1/h)}{(nh\log(p/h))^{\gamma_w(1-\gamma_w)/2}} = O(1),$$

and if $\sup_{u\in[a,b]} m_p(u)\tau_{np} = o(1)$, then uniformly in $u \in [a, b]$,

$$\left\|\hat{\Sigma}^{(st)}(u) - \Sigma(u)\right\| = O_p(m_p(u)\tau_{np}).$$

In addition to the above conditions, if Condition (C9) holds, then uniformly in $u \in [a, b]$,

$$\begin{aligned}
\left\|\hat{\Sigma}^{(st)}(u)\Sigma^{-1}(u) - I_p\right\| &= O_p\left(m_p(u)\tau_{np}s_{0p}^{-1}\right) = o_p\left(\sqrt{m_p(u)\tau_{np}}\right). \\
\left\|\Sigma(u)(\hat{\Sigma}^{(st)}(u))^{-1} - I_p\right\| &= O_p\left(m_p(u)\tau_{np}s_{0p}^{-1}\right) = o_p\left(\sqrt{m_p(u)\tau_{np}}\right). \\
\left\|(\hat{\Sigma}^{(st)}(u))^{-1} - \Sigma^{-1}(u)\right\| &= O_p\left(m_p(u)\tau_{np}s_{0p}^{-2}\right) = o_p(1).
\end{aligned}$$

Finally, in addition to the above conditions, if Condition (C10) holds, then the above results continue to hold after replacing $t_0(u)$ by $\hat{t}_0(u)$ in $\hat{\Sigma}^{(t)}(u)$ and $\hat{\Sigma}^{(st)}(u)$.

Note that if $h = c_0 n^{-1/5}$ (c_0 is a constant) and for $d = \max\{1/(2\gamma_1), 2/\gamma_1 - 1, 1/(2\gamma_2), 2/\gamma_2 - 1\}$, $(\log(p))^d/n = o(1)$, the above condition imposed on h holds. Note that the above bandwidth assumption that they have the same convergence rate to zero does not rule out these bandwidths are different. However, the cross-validation (or the so-called subset) selected bandwidths may not tend to zero. In particular, some of these bandwidths may tend to infinity when there are many zeros and a few non-zeros in the underlying covariance matrix. In this situation, simulation studies in the next section showed that the proposed estimators could reduce the bias and outperformed the DCM in terms of integrated mean squared errors. The theoretical development along this aspect will be spelled out in a future paper.

4 Numerical studies

In this section, to demonstrate the merits of the proposed estimators in finite sample settings, we applied the proposed procedures to both synthetic and real data. We presented the numerical results for the proposed procedure using m ($m = 0, 1$) band matrix factors.

To facilitate the presentation, let ${}_{\mathrm{t}}\mathrm{NCM}_0$ and ${}_{\mathrm{st}}\mathrm{NCM}_0$ denote the proposed estimators $\hat{\Sigma}^{(t)}(u)$ and $\hat{\Sigma}^{(st)}(u)$ respectively with $m = 0$. Let ${}_{\mathrm{t}}\mathrm{NCM}_1$ and ${}_{\mathrm{st}}\mathrm{NCM}_1$ denote the proposed estimators respectively with $m = 1$. Let DCM_1 and DCM_2 denote three DCM estimators defined by

$$\begin{aligned}
\mathrm{DCM}_1(u) &= (\hat{\sigma}_{1jk}(u)I(\hat{\sigma}_{1jk}(u) \geq d(u))), \\
\mathrm{DCM}_2(u) &= (\hat{\sigma}_{2jk}(u)I(\hat{\sigma}_{2jk}(u) \geq d(u))),
\end{aligned}$$

where $d(u)$ is the level of thresholding and

$$\begin{aligned}
\hat{\Sigma}_1(u) &= \sum_{i=1}^{n} w_{ih}(u)(\mathbf{y}_i - \hat{\boldsymbol{\mu}}(u_i))(\mathbf{y}_i - \hat{\boldsymbol{\mu}}(u_i))^T \hat{=} (\hat{\sigma}_{1jk}(u))_{1\leq j,k\leq p}, \\
\hat{\Sigma}_2(u) &= \sum_{i=1}^{n} w_{ih}(u)(\mathbf{y}_i - \hat{\boldsymbol{\mu}}(u))(\mathbf{y}_i - \hat{\boldsymbol{\mu}}(u))^T \hat{=} (\hat{\sigma}_{2jk}(u))_{1\leq j,k\leq p}.
\end{aligned}$$

Note that DCM_1 differs from DCM_2 in the way of estimating the residuals: The former

uses estimators $\mathbf{y}_i - \hat{\boldsymbol{\mu}}(u_i) = \Sigma(u_i)^{1/2}\boldsymbol{\varepsilon}_i + \boldsymbol{\mu}(u_i) - \hat{\boldsymbol{\mu}}(u_i)$, $1 \le i \le n$ while the latter adopts estimators $\mathbf{y}_i - \hat{\boldsymbol{\mu}}(u) = \Sigma(u_i)^{1/2}\boldsymbol{\varepsilon}_i + \boldsymbol{\mu}(u_i) - \hat{\boldsymbol{\mu}}(u)$, $1 \le i \le n$. Here, compared to $\boldsymbol{\mu}(u_i) - \hat{\boldsymbol{\mu}}(u_i)$, $\boldsymbol{\mu}(u_i) - \hat{\boldsymbol{\mu}}(u) = \boldsymbol{\mu}(u_i) - \hat{\boldsymbol{\mu}}(u_i) + \hat{\boldsymbol{\mu}}(u_i) - \hat{\boldsymbol{\mu}}(u)$ has an extra bias $\hat{\boldsymbol{\mu}}(u_i) - \hat{\boldsymbol{\mu}}(u)$. So, DCM_2 is expected to perform worse than DCM_1. Following the same procedure as in ${}_{\mathrm{st}}\mathrm{NCM}_0$, we improve DCM_1 by incorporating the effects of shrinkage on it. Let ${}_{\mathrm{s}}\mathrm{DCM}_1$ denote the optimal shrinkage estimator after replacing ${}_{\mathrm{t}}\mathrm{NCM}_0$ by DCM_1 in the definition of ${}_{\mathrm{st}}\mathrm{NCM}_0$.

4.1 Choice of tuning parameters

As is common in most smoothing methods, the choice of appropriate tuning parameters plays an important role in the performance of a regularized estimator. It is prudent to restrict attention to data-driven choice of the tuning parameters. Here we apply the cross-validation to choose the values of tuning parameters in a sequential manner as follows.

Bandwidth for estimating $\boldsymbol{\mu}(u)$. We let $h_a = \arg\min \mathrm{CV}_{\boldsymbol{\mu}}(h)$ as the optimal bandwidth for the mean kernel estimator in equation (1), where

$$\mathrm{CV}_{\boldsymbol{\mu}}(h) = \frac{1}{n}\sum_{i=1}^{n} \left\|\mathbf{y}_i - \hat{\boldsymbol{\mu}}_{h,-i}(u_i)\right\|^2 \omega(u_i).$$

Here, $\hat{\boldsymbol{\mu}}_{h,-i}(u_i)$ is a kernel mean function estimator after dropping the ith observation from the data. The trimming function $\omega(u) = I(u_{(1)} < u < u_{(n-1)})$ is used for reducing the boundary effects on $\mathrm{CV}_{\boldsymbol{\mu}}(h)$, where $u_{(k)}$ is the kth order statistic of $(u_i)_{i=1}^n$.

Bandwidth for estimating $Q_0(u)$. To select the bandwidth for $\hat{Q}(u)$, for each h, we calculate $\hat{\sigma}_{kk(-i)} : 1 \le k \le p$ after dropping the ith observation. We choose the optimal bandwidth $h_0 = \arg\min \mathrm{CV}_0(h)$ for $\hat{Q}(u)$, where $\mathrm{CV}_0(h)$ is a Stein-loss-based cross-validation function defined by

$$\mathrm{CV}_0(h) = \sum_{i=1}^{n}\sum_{k=1}^{p}\left\{\frac{(y_{ki} - \hat{\boldsymbol{\mu}}_k(u_i))^2}{\hat{\sigma}^2_{kk(-i)}(u_i)} + \log(\hat{\sigma}_{kk(-i)}(u_i))\right\}.$$

Bandwidth for estimating $Q_k(u)$, $1 \le k \le m$. We choose $h_k = \arg\min \mathrm{CV}_k(h)$ at which the following criterion attains the minimum:

$$\mathrm{CV}_k(h) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{p-k}\left(\hat{\rho}_{j(j+k)(-i)}(u_i) - \tilde{y}_{ij}\tilde{y}_{i(j+k)}\right)^2,$$

where $\hat{\rho}_{j(j+k)(-i)}(u_i)$ is the kernel estimator of the $j(j+k)$th correlation $\rho_{j(j+k)}(u_i)$ based on the leave-one-out dataset $(\tilde{\mathbf{y}}_t, u_i)_{t\neq i}$.

