DYNAMIC ORTHOGONAL SUBSERIES FOR HIGH-DIMENSIONAL AND NONSTATIONARY TIME SERIES

A Dissertation

by

XIAO WANG

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Mohsen Pourahmadi
Committee Members,	Raymond Carroll
	Suhasini Subba Rao
	Seth Murray
Head of Department,	Valen Johnson

May 2017

Major Subject: Statistics

ABSTRACT

A multivariate time series could be partitioned either horizontally (over time) to induce local stationarity or vertically (over the variables) to reduce dimension and the high computational cost. Dimension reduction for a high-dimensional time series by linearly transforming it into several lower-dimensional subseries (vertical partition) where any two subseries are uncorrelated both temporally and cross-sectionally is of central importance in the modern age of big data. It reduces the challenging multivariate estimation problem with many parameters to that of a number of disjoint lower-dimensional problems with much fewer parameters. A notable example in the previous studies is the dynamic orthogonal components (DOC) utilizing nonlinear optimization which works well for stationary and low-dimensional time series data. First we reduce the computational burden of DOC by connecting it to the time series principal components analysis (TS-PCA) method in recent studies based on eigenanalysis of a positive-definite matrix. Next, we extend DOC to nonstationary processes which can be divided into several nearly homogeneous segments. Consistency and joint asymptotic normality of the estimates of the Givens angles parameterizing orthogonal matrices in each segment are established under some regularity conditions. Applications to multivariate volatility modeling in finance are illustrated using simulated and real datasets.

DEDICATION

To my mother, my father, my grandfather, and my grandmother.

ACKNOWLEDGMENTS

First and foremost I want to thank my advisor Dr. Mohsen Pourahmadi. It has been an honor to be his Ph.D. student. He has taught me, how to apply statistical methods, especially the multivariate time series analysis techniques to analyze real-world datasets. I appreciate all his contributions of time and ideas to make my Ph.D. experience productive and stimulating. The joy and enthusiasm he has for his research was contagious and motivational for me, even during tough times in the Ph.D. pursuit. I am also thankful for the excellent example he has provided as a successful statistician and professor.

I would also like to thank Dr. Xianyang Zhang for contributing immensely to my professional time at TAMU. He has been a source of friendship as well as good advice and collaboration. We worked together (along with Dr. Mohsen Pourahmadi) on this Dynamic Orthogonal Subseries project, and I very much appreciate his enthusiasm, intensity, willingness to help me solve the problems encountered in the Asymptotic proofs and Numerical Analysis sections.

I am also thankful to all my committee members, Dr. Raymond Carroll, Dr. Suhasini Subba Rao, Dr. Seth Murray for their direction, dedication, and invaluable advice along this project.

Lastly, I would like to thank my family for all their love and encouragement. For my parents who raised me with a love of science and supported me in all my pursuits. Thank you all.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supported by a dissertation committee consisting of Professor Mohsen Pourahmadi, Raymond Carroll and Suhasini Subba Rao of the Department of Statistics and Professor Seth Murray of the Department of Soil and Crop Sciences.

All work for the dissertation was completed by the student, in collaboration with Professor Mohsen Pourahmadi and Xianyang Zhang of the Department of Statistics.

Funding Sources

Graduate study was supported by a teaching assistantship from Texas A&M University.

NOMENCLATURE

ACF	Auto-Correlation function
AIC	Akaike information criterion
AR	AutoRegression
ARCH	AutoRegressive Conditional Heteroskedasticity
ARMA	AutoRegressive Moving Average
ARIMA	AutoRegressive Integrated Moving Average
BEKK	Baba-Engle-Kraft-Kroner model
CCA	Canonical Correlation Analysis
CCC	Constant Conditional Correlations
CCF	Cross Correlation Function
CISCO	Computer Information System Company
CUC	Conditionally Uncorrelated Components
DCC	Dynamic Conditional Correlation
DOC	Dynamic Orthogonal Component
FA	Factor Analysis
FRED-MD	Federal Reserve Economic Data - Monthly Data
ICA	Independent Component Analysis
INTEL	INTegrated ELectronics Company
L_2-NED	L_2-Near-Epoch-Dependent
MD	Minimum Distance
MSE	Mean Squared Error

OC	Orthogonal Component
PC	Principal Component
PCA	Principal Component Analysis
PCV	Principal Component Volatility
S & P 500	Standard & Poor's 500 index
SD	Standard Deviation
SDE	Spectral Density Estimation
TS-PCA	Time Series PCA
TVDOC	Time-Varying Dynamic Orthogonal Component
VAR	Vector AutoRegression
VEC-GARCH	VECtorized GARCH

TABLE OF CONTENTS

Page

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

TABLE Page

1. INTRODUCTION

High-dimensional nonstationary time series data are often encountered in a variety of fields, such as management of climate risks in agriculture ([44]), electrocardiogram analysis ([2]), electroencephalography (EEG) analysis ([30], [40]), macroeconomics analysis ([11], [36]), trading volatility analysis ([27]), bond price prediction ([50]), etc.

Fitting multivariate models to high-dimensional time series data is computationally expensive and will encounter convergence problems in optimization routine due to the large number of parameters involved. For example, the standard multivariate volatility models such as VEC-GARCH([13]) or BEKK([25]), formulate the conditional covariance matrix in terms of linear combinations of the squares and cross products of the data and thus the number of parameters contained in the coefficient matrix for a d dimensional time series are $O(d^2)$ and will be problematic for large d.

Dimension reduction is thus of vital importance to the analysis of high-dimensional time series data. In the following we will review some of the most popular dimension reduction methods. Principal component analysis (PCA), developed by [45] and [31], explains the covariance structure of a set of variables through a few linear combinations of them, called principal components (PCs), with decreasing variance. An integer r ($r << d$) is chosen such that the first r PCs explain a high percentage of the total variation of the data and dimensional reduction is achieved by analyzing the r PCs instead of the original time series. For example, orthogonal GARCH (O-GARCH) [3] and generalized orthogonal GARCH (GO-GARCH) models [51] use PCA of the covariance matrix to decorrelate a multivariate series of asset returns cross-sectionally before applying volatility models to the uncorrelted components separately. Canonical correlation analysis (CCA) ([32], [14]) finds linear combinations of variables such that the "predictability" of the transformed vari-

ables are maximized sequentially. "Predictability" is measured by the cross-correlation between the current observation of the transformed variable with its own past. Factor model ([46]) finds a canonical representation of the time series using a mall number of common factors. The number of common latent factors is decided by the eigen-analysis of the lagged sample covariance matrices. Independent component analysis (ICA) ([7]) projects a multivariate time series to a new space spanned by non-Gaussian independent components. Similar to PCA, ICA also considers decorrelating the components cross-sectionally, but with respect to their higher order moments. For example, Kurtosis or the fourth-moment cumulant, is often applied to measure the non-Gaussianity in ICA. Dynamic factor models ([15], [28] and [23]) in spectral-domain, as analogues of PCA, have been proposed to explain the serial correlations in the latent factors, where the factor loadings are derived from an eigen-analysis of the spectral density matrix.

It is ironic that most of these multivariate statistics techniques are applied almost verbatim to time series data without adequate accounting of the temporal dependence. This problem can be solved by dynamic orthogonal component (DOC) method ([41]) which decorrelates the time series both contemporaneously and serially. It finds an orthogonal transformation matrix $\mathbf{M}$ such that it minimizes the sum of squares of the off-diagonal entries of the first few lagged autocovariance matrices of the transformed series. Once $\mathbf{M}$ is found, modeling a high-dimensional series is then reduced to analyzing a sequence of univariate series which is much easier to model and predict, and then the univariate results are combined and transformed back to a parsimonious model for the original high-dimensional series.

The mixing matrix $\mathbf{M}$ in DOC is parameterized using the Givens angles and finding it involves solving a non-convex optimization problem. This makes DOC computationally expensive and hence not suitable for high-dimensional multivariate time series. However, we show that the recent time series PCA (TS-PCA) method ([17]) which segments a high-

dimensional time series into several lower-dimensional *decorrelated subseries*, is of great help in managing the computational bottle-neck. In the following the phrase *decorrelated subseries* means that they are uncorrelated both contemporaneously and serially. In this setting, the subseries are modelled separately using the DOC method by taking advantage of the substantial dimension reduction. The TS-PCA method is based on an eigenanalysis of a positive definite matrix defined as a quadratic function of the first few autocovariance matrices, and can be viewed as a natural extension of the standard PCA for multiple time series. As in PCA, it finds a time-invariant matrix transforming the series into several decorrelated subseries, but unlike PCA the subseries have varying dimensions.

Covariance stationarity is a key assumption in developing the statistical theory of the DOC methodology. However, for many real life examples the stationarity feature is often violated, and it is common to consider classes of nonstationary models such as locally stationary and piecewise stationary processes([1] and [19]). For example, [20] and [21] have considered piecewise AR and GARCH models, respectively, and in [5] the nature of nonstationarity is due to time-varying covariance matrices of multivariate time series.

The primary contribution of our work is to extend the stationary DOC methodology in [41] to a high-dimensional nonstationary setup where the series can be segmented into several locally stationary segments. We rely on a change point detection method in [18] to divide the whole observed series into several nearly stationary segments, and then DOC is applied to each segment separately. Our secondary contribution is to ease the computational burden of DOC for high-dimensional series and potentially replace the nonconvex optimization in DOC by the eigenanalysis of a positive-definite function as in TS-PCA. This second goal is nearly achieved by exploring the equivalence or close connection between DOC and TS-PCA both for in mean and in volatility [16, Section 5].

The outline of the dissertation is as follows: In Section 2 the key ideas and steps of DOC and TS-PCA methods are reviewed. Also the connection between these two

methods are discussed. A methodology for time-varying DOC (TVDOC) is developed in Section 3 for a class of multivariate piecewise stationary times series. Asymptotic properties of the estimators of the Givens angles of the mixing matrices for various segments are derived under some regularity conditions on the underlying processes. The TVDOC method is illustrated using simulation and real data where the important role of TS-PCA is highlighted. Section 4 concludes the dissertation. The technical proofs and additional information about the numerical examples are provided in the Appendix.

2. DOC AND TS-PCA FOR MULTIVARIATE STATIONARY TIME SERIES

2.1 Introduction

Fitting standard multivariate time series models such as vector autoregressive (VAR) models to high-dimensional data is challenging statistically due to a large number of parameters. In the time-domain, dimension reduction methods such as canonical correlation analysis [14], factor models [46], principal component analysis [3, 48, 9] and independent component analysis [7] are based on the idea of finding instantaneous linear combinations of the variables with simpler univariate time series structures. In the spectral-domain, analogues of the principal component analysis (PCA) and factor models have been introduced by [15], [28] and [23] for stationary and nonstationary time series, respectively, where here linear combinations may involve current and lagged values of the observed multivariate series.

Dimension reduction via time-invariant linear transformations of a multivariate time series has the more ambitious goal of extending the classical PCA from sample data to the (dependent) time series data setup. Its key task is to find a matrix so that the subseries of the transformed (segmented, vertically partitioned) series are *decorrelated* [41, 17] in the sense that they are *uncorrelated both contemporaneously and serially*. This more stringent requirement is in contrast to some of the earlier approaches cited above, and those in finance like the orthogonal GARCH (O-GARCH) [3] and generalized orthogonal GARCH (GO-GARCH) models [51] where PCA of the marginal (lag-zero) covariance matrix of the data is used to *decorrelate only cross-sectionally* a multivariate series of asset returns. Another related method is the independent component anslysis (ICA) which finds a matrix such that the linearly transformed subseries are independent cross-sectionally [34].

More formally for any multivariate time series $\boldsymbol{X}_t = (x_{1,t}, x_{2,t}, \ldots, x_{d,t})'$ each $d \times d$

autocovariance matrix has $O(d^2)$ unknown parameters. Even under the covariance stationarity assumption estimating all these covariance parameters simultaneously is a challenging statistical problem. The large number of covariance parameters can be reduced considerably by assuming that the d-dimensional observed process $\boldsymbol{X}_t$ is a time-invariant linear transformation of q decorrelated latent (unobserved) subseries $\mathbf{s}_t^{(i)}$, $i = 1, 2, \ldots, q$, of dimensions d_i, $\sum_{i=1}^{q} d_i = d$. This amounts to assuming that there exist an invertible matrix $\mathbf{M}$ and a latent time series $\mathbf{s}_t$ such that

$$\boldsymbol{X}_t = \mathbf{M}\mathbf{s}_t, \;\; \mathbf{s}_t = (\mathbf{s}_t^{(1)}, \ldots, \mathbf{s}_t^{(q)})' \text{ with } \operatorname{cov}(\mathbf{s}_t^{(i)}, \mathbf{s}_s^{(j)}) = 0, \;\; i \neq j, \tag{2.1}$$

and the $\mathbf{s}_t^{(i)}$'s are referred to as the *decorrelated subseries* of $\boldsymbol{X}_t$. In modeling volatility in finance, one may require that certain transformations of $\mathbf{s}_t^{(i)}$'s are decorrelated, i.e.

$$\operatorname{cov}\left(\mathbf{h}(\mathbf{s}_t^{(i)}), \mathbf{h}(\mathbf{s}_s^{(j)})\right) = 0, \;\; i \neq j, \tag{2.2}$$

where $\mathbf{h}(\cdot)$ is a function acting componentwise on its vector argument. Popular examples of $\mathbf{h}(\cdot)$ are the identity, square and Huber functions, see (2.4).

An important advantage of (2.1)-(2.2) is that regardless of the size of the dimension d, modeling a high-dimensional series is reduced to the simpler task of modeling q disjoint (lower-dimensional) subseries. The vector of lower-dimensional models (forecasts) will then be combined and transformed back to a parsimonious model (forecast) for the original high-dimensional series $\boldsymbol{X}_t$. We note that whereas the classical PCA always ensures existence of an orthogonal matrix $\mathbf{M}$ and the principal components (PCs) for variables with finite second moments, existence of $\mathbf{M}$ and decorrelated subseries $\mathbf{s}_t^{(i)}$'s in (2.1) cannot be guaranteed due to the additional and stringent requirement of decorrelation of $\mathbf{s}_t^{(i)}$'s over time. In the recent literature, there are two important special cases of (2.1) depending on

whether all the latent subseries $\mathbf{s}_t^{(i)}$'s are required to be univariate or not.

First, when all d_i's are equal to one, then (2.1) reduces to the framework of dynamic orthogonal components (DOC) in mean in [41] which is still more general than the classical PCA in that an orthogonal matrix $\mathbf{M}$ is found so that the cross-covariances between any two pairs of univariate DOCs is zero. In this case, we denote a univariate DOC by $s_{i,t}$ to distinguish it from a low-dimensional subseries $\mathbf{s}_t^{(i)}$. As noted earlier existence of DOCs is not ensured, however, when they exist as soon as the mixing matrix $\mathbf{M}$ is found, they are computed using $\mathbf{s}_t = \mathbf{M}^{-1}\boldsymbol{X}_t$, and univariate ARMA or volatility models like the GARCH(1,1) are fitted to each DOC $s_{i,t}, i = 1, \ldots, d$, separately. Even though existence of $\mathbf{M}$ and univariate DOCs $s_{i,t}$ cannot be guaranteed, still for practical reasons one may choose an $\mathbf{M}$ so that the DOCs $s_{i,t}$'s are as close to being decorrelated as possible.

The orthogonal matrix $\mathbf{M}$ in DOC analysis [41, Section 2.3] is parameterized in terms of the Givens angles and its estimation involves optimization of a nonconvex objective function defined as the sum of squares of the off-diagonal entries of the first few lagged autocovariance matrices of the transformed data. It is computationally expensive for dimensions as low as five and hence not suitable for high-dimensional time series which are often encountered in business and economics. This computational challenge might be reduced considerably by relaxing the requirement that all the subseries be univariate. In fact, when some of the d_i's are bigger than one, the setup in (2.1) reduces to the time series PCA (TS-PCA) method in [17] which has the goal of segmenting a multivariate stationary time series into several (lower-dimensional) *decorrelated subseries*. Unlike the DOC method which finds $\mathbf{M}$ by solving a nonconvex optimization problem, the TS-PCA method relies on eigenanalysis of a positive-definite matrix defined as a quadratic function of the first few autocorrelation matrices, see (2.11). It is a natural extension of the standard PCA and the DOC for multiple time series in that as in PCA and DOC it finds an orthogonal matrix transforming a multivariate series into several decorrelated subseries, but unlike PCA and

DOC some of the subseries could have dimensions greater than one.

The vertical partitioning or the TS-PCA method seems ideal for managing the computational bottle-neck encountered in modeling high-dimensional time series. Dividing up a large computational problem into several smaller problems opens up the possibility of parallel computing. In particular, solving the nonlinear optimization problem in DOC [41, Section 2.6] can be reduced to solving many subproblems of much lower dimensions. Moreover, after applying TS-PCA to a high-dimensional time series the mere existence of low-dimensional (non-singleton) subseries is an indication that the time series is not in DOC, while its leading to all one-dimensional subseries should be taken as the indication that the time series is already in DOC. In the former case, the low-dimensional subseries can be partitioned further using the DOC method by taking advantage of the substantial dimension reduction.

