STATISTICAL INFERENCE OF MULTIVARIATE TIME SERIES AND FUNCTIONAL DATA USING NEW DEPENDENCE METRICS

BY

CHUNG EUN LEE

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Statistics
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2017

Urbana, Illinois

Doctoral Committee:

Professor Xiaofeng Shao, Chair
Professor Douglas Simpson
Associate Professor Bo Li
Assistant Professor Xiaohui Chen

Abstract

In this thesis, we focus on inference problems for time series and functional data and develop new methodologies by using new dependence metrics which can be viewed as an extension of Martingale Difference Divergence (MDD^2) [see Shao and Zhang (2014)] that quantifies the conditional mean dependence of two random vectors. For one part, the new approaches to dimension reduction of multivariate time series for conditional mean and conditional variance are proposed by applying new metrics, the so-called Martingale Difference Divergence Matrix (MDDM), Volatility Martingale Difference Divergence (VMDDM), and vec Volatility Martingale Difference Divergence (vecVMDDM). The metrics involve less user-chosen quantities and their computation and associated inference are less computationally expensive than some existing ones. Therefore, the new approaches are relatively simple to implement and computationally convenient. Also, the new methods outperform the existing methods in the presence of strong nonlinear dependence. For the other part, we propose a nonparametric conditional mean independence test for a response variable Y given a covariate variable X, both of which can be function-valued or vector-valued. The test is built upon Functional Martingale Difference Divergence (FMDD) which fully measures the conditional mean independence of Y on X. One distinct feature of our test is that it does not use any dimension reduction techniques or user-chosen parameters and is model free. The proposed test is shown to have higher power than some existing tests in theory, and favorable size and power properties in numerical simulations.

Acknowledgements

I would like to express my sincere appreciation to my advisor, Professor Xiaofeng Shao, for his tremendous guidance, enthusiasm, and patience. His wise advice and enormous support always allowed me to grow as a better researcher and person which I deeply appreciate. Above all, I am grateful to observe his respectable attitude toward research and education which become an invaluable learning for me.

I also send my gratitude to Professor Douglas Simpson, Bo Li, and Xiaohui Chen for their time, insightful comments, and for being my doctoral thesis committee. I would like to thank Professor John Marden for several discussions and writing me the recommendation letters. I further thank Professor Xianyang Zhang for his collaboration and helpful advice which enabled me to substantially improve my research.

It has been a great pleasure to be a part of the Statistics Department of the University of Illinois at Urbana-Champaign. I truly appreciate the friendly and supportive environment that I have received from all faculty, colleagues, and staff members in the department. This has made my life in graduate school more comfortable and constructive.

I would like to thank my parents for giving me courage and for their unconditional love, continuous encouragement, and belief in me. I am so blessed to have them as my parents. Because of them, I made it this far.

Without any of the above, this work would not have been possible. I am very grateful for all opportunities I have been afforded from the University of Illinois at Urbana-Champaign.

Table of Contents

Chapter 1

Introduction

Dimension reduction is a critical step for modeling large dimensional time series $Y_t \in R^p$ since the number of parameters involved in the model grows dramatically as the dimension of the data increases. The key consideration of dimension reduction is how to effectively reduce the dimension of the time series that matters in modeling the time series dynamics while losing least amount of information. In Chapter 2, we introduce a new methodology to perform dimension reduction for a stationary multivariate time series. Our method is motivated by the consideration of optimal prediction and focuses on the reduction of the effective dimension in conditional mean of time series given the past information. In particular, we seek a contemporaneous linear transformation such that the transformed time series has two parts with one part being conditionally mean independent of the past. To achieve this goal, we first propose MDDM, which can quantify the conditional mean independence of $V \in R^p$ given $U \in R^q$ and also encodes the number and form of linear combinations of V that are conditional mean independent of U. Our dimension reduction procedure is based on eigen-decomposition of the cumulative martingale difference divergence matrix, which is an extension of MDDM to the time series context. Interestingly, there is a static factor model representation for our dimension reduction framework and it has subtle difference from the existing static factor model used in the time series literature. Some theory is also provided about the rate of convergence of eigenvalue and eigenvector of the sample cumulative MDDM in the fixed-dimensional setting. Favorable finite sample performance is demonstrated via simulations and real data illustrations in comparison with some existing methods.

In the inference of econometric and finanical time series, it is vital to have a good estimation of volatility matrix. In Chapter 3, we propose VMDDM to quantify the conditional variance dependence of a random vector $Y \in R^p$ given $X \in R^q$, building on recent work on martigale difference divergence matrix that measures the conditional mean dependence. We further generalize VMDDM to the time series context and apply it to do dimension reduction for multivariate volatility, following the recent work by Hu and Tsay (2014) and Li, Gao, Li and Yao (2016). However, unlike the latter two papers, our metric is easy to compute, can fully capture nonlinear serial dependence and involves less user-chosen numbers. Furthermore, we propose a variant of VMDDM and apply it to the estimation of conditional uncorrelated components model [Fan, Wang and Yao (2008)]. Simulation and data illustration show that our method performs well in comparison with the existing ones, and can outperform others in cases of strong nonlinear dependence.

Functional data analysis (FDA) is becoming an important subarea in statistics due to the fact that many real data are in the forms of curves and images. For a response variable Y and a covariate variable X, which can be function-valued or vector-valued, it is a fundamental problem to assess the conditional mean independence of Y on X, i.e., $H_0 : E[Y|X] = E[Y]$ *a.s.* If the null is true, then there is no need to do regression modeling when the interest is on the conditional mean. In Chapter 4, we propose a new nonparametric conditional mean independence test for a response variable Y and a predictor variable X where either or both can be function-valued. Our test is built on a new metric, FMDD, which fully characterizes the conditional mean dependence of Y given X and extends the MDD proposed in Shao and Zhang (2014). We define the unbiased estimator of FMDD by using a $\mathcal{U}$-centering approach, and obtain its limiting null distribution under mild assumptions. Since the limiting null distribution is not pivotal, we adopt the wild bootstrap method to estimate the critical value and show its consistency. It turns out that our test can detect the local alternatives which approach the null at the rate of $n^{-1/2}$ with nontrivial power, where n is the sample size. Unlike the recent two tests developed by Kokoszka et al (2008) and Patilea et al. (2016), our test do not require finite dimensional projection and linear model assumption or the choice of tuning parameters. Promising finite sample performance is demonstrated via simulations in comparison with the above two tests.

Chapter 2

Dimension Reduction for Stationary Multivariate Time Series

2.1 Background

A central problem in the modeling and inference of multivariate time series is the reduction of dimensionality of parameters. In the time domain, several dimension reduction methods have been proposed, including the canonical correlation analysis of Box and Tiao (1977), the factor models of Pe$\tilde{n}$a and Box (1987), the scalar component analysis of Tiao and Tsay (1989), the independent component analysis of Back and Weigend (1997), the principal component analysis of Stock and Watson (2002), and the dynamic orthogonal component analysis of Matteson and Tsay (2011). In these works, linear combinations are sought to make linearly transformed series have simpler dynamic structure, which can be captured by parsimonious parametric models. In the spectral domain, dimension reduction methods have been developed by Geweke

(1977), Brillinger (1981), Stoffer (1999), Ombao, von Sachs and Guo (2005), Eichler, Motta and von Sachs (2011), among others.

In this paper, we propose a new methodology to perform dimension reduction for a strictly stationary multivariate time series. Our proposal is motivated by the consideration of optimal prediction. Let $Y_t \in R^p, t \in Z$ be a mean zero p-variate stationary time series, then the optimal predictor of Y_{n+1} given the past information set $\mathcal{F}_n = \sigma(Y_n, \cdots, Y_1, \cdots)$ is $E(Y_{n+1}|\mathcal{F}_n)$ in the mean squared error sense. This led us to focus on the dimension reduction of $E(Y_{n+1}|\mathcal{F}_n)$, which we intend to do in a way without imposing any parametric or linear structure. In particular, we seek for a contemporaneous linear (invertible) transformation for Y_t, say, $M \in R^{p\times p}$, such that $MY_t = Z_t = [Z_{1t}^T, Z_{2t}^T]^T$, where $Z_{1t} \in R^s$ and $Z_{2t} \in R^{p-s}$, such that $E(Z_{1(n+1)}|\mathcal{F}_n) \neq E(Z_{1(n+1)})$ and $E(Z_{2(n+1)}|\mathcal{F}_n) = E(Z_{2(n+1)})$. In other words, the transformed series can be separated into two parts with one part being conditionally mean dependent on the past and the other part being conditionally mean independent upon the past. Thus, the modeling task for the whole series Y_t is reduced to that for the lower dimensional series Z_{1t}, since

$$E(Y_{n+1}|\mathcal{F}_n) = M^{-1} \begin{pmatrix} E(Z_{1(n+1)}|\mathcal{F}_n) \\ E(Z_{2(n+1)}) \end{pmatrix},$$

and dimension reduction can be achieved without loss of prediction accuracy.

Interestingly, our new method can be formulated equivalently in a factor model framework. Representing multiple time series in terms of several static or dynamic factors is quite popular and the literature is large, see Peña and Box (1987), Forni, Hallin, Lippi and Reichlin (2000, 2005), Bai and Ng (2002), Bai (2003), Stock and Watson (2005), Pan and Yao (2008), Lam, Yao and Bathia (2011), Lam and Yao

(2012), among others. A distinction from the static factor models in the existing literature is that our error process, i.e., $e_t = Y_t - E(Y_t|\mathcal{F}_{t-1})$ is a vector martingale difference sequence, which is stronger than the usual vector white noise assumption. This implies that the effective number of factors under our model could be different from (more precisely, is equal to or larger than) the number of factors in the factor models described in Peña and Box (1987) and Pan and Yao (2008). A more detailed discussion of the difference is provided in Section 2.4.2.