Bandwidth for estimating $C_m(u)$. There are two existing cross-validation methods for selecting the bandwidth h for $C_m(u)$: One is a Stein-loss-based approach (**yinNonparametricCovarianceModel20**
which was however applicable only to low-dimensional data. The other is a subset-based approach (**chenDynamicCovarianceModels2016**) for high-dimensional data. In this paper, we choose $h_{m+1} = \arg\min \mathrm{CV}_C(h)$ at which the following criterion attains the minimum:

$$\mathrm{CV}_C(h) = \frac{1}{n}\sum_{i=1}^{n}\left\|\hat{C}_{m(-i)}(u_i) - \breve{\mathbf{y}}_i\breve{\mathbf{y}}_i^T\right\|_F^2,$$

where $\hat{C}_{m(-i)}(u_i)$ is the kernel estimator of $C_m(u)$ based on the leave-one-out dataset $(\breve{\mathbf{y}}_j, u_i)_{j\neq i}$.

Thresholding level for $\hat{C}^{(t)}(u)$. Following **bickelCovarianceRegularizationThresholding2008**, we split the sample into two sub-samples called trial and testing samples and select the threshold by minimizing the Frobenius norm of the difference between the trial-sample-based thresholded estimator and the testing-sample-based covariance matrix. Specifically,

we divide the original sample into two samples at random of size n_1 and n_2, where $n_1 = n(1 - 1/\log(n))$ and $n_2 = n/\log(n)$, and repeat this N_1 times. Here, we set $N_1 = 100$ as the default value according to our numerical experience. Let $\hat{C}_{1,s}(u)$ and $\hat{C}_{2,s}(u)$ be the plug-in estimators based on n_1 and n_2 observations respectively with the bandwidth selected by the leave-one-out cross validation. Let $\hat{C}_{1,s}^{(t)}$ be the thresholded estimator derived from $\hat{C}_{1,s}(u)$ with the thresholding level $t_0(u)$. Given u, we select $t_0(u)$ by minimizing $\sum_{s=1}^{N_1} \|\hat{C}_{1,s}^{(t)} - \hat{C}_{2,s}\|_F / N_1$.

Turning parameters for estimating DCM. The bandwidth h and the level of thresholding of the DCM estimators in (7) are determined by the so-called subset and sample-splitting approaches respectively (**chenDynamicCovarianceModels2016**).

4.2 Criteria for performance assessment

We need a criterion to measure the performance of a nonparametric covariance matrix estimator. There are multiple possible criteria, but one particularly convenient choice is integrated root-squared error (IRSE). For any estimator $\hat{\Psi}(u)$ of $\Sigma(u)$, $u \in [a, b]$ the IRSE is defined as

$$\text{IRSE}(\hat{\Psi}) = \int_a^b \left\| \hat{\Psi}(u) - \Sigma(u) \right\|_F du \approx \frac{1}{K_0} \sum_{k=1}^{K_0} \left\| \hat{\Psi}(v_k) - \Sigma(v_k) \right\|_F,$$

where $v_k, 1 \le k \le K_0$ be grids evenly distributed over the interval (a, b). In the following, we set $K_0 = 20$ for $(a, b) = (-1, 1)$. In our study, we also consider a spectral-norm based IRSE. The results are omitted as they are similar to the Frobenius version.

We also evaluate the performance of the proposed procedure in discovering zero entries in the covariance matrix. Let p_1 (p_2) be the number of non-zero (zero) entries in $\Sigma(u)$. For any estimator $\hat{\Psi}(u)$ of $\Sigma(u)$, let n_{11} be the number of true discoveries of non-zero entries in $\Sigma(u)$ by $\hat{\Psi}(u)$. Similarly, let n_{22} denote the number of true discoveries of zero entries in $\Sigma(u)$ by $\hat{\Psi}(u)$. Let SEN, SPE and ACC denote sensitivity, specificity and accuracy in the above testing, namely

$$\text{SEN} = \frac{n_{11}}{p_1}. \quad \text{SPE} = \frac{n_{22}}{p_2}. \quad \text{ACC} = \frac{n_{11} + n_{22}}{p_1 + p_2}.$$

4.3 Synthetic data

In this subsection, we carried out a set of simulation studies. We considered three settings for $\boldsymbol{\mu}(u)$ and $\Sigma(u)$ in our simulations.

Setting 1: Following **yuanReproducingKernelHilbert2010** and **chenDynamicCovarianceModels2016**, we set $\boldsymbol{\mu}(u)$ and $\Sigma(u)$ as follows. Let $\boldsymbol{\mu}(u) = (\mu_1(u), \ldots, \mu_p(u))^T$ with

$$\mu_j(u) = \sum_{k=1}^{50} \frac{(-1)^{k+1}}{k^2} Z_{jk} \cos(k\pi u), \quad 1 \le j \le p,$$

where $\{Z_{jk} : 1 \le j \le p, 1 \le k \le 50\}$ is an independent sample drawn from the uniform distribution over $[-5, 5]$. Let $\Sigma(u) = \{\sigma_{ij}(u)\}_{1 \le i,j \le p}$ with $\sigma_{ij}(u) = \exp(u/2)[\{\phi(u) + 0.1\} I(|i - j| = 1) + \phi(u) I(|i - j| = 2) + I(i = j)]$ and $\phi(u)$ is the standard normal density. Note that $\text{diag}(\Sigma(u)) = \exp(u/2) I_p$ is spherical and the correlation matrix $C_0(u) = (c_{ij}(u))_{1 \le i,j \le p}$ with $c_{ij}(u) = I(|i - j| = 1) + \phi(u) I(|i - j| = 2) + I(i = j)$ which is equal to zero when $|i - j| \ge 3$. Therefore, $C_0(u)$ is sparse as it is banded with bandwidth 2.

Setting 2: Following **zhangLinearlyConstrainedMinimum2015**, let $\boldsymbol{\mu}(u) = (\mu_1(u), \ldots, \mu_p(u))^T$ with

$$\mu_j(u) = Z_j \exp\left(\frac{(u - \tau_j)^2}{2^2} \right) \sin(2\pi(u - \tau_j)), \qquad 1 \le j \le p,$$

where $Z_j, j = 1, \ldots, p$ are independently drawn from uniform distribution $U(-5, 5)$, $\tau = (\tau_1, \ldots, \tau_p)$ is a row vector of p evenly spaced points between -1 and 1. Set $\Sigma(u) = \{\sigma_{ij}(u)\}_{1 \le i,j \le p}$ with $\sigma_{ij}(u) = \exp(u/2)\phi(u)^{|i-j|}$. Note that $\text{diag}(\Sigma(u)) = \exp(u/2)I_p$ is spherical and the correlation matrix

$C_0(u) = (c_{ij}(u))_{1 \le i,j \le p}$ with $c_{ij}(u) = \phi(u)^{|i-j|}$. Therefore, $c_{ij}(u)$ is decreasing exponentially fast but is not sparse.

Setting 3: Let $\boldsymbol{\mu}(u)$ be the same as that in **Setting 1**. Let $\Sigma(u) = A^T(u)A(u)$, where the (i, j)th entry of $A(u)$ equals

$$\begin{aligned} a_{ij}(u) &= \exp\left(\frac{u\sin(ij)}{2}\right)\Big\{[\sin(\pi u) + 0.1]\, I(|i-j| = 1) \\ &\quad + \sin(\pi u) I(|i-j| = 2) + I(i = j)\Big\}. \end{aligned}$$

Note that $\text{diag}(\Sigma(u)) = \text{diag}(\sum_{j=1}^{p} a_{ij}^2(u) : 1 \le i \le p)$ is not spherical. $C_0(u)$ is sparse as it is banded with bandwidth 4.

For each combination of (n, p) with $n = 100, 200, 500$ and $p = 50, 100,\ 150, 300, 500$, we repeated the experiment 90 times, generating 90 datasets of $(\mathbf{y}_i, u_i)$, $1 \le i \le n$. Each dataset was obtained in two steps. In Step 1, we drew $u_i, i = 1, \ldots, n$ independently from the uniform distribution $U(-1, 1)$. In Step 2, for each given u_i, we drew $\mathbf{y}_i$ from the covariance model $\mathbf{y}_i = \boldsymbol{\mu}(u_i) + \Sigma(u_i)^{1/2}\boldsymbol{\varepsilon}_i$, where $\boldsymbol{\varepsilon}_i$, $i = 1, \ldots, n$ were iteratively drawn from the vector VAR(1) model

$$\boldsymbol{\varepsilon}_0 = \xi_0, \quad \boldsymbol{\varepsilon}_i = \rho\boldsymbol{\varepsilon}_{i-1} + \xi_i, \quad i = 1, \ldots, n$$

with $0 \le \rho < 1$ and $\xi_k, k = 0, 1, \ldots$ independently sampled from the standard p-dimensional Normal $N(0, I_p)$. We considered $\rho = 0, 0.3, 0.8$.

For each combination of (n, p, ρ), we applied ${}_{\text{t}}\text{NCM}_m$, ${}_{\text{st}}\text{NCM}_m$, $m = 0, 1$, DCM_1, ${}_{\text{s}}\text{DCM}_1$ and DCM_2 to each of 90 datasets and calculated their IRSE values and (SEN, SPE, ACC) values. The mean and standard error of these values are displayed in Tables 1∼ 9 below and Tables 1∼7 in the Appendix D (the Supplementary Material) respectively. As example, for each of 90 datasets simulated in Setting 1 with $n = p = 100$, the CPU time required by DCM_1, ${}_{\text{s}}\text{DCM}_1$, DCM_2, ${}_{\text{t}}\text{NCM}_m$ and ${}_{\text{st}}\text{NCM}_m$, $m = 0, 1$ to estimate the covariance matrix function is reported in the Appendix D, the Supplementary Material.

[Put Tables 1∼ 9 here.]

The results can be summarized as follows:

- On average, the IRSE loss of each procedure was increasing in the dimension p and in the degree of serial correlation ρ while decreasing in sample size n.
- The degrees of sparsity and diagonal homogeneity in $\Sigma(u)$ had an effect on the performance of these four procedures. For example, when $(n, p, \rho) = (100, 300, 0)$, compared to in Setting 1, the IRSE loss of ${}_{\text{st}}\text{NCM}_0$ in Setting 2 increased by 84%. This is not surprising as the degrees of sparsity and diagonal homogeneity in Setting 2 lead to a higher dimensionality (i.e., the number of effective parameters in the model) than that in Setting 1.
- Among the seven procedures, ${}_{\text{st}}\text{NCM}_1$ performed best in all three settings, followed by ${}_{\text{st}}\text{NCM}_0$, ${}_{\text{t}}\text{NCM}_1$, ${}_{\text{t}}\text{NCM}_0$, ${}_{\text{s}}\text{DCM}_1$, DCM_1 and DCM_2. In particular, the performance of DCM_2 was substantially worse than its competitors. For example, for $(n, p, \rho) = (100, 300, 0)$, in Setting 1, compared to DCM_1, on average ${}_{\text{t}}\text{NCM}_0$ and ${}_{\text{st}}\text{NCM}_0$ reduced the IRSE loss by 23% and 25% respectively. ${}_{\text{t}}\text{NCM}_1$ and ${}_{\text{st}}\text{NCM}_1$ performed slightly better than ${}_{\text{t}}\text{NCM}_0$ and ${}_{\text{st}}\text{NCM}_0$ in some cases. Compared to ${}_{\text{t}}\text{NCM}_0$, on average ${}_{\text{st}}\text{NCM}_0$

reduced the IRSE loss by 3%. Compared to DCM_2, on average DCM_1 reduced the IRSE loss by 99%. In Setting 2, compared to DCM_1, on average ${}_{\text{t}}\text{NCM}_0$ and ${}_{\text{st}}\text{NCM}_0$ reduced the IRSE loss by 12% and 16% respectively. ${}_{\text{t}}\text{NCM}_1$ and ${}_{\text{st}}\text{NCM}_1$ performed slightly better than ${}_{\text{t}}\text{NCM}_0$ and ${}_{\text{st}}\text{NCM}_0$, in particular, in Setting 3. Compared to DCM_2, on average DCM_1 reduced the IRSE loss by 99%. Compared to ${}_{\text{t}}\text{NCM}$, on average ${}_{\text{st}}\text{NCM}$ reduced the IRSE loss by 5%. In Setting 3, compared to DCM_1, on average ${}_{\text{t}}\text{NCM}_0$ and ${}_{\text{st}}\text{NCM}_0$ reduced the IRSE by 14% and 15% respectively. Compared to DCM_2, on average DCM_1 reduced the loss by 94%. Compared to ${}_{\text{t}}\text{NCM}_0$, on average ${}_{\text{st}}\text{NCM}_0$ reduced the IRSE loss by 2%. ${}_{\text{t}}\text{NCM}_1$ and ${}_{\text{st}}\text{NCM}_1$ performed substantially better than their counterparts ${}_{\text{t}}\text{NCM}_0$ and ${}_{\text{st}}\text{NCM}_0$. The similar conclusion can be made for dependent samples when $\rho = 0.3$ and 0.8. In particular, the optimal shrinkage can reduce the serial correlation effect on the proposed procedures ${}_{\text{st}}\text{NCM}_0$ and ${}_{\text{st}}\text{NCM}_1$.

- Similar results were obtained in terms of ACC. See Tables 1∼7 in the Appendix D, the Supplementary Material.
- The CPU-time costs of ${}_{\text{t}}\text{NCM}_m$ and ${}_{\text{st}}\text{NCM}_m$, $m = 0, 1$, are less than those of DCM_1 and DCM_2. See Figures 2 and 3 in the Appendix D, the Supplementary Material.

4.4 Asset return data

Capital asset pricing model (CAPM) is a model that describes the relationship between systematic risk and expected return for assets, which is widely used throughout finance for the pricing of risky assets. However, the assumption that asset returns are linearly related to the market return is imposed on the model. The primary goal of this study was to extend the CAPM to the nonlinear setting. In particular, we are interested in how the volatility and co-volatility of a group of asset returns depend on the market return.

For this purpose, from the database of Yahoo Finance, we collected monthly return data of 75 assets across 8 sectors over three time-periods, namely, before-financial-crisis period from 02/2001 to 01/2007, in-financial-crisis period from 02/2007 to 01/2010 and after-financial-crisis period from 02/2010 to 12/2017. The sector distribution of these assets as follows. Technology: AAPL, AMD, HPQ, IBM, IIN, INTC, LNGY, LOGI, MSFT, NTAP, NVDA, SNE, TACT and WDC. Health care: AET, AMGN, AZN, BAX, CVS, GILD, GSK, HUM, IMMU, JNJ, LLY, MRK, NVS, PFE, TECH and UNH. Energy: BP, CVX, OXY, RDS-B, SU and XOM. Financial services: C, GS, HSBC, JPM, MS, PGR, RF and THG. Communication services: SHEN, T and TEO. Consumer defensive: BIG, DLTR, FRED, KO, TGT, TUES, UN and WMT. Consumer cyclical: AMZN, EMMS, KSS, SIRI and TM. Industrial: BA, CAJ, DY, EME, FIX, GE, GVA, IR, MMM, MTZ, PWR, SKYW, UPS, UTX and VMI. We also collected the index return of S&P500 which was treated as the market's return.

We applied the proposed ${}_{\text{st}}\text{NCM}_0$ and ${}_{\text{st}}\text{NCM}_1$ to the data for each time-period, obtaining almost the same result. Here, we reported the corresponding estimates for mean $\boldsymbol{\mu}(u)$ and covariance matrix $\Sigma(u)$. Note that the diagonals of estimated $\Sigma(u)$ show the volatility of individual returns while estimated correlation coefficient matrix $C_0(u)$ captures cross-sectional relationships in these returns.

We plotted the estimated individual mean functions and the estimated volatility functions in Figure 1, revealing a number of assets which had nonlinear relationships to the market return. The degree of this non-linearity significantly decreased after financial crisis, indicating that the CAPM fitted to the market better than before the financial crisis. See the Appendices E and F, the Supplementary Material for more details. Figure 1 also shows that the individual volatility of the assets increased a lot during the financial crisis period but returned to normal after the financial crisis. The pattern of the dependence of the volatility on the market also changed a lot after financial crisis: Changes from non-constant

volatility functions before the financial crisis to almost constant volatility functions after the financial crisis. We also investigated effects of the financial crisis on the co-volatility of the selected assets by the estimated non-zero correlation coefficient functions. By use of the estimated covariance matrix functions, in each time-period, we identified the associated pairs of assets that were of non-zero market-dependent conditional correlation coefficients (and non-zero conditional co-volatility). We further conducted asymptotic tests for significance of co-volatility for these pairs as follows. For any pair of assets (a, b), let $\mathrm{Corr}_{(a,b)}(u)$ denote its correlation coefficient as a function of u (the market's return) and with estimator $\hat{\mathrm{Corr}}_{(a,b)}(u)$. Let $\hat{F}_{(a,b)}(u) = 0.5\log(1 + \hat{\mathrm{Corr}}_{(a,b)}(u))/\log(1 - \hat{\mathrm{Corr}}_{(a,b)}(u))$ be Fisher's Z transformation. To test $H_0 : \mathrm{Corr}_{(a,b)}(u) \not\equiv 0$, we considered the test statistics

$$\mathrm{Avec}_{(a,b)} = \sum_{i=1}^{n} |\hat{F}_{(a,b)}(u_i)|/n \approx N\left(E[|F_{(a,b)}(U)|], \mathrm{var}(|F_{(a,b)}(U)|)/n\right)$$

and calculated the approximate P-value

$$P\left(\sqrt{n}\mathrm{Avec}_{(a,b)}/\sqrt{\hat{\mathrm{var}}(\mathrm{Corr}_{(a,b)}(U))}\Big|N(0,1)\right),$$

where the sample variance of $|\hat{F}_{(a,b)}(u_i)|, 1 \le i \le n$ is denoted by $\hat{\mathrm{var}}(|F_{(a,b)}(U)|)$ and $P(\cdot|N(0,1))$ is the cumulative distribution function of the standard normal $N(0,1)$. Then, even after Bonferroni correction for multiple testing, these P-values were all significant ($< 10^{-2}$) for the above selected pairs of assets. The final list of significant pairs are as follows:

- *Before-financial-crisis.* There were 1, 14, 1 pairs existed within Technology, Energy and Consumer-Defensive respectively.
- *In-financial-crisis.* There were 4, 1, 8,1, 4,1, 1,1,1 pairs of correlated assets presented within Technology, Industrial, Energy, Consumer Defensive, Health Care and Financial Services respectively. Also, there was a pair of correlated assets belonging two different sectors: Industrial and Consumer cyclical, Consumer Cyclical and Consumer Defensive, and Financial Service and Industrial.
- *After-financial-crisis.* There were 3, 2, 10, 1, 11, and 12 pairs of assets within Technology, Industrial, Energy, Consumer defensive, Health care, and Financial services. There were 1, 1, 1, 1 and 2 pairs of assets between Financial service and Industrial, between consumer defensive and Financial services, between Consumer cyclical and Consumer defensive, between Technology and Industrial, and between Health Care and Consumer Defensive.

The results indicate that before financial crisis, there were only 16 significant within-sector co-volatility connections among these assets. In particular, there were no significant cross-sectional co-volatility connections among these assets. The number of co-volatility assets within and across sectors was significantly increasing during and after financial-crisis: The number of within-sector co-volatility connections increased from 16 to 22 during the financial crisis period and to 37 after the financial crisis. The number of between-sector co-volatility connections increased from 0 to 3 during the financial crisis period and to 7 after the financial crisis. This implies that in response to the financial crisis, the financial market has been more closely integrated than before the financial crisis.

5 Discussion and conclusion

Estimating covariate-dependent covariance matrix $\Sigma(u)$ of a high-dimensional response vector poses a big challenge to contemporary statistical research. The existing kernel methods in **yinNonparametricCovarianceModel2010** and **chenDynamicCovarianceModels2016** might not be flexible enough to capture varying smoothness across key parts of the matrix as they used a single bandwidth for all entries of $\Sigma(u)$. Here, we have proposed a novel estimation procedure to overcome this obstacle, based on a variance-correlation factorization of $\Sigma(u)$, namely $\Sigma(u) = Q_0(u)C_0(u)Q_0^T(u)$, where $Q_0(u) = \text{diag}(\Sigma(u))^{1/2}$ and the correlation matrix function $C_0(u)$ is further factorized into the product of multiple band matrices. The proposal has been implemented in two steps. In Step 1, we estimate $Q_0(u)$ and $C_0(u)$ robustly by use of separate bandwidths for band matrices, followed by thresholding entries of the estimated $C_0(u)$. In Step 2, substituting these estimators in the above factorization formula to obtain a plug-in estimator, followed by an optimal shrinkage from a decision-making point of view.

We have conducted a set of simulations to demonstrate that the new proposal outperforms the existing DCM approach in terms of estimation loss and CPU-time cost. To illustrate our new proposal, we have applied it to a dataset of asset returns. We have developed a nonparametric capital asset pricing model to capture volatility and co-volatility among these risky assets. We have showed that under some sparsity conditions, the proposed estimator is consistent with the underlying covariance matrix as both the sample size and the dimension tend to infinity. There are a few important topics which are remained to address but beyond the scope of this paper, such as nonparametric nonlinear shrinkage.

Acknowledgments

The research of the second author was supported by a Graduate Teaching Assistant (GTA) scholarship at the University of Kent. We are grateful to the Co-Editor and two reviewers for their positive statements on the manuscript as well as their specific, detailed, and valuable comments that have helped to improve the paper.

Supplementary materials

The detailed proofs of the lemmas and some extra information on numerical results can be found in the Online Supplementary Material.

References

Bai, Z. and Silverstein, J. W. (2010). Spectral Analysis of Large Dimensional Random Matrices, 2nd ed. *Springer,* New York.

Bickel, P. and Levina, E. (2008). Covariance regularization by thresholding, *Ann. Stat.*, **36**, 2577–2604.

Donoho, D., Gavish, M. and Johnstone, I. (2018). Optimal shrinkage of eigenvalues in the spiked covariance model. *Ann. Stat.*, **46**, 1742-1778.

Cai, T. and Liu, W. (2012). Adaptive thresholding for sparse covariance matrix estimation. *Jour. Ameri. Stat. Assoc.*, **106**, 672-684.

Chen, Z. and Leng, C. (2016). Dynamic covariance models. *Jour. Ameri. Stat. Assoc.*, **111**, 1196-1207.

Engle, R.F., Ledoit, O. and Wolf, M. (2017). Large Dynamic Covariance Matrices. *Jour. Busi. & Econ. Stat.*, DOI: 10.1080/07350015.2017.1345683.

Fama, E. and French, K. (2004). The Capital asset pricing model: theory and evidence". *Journal of Economic Perspectives*, **18**, 25–46.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications*, Chapman & Hall, London.

Fan, J., Liao, Y. and Micheva, M.(2013). Large Covariance Estimation by Thresholding Principal Orthogonal Complements (with discussions). *Journal of Royal Statistical Society Series B*, **75**, 603-680.

Fan, J. and Yao, Q. (2003). *Nonlinear Time Series: Nonparametric and Parametric Methods* . New York: Springer.

Fox, E. and Dunson, D. (2015). Bayesian nonparametric covariance regression. *Journal of Machine Learning Research*, **16**, 2501-2542.

Jolliffe, I. (2002). Principal Component Analysis. 2nd Edition. *Springer*, New York.

Lam, C. and Yao, Q.(2012). Factor modeling for high-dimensional time series: inference for the number of factors. . *Ann. Stat.* , **40**, 694-726.

Lamusa, C., Hämäläinenc, M.S., Temereancaa,S., Browna,E.N., and Purdona, P.L. (2012). A spatiotemporal dynamic distributed solution to the MEG inverse problem. *Neuroimage*, **63**, 894-909.

Ledoit, O. and Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Jour. Multi. Analy.*, **88**, 365–411.

Merlevéde, F., Peligrad, M. and Rio, E. (2009). Bernstein inequality and moderate deviations under strong mixing conditions. In *IMS Collections. High Dimensional Probability* **V**, 273–292. Beachwood, OH.

Reich, B.J., Eidsvik, J., Guindani, M., Nail, A.J. and Schmidt, A.M. (2011). A class of covariate-dependent spatiotemporal covariance functions for the analysis of daily ozone concentration. *Ann. Appl. Stat.* , **5**, 2425–2447.

Rothman, A.J., Levina, E. and Zhu, J. (2010). A new approach to Cholesky-based covariance regularization in high dimensions. *Biometrika*, **97**, 539-550.

Yin, J., Geng, Z., Li, R. and Wang, H. (2010). Nonparametric covariance model. *Statistica Sinica*, **20**, 469-479.

Yuan, M. and Cai, T. (2010). A reproducing kernel Hilbert space approach to functional linear regression, *Ann. Stat.*, **38**, 3412–3444.

Zhang, J. and Liu, C. (2015). On linearly constrained minimum variance beamforming. *Journal of Machine Learning Research*, **16**, 2099-2145.

Zhang, J. and Su, L. (2015). Temporal autocorrelation-based beamforming With MEG neuroimaging data. *Jour. Ameri. Stat. Assoc.*, **110**, 1375-1388.

Jian Zhang
School of Mathematics, Statistics and Actuarial Science, University of Kent, Canterbury, Kent CT2 7FS, UK

E-mail: jz79@kent.ac.uk

Jie Li
School of Mathematics, Statistics and Actuarial Science, University of Kent, Canterbury, Kent CT2 7FS, UK
and
School of Mathematics and Computer Science, Dali University, Dali, 671003, China.
E-mail: jl705@kent.ac.uk

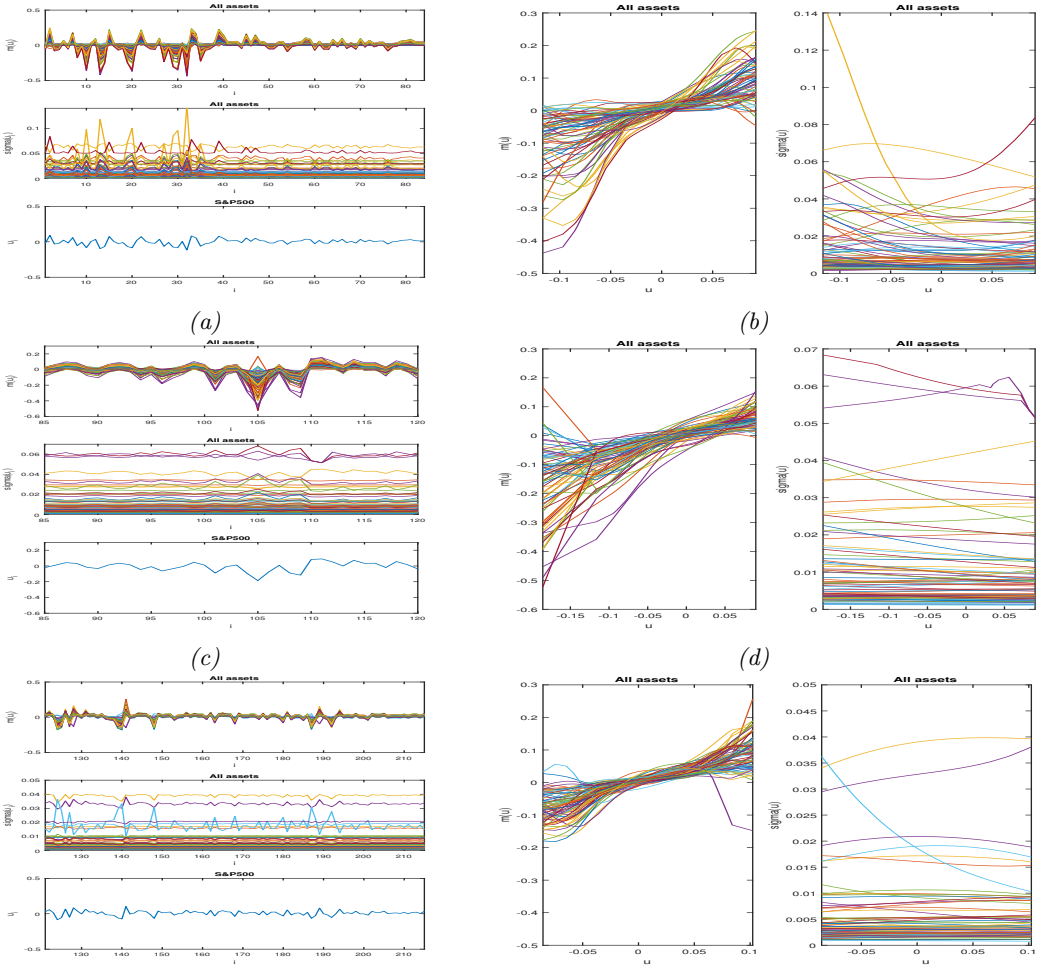


Figure 1: **Pattern changes in the three periods.** *Before-financial-crisis: (a) Plots of estimated means* $\hat{\mu}_k(u_i)$ *against* i *(top), estimated individual volatility* $\hat{\sigma}_{kk}(u_i)$ *against* i *(middle) and* u_i *against* i *(bottom). (b) Plots of estimated* $\hat{\mu}_k(u)$ *against* u *(left) and estimated individual volatility* $\hat{\sigma}_{kk}(u)$ *against* u *right. Similarly, (c) and (d) for the in-financial-crisis period while (e) and (f) for the after-financial-crisis.*

Table 1: The Average (standard error in %) of IRSE for Setting 1

n	p	DCM_2	DCM_1	sDCM_1	${}_{\mathrm{t}}\mathrm{NCM}_0$	${}_{\mathrm{st}}\mathrm{NCM}_0$	${}_{\mathrm{t}}\mathrm{NCM}_1$	${}_{\mathrm{st}}\mathrm{NCM}_1$
				$\rho = 0$				
100	50	5.13(25.0)	0.58(3.5)	0.56(3.5)	0.45(3.5)	0.45(3.6)	0.45(3.4)	0.44(3.5)
	100	16.21(47.0)	0.63(2.6)	0.61(2.5)	0.50(2.5)	0.49(2.6)	0.49(2.3)	0.48(2.4)
	150	49.43(75.6)	0.65(2.1)	0.64(2.1)	0.52(2.1)	0.51(2.2)	0.51(2.0)	0.50(2.0)
	300	78.04(48.3)	0.70(1.7)	0.69(1.7)	0.57(1.5)	0.57(1.6)	0.56(1.4)	0.55(1.4)
	500	102.88(39.1)	0.75(1.7)	0.74(1.7)	0.61(1.3)	0.60(1.4)	0.59(1.2)	0.58(1.2)
200	50	2.89(8.3)	0.37(2.7)	0.36(2.7)	0.28(2.4)	0.28(2.5)	0.28(2.4)	0.28(2.5)
	100	8.84(15.6)	0.39(1.8)	0.38(1.8)	0.30(1.6)	0.30(1.6)	0.30(1.5)	0.30(1.6)
	150	18.57(30.1)	0.39(1.6)	0.39(1.7)	0.30(1.5)	0.30(1.5)	0.30(1.5)	0.30(1.5)
	300	70.85(32.0)	0.42(1.3)	0.42(1.3)	0.32(1.0)	0.32(1.0)	0.32(1.0)	0.32(1.0)
	500	84.86(25.4)	0.45(1.1)	0.45(1.1)	0.33(0.9)	0.33(0.9)	0.33(0.9)	0.33(0.9)
500	50	1.58(3.6)	0.20(1.3)	0.20(1.4)	0.18(1.1)	0.18(1.2)	0.18(1.1)	0.18(1.2)
	100	3.37(4.4)	0.21(0.9)	0.21(0.9)	0.18(0.7)	0.18(0.8)	0.18(0.7)	0.18(0.8)
	150	6.06(6.5)	0.21(0.8)	0.21(0.8)	0.18(0.6)	0.18(0.6)	0.18(0.6)	0.18(0.6)
	300	28.71(24.4)	0.23(0.5)	0.23(0.5)	0.18(0.4)	0.19(0.4)	0.18(0.4)	0.18(0.4)
	500	91.00(20.6)	0.25(0.4)	0.25(0.4)	0.18(0.4)	0.19(0.4)	0.18(0.4)	0.19(0.4)

Table 2: The Average (standard error in %) of IRSE for Setting 1 (continued)

n	p	DCM_2	DCM_1	sDCM_1	${}_{\mathrm{t}}\mathrm{NCM}_0$	${}_{\mathrm{st}}\mathrm{NCM}_0$	${}_{\mathrm{t}}\mathrm{NCM}_1$	${}_{\mathrm{st}}\mathrm{NCM}_1$
				$\rho = 0.3$				
100	50	5.79(28.4)	0.62(3.4)	0.60(3.2)	0.49(2.9)	0.48(3.0)	0.49(2.9)	0.47(2.9)
	100	18.63(54.9)	0.67(3.0)	0.65(2.9)	0.54(2.5)	0.53(2.5)	0.53(2.3)	0.52(2.3)
	150	55.34(74.9)	0.71(2.8)	0.69(2.8)	0.56(2.1)	0.55(2.2)	0.56(1.9)	0.54(2.0)
	300	80.55(48.3)	0.76(2.1)	0.75(2.0)	0.62(1.6)	0.60(1.6)	0.60(1.5)	0.59(1.5)
	500	102.59(51.0)	0.82(3.0)	0.81(2.9)	0.65(1.1)	0.64(1.1)	0.64(1.0)	0.62(1.0)
200	50	3.05(9.4)	0.40(2.5)	0.39(2.5)	0.31(2.1)	0.31(2.2)	0.31(2.1)	0.31(2.1)
	100	8.17(15.7)	0.42(1.8)	0.41(1.8)	0.32(1.5)	0.32(1.5)	0.32(1.5)	0.32(1.5)
	150	17.63(28.1)	0.43(1.8)	0.42(1.7)	0.33(1.5)	0.33(1.5)	0.33(1.5)	0.33(1.5)
	300	72.98(32.4)	0.46(1.5)	0.46(1.5)	0.35(1.0)	0.35(1.0)	0.35(1.0)	0.35(1.0)
	500	93.09(29.2)	0.50(1.2)	0.49(1.2)	0.37(0.8)	0.37(0.9)	0.37(0.8)	0.37(0.8)
500	50	1.56(4.1)	0.22(1.2)	0.21(1.3)	0.19(1.0)	0.19(1.0)	0.19(0.9)	0.19(1.0)
	100	3.35(4.8)	0.22(1.0)	0.22(1.0)	0.19(0.7)	0.19(0.8)	0.19(0.7)	0.19(0.8)
	150	6.43(6.6)	0.22(0.9)	0.22(1.0)	0.19(0.7)	0.19(0.7)	0.19(0.7)	0.19(0.7)
	300	27.20(24.8)	0.24(0.6)	0.24(0.6)	0.19(0.5)	0.19(0.5)	0.19(0.5)	0.19(0.5)
	500	92.53(25.5)	0.26(0.4)	0.26(0.4)	0.19(0.4)	0.19(0.4)	0.19(0.4)	0.19(0.4)

Table 3: The Average (standard error in %) of IRSE for Setting 1 (continued)

n	p	DCM_2	DCM_1	sDCM_1	tNCM_0	stNCM_0	tNCM_1	stNCM_1
				$\rho = 0.8$				
100	50	5.65(36.5)	1.40(12.6)	1.19(11.0)	1.31(11.8)	1.13(10.4)	1.31(11.9)	1.13(10.5)
	100	18.08(51.4)	1.89(12.6)	1.57(11.2)	1.78(11.8)	1.49(10.5)	1.78(11.8)	1.48(10.6)
	150	55.78(61.0)	2.33(12.9)	1.93(11.7)	2.20(12.2)	1.81(10.9)	2.20(12.3)	1.80(11.0)
	300	80.82(41.6)	3.22(12.6)	2.66(11.5)	3.06(12.0)	2.48(10.9)	3.06(12.0)	2.47(10.9)
	500	102.70(42.6)	4.13(10.8)	3.38(10.0)	3.93(10.2)	3.15(9.4)	3.93(10.2)	3.15(9.4)
200	50	3.08(14.3)	1.11(7.8)	0.96(6.6)	1.03(7.1)	0.91(6.2)	1.03(7.2)	0.91(6.3)
	100	8.18(19.6)	1.50(7.3)	1.25(6.3)	1.40(6.8)	1.19(6.0)	1.40(6.8)	1.19(6.0)
	150	17.34(30.2)	1.83(6.7)	1.51(5.9)	1.71(6.2)	1.43(5.6)	1.71(6.2)	1.42(5.6)
	300	73.00(32.9)	2.54(5.9)	2.09(5.4)	2.40(5.5)	1.96(5.0)	2.40(5.6)	1.96(5.1)
	500	93.15(26.7)	3.25(6.9)	2.67(6.3)	3.08(6.4)	2.48(5.8)	3.08(6.5)	2.48(5.8)
500	50	1.62(6.2)	0.74(4.1)	0.66(3.2)	0.69(3.2)	0.63(2.7)	0.69(3.2)	0.63(2.7)
	100	3.38(6.2)	1.03(3.2)	0.88(2.5)	0.95(2.8)	0.84(2.4)	0.95(2.8)	0.84(2.4)
	150	6.48(8.9)	1.25(3.0)	1.05(2.5)	1.16(2.8)	1.00(2.5)	1.16(2.8)	1.00(2.5)
	300	26.59(28.8)	1.72(3.0)	1.42(2.7)	1.62(2.8)	1.35(2.5)	1.62(2.8)	1.35(2.5)
	500	92.73(25.0)	2.20(2.7)	1.80(2.3)	2.08(2.5)	1.71(2.2)	2.08(2.5)	1.71(2.2)

Table 4: The Average (standard error in %) of IRSE for Setting 2

n	p	DCM_2	DCM_1	sDCM_1	${}_{\mathrm{t}}\mathrm{NCM}_0$	${}_{\mathrm{st}}\mathrm{NCM}_0$	${}_{\mathrm{t}}\mathrm{NCM}_1$	${}_{\mathrm{st}}\mathrm{NCM}_1$
				$\rho = 0$				
	50	11.89(25.9)	0.53(1.9)	0.50(2.0)	0.45(2.3)	0.43(2.3)	0.43(2.5)	0.41(2.4)
	100	37.33(126.7)	0.55(1.4)	0.53(1.5)	0.48(1.5)	0.46(1.6)	0.46(1.9)	0.44(1.9)
100	150	62.90(59.7)	0.56(1.5)	0.54(1.5)	0.50(1.6)	0.48(1.7)	0.47(1.5)	0.45(1.5)
	300	87.77(59.1)	0.59(1.8)	0.57(1.9)	0.52(1.1)	0.51(1.1)	0.50(1.3)	0.48(1.3)
	500	114.85(50.7)	0.61(1.7)	0.59(1.7)	0.54(0.7)	0.53(0.8)	0.52(0.8)	0.50(0.8)
	50	6.12(21.5)	0.39(1.9)	0.37(1.8)	0.32(1.6)	0.31(1.6)	0.31(1.6)	0.30(1.6)
	100	16.99(23.0)	0.40(1.1)	0.39(1.1)	0.33(1.1)	0.32(1.1)	0.31(1.1)	0.31(1.1)
200	150	35.62(50.7)	0.40(1.2)	0.39(1.1)	0.33(1.0)	0.33(1.0)	0.32(0.9)	0.31(0.9)
	300	90.48(44.9)	0.42(1.1)	0.41(1.0)	0.34(0.8)	0.34(0.8)	0.32(0.8)	0.32(0.8)
	500	116.32(39.7)	0.44(0.9)	0.44(0.9)	0.35(0.7)	0.34(0.7)	0.33(0.7)	0.32(0.7)
	50	2.82(7.9)	0.26(0.9)	0.26(0.9)	0.24(0.9)	0.23(1.0)	0.23(0.9)	0.23(1.0)
	100	7.30(8.7)	0.27(0.5)	0.26(0.5)	0.24(0.5)	0.24(0.6)	0.24(0.5)	0.24(0.5)
500	150	13.83(9.6)	0.27(0.4)	0.27(0.5)	0.25(0.4)	0.24(0.4)	0.24(0.4)	0.24(0.4)
	300	84.40(81.0)	0.28(0.3)	0.28(0.3)	0.25(0.2)	0.25(0.3)	0.25(0.2)	0.24(0.2)
	500	117.46(26.2)	0.30(0.2)	0.30(0.2)	0.25(0.2)	0.25(0.2)	0.25(0.2)	0.25(0.2)

Table 5: The Average (standard error in %) of IRSE for Setting 2 (continued)

n	p	DCM_2	DCM_1	sDCM_1	${}_{\mathrm{t}}\mathrm{NCM}_0$	${}_{\mathrm{st}}\mathrm{NCM}_0$	${}_{\mathrm{t}}\mathrm{NCM}_1$	${}_{\mathrm{st}}\mathrm{NCM}_1$
				$\rho = 0.3$				
100	50	11.91(26.3)	0.55(2.3)	0.52(2.3)	0.48(3.0)	0.46(2.8)	0.47(3.3)	0.44(3.0)
	100	37.55(133.3)	0.56(1.6)	0.54(1.6)	0.50(1.6)	0.48(1.7)	0.49(1.9)	0.47(1.9)
	150	62.92(54.1)	0.58(1.6)	0.56(1.6)	0.52(1.4)	0.50(1.5)	0.50(1.5)	0.48(1.5)
	300	87.63(53.6)	0.61(1.9)	0.59(1.9)	0.55(0.8)	0.53(0.9)	0.53(1.0)	0.51(1.1)
	500	114.89(59.9)	0.63(2.2)	0.61(2.2)	0.56(0.7)	0.54(0.7)	0.55(0.8)	0.53(0.8)
200	50	6.22(21.0)	0.41(1.7)	0.39(1.6)	0.33(1.7)	0.32(1.6)	0.33(1.8)	0.32(1.7)
	100	16.92(21.6)	0.42(1.6)	0.40(1.6)	0.35(1.6)	0.34(1.6)	0.33(1.7)	0.32(1.6)
	150	35.65(52.2)	0.43(1.1)	0.41(1.1)	0.35(0.9)	0.34(0.9)	0.34(1.0)	0.33(1.0)
	300	90.40(44.1)	0.45(1.2)	0.44(1.2)	0.37(0.8)	0.36(0.8)	0.35(0.9)	0.34(0.9)
	500	116.35(37.4)	0.47(1.2)	0.46(1.2)	0.38(0.8)	0.37(0.8)	0.36(0.9)	0.35(0.9)
500	50	2.81(8.5)	0.27(0.8)	0.27(0.9)	0.25(0.8)	0.24(0.8)	0.24(0.8)	0.24(0.8)
	100	7.35(7.7)	0.28(0.7)	0.27(0.7)	0.25(0.5)	0.25(0.5)	0.25(0.5)	0.25(0.5)
	150	13.71(9.7)	0.28(0.5)	0.28(0.5)	0.25(0.4)	0.25(0.4)	0.25(0.4)	0.25(0.4)
	300	84.65(54.8)	0.29(0.4)	0.29(0.4)	0.25(0.3)	0.25(0.3)	0.25(0.3)	0.25(0.3)
	500	117.48(26.7)	0.30(0.3)	0.30(0.3)	0.25(0.2)	0.25(0.2)	0.25(0.2)	0.25(0.2)

Table 6: The Average (standard error in %) of IRSE for Setting 2 (continued)

n	p	DCM_2	DCM_1	sDCM_1	tNCM_0	stNCM_0	tNCM_1	stNCM_1
				$\rho = 0.8$				
	50	11.81(29.6)	1.31(10.5)	1.09(8.9)	1.23(9.3)	1.03(8.1)	1.22(9.4)	1.03(8.2)
	100	37.55(152.3)	1.78(11.2)	1.46(10.0)	1.68(10.5)	1.37(9.4)	1.67(10.6)	1.37(9.5)
100	150	62.89(57.0)	2.16(11.2)	1.76(10.1)	2.04(10.5)	1.65(9.4)	2.04(10.5)	1.65(9.5)
	300	87.59(52.0)	3.02(9.1)	2.46(8.4)	2.86(8.7)	2.29(7.9)	2.86(8.7)	2.29(8.0)
	500	114.95(56.6)	3.88(9.7)	3.15(9.0)	3.68(9.4)	2.93(8.7)	3.68(9.4)	2.93(8.7)
	50	6.00(26.4)	1.07(5.6)	0.89(4.9)	0.99(5.3)	0.84(4.8)	0.99(5.3)	0.84(4.8)
	100	16.69(29.2)	1.45(5.5)	1.18(4.8)	1.35(5.3)	1.12(4.7)	1.35(5.3)	1.12(4.7)
200	150	36.69(64.7)	1.77(5.9)	1.44(5.3)	1.66(5.6)	1.36(5.1)	1.66(5.6)	1.36(5.1)
	300	90.40(45.6)	2.46(4.9)	2.00(4.5)	2.33(4.6)	1.88(4.2)	2.33(4.6)	1.88(4.2)
	500	116.27(38.8)	3.15(6.0)	2.56(5.5)	2.99(5.7)	2.39(5.2)	2.99(5.7)	2.39(5.2)
	50	2.80(10.2)	0.73(3.2)	0.62(2.5)	0.67(2.7)	0.60(2.3)	0.67(2.7)	0.59(2.4)
	100	7.13(11.8)	1.01(2.7)	0.83(2.2)	0.93(2.5)	0.80(2.3)	0.93(2.5)	0.79(2.3)
500	150	13.40(13.8)	1.22(2.5)	1.00(2.2)	1.14(2.3)	0.95(2.1)	1.14(2.4)	0.95(2.2)
	300	82.86(86.7)	1.69(2.6)	1.37(2.3)	1.59(2.4)	1.31(2.2)	1.59(2.4)	1.30(2.2)
	500	117.47(26.9)	2.16(2.2)	1.75(1.9)	2.05(2.1)	1.66(1.8)	2.05(2.1)	1.66(1.9)

Table 7: The Average (standard error in %) of IRSE for Setting 3

n	p	DCM_2	DCM_1	sDCM_1	tNCM_0	stNCM_0	tNCM_1	stNCM_1
				$\rho = 0$				
	50	7.20(40.1)	3.52(16.4)	3.41(13.8)	3.03(13.5)	3.00(14.2)	2.99(22.5)	2.94(22.5)
	100	16.44(45.7)	3.80(7.3)	3.79(6.6)	3.25(5.5)	3.22(5.9)	3.19(5.7)	3.14(6.0)
100	150	50.77(57.9)	3.94(6.5)	3.92(6.1)	3.39(4.1)	3.36(4.3)	3.34(4.0)	3.29(4.1)
	300	78.24(68.1)	4.10(6.2)	4.08(5.9)	3.58(3.0)	3.54(2.9)	3.52(2.9)	3.47(2.9)
	500	102.97(69.5)	4.28(7.9)	4.25(7.8)	3.71(2.2)	3.67(2.2)	3.61(2.2)	3.56(2.2)
	50	4.77(22.4)	2.79(12.9)	2.72(10.9)	2.37(9.9)	2.35(10.2)	2.28(9.0)	2.25(9.1)
	100	10.37(27.3)	2.96(9.6)	2.92(8.6)	2.58(5.5)	2.57(5.7)	2.47(4.7)	2.45(4.8)
200	150	16.73(25.1)	3.50(5.8)	3.50(5.8)	2.72(4.5)	2.70(4.6)	2.59(3.5)	2.57(3.7)
	300	71.16(46.6)	3.70(3.7)	3.69(3.6)	2.99(3.0)	2.96(3.1)	2.81(2.5)	2.78(2.5)
	500	85.06(43.2)	3.98(4.5)	3.96(4.4)	3.17(2.1)	3.13(2.2)	2.95(1.8)	2.92(1.8)
	50	3.08(9.9)	2.06(6.4)	2.02(5.9)	1.61(6.2)	1.61(6.3)	1.55(5.2)	1.55(5.4)
	100	5.40(12.8)	2.18(5.8)	2.16(5.3)	1.87(4.3)	1.87(4.4)	1.74(3.8)	1.73(3.9)
500	150	7.97(9.3)	2.23(4.7)	2.21(4.5)	2.03(3.3)	2.03(3.4)	1.88(2.9)	1.87(3.0)
	300	20.07(14.9)	2.34(3.7)	2.33(3.6)	2.31(2.2)	2.31(2.3)	2.11(1.9)	2.10(2.0)
	500	90.35(26.8)	3.25(1.3)	3.26(1.3)	2.52(1.8)	2.51(1.8)	2.28(1.6)	2.27(1.6)

Table 8: The Average (standard error in %) of IRSE for Setting 3 (continued)

n	p	DCM_2	DCM_1	sDCM_1	tNCM_0	stNCM_0	tNCM_1	stNCM_1
				$\rho = 0.3$				
	50	7.97(43.3)	3.56(18.8)	3.44(15.6)	3.05(11.3)	3.01(11.9)	3.48(418.0)	3.43(412.5)
	100	18.17(47.0)	3.84(8.4)	3.81(7.8)	3.27(5.2)	3.23(5.5)	3.22(5.5)	3.17(5.7)
100	150	56.43(68.4)	3.97(7.8)	3.95(7.2)	3.41(4.3)	3.38(4.6)	3.38(4.6)	3.33(4.7)
	300	80.49(72.2)	4.13(7.5)	4.10(7.1)	3.58(2.9)	3.54(2.8)	3.56(2.7)	3.50(2.6)
	500	102.53(77.3)	4.28(9.9)	4.25(9.8)	3.70(2.3)	3.66(2.3)	3.65(2.2)	3.60(2.2)
	50	4.79(21.0)	2.83(11.5)	2.76(9.7)	2.40(8.5)	2.37(8.8)	2.32(7.6)	2.29(7.9)
	100	9.83(23.5)	3.05(10.0)	3.00(9.2)	2.60(5.8)	2.58(5.9)	2.50(4.6)	2.47(4.7)
200	150	16.20(23.9)	3.55(6.4)	3.54(6.5)	2.73(4.9)	2.71(5.1)	2.62(4.4)	2.59(4.5)
	300	73.19(46.7)	3.75(4.2)	3.74(4.1)	2.99(3.1)	2.96(3.1)	2.83(2.7)	2.80(2.7)
	500	93.02(41.0)	4.00(4.9)	3.98(4.8)	3.15(2.7)	3.12(2.8)	2.97(1.9)	2.93(2.0)
	50	2.99(12.1)	2.10(8.0)	2.06(7.1)	1.64(5.7)	1.64(5.8)	1.58(5.0)	1.57(5.0)
	100	5.38(15.9)	2.22(6.1)	2.19(5.5)	1.89(4.4)	1.88(4.6)	1.76(3.8)	1.76(3.9)
500	150	7.97(7.9)	2.27(5.0)	2.25(4.7)	2.05(3.5)	2.04(3.6)	1.90(3.0)	1.89(3.1)
	300	18.99(14.3)	2.38(4.4)	2.37(4.2)	2.32(2.4)	2.31(2.4)	2.12(2.0)	2.11(2.1)
	500	91.60(30.6)	3.28(1.8)	3.28(1.8)	2.51(1.6)	2.50(1.7)	2.29(1.5)	2.28(1.5)

Table 9: The Average (standard error in %) of IRSE for Setting 3 (continued)

n	p	DCM_2	DCM_1	sDCM_1	tNCM_0	stNCM_0	tNCM_1	stNCM_1
				$\rho = 0.8$				
	50	8.26(58.0)	4.64(51.0)	4.19(35.7)	4.18(30.2)	3.72(17.9)	4.17(30.9)	3.70(18.8)
	100	18.14(49.4)	5.98(41.0)	5.19(28.0)	5.30(31.2)	4.43(19.0)	5.28(31.0)	4.41(19.1)
100	150	50.71(628.7)	7.11(50.6)	5.99(37.3)	6.30(35.0)	5.07(22.5)	6.29(35.0)	5.06(22.9)
	300	78.96(66.1)	9.36(36.8)	7.49(28.5)	8.43(34.4)	6.43(23.2)	8.41(34.3)	6.41(23.3)
	500	101.14(67.3)	11.57(33.2)	9.07(28.0)	10.58(29.2)	7.83(22.3)	10.56(29.4)	7.81(22.7)
	50	5.54(37.8)	3.84(40.0)	3.56(29.0)	3.53(20.5)	3.13(13.0)	3.47(22.9)	3.09(15.3)
	100	10.37(25.5)	4.81(44.4)	4.46(32.0)	4.50(17.6)	3.80(10.8)	4.43(19.2)	3.73(12.4)
200	150	16.49(28.1)	6.22(44.6)	5.40(32.2)	5.31(15.4)	4.32(9.7)	5.24(16.7)	4.26(11.0)
	300	70.79(46.4)	8.23(20.6)	6.66(16.6)	7.04(15.8)	5.43(11.3)	6.97(16.5)	5.37(12.1)
	500	91.21(37.8)	10.10(20.8)	8.00(16.9)	8.80(17.8)	6.58(13.2)	8.74(18.3)	6.53(13.8)
	50	3.45(21.6)	2.64(20.0)	2.53(16.9)	2.46(14.3)	2.30(11.0)	2.43(15.3)	2.28(11.9)
	100	6.01(16.1)	3.00(20.6)	2.93(18.8)	3.26(9.8)	2.92(5.6)	3.21(10.9)	2.88(6.8)
500	150	8.25(11.4)	3.30(18.7)	3.24(17.2)	3.84(7.0)	3.33(4.4)	3.78(7.8)	3.28(5.1)
	300	19.17(15.3)	4.23(36.2)	4.15(32.2)	5.03(7.7)	4.13(5.3)	4.98(8.5)	4.10(6.2)
	500	40.05(34.0)	7.23(8.4)	5.91(6.4)	6.21(7.1)	4.91(4.9)	6.18(7.8)	4.91(5.6)

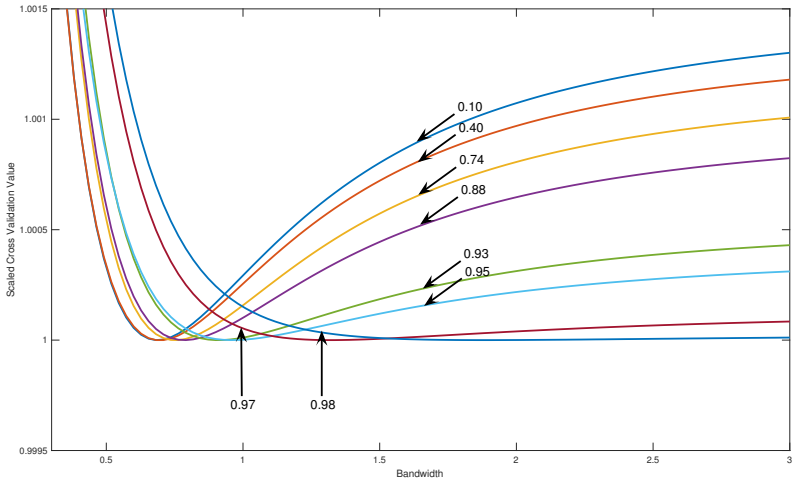
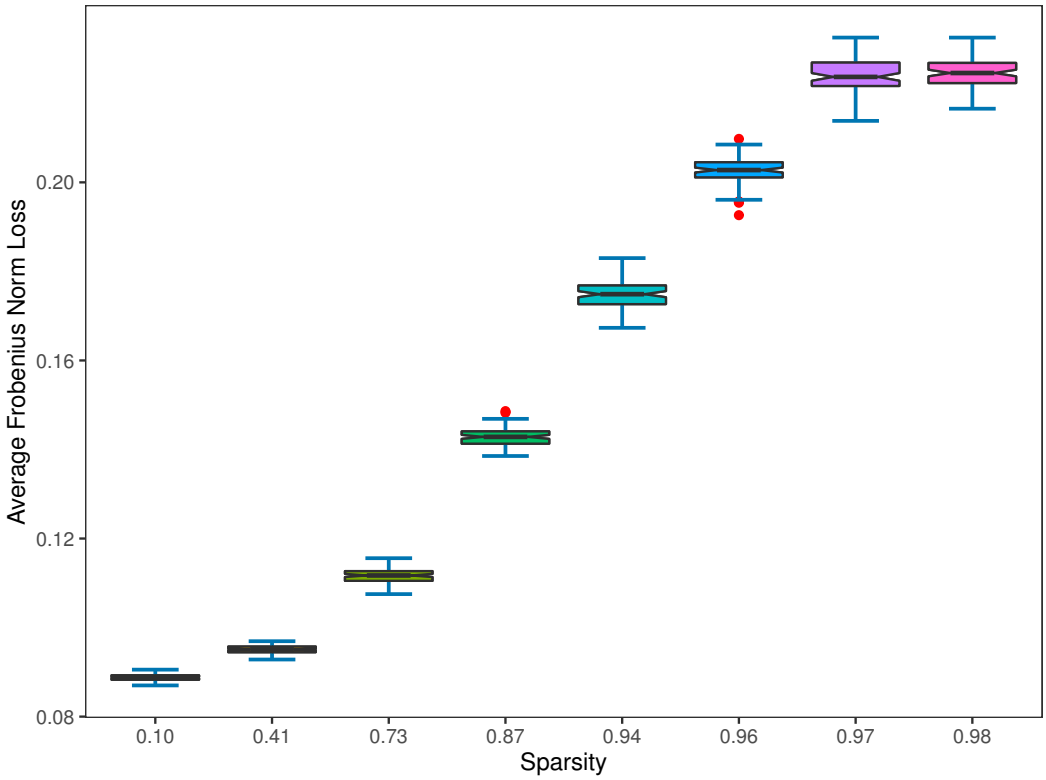
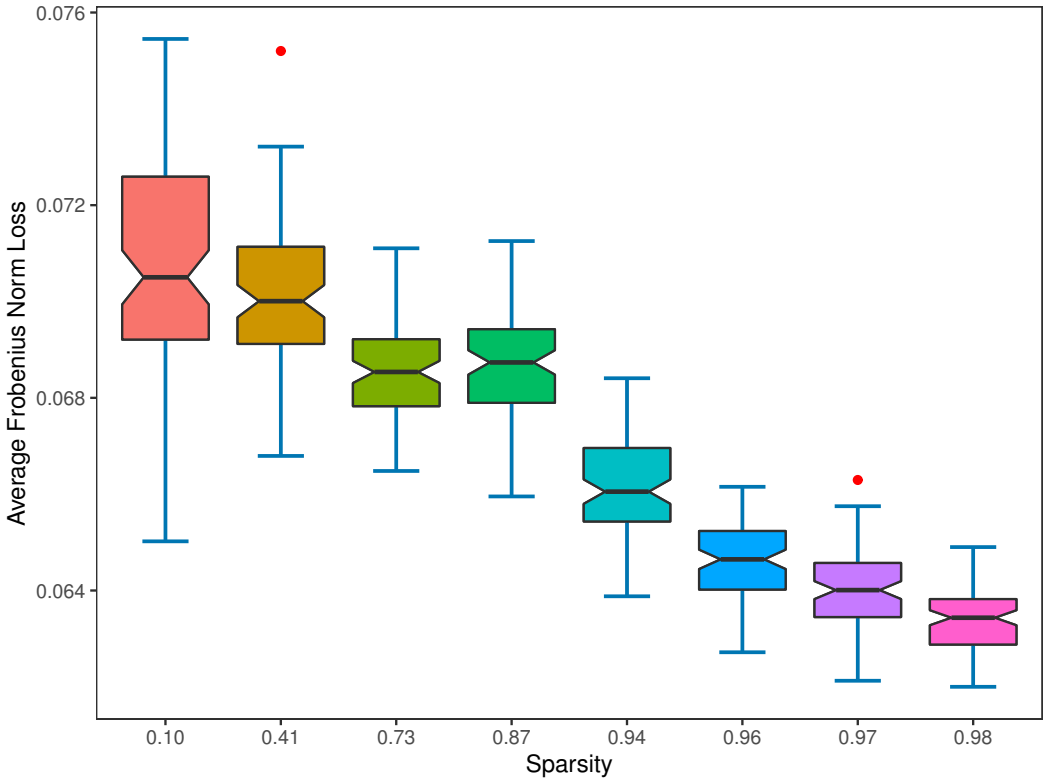


Figure 2: Cross-validation curves with different sparsities. The number attached to each curve is the associated sparsity index S_{C_0}.

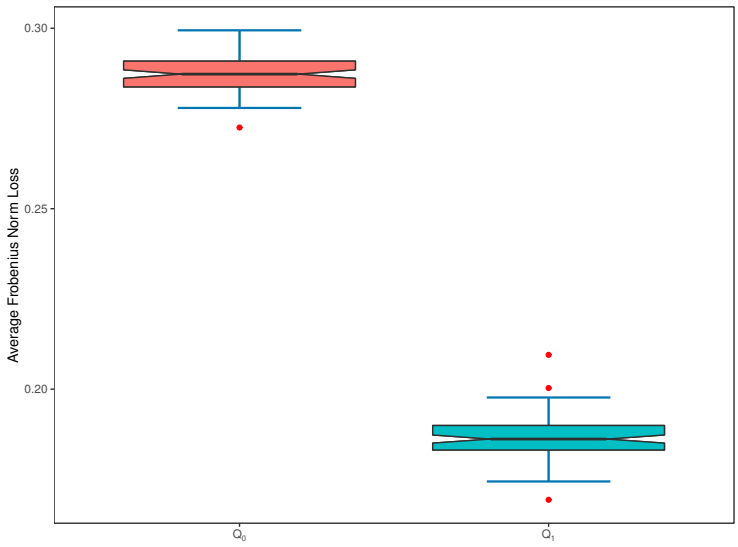


(a) Non-zero entries

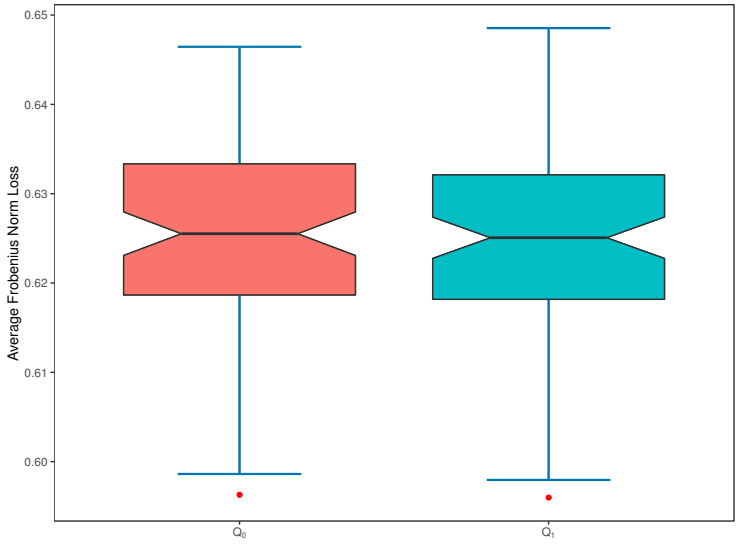


(b) Zero entries

Figure 3: Frobenius loss for different sparsities.



(a) Non-zero entries



(b) Zero entries

Figure 4: Frobenius losses for the methods based on Q_0 and Q_1Q_0. In each panel, the left box-plot for Q_0-based estimate and the right box-plot for Q_1Q_0-based estimate.