In the following sections, we provide a brief review of DOC and TS-PCA methods and discuss the connections between them. Recall that their goals are similar, but they use different objective functions and optimization methods. They transform a multivariate stationary time series into decorrelated univariate series and decorrelated (low-dimensional) subseries, respectively. The objective function of DOC is statistically meaningful and non-convex in $\mathbf{M}$ while that of TS-PCA is less so but quadratic in $\mathbf{M}$.

Let $\mathbf{Y}_t$ be a multivariate stationary time series and $\mathcal{F}_t$ be the information in its past history up to and including the current time t. The series can be decomposed as $\mathbf{Y}_t = \mu_t + e_t$, where $\mu_t = E(\mathbf{Y}_t|\mathcal{F}_{t-1})$ is the conditional mean and e_t is the serially uncorrelated noise. Let $\Sigma_t = \text{cov}(\mathbf{Y}_t|\mathcal{F}_{t-1})$ be the conditional covariance matrix of $\mathbf{Y}_t$ and $\Sigma_y = \text{cov}(\mathbf{Y}_t)$ be its unconditional (marginal) covariance matrix. The time-varying conditional covariance matrix Σ_t is also referred to as the volatility matrices of the returns of financial assets. Developing simple and interpretable dynamic models for μ_t and Σ_t is a key goal of multivariate time series analysis.

2.2 DOC for stationary time series

In [41] a DOC in mean for μ_t and a separate DOC in volatility for Σ_t are introduced. Here we work with $\boldsymbol{X}_t = \mathbf{Y}_t - \mu_t$ or take $\mu_t = 0$ and focus on DOC in volatility for Σ_t, unless stated otherwise.

The goal of DOC in volatility is to find a nonsingular mixing matrix $\mathbf{M}$ such that the latent time series $\{\mathbf{s}_t\}_{t=1}^n$ enjoys quadratic orthogonality in the sense that the autocovariance matrices

$$\mathbf{\Gamma}(\ell) = \text{cov}(\mathbf{s}_t^2, \mathbf{s}_{t-l}^2), \ell = 0, \ \pm 1, \pm 2, \cdots \tag{2.3}$$

are diagonal and $\text{cov}(s_{i,t}s_{j,t}, s_{i,t-\ell}s_{j,t-\ell}) = 0$ for $i \neq j$ and for all lags $\ell = 1, 2, \cdots$. In developing the DOC methodology, it is convenient to have the components of $\boldsymbol{X}_t$ uncorrelated at each time point. A way to do this is by setting $\mathbf{z}_t = \mathbf{U}\boldsymbol{X}_t$, where $\mathbf{U} = \mathbf{\Lambda}^{-1/2}\mathbf{P}'$ and $\mathbf{\Lambda}, \mathbf{P}$ are the diagonal matrix of eigenvalues and the orthogonal matrix of the eigenvectors of the sample covariance matrix of the data. Then, using (2.1) the latent vector of DOCs has the representation

$$\mathbf{s}_t = \mathbf{M}^{-1}\boldsymbol{X}_t = \mathbf{M}^{-1}\mathbf{U}^{-1}\mathbf{z}_t = \mathbf{W}\mathbf{z}_t, \ \text{ where } \ \mathbf{W} = (\mathbf{U}\mathbf{M})^{-1},$$

which implies that $\text{cov}(\mathbf{s}_t) = \mathbf{W}\ \text{cov}(\mathbf{z}_t)\mathbf{W}'$. Since cov $(\mathbf{z}_t) = \mathbf{I}$, it follows that the separating matrix $\mathbf{W}$ is an orthogonal matrix provided that cov $(\mathbf{s}_t) = \mathbf{I}$, which we assume from here on. When the determinant of $\mathbf{W}$ is equal to one, its $d_0 := d(d-1)/2$ free entries can be reparameterized using the Givens angles [41, Section 2.3], conveniently collected in a vector $\boldsymbol{\theta}$ of dimension p where its components take value in $(-\pi, \pi]$. From here on, $\mathbf{W}$ is denoted by $\mathbf{W}_{\boldsymbol{\theta}}$ and $\mathbf{s}_t$ by $\mathbf{s}_t(\boldsymbol{\theta}) = \mathbf{W}_{\boldsymbol{\theta}}\mathbf{z}_t$, to emphasize their dependence on the vector $\boldsymbol{\theta}$, see [41] for more details on the structure of $\mathbf{W}_{\boldsymbol{\theta}}$.

Given the time series data $\boldsymbol{X}_t, t = 1, \ldots, n$ or the zero-mean, uncorrelated $\mathbf{z}_t, t = 1, \ldots, n$ and a vector function $\mathbf{h}(\cdot)$, the sample cross-covariance function of the componentwise transformed latent process is defined by

$$\widehat{\mathbf{\Gamma}}^{\mathbf{h}(\mathbf{s}(\boldsymbol{\theta}))}(\ell) = \widehat{E}\{\mathbf{h}(\mathbf{s}_t(\boldsymbol{\theta}))\mathbf{h}(\mathbf{s}_{t-\ell}(\boldsymbol{\theta}))'\} - \widehat{E}\{\mathbf{h}(\mathbf{s}_t(\boldsymbol{\theta}))\}\widehat{E}\{\mathbf{h}(\mathbf{s}_{t-\ell}(\boldsymbol{\theta}))\}', \quad \ell = 0, 1, 2 \ldots,$$

where $\widehat{E}(\cdot)$ is the sample expectation operator. Ideally, one should choose $\boldsymbol{\theta}$ so that these covariance matrices are as close to being diagonal as possible at all lags, but here for a given positive integer m_1 we restrict attention to a prespecified set of lags $\ell \in \overline{\mathbb{N}}_0 := \{0 \leq \ell \leq m_1\}$ which always includes the lag 0.

An objective function for estimating $\mathbf{W}_{\boldsymbol{\theta}}$ or $\boldsymbol{\theta}$ in the modeling time-varying volatility would naturally rely on the dynamic structure or cross-correlations which amounts to taking $\mathbf{h}(\mathbf{s}) = \mathbf{s}^2$, namely the square of the entries of the latent vector. However, since asset returns usually exhibit heavy tails applying the following Huber's function

$$h_c(s) = \begin{cases} s^2 & \text{if } |s| \leq c, \\ 2|s|c - c^2 & \text{if } |s| > c, \end{cases} \tag{2.4}$$

to each $s_{i,t}$ would make the procedure more robust to outliers [41]. We use $c = 2.25$ in our computations in Section 4.

To define the objective function, we vectorize and arrange all the off-diagonal elements of $\{\widehat{\mathbf{\Gamma}}^{\mathbf{h}(\mathbf{s}(\boldsymbol{\theta}))}(\ell) | \ell \in \overline{\mathbb{N}}_0\}$ in the vector $\bar{\mathbf{f}}_n(\boldsymbol{\theta}) = \widehat{E}\{\mathbf{f}(\mathbf{z}_t, \boldsymbol{\theta})\}$ where $\mathbf{f}(\mathbf{z}_t, \boldsymbol{\theta})$ is a vector with entries

$$f_{ij}^{\ell}(\mathbf{z}_t, \boldsymbol{\theta}) = \mathbf{h}_i(\mathbf{s}_t(\boldsymbol{\theta}))\mathbf{h}_j(\mathbf{s}_{t-\ell}(\boldsymbol{\theta})) - \widehat{E}\{\mathbf{h}_i(\mathbf{s}_t(\boldsymbol{\theta}))\}\widehat{E}\{\mathbf{h}_j(\mathbf{s}_{t-\ell}(\boldsymbol{\theta}))\}, \tag{2.5}$$

indexed by $i < j$ for $\ell = 0$, and by $i \neq j$ for $\ell > 0$. Since the lagged cross-dependence is typically strongest at lower lags, we use the following larger weights for the lower-lag cross-covariance matrices:

$$\phi_\ell = \frac{1 - \ell/|\overline{\mathbb{N}}_0|}{\Sigma_\ell(1 - \ell/|\overline{\mathbb{N}}_0|)}/(d_0 + d_0\mathbf{I}_{\{\ell\neq 0\}}) \text{ for } \ell \in \overline{\mathbb{N}}_0,$$

where $|\overline{\mathbb{N}}_0|$ is the cardinality of the set $\overline{\mathbb{N}}_0$ and $\mathbf{I}_{\{\cdot\}}$ denotes an indicator function. Arranging these weights into the following diagonal matrix

$$\mathbf{\Phi} = \text{diag}\{\phi_{\ell_1}, \cdots, \phi_{\ell_1}, \phi_{\ell_2}, \cdots, \phi_{\ell_2}, \cdots, \phi_{\ell_{|\bar{\mathbf{N}}_0|}}, \cdots, \phi_{\ell_{|\bar{\mathbf{N}}_0|}}\},$$

the objective function is then defined as a quadratic form in the off-diagonal entries of the cross-covariance matrices:

$$\mathcal{J}_n(\boldsymbol{\theta}) = \bar{\mathbf{f}}_n(\boldsymbol{\theta})'\mathbf{\Phi}\bar{\mathbf{f}}_n(\boldsymbol{\theta}). \tag{2.6}$$

An estimator of $\widehat{\boldsymbol{\theta}}_n$ is defined as its minimizer: $\widehat{\boldsymbol{\theta}}_n = \text{argmin}_{\boldsymbol{\theta}}\mathcal{J}_n(\boldsymbol{\theta})$. Finally, the separating matrix is estimated as $\mathbf{W}_{\widehat{\boldsymbol{\theta}}_n}$ and the estimated DOC series is given by $\widehat{\mathbf{s}}_t = \mathbf{W}_{\widehat{\boldsymbol{\theta}}_n}\mathbf{z}_t$.

There are three sources of nonuniqueness in estimating $\mathbf{M}$ and $\mathbf{s}_t$, related to the scale, sign and the order of the DOCs. These stem from the matrix product on the right hand side of (2.1) where $\mathbf{M}\mathbf{s}_t = \mathbf{M}\mathbf{H}\mathbf{H}^{-1}\mathbf{s}_t$, for any invertible matrix $\mathbf{H}$. The scale of DOCs can be fixed by assuming $\text{cov}(\mathbf{s}_t) = I$, then taking $\mathbf{H}$ to be a signed permutation matrix allows identification of the DOCs up to a signed permutation which is sufficient for forecasting purposes for several situations discussed in [41, Section 2.4]. In addition, since the objective function is nonconvex its numerical optimization requires special attention to avoid getting stuck at the local minima. A way to address this issue is to work with several initial values in the high-dimensional parameter space as in [47].

Existence of DOCs implies that the off-diagonal elements of $\mathbf{\Gamma}^{\mathbf{h}(\mathbf{s}(\boldsymbol{\theta}))}(\ell)$ are zero for $\ell \geq 0$, so that one may develop a Ljung-Box type test for their existence by testing the hypothesis that all these off-diagonal elements are zero. Let $h_{i,t-\ell} = \mathbf{h}_i(\mathbf{s}_{t-\ell})$ and $\rho^h_{i,j}(\ell) = \text{corr}\{h_{i,t}, h_{j,t-\ell}\}$ where $\mathbf{h}(\cdot)$ is the square function for DOC in volatility. The null and alternative hypothesis to test for the existence of DOCs are,

$$H_0 \quad : \rho^h_{i,j}(\ell) = 0, \ for\ all\ i \neq j, \ell = 0, 1, 2, \cdots, m; \tag{2.7}$$

$$H_a \quad : \rho^h_{i,j}(\ell) \neq 0, \ for\ some\ i \neq j, \ell = 0, 1, 2, \cdots, m. \tag{2.8}$$

The test statistic used is

$$Q^0_d(m) = n\Sigma_{i<j}\rho^h_{i,j}(0)^2 + n(n+2)\Sigma^m_{k=1}\Sigma_{i\neq j}\rho^h_{i,j}(k)^2/(n-k), \tag{2.9}$$

which under H_0, is asymptotically distributed as a χ^2 distribution with $d(d-1)/2 + md(d-1)$ degrees of freedom [39]. The null hypothesis is rejected for larger values of the test statistic.

2.3 TS-PCA for high-dimensional time series

Estimation of the mixing matrix in DOC involves nonlinear optimization, is computationally expensive and hence not suited for high-dimensional data [41]. In this section, we review the computationally attractive TS-PCA method [17] involving eigenanalysis of a suitable positive-definite matrix in the spirit of classical PCA. We discuss its potential connection with and role in reducing the computational burden encountered in DOC analysis.

To describe the TS-PCA methodology, it is convenient to assume that the two time

series in (2.1) are standardized, namely

$$\text{var}(\boldsymbol{X}_t) = \mathbf{I}_d \quad \text{and} \quad \text{var}(\mathbf{s}_t) = \mathbf{I}_d. \tag{2.10}$$

For a pre-selected positive integer k_0, consider the positive-definite matrix

$$\mathbf{W}_x = \sum_{k=0}^{k_0} \mathbf{\Gamma}_x(k)\mathbf{\Gamma}_x(k)' = \mathbf{I}_d + \sum_{k=1}^{k_1} \mathbf{\Gamma}_x(k)\mathbf{\Gamma}_x(k)', \tag{2.11}$$

where $\mathbf{\Gamma}_x(k) = \text{corr}(\boldsymbol{X}_{t+k}, \boldsymbol{X}_t)$ is the cross-correlation matrix of the standardized time series. In contrast to using nonconvex optimization in DOC, TS-PCA finds the mixing matrix $\mathbf{M}$ in (2.1) by relying on the simpler task of eigenanalysis of the matrix $\mathbf{W}_x$. Let $\mathbf{\Gamma}_x$ be the $d \times d$ orthogonal matrix of the eigenvectors of $\mathbf{W}_x$. Then, the matrix $\mathbf{M}$ is identified as a column-permutation of an estimator of $\mathbf{\Gamma}_x$. The permutation is designed to group the transformed series $\hat{\mathbf{s}}_t = \hat{\mathbf{\Gamma}}_x' \boldsymbol{X}_t$ into a number of decorrelated subseries of lower dimensions so that the within-subseries correlations are significant while those of the between-subseries are not. The following is the two-step TS-PCA procedure in [17, Section 2.2]:

1. Find a consistent estimator of $\mathbf{W}_x$, and let $\hat{\mathbf{\Gamma}}_x$ be the orthogonal matrix obtained from its eigenanalysis (spectral decomposition).

2. Obtain the matrix $\hat{\mathbf{M}} = (\hat{\mathbf{M}}_1, \ldots, \hat{\mathbf{M}}_q)$ by permutating the columns of $\hat{\mathbf{\Gamma}}_x$ so that $\hat{\mathbf{s}}_t = \hat{\mathbf{\Gamma}}_x' \boldsymbol{X}_t$ is segmented into q decorrelated subseries as in (2.1).

As for a consistent estimator of $\mathbf{W}_x$, it is known that for large dimensions d, the sample autocovariance matrix $\widehat{\mathbf{\Gamma}}_x(k) = \frac{1}{n}\sum_{t=1}^{n-k}(\boldsymbol{X}_{t+k} - \bar{\boldsymbol{X}})(\boldsymbol{X}_t - \bar{\boldsymbol{X}})'$ with $\bar{\boldsymbol{X}} = \frac{1}{n}\sum_{t=1}^{n} \boldsymbol{X}_t$, is not a consistent estimator for $\mathbf{\Gamma}_x(k)$. Consider a regularized estimator such as the threshold

estimator

$$T_u(\widehat{\mathbf{\Gamma}}_x(k)) = (\hat{\Gamma}_{i,j}^{(k)} I\{|\hat{\Gamma}_{i,j}^{(k)}| \geq u\})_{i,j=1,2,\cdots,d},$$

where $u = \lambda(\log d/n)^{1/2}$ is the threshold level and $\lambda > 0$ is a tuning parameter, $I(\cdot)$ is the indicator function and $\hat{\Gamma}_{i,j}^{(k)}$ represents the (i,j)-th entry of $\widehat{\mathbf{\Gamma}}_x(k)$ [12]. Then, the threshold estimator defined by

$$\widehat{\mathbf{W}}_x^{(thres)} = \mathbf{I}_d + \sum_{k=1}^{k_0} T_u(\widehat{\mathbf{\Gamma}}_x(k)) T_u(\widehat{\mathbf{\Gamma}}_x(k))',$$

is known to be consistent for $\mathbf{W}_x$ [17, Lemma 8] for a suitable choice of the tuning parameter (m, k_0, λ), where m is the number of lags in the multiple null hypothesis in (2.12).

The consistent estimator above provides an estimator for $\mathbf{M}$ up to a regrouping of its columns. Intuitively, the permutation in Step 2 is found by visually examining the cross-correlograms of pairs of components of $\hat{\mathbf{z}}_t = \hat{\mathbf{\Gamma}}_x' \boldsymbol{X}_t$, and putting in the same group those components which have significant cross-correlations at all lags. This amounts to obtaining $\mathbf{M}$ by rearranging the columns of $\hat{\mathbf{\Gamma}}_x$ according to the grouping suggested by the cross-correlograms. Though the idea of visual inspection of pairwise cross-correlations is not practical for high-dimensional time series, its core insight is used to develop automatic permutation rules based on certain functionals of the cross-correlations. More precisely, with $\rho(k)$ as the lag-k cross-correlation between two component series of $\hat{\mathbf{z}}_t$, we say these two components are *connected* if the multiple null hypothesis

$$H_0 \;:\; \rho(k) = 0, \;\; \text{for any} \;\; k = 0, \pm 1, \ldots, \pm m, \tag{2.12}$$

is rejected. Evidently, connected components with significant cross-correlation should belong to the same group. Thus, the permutation in Step 2 starts with d groups of singletons,

then two groups are combined if connected pairs in $\hat{\mathbf{z}}_t$ are split over two groups, and the process is continued until all connected pairs are united in one group.

A method for identifying the connected pairs using cross-correlations $\hat{\rho}_{i,j}(h)$ of the series $\hat{z}_t$, for any pair $(1 \leq i < j \leq d)$, is based on their maximum,

$$\widehat{L}_n(i,j) = \max_{|l|\leq m} |\hat{\rho}_{i,j}(l)|. \tag{2.13}$$

The null hypothesis of significant cross-correlations would be rejected for the (i,j) pair if this statistic is greater than a specified threshold. To avoid multiple tests for $d_0 = d(d-1)/2$ pairs, a ratio statistic is used to single out those pairs for which H_0 will be rejected. It is based on the rearrangement in descending order: $\hat{L}_1 \geq \hat{L}_2 \geq ... \geq \hat{L}_{d_0}$ and defining $\hat{r}$ as

$$\hat{r} = \arg \max_{1\leq j\leq c_0.p} \frac{\hat{L}_j}{\hat{L}_{j+1}}, \tag{2.14}$$

where $c_0 \in (0,1)$ is used to guard against dividing by 0. Once $\hat{r}$ is determined, the pairs corresponding to the first $\hat{r}$ maximum cross-correlations are declared connected or significantly cross-correlated and groups are formed based on these pairs.

An extension of TS-PCA to segment multivariate volatility processes in [17, Section 5] amounts to applying the above procedure to the target matrix:

$$\mathbf{W}_x = \sum_{B\in\mathcal{B}_{t-1}} \left[\mathrm{E}\{\boldsymbol{X}_t \boldsymbol{X}_t' \mathrm{I}(B)\}\right]^2, \tag{2.15}$$

where $\mathcal{B}_{t-1}$ is a π-class and the σ-algebra it generate is $\mathcal{F}_{t-1} = \sigma(\boldsymbol{X}_{t-1}, \boldsymbol{X}_{t-2}, \cdots)$. The target matrix $\mathbf{W}_x$ is estimated by

$$\widehat{\mathbf{W}}_x = \sum_{\ell=1}^{n} \sum_{k=1}^{k_0} \{\frac{1}{n-k} \sum_{t=k+1}^{n} \boldsymbol{X}_t \boldsymbol{X}_t' \mathrm{I}(\boldsymbol{X}_{t-k}' \boldsymbol{X}_{t-k} \leq \boldsymbol{X}_\ell' \boldsymbol{X}_\ell)\}^2,$$

see [26].

2.4 Connection between DOC and TS-PCA

The connection between DOC and TS-PCA is not evident and has not been studied. The following result sheds some light on the possible connections between their objective functions.

Lemma 1. *Let* $\{\boldsymbol{X}_t\}$ *be a mean-zero stationary process satisfying* $\boldsymbol{X}_t = \mathbf{M}\mathbf{s}_t$. *Then,*

(a)

$$cov(\mathbf{s}_t, \mathbf{s}_{t-\ell}) = \mathbf{M}\boldsymbol{\Gamma}_x(\ell)\mathbf{M}'. \tag{2.16}$$

(b) provided that $\{\boldsymbol{X}_t\}$ *is Gaussian,we have*

$$cov(\mathbf{s}_t^2, \mathbf{s}_{t-\ell}^2) = 2(\mathbf{M}\boldsymbol{\Gamma}_x(\ell)\mathbf{M}') \circ (\mathbf{M}\boldsymbol{\Gamma}_x(\ell)\mathbf{M}'), \tag{2.17}$$

where $\circ$ *denotes the Hadamard product two matrices.*

The equivalence between DOC in mean and TS-PCA is immediate from the identity in (A.1). However, if one could replace the Hadamard product in the identity (A.2) by the usual matrix product, then the equivalence between DOC in volatility and TS-PCA for volatility processes would be immediate. Unfortunately, this does not seem to be possible mostly because of the following observation: A key difference between DOC in volatility and TS-PCA in volatility lies in their assumptions on the latent process $\mathbf{s}_t$. Recall that DOC assumes that $\text{cov}(\mathbf{s}_t^2, \mathbf{s}_{t-\ell}^2)$ is diagonal for $\ell = 0, \pm 1, \pm 2, \ldots$, and $\text{cov}(s_{ti}s_{tj}, s_{t-\ell,i}s_{t-\ell,j}) = 0$ for $i \neq j$ and $\ell = 1, 2, \ldots$; while TS-PCA requires the conditional covariance $\text{cov}(\mathbf{s}_t|\mathcal{F}_{t-1})$ to be block diagonal. Without any additional assumption, these two sets of conditions do not nest each other. When $\mathbf{s}_t$ follows the multivariate GARCH model, as pointed out in Matteson and Tsay (2011), the assumption for DOC in volatility implies that $\text{cov}(\mathbf{s}_t|\mathcal{F}_{t-1})$ is diagonal.

Nevertheless, we provide further support on their close connections through a simulation study reported in the Appendix. Furthermore, the recent papers by [33] on principal component volatility (PCV) and the paper by [38] are closely related to this topic.

3. DOC FOR NONSTATIONARY TIME SERIES

3.1 Introduction

The key assumptions in both TS-PCA and DOC methods are stationarity of the data and existence of time-invariance of the linear transformations. For many real life examples, however, the stationarity feature is often violated for reasonably long time series. In such situations, it is natural and common to work with locally homogeneous or piecewise stationary processes ([1]). For example, [19] presented a class of nonstationary time series models with evolutionary spectral representation which can be approximated arbitrarily closely by AR models with time-varying coefficients. [20] considered piecewise AR processes, and then piecewise GARCH and stochastic volatility models in [21]. [37] studied structural breaks in spectral distribution of piece-wise stationary time series, and [5] considered change point detections in covariance matrices of nonstationary time series.

We extend the TS-PCA and DOC methods to the nonstationary setup where the series is composed of several locally homogeneous segments due to changes in its volatility or other features. The key challenge is the identification of the change points or finding the locally homogeneous intervals. A parametric approach to the problem will assume a subjective global model for the series which may lack the flexibility to deal with sudden changes, see [10]. A local parametric approach due to [43], which assumes that the volatility process is approximately constant locally but time-varying over longer stretch of time, is ideal for the problem at hand. It allows developing a data-based method to select the change-points, the ensuing intervals of homogeniety ([18]) and estimation of the mixing matrices. Our proposed time-varying TS-PCA and DOC methods combines the merits of the decorrelating methods restricted by stationarity assumption and the change point detection method in [18] such that the TS-PCA or DOC could be applied to analyze the

partitioned stationary segments of the data separately.

3.2 TVDOC for piece-wise stationary data

Stationarity assumption central to DOC analysis [41] may not be tenable in practice especially for asset returns where there could be drastic structural changes over time. Thus, it is desirable to study and extend the DOC methodology to nonstationary data where the mixing matrix is time-varying.

In this section, we propose a time-varying DOC (TVDOC) methodology for multivariate piecewise stationary data generalizing the stationary DOC. It relies on a change point detection method, reviewed in Section 3.2, to divide the whole series into a number of locally homogeneous segments, and then stationary DOC technique is applied to each segment separately.

For simplicity, we consider the case where the original series $\boldsymbol{X}_t$ is standardized with $E(\boldsymbol{X}_t) = 0$ and cov($\boldsymbol{X}_t$) = $\mathbf{I}$, has only one change point at the known time point $k_0(n)$. It is straightforward to generalize the results to the multiple change points situation. Let us denote the two stationary segments by $\boldsymbol{X}_t^{(1)} = \boldsymbol{X}_t$ if $t \leq k_0(n)$ and $\boldsymbol{X}_t^{(2)} = \boldsymbol{X}_t$ if $k_0(n) + 1 \leq t \leq n$, and their angle parameters for each segment by $\boldsymbol{\theta}_i$ and $\mathbf{s}_t(\boldsymbol{\theta}_i)$, respectively. Recall that $\mathbf{s}_t(\boldsymbol{\theta}_i) = \mathbf{W}_{\boldsymbol{\theta}_i}\boldsymbol{X}_t^{(i)}$ for $i = 1, 2$, and that in TVDOC we estimate $\boldsymbol{\theta}_i$ by minimizing the two separate objective functions $\mathcal{J}^{(i)}(\boldsymbol{\theta_i}) = \bar{f}^{(i)}(\boldsymbol{\theta_i})'\Phi_n\bar{f}^{(i)}(\boldsymbol{\theta_i}), i = 1, 2$, where as before, $\bar{f}^{(i)}(\boldsymbol{\theta_i}) = \hat{E}f(\boldsymbol{X}_t, \boldsymbol{\theta}_i)$ and $f(\boldsymbol{X}_t, \boldsymbol{\theta}_i)$ is the vector that stacks up all the off-diagonal elements in the lagged autocovariance matrices $\widehat{\text{cov}}(\mathbf{s}_t^2(\boldsymbol{\theta}_i), \mathbf{s}_{t-\ell}^2(\boldsymbol{\theta}_i))$. Set $\boldsymbol{\theta} = (\boldsymbol{\theta}_1', \boldsymbol{\theta}_2')'$, $\bar{g}(\boldsymbol{\theta}) = (\bar{f}^{(1)}(\boldsymbol{\theta}_1)', \bar{f}^{(2)}(\boldsymbol{\theta}_2)')'$, and let

$$\boldsymbol{\Theta} = \{\theta_i^j, 2 \leq j \leq d, 1 \leq i \leq j-1, \text{ where } \theta_i^j \in (-\pi, \pi]\},$$

be the parameter space and $\bar{\boldsymbol{\Theta}}$ be a sufficiently large compact subset of $\boldsymbol{\Theta}$.

3.3 Asymptotic properties of TVDOC

In this section, we establish strong consistency of the estimator $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\theta}}_1', \hat{\boldsymbol{\theta}}_2')'$ in the TVDOC setup under the Conditions C1-C4 which are similar to those in [41]. Then, the joint asymptotic normality for $(\hat{\boldsymbol{\theta}}_1', \hat{\boldsymbol{\theta}}_2')'$ is proved using the concept of near-epoch-dependence for triangular array of random variables. In what follows, $\boldsymbol{\theta}_0 = (\boldsymbol{\theta}_{01}', \boldsymbol{\theta}_{02}')'$ stands for the true value of the parameter. With suitable modifications, our theoretical results are still valid when the Huber's function is used.

C1. There exists unique minimizer $\boldsymbol{\theta}_{0i} \in \bar{\boldsymbol{\Theta}}$ for $\mathcal{J}^{(i)}(\boldsymbol{\theta})$, $i = 1, 2$.

C2. The process $\boldsymbol{X}_t^{(i)}$ is stationary and ergodic with $E||\boldsymbol{X}_t^{(i)}||^2 < \infty$ for $i = 1, 2$.

C3. $\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \mathrm{E}||\mathbf{s}_t(\boldsymbol{\theta})^4||^2 < \infty$ and $\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \mathrm{E}||\frac{\partial \mathbf{s}_t(\boldsymbol{\theta})^2}{\partial \boldsymbol{\theta}_i}\boldsymbol{X}_t^{(i)}||^2 < \infty$.

C4. $\mathbf{W}_{\boldsymbol{\theta}_{0i}}$ has a unique continuous inverse.

Theorem 1. *(Strong Consistency) Under Conditions* $C1 - C4$, $\hat{\boldsymbol{\theta}} \xrightarrow{a.s.} \boldsymbol{\theta}_0$ *as* $n \to \infty$.

To establish the joint asymptotic normality for $(\hat{\boldsymbol{\theta}}_1', \hat{\boldsymbol{\theta}}_2')'$, we introduce the concept of near-epoch-dependence for triangular array of random variables, which is one of the most general concepts of weak temporal dependence for nonlinear models. Its origin can be traced to as far back as [35] and it has been widely used in the econometrics literature, see e.g. [52], [4], and [22] among others.

Definition A triangular array of random variables $\{\boldsymbol{X}_{n,t}\}$ is L_2-near-epoch-dependent (L_2-NED) on a triangular array of random variables $\{U_{n,t}\}$ if for $k \geq 0$,

$$\sup_n \sup_t ||\boldsymbol{X}_{n,t} - E[\boldsymbol{X}_{n,t}|U_{n,t-k}, \cdots, U_{n,t+k}]|| \leq v(k),$$

and $v(k) \to 0$ as $k \to \infty$.

Definition A sequence δ_k is of size $-\lambda$ if $\delta_k = O(k^{-\lambda-\epsilon})$ for some $\epsilon > 0$.

The following assumptions are imposed to facilitate our theoretical derivations.

C5. $\lim_{n\to+\infty} \frac{k_0(n)}{n-k_0(n)} = c$, for a positive constant c.

C6. Let $\mathbf{s}_{n,t} = \mathbf{s}_t(\boldsymbol{\theta}_{01})$ if $t \leq k_0(n)$ and $\mathbf{s}_{n,t} = \mathbf{s}_t(\boldsymbol{\theta}_{02})$ if $t > k_0(n)$. Then, for some $r > 2$, $\{\mathbf{s}_{n,t}\}$ is a triangular array of mean zero random vectors that is L_2-NED of size -1 on an α-mixing base $\{U_{n,t}\}$ of size $-r/(r-2)$ and $\sup_n \sup_t E||\mathbf{s}_{n,t}||^{4r} < \infty$.

C7. Let $\bar{g}(\boldsymbol{\theta}_0) = (\bar{f}^{(1)}(\boldsymbol{\theta}_{01})', \bar{f}^{(2)}(\boldsymbol{\theta}_{02})')'$. $\lim_{n\to\infty} \text{var}(\sqrt{n}\bar{g}(\boldsymbol{\theta}_0)) = \mathbf{V}_0 = \text{diag}(\mathbf{V}_{1,1}, \mathbf{V}_{2,2})$ for some positive definite matrix $\mathbf{V}_0$.

C8. There exists a weakly consistent estimator $\hat{\mathbf{V}}_n := \text{diag}(\hat{\mathbf{V}}_{1,1}, \hat{\mathbf{V}}_{2,2})$ for $\mathbf{V}_0$, namely $\hat{\mathbf{V}}_n - \mathbf{V}_0 \xrightarrow{\mathcal{P}} 0$ as $n \to \infty$.

For $i = 1, 2$, note that $\bar{f}^{(i)}(\boldsymbol{\theta}_i)$ is continuously differentiable with respect to $\boldsymbol{\theta}_i$ on $\boldsymbol{\Theta}$. We denote its matrix gradient by $\bar{F}^{(i)}(\theta_i)$ and define the matrices

$$G_i = \bar{F}^{(i)}(\hat{\theta}_i)'\Phi_n\hat{\mathbf{V}}_{i,i}\Phi_n\bar{F}^{(i)}(\hat{\theta}_i), \quad H_i = \bar{F}^{(i)}(\hat{\theta}_i)'\Phi_n\bar{F}^{(i)}(\hat{\theta}_i),$$

and

$$A_n = \text{diag}(G_1^{-1/2}, G_2^{-1/2})\text{diag}(H_1, H_2).$$

Theorem 2. *(Asymptotic Normality) Under Conditions $C1 - C8$, as $n \to \infty$,*

$$A_n \times \sqrt{n} \begin{pmatrix} (\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1) \\ (\hat{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}_2) \end{pmatrix} \xrightarrow{D} N\left(\mathbf{0}_{2p}, \mathbf{I}_{2p}\right), \text{ where } p := d(d-1)/2.$$

Some remarks on the assumptions are in order: Assumption C5 controls the two segment lengths so that both grow to infinity. Assumption C6 allows for general serial correlation in $\mathbf{s}_{n,t}$ and it accounts for potential heteroscedasticity across different segments of the latent process. Similar conditions have been considered in the change-point detection literature, see e.g. [4] and [8]. Assumption C7 ensures the existence of a positive

definite asymptotic covariance matrix for $\bar{g}(\boldsymbol{\theta}_0)$. As $\{\mathbf{s}_{n,t}\}$ is L_2-NED of size -1, it can be shown that $\bar{f}^{(1)}(\boldsymbol{\theta}_{01})$ and $\bar{f}^{(2)}(\boldsymbol{\theta}_{02})$ are asymptotically uncorrelated, which implies the asymptotic independence between $\hat{\boldsymbol{\theta}}_1$ and $\hat{\boldsymbol{\theta}}_2$. Assumption C8 requires the existence of a consistent covariance estimator for $\mathbf{V}$, which can be constructed based on the classical kernel-window estimation.

We note that, in practice, the change point $k_0(n)$ is unknown and needs to be estimated from the data as described in the next subsection. Let $\hat{k}(n)$ be a consistent estimator such that

$$n^{-1}|\hat{k}(n) - k_0(n)| = o_p(1). \tag{3.1}$$

Then following the arguments in Corollary 1 of [8], we expect that the conclusion in Theorem 4 remains valid if $k_0(n)$ is replaced by $\hat{k}(n)$.

3.4 Change point detection

Identification of change points or segmenting a nonstationary series into locally homogeneous intervals is an important step in the development of TVDOC. There are diverse change point detection methods in the literature. A parametric approach which usually assumes a subjective global model for the series may lack the flexibility to deal with sudden changes [10].

We rely on a local parametric approach due to [43] as implemented by [18, Section 2] which assumes that the volatility process is approximately constant locally but time-varying over longer stretch of time. For our goals here, it leads to an ideal data-based and sequential testing method to detect the change-points. More specifically, for a given t one starts with a set of K candidate intervals of increasing lengths of the form $I_{t,k} = (t-m_k, t]$ with $m_k = m_0 a^k$, $1 \leq k \leq K$, with prespecified m_0 and a multiplier $a > 1$. The shortest interval $I_{t,0}$ is always accepted due its smaller modeling bias relative to others. Next, for a longer interval $I_{t,k}$ with $k = 1, \ldots, K$ which nests the previously accepted inter-

val $I_{t,k-1}$, the focus will be on testing the status of the new time points in the interval $J_{t,k} = [t - m_k, t - m_{k-1}]$ as potential change points. A log-likelihood ratio test in [18] is used to sequentially screen all the new time points in the interval $J_{t,k}$. One accepts the interval $I_{t,k}$ if all the time points are found to be insignificant as potential change point. The procedure is then continued in the next longer interval until either a change point is detected or the longest interval $I_{t,K}$ is reached. Otherwise, the procedure terminates and the last accepted interval is selected.

The choice of m_0 is delicate and it is recommended to be chosen small as compared to the sample size so that smaller candidate intervals are constructed to capture all potential change points. In our experience, we found satisfactory results when m_0 was around $(1/8)^{th}$ of the sample size, and for fixed $a = 1.25$ and $K = 5$ as suggested in [18].

3.5 Simulation and data analysis

In this section, we illustrate the TVDOC method and compare its performance with DOC, PCA and TVPCA using simulated and real datasets with dimensions ranging from $d = 3$ to 135. The latter high-dimensional dataset highlights the important role of TS-PCA as a tool to vertically partition a high-dimensional time series into lower-dimensional decorrelated subseries suitable for further analysis by the DOC method.

3.5.1 A simulation study

Using a simulation experiment we illustrate the TVDOC methodology and assess its performance relative to DOC and other methods when the assumption of stationarity or constant mixing matrix is violated.

We consider the GARCH$(1,1) - t_\nu$ model for each volatility component $s_{it}, i = 1, 2, \cdots, d$, where $t_\nu(0,1)$ denote the standardized Student-t distribution with ν degrees

of freedom. The multivariate volatility model for the original time series of innovations is:

$$\begin{aligned} \mathbf{e}_t &= \mathbf{M}\mathbf{s}_t = \mathbf{M}\mathbf{V}_t^{1/2}\boldsymbol{\epsilon}_t, \\ \mathbf{V}_t &= \text{diag}\{\sigma_{1t}^2, \cdots, \sigma_{dt}^2\}, \qquad \epsilon_{it} \overset{iid}{\sim} t_{v_i}(0,1), \\ \boldsymbol{\Sigma}_t &= \mathbf{M}\mathbf{V}_t\mathbf{M}', \qquad \sigma_{it}^2 = \omega_i + \alpha_i s_{i,t-1}^2 + \beta_i \sigma_{i,t-1}^2, \end{aligned} \tag{3.2}$$

where $\omega_i > 0$, and $\alpha_i, \beta_i \geq 0$ to ensure positiveness of the variances. It is further assumed that $\nu_i > 2$ and $\alpha_i + \beta_i < 1$, to ensure second order stationarity and ergodicity of the process see [29, Theorem 2.5].

In each simulation experiment for series of length $n = 1000, 2000$, the DOCs $\mathbf{s}_t^{(1)}, t = 1, \ldots, n/2$, and $\mathbf{s}_t^{(2)}, t = n/2+1, \ldots, n$, are simulated as in (3.2) with $v_i = 6, \omega_i = 0.01, \alpha_i = 0.09, \beta_i = 0.90$ for $i = 1, 2$. Two fixed $d \times d$ mixing matrices $\mathbf{M}_1$ and $\mathbf{M}_2$ are constructed whose entries are iid draws from a standard normal distribution (these matrices are presented in the Appendix). We set $\mathbf{M}_t = \mathbf{M}_1$ for the first $n/2$ observations and $\mathbf{M}_t = \mathbf{M}_2$ for the rest, denote the first segment of the series as $\boldsymbol{X}_t^{(1)}$ and the second as $\boldsymbol{X}_t^{(2)}$, then $\boldsymbol{X}_t^{(i)} = \mathbf{M}_i \mathbf{s}_t^{(i)}$ for $i = 1, 2$.

We use the following Amari metric [6] to assess the performance or accuracy of an estimator $\widehat{\mathbf{M}}_1$ with the true $\mathbf{M}_1$:

$$A(\mathbf{M}_1, \widehat{\mathbf{M}}_1) = \frac{1}{2d}\sum_{i=1}^{d}\left(\frac{\sum_{j=1}^{d}|\tilde{m}_{ij}|}{\max_j|\tilde{m}_{ij}|} - 1\right) + \frac{1}{2d}\sum_{j=1}^{d}\left(\frac{\sum_{i=1}^{d}|\tilde{m}_{ij}|}{\max_i|\tilde{m}_{ij}|} - 1\right),$$

where $\tilde{m}_{ij} = (\mathbf{M}_1\widehat{\mathbf{M}}_1^{-1})_{ij}$ and d is the dimension of a square matrix. It takes values between 0 and $d-1$, and is equal to zero if and only if $\mathbf{M}_1$ and $\widehat{\mathbf{M}}_1$ represent permutations of the same components. The metric is invariant to permutation and scaling of the matrices, and is thus ideal for comparing various estimated mixing matrices.

We conduct simulation experiments to assess the performance of the following four

methods for estimating the mixing matrix $\mathbf{M}_t$ for dimensions $d = 5, 10$ where the change-point is known to be $k_0 = n/2 + 1$. We note that for TVPCA (time-varying PCA) and TVDOC methods, the mixing matrices $\hat{\mathbf{M}}_1$ and $\hat{\mathbf{M}}_2$ are estimated separately for the two segments $\boldsymbol{X}_t^{(1)}$ and $\boldsymbol{X}_t^{(2)}$, while the PCA and DOC method are applied to the whole series $\boldsymbol{X}_t$ obtaining a single mixing matrix denoted by $\hat{\mathbf{M}}$. The notation TVDOC(k) or DOC(k) in Table 3.1 corresponds to using $\overline{\mathbb{N}}_0 = \{0, 1, \cdots, k\}$ or including the first k lags in the objective function (2.6). Table 3.1 shows the means and standard deviations of the Amari errors based on 10000 runs of the simulation experiments. The Amari error here is the distance between the matrices $\mathbf{M}_1$ and $\mathbf{M}_2$ and their estimated counterparts $\hat{\mathbf{M}}_1, \hat{\mathbf{M}}_2$ for TVPCA and TVDOC. However, for the PCA and DOC the Amari error is computed noting that $A(\mathbf{M}_i, \widehat{\mathbf{M}}_i) = A(\mathbf{M}_i, \widehat{\mathbf{M}}), i = 1, 2$.

It can be seen from Table 3.1 that TVDOC method outperforms DOC/PCA/TVPCA

	d=5				d=10			
	n=1000		n=2000		n=1000		n=2000	
	$A(\mathbf{M}_1, \widetilde{\mathbf{M}}_1)$	$A(\mathbf{M}_2, \widetilde{\mathbf{M}}_2)$	$A(\mathbf{M}_1, \widetilde{\mathbf{M}}_1)$	$A(\mathbf{M}_2, \widetilde{\mathbf{M}}_2)$	$A(\mathbf{M}_1, \widetilde{\mathbf{M}}_1)$	$A(\mathbf{M}_2, \widetilde{\mathbf{M}}_2)$	$A(\mathbf{M}_1, \widetilde{\mathbf{M}}_1)$	$A(\mathbf{M}_2, \widetilde{\mathbf{M}}_2)$
PCA	1.39(0.18)	1.89(0.22)	1.43(0.13)	1.93(0.18)	3.27(0.32)	3.95(0.44)	3.24(0.29)	4.02(0.40)
DOC(1)	1.21(0.16)	1.72(0.16)	1.18(0.22)	1.69(0.15)	3.21(0.27)	3.36(0.45)	3.18(0.26)	3.20(0.36)
DOC(2)	1.27(0.19)	1.72(0.15)	1.19(0.18)	1.74(0.16)	3.21(0.27)	3.38(0.52)	3.25(0.24)	3.21(0.39)
DOC(3)	1.23(0.17)	1.77(0.15)	1.18(0.18)	1.74(0.18)	3.30(0.29)	3.46(0.52)	3.21(0.30)	3.29(0.47)
TVPCA	1.31(0.23)	1.79(0.21)	1.29(0.26)	1.83(0.21)	2.93(0.28)	3.20(0.29)	3.01(0.33)	3.32(0.29)
TVDOC(1)	0.69(0.16)	0.72(0.17)	0.51(0.11)	0.51(0.10)	2.18(0.29)	2.19(0.32)	1.57(0.24)	1.55(0.24)
TVDOC(2)	0.80(0.20)	0.81(0.18)	0.55(0.13)	0.58(0.15)	2.34(0.30)	2.42(0.28)	1.75(0.24)	1.77(0.29)
TVDOC(3)	0.83(0.19)	0.84(0.19)	0.59(0.16)	0.61(0.16)	2.45(0.28)	2.49(0.29)	1.90(0.25)	1.87(0.27)

Table 3.1: Mean (SD) of the values of the Amari error between the true and estimated mixing matrices of indicated methods, dimensions and sample sizes.

algorithms for different choices of the dimension d and the sample size n. For example, when d=5 and n=1000, the Amari measure of 0.69 for TVDOC(1) is about 49.6% of the measure for PCA, 52.7% for TVPCA and 57.1% for DOC(1). The smaller averages and standard deviations of the Amari metric for the TVDOC estimator indicate its mixing matrix estimator is less biased and more stable than the other methods. The better perfor-

mance of the TVDOC indicates that, for the analysis of nonstationary multivariate time series, it is plausible to divide the whole time span into several homogeneous segments and then apply DOC to each segment separately.

3.5.2 Real data analysis

We illustrate the details of implementing TVDOC and TS-PCA methods by analyzing two real datasets of dimensions 3 (Example 1) and 135 (Example 2), respectively.

Example 1. First, we consider a three-dimensional time series of the daily log returns in percentage of the S&P 500 Index, Cisco System and Intel Corporation stocks from January 2, 2007 through January 2, 2012, with n = 1259 observations. This dataset is a shorter, but a more recent segment of the same three series analyzed in [41] and includes data for 2008, the year of financial crisis. The return series plotted in Figure 3.1 show presence of

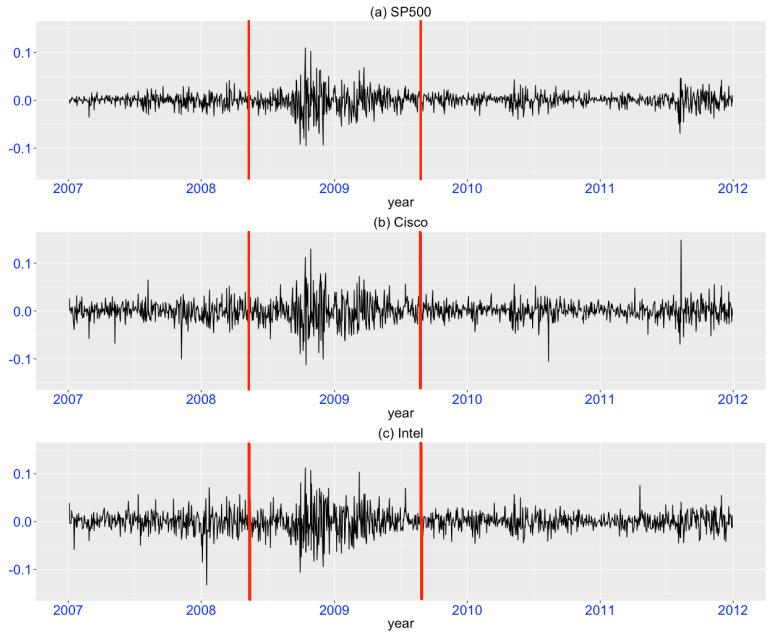

Figure 3.1: Daily log returns of (a) S&P 500 Index, (b) Intel Corporation stock and (c) Cisco Systems stock. The vertical lines indicate the locations of the two change points.

volatility clusters where the volatilities generally move together, and as might be expected there is increased volatility in the fourth quarter of 2008 in each series due to the financial crisis. We note that the three series are pairwise correlated and their (contemporaneous) sample correlations are about 0.5. It is expected and we show that TVDOC outperforms DOC in such a dataset with changing volatility and a nonstationary pattern.

In applying the TVDOC to the data we use the method in [18] for detecting change points in the series. As noted earlier the number of detected change points depends on the tuning parameter m_0. For example, it segments the time series into two parts over the time ranges from 1 to 610 and 611 to 1259 when $m_0 = 200$, while reducing m_0 to 150 it divides the series into three segments over the time ranges 1 to 290, 291 to 656 and 657 to 1259. Note that for the latter segmentation the middle segment has larger volatility while the other two seem reasonably homogeneous.

A VAR (3) model, with order selected using the AIC with the upper bound of 5, is fitted to the whole series to prewhiten it. Let its residual series be denoted by $\hat{\mathbf{e}}_t$ and divided into three segments $\hat{\mathbf{e}}_{1,t}, \hat{\mathbf{e}}_{2,t}$ and $\hat{\mathbf{e}}_{3,t}$, respectively. The multivariate Ljung-Box statistics and the p-values for $\hat{\mathbf{e}}_{i,t}$ and $\hat{\mathbf{e}}_{i,t}^2$ in Table 3.2 reveal that, indeed, the VAR(3) model has removed the serial correlation, but significant serial correlation remains in the squared residuals in each segment, indicating conditional heteroscedasticity.

Next, for each $\hat{\mathbf{e}}_{i,t}, i = 1, 2, 3$, we check whether it is already DOC in volatility. In Table 3.2, the observed DOC test statistics $Q_3^0(\hat{\mathbf{e}}_{i,t}^2, 10)$ for the three segments are very large relative to a χ^2 with 63 degrees of freedom, indicating that $\hat{\mathbf{e}}_{i,t}$ are not DOC in volatility. Thus, one may model them as linear combinations of their respective DOCs, namely as $\hat{\mathbf{e}}_{i,t} = \mathbf{M}_i \mathbf{s}_{i,t}$. To this end, we first decorrelate $\hat{\mathbf{e}}_{i,t}$ using $\hat{\mathbf{z}}_{i,t} = \hat{\mathbf{\Lambda}}_i^{-1/2} \hat{\mathbf{P}}_i' \hat{\mathbf{e}}_{i,t}$, where $\hat{\mathbf{\Lambda}}_i$ and $\hat{\mathbf{P}}_i$ are the diagonal matrix of eigenvalues and the orthogonal matrix of eigenvectors of their sample covariance matrices, and apply DOC to $\hat{\mathbf{z}}_{i,t}$ by estimating the mixing matrix $\mathbf{M}_i$ and the DOCs $\hat{\mathbf{s}}_{i,t}$. For example, the Ljung-Box type test statistic $Q_3^0(\hat{\mathbf{s}}_{1,t}^2, 10) = 95.34$

	m			
segment1	5	10	15	20
$\hat{e}_{1,t}$	62.21	103.81	146.23	186.81
	0.05	0.15	0.24	0.35
$\hat{e}^2_{1,t}$	64.69	126.75	179.66	200.99
	0.03	0.01	0.01	0.14
$\hat{\epsilon}_{1,t}$	58.68	98.69	140.1	180.58
	0.08	0.25	0.36	0.47
$\hat{\epsilon}^2_{1,t}$	24.42	53.36	69.91	80.74
	0.99	1	1	1
segment2	5	10	15	20
$\hat{e}_{2,t}$	49.31	98.85	142.41	206.43
	0.3	0.25	0.31	0.09
$\hat{e}^2_{2,t}$	210.11	390.44	585.29	717.24
	0	0	0	0
$\hat{\epsilon}_{2,t}$	30.93	70.32	104.8	163.4
	0.95	0.94	0.97	0.81
$\hat{\epsilon}^2_{2,t}$	43.34	93.89	155.93	216.16
	0.54	0.37	0.1	0.03
segment3	5	10	15	20
$\hat{e}_{3,t}$	74.99	111.52	160.69	191.73
	0	0.06	0.06	0.26
$\hat{e}^2_{3,t}$	211.65	308.34	352.35	377.32
	0	0	0	0
$\hat{\epsilon}_{3,t}$	65.51	103.99	152.43	183.76
	0.02	0.15	0.14	0.41
$\hat{\epsilon}^2_{3,t}$	48.98	91.74	118.61	132.15
	0.32	0.43	0.84	1

Table 3.2: Ljung-Box statistics and p-values for (a) the residual and the squared residuals of the fitted VAR(3) and (b) the standardized residual and their squares of the fitted DOC-GARCH$(1,1)-t$ model for S&P 500 Index, Cisco and Intel stock's daily percentage log-returns.

indicates that $\hat{s}^2_{1,t}$ is already DOC in volatility. Finally, we apply the GARCH$(1,1)-t$ model in (3.2) to each DOC in each segment, i.e. $\hat{s}_{i,k,t}, i, k = 1,2,3$. Denoting the estimated GARCH residuals for each segment as $\hat{\epsilon}_{i,t}$, it can be seen from Table 3.2 that this TVDOC-GARCH$(1,1)-t$ model has successfully decorrelated the original time series $\mathbf{X}_t$.

Next, we use Figure 3.2 to compare the performances of TVDOC, DOC, O-GARCH

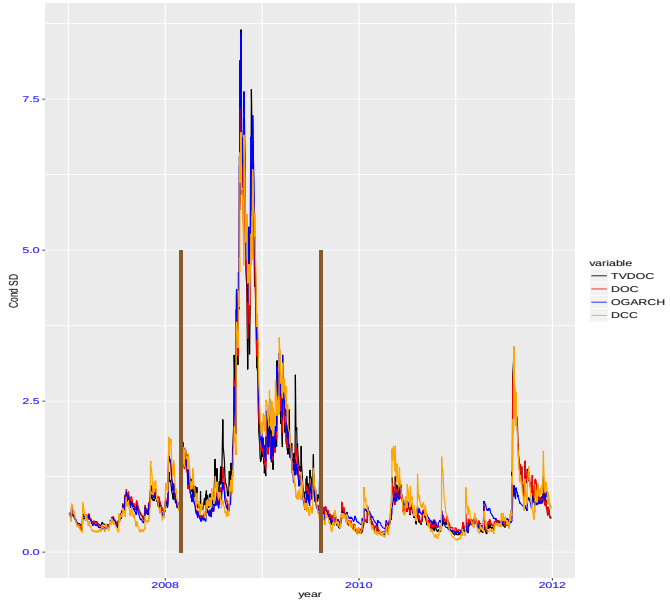

Figure 3.2: Conditional standard deviation fitted using 4 different models, TVDOC-GARCH, DOC-GARCH, O-GARCH and DCC-GARCH for S&P 500 Index daily percentage log returns.

([3]) and DCC ([24]) models by fitting them to the three segments of the series, separately. Figure 3.2 shows the results for the fitted conditional standard deviations for the S&P 500 series, while Figure 3.3 shows the results for the estimated conditional correlations between the S&P 500 Index and Intel Corporation returns, respectively. A rolling window correlation estimator with a 6-month window is also plotted for comparison.

It can be seen from Figure 3.2 that for the conditional standard deviation estimation

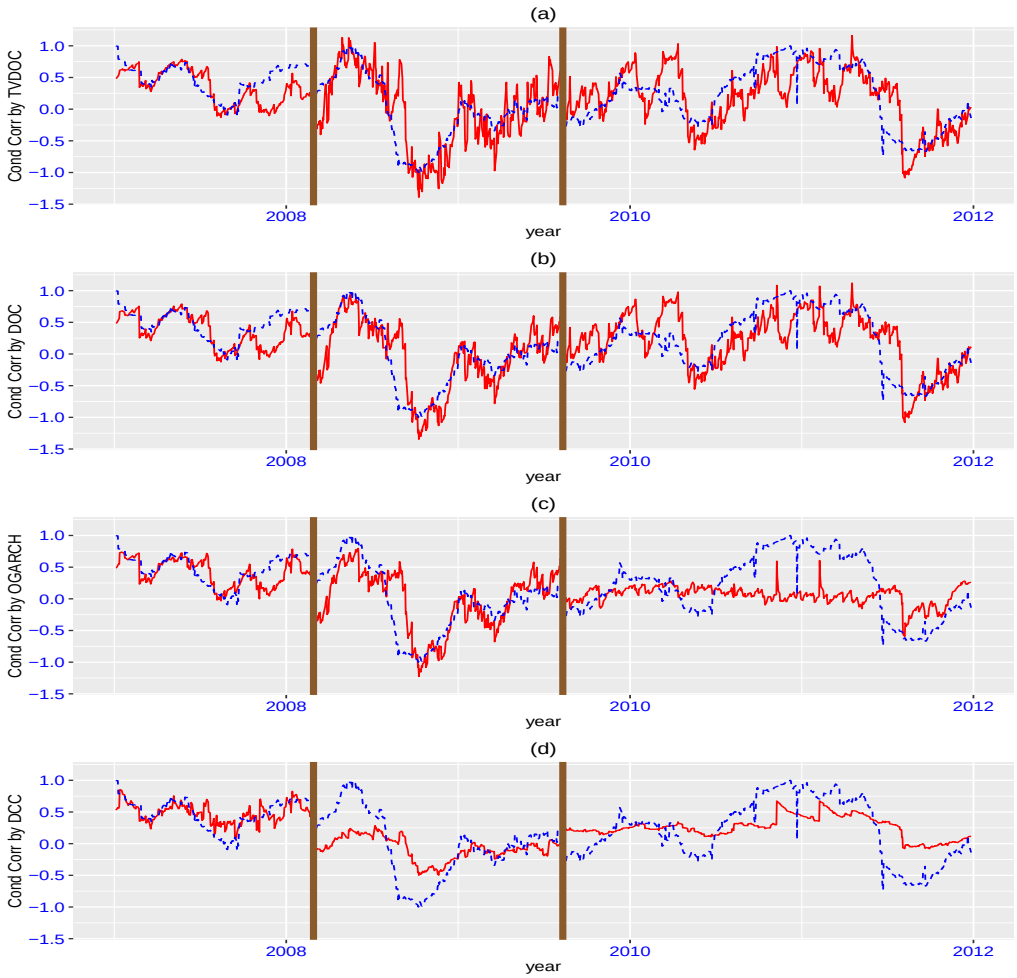

Figure 3.3: Conditional correlations fitted using 4 different models, TVDOC-GARCH, DOC-GARCH, O-GARCH and DCC-GARCH for S&P 500 Index and Intel Corporation daily percentage log returns. A rolling window correlation estimator with a 6-month window is plotted with brown solid lines.

on the first and second segments of data, the performances of the four models are similar to each other. Figure 3.2 also shows that on the third segment, the O-GARCH model performs worse than the other three in the sense that it fails to capture the large volatil-

ity. For the fitted correlations, the results for TVDOC and DOC are comparable and they outperform the other two. For example, in the second segment, it can be seen from Figure 3.3 that the correlations estimated using TVDOC and DOC match up closely with the rolling window correlation estimator (viewed as a proxy for the true correlation). However, the conditional correlation fitted by DCC is oscillating slightly around zero and it doesn't match up with the rolling window estimator. Figure 3.3 also shows that, for the third segment, the correlations estimated by TVDOC and DOC match up with the rolling window estimator more closely than the other two methods.

Next, we illustrate the use of TS-PCA in partitioning vertically a high-dimensional macroeconomic time series and highlight its potential role in DOC analysis.

Example 2. We apply the TS-PCA to FRED-MD [42], a large monthly macroeconomic data available at `research.stlouisfed.org/econ/mccracken`. The latest available version is from January 01, 1959 to August 01, 2015, with $n = 680$ observations for $d = 135$ series, with some missing values. The dimension of the series here is much larger than $d = 25$ in Example 4 in [16]. We present results for various choice of the tuning parameters (λ, k_0, m) appearing in TS-PCA. The tuning parameter λ seems to have the most influence on the number of non-singleton subseries as seen in Table 3.3.

We also assess the impact of k_0 by fixing $\lambda = 2, m = 25$ and varying k_0 from 1 to 5. The resulting subseries with more than 1 components are shown in the Appendix where it can be seen that its impact is minimal, only for $k_0 = 3$ a four-dimensional subseries appears in the list.

λ	Time	Grouping
1	0.14	{115,127}, {116,126}, {122,130}, {118,124,129}, {125,128},
2	0.12	{61,125}, {64,126}, {75,128}, {86,127}, {89,129},
3	0.11	{17,96}, {37,111}, {47,108}, {56,112}, {77,106,109,110}, {100,103}, {105,113}
4	0.12	{32,67}, {46,74}, {59,65,66,68,69}, {71,87}
5	0.10	{31,39}, {33,41}, {36,38}, {37,108}, {40,42,65,66,67,68,69,70,71,72}, {107,110}

Table 3.3: Computation times (in minutes) and the non-singleton subseries from applying TS-PCA to the FRED-MD data, with $k_0 = 5, m = 25$ fixed and varying λ from 1 to 5.

4. SUMMARY AND CONCLUSIONS

4.1 Challenges

We have extended the stationary DOC analysis to the high-dimensional nonstationary setup, and have explored the connections between the DOC and the TS-PCA methods. Computationally, TS-PCA is much faster than DOC, but its objective function is less statistically interpretable than that in DOC. Nevertheless, TS-PCA has the potential to overcome the computational bottle-neck encountered in optimizing the DOC's nonconvex objective function.

4.2 Further study

A number of problems remain unsolved for our proposed method. The first one is that of existence of the mixing matrix M for a given a high dimensional data series $\boldsymbol{X}_t$ or establishing a valid test for deciding when it exists. Only after the test shows that the series could be grouped into lower-dimensional subseries, one can apply TS-PCA method to preprocess the data. If it cannot be grouped, then the second question is, how to design an alternative method to reduce the data dimension and then make it possible to apply the DOC model? The third problem is that of studying the impact of estimating the change points on our asymptotic results.

REFERENCES

[1] S. Adak. Time-dependent spectral analysis of nonstationary time series. *Journal of the American Statistical Association*, 93(444):1488–1501, 1998.

[2] S. Aghabozorgi, A. Seyed Shirkhorshidi, and T. Ying Wah. Time-series clustering - a decade review. *Information Systems*, 53(C):16–38, 2015.

[3] C. Alexander. *Market Models: A Guide to Financial Data Analysis*. John Wiley and Sons, Chichester, 2001.

[4] D. W. K. Andrews. Tests for parameter instability and structural change with unknown change point. *Econometrica*, 61(4):821–856, 1993.

[5] A. Aue, S. Hörmann, L. Horváth, and M. Reimherr. Break detection in the covariance structure of multivariate time series models. *The Annals of Statistics*, 37(6B):4046–4087, 2009.

[6] F. R. Bach and M. I. Jordan. Kernel independent component analysis. *Journal of Machine Learning Research*, 3(1):1–48, 2003.

[7] A. D. Back and A. S. Weigend. A first application of independent component analysis to extracting structure from stock returns. *International Journal of Neural Systems*, 8(4):473–484, 1997.

[8] J. Bai. Estimation of a change point in multiple regression models. *Review of Economics and Statistics*, 79(4):551–563, 1997.

[9] J. Bai. Inferential theory for factor models of large dimensions. *Econometrica*, 71(1):135–171, 2003.

[10] R. T. Baillie and C. Morana. Modelling long memory and structural breaks in conditional variances: an adaptive FIGARCH approach. *Journal of Economic Dynamics and Control*, 33(8):1577–1592, 2009.

[11] D. Bianchi, M. Guidolin, and F. Ravazzolo. Macroeconomic factors strike back: a bayesian change-point model of time-varying risk exposures and premia in the U.S. cross-section. *Journal of Business and Economic Statistics*, 35(1):110–129, 2017.

[12] P. J. Bickel and E. Levina. Covariance regularization by thresholding. *The Annals of Statistics*, 36(6):2577–2604, 2008.

[13] T. Bollerslev, R. F. Engle, and J. M. Wooldridge. A capital asset pricing model with time-varying covariances. *Journal of Political Economy*, 96(1):116–131, 1988.

[14] G. E. Box and G. C. Tiao. A canonical analysis of multiple time series. *Biometrika*, 64(2):355–365, 1977.

[15] D. R. Brillinger. *Time Series: Data Analysis and Theory*. Society for Industrial and Applied Mathematics, Philadelphia, 2001.

[16] J. Chang, B. Guo, and Q. Yao. High dimensional stochastic regression with latent factors, endogeneity and nonlinearity. *Journal of Econometrics*, 189(2):297–312, 2015.

[17] J. Chang, B. Guo, and Q. Yao. Principal component analysis for second-order stationary vector time series. 2016. To appear, available at `https://arxiv.org/pdf/1410.2323.pdf`.

[18] R. B. Chen, Y. Chen, and W. K. Härdle. TVICA - time varying independent component analysis and its application to financial data. *Computational Statistics and Data Analysis*, 74(54):95–109, 2014.

[19] R. Dahlhaus. Fitting time series models to nonstationary processes. *The Annals of Statistics*, 25(1):1–37, 1997.

[20] R. A. Davis, T. C. M. Lee, and G. A. Rodriguez-Yam. Structural break estimation for nonstationary time series models. *Journal of the American Statistical Association*, 101(473):223–239, 2006.

[21] R. A. Davis, T. C. M. Lee, and G. A. Rodriguez-Yam. Break detection for a class of nonlinear time series models. *Journal of Time Series Analysis*, 29(5):834–867, 2008.

[22] R. M. De Jong. Central limit theorems for dependent heterogeneous random variables. *Econometric Theory*, 13(3):353–367, 1997.

[23] M. Eichler, G. Motta, and R. Von Sachs. Fitting dynamic factor models to nonstationary time series. *Journal of Econometrics*, 163(1):51–70, 2011.

[24] R. Engle. Dynamic conditional correlation: a simple class of multivariate generalized autoregressive conditional heteroskedasticity models. *Journal of Business and Economic Statistics*, 20(3):339–350, 2002.

[25] R. F. Engle and K. F. Kroner. Multivariate simultaneous generalized ARCH. *Econometric Theory*, 11(1):122–150, 1995.

[26] J. Fan, M. Wang, and Q. Yao. Modelling multivariate volatilities via conditionally uncorrelated components. *Journal of the Royal Statistical Society: Series B*, 70(4):679–702, 2008.

[27] X. Fan, Y. Yuan, X. Zhuang, and X. Jin. Long memory of abnormal investor attention and the cross-correlations between abnormal investor attention and trading volume, volatility respectively. *Physica A: Statistical Mechanics and its Applications*, 469(1):323–333, 2017.

[28] M. Forni, M. Hallin, M. Lippi, and L. Reichlin. The generalized dynamic factor model: one-sided estimation and forecasting. *Journal of the American Statistical Association*, 100(471):830–840, 2005.

[29] C. Francq and J. Zakoian. *GARCH Models: Structure, Statistical Inference and Financial Applications*. John Wiley and Sons, New York, 2010.

[30] T. Gasser, J. Möcks, and R. Verleger. SELAVCO: a method to deal with trial-to-trial variability of evoked potentials. *Electroencephalography and Clinical Neurophysiology*, 55(6):717–723, 1983.

[31] H. Hotelling. Analysis of a complex statistical variables into principal components. *Journal of Educational Psychology*, 24(6):417–441, 1933.

[32] H. Hotelling. Relations between two sets of variates. *Biometrika*, 28(3-4):321–377, 1936.

[33] Y. P. Hu and R. S. Tsay. Principal volatility component analysis. *Journal of Business and Economic Statistics*, 32(2):153–164, 2014.

[34] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. John Wiley and Sons, New York, 2004.

[35] I. A. Ibragimov. Some limit theorems for stationary processes. *Theory of Probability and Its Applications*, 7(4):349–382, 1962.

[36] H. Kato, S. Naniwa, and M. Ishiguro. A bayesian multivariate nonstationary time series model for estimating mutual relationships among variables. *Journal of Econometrics*, 75(1):147–161, 1996.

[37] M. Lavielle and C. Ludeña. The multiple change-points problem for the spectral distribution. *Bernoulli*, 6(5):845–869, 2000.

[38] W. Li, J. Gao, K. Li, and Q. Yao. Modeling multivariate volatilities via latent common factors. *Journal of Business and Economic Statistics*, 34(4):564–573, 2016.

[39] W. K. Li. *Diagonostic Checks in Time Series*. Chapman and Hall, Boca Raton, 2004.

[40] O. V. Lie and P. Van Mierlo. Seizure-onset mapping based on time-variant multivariate functional connectivity analysis of high-dimensional intracranial EEG: a kalman filter approach. *Brain Topography*, 30(1):46–59, 2017.

[41] D. S. Matteson and R. S. Tsay. Dynamic orthogonal components for multivariate time series. *Journal of the American Statistical Association*, 106(496):1450–1463, 2011.

[42] M. W. McCracken and S. Ng. Fred-md: a monthly database for macroeconomic research. *Journal of Business and Economic Statistics*, 34(4):574–589, 2016.

[43] D. Mercurio and V. Spokoiny. Statistical inference for time-inhomogeneous volatility models. *Annals of Statistics*, 32(2):577–602, 2004.

[44] M. Odening, E. Berg, and C. Turvey. Management of climate risks in agriculture. *Agricultural Finance Review*, 68(1):83–97, 2008.

[45] K. Peason. On lines and planes of closest fit to systems of point in space. *Philosophical Magazine*, 2(11):559–572, 1901.

[46] D. Peña and G. E. Box. Identifying a simplifying structure in time series. *Journal of the American Statistical Association*, 82(399):836–843, 1987.

[47] B. B. Risk, D. S. Matteson, D. Ruppert, A. Eloyan, and B. S. Caffo. An evaluation of independent component analyses with an application to resting-state fMRI. *Biometrics*, 70(1):224–236, 2014.

[48] J. H. Stock and M. W. Watson. Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association*, 97(460):1167–1179, 2002.

[49] W. Stout. *Almost Sure Convergence*. Academic Press, London, 1974.

[50] Y. Tu and Y. Yi. Forecasting cointegrated nonstationary time series with time-varying variance. *Journal of Econometrics*, 196(1):83–98, 2017.

[51] R. Van Der Weide. GO-GARCH: a multivariate generalized orthogonal GARCH model. *Journal of Applied Econometrics*, 17(5):549–564, 2002.