To quantify conditional mean (in)dependence for a multivariate time series, we extend the notion of Martingale Difference Divergence (MDD) recently proposed by Shao and Zhang (2014), which is used to measure the conditional mean dependence of a univariate response Y with respect to a vector covariate X, in several aspects. First we consider multivariate response variable, and generalize MDD to a matrix-valued quantity called MDDM (Martingale Difference Divergence Matrix). Second we define the cumulative MDDM by taking the sum of MDDM at several lags to account for the underlying time series structure, either jointly or in a pairwise fashion. In order to determine the number and the form of linear combinations that are conditional mean independent of the past, we perform the eigen-decomposition of the sample cumulative MDDM and use ratio-based estimator, as adopted in Lam, Yao and Bathia (2011), Lam and Yao (2012). Note that the inference in the latter work is based on a linear analogue of cumulative MDDM, which only measures linear dependence. Since nonGaussian and nonlinear time series are prevalent in various applied areas, our methodology has a built-in advantage over the ones that rely on linear dependence measures for the dimension reduction of multivariate nonlinear time series.

The rest of the paper is organized as follows. Section 2.2 provides a review of martingale difference divergence and its sample estimate. In Section 2.3, we introduce the definition of martingale difference divergence matrix and its properties. Our dimension reduction methodology for conditional mean is presented in Section 2.4, which includes an extension of principal component analysis to principal conditional mean component analysis, and factor model representation as well as some discussion of practical issues and related work. Simulation results are gathered in Section 4.4.1. Section 2.6 presents two real data illustrations and Section 2.5 concludes. Technical details are included in Appendix.

A word on notation. Let $i = \sqrt{-1}$ be the imaginary unit. For $x \in \mathbf{C}^p$, we use x^* for "x-conjugate-transpose" (conjugate for scalars). The scalar product of vectors x and y is denoted by $< x, y >$. For a complex-valued function $f(\cdot)$, the complex conjugate of f is denoted by f^* and $|f|^2 = ff^*$. Denote the Euclidean norm of $x = (x_1, \cdots, x_p) \in \mathbf{C}^p$ as $|x|_p$, where $|x|_p^2 = x_1x_1^* + \cdots + x_px_p^*$, and if $x = (x_1, \cdots, x_p) \in R^p$, it is sometimes denoted as $\|x\|$, where $\|x\|^2 = x_1^2 + \cdots x_p^2$. For a square matrix A, spectral norm of A is denoted as $\|A\|_2$, where $\|A\|_2 = \sqrt{\lambda_{max}(A^TA)}$ and Frobenius norm of A is denoted by $\|A\|_F$, where $\|A\|_F = \sqrt{tr(A^TA)}$ and $tr(A) = \sum_{i=1}^p A_{i,i}$. A random vector $X \in \mathcal{L}^s$ if $E|X|_p^s < \infty$.

2.2 Review of Martingale Difference Divergence

For U $\in R^q$ and V$\in R$, where q is a fixed positive integer, Shao and Zhang (2014) proposed the so-called martingale difference divergence (MDD) and its standardized version martingale difference correlation (MDC) to measure the conditional mean independence of V on U, i.e.,

$$E(V|U) = E(V), \text{ almost surely.} \tag{2.2.1}$$

Specifically MDD$(V|U)$ is defined as the nonnegative number that satisfies

$$MDD(V|U)^2 \ = \ \frac{1}{c_q}\int_{R^q}\frac{|g_{V,U}(s)-g_V g_U(s)|^2}{|s|_q^{1+q}}ds, \tag{2.2.2}$$

where $g_{V,U}(s) = E(Ve^{i<s,U>})$, $g_V = E(V)$, $g_U(s) = E(e^{i<s,U>})$ and $c_q = \pi^{(1+q)/2}/\Gamma((1+q)/2)$. The definition can be regarded as an extension of distance covariance [Székeley, Rizzo and Bakirov (2007)] since a similar weighting function is used and it inherits many desirable properties of distance covariance. For example, MDD$(V|U)^2 = 0$ if and only if (4.2.1) holds. Furthermore, if $E(|V|^2 + |U|_q^2) < \infty$, then

$$MDD(V|U)^2 = -E[(V-E(V))(V^{'}-E(V^{'}))|U-U^{'}|_q],$$

where $(V^{'}, U^{'})$ is an independent copy of (V, U).

Recently, Park, Shao and Yao (2015) made an extension of MDD to allow multivariate response. If $V \in R^p$, $p \geq 1$, then the characteristic function based definition (4.2.2) still applies. Under the assumption that $E(|V|_p^2 + |U|_q^2) < \infty$, Park et al. showed

$$MDD(V|U)^2 = -E[(V-E(V))^T(V^{'}-E(V^{'}))|U-U^{'}|_q], \tag{2.2.3}$$

which is a scalar-valued quantity. Most of the properties mentioned in Shao and Zhang (2014) still hold for this more general definition.

Assume that we have a random sample $(U_k, V_k)_{k=1}^n$ from the joint distribution of (U, V). Let $\overline{V_n} = \frac{1}{n}\sum_{k=1}^n V_k$, $a_{kl} = V_k V_l$, $\overline{a_{k\cdot}} = \frac{1}{n}\sum_{l=1}^n a_{kl} = V_k\overline{V_n}$, $\overline{a_{\cdot l}} = \frac{1}{n}\sum_{k=1}^n a_{kl} = \overline{V_n}V_l$, $\overline{a_{\cdot\cdot}} = \frac{1}{n^2}\sum_{k,l=1}^n a_{kl} = \overline{V_n}\ \overline{V_n}$ and $A_{kl} = a_{kl} - \overline{a_{k\cdot}} - \overline{a_{\cdot l}} + \overline{a}_{\cdot\cdot} = (V_k - \overline{V_n})(V_l - \overline{V_n})$ for $k, l = 1, \cdots, n$. Similarly, let $b_{kl} = |U_k - U_l|_q$, $\overline{b_{k\cdot}} = \frac{1}{n}\sum_{l=1}^n b_{kl}$, $\overline{b_{\cdot l}} = \frac{1}{n}\sum_{k=1}^n b_{kl}$, $\overline{b_{\cdot\cdot}} = \frac{1}{n^2}\sum_{k,l=1}^n b_{kl}$ and $B_{kl} = b_{kl} - \overline{b_{k\cdot}} - \overline{b_{\cdot l}} + \overline{b_{\cdot\cdot}}$, for $k, l = 1, \cdots, n$. Based on the above quantities, sample martingale difference divergence MDD_n [Shao and Zhang (2014)] is defined as the nonnegative number that satisfies

$$MDD_n(V|U)^2 \;\; = \;\; -\frac{1}{n^2}\sum_{k,l=1}^{n} A_{kl}B_{kl} = \frac{1}{c_q}\int_{R^q}\frac{|g_{V,U}^n(s) - g_V^n g_U^n(s)|^2}{|s|_q^{1+q}}ds,$$

where $g_{V,U}^n(s) = \frac{1}{n}\sum_j V_j e^{i<s,U_j>}$, $g_V^n = \frac{1}{n}\sum_j V_j$ and $g_U^n(s) = \frac{1}{n}\sum_j e^{i<s,U_j>}$. The second equality in the above equation is shown in Theorem 2 of Shao and Zhang (2014) and it implies that the simpler algebraic form is equivalent to an empirical plug-in version. The above definition applies to the case $p = 1$. When $p > 1$, the sample MDD is defined as the nonnegative number that satisfies

$$MDD_n(V|U)^2 \;\; = \;\; -\frac{1}{n^2}\sum_{k,l=1}^{n}(V_k - \overline{V}_n)^T(V_l - \overline{V}_n)B_{kl}. \tag{2.2.4}$$

It turns out that the above definition can be further simplified as

$$MDD_n(V|U)^2 = -\frac{1}{n^2}\sum_{k,l=1}^{n}(V_k - \overline{V}_n)^T(V_l - \overline{V}_n)|U_k - U_l|_q,$$

which can be shown by a straightforward calculation, and the details are omitted. Note that in general $MDD_n(V|U)^2$ is a biased estimator of $MDD(V|U)^2$, however the bias is expected to be asymptotically negligible when p is fixed. It is indeed possible to adopt the U-centering idea [Székeley and Rizzo (2014), Park et al. (2015)] to define an unbiased estimator, but it unfortunately complicates the asymptotic analysis in Section 2.4.

2.3 Martingale Difference Divergence Matrix

For $U \in R^q$ and $V \in R^p$, it is possible that there exists a linear combination of V, say $\alpha \in R^p$, such that $E(\alpha^T V|U) = E(\alpha^T V)$ although V is not necessarily conditionally mean independent of U (i.e., $E(V|U) \neq E(V)$). In the presence of such a relationship, the modeling of conditional mean of V as a function of U can be simplified, as the effective dimension of $E(V|U)$ can be reduced via linear transformation and separating out the part that is conditionally mean independent of U. To this end, we introduce a new matrix object, the so-called martingale difference divergence matrix (MDDM), which can be viewed as an extension of martingale difference divergence from a scalar to a matrix.

Definition **2.3.1**. *Martingale Difference Divergence Matrix*
Given $V = (V_1, \cdots, V_p)^T \in R^p, U \in R^q$,

$$MDDM(V|U) \quad = \quad \frac{1}{c_q}\int_{R^q} \frac{G(s)G(s)^*}{|s|_q^{1+q}} ds,$$

where $G(s) = cov(V, e^{i<s,U>}) = (G_1(s), \cdots, G_p(s))^T$ *for* $s \in R^q$, $G_j(s) = cov(V_j, e^{i<s,U>})$.
Note that the (i,i)th entry of the $p \times p$ matrix MDDM equals to $MDD(V_i|U)^2$, whereas the (i,j)th entry is

$$MDDM(V|U)_{ij} = \frac{1}{c_q}\int_{R^q} \frac{G_i(s)G_j(s)^*}{|s|_q^{1+q}} ds = MDDM(V|U)_{ji}^*$$

i.e., MDDM$(V|U)$ is a hermitian matrix and thus has real eigenvalues. Here for the notational simplicity, we opt to use the notation $MDDM$ instead of $MDDM^2$ in our definition. Below we provide a simple and equivalent expression for MDDM.

Lemma **2.3.1**. *If $E(|V|_p^2 + |U|_q^2) < \infty$, then*

$$MDDM(V|U) = -E[(V - E(V))(V^{'} - E(V^{'}))^T |U - U^{'}|_q],$$

where $(V^{'}, U^{'})$ is an iid copy of (V, U). Therefore MDDM($V|U$) is a real, symmetric and positive semidefinite matrix.

Lemma 2.3.1 implies that MDDM($V|U$) is a $p \times p$ matrix with the (i, j)th entry equal to $\text{MDDM}_{i,j}(V|U) = -E[(V_i - E(V_i))(V_j^{'} - E(V_j^{'}))^T |U - U^{'}|_q]$, provided that $E(|V|_p^2 + |U|_q^2) < \infty$. Since $G_j(s) = 0$, $\forall s \Leftrightarrow E(V_j|U) = E(V_j)$, we have that $MDDM_{i,j}(V|U) = 0$, provided that $E(V_j|U) = E(V_j)$ or $E(V_i|U) = E(V_i)$. It is also worth noting that $tr(MDDM(V|U)) = MDD(V|U)^2$ in (2.2.3).

By elementary matrix algebra, it is not difficult to show that Lemma 2.3.1 implies the following theorem, which states that the rank of the MDDM is closely tied to the number of linear combinations of V that are conditionally mean independent of U.

Theorem **2.3.1**. *For $V \in R^p$ and $U \in R^q$, if $E(|V|_p^2 + |U|_q^2) < \infty$, then for any real $p \times s$ matrix D, $MDDM(D^T V|U) = D^T MDDM(V|U)D$; Subsequently, there exist $p - s$ linearly independent combinations of V such that they are conditionally mean independent of U, if and only if rank$(MDDM(V|U)) = s$.*

Remark **2.3.1**. We shall provide a discussion on a possible analogue of MDDM to measure the linear dependence between two vectors $V \in R^p$ and $U \in R^q$. Define

$$L(V|U) = \text{cov}(V, U)\text{cov}(V, U)^T,$$

where cov(V, U) is a $p \times q$ matrix with the (i, j)th entry being cov(V_i, U_j). It is easy to show that $L(V|U)$ is a real, symmetric and positive semidefinite matrix. Then there exists a nonzero $\alpha \in R^p$, such that cov$(\alpha^T V, U) = 0$ (i.e., a linear combination of V is uncorrelated with U), if and only if $L(V|U)$ is singular. Further, suppose the number of linearly independent combinations of V that are uncorrelated with U is $p-r$, then $r = \text{rank}(L(V|U))$. Since conditional mean independence implies uncorrelatedness, it is not difficult to show that $\text{rank}(MDDM(V|U)) \geq \text{rank}(L(V|U))$.

Given a random sample $(U_k, V_k)_{k=1}^n$ from the joint distribution of (U, V), sample martingale difference divergence matrix MDDM_n can be defined as

$$MDDM_n(V|U) = -\frac{1}{n^2}\sum_{h,l=1}^{n}(V_h - \overline{V}_n)(V_l - \overline{V}_n)^T|U_h - U_l|_q.$$

2.4 Dimension Reduction for Conditional Mean

As we mentioned in Section 2.1, our goal is to seek linear transformation of Y_t such that linear transformed series can be separated into two parts with one part being conditionally mean independent of the past. Mathematically, we look for linear combinations of Y_t, say $\alpha^T Y_t$, that are conditionally mean independent of $\mathcal{F}_{t-1} = \sigma(Y_{t-1}, Y_{t-2}, \cdots)$. As we only have a finite stretch of observations from the process $Y_t, t \in Z$, we shall approximate the conditional mean independence of $\alpha^T Y_t$ on $\mathcal{F}_{t-1}$ by that on $\mathcal{F}_{t-1,t-k_0} = \sigma(Y_{t-1}, \cdots, Y_{t-k_0})$, where k_0 is a pre-specified fixed integer. This practice is quite common in time series analysis, and it is consistent with the notion that for weakly dependent time series the main dependence is at short lags. The approximation can be in fact supported by certain time series models. For example, if the time series model is $VAR(k_0)$, then the conditional distribution of Y_t given $\mathcal{F}_{t-1}$ is identical to the conditional distribution of Y_t given $\mathcal{F}_{t-1,t-k_0}$, thus there is no loss of information in this approximation. In the sequel, we define the so-called cumulative MDDM to quantify the conditional mean independence of Y_t on its recent past $\mathcal{F}_{t-1,t-k_0}$.

Definition **2.4.1**. *The cumulative MDDM is defined as*

$$\Gamma_{k_0} \;=\; \Gamma_{k_0}^{(1)} := \sum_{j=1}^{k_0} MDDM(Y_t|Y_{t-j}). \tag{2.4.1}$$

Since $MDDM$ is a positive semidefinite matrix, Γ_{k_0} is also a positive semidefinite matrix. Note that $MDDM(Y_t|Y_{t-j})$ depends on the time lag j but not on t due to strict stationarity. The sample estimate of Γ_{k_0} is given by $\widehat{\Gamma}_{k_0} = \sum_{j=1}^{k0} MDDM_n(Y_t|Y_{t-j})$.

Remark **2.4.1**. Our definition follows the common practice in time series analysis, where the cumulative contribution from various lags are added up in a pairwise fashion. Alternatively, we could adopt a joint approach, i.e., we can define

$$\Gamma_{k_0}^{(2)} = MDDM(Y_t|(Y_{t-1}^T, \cdots, Y_{t-k_0}^T)^T)$$

and its sample estimate as $\widehat{\Gamma}_{k_0}^{(2)} = MDDM_n(Y_t|(Y_{t-1}^T, \cdots, Y_{t-k_0}^T)^T)$. For a given k_0, it seems difficult to know whether inference based on $\widehat{\Gamma}_{k_0}^{(2)}$ is preferred for the given series at hand. We shall compare the finite sample performance for inferences based on $\widehat{\Gamma}_{k_0}^{(j)}$, $j = 1, 2$ in simulation studies.

2.4.1 Principal Conditional Mean Components (PCMC)

As outlined in Section 2.1, dimension reduction for conditional mean can be achieved once we identify the number and the form of linear combinations of Y_t that are conditionally mean independent of the past. It turns out that such information is encoded in Γ_{k_0} (or $\Gamma_{k_0}^{(2)}$); see Theorem 2.3.1. Inspired by the work of Hu and Tsay (2014), who proposed the concept of principal volatility component analysis, we shall introduce the so-called principal conditional mean component analysis.

Since Γ_{k_0} is a real symmetric positive semidefinite matrix, its eigenvalues $\{\lambda_j\}_{j=1}^p$ are either zero or positive. We shall assume that

(C1), $\lambda_1 > \cdots > \lambda_2 > \cdots > \lambda_s > 0 = \lambda_{s+1} = \cdots = \lambda_p$.

Let γ_j be an orthonormal eigenvector of Γ_{k_0} corresponding to the eigenvalue λ_j. Then we have

$$\gamma_j^T \Gamma_{k_0} \gamma_j = \lambda_j, \ j = 1, \cdots, p.$$

If we let $\mathbf{M} = [\gamma_1, \cdots, \gamma_p]$ and $\Lambda = diag(\lambda_1 > \cdots \geq \lambda_p)$, then $\Gamma_{k_0}\mathbf{M} = \mathbf{M}\Lambda$ by spectral decomposition of Γ_{k_0}. Therefore the rank of Γ_{k_0} is s, which means that there exist $p-s$ linearly independent combinations $(\gamma_{s+1}, \cdots \gamma_p)$ which make $MDD(\gamma_i^T Y_t | Y_{t-j})^2 = 0$, $j = 1, \cdots, k_0$, $i = s+1, \cdots, p$. Since all these linear combinations live in the null space of Γ_{k_0}, they can be readily estimated based on eigen-decomposition of $\widehat{\Gamma}_{k_0}$.

Let $(\widehat{\lambda}_j, \widehat{\gamma}_j)_{j=1}^p$ be the p pairs of eigenvalues and eigenvectors of $\widehat{\Gamma}_{k_0}$, where $\widehat{\lambda}_1 \geq \widehat{\lambda}_2 \geq \cdots \geq \widehat{\lambda}_p$ and the eigenvectors $\{\widehat{\gamma}_j\}_{j=1}^p$ are orthonormal. To estimate s, the rank of Γ_{k_0}, we adopt a ratio-based estimator following the practice of Lam, Yao and Bathia (2011), Lam and Yao (2012), i.e.,

$$\widehat{s} = argmin_{1 \leq i \leq R} \frac{\widehat{\lambda}_{i+1}}{\widehat{\lambda}_i} \tag{2.4.2}$$

where R is an integer satisfying $s \leq R < p$. There are other methods developed for the estimation of s, see e.g., Bai and Ng (2002, 2007) and Hallin and Liška (2007) for information criteria based approach, Bathia, Yao and Ziegelmann (2010) for the bootstrap approach, and Hu and Tsay (2014) for a sequential testing approach. As mentioned in Lam and Yao (2012), the ratio-based method can be viewed as an enhanced scree test [Cattell (1966)], and it is very easy to implement. We use the ratio-based estimator in part because of the connection of our dimension reduction framework with the ones in Lam, Yao and Bathia (2011) and Lam and Yao (2012); see Section 2.4.2 for details. It is also worth noting that Lam and Yao (2012) showed that the ratio-based estimator can still work even when p is large and grows to infinity in the asymptotics.

For $j = 1, \cdots, p$, we can then estimate γ_j by $\widehat{\gamma}_j$. Since $\widehat{\gamma}_j$ might be replaced by $-\widehat{\gamma}_j$, the results stated below concerning a comparison of eigenvectors implicitly assume that signs have been chosen appropriately. Below we further impose suitable moment and weak dependence assumptions on Y_t.

(C2) Let $(Y_t)_{t\in N}$ be a strictly stationary and β-mixing process. Assume that there exist $\delta > 0$ such that $E[|Y_t|^{6+3\delta}] < \infty$. For a $\delta^{'} \in (0, \delta) : \beta(n) = O(n^{-\frac{2+\delta^{'}}{\delta^{'}}})$.

(C2') Let $(Y_t)_{t\in N}$ be a strictly stationary and m-dependent process. Assume that $E|Y_t|^6 < \infty$.

Theorem **2.4.1**. *Let conditions (C1)-(C2) hold. Then as $n \to \infty$, it holds that (i) $\widehat{\lambda}_j - \lambda_j = O_p(n^{-1/2})$ and $\|\widehat{\gamma}_j - \gamma_j\| = O_p(n^{-1/2})$ for $j = 1, \cdots, s$. Under the conditions (C1)-(C2'), we have that (ii) $\widehat{\lambda}_j = O_p(n^{-1})$ for $j = s + 1, \cdots, p$.*

Theorem 2.4.1 is analogous to Proposition 1 in Lam and Yao (2012), but is stated for our cumulative MDDM. It suggests that the empirical eigenvalues and eigenvectors obtained by the eigen-decomposition of sample cumulative MDDM are indeed reasonable estimators of their population counterparts for large sample size. The above theory is developed for the fixed p case, and theory for the growing p setting seems very challenging and is left for future research. Nevertheless, we examine the finite sample performance of the ratio-based estimator in the large p setting in simulation studies.

2.4.2 Factor Model Representation

In this subsection, we shall provide a factor model representation for our dimension reduction framework. Our static factor model is closely related to the one used in Peña and Box (1987), Pan and Yao (2008), Lam, Yao and Bathia (2011), as well as Lam and Yao (2012). We provide a brief review of the latter first.

Factor Model with White Noise Error. (FM-WNE)

$$Y_t = AX_t + \epsilon_t,$$

where Y_t is a $p \times 1$ observed vector of time series, X_t is a $r \times$ 1 latent factor time series which is usually assumed to be stationary and not a white noise, A is a $p \times r$ constant factor loading matrix, $r \leq p$ is the number of factors, $\{\epsilon_t\}$ is a $p \times 1$ white noise sequence with mean zero.

It is important to note that matrix A is not uniquely identified. For instance, if we replace A and X_t by AQ and $Q^{-1}X_t$ for any invertible matrix Q, we still get the same Y_t. Let B be a $p \times (p-r)$ matrix for which (A, B) forms a $p \times p$ orthogonal matrix. Following the practice in Pan and Yao (2008), Lam, Yao and Bathia (2011), Lam and Yao (2012), we assume that

Assumption **2.4.1**. *$A^TA = I_r$, $A^TB = 0$ and $B^TB = I_{p-r}$.*

Even with the above assumption, the matrix A is not unique, but the factor loading space $\mathcal{M}(A)$, which is the linear space spanned by the columns of A, is uniquely defined. Note that $\mathcal{M}(B)$ can be interpreted as the null space of the factor loading space $\mathcal{M}(A)$. The main inferential task is to estimate r, the number of factors, and the factor loading space $\mathcal{M}(A)$ or A. Once an estimator of A, say $\widehat{A}$ is obtained, it is natural to estimate the factor process X_t by $\widehat{X}_t = \widehat{A}^T Y_t$ and the residuals are $\widehat{\epsilon}_t = (I_p - \widehat{A}\widehat{A}^T)Y_t$. Then the next step is to fit a parsimonious model to $\widehat{X}_t$, which may be achieved by rotating $\widehat{X}_t$ appropriately, or equivalently modeling $H^T \widehat{X}_t$, where H is an $r \times r$ orthogonal matrix.

Based on the above model assumptions, we derive that

$$\begin{aligned} B^T Y_t &= B^T A X_t + B^T \epsilon_t = B^T \epsilon_t & (2.4.3) \\ A^T Y_t &= A^T A X_t + A^T \epsilon_t = X_t + A^T \epsilon_t, & (2.4.4) \end{aligned}$$

where $\{B^T \epsilon_t\}$ and $\{A^T \epsilon_t\}$ are both white noise sequences. This implies that certain linear combinations of Y_t (i.e., those corresponding to $B^T Y_t$) are white noise. Assuming that the cross-correlations between X_t and ϵ_t are zero at all lags, we can derive that $\text{cov}(Y_{t+k}, Y_t) = A\text{cov}(X_{t+k}, X_t)A^T$, for $k = 1, 2, \cdots$, thus the columns of B are the orthonormal eigenvectors of $\text{cov}(Y_{t+k}, Y_t)$ corresponding to zero eigenvalues. This observation led them to focus on the following matrix

$$\mathcal{L}_{k_0} = \sum_{j=1}^{k_0} \text{cov}(Y_{t+j}, Y_t)\text{cov}(Y_{t+j}, Y_t)^T = \sum_{j=1}^{k_0} L(Y_t | Y_{t-j}) \qquad (2.4.5)$$

and their estimation of r and $\mathcal{M}(A)$ is based on the eigen-analysis of sample estimate of the positive semidefinite matrix $\mathcal{L}_{k_0}$. It is interesting to note that the matrix $\mathcal{L}_{k_0}$ they defined is basically a linear counterpart of our cumulative MDDM Γ_{k_0}; see Remark 2.3.1. Thus the main difference between the two is that $\mathcal{L}_{k_0}$ encodes the information about the number and form of linear combinations of Y_t that are uncorrelated with $Y_{t-1}, \cdots, Y_{t-k_0}$ in a pairwise fashion, whereas Γ_{k_0} encodes the number and form of linear combinations of Y_t that are conditionally mean independent of $Y_{t-1}, \cdots, Y_{t-k_0}$ in a pairwise fashion.

In time series analysis, $\mathrm{cov}(Y_{t+j}, Y_t)$ is used to measure the linear dependence at lag j, whereas $MDDM(Y_{t+j}|Y_t)$ is used to measure conditional mean dependence at lag j. If the time series is Gaussian, then the second order property fully characterizes the joint distribution and autocovariances at all lags are sufficient to characterize the joint dependence. However, for non-Gaussian and nonlinear time series, the second order property may not be sufficient to characterize the serial dependence, which has motivated the development of various nonlinear dependence measures in the literature; see Hong (1999), Granger, Maasoumi and Racine (2004) and Zhou (2012) among others. Most of these nonlinear dependent measures are however scalar-valued, and do not seem directly useful for dimension reduction.

Factor Model with Martingale Difference Error. (FM-MDE)

$$Y_t = E(Y_t|\mathcal{F}_{t-1}) + \eta_t, \tag{2.4.6}$$

where $\eta_t = Y_t - E(Y_t|\mathcal{F}_{t-1})$ is a martingale difference sequence by construction. Assume that $E(Y_t|\mathcal{F}_{t-1}) = \mathcal{A}Z_t$, where $\mathcal{A}$ is the $p \times s$ $(s \leq p)$ factor loading matrix and Z_t is the s-dimensional latent factor process.

Similar to the factor model with white noise error, only the factor loading space $\mathcal{M}(\mathcal{A})$ is uniquely defined. Note that $\mathcal{M}(\mathcal{A})$ is the column space of $\mathbf{M} = (\gamma_1, \cdots, \gamma_s)$. For the convenience of discussion, we assume that there is a $p \times (p-s)$ matrix $\mathcal{B}$, such that

Assumption **2.4.2**. *$\mathcal{A}^T\mathcal{A} = I_s$, $\mathcal{A}^T\mathcal{B} = 0$ and $\mathcal{B}^T\mathcal{B} = I_{p-s}$.*

Following the argument in the derivation of (2.4.3), we have that

$$\mathcal{B}^T Y_t = \mathcal{B}^T E(Y_t|\mathcal{F}_{t-1}) + \mathcal{B}^T\eta_t = \mathcal{B}^T\mathcal{A}Z_t + \mathcal{B}^T\eta_t = \mathcal{B}^T\eta_t, \tag{2.4.7}$$

$$\mathcal{A}^T Y_t = \mathcal{A}^T E(Y_t|\mathcal{F}_{t-1}) + \mathcal{A}^T\eta_t = \mathcal{A}^T\mathcal{A}Z_t + \mathcal{A}^T\eta_t = Z_t + \mathcal{A}^T\eta_t, \tag{2.4.8}$$

where $\{\mathcal{B}^T\eta_t\}$ and $\{\mathcal{A}^T\eta_t\}$ are martingale difference sequences. Based on (2.4.8), it is easy to see that $Z_t = E[\mathcal{A}^T Y_t|\mathcal{F}_{t-1}]$. Suppose we can obtain good estimates of $\mathcal{A}$ and $\mathcal{B}$ (or the corresponding column spaces), say $\widehat{\mathcal{A}}$ and $\widehat{\mathcal{B}}$, then we can estimate Z_t by $\widehat{Z}_t = \widehat{\mathcal{A}}^T Y_t$ and the residuals are $\widehat{\eta}_t = (I_p - \widehat{\mathcal{A}}\widehat{\mathcal{A}}^T)Y_t$. A lower dimensional model can be fitted to $\{\widehat{Z}_t\}$ so dimension reduction is achieved.

As the two factor models (FM-WNE and FM-MDE) appear to have the same form, it pays to mention their differences. On one hand, our latent factor process Z_t is measurable with respect to $\mathcal{F}_{t-1}$, and its contemporary linear combination $\mathcal{A}Z_t$ is the conditional mean of Y_t given the past information by definition. Since $E(Y_t|\mathcal{F}_{t-1})$ has the interpretation of being the optimal predictor of Y_t given $\mathcal{F}_{t-1}$ (in the mean squared error sense), our dimension reduction is well motivated by the consideration of optimal prediction. By contrast, the process X_t in FM-WNE is not necessarily measurable with respect to $\mathcal{F}_{t-1}$ and AX_t may not be the conditional mean. On the other hand, the estimation methods are different for these two factor models. Under the FM-WNE, we seek to find contemporary linear combinations (i.e., B) that make the linear transformed sequence $B^T Y_t$ a white noise sequence, whereas under the FM-MDE, contemporary linear transformations (i.e., $\mathcal{B}$) are sought to make $\mathcal{B}^T Y_t$ a martingale difference sequence; compare (2.4.3) and (2.4.8). Due to different requirements on the error sequence, the matrix objects that encode the information about the dimension of latent factor process and the factor loading space are different. Under the FM-WNE, we take advantage of the assumptions on the second order property of (X_t, ϵ_t), and the inference is based on the cumulative linear matrix L_{k_0}, which encodes the linear dependence, whereas under the FM-MDE, we naturally focus on cumulative MDDM Γ_{k_0}, which characterizes conditional mean independence. To shed some light on the difference, we consider the following state space model.

Table 2.1: Factor model representations for the state space model

		Example 2.4.1	
FM-WNE	$A = D_1$	$X_t = W_t$	$\epsilon_t = \epsilon_{1t}$
FM-MDE	$\mathcal{A} = D_1$	$Z_t = E(W_t \vert \mathcal{F}_{t-1})$	$\eta_t = Y_t - D_1 E(W_t \vert \mathcal{F}_{t-1})$
		Example 2.4.2	
FM-WNE	$A = A_1$	$X_t = W_{1t}$	$\epsilon_t = [A_1, A_2] W_{2t} + \epsilon_{1t}$
FM-MDE	$\mathcal{A} = [A_1, A_2]$	$Z_t = (E(W_{1t} + W_{3t} \vert \mathcal{F}_{t-1})^T$ $, E(W_{4t} \vert \mathcal{F}_{t-1})^T)^T$	$\eta_t = Y_t - A_1 E(W_{1t} \vert \mathcal{F}_{t-1})$ $-[A_1, A_2] E(W_{2t} \vert \mathcal{F}_{t-1})$

Example **2.4.1**. Let $Y_t = D_1 W_t + \epsilon_{1t}$, where Y_t is $p \times 1$ time series, D_1 is a $p \times r$ constant factor loading matrix and $\{\epsilon_{1t}\}$ are iid mean zero error process. Let $W_t = h(\epsilon_{2t}, \epsilon_{2(t-1)}, \cdots)$, $t \in Z$ be a r-dimensional nonlinear stationary causal process, where $\{\epsilon_{2t}\}_{t \in Z}$ is the r-dimensional mean zero iid innovation process that is independent of the p-dimensional error process $(\epsilon_{1t})_{t \in Z}$. Assume that W_t is not a white noise sequence. Note that several models used in simulation studies of Lam, Yao and Bathia (2011) and Lam and Yao (2012) fall into the above framework. Table 2.1 shows the detailed representation under the two factor models. Although the latent processes under the two models (i.e., X_t and Z_t) are different, it is easy to see that $r = s$ and $\mathcal{M}(A) = \mathcal{M}(\mathcal{A})$, i.e., the two factor loading spaces are identical. It would be interesting to see which inference method (i.e., the one based on $\mathcal{L}_{k_0}$ versus the one based on Γ_{k_0}) delivers a better estimate of the factor loading space in this case and we shall address this question in our simulations.

In general, a white noise sequence is not necessarily a martingale difference sequence but a martingale difference sequence has to be a white noise sequence under finite second moment assumption. This fact implies that for a stationary time series Y_t that admits both representations (i.e., FM-WNE and FM-MDE), the two could coincide as demonstrated in the following proposition.

Proposition **2.4.1**. *Suppose that Assumptions (2.4.1) and (2.4.2) hold. If $(\epsilon_t, \mathcal{F}_t)$ is a martingale difference sequence, then we have $\mathcal{M}(A) = \mathcal{M}(\mathcal{A})$.*

In some cases, s can be strictly larger than r, as shown in the following example.

EXAMPLE **2.4.2**. Let $Y_t = A_1 W_{1t} + [A_1, A_2] W_{2t} + \epsilon_{1t}$, where A_1 is a $p \times r$ matrix and A_2 is a $p \times (q-r)$ matrix. Set $p > q > r$. We assume that (i), $A_1^T A_1 = I_r$, $A_1^T A_2 = 0$ and $A_2^T A_2 = I_{q-r}$; (ii), W_{1t} is a r-dimensional stationary causal process as defined in Example 2.4.1 and is not a white noise sequence, W_{2t} is a q-dimensional vector white noise sequence but not martingale difference sequence, and ϵ_{1t} are iid mean zero. (iii), The three processes $\{W_{1t}\}$, $\{W_{2t}\}$ and $\{\epsilon_{1t}\}$ are mutually independent.

Let $W_{2t} = (W_{3t}^T, W_{4t}^T)^T$, where W_{3t} is of dimension r and W_{4t} is of dimension $q-r$. Then the model can be reformulated as

$$Y_t = [A_1, A_2] \begin{pmatrix} (W_{1t} + W_{3t}) \\ W_{4t} \end{pmatrix} + \epsilon_{1t}.$$

It is easy to see that under the framework of FM-WNE, $[A_1, A_2] W_{2t} + \epsilon_{1t}$ is a vector white noise so $\mathcal{M}(A) = \mathcal{M}(A_1)$, whereas under the framework of FM-MDE, $E(Y_t | \mathcal{F}_{t-1}) = A_1 E(W_{1t} | \mathcal{F}_{t-1}) + [A_1, A_2] E(W_{2t} | \mathcal{F}_{t-1})$. Then $s = q > r$ and $\mathcal{M}(\mathcal{A}) = \mathcal{M}([A_1, A_2])$; see Table 2.1. In the univariate case, examples for white noise but not martingale difference can be found in Shao (2011). We shall examine the performance of our dimension reduction method for this example in Section 6.

2.4.3 Related Work and Practical Issues

As pointed out by a referee, our work is to some extent related to Park, Sriram and Yin (2009, 2010), who have extended the sufficient dimension reduction framework from random sample to the time series setting. Specifically, the latter authors focused on the univariate time series and considered the estimation of central subspace [Park et al. (2010)] and central mean subspace [Park et al. (2009)] of the conditional distribution of Y_t given $(Y_{t-1}, \cdots, Y_{t-d})$, where d is assumed to be fixed and possibly unknown. For the central subspace, they estimated it by minimizing Kullback-Leibler distance and for the central mean subspace, they used a variant of MAVE (minimum average variance estimation), which was proposed by Xia et al. (2002) and shown to be applicable to time series data. While the work by Park et al. (2009, 2010) mainly focuses on the dimension reduction of covariates, which are naturally defined as the lagged observations $(Y_{t-1}, \cdots, Y_{t-d})$ in the time series setting, our work focuses on the dimension reduction of the multivariate response Y_t, and thus are quite different in terms of the goal and the setting. In particular, (1) the parameter Park et al. targets is the column space associated with the central subspace or central mean subspace, whereas we want to estimate the space spanned by linear combinations that make the response conditional mean independent of the past information. In a sense, our dimension reduction is closer in spirit to the recently developed envelop models by Cook, Li and Chiaromonte (2000), which also remove the redundant linear combinations of the response that are not related to covariates. But the latter was done under a linear model and for random sample, whereas we do not have any parametric/linear assumptions and our reduction is formulated in a time series setting; (2) our dimension reduction is based on spectral decomposition of a sample matrix, and no smoothing is involved; whereas nonparametric estimation and smoothing is required in Park et al. (2009, 2010) since the targeted space is different.

In practical implementation, we need to come up with a choice of k_0. In theory, k_0 can be chosen as the smallest positive integer that makes

$$E(Y_t|\mathcal{F}_{t-1}) = E(Y_t|\mathcal{F}_{t-1,t-k_0}),$$

i.e., given $(Y_{t-1}, \cdots, Y_{t-k_0})$, Y_t is conditionally mean independent of $(Y_{t-k_0-1}, Y_{t-k_0-2}, \cdots)$.

Thus the determination of k_0 itself is a nontrivial task. If the series follows a vector autoregressive model, then partial autocorrelation function provides an indication about the magnitude of k_0. Alternatively, we can first look at the lag j sample MDD, i.e., $MDD_n(Y_t|Y_{t-j})^2$, accompanied by the standard error estimated from, say, the moving block bootstrap, and choose k_0 as the largest j such that jth MDD is still significant from zero. For time series that exhibits seasonal dependence patterns, we often want to let k_0 to be an integer multiple of the period (see Section 2.6.2) to capture the conditional mean dependence at seasonal lags. We leave a more careful and data-driven choice of k_0 and their impact to dimension reduction to future work.

An additional practical and methodological issue is that after we obtain the rank of Γ_{k_0}, say by $\widehat{s}$, and the null space of Γ_{k_0}, by $M_1 = [\widehat{\gamma}_{\widehat{s}+1}, \cdots, \widehat{\gamma}_p]$, it would be useful to verify that the transformed series $M_1 Y_t$ are indeed conditionally mean independent of $\mathcal{F}_{t-1}$ (or $\mathcal{F}_{t-1,t-k_0}$ in a pairwise fashion). To this end, one can look at the magnitude of $H_n = \sum_{j=1}^{k_0} MDD_n^2(M_1 Y_t|Y_{t-j})$ and perform a significance test. Under the null that the transformed series are conditionally mean independent of the past, H_n is expected to be small, and its significance can presumably be assessed by using a block bootstrap approach. We shall also leave a rigorous investigation of this topic to future study.

2.5 Numerical Simulations

In this section, we examine the finite sample performance of our MDDM-based dimension reduction approach via simulations. In particular, we compare with the method in Lam and Yao (2012), which is based on the linear dependence metric $\mathcal{L}_{k_0}$ (see (2.4.5)). In our simulations, we tried $\Gamma_{k_0}^{(j)}$, $j = 1, 2$ for several k_0s to assess the sensitivity of our dimension reduction method with respect to the choice of k_0 and cumulative MDDM employed. Even though our theory is developed for the fixed p case, we also investigate the finite sample performance for the large p case by Monte-Carlo experiments.

Recall that for both methods, two steps are involved in the estimation. The first step corresponds to the estimation of the true number of factors, i.e., r or s using ratio-based estimator (see (2.4.2)), where we set $R = p - 1$, $R = [p/2]$ or $[p/3]$ for small p and large p setting respectively. The second step refers to the estimation of the factor loading space, i.e., $\mathcal{M}(\mathcal{A})$ (or $\mathcal{M}(A)$) once s (or r) is estimated. This can be achieved by performing principal component analysis on the sample cumulative MDDM as described in Section 2.4.1. For each example, we replicate the simulation 1000 times and use the following criteria to measure the accuracy of our dimension reduction method.

- D-distance ($D_1(\cdot, \cdot)$) [Pan and Yao (2008)]

$$D_1(\mathcal{A}, \hat{\mathcal{A}}) = [\{tr(\hat{\mathcal{A}}^T(I_p - \mathcal{A}\mathcal{A}^T)\hat{\mathcal{A}}) + tr(\hat{\mathcal{B}}^T\mathcal{A}\mathcal{A}^T\hat{\mathcal{B}})\}/p]^{1/2},$$

  where $\hat{\mathcal{B}}$ is a basis of an orthogonal complement of the column space spanned by $\hat{\mathcal{A}}$. $D_1(\mathcal{A}, \hat{\mathcal{A}})$ is used to measure the distance between $\mathcal{M}(\mathcal{A})$ and $\mathcal{M}(\hat{\mathcal{A}})$. Note that under Assumption 2.4.2, $\mathcal{A}\mathcal{A}^T$ is a projection matrix onto the linear space $\mathcal{M}(\mathcal{A})$ and $D_1(\mathcal{A}, \hat{\mathcal{A}}) \in [0, 1]$. $D_1(\mathcal{A}, \hat{\mathcal{A}}) = 0$ if and only if $\mathcal{M}(\mathcal{A}) = \mathcal{M}(\hat{\mathcal{A}})$, and $D_1(\mathcal{A}, \hat{\mathcal{A}}) = 1$ if and only if $\mathcal{M}(\mathcal{A}) = \mathcal{M}(\hat{\mathcal{B}})$.

- Root Mean Squared Error (RMSE) [Lam, Yao and Bathia (2011)]

$$RMSE = (\sum_{t=1}^{n} \frac{\|\widehat{\mathcal{A}}\widehat{X}_t - \mathcal{A}X_t\|^2}{pn})^{1/2},$$

  which measures the overall closeness of the estimated signal $\widehat{\mathcal{A}}\widehat{X}_t$ to the true signal $\mathcal{A}X_t$. Smaller value of RMSE indicates more accurate estimation of underlying factor series.

We shall investigate the following examples.

Example **2.5.1**. Example 2.5.1 is adopted from the simulation study of Pan and Yao (2008) with slight modification so the model falls into the framework of Example 2.4.1. We define the factor series $X_t = (x_{1t}, x_{2t}, x_{3t})^T$ as

$$x_{1,t} = 0.8x_{1,t-1} + e_{1,t}, x_{2,t} = e_{2,t} + 0.9e_{2,t-1} + 0.3e_{2,t-2}, x_{3,t} = -0.5x_{3,t-1} - e_{3,t} + 0.8e_{3,t-1}$$

where $e_{i,t}$, $i = 1, 2, 3$ are all iid standard normal variables. The observed data $Y_t = (Y_{1,t}, \cdots, Y_{p,t})^T$ is defined by

$$Y_{i,t} = \begin{cases} x_{i,t} + \epsilon_{i,t} & \text{for } i = 1, 2, 3 \\ \epsilon_{i,t} & \text{for } i = 4, \cdots p \end{cases}$$

where $\epsilon_{i,t}$, $i = 1, 2, \cdots, p$ are iid standard normal variables and independent from $\{e_{j,t}\}, j = 1, 2, 3$. We consider several different combinations of (p, n, k_0), i.e., $p = 5, 10, 20$, $n = 300, 600, 1000$ and $k_0 = 1, 3$. For the above data generating process, the true number of factors r and s are 3 and the factor loading matrix, $A = \mathcal{A} = (I_3, 0_{p-3})^T$. Note that when $k_0 = 1$, Γ_{k_0} and $\Gamma_{k_0}^{(2)}$ become the same so some results are identical in Table 2.2.

Table 2.2: Mean, standard error (*in the bracket*) of D-distance and $\widehat{r}, \widehat{s}$ with 1000 replicates for Example 2.5.1

	$\mathcal{L}_{k_0}$				Γ_{k_0}				$\Gamma_{k_0}^{(2)}$			
	$D(\hat{\mathcal{A}}, \mathcal{A})$	$\widehat{r}$			$D(\hat{\mathcal{A}}, \mathcal{A})$	$\widehat{s}$			$D(\hat{\mathcal{A}}, \mathcal{A})$	$\widehat{s}$		
		$\widehat{r} < 3$	$\widehat{r} = 3$	$\widehat{r} > 3$		$\widehat{s} < 3$	$\widehat{s} = 3$	$\widehat{s} > 3$		$\widehat{s} < 3$	$\widehat{s} = 3$	$\widehat{s} > 3$
$k_0 = 1$												
$p = 5, n = 300$	0.171 (0.15)	0	0.794	0.206	0.099 (0.05)	0.009	0.991	0	0.099 (0.05)	0.009	0.991	0
$p = 5, n = 600$	0.116 (0.13)	0	0.872	0.128	0.067 (0.02)	0	1	0	0.067 (0.02)	0	1	0
$p = 5, n = 1000$	0.088 (0.12)	0	0.909	0.091	0.052 (0.02)	0	1	0	0.052 (0.02)	0	1	0
$p = 10, n = 300$	0.415 (0.32)	0.011	0.549	0.440	0.135 (0.05)	0.042	0.958	0	0.135 (0.05)	0.042	0.958	0
$p = 10, n = 600$	0.287 (0.31)	0	0.713	0.287	0.090 (0.02)	0.001	0.999	0	0.090 (0.02)	0.001	0.999	0
$p = 10, n = 1000$	0.240 (0.30)	0	0.76	0.24	0.071 (0.01)	0	1	0	0.071 (0.01)	0	1	0
$p = 20, n = 300$	0.624 (0.36)	0.022	0.335	0.643	0.146 (0.04)	0.132	0.868	0	0.146 (0.04)	0.132	0.868	0
$p = 20, n = 600$	0.479 (0.40)	0	0.521	0.479	0.098 (0.02)	0.01	0.99	0	0.098 (0.02)	0.01	0.99	0
$p = 20, n = 1000$	0.362 (0.39)	0	0.652	0.348	0.076 (0.01)	0	1	0	0.076 (0.01)	0	1	0
$k_0 = 3$												
$p = 5, n = 300$	0.132 (0.10)	0.078	0.914	0.008	0.142 (0.12)	0.129	0.871	0	0.139 (0.12)	0.124	0.876	0
$p = 5, n = 600$	0.071 (0.03)	0.001	0.996	0.003	0.072 (0.04)	0.01	0.99	0	0.071 (0.04)	0.01	0.99	0
$p = 5, n = 1000$	0.055 (0.02)	0	0.999	0.001	0.054 (0.02)	0	1	0	0.054 (0.02)	0	1	0
$p = 10, n = 300$	0.188 (0.09)	0.281	0.719	0	0.184 (0.10)	0.295	0.705	0	0.181 (0.10)	0.283	0.717	0
$p = 10, n = 600$	0.105 (0.05)	0.041	0.959	0	0.107 (0.06)	0.071	0.929	0	0.107 (0.06)	0.071	0.929	0
$p = 10, n = 1000$	0.075 (0.02)	0.005	0.995	0	0.075 (0.03)	0.014	0.986	0	0.075 (0.03)	0.013	0.987	0
$p = 20, n = 300$	0.195 (0.06)	0.523	0.477	0	0.185 (0.06)	0.491	0.509	0	0.186 (0.07)	0.489	0.510	0.001
$p = 20, n = 600$	0.132 (0.06)	0.225	0.775	0	0.131 (0.06)	0.26	0.74	0	0.130 (0.06)	0.249	0.751	0
$p = 20, n = 1000$	0.089 (0.04)	0.059	0.941	0	0.090 (0.04)	0.085	0.915	0	0.090 (0.04)	0.088	0.912	0

It appears from Table 2.2 that when p increases or n decreases, the ability of correctly identifying the true number of factors diminishes and the D-distance gets larger for all methods. It might be expected that the method based on $\mathcal{L}_{k_0}$ performs the best since Y_t is generated from Gaussian linear time series and all dependence of Y_t can be effectively captured by autocovariance matrices. However, when $k_0 = 1$, our MDDM-based approach is superior to Lam and Yao's $\mathcal{L}_{k_0}$-based counterpart in terms of the probability of correctly estimating the true number of factors and D-distance. For $k_0 = 3$, the performance of the $\mathcal{L}_{k_0}$-based one and MDDM-based one is similar. It is interesting to note that when k_0 increases from 1 to 3, the performance of $\mathcal{L}_{k_0}$-based method improves substantially, showing its sensitivity with respect to the choice of k_0, whereas for Γ_{k_0} (or $\Gamma_{k_0}^{(2)}$), the performance is quite stable with respect to k_0.

Example **2.5.2**. In this example, the linear ARMA model for X_t in Example 2.5.1 is replaced by a nonlinear model, where $X_t = (x_{1,t}, x_{2,t}, x_{3,t})^T$ is defined as

$$x_{1,t} = -(0.9e^{-0.2x_{1,t-1}^2})x_{1,t-1} + e_{1,t}, x_{2,t} = (0.5e^{-0.4x_{2,t-1}^2} + 0.4)x_{2,t-1} + e_{2,t},$$

$$x_{3,t} = (0.1e^{-x_{3,t-1}^2} + 0.7)x_{3,t-1} + e_{3,t}$$

Then the data $Y_t = \mathcal{A}X_t + \epsilon_t$, where $\mathcal{A} = (I_3, 0_{p-3})^T$, $\epsilon_{i,t}, e_{i,t}$ are iid standard normal and independent from each other. Like Example 2.5.1, we consider $n = 300, 600, 1000$, $p = 5, 10, 20$ and $k_0 = 1, 3$. According to Theorem 5.1 and Example 5.1 in Shao and Wu (2007), X_t admits a stationary solution and the model falls into the framework in Example 2.4.1. For this example, the true number of factors r and s are still 3 and $\mathcal{M}(A) = \mathcal{M}(\mathcal{A})$.

Table 2.3: Mean, standard error (*in the bracket*) of $D-$distance and $\widehat{r}, \widehat{s}$ with 1000 replicates for Example 2.5.2

	$\mathcal{L}_{k_0}$				Γ_{k_0}				$\Gamma_{k_0}^{(2)}$			
	$D(\hat{A}, A)$	$\widehat{r}$			$D(\hat{A}, A)$	$\widehat{s}$			$D(\hat{A}, A)$	$\widehat{s}$		
		$\widehat{r} < 3$	$\widehat{r} = 3$	$\widehat{r} > 3$		$\widehat{s} < 3$	$\widehat{s} = 3$	$\widehat{s} > 3$		$\widehat{s} < 3$	$\widehat{s} = 3$	$\widehat{s} > 3$
$k_0 = 1$												
$p = 5, n = 300$	0.277 (0.15)	0.006	0.629	0.365	0.178 (0.11)	0.058	0.941	0.001	0.178 (0.11)	0.058	0.941	0.001
$p = 5, n = 600$	0.198 (0.15)	0	0.749	0.251	0.109 (0.04)	0	1	0	0.109 (0.04)	0	1	0
$p = 5, n = 1000$	0.16 (0.15)	0	0.806	0.194	0.087 (0.03)	0	1	0	0.087 (0.03)	0	1	0
$p = 10, n = 300$	0.608 (0.25)	0.033	0.268	0.699	0.238 (0.10)	0.182	0.818	0	0.238 (0.10)	0.182	0.818	0
$p = 10, n = 600$	0.497 (0.31)	0.001	0.442	0.557	0.151 (0.05)	0.018	0.982	0	0.151 (0.05)	0.018	0.982	0
$p = 10, n = 1000$	0.391 (0.32)	0	0.586	0.414	0.117 (0.02)	0	1	0	0.117 (0.02)	0	1	0
$p = 20, n = 300$	0.828 (0.19)	0.043	0.054	0.903	0.245 (0.07)	0.417	0.583	0	0.245 (0.07)	0.417	0.583	0
$p = 20, n = 600$	0.736 (0.3)	0.18	0.198	0.784	0.172 (0.06)	0.133	0.867	0	0.172 (0.06)	0.133	0.867	0
$p = 20, n = 1000$	0.62 (0.37)	0	0.355	0.645	0.126 (0.03)	0.011	0.989	0	0.126 (0.03)	0.011	0.989	0
$k_0 = 3$												
$p = 5, n = 300$	0.243 (0.18)	0.159	0.798	0.043	0.262 (0.21)	0.258	0.742	0	0.258 (0.21)	0.25	0.75	0
$p = 5, n = 600$	0.123 (0.08)	0.014	0.974	0.012	0.131 (0.12)	0.053	0.947	0	0.131 (0.12)	0.053	0.947	0
$p = 5, n = 1000$	0.089 (0.03)	0	1	0	0.085 (0.03)	0	1	0	0.085 (0.03)	0	1	0
$p = 10, n = 300$	0.316 (0.13)	0.478	0.52	0.002	0.301 (0.14)	0.466	0.534	0	0.306 (0.14)	0.481	0.519	0
$p = 10, n = 600$	0.197 (0.11)	0.168	0.832	0	0.2 (0.13)	0.216	0.784	0	0.202 (0.13)	0.218	0.782	0
$p = 10, n = 1000$	0.124 (0.05)	0.019	0.981	0	0.124 (0.07)	0.041	0.959	0.0	0.123 (0.07)	0.039	0.961	0
$p = 20, n = 300$	0.285 (0.07)	0.691	0.306	0.003	0.267 (0.07)	0.633	0.367	0	0.267 (0.08)	0.628	0.368	0.004
$p = 20, n = 600$	0.224 (0.08)	0.441	0.559	0	0.215 (0.09)	0.433	0.567	0	0.216 (0.09)	0.441	0.559	0
$p = 20, n = 1000$	0.157 (0.07)	0.163	0.837	0	0.158 (0.08)	0.209	0.791	0	0.158 (0.08)	0.207	0.793	0

From Table 2.3, we can see that when k_0 is 1, our method outperforms the $\mathcal{L}_{k_0}$-based one with smaller D-distance and higher proportion of estimating the number of factors correctly. If k_0 is 3, the performance of all methods are fairly comparable in terms of estimating the true number of factors and the factor loading matrix. Overall, the finding is similar to that in Example 2.5.1.

Example **2.5.3**. This example is from Lam, Yao and Bathia (2011) and it addresses the large p case, where p can exceed n. More specifically, three different factors $X_t = (x_{1t}, x_{2t}, x_{3t})^T$ are generated by the following model,

$$x_{1t} = w_t, x_{2t} = w_{t-1}, x_{3t} = w_{t-2}, w_t = 0.2z_{t-1} + z_t, z_t \sim N(0,1)$$

For the factor loading matrix $\mathcal{A}$, first $p/2$ elements of each column are iid U(-2, 2) and are kept fixed once generated and the other elements are set to be 0. The data Y_t is defined as $Y_t = \mathcal{A}X_t + \epsilon_t$ where ϵ_t is a random sample of $N(0, \Sigma)$ and is independent of X_t, where $\Sigma = (\sigma_{i,j})_{i,j=1}^p$ is defined as

$$\sigma_{i,j} = \begin{cases} \frac{1}{2}[(|i-j|+1)^{2H} - 2|i-j|^{2H} + (|i-j|-1)^{2H}],\ H = 0.9, & \text{if } i \neq j \\ 1 & \text{if } i = j \end{cases}$$

We consider $(p, n) = (100, 100), (100, 200), (400, 200)$ with $k_0 = 1$ and 5. Again the true number of factors r and s are 3 and the model falls into the framework in Example 2.4.1. When $p = 100$, R is set to be $p/2$ to estimate the number of factors and $R = [p/3]$ is used when $p = 400$ (see (2.4.2)). According to Table 2.4, the performance of $\mathcal{L}_{k_0}$-based and Γ_{k_0}-based methods are very much comparable for $k_0 = 1$ and 5. It appears that when $p = 400$ and $n = 200$, none of the methods succeed, since the proportion of estimating the true number of factors correctly is low. This phenomenon might be due to the use of ratio-based estimator. More sophisticated method of estimating the number of factors in the large-p setting has been developed in Li, Wang and Yao (2017) recently.

Example **2.5.4**. In this example, we replace the data generating process of w_t in

Example 2.5.3 by a nonlinear one as follows.

$$w_t = \begin{cases} 0.5 + (0.05e^{-0.01w_{t-1}^2} + 0.9)w_{t-1} + z_t, & \text{if } w_{t-1} < 5 \\ (0.9e^{-10w_{t-1}^2})w_{t-1} + z_t & \text{if } w_{t-1} \geq 5 \end{cases}, \ z_t \sim N(0,1)$$

Furthermore, Y_t is defined as $Y_t = \mathcal{A}X_t + \epsilon_t$, where error ϵ_t is generated from $N(0, 0.25\Sigma)$ and Σ is defined in Example 2.5.3. Other parts of the model, such as $\mathcal{A}$ are exactly the same as Example 2.5.3, along with the combinations of (p, n, k_0). From Table 2.4, our Γ_{k_0}-based approach outperforms $\mathcal{L}_{k_0}$-based counterpart in all cases and under both criteria with the advantage more pronounced when $k_0 = 1$. When $k_0 = 5$, $\Gamma_{k_0}^{(2)}$-based approach is slightly inferior to Γ_{k_0}-based counterpart, but is still superior to $\mathcal{L}_{k_0}$-based one, especially for the case $(p, n) = (400, 200)$. We speculate that part of the reason $\mathcal{L}_{k_0}$-based approach does not perform well is that it only captures linear auto-dependence. In the presence of strong nonlinearity in the factor series and relatively low noise (compared to Example 2.5.3), the inability of $\mathcal{L}_{k_0}$-based method to accommodate nonlinear dependence is amplified. It is also worth noting that for Γ_{k_0}, $\Gamma_{k_0}^{(2)}$ and $\mathcal{L}_{k_0}$, the performance can depend on k_0 to some extent.

Table 2.4: Mean, standard error (*in the bracket*) of RMSE and $\widehat{r}, \widehat{s}$ with 1000 replicates for Examples 2.5.3 and 2.5.4

	$\mathcal{L}_{k_0}$				Γ_{k_0}				$\Gamma_{k_0}^{(2)}$			
	RMSE	$\widehat{r}$			RMSE	$\widehat{s}$			RMSE	$\widehat{s}$		
		$\widehat{r} < 3$	$\widehat{r} = 3$	$\widehat{r} > 3$		$\widehat{s} < 3$	$\widehat{s} = 3$	$\widehat{s} > 3$		$\widehat{s} < 3$	$\widehat{s} = 3$	$\widehat{s} > 3$
EX 2.5.3					$k_0 = 1$							
$p = 100, n = 100$	0.654 (0.17)	0.011	0.936	0.053	0.557 (0.21)	0.047	0.925	0.028	0.557 (0.21)	0.047	0.925	0.028
$p = 100, n = 200$	0.641 (0.18)	0	0.986	0.014	0.52 (0.21)	0.001	0.997	0.002	0.52 (0.21)	0.001	0.997	0.002
$p = 400, n = 200$	0.805 (0.05)	0.995	0.005	0	0.807 (0.04)	1	0	0	0.807 (0.04)	1	0	0
					$k_0 = 5$							
$p = 100, n = 100$	0.811 (0.21)	0.563	0.294	0.143	0.843 (0.14)	0.467	0.026	0.507	0.846 (0.15)	0.496	0.051	0.453
$p = 100, n = 200$	0.713 (0.24)	0.319	0.643	0.038	0.805 (0.21)	0.514	0.167	0.319	0.803 (0.22)	0.524	0.207	0.269
$p = 400, n = 200$	0.78 (0.12)	0.937	0.054	0.009	0.699 (0.14)	0.561	0.029	0.41	0.724 (0.14)	0.675	0.033	0.292
EX 2.5.4					$k_0 = 1$							
$p = 100, n = 100$	1.085 (0.34)	0.856	0.141	0.003	0.598 (0.59)	0.346	0.652	0.002	0.598 (0.59)	0.346	0.652	0.002
$p = 100, n = 200$	1.062 (0.28)	0.894	0.106	0	0.274 (0.4)	0.112	0.888	0	0.274 (0.4)	0.112	0.888	0
$p = 400, n = 200$	1.167 (0.39)	0.837	0.163	0	0.111 (0.04)	0.001	0.999	0	0.111 (0.04)	0.001	0.999	0
					$k_0 = 5$							
$p = 100, n = 100$	1.399 (0.52)	0.841	0.156	0.003	1.252 (0.61)	0.761	0.235	0.004	1.291 (0.61)	0.772	0.223	0.005
$p = 100, n = 200$	1.405 (0.49)	0.861	0.139	0	1.252 (0.62)	0.762	0.237	0.001	1.331 (0.59)	0.805	0.194	0.001
$p = 400, n = 200$	1.12 (0.57)	0.699	0.3	0.001	0.383 (0.56)	0.172	0.827	0.001	0.416 (0.58)	0.184	0.814	0.002

EXAMPLE **2.5.5**. Example 2.5.5 is constructed by following Example 2.4.2 where r and s are different. The factor loading matrix $\mathcal{A} = ([A_1]_{10\times 2}, [A_2]_{10\times 1})$ is a 10×3 matrix. For each columns of $\mathcal{A}$, the first 5 elements are iid U(-2,2) and the rest 5 elements are set to 0. The time series $X_t = ((W_{1t} + W_{3t})^T, W_{4t}^T)^T$, where $W_{1t} = (W_{1t,1}^T, W_{1t,2}^T)^T$ and $W_{3t} = (W_{3t,1}^T, W_{3t,2}^T)^T$. They are

$$\begin{aligned} W_{1t,1} &= v_t,\ W_{1t,2} = v_{t-1}, v_t = 0.5e_{1,t-1} + e_{1,t}, \\ W_{3t,1} &= w_t,\ W_{3t,2} = w_{t-1},\ W_{4t} = w_{t-2}, w_t = e_{3,t-2}e_{3,t-1} + e_{3,t}, \end{aligned}$$

where $e_{1,t}$ follows $N(0, 8^2)$ and $e_{3,t}$ follows $N(0, 1.5^2)$. Then the data is generated by $Y_t = \mathcal{A}X_t + \epsilon_{1t} = A_1(W_{1t} + W_{3t}) + A_2 W_{4t} + \epsilon_{1t}$, where $\epsilon_{1t} \sim N(0, 0.5 \times I_{10})$ and independent from $\{e_{i,t}\}, i = 1, 3$. Note that W_{1t} is from a stationary MA(1) model and $W_{it}, i = 3, 4$ are consecutive observations from a stationary nonlinear MA model. Here $p = 10$, $n = 50, 100, 200$ and k_0 is either 1 or 3. Theoretically, r is equal to 2 but s is 3 therefore the true number of factors, r and s, are different for this example. Not only the true number of factors are different but also factor loading spaces are different i.e., $\mathcal{M}(A) = \mathcal{M}(A_1)$, $\mathcal{M}(\mathcal{A}) = \mathcal{M}([A_1, A_2])$. This is due to the fact that W_{3t} and W_{4t} are white noise sequence but not martingale difference sequence; see Example 2.3 in Shao (2011).

Table 2.5: $\widehat{r}, \widehat{s}$ with 1000 replicates for Example 2.5.5

	$\mathcal{L}_{k_0}$				Γ_{k_0}			$\Gamma_{k_0}^{(2)}$		
	$\widehat{r}$				$\widehat{s}$			$\widehat{s}$		
	$\widehat{r} = 1$	$\widehat{r} = 2$	$\widehat{r} = 3$	$\widehat{r} > 3$	$\widehat{s} < 3$	$\widehat{s} = 3$	$\widehat{s} > 3$	$\widehat{s} < 3$	$\widehat{s} = 3$	$\widehat{s} > 3$
$k_0 = 1$										
$n = 50$	0.072	0.501	0.331	0.096	0.339	0.65	0.011	0.339	0.65	0.011
$n = 100$	0.029	0.558	0.338	0.075	0.251	0.747	0.002	0.251	0.747	0.002
$n = 200$	0.004	0.606	0.324	0.066	0.216	0.784	0	0.216	0.784	0
$k_0 = 3$										
$n = 50$	0.067	0.433	0.482	0.018	0.26	0.722	0.018	0.283	0.692	0.025
$n = 100$	0.043	0.446	0.5	0.011	0.186	0.812	0.002	0.193	0.789	0.018
$n = 200$	0.023	0.504	0.469	0.004	0.154	0.846	0	0.144	0.854	0.002

According to Table 2.5, we can clearly see that both methods are targeting different number of factors and the proportion of estimating the true number of factor correctly increases as n increases. This example confirms that our MDDM-based method is seeking different linear combinations of Y_t from the $\mathcal{L}_{k_0}$-based counterpart and the true number of factors inferred on the basis of $\mathcal{L}_{k_0}$ or Γ_{k_0} can be different.

Our limited simulation results suggest that (1) For dimension reduction of conditional mean, MDDM-based approach can outperform $\mathcal{L}_{k_0}$-based one in both the case of linear Gaussian time series and the nonlinear case in the small-p setting. The performance of the $\mathcal{L}_{k_0}$-based approach seems noticeably inferior when k_0 is small or when nonlinearity is strong in the series; (2) In the large-p setting, our MDDM-based method can still be effective but it depends on the combination of p and n and the data generating process; (3) The performance of $\Gamma_{k_0}^{(1)}$ and $\Gamma_{k_0}^{(2)}$-based ones seem fairly close, as the dependence two cumulative MDDMs capture are quite overlapping after all; (4) The $\mathcal{L}_{k_0}$ and $\Gamma_{k_0}^{(1)}$-based approaches target their respective number of factors and factor loading spaces and their targets could be quite different, as demonstrated in Example 2.5.5. Overall, the finite sample performance of MDDM-based method is quite encouraging.

2.6 Data Illustrations

In this section, we illustrate the usefulness of MDDM-based dimension reduction approach in the context of prediction using two real data sets. The prediction error is measured by

- Forecasting Error (FE) [Lam, Yao and Bathia (2011)]

$$FE = \frac{\|\widehat{Y}_{n+1}^{(1)} - