risks

MDPI

Article

L_1 Regularization for High-Dimensional Multivariate GARCH Models

Sijie Yao [1], Hui Zou [2] and Haipeng Xing [3,*]

[1] Department of Biostatistics and Bioinformatics, H. Lee Moffitt Cancer Center & Research Institute, Tampa, FL 33612, USA; sijie.yao@moffitt.org
[2] School of Statistics, University of Minnesota, Minneapolis, MN 55455, USA
[3] Department of Applied Mathematics and Statistics, State University of New York at Stony Brook, Stony Brook, NY 11733, USA
* Correspondence: haipeng.xing@stonybrook.edu; Tel.: +1-631-632-1892

Abstract: The complexity of estimating multivariate GARCH models increases significantly with the increase in the number of asset series. To address this issue, we propose a general regularization framework for high-dimensional GARCH models with BEKK representations, and obtain a penalized quasi-maximum likelihood (PQML) estimator. Under some regularity conditions, we establish some theoretical properties, such as the sparsity and the consistency, of the PQML estimator for the BEKK representations. We then carry out simulation studies to show the performance of the proposed inference framework and the procedure for selecting tuning parameters. In addition, we apply the proposed framework to analyze volatility spillover and portfolio optimization problems, using daily prices of 18 U.S. stocks from January 2016 to January 2018, and show that the proposed framework outperforms some benchmark models.

Keywords: Markov chain Monte Carlo; multivariate GARCH; spillover; stochastic approximation

Citation: Yao, Sijie, Hui Zou, and Haipeng Xing. 2024. L_1 Regularization for High-Dimensional Multivariate GARCH Models. *Risks* 12: 34. https://doi.org/10.3390/risks12020034

Academic Editor: Mogens Steffensen

Received: 23 December 2023
Revised: 27 January 2024
Accepted: 31 January 2024
Published: 4 February 2024

1. Introduction

Modeling the dynamics of high-dimensional variance–covariance matrices is a challenging problem in high-dimensional time series analysis and has wide applications in financial econometrics. Classical time series models for variance–covariance matrices assume that the number of component time series is low with respect to the number of observed samples. However, many financial and economic applications these days need to model the dynamics of high-dimensional variance–covariance matrices. For example, in modern portfolio management, the number of assets can easily be more than thousands and be larger or on the same order as the observed historical prices of the assets; in analyzing the movements in the financial markets of different products in different countries, it is critical to understand the interdependence and contagion effects of price movements over thousands of markets, while the amounts of jointly observed financial data are only available in decades.

In this paper, we propose an inference procedure with L_1 regularization for high-dimensional BEKK representations and obtain a class of penalized quasi-maximum likelihood (PQML) estimators. The L_1 regularization allows us to identify important parameters and shrink the non-essential ones to zero, hence providing an estimate of sparse parameters in BEKK representations. Under some regularity conditions, we establish some theoretical properties, such as the sparsity and the consistency, of the PQML estimator for BEKK representations. The proposed procedure is a fairly general framework that can be applied to a large class of high-dimensional MGARCH models; by applying our regularization techniques, the complexity of making inferences from high-dimensional MGARCH models can be greatly reduced and the intrinsic sparse model structures can be uncovered. We carried out simulation studies to show the performance of the proposed inference

framework and the procedure for selecting tuning parameters. In addition, we applied the proposed framework to analyze volatility spillover and portfolio optimization problems, using daily prices of 18 U.S. stocks from January 2016 to January 2018. In the comparison of portfolio optimization based on different MGARCH models, we show that the proposed framework outperforms three benchmark models, i.e., the constant covariance model, the factor MGARCH model, and the dynamic conditional correlation model.

The proposed framework can be viewed as an extension of the literature on regularization techniques for converting high-dimensional linear models to nonlinear time series models. Since Tibshirani (1996) introduced LASSO for linear regression models, various regularization techniques concerning high-dimensional statistical inference have been studied for various problems in linear models. For example, Fan and Li (2001) proposed the smoothly clipped absolute deviation (SCAD) penalty that generates sparse estimation of regression coefficients with reduced bias and explored the so-called "oracle property", in which the estimator has asymptotic properties that are equivalent to the maximum likelihood estimator in the non-penalized model. Zou (2006) proposed adaptive LASSO by adding adaptive weights for different parameters in the L_1 penalty to obtain better estimator performance. Yuan and Lin (2006) proposed a group LASSO penalty to solve the problem of selecting grouped factors in regression models. Zhang (2010) proposed a minimax concave penalty that gives nearly unbiased variable selection in linear regression. In addition to discussions on regularized estimation in high-dimensional statistics, which relies primarily on independent and identically distributed (i.i.d.) samples and linear models, regularization techniques have also been applied to study inference problems in high-dimensional linear time-series models. For instance, Uematsu (2015) studied a class of penalty functions and showed the oracle properties for the estimators in high-dimensional vector autoregressive (VAR) models. Basu and Michailidis (2015) investigated the theoretical properties of L_1-regularized estimates in high-dimensional stochastic regressions with serially correlated errors and transition matrix estimation in high-dimensional VAR models. Sun and Lin (2011) developed a regularization framework for full-factor MGARCH models (Vrontos et al. 2003), in which the dynamics of covariance matrices are determined by the dynamics of univariate GARCH processes for orthogonal factors. Using the group LASSO technique, Poignard (2017) studied the inference problem for MGARCH models with vine structure, an alternative to dynamic conditional correlation MGARCH models.

The proposed regularization framework is also related to the problem of estimating $p \times p$ covariance matrices using various shrinkage and regularization methods. For instance, Ledoit and Wolf (2004) proposed an optimal linear shrinkage method to estimate constant covariance matrices of p-dimensional i.i.d. vectors, and, later on, Ledoit and Wolf (2012) extended the method and developed nonlinear shrinkage estimators for high-dimensional covariance matrices. Bickel and Levina (2008) and Cai and Liu (2011) proposed covariance regularization procedures that are based on the thresholding of sample covariance matrices to estimate inverse covariance matrices. Lam and Fan (2009) studied sparsistency and rates of convergence for estimating covariance based on penalized likelihood with nonconcave penalties, and Ravikumar et al. (2011) estimated high-dimensional inverse covariance by minimizing L_1-penalized log-determinant divergence. This method is also called graphical LASSO and was studied in Yuan and Lin (2006) and Friedman et al. (2007). We note that all these discussions focus on high-dimensional constant covariance matrices; thus, they do not involve the dynamics of covariance matrices.

The remainder of the paper is organized as follows. Section 2 provides a literature review of MGARCH models and their applications in volatility spillover. Section 3 explains the BEKK model with L_1-penalty functions in detail. In Section 4, we provide theoretical properties and implementation procedures for the regularized BEKK model. Simulation results and real data analysis are presented in Sections 5 and 6, respectively. Section 7 gives concluding remarks.

2. Literature Review

Inspired by the idea of univariate generalized autoregressive conditionally heteroskedastic (GARCH) models Bollerslev (1986); Engle (1982); Francq and Zakoian (2019); Hafner et al. (2022), various multivariate GARCH (MGARCH) models were proposed to characterize the dynamics of covariance matrices during the last three decades. Among these MGARCH models, the Baba–Engle–Kraft–Kroner (BEKK) model (Engle and Kroner 1995) uses a general specification to describe the dynamics of covariance matrices of an n-dimensional multivariate time series. Since such a specification contains unknown parameters of order $O(n^2)$, inference on the BEKK model becomes complicated, even for not very large ns. When n^2 increases with the same order as, or larger order than, the length of the time series, inference on the MGARCH–BEKK representation becomes even more difficult due to "the curse of dimensionality".

To reduce the complexity of inference procedures for unknown parameters in MGARCH models, other forms of MGARCH specifications were proposed to reduce the number of unknown parameters in the model. An important improvement to MGARCH models is the dynamic conditional correlation (DCC) model (Aielli 2013; Bauwens and Laurent 2005; Boudt et al. 2013; Engle 2002). The DCC model allows for time-varying conditional correlations and reduces the dimensionality by factorizing the conditional covariance matrix into the product of a diagonal matrix of conditional standard deviations and a correlation matrix that evolves dynamically over time. Other forms of MGARCH specifications make more assumptions on structures and dynamics of covariance matrices and include, for example, the MGARCH in mean model (Bollerslev et al. 1988), the constant conditional correlation GARCH model (Bollerslev 1990; Ling and McAleer 2003; McAleer et al. 2009), the time-varying conditional correlation MGARCH model (Tse and Tsui 2002), the orthogonal factor MGARCH model (Hafner and Preminger 2009; Lanne and Saikkonen 2007), and so on. Although these MGARCH models provide relatively simple inference procedures, the assumptions on dynamics of covariance matrices are usually too specific to capture the complexity of dynamics of covariance matrices. Furthermore, these models still fail to address the issue of making inference on high-dimensional MGARCH models.

In addition to modeling the joint behavior of volatilities for a set of returns, another aspect of MGARCH models is to characterize volatility spillover in financial markets. Volatility spillover refers to as the process and magnitude by which the instability in one market affects other markets. Volatility spillover is widely observed in equity markets (Hamao et al. 1990), bond markets (Christiansen 2007), futures markets (Pan and Hsueh 1998), exchange markets (Baillie and Bollerslev 1990), markets of equities and exchanges (Apergis and Rezitis 2001), various industries and commodities (Apergis and Rezitis 2003; Kaltenhäuser 2002), and so on. Understanding volatility spillover can provide an insight into financial vulnerabilities, as well as the source and nature of financial exposures, for academic researchers, financial practitioners, and regulatory authorities. For investors, as significant volatility spillover may increase non-systemic risk, understanding volatility spillover can help them diversify the risks associated with their investment. For financial sector regulators, understanding volatility spillover can help them formulate appropriate policies to maintain financial stability, especially when stress from a particular market is transmitted to other markets, such that the risk of systemic instability increases. MGARCH models are generally used to characterize volatility spillover in the markets, which are represented via a low-dimensional multivariate series; see Hamao et al. (1990), Christiansen (2007), Pan and Hsueh (1998), Engle et al. (1990), and Baillie and Bollerslev (1990). In particular, Theodossiou and Lee (1993) used multivariate GARCH-in-mean model to study the economic spillover effect across five countries, Worthington and Higgs (2004) applied a BEKK(1,1) model to study transmission of weekly equity returns and volatility in nine Asian countries from 1988 to 2000, and Hassan and Malik (2007) employed the BEKK(1,1) specification to study three-dimensional US sector indices. Spillover effect has also been explored recently for other financial markets, such as cryptocurrency markets (Billio et al. 2023) and European banks with GARCH models (Giacometti et al. 2023). Additionally,

there has been an investigation into the spillover effects using network representations derived from GARCH models in recent studies (Ampountolas 2022; Hong et al. 2023).

The aforementioned studies on spillover effects rely on the foundational structures of the DCC model for analysis (Ampountolas 2022; Shiferaw 2019; Siddiqui and Khan 2018). Although these MGARCH models provide relatively simple inference procedures, the assumptions on dynamics of covariance matrices are usually too specific to capture the complexity of the dynamics of covariance matrices. Moreover, these models still fail to address the issue of making inference on high-dimensional MGARCH models. Under these constraints, the performance and accuracy of these simplified MGARCH models need further investigation in real markets (Engle and Colacito 2006).

3. The MGARCH–BEKK Representations with L_1 Regularization

We first introduce the following notations. Given a vector x and a matrix A, the ith component of x and the ijth elements of A are written as x_i and A_{ij}, respectively. The jth column and the ith row vectors of A are denoted as $A_{\cdot j}$ and $A_{i\cdot}$, respectively. $||x||$ is the Euclidean norm for vector x. $||x||_\infty$ is the largest element of x in the modulus. $\rho(A)$ is the spectral radius of A, i.e., the largest modulus of eigenvalues of A. $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ are the minimum and maximum eigenvalues of A, respectively. $||A||$ is the spectral norm, i.e., a square root of $\rho(A^TA)$. $||A||_\infty$ represents the operator norm induced by $||x||_\infty$, or the largest absolute row sum. For any matrix A and vector x such that Ax is well defined, let $||A||_{2,\infty} := \max_{||x||=1} ||Ax||_\infty$. We use $\text{sign}(x)$ to denote the sign of x : $\text{sign}(x) = x/|x|$ if $x \neq 0$, and $\text{sign}(x) = 0$ otherwise.

3.1. The MGARCH–BEKK Representation

Let r_t be the vector of returns on n assets in period t. Let ϵ_t be i.i.d. n-dimensional standard normal random vectors. Let $\mathcal{F}_t$ be the sigma field generated by the past information from r_ts. Then, Σ_t is measurable with respect to $\mathcal{F}_{t-1}$; the distribution of r_t can be specified as

$$r_t = \Sigma_t^{-\frac{1}{2}}\epsilon_t, \qquad \epsilon_t \sim N(0, I_n), \tag{1}$$

where I_n is an $n \times n$ identity matrix. Denote the conditional covariance matrix of r_t given $\mathcal{F}_{t-1}$ as Σ_t, i.e., $\Sigma_t = \text{Cov}(r_t|\mathcal{F}_{t-1})$. Engle and Kroner (1995) proposed the following BEKK(a, b) model to characterize the dynamics of Σ_t:

$$\Sigma_t = C'C + \sum_{k=1}^{K}\sum_{i=1}^{a} A_{ik} r_{t-1} r'_{t-1} A'_{ik} + \sum_{k=1}^{K}\sum_{i=1}^{b} B_{ik}\Sigma_{t-1}B'_{ik}, \tag{2}$$

where A_{ik}, and B_{ik} are $n \times n$ parameter matrices, C is an $n \times n$ triangular matrix, and the summation limit K determines the generality of the process.

To illustrate the idea, we consider BEKK(1,1) in our examples with $K = 1$ in this paper, which can be written as

$$\Sigma_t = C'C + \sum_{i=1}^{a} A_i r_{t-i} r'_{t-i} A'_i + \sum_{i=1}^{b} B_i \Sigma_{t-i} B'_i. \tag{3}$$

in which A_i, B_i, and C are real $n \times n$ matrices. And, without loss of generality, we choose $\Sigma_t^{-1/2}$ to be symmetric. For identification purposes, Engle and Kroner (1995) showed the following property for the BEKK model.

Proposition 1. *Suppose that the diagonal elements in C, a_{11}, and b_{11} are positive. Then, there exists no other C, A, or B in Model (3) that gives an equivalent representation.*

Proposition 1 is also known as the identification condition (Comte and Lieberman 2003).

Let *vec* and *vech* be the vector operators that stack the columns of a matrix and the lower triangular part of a matrix, respectively. That is, if

$$Y = \begin{pmatrix} y_{11} & \cdots & y_{1n} \\ \vdots & \ddots & \vdots \\ y_{n1} & \cdots & y_{nn} \end{pmatrix},$$

then

$$vec(Y) = (y_{11}, \cdots, y_{n1}, y_{12}, \cdots, y_{n2}, \cdots, y_{1n}, \cdots, y_{nn})',$$

and

$$vech(Y) = (y_{11}, \cdots, y_{n1}, y_{22}, \cdots, y_{n2}, \cdots, y_{ii}, \cdots, y_{in}, \cdots y_{nn})'.$$

Then, Model (3) can be rewritten in a vector form:

$$vec(\Sigma_t) = vec(C'C) + \sum_{i=1}^{a} \mathring{A}_i vec(r_{t-i} r'_{t-i}) + \sum_{i=1}^{b} \mathring{B}_i vec(\Sigma_{t-i}). \tag{4}$$

in which $\mathring{A}_i = A_i \otimes A_i$, $\mathring{B}_i = B_i \otimes B_i$, and $\otimes$ is the Kronecker product. Since the covariance matrices Σ_t are symmetric, we can also write (3) in the vector-half form:

$$vech(\Sigma_t) = vech(C'C) + \sum_{i=1}^{a} \widetilde{A}_i vech(r_{t-i} r'_{t-i}) + \sum_{i=1}^{b} \widetilde{B}_i vech(\Sigma_{t-i}). \tag{5}$$

where $\widetilde{A_i} = L_n \mathring{A}_i K'_n$, $\widetilde{B}_i = L_n \mathring{B}_i K'_n$, and L_n and K_n are matrices of dimension $n(n+1) \times n^2$ extracting the upper triangular parts of symmetric matrices $\mathring{A}_i$ and $\mathring{B}_i$. Note that $\dim(vec(\Sigma_t)) = n^2$ and $\dim(vech(\Sigma_t)) = n(n+1)/2$. For convenience, we denote $\theta = (\theta_1, \ldots, \theta_p)'$ by the parameter vector in Model (3), in which $p = 2(a+b)n^2 + n(n+1)/2$, so that the matrices C, A_i, and B_i are functions of θ: $C = C(\theta), A_i = A_i(\theta), B_i = B_i(\theta)$. And we denote by θ^0 the true parameter vector of the model.

We assume that the values of r_t in (1) are stationary; then, the following stationary condition should be imposed for the BEKK(a, b) Model (5) (see Engle and Kroner (1995) and Comte and Lieberman (2003)).

Condition 1 (**Stationary Condition**). *The p-dimensional return series r_t in (1) is stationary if the following conditions hold for Model (3):*

(i) $C^*(\theta) = C'C$ *is a continuous function of* θ, *and there exists* $C_0 > 0$, $\det(C^*(\theta)) \geq C_0$, *where* $\det(\cdot)$ *represents the determinant of a matrix;*

(ii) *For any* θ, $\widetilde{A}_i(\theta)$ *and* $\widetilde{B}_i(\theta)$ *are continuous functions of* θ;

(iii) *For any* θ, $\rho(\sum_{i=1}^{a} \widetilde{A}_i(\theta) + \sum_{i=1}^{b} \widetilde{B}_i(\theta)) < 1$, *i.e., the largest modulus of eigenvalues of* $\sum_{i=1}^{a} \widetilde{A}_i(\theta) + \sum_{i=1}^{b} \widetilde{B}_i(\theta)$ *is less than 1.*

3.2. Likelihood Function

In this section, we discuss some properties of the likelihood of the BEKK(a, b) model. Assume that ϵ_t follows a standard n-dimensional Gaussian distribution. Ignoring constants, we can write the quasi-log-likelihood as

$$\ell_T(\theta) = \frac{1}{2T} \sum_{t=1}^{T} l_t(\theta), \qquad l_t(\theta) = -(\log[\det(\Sigma_t)] + r'_t \Sigma_t^{-1} r_t). \tag{6}$$

Taking the derivative on Σ_t with respect to the ith element of θ, we obtain

$$\frac{\partial \Sigma_t}{\partial \theta_i} = \frac{\partial C'C}{\partial \theta_i} + \sum_{j=1}^{a}\left(\frac{\partial A_j}{\partial \theta_i} r_{t-j} r'_{t-j} A'_j + A_j r_{t-j} r'_{t-j} \frac{\partial A'_j}{\partial \theta_i}\right) + \sum_{j=1}^{b}\left(\frac{\partial B_j}{\partial \theta_i}\Sigma_{t-j} B_j + B_j \Sigma_{t-j} \frac{\partial B'_j}{\partial \theta_i} + B_j \frac{\partial \Sigma_{t-j}}{\partial \theta_i} B'_j\right), \tag{7}$$

which can be computed recursively. The derivative in (7) has the following property (the proof is given in Appendix A).

Proposition 2. *Let* $\mathfrak{R}_t = vech(r_t r'_t)$*; then,*

$$\left\| \frac{\partial \Sigma_t}{\partial \theta_i} \right\| \leq \Psi_1 + \Psi_2 \cdot \sup ||\mathfrak{R}_t||.$$

where Ψ_1 *and* Ψ_2 *are two constants.*

Assume that $L_T(\theta)$ is twice continuously differentiable in a neighborhood $\Theta^0 \in \Theta$ of θ^0. We define the averages of the score vector and Hessian matrix as follows:

$$S_T(\theta) = T^{-1}\sum_{t=1}^{T} s_t(\theta) \quad \text{and} \quad H_T(\theta) = T^{-1}\sum_{t=1}^{T} h_t(\theta),$$

where $s_t(\theta) = \partial l_t(\theta)/\partial\theta$ and $h_t(\theta) = \partial^2 l_t(\theta)/\partial\theta\partial\theta'$. Taking the derivative of (6) with respect to θ_i yields

$$\frac{\partial l_t}{\partial \theta_i} = -Tr\left(\frac{\partial \Sigma_t}{\partial \theta_i}\Sigma_t^{-1} - r_t r'_t \Sigma_t^{-1}\frac{\partial \Sigma_t}{\partial \theta_i}\Sigma_t^{-1}\right),$$

$$\begin{aligned}\frac{\partial^2 l_t(\theta)}{\partial \theta_j \partial \theta_i} &= -Tr\left(\frac{\partial^2 \Sigma_t}{\partial \theta_j \partial \theta_i}\Sigma_t^{-1} - \frac{\partial \Sigma_t}{\partial \theta_i}\Sigma_t^{-1}\frac{\partial \Sigma_t}{\partial \theta_j}\Sigma_t^{-1} + r_t r'_t \Sigma_t^{-1}\frac{\partial \Sigma_t}{\partial \theta_j}\Sigma_t^{-1}\frac{\partial \Sigma_t}{\partial \theta_i}\Sigma_t^{-1}\right.\\ &\quad \left. - r_t r'_t \Sigma_t^{-1}\frac{\partial^2 \Sigma_t}{\partial \theta_j \partial \theta_i}\Sigma_t^{-1} + r_t r'_t \Sigma_t^{-1}\frac{\partial \Sigma_t}{\partial \theta_i}\Sigma_t^{-1}\frac{\partial \Sigma_t}{\partial \theta_j}\Sigma_t^{-1}\right),\end{aligned}$$

in which $Tr(\cdot)$ represents the trace of a matrix. Comte and Lieberman (2003) showed the following property for $l_t(\theta)$.

Proposition 3. *Under Condition 1, the following properties hold:*

(i) *When* $T \to +\infty$, $-H_T^0 := -\frac{1}{T}\sum_{t=1}^{T}\frac{\partial l_t^2(\theta^0)}{\partial\theta\partial\theta'} \xrightarrow{P} H$ *for a nonrandom positive-definite matrix* H*;*

(ii) *For the Fisher information matrix* $I_0 := E\left(\frac{\partial l_t(\theta^0)}{\partial\theta}\cdot\frac{\partial l_t(\theta^0)}{\partial\theta'}\right) = E(S_T^0(S_T^0)')$, $||I_0||_\infty < \infty$*;*

(iii) *For* $\theta \in \Theta$, $E\left(\sup_{||\theta-\theta^0||\leq\epsilon}\left|\frac{\partial^3 l_t(\theta)}{\partial\theta_i\partial\theta_j\partial\theta_k}\right|\right)$ *is bounded for all* $\epsilon > 0$ *and* $i, j, k = 1, \ldots, p$*.*

In the sparse representation, the majority elements of the true parameter vector θ^0 are exactly 0. Hence, we could partition θ^0 into two sub-vectors. Let $\mathscr{U}_0$ be the set of indices $\{j \in \{1, \ldots, p\} : \theta_j^0 \neq 0\}$ and $\theta^0_{\mathscr{U}_0}$ be the q-dimensional vector composed of the nonzero elements $\{\theta_j^0 \neq 0 : j \in \mathscr{U}_0\}$. Similarly, we define $\theta^0_{\mathscr{U}_0^c}$ as a $(p-q)$-dimensional zero vector. Without loss of generality, θ^0 is stacked as $\theta^0 = ((\theta^0_{\mathscr{U}_0})', 0') = ((\theta^0_{\mathscr{U}_0})', (\theta^0_{\mathscr{U}_0^c})')$. For convenience, we define the average of the "score subvector" $S_{\mathscr{U}_0,T}(\theta)$ and the "Hessian sub-

matrix" $H_{\mathscr{U}_0,T}(\theta)$ by $s_{\mathscr{U}_0,t}(\theta) = \partial l_t(\theta)/\partial\theta_{\mathscr{U}_0}$ and $h_{\mathscr{U}_0,t}(\theta) = \partial^2 l_t(\theta)/\partial\theta_{\mathscr{U}_0}\partial\theta'_{\mathscr{U}_0}$. Similarly, we define $S_{\mathscr{U}_0^c,T}(\theta)$. We also denote $S_T(\theta^0) = S_T(\theta^0_{\mathscr{U}_0}, 0)$ as S_T^0.

Proposition 4. *The quasi-log-likelihood function L_T for the BEKK(1,1) has the following properties:*

(i) *For $i = 1, \ldots, p$, $E\left(|T \cdot S^0_{T,i}|^4\right) < \infty$, where $S^0_{T,i}$ is the ith element of S^0_T;*

(ii) *For a sufficiently large T, $-H^0_{\mathscr{U}_0,T}$ is almost surely positive definite, and $\lambda_{\min}(-H^0_{\mathscr{U}_0,T}) = O_p(1)$;*

(iii) *There exists a neighborhood $\Theta^0_{\mathscr{U}_0} \subset \Theta$ of $\theta^0_{\mathscr{U}_0}$ such that, for all $\theta^{(1)}$ and $\theta^{(2)} \in \Theta^0_{\mathscr{U}_0}$ and some $K_T = O_p(1)$,*

$$||H_{\mathscr{U}_0,T}(\theta^{(1)}, 0) - H_{\mathscr{U}_0,T}(\theta^{(2)}, 0)|| \leq K_T||\theta^{(1)} - \theta^{(2)}||.$$

Here, $a_T = O_p(1)$ means that $|a_T| \leq c$ with probability 1 when $T \to \infty$ and c is a constant. Proposition 4(i) shows that the fourth moment of the score function S_T is always finite. Proposition 4(ii) indicates that $\lambda_{\min}(-H^0_{\mathscr{U}_0,T})$ is almost surely positive and bounded away from 0. Hence, when the L_1 penalty is combined with the quasi-likelihood function, the concavity around θ^0 can be ensured, so that a local maximizer can be obtained. Proposition 4(iii) is trivial in linear models, but not in our case. The proof of Proposition 4 is given in Appendix A.

3.3. L_1 Penalty Function and Penalized Quasi-Likelihood

Before discussing the consistency of the sparse estimator, we introduce the following condition, by following the strong irrepresentable condition for LASSO-regularized linear regression models in Zhao and Yu (2006).

Condition 2 (Irrepresentable condition). *There exists a neighborhood $\Theta^0_{\mathscr{U}_0} \subset \Theta$ of $\theta^0_{\mathscr{U}_0}$, such that*

$$\sup_{\theta^{(1)},\theta^{(2)} \in \Theta^0_{\mathscr{U}_0}} ||[(\partial/\partial\theta^T_{\mathscr{U}_0})S_{\mathscr{U}_0^c,T}(\theta^{(1)}, 0)][H_{\mathscr{U}_0,T}(\theta^{(2)}, 0)]^{-1}||_\infty \leq c$$

for a constant c that takes its value in (0,1) almost surely.

Definition 1. *The half of the minimum signal d is defined as*

$$d(d = d_T) = \frac{1}{2}\min\{|\theta^0_j| : \theta^0_j \neq 0\} = \frac{1}{2}\min_{j \in \mathscr{U}_0}|\theta^0_j|. \tag{8}$$

Assume that $p_\lambda(x)$ is an L_1 penalty function, i.e., $p_\lambda(|x|) = \lambda|x|$. We consider the following *penalized quasi-likelihood* (PQL):

$$Q_T(\theta) = L_T(\theta) - P_T(\theta) \tag{9}$$

in which $P_T(\theta) = \sum_{j=1}^p p_\lambda(|\theta_j|) = \lambda\sum_{j=1}^p |\theta_j|$ is the penalty term and $\lambda(= \lambda_T) \geq 0$ is the regularization parameter determining the size of the model. If $\widehat{\theta}$ maxmizes the PQL, i.e.,

$$\widehat{\theta} = \arg\max_{\theta \in \Theta} Q_T(\theta).$$

we say that $\widehat{\theta}$ is a penalized quasi-maximum likelihood estimator (PQMLE).

Similar to Fan and Lv (2011) and Uematsu (2015), we add some conditions on the penalty function $p_\lambda(\cdot)$ and the half minimum signal.

Condition 3. *The penalty function p_λ satisfies the following properties:*

(i) $\lambda = \min\{O(T^{-\alpha}), o(q^{-\frac{1}{2}}T^{-\gamma}\log T)\}$ *for some* $\alpha \in (\delta_0 + \gamma, \frac{1}{2} - \frac{\delta_0}{4})$, $\gamma \in (0, \frac{1}{2}]$ *and large* T. *Here,* $a = O(f(T))$ *means* $|a/f(T)|$ *is bounded by a constant and* $b = o(g(T))$ *means* $|b/g(T)| \to$ *when* $T \to \infty$;

(ii) $d \geq T^{-\gamma}\log T$ *for some* $\gamma \in (0, \frac{1}{2}]$ *and large* T, *where* d *is the half-minimum signal we defined before.*

4. Properties of the PQML Estimator and Implementation

This section studies the sparsity and the consistency of the PQML estimator and discuss some implementation issues.

4.1. Sparsity of the PQML Estimator

First, we introduce three lemmas whose proofs are given in Appendix A. For convenience, we denote $\widehat{\mathscr{U}} := \text{supp}(\widehat{\theta})$, which is a set of indices corresponding to all nonzero components of $\widehat{\theta}$, where supp is the notation of support set and $\widehat{\theta}_{\widehat{\mathscr{U}}}$ is a subvector of $\widehat{\theta}$, formed by its restriction to $\widehat{\mathscr{U}}$. Then, $\widehat{\mathscr{U}}^c$ represents a set of indices corresponding to all 0 components in $\widehat{\theta}$. We also denote $\odot$ as the Hadmard product.

Lemma 1. *When the penalty function* p_λ *satisfies Condition 3,* $\widehat{\theta}$ *is a strict local maximizer of the* L_1*-PQL* $Q_T(\theta)$ *defined in (9) if*

$$S_{\widehat{\mathscr{U}},T}(\widehat{\theta}) - \lambda_T \mathbf{1} \odot sign(\widehat{\theta}_{\widehat{\mathscr{U}}}) = 0, \tag{10}$$

$$||S_{\widehat{\mathscr{U}}^c,T}(\widehat{\theta})||_\infty < \lambda_T, \tag{11}$$

$$\lambda_{min}[-H_{\widehat{\mathscr{U}},T}(\widehat{\theta})] > 0, \tag{12}$$

in which $\mathbf{1}$ *represents the vector with all elements equaling to 1 and* $sign(\cdot)$ *is as defined at the beginning of Section 3.*

To show the weak oracle property of the PQML estimator, we also need the following lemma.

Lemma 2. *Let* w_t *be a martingale difference sequence with* $E|w_t|^m \leq C_w$ *for all* t, *where* $m > 2$ *and* C_w *is a constant. Then, we have*

$$T^{-\frac{m}{2}} E\Big(\sum_{t=1}^{T} w_t\Big)^m < \infty.$$

Then, the weak oracle property of the PQML estimator can be established by the following theorem, whose proof is provided in Appendix A.

Theorem 1. *(L_1-PQML estimator) Under Conditions 2 and 3, for the* L_1 *penalty function* $P_T(\theta) = \lambda \sum_{i=1}^{p} |\theta|$, *in which* $p = O(T^\delta)$ *and* $q = O(T^{\delta_0})$, *if*

$$\delta \in [0, 4(\frac{1}{2} - \alpha)), \qquad 0 < \delta_0 < \min\{\frac{2}{3}(1 - 2\gamma), \gamma\},$$

with $\alpha \in (\delta_0 + \gamma, \frac{1}{2} - \frac{\delta_0}{4})$, $\gamma \in (0, \frac{1}{2}]$, *and* $\delta > \delta_0$, *then there exists a local maximizer* $\widehat{\theta} = ((\widehat{\theta}_{\mathscr{U}_0})', (\widehat{\theta}_{\mathscr{U}_0^c})')'$ *for* $Q_T(\theta)$, *such that the following properties are satisfied:*

(i) (Sparsity) $\widehat{\theta}_{\mathscr{U}_0^c} = 0$ *with probability approaching one;*

(ii) (Rate of convergence) $||\widehat{\theta}_{\mathscr{U}_0} - \theta^0_{\mathscr{U}_0}||_\infty = O_p(T^{-\gamma}\log T)$.

$p = O(T^\delta)$ is equivalent to $\frac{p}{T^\delta} \leq c$ when $T \to \infty$. The growth rate of p is controlled by T^δ and q is slower than with T^{δ_0}. For example, to make this growth rate of q much slower than p, we can find a set of values for $\delta = \frac{3}{2}$, $\delta_0 = \frac{1}{20}$, $\gamma = \frac{1}{30}$, and $\alpha = \frac{1}{5}$ that satisfy

the conditions above. Since, in our case, $p \sim O(n^2)$, we have $n = O(T^{\frac{3}{4}})$ and, hence, it is possible for n to exceed the sample size T. Although the difference between the rates of p and n is not as large as that in (Fan and Lv 2011), in which $\log p = O(T^{1-2\alpha})$ and $q = o(T)$, it is enough to be applied in most cases in practice.

4.2. Implementation and Selection of λ

To compute the whole regularization path of L_1-PQML estimators, we note that several algorithms have been proposed to solve penalized optimization problems. For example, Efron et al. (2004) proposed the least-angle regression (LARS) algorithm to compute an efficient solution to the optimization problem for LASSO. Later on, pathwise coordinate descent methods were proposed to solve the LASSO-type problem efficiently; see Friedman et al. (2007) and Wu and Lange (2008). For the PQML estimator, we used an algorithm inspired by the BLasso algorithm (Zhao and Yu 2006, 2007) with some necessary modifications since the BLasso algorithm does not need to explicitly calculate the first derivatives and second derivatives of the likelihood function, which are complicated in our case. We note that the original BLasso algorithm uses 0 as initial values for all parameters, but the diagonal elements of A and B are positive by definition, so we make the following modification. We set 0 as the initial values for all off-diagonal elements in A, B, and C, and set the estimated values of fitting the component series into a univariate GARCH model as the initial values of the diagonal elements in parameter matrices A, B, and C.

Another issue in the implementation is to select the tuning parameter λ, which leads to the problem of model selection. The tuning parameter λ can be chosen by several criteria. For example, it is usually easy to consider the Akaike information criterion (AIC), the small-sample corrected AIC (AICC), and Bayesian information criterion (BIC) criteria to select the tuning parameter. In addition, Wang et al. (2009) proposed a modified BIC criterion and Fan and Tang (2013) extended it for the case $p > T$. Sun et al. (2013) proposed using Cohen's kappa coefficient, which measures the agreement between two sets. Another method for model selection is to use cross-validation (CV). Zhang and Yang (2015) used CV to choose the best model among model selection procedures such as AIC and BIC. In our study, we apply the AIC and BIC criteria on the testing data and select the best tuning parameters. Note that, since our data are ordered, k-fold CV is not applicable here and the data are split in time order.

5. Simulation

In this section, we study the performance of the regularized BEKK models on some simulated examples. Consider Model (3) with $n = 4$ and $a = b = 1$. Note that we then have $p = 42$ parameters, as matrix C is lower triangular. We assume that the parameter matrices satisfy the stationary condition, Condition 2, and, for identification purposes, we assume that the diagonal elements in C are positive, $a_{11} > 0$, and $b_{11} > 0$. We consider two cases for matrices A, B, and C, which are summarized in Table 1. In both cases, the indices of nonzero elements in coefficient matrices A, B, and C are randomly generated. To ensure that the matrices satisfy Condition 1, values of the diagonal elements in A and B are randomly generated from a uniform distribution on $\mathrm{U}(-0.45, 0.45)$, and the off-diagonal nonzero elements in A and B are generated from $\mathrm{U}(-0.5, 0.5)$. All the nonzero elements in C are generated from $\mathrm{U}(-0.1, 0.1)$.

For each case, we simulate the data r_t $(1 \le t \le T)$ with $T = 600$, and then use the proposed regularized procedure to make inference on the model. Since the diagonal elements in A, B, and C cannot be zero, we do not shrink the diagonal elements in A, B, and C. Additionally, we set the estimates of parameters in univariate GARCH models for each component series as the initial values of diagonal elements in A, B, and C.

To demonstrate the performance of our estimates, we consider three measurements. The first is the success rate in estimating zero and nonzero elements in θ or parameter matrices:

$$\tau_0 = \frac{\sum_{i=1}^{p} I(\theta_i^0 = 0 \wedge \widehat{\theta}_i = 0)}{\sum_{i=1}^{p} I(\theta_i^0 = 0)}, \quad \tau_{0^C} = \frac{\sum_{i=1}^{p} I(\theta_i^0 \neq 0 \wedge \widehat{\theta}_i \neq 0)}{\sum_{i=1}^{p} I(\theta_i^0 \neq 0)}.$$

The second measure is the root of mean squared errors, which is defined as $\nu = ||\theta^0 - \widehat{\theta}_\lambda||_2$. The third measure is the Kullback–Leibler information, which is given by

$$\kappa = \frac{1}{2T} \sum_{t=1}^{T} \left(|\Sigma_t \widehat{\Sigma}_t^{-1}| - \log |\Sigma_t \widehat{\Sigma}_t^{-1}| \right),$$

where $\widehat{\Sigma}_t = \widehat{C}'\widehat{C} + \widehat{A} r_{t-1} r_{t-1}' \widehat{A}' + \widehat{B} \Sigma_{t-1} \widehat{B}'$. We run $N = 500$ simulations for each case, and present the performance measures and their standard errors (in parentheses) for different λs in Table 2.

Table 1. Parameter matrices in simulations.

	Case 1	**Case 2**
A	$\begin{pmatrix} 0.1268 & 0 & 0.0358 & -0.0618 \\ 0 & 0.1737 & 0 & 0 \\ 0 & 0 & 0.2621 & 0 \\ 0 & 0 & 0 & 0.4096 \end{pmatrix}$	$\begin{pmatrix} 0.4040 & -0.0200 & 0 & 0 \\ 0 & 0.4434 & -0.0752 & 0 \\ 0 & 0 & 0.0406 & 0.0684 \\ 0 & 0 & 0 & 0.2226 \end{pmatrix}$
B	$\begin{pmatrix} 0.4257 & 0 & 0 & 0 \\ 0 & 0.3008 & 0 & 0 \\ 0.0912 & 0 & 0.2868 & 0 \\ 0 & 0 & 0 & 0.0372 \end{pmatrix}$	$\begin{pmatrix} 0.2453 & 0 & 0 & 0 \\ 0 & 0.2401 & 0 & 0 \\ 0.2398 & 0 & 0.2157 & 0 \\ 0 & 0 & 0 & 0.3996 \end{pmatrix}$
C	$\begin{pmatrix} 0.0324 & 0 & 0 & 0 \\ 0 & 0.0681 & 0 & 0 \\ 0 & 0.0349 & 0.0469 & 0 \\ 0.0728 & 0 & 0 & 0.0739 \end{pmatrix}$	$\begin{pmatrix} 0.0804 & 0 & 0 & 0 \\ 0 & 0.0473 & 0 & 0 \\ 0 & 0 & 0.0521 & 0 \\ 0.0200 & 0 & 0 & 0.0628 \end{pmatrix}$

Table 2. Performance measures in two cases.

Case	λ	**6**	**5**	**4**	**3**	**2**	**1**	**0.64**	**0.32**	**0.16**	**0.08**
1	τ_0	0.988 (0.031)	0.988 (0.031)	0.984 (0.035)	0.981 (0.037)	0.974 (0.040)	0.954 (0.046)	0.934 (0.051)	0.890 (0.058)	0.841 (0.066)	0.756 (0.083)
	τ_{0^C}	0.755 (0.028)	0.755 (0.033)	0.767 (0.035)	0.786 (0.036)	0.814 (0.024)	0.821 (0.014)	0.822 (0.013)	0.834 (0.026)	0.864 (0.034)	0.897 (0.038)
	ν	0.151 (0.029)	0.151 (0.029)	0.150 (0.029)	0.147 (0.029)	0.142 (0.030)	0.133 (0.032)	0.132 (0.032)	0.131 (0.031)	0.128 (0.030)	0.125 (0.029)
	100κ	3.104 (6.463)	3.533 (7.124)	2.370 (5.471)	1.611 (4.158)	0.735 (1.051)	0.461 (0.460)	0.359 (0.356)	0.325 (0.294)	0.289 (0.242)	0.282 (0.257)
2	τ_0	0.985 (0.030)	0.985 (0.030)	0.983 (0.034)	0.977 (0.034)	0.960 (0.043)	0.918 (0.048)	0.889 (0.052)	0.841 (0.052)	0.805 (0.054)	0.767 (0.065)
	τ_{0^C}	0.740 (0.029)	0.742 (0.029)	0.745 (0.028)	0.752 (0.025)	0.759 (0.019)	0.766 (0.011)	0.765 (0.014)	0.767 (0.013)	0.768 (0.014)	0.782 (0.030)
	ν	0.151 (0.022)	0.151 (0.022)	0.151 (0.022)	0.149 (0.021)	0.145 (0.021)	0.136 (0.023)	0.136 (0.023)	0.135 (0.022)	0.132 (0.021)	0.126 (0.022)
	100κ	2.004 (8.525)	2.040 (9.024)	1.818 (7.710)	0.330 (1.381)	0.323 (1.936)	0.369 (1.829)	0.402 (2.388)	0.278 (0.163)	0.275 (0.163)	0.209 (0.180)

To select the tuning parameter λ, we use the first 500 samples as the training data and the last 100 samples as the test data. The training data are used to estimate model

parameters θ_λ for a given λ, and the test data are used to choose the best λ, i.e., the one that gives the minimum AICs and BICs. That is,

$$\widehat{\lambda}_{\text{BIC}} = \arg\min_{\lambda} \text{BIC}_\lambda, \qquad \widehat{\lambda}_{\text{AIC}} = \arg\min_{\lambda} \text{AIC}_\lambda,$$

in which the AIC_λ and BIC_λ are defined as

$$\text{BIC}_\lambda = -2L_{T_{test}}(\widehat{\theta}_\lambda) + \frac{k\log(T_{test})}{T_{test}} \qquad \text{AIC}_\lambda = -2L_{T_{test}}(\widehat{\theta}_\lambda) + \frac{2k}{T_{test}},$$

where, in this case, $k = \sum_{i=1}^{p} I(\widehat{\theta}_i^\lambda \neq 0)$, $T_{test} = 100$, and

$$L_{T_{test}}(\theta) = \frac{1}{2T_{test}} \sum_{t=501}^{600} -(\log[\det(\Sigma_t)] + r_t'\Sigma_t^{-1}r_t).$$

Figure 1 shows the histograms of selected λs via BIC and AIC with CV for Cases 1 and 2. In general, we can see from Figure 1 that λ is favored by BIC and AIC when its value is between 0.64 and 2. However, slight differences between these two cases can be found. For Case 1, λs around 1 are most favored by both BIC and AIC, while, for Case 2, λs around 1 and 2 are most favored by BIC and AIC, respectively.

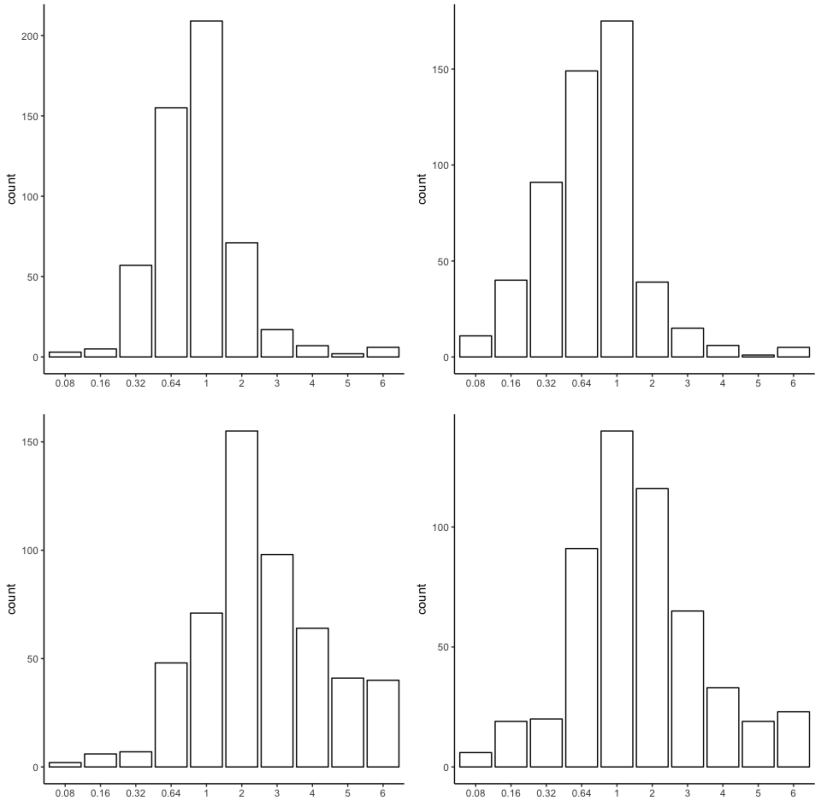

Figure 1. Histograms of selected λ in Cases 1 (**top**) and 2 (**bottom**) via BIC (**left**) and AIC (**right**).

6. Real Data Applications

In this section, we use the regularized BEKK representation to study the volatility spillover effect and find optimal Markowitz's mean–variance portfolios. The data we studied consist of daily log-returns of 18 stocks during the period 4 January 2016–31 January 2018, which are listed in Table 3 (NASDAQ Stock Symbols n.d.). Figure 2 shows the time series of these 18 stocks and Table 4 summarizes the sample mean, the sample standard deviation, the sample skewness, the sample kurtosis, and the correlations of these 18 series.

All the correlations are positive for every two stocks in the selected period, and, except for IPG, all the stocks have a positive mean. The sample kurtosis for some stocks is way larger than 3, which indicates that we cannot simply assume that those returns are following normal distributions individually. Hence, it is natural to employ a suitable time series model to examine the data.

Table 3. Full names of 18 tickers.

Ticker	Company	Ticker	Company
GOOG	Alphabet Inc., Mountain View, CA, USA	GWW	W.W. Grainger, Inc., Lake Forest, FL, USA
IBM	International Business Machines Corporation, Armonk, NY, USA	JPM	JPMorgan Chase & Co., New York, NY, USA
MSFT	Microsoft Corporation, Redmond, WA, USA	NKE	Nike Inc., Beaverton, OR, USA
ORCL	Oracle Corporation, Austin, TX, USA	TIF	Tiffany & Co., New York, NY, USA
IPG	The Interpublic Group of Companies, New York, NY, USA	MAS	Masco Corporation, Livonia, MI, USA
MCD	Mcdonald's Corp., Chicago, IL, USA	NFLX	Netflix, Inc., Los Gatos, CA, USA
RL	Ralph Lauren Corporation, New York, NY, USA	TXT	Textron Inc., Providence, RI, USA
LNC	Lincoln National Corporation, Radnor, PA, USA	MRO	Marathon Oil Corporation, Houston, TX, USA
TGT	Target Corporation, Minneapolis, MN, USA	WMT	Walmart Inc., Bentonville, AR, USA

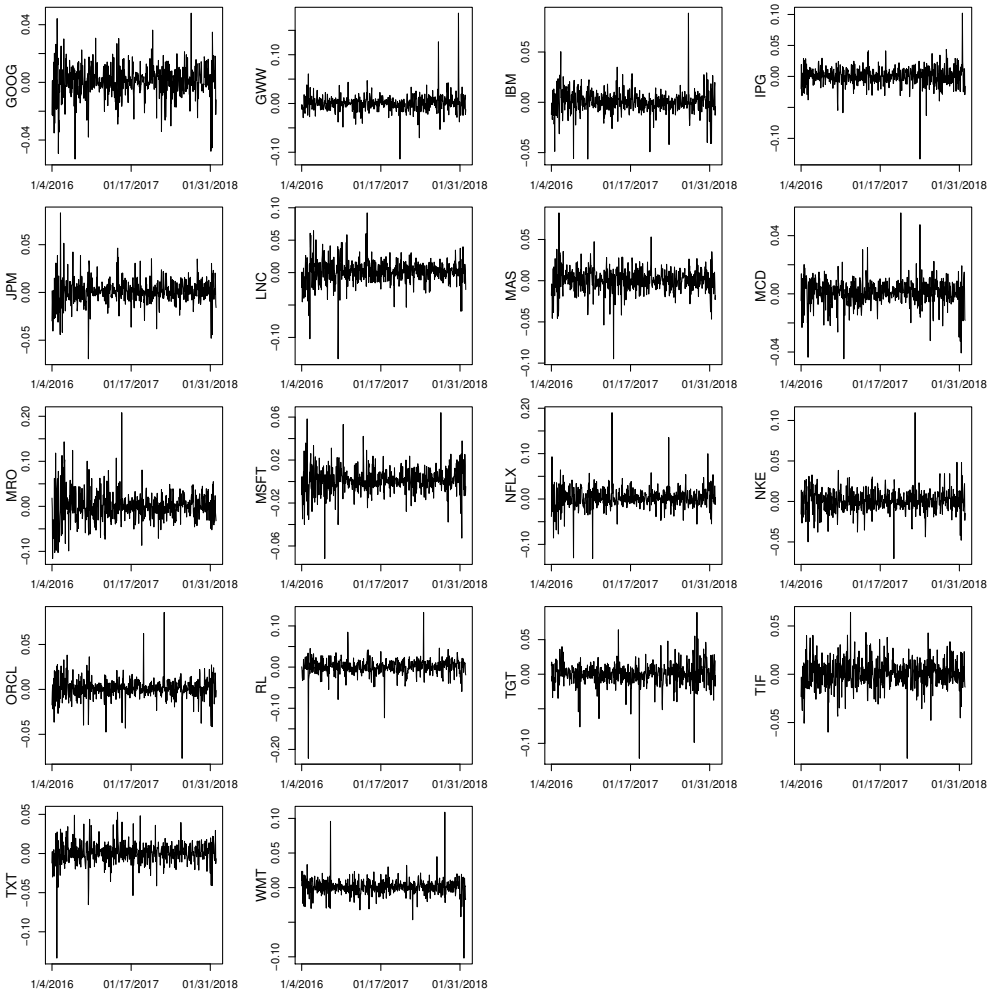

Figure 2. Daily returns of 18 stocks from 4 January 2016 to 31 January 2018.

Table 4. Correlation and statistical features of 18 stocks for 2016–2017.

	GOOG	**GWW**	**IBM**	**JPM**	**MSFT**	**NKE**	**ORCL**	**TIF**	**IPG**	**MAS**	**MCD**	**NFLX**	**RL**	**TXT**	**LNC**	**MRO**	**TGT**	**WMT**
Mean	7.037×10^4	5.164×10^4	4.297×10^4	1.150×10^3	1.033×10^3	1.420×10^4	6.439×10^4	8.143×10^4	-7.648×10^5	1.029×10^3	8.891×10^4	1.304×10^3	1.610×10^4	7.103×10^4	1.148×10^3	1.217×10^3	5.523×10^5	1.116×10^3
Std. Dev.	0.013	0.016	0.011	0.013	0.012	0.014	0.012	0.015	0.015	0.014	0.009	0.023	0.021	0.015	0.019	0.034	0.017	0.012
Skewness	−0.440	0.124	0.291	0.348	0.040	0.509	0.097	−0.391	−1.779	−0.456	0.005	0.635	−2.332	−1.438	−0.779	0.640	−1.108	2.291
Kurtosis	6.120	16.399	14.446	8.907	8.926	11.683	15.022	6.274	16.842	9.409	9.640	17.196	33.455	17.970	10.830	7.383	13.600	24.910
GOOG		0.04	0.24	0.26	0.67	0.28	0.36	0.22	0.28	0.33	0.29	0.44	0.15	0.23	0.27	0.11	0.06	0.13
GWW			0.38	0.34	0.13	0.17	0.16	0.24	0.29	0.27	0.03	0.06	0.09	0.34	0.28	0.21	0.16	0.09
IBM				0.38	0.30	0.15	0.42	0.27	0.30	0.31	0.12	0.13	0.15	0.39	0.37	0.19	0.15	0.14
JPM					0.37	0.23	0.38	0.41	0.31	0.44	0.23	0.19	0.28	0.52	0.78	0.36	0.16	0.09
MSFT						0.26	0.45	0.29	0.29	0.38	0.35	0.36	0.19	0.32	0.33	0.20	0.08	0.12
NKE							0.20	0.23	0.30	0.30	0.18	0.18	0.33	0.23	0.30	0.12	0.27	0.16
ORCL								0.34	0.26	0.36	0.26	0.26	0.15	0.30	0.35	0.19	0.09	0.12
TIF									0.23	0.33	0.14	0.18	0.25	0.33	0.42	0.26	0.27	0.11
IPG										0.38	0.09	0.12	0.21	0.26	0.32	0.15	0.21	0.10
MAS											0.28	0.23	0.26	0.36	0.46	0.20	0.24	0.12
MCD												0.14	0.13	0.12	0.16	0.07	0.09	0.16
NFLX													0.10	0.22	0.19	0.07	0.02	0.09
RL														0.22	0.35	0.16	0.31	0.09
TXT															0.53	0.26	0.17	0.09
LNC																0.39	0.19	0.05
MRO																	0.08	0.04
TGT																		0.36

6.1. Volatility Spillovers

To use the MGARCH–BEKK representation to analyze the market, consisting of 18 stocks, we should realize that certain types of regularization or shrinkage are necessary, due to the complexity of the volatility dynamics. In particular, we use the proposed L_1-regularized BEKK(1,1) model and procedure to study the volatility spillover among the 18 stocks. We first compute the PQML estimates of the model for different λs. Figure 3 shows the estimated structures of estimated coefficient matrices $\widehat{A}_\lambda$ and $\widehat{B}_\lambda$ for $\lambda = 4, 2, 1, 0.5, 0.3$, in which the nonzero values of $\widehat{A}_\lambda$ and $\widehat{B}_\lambda$ are represented as the directional lines among stocks. Since matrices A and B in the model are not symmetric before the quadratic forms, we use the directional lines to distinguish the nonzero elements between upper-diagonal and lower-diagonal elements. Specifically, if $a_{ij} \neq 0$, the directional line progresses from i to j. As the PQML estimates $\widehat{A}_\lambda$ and $\widehat{B}_\lambda$ tell us the significant interdependence and contagion effects of the 18 stocks, the network structures in Figure 3 provide a clear representation on volatility spillover. Furthermore, we notice that, for some moderate values of λ, for example, $\lambda = 0.5$, $\widehat{A}_\lambda$ is very sparse, whereas $\widehat{B}_\lambda$ demonstrates more interdependence among stocks. When larger values of λ are used in the regularization procedure, the PQML estimates $\widehat{A}_\lambda$ are quickly shrunk into diagonal matrices, and $\widehat{B}_\lambda$ also become more sparse than for the case $\lambda = 0.5$.

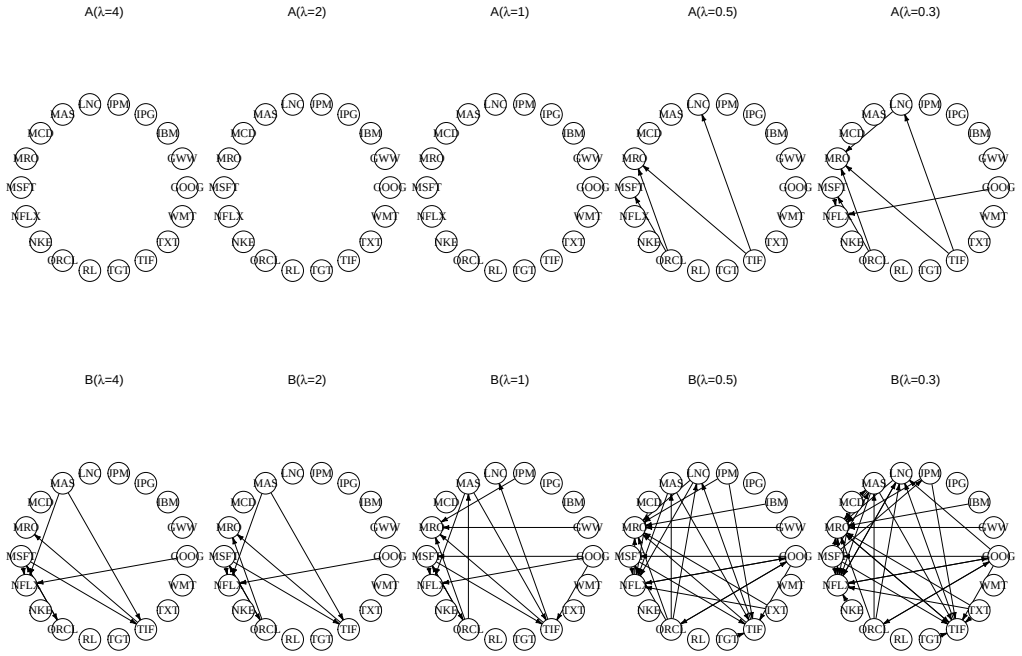

Figure 3. The network structure of estimated matrices A (**top**) and B (**bottom**) under different λs.

Using the PQML estimates of $\widehat{A}_\lambda$, $\widehat{B}_\lambda$, and $\widehat{C}_\lambda$ and the BEKK(1, 1) representation, we compute the estimated volatilities and the dynamic correlations among 18 stocks. Figure 4 shows the volatilities estimated by the regularized BEKK(1,1) model with $\lambda = 2, 0.5$ and univariate GARCH models. Note that most volatility series estimated by the three models are similar, except for stocks `NFLX`, `ORCL`, and `TIF`. We also show the estimated dynamic correlations among 18 stocks in a regularized BEKK(1,1) model with $\lambda = 1$ in Figure 5. We note that most correlations among the 18 stocks are positive during the sample period.

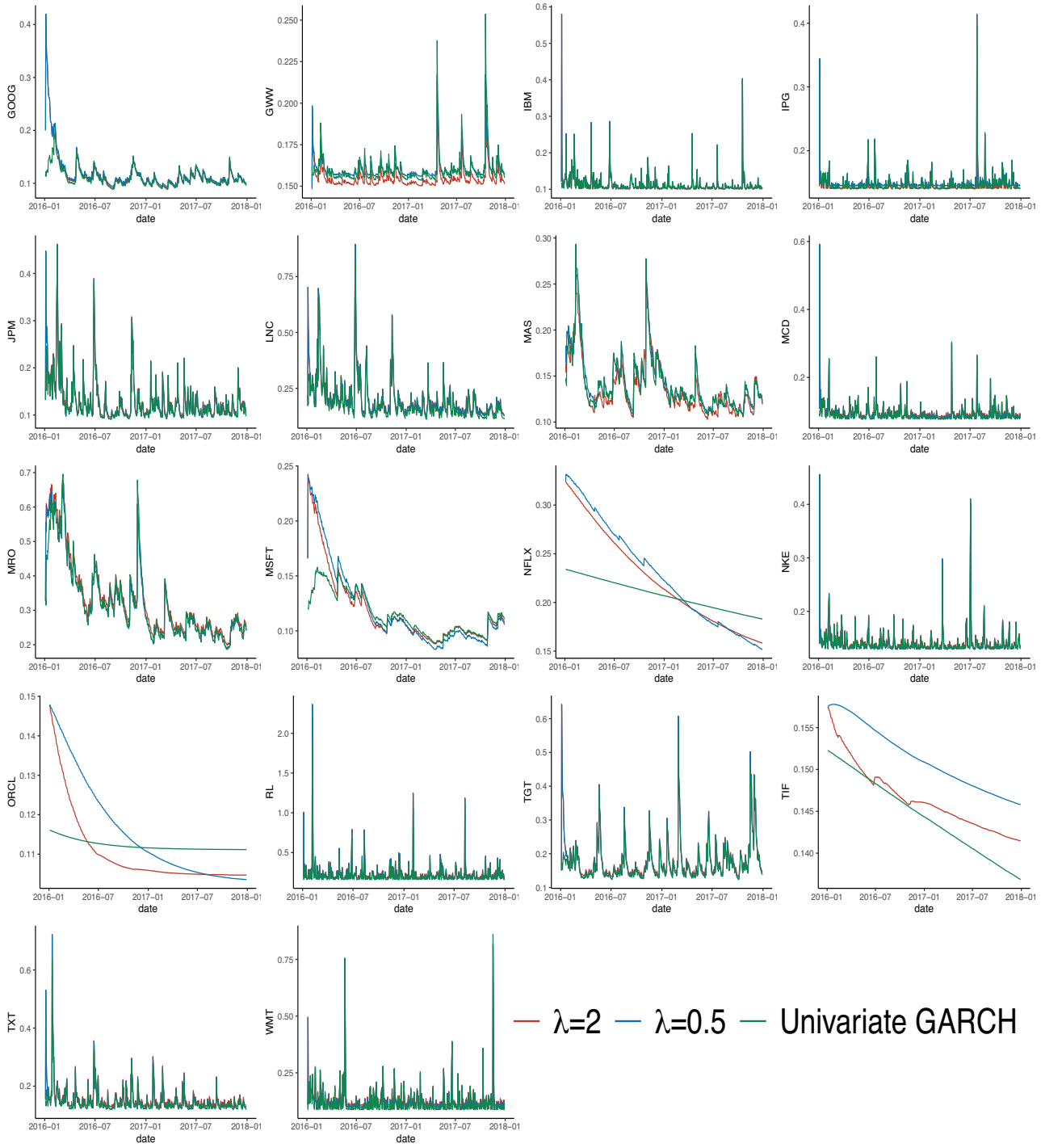

Figure 4. Estimated volatilities by regularized BEKK(1,1) with $\lambda = 2$ (red lines), $\lambda = 0.5$ (blue lines), and univariate GARCH models (green lines).

To show the overall volatility spillover, we extend the idea of the spillover index in Diebold and Yilmaz (2009). Specifically, note that $E[\epsilon_{t+1}\epsilon_{t+1}'] = \Sigma_{t+1} = \Sigma_{t+1}^{\frac{1}{2}}(\Sigma_{t+1}^{\frac{1}{2}})'$, where $\Sigma_{t+1}^{\frac{1}{2}}$ is the unique lower-triangular Cholesky factor of Σ_{t+1}. We denote elements of $\Sigma_t^{\frac{1}{2}}$ by $\sigma_{\frac{1}{2},i,j,t}$; then, the Spillover Index S_{t+1} is defined as

$$S_{t+1} = \frac{\sum_{i,j=1,i\neq j}^{n} \widehat{\sigma}^2_{\frac{1}{2},i,j,t+1}}{trace(\widehat{\Sigma}_{t+1})} \times 100\%,$$

where n is the number of stocks, which is equal to 18. We plot the daily spillover indices of 18 stocks for $\lambda = 2$ and 0.5. The spillover indices during the sample period vary between 5% and 80%, and smaller λs seem to generate more correlations among stocks. In particular, three big spikes can be found on 4 February 2016, 24 June 2016, and 9 November 2016. In addition to finding the PQML estimates for different λs, we also find the whole $Ł_1$ regularization path. Note that the number of parameters in the BEKK(1,1) model for 18 stocks is $p = 819$, and we only show the regularized path for $819 - 18 \times 3 = 765$ off-diagonal elements in $\widehat{A}_\lambda$, $\widehat{B}_\lambda$, and $\widehat{C}_\lambda$. And both plots are shown in Figure 6.

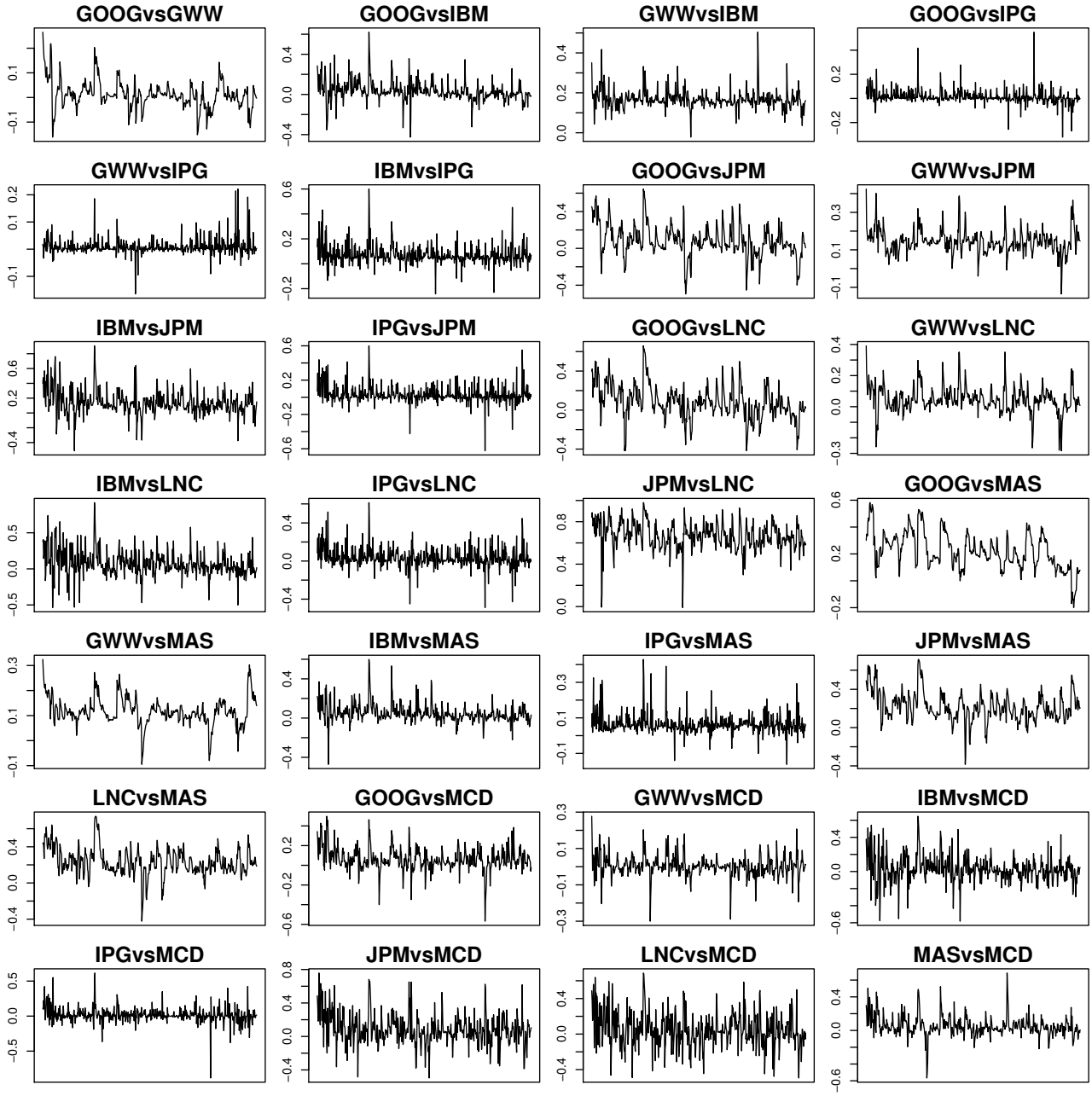

Figure 5. Daily estimated conditional correlations when $\lambda = 1$.

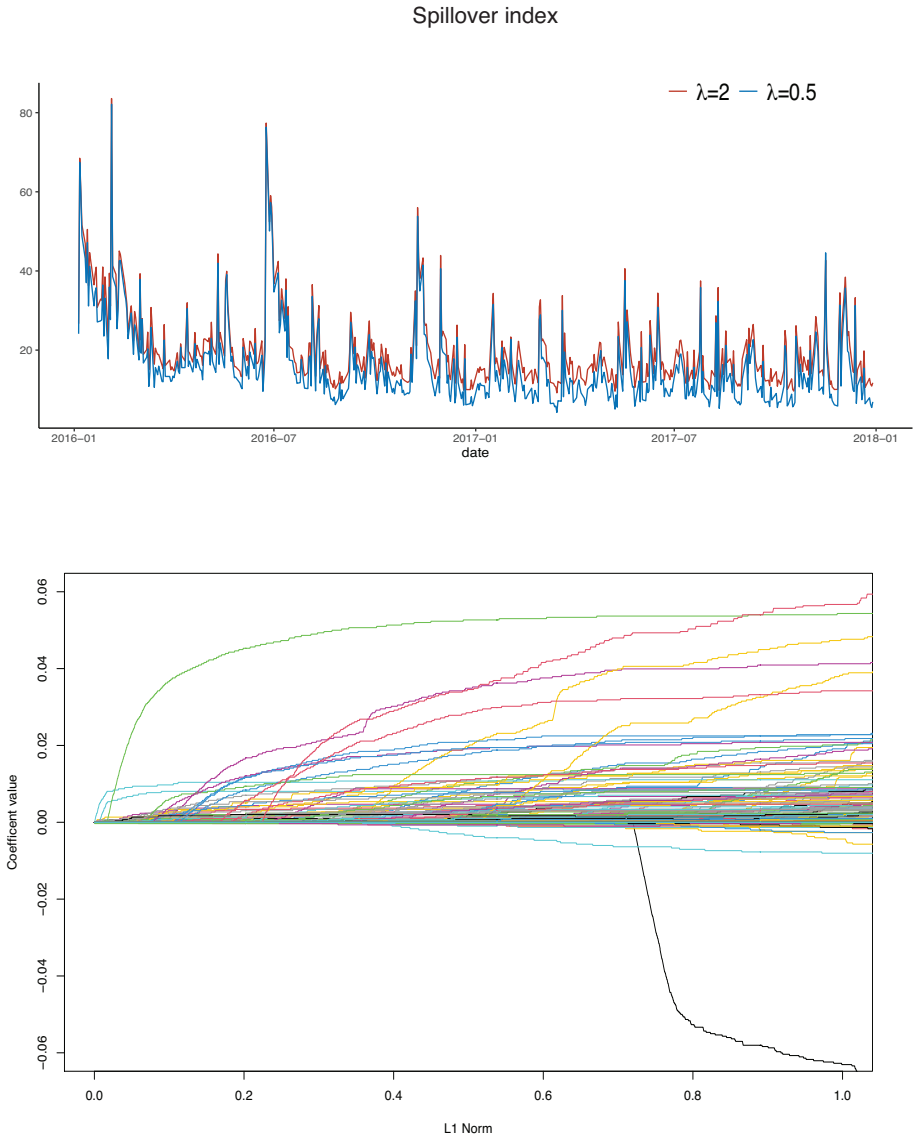

Figure 6. Daily spillover index (**top**) and regularization paths of estimated off-diagonal parameters in BEKK regularization Model represented by different colors (**bottom**).

6.2. Portfolio Optimization

We further apply the regularized BEKK model to Markowitz mean–variance portfolio optimization (Markowitz 1952). Using portfolio variance as a measure of the risk, Markowitz portfolio optimization theory provides an optimal pay-off between the profit and the risk. Since the means and covariance matrix of assets are assumed to be known in the theory, they need to be estimated before being plugged into the framework. For high-dimensional portfolios, regularized methods are commonly used to achieve better performance. For instance, Brodie et al. (2009) and Fastrich et al. (2015) used an L_1 penalty function for sparse portfolios, and Di Lorenzo et al. (2012) used a concave optimization-based approach to estimate the optimal portfolio. In our case, we use the regularized BEKK model to predict the covariance matrices in the next period, and then apply Markowitz portfolio theory to find the optimal portfolios.

In particular, we assume that the portfolio consists of $n = 18$ risky assets and denote μ_t and Σ_t as the mean and covariance matrix, respectively, of the n risk assets at time t. Let $\mathbf{1} = (1, \ldots, 1)'$ be an n-dimensional vector of ones. Markowtiz mean–variance portfolio theory minimizes the variance of the portfolio $\min_{w_t} w_t' \Sigma_t w_t$, subject to the constraint

$w_t'\mathbf{1} = 1$ and $w_t'\mu_t = \mu_*$, where μ_* is the target return. When short selling is allowed, the efficient portfolio can be explicitly expressed as

$$w_{\text{effi},t} = \frac{\widetilde{b}}{\widetilde{d}}\Sigma_t^{-1}\mathbf{1} - \frac{\widetilde{a}}{\widetilde{d}}\Sigma_t^{-1}\mu_t + \mu_*\Big(\frac{\widetilde{c}}{\widetilde{d}}\Sigma_t^{-1}\mu_t - \frac{\widetilde{a}}{\widetilde{d}}\Sigma_t^{-1}\mathbf{1}\Big),$$

where $\widetilde{a} = \mu_t\Sigma_t^{-1}\mathbf{1}$, $\widetilde{b} = \mu_t'\Sigma_t\mu_t$, $\widetilde{c} = \mathbf{1}'\Sigma_t\mathbf{1}$, and $\widetilde{d} = \widetilde{b}\widetilde{c} - \widetilde{a}^2$. When the target return μ_* is chosen to minimize the variance of the efficient portfolio, we obtain the global minimum variance (GMV) portfolio:

$$w_{\text{minvar},t} = \Sigma_t^{-1}\mathbf{1}/(\mathbf{1}'\Sigma_t\mathbf{1}).$$

For comparison purposes, we also use another three multivariate volatility models to predict the covariance matrices of $n = 18$ stocks. The first is very simple, and it assumes a constant covariance matrix for n stocks. The second is a factor-GARCH model (Alexander 2000; Engle 1990; van der Weide 2002; Vrontos et al. 2003), which assumes the following for asset return vector r_t, factors f_t, and volatilties of k independent factors:

$$r_t = Wf_t, \quad \text{Cov}(f_t) = \Sigma_t = \text{diag}\{\sigma_{1t}^2, \sigma_{2t}^2, \ldots, \sigma_{kt}^2\},$$

$$\sigma_{it}^2 = 1 + \beta_i x_{i,t-1}^2 + \gamma_i\sigma_{i,t-1}^2,$$

where W is a $k \times k$ lower-triangular matrix with diagonal elements equal to 1 and $x_t = (x_{1t}, \ldots, x_{kt})'$ is a vector of k independent factors. The third covariance model is a dynamic conditional correlation GARCH (DCC–GARCH) model (Engle 2002), which has the form

$$r_t = \Sigma_t^{\frac{1}{2}}\epsilon_t, \quad \epsilon_t \sim N(0, I_n), \quad \Sigma_t = D_tR_tD_t,$$

$$Q_t = (1 - \alpha - \beta)C + \alpha s_{t-1}s_{t-1}' + \beta Q_{t-1}, \qquad R_t = \text{diag}(Q_t)^{-\frac{1}{2}}Q_t\text{diag}(Q_t)^{-\frac{1}{2}},$$

where $D_t = \text{diag}(d_{1t}, \ldots, d_{nt})$, $s_{i,t} = r_{i,t}/d_{i,t}$, $s_t = (s_{1,t}, \ldots, s_{T,t})'$, and R_t is the conditional correlation matrix at time t, that is, $R_t = \text{Corr}(r_t|\mathcal{F}_{t-1})$. And C is the unconditional correlation matrix, i.e., $C = E(R_t)$. The matrix Q_t can be interpreted as a conditional covariance matrix of devolatilized residuals. For the dynamics of the univariate volatilities, $d_{i,t}$s are assumed to follow a GARCH(1,1) process:

$$d_{i,t} = \omega_i + a_i r_{i,t-1}^2 + b_i d_{i,t-1}^2,$$

where (w_i, a_i, b_i) are GARCH(1,1) parameters.

Let $t =$ 2 January 2018, ..., 31 January 2018; we first fit 4 covariance models to the returns of 18 stocks from 4 January 2016 to t, and then compute the 1-day-ahead prediction of covariance matrices. Using the predicted covariance matrices, we compute the efficient portfolios $w_{\text{minvar},t+1}$ and $w_{\text{effi},t+1}$ for $\mu_* = 0.15\%, 0.10\%$, and 0.05%. Table 5 shows the means, standard deviations (SD), and the *information ratios* (IR, i.e., ratio of means and standard deviations) for realized portfolio returns in the month of January 2018. As argued by Engle and Colacito (2006) and Engle et al. (2019), these statistics are good measurements of the out-of-sample performance of Markowitz portfolios. As DeMiguel et al. (2007) claimed that it is difficult to outperform equally weighted portfolios in terms of the out-of-sample mean for Markowitz portfolios, we also include the performance of equally weighted portfolios as a benchmark in Table 5. We note that all the means generated from four covariance models are smaller than that from equally weighted portfolios (0.430%), and the standard deviations of covariance models, except the factor GARCH, are smaller than that of equally weighted portfolios. Notably, the regularized BEKK model consistently maintains the second-best mean performances at 0.39%, 0.352%, 0.382%, and 0.416% for GMV, and μ_* values of 0.15%, 0.10%, and 0.05%. However, the information ratio of the regularized BEKK model surpasses that of all other portfolios. It achieves the highest values across all scenarios—0.601, 0.540, 0.654, and 0.657—for GMV and $\mu_* = 0.15\%, 0.10\%$, and 0.05%. These results show the robustness and efficiency of the regularized BEKK model in

portfolio optimization, consistently delivering competitive mean performance and superior risk-adjusted returns compared to other covariance models.

Table 5. Performance of portfolios using different covariance models.

Model	Mean (%)	SD. (%)	IR	Mean (%)	SD. (%)	IR
Equally weighted	0.430	0.761	0.565			
		GMV			$\mu_* = 0.15\%$	
Regularized BEKK	0.390	0.650	0.601	0.352	0.652	0.540
Factor GARCH	0.326	0.885	0.368	0.223	1.200	0.186
DCC–GARCH	0.244	0.665	0.367	0.302	0.677	0.446
Constant covariance	0.261	0.658	0.397	0.165	0.777	0.212
		$\mu_* = 0.10\%$			$\mu_* = 0.05\%$	
Regularized BEKK	0.382	0.585	0.654	0.416	0.633	0.657
Factor GARCH	0.210	1.169	0.180	0.221	1.321	0.167
DCC–GARCH	0.286	0.631	0.452	0.316	0.660	0.479
Constant covariance	0.219	0.669	0.327	0.273	0.668	0.409

7. Discussion and Conclusive Remarks

Modeling the dynamics of high-dimensional covariance matrices is an interesting and challenging problem in both financial econometrics and high-dimensional time series analysis. To address this issue, this paper proposes an inference procedure with L_1 regularization for the sparse representation of high-dimensional BEKK and to obtain a class of penalized quasi-maximum likelihood estimators. The proposed regularization allows us to find significant parameters in the BEKK representation and shrink the non-essential ones to zero, hence providing a sparse estimate of the BEKK representations. We show that the sparse BEKK representation has suitable theoretical properties and is promising for applications in portfolio optimization and volatility spillover.

The proposed sparse BEKK representation also contributes to the application of machine learning methods in time series modeling. As most discussion on applying regularization methods to time series modeling focuses on regularizing high-dimensional vector autoregressive models and their variants (Nicholson et al. 2017; Sánchez García and Cruz Rambaud 2022), it seems that the sparse representation of dynamics of high-dimensional variance–covariance matrices has been ignored in the literature. While obtaining a sparse representation of the dynamics within high-dimensional variance–covariance matrices is crucial to enhance interpretability in time series modeling, our study bridges this gap by considering a basic L_1 regularization method. One obvious extension from our current study is to replace the L_1 penalty with other types of penalty for high-dimensional MGARCH models, for instance, the SCAD penalty (Fan and Li 2001), the adaptive LASSO (Zou 2006), and the group LASSO (Yuan and Lin 2006). With different types of penalty functions, one can regularize the assets in the model with different requirements, hence causing the estimates to have different kinds of asymptotic properties.

As the proposed sparse BEKK representation simplifies the dynamics of covariance matrices of high-dimensional time series, it has advantages over existing MGARCH models in some financial applications. In particular, the sparse BEKK representation can capture significant volatility spillover effects in high-dimensional financial time series, which usually cannot be analyzed using other MGARCH models. Since significant volatility spillover is captured, the proposed method also improves the performance of portfolio optimization based on the dynamics of high-dimensional covariance matrices. The proposed procedure can certainly be extended to incorporate more empirical aspects of financial time series. Taking the leverage effect as an example, one may modify the regularization procedure to

obtain sparse representation of high-dimensional multivariate exponential or threshold GARCH models.

Although the proposed framework shows advantages in modeling dynamics of high-dimensional covariance matrices, the computational challenge is not completely resolved. The main reason is that the proposed inference procedure involves a step of computing derivatives via the Kronecker product of parameter matrices. Since the Kronecker product turns two $n \times n$ matrices into an $n^2 \times n^2$ matrix, the requirement for computational memory resources increases significantly. Hence, the proposed procedure is suitable for problems in which the number of component time series ranges from several to 100. If the number of assets progresses beyond 200, the computational cost is still a major concern. One possible remedy for this is training a neural network to approximate the regularized likelihood of the high-dimensional model. In such a way, the proposed regularization using the high-dimensional MGARCH model can be extended to characterize the dynamics of covariance matrices of larger size.

Author Contributions: Conceptualization, H.X.; methodology, H.X., H.Z. and S.Y.; software, S.Y.; validation, S.Y., H.Z. and H.X.; formal analysis, S.Y.; investigation, S.Y., H.X. and H.Z.; resources, S.Y.; data curation, S.Y.; writing—original draft preparation, S.Y., H.X. and H.Z.; writing—review and editing, H.X.; visualization, S.Y.; supervision, H.X.; project administration, H.X. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data that support the findings of this study are available by request.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AIC	Akaike information criterion
BEKK	Baba–Engle–Kraft–Kroner
BIC	Bayesian information criterion
CV	Cross-validation
DCC	Dynamic conditional correlation
GARCH	Generalized autoregressive conditionally heteroskedastic
GMV	Global minimum variance
IR	Information ratio
LARS	Least-angle regression
LASSO	Least absolute shrinkage and selection operator
MGARCH	Multivariate GARCH
PQL	Penalized quasi-likelihood
PQML	Penalized quasi-maximum likelihood
SCAD	Smoothly clipped absolute deviation
SD	Standard deviation

Appendix A. Proofs of Propositions, Lemmas, and Theorems

Appendix A.1. Proof of Propostion 2

Let $\mathfrak{R}_t$,$\mathfrak{C}$, and Σ_t^* be defined by

$$\mathfrak{R}_t = (vech(r_t r_t')', \ldots, vech(r_{t-m+1} r_{t-m+1}')')', \qquad \Sigma_t^* = (vech(\Sigma_t)', \ldots, vech(\Sigma_{t-m+1})')',$$

where $m = max(a, b)$, and let

$$\mathfrak{C} = (vech(C'C)', 0, \ldots, 0)',$$

with dimensions $mn(n+1)/2$.

$$\text{Define } A = \begin{pmatrix} \widetilde{A}_1 & \dots & \dots & \dots & \widetilde{A}_m \\ I & 0 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & I & 0 \end{pmatrix} \text{ and } B = \begin{pmatrix} \widetilde{B}_1 & \dots & \dots & \dots & \widetilde{B}_m \\ I & 0 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & I & 0 \end{pmatrix},$$

with convention $\widetilde{A}_i = 0$ if $i > a$ and $\widetilde{B}_i = 0$ if $i > b$. Then, the model can be written as

$$\Sigma_t^* = \mathfrak{C} + A\mathfrak{R}_t + B\Sigma_{t-1}^* = \sum_{k=0}^{t-1}[B^k(\theta)\mathfrak{C}(\theta)] + B^t(\theta)\Sigma_0^* + \sum_{k=0}^{t-1} B^k(\theta)A(\theta)L^k\mathfrak{R}_{t-1}(\theta_0)$$

where L is the backshift operator $Lr_t = r_{t-1}$. Here, Σ_0^* is fixed and $\mathfrak{R}_t$ depends on θ_0 but is not a function of θ. Then, we have

$$\frac{\partial \Sigma_t^*}{\partial \theta_i} = \frac{\partial}{\partial \theta_i}(\sum_{k=0}^{t-1} \widetilde{B}^k\mathfrak{C}) + \frac{\partial}{\partial \theta_i}(\widetilde{B}^t)\Sigma_0^* + \frac{\partial}{\partial \theta_i}(\sum_{k=0}^{t-1} \widetilde{B}^k\widetilde{A}L)\mathfrak{R}_{t-1} \tag{A1}$$

Since

$$\frac{\partial \widetilde{B}^k}{\partial \theta} = \sum_{j=0}^{k-1} \widetilde{B}^j \frac{\partial \widetilde{B}}{\partial \theta_j} \widetilde{B}^{k-1-j},$$

we have

$$\left\| B^j \frac{\partial B}{\partial \theta_j} B^{k-1-j} \right\| \leq ||B^j|| \left\| \frac{\partial B}{\partial \theta_i} \right\| ||B^{k-1-j}||, \qquad j = 0, \dots, k-1.$$

Applying Lemma A.3. from Comte and Lieberman (2003), $||B^k|| \leq \Psi k^{n_0}\rho_0^k$ for all k, we have

$$\left\| B^j \frac{\partial B}{\partial \theta_j} B^{k-1-j} \right\| \leq \Psi^2 k^{n_0}\rho_0^k \left\| \frac{\partial B}{\partial \theta_i} \right\|.$$

in which n_0 is a fixed number, Ψ is a constant independent of θ, and $-1 < \rho_0 < 1$.
To bound (A1), there are three terms to bound:

$$\begin{aligned} \left\| \frac{\partial}{\partial \theta_i}(\sum_{k=0}^{t-1} \widetilde{B}^k\mathfrak{C}) \right\| &= \left\| \sum_{k=1}^{t-1} \frac{\partial \widetilde{B}^k}{\partial \theta_i}\mathfrak{C} + \sum_{k=0}^{t-1} \widetilde{B}^k \frac{\partial \mathfrak{C}}{\partial \theta_i} \right\| \leq \\ \sum_{k=1}^{t-1} \left\| \frac{\partial \widetilde{B}^k}{\partial \theta_i} \right\| ||\mathfrak{C}|| + \sum_{k=0}^{t-1} ||\widetilde{B}^k|| \left\| \frac{\partial \mathfrak{C}}{\partial \theta_i} \right\| & \\ \leq \Psi^2 ||\mathfrak{C}|| \sum_{k=1}^{t-1} k^{n_0}\rho_0^k \left\| \frac{\partial \widetilde{B}}{\partial \theta_i} \right\| + \Psi \left\| \frac{\partial \mathfrak{C}}{\partial \theta_i} \right\| \sum_{k=0}^{t-1} k^{n_0}\rho_0^k & \\ \leq \pi(n_0)\Psi\left(\Psi||\mathfrak{C}|| \cdot \left\| \frac{\partial \widetilde{B}}{\partial \theta_i} \right\| + \left\| \frac{\partial \mathfrak{C}}{\partial \theta_i} \right\|\right), & \end{aligned}$$

using $\sum_{k=0}^{t-1} k^{n_0}\rho_0^k \leq \sum_{k=0}^{t} k^{n_0}\rho_0^{k-1} \leq \sum_{k=0}^{\infty} k^{n_0}\rho_0^{k-1} = \pi(n_0)$, where $\pi(n_0)$ is a constant that only depends on n_0. And, if $\rho_0 = 0$, this term is then easily bounded because $\widetilde{B}$ is the nilpotent and all sums are finite. In the same way,

$$\left\| \frac{\partial}{\partial \theta_i}(\widetilde{B}^t)\Sigma_0^* \right\| \leq \Psi\pi(n_0) \left\| \frac{\partial \widetilde{B}}{\partial \theta_i} \right\| ||\Sigma_0^*||.$$

Finally,

$$||\frac{\partial}{\partial\theta_i}(\sum_{k=0}^{t-1}\widetilde{B}^k L^k \widetilde{A})\mathfrak{R}_{t-1}|| \le ||\sum_{k=0}^{t-1}(\frac{\partial}{\partial\theta_i}\widetilde{B}^k L^k \widetilde{A})\mathfrak{R}_{t-1}|| + ||[\sum_{k=0}^{t-1}\widetilde{B}^k L^k(\frac{\partial}{\partial\theta_i}\widetilde{A})\mathfrak{R}_{t-1}||.$$

Denote the first and second sums on the right-hand side of the inequality as T_1 and T_2, respectively, we have

$$||T_1|| \le \Psi^2\Big(\sum_{k=1}^{t-1} k^{n_0+1}\rho_0^{k-1}||\widetilde{A}||\cdot||\mathfrak{R}_{t-k-1}||\Big)\cdot||\frac{\partial\widetilde{B}}{\partial\theta_i}||$$
$$\le \Psi^2||\widetilde{A}||\cdot||\frac{\partial\widetilde{B}}{\partial\theta_i}||\cdot\Big(\sum_{k=1}^{t-1} k^{n_0+1}\rho_0^{k-1}\Big)\cdot\sup_t||\mathfrak{R}_t|| \le \pi(n_0+1)\Psi^2||\widetilde{A}||\cdot||\frac{\partial\widetilde{B}}{\partial\theta_i}||\cdot\sup_t||\mathfrak{R}_t||,$$

and

$$||T_2|| \le \Psi^2\pi(n_0)||\frac{\partial\widetilde{A}}{\partial\theta_i}||sup_t||\mathfrak{R}_t||.$$

By our assumption, $||\mathfrak{C}||, ||\frac{\partial\mathfrak{C}}{\partial\theta_i}||, ||\widetilde{A}||, ||\frac{\partial\widetilde{A}}{\partial\theta_i}||, ||\frac{\partial\widetilde{B}}{\partial\theta_i}||, ||\Sigma_0^*||$ are all bounded. And there exists a constant w such that $||\frac{\partial\Sigma_t^*}{\partial\theta_i}|| = m||\frac{\partial\Sigma_t}{\partial\theta_i}||$. Hence,

$$||\frac{\partial\Sigma_t}{\partial\theta_i}|| \le \Psi_1 + \Psi_2\sup_t||\mathfrak{R}_t||.$$

where $\Psi_1 = \Psi\pi(n_0)(\Psi||\mathfrak{C}||\cdot||\frac{\partial\widetilde{B}}{\partial\theta_i}|| + ||\frac{\partial\mathfrak{C}}{\partial\theta_i}||) + \Psi\pi(n_0)||\frac{\partial\widetilde{B}}{\partial\theta_i}||\cdot||\Sigma_0^*||$, and $\Psi_2 = \Psi^2\pi(n_0+1)||\widetilde{A}||\cdot||\frac{\partial\widetilde{B}}{\partial\theta_i}|| + \Psi^2\pi(n_0)||\frac{\partial\widetilde{A}}{\partial\theta_i}||$. □

Appendix A.2. Proof of Proposition 4

As $\frac{\partial l_t(\theta)}{\partial\theta_i} = Tr(\frac{\partial\Sigma_t}{\partial\theta_i}\Sigma_t^{-1} - r_t r_t'\Sigma_t^{-1}\frac{\partial\Sigma_t}{\partial\theta_i}\Sigma_t^{-1})$, where $Tr(\cdot)$ denote the trace of a matrix, and $E[r_t r_t'|\mathscr{F}_{t-1}] = \Sigma_t$, we, hence, have $E[\frac{\partial l_t(\theta)}{\partial\theta_i}|\mathscr{F}_{t-1}] = 0$, which means that $\frac{\partial l_t(\theta)}{\partial\theta_i}$ is a martingale difference. Then, we want to prove that $E[|T^{1/2}T^{-1}\sum_{t=1}T\frac{\partial l_t(\theta^0)}{\partial\theta_i}|^m] = E[|T^{-1/2}\sum_{t=1}^{T}\frac{\partial l_t(\theta^0)}{\partial\theta_i}|^m] < \infty$ holds for $m = 4$. By Lemma 2, this proof is thus completed if we show that $E[|\frac{\partial l_t(\theta^0)}{\partial\theta_i}|^4] < \infty$. By Proposition 2, $\left\|\frac{\partial\Sigma_t}{\partial\theta_i}\right\| \le \Psi_1 + \Psi_2\sup||vech(r_t r_t')||$. Since

$$\left\|\frac{\partial\Sigma_t}{\partial\theta_i}\Sigma_t^{-1} - r_t r_t'\Sigma_t^{-1}\frac{\partial\Sigma_t}{\partial\theta_i}\Sigma_t^{-1}\right\| = \left\|(I - r_t r_t'\Sigma_t^{-1})\frac{\partial\Sigma_t}{\partial\theta_i}\Sigma_t^{-1}\right\|$$
$$\le ||(I - r_t r_t'\Sigma_t^{-1})||\left\|\frac{\partial\Sigma_t}{\partial\theta_i}\Sigma_t^{-1}\right\| \le ||(I - r_t r_t'\Sigma_t^{-1})||\left\|\frac{\partial\Sigma_t}{\partial\theta_i}\right\||\Sigma_t^{-1}||,$$

it is equivalent to show that

$$E\left[\left|\frac{\partial l_t(\theta^0)}{\partial\theta_i}\right|^4\right] = E\left[Tr^4\left(\frac{\partial\Sigma_t}{\partial\theta_i}\Sigma_t^{-1} - r_t r_t'\Sigma_t^{-1}\frac{\partial\Sigma_t}{\partial\theta_i}\Sigma_t^{-1}\right)\right] < \infty.$$

Since $Tr(AB) \le ||A||\cdot||B||$ and $||\Sigma_t^{-1}||$ is finite, there exists a constant M such that $||\Sigma_t^{-1}|| \le M$ for all t. Additionally, $||(I - r_t r_t'\Sigma_t^{-1})|| \le ||I|| + ||r_t r_t'||\cdot||\Sigma_t||^{-1} \le 1 + M||r_t r_t'||$, then

$$Tr\left(\frac{\partial\Sigma_t}{\partial\theta_i}\Sigma_t^{-1} - r_t r_t'\Sigma_t^{-1}\frac{\partial\Sigma_t}{\partial\theta_i}\Sigma_t^{-1}\right) \le ||(I - r_t r_t'\Sigma_t^{-1})||||\Sigma_t^{-1}||\left\|\frac{\partial\Sigma_t}{\partial\theta_i}\right\|$$

$$E\left[Tr^4\left(\frac{\partial\Sigma_t}{\partial\theta_i}\Sigma_t^{-1} - r_t r_t'\Sigma_t^{-1}\frac{\partial\Sigma_t}{\partial\theta_i}\Sigma_t^{-1}\right)\right] \le E[(1 + M\sup||r_t r_t'||)^4(\Psi_1 + \Psi_2\sup_t||vech(r_t r_t')||)^4].$$

Because $||A|| \leq ||vech(A)|| \leq rank(A)||A||$, there exists a constant k such that $||r_t r_t'|| = k||vech(r_t r_t')||$. Hence, if we let $||vech(r_t r_t')|| = ||\mathfrak{R}_t||$,

$$\begin{aligned} & E[(1 + M\sup_t ||r_t r_t'||)^4 (\Psi_1 + \Psi_2 \sup_t ||\mathfrak{R}_t||)^4] \\ = \; & E[(1 + kM\sup_t ||\mathfrak{R}_t||)^4 (\Psi_1 + \Psi_2 \sup_t ||\mathfrak{R}_t||)^4] = E(\sum_{i=0}^{8} a_i ||\mathfrak{R}_t||^i) \end{aligned}$$

where $a_i s$ are constants. Since $r_t \sim \Sigma_t^{\frac{1}{2}} \epsilon_t$, where $\epsilon_t s$ follow a normal distribution, $r_t s$, hence, admit 16 moments of order. Hence, $E||\mathfrak{R}_t||^i < \infty$, for i from 0 to 8. Hence, $E(\sum_{i=0}^{8} a_i ||\mathfrak{R}_t||^i) < \infty$; then, $E[|\frac{\partial l_t(\theta^0)}{\partial \theta_i}|^4] < \infty$.

Next, we check (c) and (d). (c) is clear, as we said before. By (III) in Lemma 1, the derivative of $H_{\mathscr{U}_0,T}(\theta)$ is bounded. By the mean-value theorem,

$$vec(H_{\mathscr{U}_0,T}(\theta^{(1)}, 0) - H_{\mathscr{U}_0,T}(\theta^{(2)}, 0)) = \frac{\partial H_{\mathscr{U}_0,T}(\theta, 0)}{\partial \theta}\Big|_{\theta=\theta^*} \cdot (\theta^{(1)} - \theta^{(2)}).$$

Hence,

$$\begin{aligned} & ||H_{\mathscr{U}_0,T}(\theta^{(1)}, 0) - H_{\mathscr{U}_0 T}(\theta^{(2)}, 0)|| \leq ||vec(H_{\mathscr{U}_0,T}(\theta^{(1)}, 0) - H_{\mathscr{U}_0,T}(\theta^{(2)}, 0)|| \\ = \; & \left\| \frac{\partial H_{\mathscr{U}_0,T}(\theta, 0)}{\partial \theta}\Big|_{\theta=\theta^*} \cdot (\theta^{(1)} - \theta^{(2)}) \right\| \leq \left\| \frac{\partial H_{\mathscr{U}_0,T}(\theta, 0)}{\partial \theta}\Big|_{\theta=\theta^*} \right\| \cdot ||\theta^{(1)} - \theta^{(2)}|| \\ \leq \; & \widetilde{K} ||\theta^{(1)} - \theta^{(2)}|| \end{aligned}$$

where $\widetilde{K}$ is bounded by (iii) in Proposition 3. Hence, $\widetilde{K} = O_p(1)$ and θ^* lies between $\theta^{(1)}$ and $\theta^{(2)}$.

Next, we verify (e) with $\beta = \delta_0/2$. For every $i \in \{1, \ldots, p\}$, it is sufficient to show that $\max_{||v||=1} |(H^0_{i1,T}, \ldots, H^0_{iq,T})v| = O_p(T^{\delta_0/2})$ for a vector $v \in \mathbb{R}^q$. Using the Cauchy–Schwarz inequality and property of the norm, the left-hand side is bounded by $||(H^0_{i1,T}, \ldots, H^0_{iq,T})|| \leq q^{1/2} \max_{1 \leq j \leq q} |H_{ij,T}|$. Since, from (I) and (II) in Lemma 1, $H^0_{ij,T} = O_p(1)$ and $q = O(T^{\delta_0})$, the result follows. □

Appendix A.3. Proof of Lemma 1

First, consider the PQL $Q_T(\theta)$, as defined in (5), in the constrained $||\widehat{\theta}||_0$-dimensional subspace $\mathbb{S} := \{\theta \in \mathbb{R}^p : \theta^c = 0\}$ of $\mathbb{R}^p$, where θ^c denotes the subvector of θ formed by the components in $\widehat{\mathscr{U}^c}$. It follows from (12) that $Q_T(\theta)$ is strictly concave in a ball $\mathbb{N}_0 \in \mathbb{S}$ centered at $\widehat{\theta}$. This, along with (10), entails that $\widehat{\theta}$, as a critical point of $Q_T(\theta)$ in $\mathbb{S}$, is the unique maximizer of $Q_T(\theta)$ in $\mathbb{N}_0$.

Now, we show that $\widehat{\theta}$ is indeed a strict local maximizer of $Q_T(\theta)$ on the whole space $\mathbb{R}^p$. Take a small ball $\mathbb{N}_1 \subset \mathbb{R}^p$ centered at $\widehat{\theta}$ such that $\mathbb{N}_1 \cap \mathbb{S} \subset \mathbb{N}_0$. We then need to show that $Q_T(\widehat{\theta}) > Q_T(\gamma_1)$ for any $\gamma_1 \in \mathbb{N}_1 \setminus \mathbb{N}_0$. Let γ_2 be the projection of γ_1 onto $\mathbb{S}$, such that $\gamma_2 \in \mathbb{N}_0$. Thus, it suffices to prove that $Q_T(\gamma_2) > Q_T(\gamma_1)$. By the mean value theorem, we have

$$Q_T(\gamma_1) - Q_T(\gamma_2) = \frac{\partial Q_T(\gamma_0)}{\partial \gamma^T}(\gamma_1 - \gamma_2),$$

where the vector γ_0 lies between γ_1 and γ_2. Note that the components of $\gamma_1 - \gamma_2$ are zero for their indices in $\widehat{\mathscr{U}}$ and $(\gamma_{0j}) = sgn(\gamma_{1j})$ for $j \in \widehat{\mathscr{U}^c}$. Therefore, we have

$$\begin{aligned} \frac{\partial Q_T(\gamma_0)}{\partial \gamma^T}(\gamma_1 - \gamma_2) &= S_T(\gamma_0)^T(\gamma_1 - \gamma_2) - \lambda_T[\mathbf{1} \odot sgn(\gamma_0)]^T(\gamma_1 - \gamma_2) \\ &= S_{\widehat{\mathscr{U}^c}T}(\gamma_0)^T \gamma_{1\widehat{\mathscr{U}^c}} - \lambda_T \sum_{j \in \widehat{\mathscr{U}^c}} |\gamma_{1j}| \end{aligned} \tag{A2}$$

where $\gamma_{1\widehat{\mathscr{U}^c}}$ is a subvector of γ_1 formed by the components in $\widehat{\mathscr{U}^c}$. By (10), there exists some $\delta > 0$ such that, for any θ in a ball in $\mathbb{R}^p$ centered at $\widehat{\theta}$ with radius δ,

$$||S_{\widehat{\mathscr{U}^c}T}(\theta)||_\infty < \lambda_T \tag{A3}$$

We further shrink the radius of ball $\mathbb{N}_1$ to less than δ, so that $|\gamma_{0j}| \leq |\gamma_{1j}| < \delta$ for $j \in \widehat{\mathscr{U}^c}$ and (A3) holds for any $\theta \in \mathbb{N}_1$. Since $\gamma_0 \in \mathbb{N}_1$, it follows from (A3) that (A2) is strictly less than

$$\lambda_T||\gamma_{1\widehat{\mathscr{U}^c}}||_1 - \lambda_T||\gamma_{1\widehat{\mathscr{U}^c}}||_1 = 0$$

Since $||S_{\widehat{\mathscr{U}^c}T}(\gamma_0)||_\infty < \lambda_T$, $S_{\widehat{\mathscr{U}^c}T}(\gamma_0)^T\gamma_{1\widehat{\mathscr{U}^c}} \leq \lambda_T||\gamma_{1\widehat{\mathscr{U}^c}}||_1$, and

$$\lambda_T \sum_{j\in\widehat{\mathscr{U}^c}} |\gamma_{1j}| \geq \lambda_T \sum_{j\in\widehat{\mathscr{U}^c}} |\gamma_{1j}| = \lambda_T \sum_{j\in\widehat{\mathscr{U}^c}} ||\gamma_{1\widehat{\mathscr{U}^c}}||_1.$$

we have $\frac{\partial Q_T(\gamma_0)}{\partial \gamma^T}(\gamma_1 - \gamma_2) \leq 0$ and $Q_T(\gamma_1) \leq Q_T(\gamma_2)$. □

Appendix A.4. Proof of Lemma 2

A Marcinkiewicz–Zygmund inequality for martingales (Rio 2017) states that

$$E(\sum_{t=1}^{T} w_t)^m \leq \{4m(m-1)\}^{m/2}T^{(m-2)/2}\sum_{t=1}^{T} E|w_t|^m \tag{A4}$$

holds for $m > 2$. Because $E|w_t|^m \leq C_w$ for all t, we have

$$T^{-m/2}E(\sum_{t=1}^{T} w_t)^m \leq \{4m(m-1)\}^{m/2}T^{-1}\sum_{t=1}^{T} E|w_t|^m \leq \{4m(m-1)\}^{m/2}C_w. \tag{A5}$$

Thus, the result follows. □

Appendix A.5. Proof for Theorem 1

For notational simplicity, we write, for example, $Q_T(((\theta_{\mathscr{U}_0})', (\theta^c_{\mathscr{U}_0})')')$ as $Q_T(\theta_{\mathscr{U}_0}, \theta^c_{\mathscr{U}_0})$. Consider events

$$\mathscr{E}^1_T = \{||S^0_{\mathscr{U}_0,T}||_\infty \leq (q^{1/2}/T)^{1/2}log^{1/4}T\}, \qquad \mathscr{E}^2_T = \{||S^0_{\mathscr{U}^c_0,T}||_\infty \leq \lambda log^{-1}T\},$$

where $q = O(T^{\delta_0})$ and $\lambda = O(T^{-\alpha})$. It follows from Bonferroni's inequality and Markov's inequality, together with Proposition 4(i), that

$$\begin{aligned} P(\mathscr{E}^1_T \cap \mathscr{E}^2_T) &\geq 1 - \sum_{i\in\mathscr{U}_0} P(|T^{1/2}S^0_{i,T}| > q^{1/4}(\log T)^{1/4}) - \sum_{i\in\mathscr{U}^c_0} P(|T^{1/2}S^0_{i,T}| > T^{1/2-\alpha}) \\ &\geq 1 - \frac{\max_{i\in\mathscr{U}_0} E(|T^{1/2}S^0_{i,T}|^4)}{q\log T} - (p-q)\frac{\max_{i\in\mathscr{U}^c_0} E(|T^{1/2}S^0_{i,T}|^4)}{T^{4(1/2-\alpha)}(\log T)^{-4}} \\ &= 1 - O(\log^{-1}T) - O(T^{\delta-4(1/2-\alpha)}(\log T)^4), \end{aligned} \tag{A6}$$

where the last two terms are $o(1)$ because of the condition $\delta < 4(1/2-\alpha)$. Under the event $\mathscr{E}^1_T \cap \mathscr{E}^2_T$, we will that there exists a solution $\widehat{\theta} \in \mathbb{R}^p$ to (10)–(12) with $sgn(\widehat{\theta}) = sgn(\theta^0)$ and $||\widehat{\theta} - \theta^0||_\infty = O(T^{-\gamma}\log T)$ for some $\gamma \in (0, 1/2]$.

First, we prove that, for a sufficiently large T, Equation (10) has a solution $\widehat{\theta}_{\mathscr{U}_0}$ inside the hypercube $\mathbb{N} = \{\theta_{\mathscr{U}_0} \in \mathbb{R}^q : ||\theta_{\mathscr{U}_0} - \theta^0_{\mathscr{U}_0}||_\infty = T^{-\gamma}\log T\}$, when we suppose $\widehat{\mathscr{U}} = \mathscr{U}_0$. Define the function $\Psi : \mathbb{R}^q \to \mathbb{R}^q$ by

$$\Psi(\theta_{\mathscr{U}_0}) = S_{\mathscr{U}_0,T}(\theta_{\mathscr{U}_0}, 0) - \lambda\mathbf{1} \odot sgn(\theta_{\mathscr{U}_0}). \tag{A7}$$

Then, (10) is equivalent to $\Psi(\widehat{\theta}_{\mathscr{U}_0}) = 0$. To show that the solution is in the hypercube $\mathbb{N}$, we expand $\Psi(\theta_{\mathscr{U}_0})$ around $\theta^0_{\mathscr{U}_0}$. Function (A7) is written as

$$\begin{aligned}
&\Psi(\theta_{\mathscr{U}_0}) = S^0_{\mathscr{U}_0,T} + H_{\mathscr{U}_0 T}(\theta^*_{\mathscr{U}_0}, 0)(\theta_{\mathscr{U}_0} - \theta^0_{\mathscr{U}_0}) - \lambda \mathbf{1} \odot sgn(\theta_{\mathscr{U}_0}) \\
&= H^0_{\mathscr{U}_0,T}(\theta_{\mathscr{U}_0} - \theta^0_{\mathscr{U}_0}) + [S^0_{\mathscr{U}_0,T} - \lambda \mathbf{1} \odot sgn(\theta_{\mathscr{U}_0})] + [H_{\mathscr{U}_0,T}(\theta^*_{\mathscr{U}_0}, 0) - H_{\mathscr{U}_0,T}](\theta_{\mathscr{U}_0} - \theta^0_{\mathscr{U}_0}) \\
&= H^0_{\mathscr{U}_0,T}(\theta_{\mathscr{U}_0} - \theta^0_{\mathscr{U}_0}) + v_T + w_T
\end{aligned} \tag{A8}$$

where $\theta^*_{\mathscr{U}_0}$ lies on the line segment that joins $\theta_{\mathscr{U}_0}$ and $\theta^0_{\mathscr{U}_0}$. Since the matrix $H^0_{\mathscr{U}_0}$ is invertible by Proposition 4(ii), (A8) is further written as

$$\begin{aligned}
\widetilde{\Psi}(\theta_{\mathscr{U}_0}) &:= (H^0_{\mathscr{U}_0,T})^{-1}\Psi(\theta_{\mathscr{U}_0}) = \theta_{\mathscr{U}_0} - \theta^0_{\mathscr{U}_0} + (H^0_{\mathscr{U}_0,T})^{-1}v_T + (H^0_{\mathscr{U}_0,T})^{-1}w_T \\
&= \theta_{\mathscr{U}_0} - \theta^0_{\mathscr{U}_0} + \widetilde{v}_T + \widetilde{w}_T
\end{aligned} \tag{A9}$$

We now derive bounds for the last two terms in (A8). We consider $\widetilde{v}_T$ first. For any $\theta_{\mathscr{U}_0} \in \mathbb{N}$,

$$\min_{j \in \mathscr{U}_0} |\theta_j| \geq \min_{j \in \mathscr{U}_0} |\theta^0_j| - d_T = d_T \geq T^{-\gamma} \log T \tag{A10}$$

by Condition 3(ii), and $sgn(\theta_{\mathscr{U}_0}) = sgn(\theta^0_{\mathscr{U}_0})$. Using Condition 3(i), we have

$$||\lambda \mathbf{1} \odot sgn(\theta_{\mathscr{U}_0})||_\infty = \lambda \leq o(q^{-1/2}T^{-\gamma} \log T).$$

This, along with the property of matrix norms and Proposition 4(ii), entails that, during the event $\mathscr{E}^1_T$,

$$\begin{aligned}
||\widetilde{v}_T||_\infty &= ||H^{0-1}_{\mathscr{U}_0 T}[S^0_{\mathscr{U}_0 T} - \lambda \mathbf{1} \odot sgn(\theta_{\mathscr{U}_0})]||_\infty \\
&\leq ||H^{0-1}_{\mathscr{U}_0 T}||_\infty ||S^0_{\mathscr{U}_0 T} - \lambda \mathbf{1} \odot sgn(\theta_{\mathscr{U}_0})||_\infty \\
&\leq q^{1/2}||H^{0-1}_{\mathscr{U}_0 T}||(||S^0_{\mathscr{U}_0 T}||_\infty + ||\lambda \mathbf{1} \odot sgn(\theta_{\mathscr{U}_0})||_\infty) \\
&\leq q^{1/2}O_p(1)((q^{2/4}/T)^{1/2}(\log T)^{1/2} + o(q^{-1/2}T^{-\gamma} \log T)) \\
&= o_p(T^{-\gamma} \log T)
\end{aligned} \tag{A11}$$

where the last equality follows from $q = O(T^{\delta_0})$ and $\delta_0 < \frac{2}{3}(1 - 2\gamma)$. Next, we consider $\widetilde{w}_T$. By the property of norms and Propositions 4(ii) and (iii),

$$\begin{aligned}
||\widetilde{w}_T||_\infty &= ||(H^0_{\mathscr{U}_0,T})^{-1}(\theta^*_{\mathscr{U}_0}, 0)[H_{\mathscr{U}_0,T}(\theta^*_{\mathscr{U}_0}, 0) - H^0_{\mathscr{U}_0,T}](\theta_{\mathscr{U}_0} - \theta^0_{\mathscr{U}_0})||_\infty \\
&\leq q^{1/2}||(H^0_{\mathscr{U}_0,T})^{-1}||||[H_{\mathscr{U}_0,T}(\theta^*, 0) - H^0_{\mathscr{U}_0,T}](\theta_{\mathscr{U}_0} - \theta^0_{\mathscr{U}_0})||_\infty \\
&\leq qO_p(1)||H_{\mathscr{U}_0,T}(\theta^*_{\mathscr{U}_0}, 0) - H^0_{\mathscr{U}_0,T}||||\theta_{\mathscr{U}_0} - \theta^0_{\mathscr{U}_0}||_\infty \\
&\leq qO_p(1)K_T||\theta^*_{\mathscr{U}_0} - \theta^0_{\mathscr{U}_0}||_\infty||\theta_{\mathscr{U}_0} - \theta^0_{\mathscr{U}_0}||_\infty,
\end{aligned}$$

Since $K_T = O_p(1)$ and $q = O(T^{\delta_0})$ with $\delta_0 < \gamma$,

$$||\widetilde{w}_T||_\infty = qO_p(T^{-2\gamma}(\log T)^2) = o_p(T^{-\gamma} \log T), \tag{A12}$$

with $\theta_i - \theta^0_i = T^{-\gamma} log T$ for all $i \in \mathscr{U}_0$. By (A9), (A11), and(A12), for sufficiently large T, and for all $i \in \mathscr{U}_0$,

$$\widetilde{\Psi}_i(\theta_{\mathscr{U}_0}) \geq T^{-\gamma} log T - ||\widetilde{v}_T||_\infty - ||\widetilde{w}_T||_\infty \geq 0, \tag{A13}$$

if $\theta_i - \theta^0_i = T^{-\gamma} \log T$, and

$$\widetilde{\Psi}_i(\theta_{\mathscr{U}_0}) \leq -T^{-\gamma} log T + ||\widetilde{v}_T||_\infty + ||\widetilde{w}_T||_\infty \leq 0, \tag{A14}$$

if $\theta_i - \theta^0_i = -T^{-\gamma} \log T$.

By the continuity of $\widetilde{\Psi}$ and inequalities (A13) and (A14), an application of Miranda's existence theorem tells us that $\widetilde{\Psi}(\theta_{\mathscr{U}_0}) = 0$ has a solution $\widehat{\theta}_{\mathscr{U}_0}$ in $\mathbb{N}$. Clearly, $\widehat{\theta}_{\mathscr{U}_0}$ also solves the equation $\Psi(\theta_{\mathscr{U}_0}) = 0$ with regard to the first equality in (A8). Thus, we have shown that (10) indeed has a solution in $\mathbb{N}$.

Second, let $\widehat{\theta} = (\widehat{\theta}'_{\mathscr{U}_0}, \widehat{\theta}'_{\mathscr{U}_0^c})' \in \mathbb{R}^p$, with $\widehat{\theta}_{\mathscr{U}_0} \in \mathbb{N}$ as a solution to (10), and $\widehat{\theta}_{\mathscr{U}_0^c} = 0$. Next, we show that $\widehat{\theta}$ satisfies (11) for the event $\mathscr{E}_T^2$. By the triangle inequality and mean value theorem, we have

$$\begin{aligned} \lambda^{-1}||S_{\mathscr{U}_0^c T}(\widehat{\theta})||_\infty &\leq \lambda^{-1}||S^0_{\mathscr{U}_0^c T}||_\infty + \lambda^{-1}||S_{\mathscr{U}_0^c T}(\widehat{\theta}) - S^0_{\mathscr{U}_0^c T}||_\infty \\ &\leq (\log T)^{-1} + \lambda^{-1}||(\partial/\partial\theta_{\mathscr{U}_0})S_{\mathscr{U}_0^c T}(\widehat{\theta}^{**}_{\mathscr{U}_0}, 0)(\widehat{\theta}_{\mathscr{U}_0} - \theta^0_{\mathscr{U}_0})||_\infty, \end{aligned} \tag{A15}$$

where $\widehat{\theta}^{**}_{\mathscr{U}_0}$ lies on the line segment joining $\widehat{\theta}_{\mathscr{U}_0}$ and $\theta^0_{\mathscr{U}_0}$. The first term of the upper bound in (A15) is negligible, so that it suffices to show that the second term is less than $g'(0+) = 1$. Since $\widehat{\theta}_{\mathscr{U}_0}$ solves the equation $\Psi(\theta_{\mathscr{U}_0}) = 0$ in (12), we obtain

$$S^0_{\mathscr{U}_0 T} + H_{\mathscr{U}_0 T}(\widehat{\theta}^*_{\mathscr{U}_0}, 0)(\widehat{\theta}_{\mathscr{U}_0} - \theta^0_{\mathscr{U}_0}) - \lambda \mathbf{1} \odot sgn(\widehat{\theta}_{\mathscr{U}_0}) = 0$$

with $\widehat{\theta}^*_{\mathscr{U}_0}$ lying between $\widehat{\theta}_{\mathscr{U}_0}$ and $\theta^0_{\mathscr{U}_0}$. From Proposition 4(ii),(iii) and Condition 1, the last term in (A15) can be expressed as

$$\begin{aligned} & \lambda^{-1}||(\partial/\partial\theta_{\mathscr{U}_0})S_{\mathscr{U}_0^c T}(\widehat{\theta}^{**}_{\mathscr{U}_0}, 0)[H_{\mathscr{U}_0 T}(\widehat{\theta}^*_{\mathscr{U}_0}, 0)]^{-1}[S^0_{\mathscr{U}_0 T} - \lambda \mathbf{1} \odot sgn(\widehat{\theta}_{\mathscr{U}_0})]||_\infty \\ \leq\ & \lambda^{-1} \sup_{\theta, \theta' \in \mathbb{N}} ||(\partial/\partial\theta_{\mathscr{U}_0})S_{\mathscr{U}_0^c T}(\theta, 0)[H_{\mathscr{U}_0 T}(\theta', 0)]^{-1}||_\infty(||S^0_{\mathscr{U}_0 T}||_\infty + \lambda) \\ \leq\ & \lambda^{-1} c(q^{1/2}/T)^{1/2} log^{1/2} T + \lambda \\ =\ & \lambda^{-1} c(q^{1/2}/T)^{1/2} \log^{1/2} T + c. \end{aligned} \tag{A16}$$

By Condition 3(i), the first term in the last equation of (A16) is $o_p(1)$; hence, (A16) is eventually less than 1. This verifies (11).

Finally, (12) is guaranteed by Lemma 1: we have $\widehat{\theta}$ as a strict local maximizer of $Q_T(\theta)$ with $||\widehat{\theta} - \theta^0||_\infty = O(T^{-\gamma} \log T)$ and $\widehat{\theta}_{\mathscr{U}_0^c} = 0$ in the event that $\mathscr{E}_T^1 \cap \mathscr{E}_T^2$. Thus, the proofs of Theorems 1(a) and (b) are complete, by (A6). □

References

Aielli, Gian Piero. 2013. Dynamic conditional correlation: On properties and estimation. *Jouranl of Business and Economic Statistics* 31: 282–99. [CrossRef]

Alexander, Carol. 2000. Orthogonal methods for generating large positive semi-definite covariance matrices. In *ICMA Centre Discussion Papers in Finance icma-dp2000-06*. London: Henley Business School, Reading University.

Ampountolas, Apostolos. 2022. Cryptocurrencies intraday high-frequency volatility spillover effects using univariate and multivariate GARCH models. *International Journal of Financial Studies* 10: 51. [CrossRef]

Apergis, Nicholas, and Anthony Rezitis. 2001. Asymmetric cross-market volatility spillovers: Evidence from daily data on equity and foreign exchange markets. *The Manchester School* 69: 81–96. [CrossRef]

Apergis, Nicholas, and Anthony Rezitis. 2003. An examination of okun's law: Evidence from regional areas in greece. *Applied Economics* 35: 1147–51. [CrossRef]

Baillie, Rechard T., and Tim Bollerslev. 1990. A multivariate generalized arch approach to modeling risk premia in forward foreign exchange rate markets. *Journal of International Money and Finance* 9: 309–24. [CrossRef]

Basu, Sumanta, and George Michailidis. 2015. Regularized estimation in sparse high-dimensional time series model. *The Annals of Statistics* 43: 1535–67. [CrossRef]

Bauwens, Luc, and Sébastien Laurent. 2005. A new class of multivariate skew densities, with application to generalized autoregressive conditional heteroscedasticity models. *Journal of Business and Economic Statistics* 23: 346–54. [CrossRef]

Bickel, Peter J. and Elizaveta Levina. 2008. Covariance regularization by thresholding. *The Annals of Statistics* 36: 2577–604. [CrossRef] [PubMed]

Billio, Monica, Massimiliano Caporin, Lorenzo Frattarolo, and Loriana Pelizzon. 2023. Networks in risk spillovers: A multivariate GARCH perspective. *Econometrics and Statistics* 28: 1–29. [CrossRef]

Bollerslev, Tim. 1986. Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics* 31: 307–27. [CrossRef]

Bollerslev, Tim. 1990. Comparing predictive accuracy modelling the coherence in short-run nominal exchange rates: A multivariate generalized arch model. *The Review of Economics and Statistics* 72: 498–05. [CrossRef]
Bollerslev, Tim, Robert Engle, and Jeffrey Wooldridge. 1988. A capital asset pricing model with time-varying covariances. *Journal of Political Economy* 96: 116–31. [CrossRef]
Boudt, Kris, Jon Danielsson, and Sébastien Laurent. 2013. Robust forecasting of dynamic conditional correlation garch models. *International Journal of Forecasting* 29: 244–57. [CrossRef]
Brodie, Joshua, Ingrid Daubechies, Christine De Mol, Domenico Giannone, and Ignace Loris. 2009. Sparse and stable markowitz portfolios. *Proceedings of the National Academy of Sciences of the United States of America* 106: 12267–72. [CrossRef]
Cai, Tony, and Weidong Liu. 2011. Adaptive thresholding for sparse covariance matrix estimation. *Journal of the American Statistical Association* 106: 672–84. [CrossRef]
Christiansen, Charlotte. 2007. Volatility-Spillover Effects in European Bond Markets. *European Financial Management* 13: 923–948. [CrossRef]
Comte, Fabienne, and Offer Lieberman. 2003. Asymptotic theory for multivariate garch processes. *Journal of Multivariate Analysis* 84: 61–84. [CrossRef]
DeMiguel, Victor, Lorenzo Garlappi, and Raman Uppal. 2007. Optimal versus naive diversification: How inefficient is the 1/N portfolio strategy? *The Review of Financial Studies* 22: 1915–53. [CrossRef]
Diebold, Francis X., and Kamil Yilmaz. 2009. Measuring financial asset return and volatitliy spillovers, with application to global equity markets. *Economic Journal* 199: 158–71. [CrossRef]
Di Lorenzo, David, Giampalo Liuzzi, Francesco Rinaldi, Fabio Schoen, and Marco Sciandrone. 2012. A concave optimization-based approach for sparse portfolio selection. *Optimization Methods and Software* 27: 983–1000. [CrossRef]
Efron, Bradley, Trevor Hastie, and Robert Tibshirani. 2004. Least angle regression. *The Annals of Statistics* 32: 407–99. [CrossRef]
Engle, Rober. 1982. Autoregressive conditional heteroskedasticity with estimates of the variance of united kingdom inflation. *Econometrica* 50: 987–1007. [CrossRef]
Engle, Robert. 1990. Asset pricing with a factor-arch covariance structure: Empirical estimates for treasury bills. *Journal of Econometrics* 45: 213–37. [CrossRef]
Engle, Robert. 2002. Dynamic conditional correlation: A simple class of multivariate generalized autoregressive conditional heteroskedasticity models. *Journal of Business and Economic Statistics* 20: 339–50. [CrossRef]
Engle, Robert, and Kenneth Kroner. 1995. Multivariate simultaneous generalized arch. *Econometric Theory* 11: 122–50. [CrossRef]
Engle, Robert, and Riccardo Colacito. 2006. Testing and valuing dynamic correlations for asset allocation. *Journal of Business and Economic Statistics* 24: 238–53. [CrossRef]
Engle, Robert, Olivier Ledoit, and Michael Wolf. 2019. Large dynamic covariance matrices. *Journal of Business and Economic Statistics* 37: 363–75. [CrossRef]
Engle, Robert, Takatoshi Ito, and Wen-Ling Lin. 1990. Meteor showers or heat waves? Heteroskedastic intra-daily volatility in the foreign exchange market. *Econometrica* 58: 525–42. [CrossRef]
Fan, Jianqing, and Jinchi Lv. 2011. Noncave penalized likelihood with np-dimensionality. *IEEE Transactions on Information Theory* 57: 5467–84. [CrossRef]
Fan, Jianqing, and Runze Li. 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* 96: 1348–60. [CrossRef]
Fan, Yingying, and Cheng Yong Tang. 2013. Tuning parameter selection in high dimensional penalized likelihood. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 75: 531–52. [CrossRef]
Fastrich, Björn, Sandra Paterlini, and Peter Winker. 2015. Constructing optimal sparse portfolios using regularization methods. *Computational Management Science* 12: 417–34. [CrossRef]
Francq, Christian, and Jean-Michel Zakoian. 2019. *GARCH Models: Structure, Statistical Inference and Financial Applications*. Hoboken: John Wiley & Sons.
Friedman, Jerome, Trevor Hastie, Holger Höfling, and Robert Tibshirani. 2007. Pathwise coordinate optimization. *The Annals of Applied Statistics* 1: 302–32. [CrossRef]
Giacometti, Rosella, Gabriele Torri, Kamonchai Rujirarangsan, and Michela Cameletti. 2023. Spatial Multivariate GARCH Models and Financial Spillovers. *Journal of Risk and Financial Management* 16: 397. [CrossRef]
Hamao, Yasushi, Ronald W. Masulis, and Victor Ng. 1990. Correlations in price changes and volatility across international stock markets. *The Review of Financial Studies* 3: 281–307. [CrossRef]
Hafner, Christian M., and Arie Preminger. 2009. Asymptotic theory for a factor GARCH model. *Econometric Theory* 25: 336–63. [CrossRef]
Hafner, Christian M., Helmut Herwartz, and Simone Maxand. 2022. Identification of structural multivariate GARCH models. *Journal of Econometrics* 227: 212–27. [CrossRef]
Hassan, Syed Aun, and Farooq Malik. 2007. Multivariate garch modeling of sector volatility transmission. *The Quarterly Review of Economics and Finance* 47: 470–80. [CrossRef]
Hong, Junping, Yi Yan, Ercan Engin Kuruoglu, and Wai Kin Chan. 2023. Multivariate Time Series Forecasting With GARCH Models on Graphs. *IEEE Transactions On Signal And Information Processing Over Networks* 9: 557–68. [CrossRef]

Kaltenhäuser, Bernd. 2002. *Return and Volatility Spillovers to Industry Returns: Does EMU Play a Role?* CFS Working Paper Series 2002/05. Frankfurt a. M.: Center for Financial Studies (CFS).
Lam, Clifford, and Jianqing Fan. 2009. Sparsistency and rates of convergence in large covariance matrix estimation. *The Annals of Statistics* 37: 4254–78. [CrossRef] [PubMed]
Lanne, Markku, and Pentti Saikkonen. 2007. A multivariate generalized orthogonal factor GARCH model. *Journal of Business & Economic Statistics* 25: 61–75.
Ledoit, Olivier, and Michael Wolf. 2004. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis* 88: 365–411. [CrossRef]
Ledoit, Olivier, and Michael Wolf. 2012. Nonlinear shrinkage estimation of large-dimensional covariance matrices. *The Annals of Statistics* 40: 1024–60. [CrossRef]
Ling, Shiqing, and Michael McAleer. 2003. Asymptotic theory for a vector arma-garch model. *Econometric Theory* 19: 280–310. [CrossRef]
Markowitz, Harry. 1952. Portfolio selection. *The Journal of Finance* 7: 77–91.
McAleer, Michael, Suhejia Hoti, and Felix Chan. 2009. Structure and asymptotic theory for multivariate asymmetric conditional volatility. *Econometric Reviews* 28: 422–40. [CrossRef]
NASDAQ Stock Symbols. n.d. Stock Symbol. Available online: https://www.nasdaq.com/market-activity/stocks/ (accessed on 24 January 2024).
Nicholson, William B., David S. Matteson, and Jacob Bien. 2017. VARX-L: Structured regularization for large vector autoregressions with exogenous variables. *International Journal of Forecasting* 33: 627–51. [CrossRef]
Pan, Ming-Shiun, and L. Paul Hsueh. 1998. Transmission of stock returns and volatility between the U.S. and Japan: Evidence from the stock index futures markets. *Asia-Pacific Financial Markets* 5: 211–25. [CrossRef]
Poignard, Benjamin. 2017. New Approaches for High-Dimensional Multivariate Garch Models. General Mathematics [math.GM]. Ph.D. thesis, Université Paris Sciences et Lettres, Paris, France
Ravikumar, Pradeep, Martin J. Wainwright, Garvesh Raskutti, and Bin Yu. 2011. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electronic Journal of Statistics* 5: 935–80. [CrossRef]
Rio, Emmanuel. 2017. *Asymptotic Theory of Weakly Dependent Random Processes*. Berlin: Springer Nature.
Sánchez García, Javier, and Salvador Cruz Rambaud. 2022. Machine Learning Regularization Methods in High-Dimensional Monetary and Financial VARs. *Mathematics* 10: 877. [CrossRef]
Shiferaw, Yegnanew A. 2019. Time-varying correlation between agricultural commodity and energy price dynamics with Bayesian multivariate DCC-GARCH models. *Physica A: Statistical Mechanics and Its Applications* 526: 120807. [CrossRef]
Siddiqui, Taufeeque Ahmad, and Mazia Fatima Khan. 2018. Analyzing spillovers in international stock markets: A multivariate GARCH approach. *IMJ* 10: 57–63.
Sun, Wei, Junhui Wang, and Yixin Fang. 2013. Consistent selection of tuning parameters via variable selection stability. *Journal of Machine Learning Research* 14: 3419–40.
Sun, Yan, and Xiaodong Lin. 2011. Regularization for stationary multivariate time series. *Quantitative Finance* 12: 573–86. [CrossRef]
Theodossiou, Panayiotis, and Unro Lee. 1993. Mean and volatility spillovers across major national stock markets: Further empirical evidence. *The Journal of Financial Research* 16: 337–50. [CrossRef]
Tibshirani, Robert. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* 58: 267–88. [CrossRef]
Tse, Yiu Kuen, and Albert K. C. Tsui. 2002. A multivariate generalized autoregressive conditional heteroscedasticity model with time-varying correlations. *Journal of Business & Economic Statistics* 20: 351–62.
Uematsu, Yoshimasa. 2015. Penalized likelihood estimation in high-dimensional time series models and its application. *arXiv* arXiv:1504.06706.
van der Weide, Roy. 2002. Go-garch: A multivariate generalized orthogonal garch model. *Journal of Applied Econometrics* 17: 549–64. [CrossRef]
Vrontos, Ioannis, Petros Dellaportas, and Dimitris N. Politis. 2003. A full-factor multivariate garch model. *The Econometrics Journal* 6: 312–34. [CrossRef]
Wang, Hansheng, Bo Li, and Chenlei Leng. 2009. Shrinkage tuning parameter selection with a diverging number of parameters. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 71: 671–83. [CrossRef]
Worthington, Andrew, and Helen Higgs. 2004. Transmission of equity returns and volatility in asian developed and emerging markets: A multivariate garch analysis. *International Journal of Finance & Economics* 9: 71–80.
Wu, Tong Tong, and Kenneth Lange. 2008. Coordinate descent algorithms for lasso penalized regression. *Annals of Applied Statistics* 2: 224–44. [CrossRef]
Yuan, Ming, and Yi Lin. 2006. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 68: 49–67. [CrossRef]
Zhang, Cun-Hui. 2010. Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics* 38: 894–942. [CrossRef]
Zhang, Yongli, and Yuhong Yang. 2015. Cross-validation for selecting a model selection procedure. *Journal of Econometrics* 187: 95–112. [CrossRef]
Zhao, Peng, and Bin Yu. 2006. On model selection consistency of lasso. *Journal of Machine Learning Research* 7: 2541–67.

Zhao, Peng, and Bin Yu. 2007. Stagewise lasso. *Journal of Machine Learning Research* 8: 2701–26.
Zou, Hui. 2006. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* 101: 1418–29. [CrossRef]