[52] J. M. Wooldridge and H. White. Some invariance principles and central limit theorems for dependent heterogeneous processes. *Econometric Theory*, 4(2):210–230, 1988.

APPENDIX A

SUPPLEMENTARY MATERIALS

A.1 DOC vs TS-PCA

We illustrate that DOC (in mean and volatility) and TS-PCA are closely related to each other, such a connection is helpful in replacing the challenging nonconvex optimization problem in DOC by the simpler eigenanalysis of the two target positive-definite matrices for TS-PCA in mean and volatility. It is easier to see that DOC in mean and TS-PCA in mean are doing nearly the same thing, but using different objective functions. However, connecting the TS-PCA for volatility processes in [16, Section 5] to DOC in volatility does not seem straightforward. A good starting point might be to apply TS-PCA directly to $\mathbf{s}_t^2$.

First, we assess the tendency of TS-PCA in mean in segmenting a high-dimensional time series into lower-dimensional subseries. To this end, we simulate multivariate time series data which are DOC in mean. We use the VAR(2) model $\mathbf{s}_t = C_1\mathbf{s}_{t-1} + C_2\mathbf{s}_{t-2} + \mathbf{e}_t$ to generate the d-dimensional orthogonal component $\mathbf{s}_t$ of length $n = 500$ where $\mathbf{e}_t$ is a d-dimensional white noise $N(\mathbf{0}_d, \mathbf{I}_d)$, and C_1, C_2 are given diagonal matrices. Then we simulate data $\boldsymbol{X}_t = M\mathbf{s}_t$ using a $d \times d$ mixing matrix M generated with entries drawn iid from a standard normal distribution.

For $d = 5, 10$, the simulated coefficient matrices and the mixing matrix are, respectively,

$$
\begin{aligned}
C1 &= \text{diag}(-0.458, 0.007, -0.366, -0.242, -0.149), \\
C2 &= \text{diag}(0.296, 0.263, 0.392, 0.632, 0.246), \\
\mathbf{M} &= \begin{pmatrix}
-0.59 & 0.14 & 0.06 & 0.48 & 0.22 \\
1.59 & 1.10 & 0.77 & 0.44 & -0.96 \\
-2.32 & -0.41 & -0.81 & 0.35 & -1.10 \\
0.63 & 0.05 & -0.84 & -0.34 & 1.22 \\
-2.25 & -0.24 & -0.24 & 1.85 & 0.07
\end{pmatrix},
\end{aligned}
$$

and

$$
\begin{aligned}
C1 &= \text{diag}(-0.009, 0.039, -0.185, -0.053, 0.0159, -0.114, -0.076, -0.17, -0.222, -0.308), \\
C2 &= \text{diag}(0.259, 0.593, 0.186, 0.509, 0.274, 0.329, 0.251, 0.2661, 0.298, 0.228), \\
\mathbf{M} &= \begin{pmatrix}
0.78 & 0.16 & 0.14 & -1.0 & 0.87 & 2.07 & 0.20 & 1.02 & -0.61 & -0.56 \\
-0.68 & -0.11 & 2.26 & 0.03 & -1.29 & 0.24 & 0.95 & -0.92 & -1.03 & 1.56 \\
0.71 & -0.17 & -0.94 & -1.21 & -1.45 & -0.04 & -0.15 & 0.91 & -1.33 & -0.78 \\
-0.87 & 0.51 & -1.20 & -0.86 & 2.07 & -0.65 & 0.10 & -1.11 & 0.67 & -1.49 \\
0.18 & -0.58 & 0.81 & 0.33 & -1.08 & -2.26 & 0.25 & 0.03 & -0.22 & 0.58 \\
1.56 & 0.31 & -1.50 & -1.41 & 1.15 & -0.60 & 0.03 & 0.58 & 0.97 & -0.83 \\
-2.02 & -0.64 & -0.03 & -0.66 & -0.26 & -0.95 & -0.43 & 1.44 & -0.08 & -0.47 \\
-0.31 & -0.24 & 0.59 & 0.34 & 1.56 & -1.70 & -0.19 & 1.31 & -0.01 & -0.41 \\
2.58 & 0.56 & 0.64 & -1.32 & -0.07 & 0.09 & -0.13 & 0.69 & 2.64 & -1.35 \\
1.55 & 0.81 & 0.89 & 0.61 & -0.37 & -1.24 & 1.25 & 0.57 & -0.65 & -1.30
\end{pmatrix}.
\end{aligned}
$$

To compare the performance of DOC in volatility and TS-PCA for volatility processes in [16], we simulate from the GARCH$(1,1) - t_\nu$ model for each component $s_{it}, i = 1, 2, \cdots, d$, with length n = 500, where $t_\nu(0,1)$ denote the standardized Student-t distribution with ν degrees of freedom. The two mixing matrices M in (A.1) and (A.1) are used to simulate the data $\boldsymbol{X}_t = M\mathbf{s}_t$ for the dimensions $d = 5, 10$, respectively. TS-PCA in volatility is applied to obtain the estimate $\hat{\mathbf{s}}_t$ of $\mathbf{s}_t$. Then we fit a GARCH(1,1) model to each of the component series of $\mathbf{s}_t$ and calculate the residuals $\widehat{\epsilon}_t$. The simulation is repeated 1000 times and the tuning parameters for TS-PCA in Volatility are set at m = 25, $\lambda = 2$ and $k_0 = 5$. The simulation results reported in Table A.1, the average time cost for the methods during each simulation, the Amari error between the estimated and the true M for both the TS-PCA and DOC method. It can be seen from Table A.1 that DOC has smaller Amari errors than the TS-PCA while its time cost is larger than that of TS-PCA.

We also compare the 1-step ahead out-of-sample prediction performances between TS-PCA and DOC. For volatility methods, during each of the 1000 simulations, we repeat the

	mean			volatility		
	$A(\hat{\mathbf{M}}, \mathbf{M})$	Time	MSE	$A(\hat{\mathbf{M}}, \mathbf{M})$	Time	MSE
	d=5					
TSPCA	1.29(0.13)	0.04	5.14(2.32)	1.44(0.04)	0.51	127.29(200.11)
DOC	0.56(0.19)	0.38	4.91(2.15)	0.92(0.21)	0.45	97.72(100.49)
	d=10					
TSPCA	3.41(0.19)	0.07	10.62(2.87)	3.47(0.19)	1.18	176.02(120.34)
DOC	2.06(0.34)	6.45	10.49(3.08)	2.59(0.31)	6.76	179.84(124.56)

Table A.1: The Amari error between the estimated and the true M, the time cost (in second), as well as the 1-step ahead out-of-sample prediction mean squared error for the TS-PCA and DOC methods.

following steps 1 to 4 for $k = 1, 2, \cdots, 5$:

Step 1: Apply TS-PCA or DOC to estimate $\hat{\mathbf{M}}$ and $\hat{\mathbf{s}}_t$ so that $\boldsymbol{X}_t = \hat{\mathbf{M}}\hat{\mathbf{s}}_t$ for $t = 1, 2, \cdots, T-k$.

Step 2: In TS-PCA, fit GARCH(1,1) to the m-th segmented subseries $\hat{\mathbf{s}}_t^{(m)}$ if it is univariate and BEKK(1,1) otherwise which is defined as,

$$\text{cov}(\hat{\mathbf{s}}_t^{(m)}|\mathcal{F}_{t-1}) = \mathbf{A}_0\mathbf{A}_0' + \mathbf{A}_1\hat{\mathbf{s}}_{t-1}^{(m)}\hat{\mathbf{s}}_{t-1}^{(m)\prime}\mathbf{A}_1' + \mathbf{B}_1\text{cov}(\hat{\mathbf{s}}_{t-1}^{(m)}|\mathcal{F}_{t-2})\mathbf{B}_1'.$$

In DOC, fit GARCH(1,1) to each component of $\mathbf{s}_t$.

Step 3: Make 1-step ahead prediction using the fitted models and transform back to get

$$\widehat{\text{cov}}(\boldsymbol{X}_{T-k+1}|\mathcal{F}_{T-k}) = \hat{\mathbf{M}}\text{cov}(\hat{\mathbf{s}}_{T-k+1}|\mathcal{F}_{T-k})\hat{\mathbf{M}}'$$

Step 4: $\text{E}(\text{cov}(\boldsymbol{X}_{T-k+1}|\mathcal{F}_{T-k})) = \text{E}(\text{E}(\boldsymbol{X}_{T-k+1}\boldsymbol{X}_{T-k+1}'|\mathcal{F}_{T-k})) = \text{E}(\boldsymbol{X}_{T-k+1}\boldsymbol{X}_{T-k+1}')$, thus $\boldsymbol{X}_{T-k+1}\boldsymbol{X}_{T-k+1}'$ is a good approximation of $\text{E}(\boldsymbol{X}_{T-k+1}\boldsymbol{X}_{T-k+1}'|\mathcal{F}_{T-k})$. Now calculate the mean squared error (MSE) between this approximated conditional covariance

matrix and the predicted one by,

$$\frac{1}{d^2}||\boldsymbol{X}_{T-k+1}\boldsymbol{X}'_{T-k+1} - \widehat{\text{cov}}(\boldsymbol{X}_{T-k+1}|\mathcal{F}_{T-k})||_2^2.$$

Finally, calculate the average of the MSE for the 5 repeats. The method with the smaller such average has better performance in predicting the conditional covariance. The prediction procedures for the mean methods are similar. The only differences are that the GARCH(1,1) is replaced by AR(p), BEKK(1,1) is replaced by VAR(p), with p selected by AIC and back-transformation is replaced by $\hat{\boldsymbol{X}}_{T-k+1} = \hat{\mathbf{M}}\hat{\mathbf{s}}_{T-k}$. The prediction MSEs for the simulated data mentioned above are listed in Table A.1. It can be seen that for most of the cases, the out-of-sample prediction performance for DOC and TSPCA are similar. When $d = 5$, the performance for DOC in volatility is better than that of TSPCA in volatility by having a smaller MSE.

Deeper connections between DOC and TS-PCA in volatility is revealed by the following surprising identity involving the covariance matrices of $\mathbf{s}_t^2$ and $\boldsymbol{X}_t$ and the mixing matrix $\mathbf{M}$.

Lemma 2. *Let $\{\boldsymbol{X}_t\}$ be a mean-zero stationary process satisfying $\boldsymbol{X}_t = \mathbf{M}\mathbf{s}_t$. Then,*
(a)

$$cov(\mathbf{s}_t, \mathbf{s}_{t-h}) = \mathbf{M}\boldsymbol{\Gamma}_x(h)\mathbf{M}'. \tag{A.1}$$

(b) provided that $\{\boldsymbol{X}_t\}$ is Gaussian, we have

$$cov(\mathbf{s}_t^2, \mathbf{s}_{t-h}^2) = 2(\mathbf{M}\boldsymbol{\Gamma}_x(h)\mathbf{M}') \circ (\mathbf{M}\boldsymbol{\Gamma}_x(h)\mathbf{M}'), \tag{A.2}$$

where $\circ$ denotes the Hadamard product two matrices.

Proof. (a) The conclusion holds since for any $i \neq j$,

$$\begin{aligned}
\operatorname{cov}(\mathbf{s}_{t,i}, \mathbf{s}_{t-h,j}) &= \mathrm{E}\mathbf{s}_{t,i}\mathbf{s}_{t-h,j} = \mathrm{E}\left[\left(\sum_{k=1}^{d} \mathbf{M}_{i,k}\boldsymbol{X}_{t,k}\right)\left(\sum_{k=1}^{d} \mathbf{M}_{j,k}\boldsymbol{X}_{t-h,k}\right)\right] \\
&= \sum_{m=1}^{d}\sum_{n=1}^{d} \mathbf{M}_{i,m}\mathbf{M}_{j,n}\mathrm{E}\left(\boldsymbol{X}_{t,m}\boldsymbol{X}_{t-h,n}\right) \\
&= \sum_{m=1}^{d}\sum_{n=1}^{d} \mathbf{M}_{i,m}\mathbf{M}_{j,n}\boldsymbol{\Gamma}_x(h)_{m,n} \\
&= \left[(\mathbf{M}_{i,1}, \cdots, \mathbf{M}_{i,d})\boldsymbol{\Gamma}_x(h)(\mathbf{M}_{j,1}, \cdots, \mathbf{M}_{j,d})'\right].
\end{aligned}$$

(b) $\mathbf{s}_t = \mathbf{M}\boldsymbol{X}_t$ implies that $\mathbf{s}_{t,i} = \sum_{k=1}^{d} \mathbf{M}_{i,k}\boldsymbol{X}_{t,k}$. Thus,

$$\begin{aligned}
\mathrm{E}\mathbf{s}_{t,i}^2\mathbf{s}_{t-h,j}^2 &= \mathrm{E}\left[\left(\sum_{k=1}^{d} \mathbf{M}_{i,k}\boldsymbol{X}_{t,k}\right)^2\left(\sum_{k=1}^{d} \mathbf{M}_{j,k}\boldsymbol{X}_{t-h,k}\right)^2\right] \\
&= \mathrm{E}\left[\sum_{m=1}^{d}\sum_{n=1}^{d} \mathbf{M}_{i,m}\mathbf{M}_{i,n}\boldsymbol{X}_{t,m}\boldsymbol{X}_{t,n}\sum_{\alpha=1}^{d}\sum_{\beta=1}^{d} \mathbf{M}_{j,\alpha}\mathbf{M}_{j,\beta}\boldsymbol{X}_{t-h,\alpha}\boldsymbol{X}_{t-h,\beta}\right] \\
&= \sum_{m=1}^{d}\sum_{n=1}^{d}\sum_{\alpha=1}^{d}\sum_{\beta=1}^{d} \mathbf{M}_{i,m}\mathbf{M}_{i,n}\mathbf{M}_{j,\alpha}\mathbf{M}_{j,\beta}\mathrm{E}\boldsymbol{X}_{t,m}\boldsymbol{X}_{t,n}\boldsymbol{X}_{t-h,\alpha}\boldsymbol{X}_{t-h,\beta} \\
&= \sum_{m=1}^{d}\sum_{n=1}^{d}\sum_{\alpha=1}^{d}\sum_{\beta=1}^{d} \mathbf{M}_{i,m}\mathbf{M}_{i,n}\mathbf{M}_{j,\alpha}\mathbf{M}_{j,\beta}(\mathrm{E}\boldsymbol{X}_{t,m}\boldsymbol{X}_{t,n}\mathrm{E}\boldsymbol{X}_{t-h,\alpha}\boldsymbol{X}_{t-h,\beta} \\
&+ \mathrm{E}\boldsymbol{X}_{t,m}\boldsymbol{X}_{t-h,\alpha}\mathrm{E}\boldsymbol{X}_{t,n}\boldsymbol{X}_{t-h,\beta} + \mathrm{E}\boldsymbol{X}_{t,m}\boldsymbol{X}_{t-h,\beta}\mathrm{E}\boldsymbol{X}_{t,n}\boldsymbol{X}_{t-h,\alpha}),
\end{aligned}$$

where the last equal is based on the conclusion stated in [15, Equation (2.3.8)]. Then we

index	2	3	4	5	22	23	56	57	58
count	1	1	1	2	12	12	12	12	12
index	59	60	65	67	69	70	79	80	84
count	12	12	398	109	1	1	1	1	5
index	102	124	125	126	127	131	133	134	136
count	168	1	1	1	1	154	2	2	42

Table A.2: The index and the number of missing values in the FRED-MD data.

have,

$$\begin{aligned}
\mathrm{cov}(\mathbf{s}_{t,i}^2, \mathbf{s}_{t-h,j}^2) &= \mathrm{E}\mathbf{s}_{t,i}^2\mathbf{s}_{t-h,j}^2 - \mathrm{E}\mathbf{s}_{t,i}^2\mathrm{E}\mathbf{s}_{t-h,j}^2 \\
&= \sum_{m=1}^{d}\sum_{n=1}^{d}\sum_{\alpha=1}^{d}\sum_{\beta=1}^{d}\mathbf{M}_{i,m}\mathbf{M}_{i,n}\mathbf{M}_{j,\alpha}\mathbf{M}_{j,\beta}(\mathbf{\Gamma}_x(0)_{m,n}\mathbf{\Gamma}_x(0)_{\alpha,\beta} \\
&+ \mathbf{\Gamma}_x(h)_{m,\alpha}\mathbf{\Gamma}_x(h)_{n,\beta} + \mathbf{\Gamma}_x(0)_{m,\beta}\mathbf{\Gamma}_x(0)_{\alpha,n}) \\
&- \sum_{m=1}^{d}\sum_{n=1}^{d}\sum_{\alpha=1}^{d}\sum_{\beta=1}^{d}\mathbf{M}_{i,m}\mathbf{M}_{i,n}\mathbf{M}_{j,\alpha}\mathbf{M}_{j,\beta}\left(\mathbf{\Gamma}_x(0)_{m,n}\mathbf{\Gamma}_x(0)_{\alpha,\beta}\right) \\
&= 2\sum_{m=1}^{d}\sum_{n=1}^{d}\sum_{\alpha=1}^{d}\sum_{\beta=1}^{d}\mathbf{M}_{i,m}\mathbf{M}_{i,n}\mathbf{M}_{j,\alpha}\mathbf{M}_{j,\beta}\left(\mathbf{\Gamma}_x(h)_{m,\alpha}\mathbf{\Gamma}_x(h)_{n,\beta}\right) \\
&= 2\left[\sum_{m=1}^{d}\sum_{n=1}^{d}\mathbf{M}_{i,m}\mathbf{M}_{j,\alpha}\mathbf{\Gamma}_x(h)_{m,\alpha}\right]\left[\sum_{\alpha=1}^{d}\sum_{\beta=1}^{d}\mathbf{M}_{i,n}\mathbf{M}_{j,\beta}\mathbf{\Gamma}_x(h)_{n,\beta}\right] \\
&= 2\left[(\mathbf{M}_{i,1},\cdots,\mathbf{M}_{i,d})\mathbf{\Gamma}_x(h)(\mathbf{M}_{j,1},\cdots,\mathbf{M}_{j,d})'\right]^2.
\end{aligned}$$

Thus, the conclusion holds. □

A.2 More on FRED-MD in 3.5.2

There are 27 series in FRED-MD data that contain missing values, the index and number of missing values are shown in Table A.2. For each series, we replaced its missing values by the mean of its observed values, and have used the imputed series in the subsequent analysis. In Section 4 of the paper, the TS-PCA method was applied to the FRED-MD data for various choices of the tuning parameter λ and fixed $m = 25, k_0 = 5$. Here, since

k_0	Time	Grouping
1	0.07	{15,120}, {56,116,117}, {61,123}, {91,125}
2	0.07	{17,126}, {60,128}, {76,121}, {92,127}
3	0.08	{65,128}, {71,125}, {83,119,122,126}, {87,127}
4	0.09	{65,128}
5	0.10	{61,125}, {64,126}, {75,128}, {86,127}, {89,129},

Table A.3: Computational times (in minutes) and the resulting non-singleton groups by applying TS-PCA to the FRED-MD data, with $\lambda = 2, m = 25$ and k_0 varying from 1 to 5.

$d = 135$ is reasonably large it of interest to assess the impact of k_0 on the dimensions of the subseries. We fix $\lambda = 2, m = 25$ and vary k_0 from 1 to 5, the resulting subseries with more than 1 components are shown in the Table A.3. It can be seen that, as in the low-dimensional cases, the impact of k_0 is minimal in the sense that the non-univariate subseries in the table have two components, except for $k_0 = 1, 3$ where four-component and three-component subseries appear in the list.

A.3 Theorems

As $\bar{f}^{(i)}(\boldsymbol{\theta}_i)$ is continuously differentiable with respect to $\boldsymbol{\theta}_i$, its matrix gradient exists and is denoted by $\bar{F}^{(i)}(\boldsymbol{\theta}_i) := \frac{\partial \bar{f}^{(i)}(\boldsymbol{\theta}_i)}{\partial \boldsymbol{\theta}_i}$. Let $F^{(i)}$ denote the a.s. limit of $\bar{F}^{(i)}(\boldsymbol{\theta}_{0i})$. Then [41, Lemma A4 in supplementary material] implies $F^{(i)}$ exists and $F^{(i)} = E\{\frac{\partial}{\partial \boldsymbol{\theta}} \bar{F}^{(i)}(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}_{0i}}\}$. Recall that $\bar{g}(\boldsymbol{\theta}) = (\bar{f}^{(1)}(\boldsymbol{\theta}_1)', \bar{f}^{(2)}(\boldsymbol{\theta}_2)')'$, then define $\bar{G}_n = \frac{\partial \bar{g}_n(\theta)}{\partial \theta}$. Let $SO(d)$ be the subgroup of $O(d)$ with determinant 1. It is compact and closed under matrix multiplication and inversion.

Lemma 3. *Suppose $\{U_{n,t}\}$ is an α-mixing sequence of size $-r/(r-2)$. If $\{\boldsymbol{X}_{n,t}\}$ is L_2-NED of size $-\lambda$ on $\{U_{n,t}\}$ and $\sup_n \sup_t ||\boldsymbol{X}_{n,t}|| < \infty$, then for any $\ell \geq 0$, $\{\boldsymbol{X}_{n,t}\boldsymbol{X}_{n,t-\ell}\}$ is L_2-NED on $\{U_{n,t}\}$ of size $-\lambda$.*

Lemma 4. *Under Conditions C1-C9,*

$$|\bar{G}_n(\theta) - G(\theta)| \xrightarrow{a.s.} 0.$$

as $n \to \infty$ *for any* $W_\theta \in SO(d)$.

Lemma 5. *Under Conditions C1-C9,*

(i) As $n \to \infty$, $\frac{\partial \bar{G}_n(\theta)}{\partial \theta} \xrightarrow{a.s.} \frac{\partial G(\theta)}{\partial \theta}$.

(ii) There exists finite, point-wise, uniform bounds such that

$$\left|\left|\frac{\partial \bar{G}_n(\theta)}{\partial \theta}\right|\right|_F \leq B_n \quad \textit{and} \quad \left|\left|\frac{\partial G(\theta)}{\partial \theta}\right|\right|_F \leq B$$

for any $W_\theta \in SO(d)$ *and* $B_n \xrightarrow{a.s.} B$ *as* $n \to \infty$.

Lemma 6. *Under Conditions C1-C9, as* $n \to \infty$, $\sup_{W_\theta \in SO(d)} |\bar{G}_n(\theta) - G(\theta)| \xrightarrow{a.s.} 0$.

Theorem 3. *(Strong Consistency) Under* $C1 - C4$, $\hat{\boldsymbol{\theta}} \xrightarrow{a.s.} \boldsymbol{\theta}_0$ *as* $n, n - k_0(n) \to \infty$.

Theorem 4. *(Asymptotic Normality) Under* $C1 - C8$, *as* $n \to \infty$,

$$A_n \times \sqrt{n} \begin{pmatrix} (\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1) \\ (\hat{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}_2) \end{pmatrix} \xrightarrow{D} N\left(\mathbf{0}_{2d_0}, \mathbf{I}_{2d_0}\right),$$

where $d_0 := d(d-1)/2$.

A.4 Proofs of the results in A.3

Proof of Lemma 3. Let $\mathcal{F}_{t-m}^{t+m} = \sigma(U_{n,t-m}, \cdots, U_{n,t+m})$ and note that,

$$
\begin{aligned}
& ||\boldsymbol{X}_{n,t}\boldsymbol{X}_{n,t-\ell} - E[\boldsymbol{X}_{n,t}\boldsymbol{X}_{n,t-\ell}|\mathcal{F}_{t-m-2\ell}^{t+m}]|| \\
= \; & ||\boldsymbol{X}_{n,t}\boldsymbol{X}_{n,t-\ell} - \boldsymbol{X}_{n,t}E[\boldsymbol{X}_{n,t-\ell}|\mathcal{F}_{t-m-2\ell}^{t+m}] + \boldsymbol{X}_{n,t}E[\boldsymbol{X}_{n,t-\ell}|\mathcal{F}_{t-m-2\ell}^{t+m}] \\
- \; & E[\boldsymbol{X}_{n,t}|\mathcal{F}_{t-m}^{t+m}]E[\boldsymbol{X}_{n,t-\ell}|\mathcal{F}_{t-m-2\ell}^{t+m}] + E[\boldsymbol{X}_{n,t}|\mathcal{F}_{t-m}^{t+m}]E[\boldsymbol{X}_{n,t-\ell}|\mathcal{F}_{t-m-2\ell}^{t+m}] \\
- \; & E[\boldsymbol{X}_{n,t}\boldsymbol{X}_{n,t-\ell}|\mathcal{F}_{t-m-2\ell}^{t+m}]|| \\
\leq \; & ||\boldsymbol{X}_{n,t}\boldsymbol{X}_{n,t-\ell} - \boldsymbol{X}_{n,t}E[\boldsymbol{X}_{n,t-\ell}|\mathcal{F}_{t-m-2\ell}^{t+m}]|| \\
+ \; & ||\boldsymbol{X}_{n,t}E[\boldsymbol{X}_{n,t-\ell}|\mathcal{F}_{t-m-2\ell}^{t+m}] - E[\boldsymbol{X}_{n,t}|\mathcal{F}_{t-m}^{t+m}]E[\boldsymbol{X}_{n,t-\ell}|\mathcal{F}_{t-m-2\ell}^{t+m}]|| \\
+ \; & ||E[\boldsymbol{X}_{n,t}|\mathcal{F}_{t-m}^{t+m}]E[\boldsymbol{X}_{n,t-\ell}|\mathcal{F}_{t-m-2\ell}^{t+m}] - E[\boldsymbol{X}_{n,t}\boldsymbol{X}_{n,t-\ell}|\mathcal{F}_{t-m-2\ell}^{t+m}]|| \\
\leq \; & ||\boldsymbol{X}_{n,t}||||\boldsymbol{X}_{n,t-\ell} - E[\boldsymbol{X}_{n,t-\ell}|\mathcal{F}_{t-m-2\ell}^{t+m}]|| \\
+ \; & ||\boldsymbol{X}_{n,t-\ell}||||\boldsymbol{X}_{n,t} - E[\boldsymbol{X}_{n,t}|\mathcal{F}_{t-m}^{t+m}]|| \\
+ \; & ||E[E[\boldsymbol{X}_{n,t}|\mathcal{F}_{t-m}^{t+m}]\boldsymbol{X}_{n,t-\ell} - \boldsymbol{X}_{n,t}\boldsymbol{X}_{n,t-\ell}|\mathcal{F}_{t-m-2\ell}^{t+m}]|| \\
\leq \; & ||\boldsymbol{X}_{n,t}||v(m+\ell) + ||\boldsymbol{X}_{n,t-\ell}||v(m) + ||E[\boldsymbol{X}_{n,t}|\mathcal{F}_{t-m}^{t+m}]\boldsymbol{X}_{n,t-\ell} - \boldsymbol{X}_{n,t}\boldsymbol{X}_{n,t-\ell}|| \\
\leq \; & ||\boldsymbol{X}_{n,t}||v(m+\ell) + 2||\boldsymbol{X}_{n,t-\ell}||v(m),
\end{aligned}
$$

where we have used the fact that $\mathcal{F}_{t-m}^{t+m} \subset \mathcal{F}_{t-m-2\ell}^{t+m}$. Therefore, $\boldsymbol{X}_{n,t}\boldsymbol{X}_{n,t-\ell}$ is L_2-NED of the same size as $\{\boldsymbol{X}_{n,t}\}$ provided that $\sup_n \sup_t ||\boldsymbol{X}_{n,t}|| < \infty$. □

Proof of Lemma 4. Denote $\text{diag}(A, B)$ as the block diagonal matrices where the upper left and lower right block are matrx A and B, respectively. Notice that

$$
\bar{G}_n = \frac{\partial \bar{g}_n(\theta)}{\partial \theta} = \text{diag}\left(\frac{1}{k_0}\sum_{t=1}^{k_0} \frac{\partial f(x_t, \theta_1)}{\partial \theta_1}, \frac{1}{n-k_0}\sum_{t=k_0+1}^{n} \frac{\partial f(x_t, \theta_2)}{\partial \theta_2} \right).
$$

Define $\bar{f}_n^{(1)}(\theta_1) = \frac{1}{k_0}\sum_{t=1}^{k_0} f(x_t, \theta_1), f^{(1)}(\theta_1) = Ef(x_t, \theta_1)$ for $1 \leq t \leq k_0$ and

$\bar{f}_n^{(2)}(\theta_2) = \frac{1}{n-k_0}\sum_{t=k_0+1}^{n} f(x_t,\theta_2)$, $f^{(2)}(\theta_2) = Ef(x_t,\theta_2)$ for $k_0+1 \le t \le n$. It suffices to show that as $n \to \infty$,

$$\frac{\partial \bar{f}_n^{(1)}}{\partial \theta_1} \xrightarrow{a.s.} \frac{\partial f^{(1)}(\theta_1)}{\partial \theta_1} \quad \text{and} \quad \frac{\partial \bar{f}_n^{(2)}}{\partial \theta_2} \xrightarrow{a.s.} \frac{\partial f^{(2)}(\theta_2)}{\partial \theta_2}.$$

As the proofs for both terms are similar, we consider only the first one. Denote the (i,j)-th element of $\bar{f}_n^{(1)}(\theta_1)$ as

$$\bar{f}_{n,k}^{(1)} = \widehat{\text{cov}}(s_{t,i}^2, s_{t-\ell,j}^2) = \hat{E}s_{t,i}^2 s_{t-\ell,j}^2 - \hat{E}s_{t,i}^2 \hat{E}s_{t-\ell,j}^2, \tag{A.3}$$

To show the convergence $\frac{\partial \bar{f}_n^{(1)}(\theta_1)}{\partial \theta_1} \to \frac{\partial f^{(1)}(\theta_1)}{\partial \theta_1}$, it suffices to show it only for the first term in (A.3) since the proof for the second one is very similar. Defining $f_{n,i,j,t}^{(1)} = s_{t,i}^2 s_{t-\ell,j}^2$, then $f_{n,i,j}^{(1)} = \frac{1}{k_0}\sum_{t=1}^{k_0} f_{n,i,j,t}^{(1)}$. Notice that

$$\frac{\partial \bar{f}_{n,i,j,t}^{(1)}(\theta_1)}{\partial \theta_{1,a,b}} = \text{Tr}\left[\left(\frac{\partial \bar{f}_{n,i,j,t}^{(1)}(\theta_1)}{\partial W_{\theta_1}}\right)' \frac{\partial W_{\theta_1}}{\partial \theta_{1,a,b}}\right],$$

where

$$\begin{aligned}
\frac{\partial \bar{f}_{n,i,j,t}^{(1)}(\theta_1)}{\partial W_{\theta_1,i',j'}} &= s_{t,i}^2 \frac{\partial s_{t-\ell,j}^2}{\partial W_{\theta_1,i',j'}} + s_{t-\ell,j}^2 \frac{\partial s_{t,i}^2}{\partial W_{\theta_1,i',j'}} \\
&= s_{t,i}^2 \frac{\partial (\sum_{p=1}^d W_{\theta_1,j,p} x_{t-\ell,p})^2}{\partial W_{\theta_1,i',j'}} + s_{t-\ell,j}^2 \frac{\partial (\sum_{p=1}^d W_{\theta_1,i,p} x_{t,p})^2}{\partial W_{\theta_1,i',j'}} \\
&= 2s_{t,i}^2 s_{t-\ell,j'} x_{t-\ell,j'} \mathbf{1}(j=i') + 2s_{t-\ell,j}^2 s_{t,i} x_{t,j'} \mathbf{1}(i=i'),
\end{aligned}$$

and

$$\frac{\partial W_{\theta_1}}{\partial \theta_{1,a,b}} = Q_{1,2}(\theta_{1,1,2}) \cdots Q_{a-1,b}(\theta_{1,a-1,b}) \frac{\partial Q_{a,b}(\theta_{1,a,b})}{\partial \theta_{1,a,b}} Q_{a+1,b}(\theta_{1,a+1,b}) \cdots Q_{d-1,d}(\theta_{1,d-1,d}).$$

Obviously, $\frac{\partial \bar{f}^{(1)}_{n,i,j,t}(\theta_1)}{\partial \theta_{1,a,b}}$ is a measurable function of $\{s_t\}$. Also notice that $\{s_t\}$ is stationary and ergodic. Thus it follows from [49, TH 3.5.7] that $\{\frac{\partial \bar{f}^{(1)}_{n,i,j,t}(\theta_1)}{\partial W_{\theta_1,i',j'}}\}$ is stationary and ergodic.

Since $s^2_{t,i}s_{t-\ell,j}x_{t-\ell,j'} \leq \frac{1}{2}s^4_{t,i} + \frac{1}{2}(s_{t-\ell,j}x_{t-\ell,j'})^2 \leq \frac{1}{2}s^4_{t,i} + \frac{1}{4}s^4_{t-\ell,j} + \frac{1}{4}x^4_{t-\ell,j'}$, then,

$$E|s^2_{t,i}s_{t-\ell,j}x_{t-\ell,j'}| \leq \frac{1}{2}Es^4_{t,i} + \frac{1}{4}Es^4_{t-\ell,j} + \frac{1}{4}Ex^4_{t-\ell,j'} \overset{C2}{<} \infty.$$

Similarly, $E|s^2_{t-\ell,j}s_{t,i}x_{t,j'}| < \infty$ and thus $E||\frac{\partial \bar{f}^{(1)}_{n,i,j,t}(\theta_1)}{\partial W_{\theta_1,i',j'}}|| < \infty$. Finally, it follows from [49, TH 3.5.8] that

$$\frac{\partial \bar{f}^{(1)}_n(\theta_1)}{\partial \theta_1} \xrightarrow{\text{a.s.}} \frac{\partial f^{(1)}(\theta_1)}{\partial \theta_1}.$$

□

Proof of Lemma 5. Notice that

$$\begin{aligned}
\frac{\partial^2 f^{(1)}_{n,i,j,t}}{\partial \theta^2_{1,a,b}} &= \frac{\partial \sum_{i_1,j_1} \frac{\partial f^{(1)}_{n,i,j,t}(\theta_1)}{\partial W_{\theta_1,i_1,j_1}} \frac{\partial W_{\theta_1,i_1,j_1}}{\partial \theta_{1,a,b}}}{\partial \theta_{1,a,b}} \\
&= \sum_{i_1,j_1,i_2,j_2} \frac{\partial^2 f^{(1)}_{n,i,j,t}(\theta_1)}{\partial W_{\theta_1,i_1,j_1}\partial W_{\theta_1,i_2,j_2}} \frac{\partial W_{\theta_1,i_1,j_1}}{\partial \theta_{1,a,b}} \frac{\partial W_{\theta_1,i_2,j_2}}{\partial \theta_{1,a,b}} \\
&+ \sum_{i_1,j_1} \frac{\partial f^{(1)}_{n,i,j,t}(\theta_1)}{\partial W_{\theta_1,i_1,j_1}} \frac{\partial^2 W_{\theta_1,i_1,j_1}}{\partial \theta^2_{1,a,b}}.
\end{aligned}$$

Notice also that

$$\frac{\partial^2 f^{(1)}_{n,i,j,t}(\theta_1)}{\partial W_{\theta_1,i_1,j_1}\partial W_{\theta_1,i_2,j_2}} = 2\frac{\partial(s^2_{t-\ell,j}s_{t,i}x_{t,j}\mathbf{1}(i=i_1) + s^2_{t,i}s_{t-\ell,j}x_{t-\ell,j_1}\mathbf{1}(j=i_1))}{\partial W_{\theta_1,i_2,j_2}}.$$

WLOG, consider only the first term, which could be further written as,

$$\begin{aligned} & 4\frac{\partial s_{t-\ell,j}}{\partial W_{\theta_1,i_2,j_2}} s_{t-\ell,j} s_{t,i} x_{t,j_1} \mathbf{1}(i=i_1) + 2\frac{\partial s_{t,i}}{\partial W_{\theta_1,i_2,j_2}} s_{t-\ell,j}^2 x_{t,j_1} \mathbf{1}(i=i_1) \\ = \;& 4x_{t-\ell,j_2} s_{t-\ell,j} s_{t,i} x_{t,j_1} \mathbf{1}(i=i_1)\mathbf{1}(j=i_2) + 2x_{t,j_2} s_{t-\ell,j}^2 x_{t,j_1} \mathbf{1}(i=i_1)\mathbf{1}(i=i_2). \end{aligned}$$

Also we have,

$$\frac{\partial^2 W_{\theta_1}}{\partial \theta_{1,a,b}^2} = Q_{1,2}(\theta_{1,1,2}) \cdots Q_{a-1,b}(\theta_{1,a-1,b}) \frac{\partial^2 Q_{a,b}(\theta_{1,a,b})}{\partial \theta_{1,a,b}^2} \cdots Q_{d-1,d}(\theta_{1,d-1,d}).$$

Thus $\frac{\partial^2 f_{n,i,j,t}^{(1)}}{\partial \theta_{1,a,b}^2}$ is a measurable function of $\{s_t\}$. Notice that

$$E|x_{t-\ell,j_2} s_{t-\ell,j} s_{t,i} x_{t,j_1}| \overset{\textbf{Holder's Ineq}}{\leq} (E|x_{t-\ell,j_2}|^4 E|s_{t-\ell,j}|^4 E|s_{t,i}|^4 E|x_{t,j_1}|^4)^{1/4} \overset{C2}{<} \infty.$$

Similarly, $E|x_{t,j_2} s_{t-\ell,j}^2 x_{t,j_1}| < \infty$. Thus, $E|\frac{\partial^2 f_{n,i,j,t}^{(1)}(\theta_1)}{\partial W_{\theta_1,i_1,j_1} \partial W_{\theta_1,i_2,j_2}}| < \infty$ and also we have $E|\frac{\partial^2 f_{n,i,j,t}^{(1)}(\theta_1)}{\partial W_{\theta_1,i_1,j_1} \partial W_{\theta_1,i_2,j_2}}| < \infty$. The rest of the proof follows exactly from the proof of Lemma A4 in the supplementary material for [41]. □

Proof of Lemma 6. By applying the results in Lemmas 4-5, the proof follows exactly from the proof for Lemma A5 in the supplementary material of [41]. □

Proof of Theorem 3. [41, Theorem 1] verifies that $\hat{\boldsymbol{\theta}}_i \xrightarrow{a.s.} \boldsymbol{\theta}_{0i}$. Thus $\hat{\boldsymbol{\theta}} \xrightarrow{a.s.} \boldsymbol{\theta}_0$. □

Proof of Theorem 4. For clarity, we drop the subscript n in $\{\mathrm{s}_{n,t}\}$ in the proof below. Recall that $\bar{g}(\boldsymbol{\theta}) = (\bar{f}^{(1)}(\boldsymbol{\theta}_1)', \bar{f}^{(2)}(\boldsymbol{\theta}_2)')'$ and let

$$\Psi_n = \begin{pmatrix} \Phi_n & 0 \\ 0 & \Phi_n \end{pmatrix}.$$

Applying DOC to each segment to estimate $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ is equivalent to solving the

problem,

$$\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2) = \underset{(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)\in\boldsymbol{\Theta}\times\boldsymbol{\Theta}}{\operatorname{argmin}} (\mathcal{J}^{(1)}(\boldsymbol{\theta}_1) + \mathcal{J}^{(2)}(\boldsymbol{\theta}_2)) = \underset{(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)\in\boldsymbol{\Theta}\times\boldsymbol{\Theta}}{\operatorname{argmin}} \bar{g}(\boldsymbol{\theta})'\Psi_n\bar{g}(\boldsymbol{\theta}),$$

where $\mathcal{J}^{(i)}(\boldsymbol{\theta}_i)$ is the objective function for the i-th segment. Define

$$g_t(\boldsymbol{\theta}) = \begin{pmatrix} f(\boldsymbol{X}_t, \boldsymbol{\theta}_1) \\ 0 \end{pmatrix} \times \mathbf{1}\{1 \le t \le k_0(n)\} + \begin{pmatrix} 0 \\ f(\boldsymbol{X}_t, \boldsymbol{\theta}_2) \end{pmatrix} \times \mathbf{1}\{k_0(n) < t \le n\}.$$

Then $\bar{g}(\boldsymbol{\theta}) = \frac{1}{k_0}\sum_{t=1}^{k_0} g_t(\boldsymbol{\theta}) + \frac{1}{n-k_0}\sum_{t=k_0+1}^{n} g_t(\boldsymbol{\theta})$. By the mean value theorem, we have

$$\bar{g}(\hat{\boldsymbol{\theta}}) = \bar{g}(\boldsymbol{\theta}_0) + \bar{G}(\hat{\boldsymbol{\theta}}^*)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \tag{A.4}$$

where $\bar{G}(\boldsymbol{\theta}) := \frac{\partial\bar{g}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}^*$ is between $\boldsymbol{\theta}_0$ and $\hat{\boldsymbol{\theta}}$. Since $\bar{G}(\hat{\boldsymbol{\theta}})'\Psi_n\bar{g}(\hat{\boldsymbol{\theta}}) = 0$, multiplying both sizes of (A.4) by $\bar{G}(\hat{\boldsymbol{\theta}})'\Psi_n$ and rearranging the terms, we obtain

$$\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 = -(\bar{G}(\hat{\boldsymbol{\theta}})'\Psi_n\bar{G}(\hat{\boldsymbol{\theta}}^*))^{-1}\bar{G}(\hat{\boldsymbol{\theta}}^*)'\Psi_n\bar{g}(\boldsymbol{\theta}_0). \tag{A.5}$$

By Theorem 3, $\hat{\boldsymbol{\theta}}^* \xrightarrow{a.s.} \boldsymbol{\theta}_0$. By Lemma 6, we have $\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} |\bar{G}(\boldsymbol{\theta}) - G(\boldsymbol{\theta})| \xrightarrow{a.s.} 0$ with $G(\boldsymbol{\theta}) := E\{\frac{\partial}{\partial\boldsymbol{\theta}} g_t(\boldsymbol{\theta})\}$. Therefore we get

$$\begin{aligned} |\bar{G}(\hat{\boldsymbol{\theta}}^*) - G(\boldsymbol{\theta}_0)| \le & |\bar{G}(\hat{\boldsymbol{\theta}}^*) - G(\hat{\boldsymbol{\theta}}^*)| + |G(\hat{\boldsymbol{\theta}}^*) - G(\boldsymbol{\theta}_0)| \\ & \le \sup_{\boldsymbol{\theta}^*\in\boldsymbol{\Theta}} |\bar{G}(\boldsymbol{\theta}^*) - G(\boldsymbol{\theta}^*)| + |G(\hat{\boldsymbol{\theta}}^*) - G(\boldsymbol{\theta}_0)| \xrightarrow{a.s.} 0. \end{aligned}$$

Let $\Phi := \lim_{n\to\infty} \Phi_n$ and define Ψ by replacing Φ_n with Φ in Ψ_n. Then $\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0$ has the same asymptotic distribution as that of $-(G(\boldsymbol{\theta}_0)'\Psi G(\boldsymbol{\theta}_0))^{-1}G(\boldsymbol{\theta}_0)\Psi\bar{g}(\boldsymbol{\theta}_0)$. Denote

$B = \text{diag}(F^{(1)'}\Phi, F^{(2)'}\Phi)$ and $C = \text{diag}(F^{(1)'}\Phi F^{(1)}, F^{(2)'}\Phi F^{(2)})$. If $\sqrt{n}\bar{g}(\boldsymbol{\theta}_0)$ converges to $N(0, \mathbf{V}_0)$, then by (A.5),

$$(\sqrt{n}(\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_{01})', \sqrt{n}(\hat{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}_{02})')' \to C^{-1}B\mathbf{V}_0^{1/2}\mathbf{N}(0, \mathbf{I}).$$

Notice that $A_n \to \left[B\mathbf{V}_0B'\right]^{-1/2}C$ under Assumption C8. Then the conclusion follows from the Slutsky's theorem.

Below we prove that $\sqrt{n}\bar{g}(\boldsymbol{\theta}_0)$ converges to $N(0, \mathbf{V}_0)$. By the Wold-device, it is equivalent to show that for any $b = (b^{(1)'}, b^{(2)'})'$,

$$b^{(1)'}\frac{\sqrt{n}}{k_0(n)}\sum_{t=1}^{k_0(n)} f(\boldsymbol{X}_t, \boldsymbol{\theta}_{01}) + b^{(2)'}\frac{\sqrt{n}}{n - k_0(n)}\sum_{t=k_0(n)+1}^{n} f(\boldsymbol{X}_t, \boldsymbol{\theta}_{02}) \tag{A.6}$$

converges to a normal distribution. Recall that for any $p \in \{1, 2, \cdots, d\}$, $E\mathbf{s}_{t,p} = 0$ and $E\mathbf{s}_{t,p}^2 = 1$, where $\mathbf{s}_{t,p}$ denotes the pth component of $\mathbf{s}_t$. Define $\hat{E}\mathbf{s}_{\ell,k}^2 := \frac{1}{n-\ell}\sum_{t=\ell+1}^{n}\mathbf{s}_{t-\ell,k}^2$.

Then we have

$$\begin{aligned}
&\frac{\sqrt{n}}{k_0(n)}b^{(1)'}\sum_{t=1}^{k_0(n)} f(\boldsymbol{X}_t, \boldsymbol{\theta}_1) = \frac{\sqrt{n}}{k_0(n)}\sum_{t=\ell+1}^{k_0(n)}\sum_{(p,q,\ell)\in\mathcal{H}} b_{p,q,\ell}^{(1)}(\mathbf{s}_{t,p}^2\mathbf{s}_{t-\ell,q}^2 - \hat{E}\mathbf{s}_{0,p}^2\hat{E}\mathbf{s}_{\ell,q}^2) \\
= \;& \frac{\sqrt{n}}{k_0(n)}\sum_{t=\ell+1}^{k_0(n)}\sum_{(p,q,\ell)\in\mathcal{H}} b_{p,q,\ell}^{(1)}\{(\mathbf{s}_{t,p}^2\mathbf{s}_{t-\ell,q}^2 - 1) - (\mathbf{s}_{t,p}^2 - 1) - (\mathbf{s}_{t-\ell,q}^2 - 1) - 1\} \\
- \;& \sqrt{n}\sum_{(p,q,\ell)\in\mathcal{H}} b_{p,q,\ell}^{(1)}\{(\hat{E}\mathbf{s}_{0,p}^2\hat{E}\mathbf{s}_{\ell,q}^2 - 1) - (\hat{E}\mathbf{s}_{0,p}^2 - 1) - (\hat{E}\mathbf{s}_{\ell,q}^2 - 1) - 1\} \\
= \;& \sum_{(p,q,\ell)\in\mathcal{H}}\frac{n}{k_0(n)}b_{p,q,\ell}^{(1)}\frac{1}{\sqrt{n}}\sum_{t=\ell+1}^{k_0(n)}(1,1,1)(\mathbf{s}_{t,p}^2\mathbf{s}_{t-\ell,q}^2 - 1, \mathbf{s}_{t,p}^2 - 1, \mathbf{s}_{t-\ell,q}^2 - 1)' \\
- \;& \sqrt{n}\sum_{(p,q,\ell)\in\mathcal{H}} b_{p,q,\ell}^{(1)}(\hat{E}\mathbf{s}_{0,p}^2 - 1)(\hat{E}\mathbf{s}_{\ell,q}^2 - 1)
\end{aligned} \tag{A.7}$$

Let $\tilde{b}_{p,q,\ell}^{(1)} := \frac{n}{k_0(n)}b_{p,q,\ell}^{(1)}$, which is bounded by C5. Since $\sqrt{n}(\hat{E}\mathbf{s}_{0,p}^2 - 1) = O_{\mathcal{P}}(1)$ and

$(\hat{E}\mathbf{s}_{\ell,q}^2 - 1) = o_{\mathcal{P}}(1)$ under C6. Then $\sqrt{n}\sum_{(p,q,\ell)\in\mathcal{H}} b_{p,q,\ell}^{(1)}(\hat{E}\mathbf{s}_{0,p}^2 - 1)(\hat{E}\mathbf{s}_{\ell,q}^2 - 1) = o_{\mathcal{P}}(1)$ and (A.7) can be rewritten as,

$$(1,1,1)\frac{1}{\sqrt{n}}\sum_{t=\ell+1}^{k_0(n)}\sum_{(p,q,\ell)}\tilde{b}_{p,q,\ell}^{(1)}(\mathbf{s}_{t,p}^2\mathbf{s}_{t-\ell,q}^2 - 1, \mathbf{s}_{t,p}^2 - 1, \mathbf{s}_{t-\ell,q}^2 - 1)' + o(1).$$

Letting $\tilde{b}_{t,p,q,\ell}^{(2)} := \frac{n}{n-k_0(n)} b_{p,q,\ell}^{(2)}$, the second term of (A.6) can be similarly rewritten as,

$$(1,1,1)\frac{1}{\sqrt{n}}\sum_{t=k_0(n)+1}^{n}\sum_{(p,q,\ell)}\tilde{b}_{p,q,\ell}^{(2)}(\mathbf{s}_{t,p}^2\mathbf{s}_{t-\ell,q}^2 - 1, \mathbf{s}_{t,p}^2 - 1, \mathbf{s}_{t-\ell,q}^2 - 1)' + o(1).$$

Define $\boldsymbol{Z}_{n,t} = \sum_{(p,q,\ell)\in\mathcal{H}} \tilde{b}_{p,q,\ell}^{(1)}(\mathbf{s}_{t,p}^2\mathbf{s}_{t-\ell,q}^2 - 1, \mathbf{s}_{t,p}^2 - 1, \mathbf{s}_{t-\ell,q}^2 - 1)'$ if $t \le k_0(n)$ and $\boldsymbol{Z}_{n,t} = \sum_{(p,q,\ell)\in\mathcal{H}} \tilde{b}_{p,q,\ell}^{(2)}(\mathbf{s}_{t,p}^2\mathbf{s}_{t-\ell,q}^2 - 1, \mathbf{s}_{t,p}^2 - 1, \mathbf{s}_{t-\ell,q}^2 - 1)'$ if $t > k_0(n)$. Then (A.6) is equal to $(1,1,1)\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\boldsymbol{Z}_{n,t}$. By the continuous mapping theorem, it suffices to show that $\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\boldsymbol{Z}_{n,t}$ is asymptotically normal.

To this end, notice that $E|\boldsymbol{Z}_{n,t}|^r \le C\max_{t,p,q}\{E(\mathbf{s}_{t,p}^2\mathbf{s}_{t-\ell,q}^2)^r, E(\mathbf{s}_{t,p}^2)^r, 1\}$ for some positive constant C. Besides, $E(\mathbf{s}_{t,p}^2)^r < \infty$ and $E(\mathbf{s}_{t,p}^{2r}\mathbf{s}_{t-\ell,q}^{2r}) \le \{E(\mathbf{s}_{t,p}^{4r})E(\mathbf{s}_{t-\ell,q}^{4r})\}^{1/2} < \infty$, where we have used C6 and the Cauchy-Schwarz inequality. Thus, $\{\boldsymbol{Z}_{n,t}\}$ has finite r-th moment. By C6, $\{\mathbf{s}_{n,t}\}$ is L_2-NED on $\{U_{n,t}\}$. Applying Lemma 3, $\{\boldsymbol{Z}_{n,t}\}$ is also L_2-NED on $\{U_{n,t}\}$ with the same size as that for $\{\mathbf{s}_{t,p}\}$. Therefore, by [22, Corollary 1], $\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\boldsymbol{Z}_{n,t}$ converges to a normal distribution and the proof is thus completed. □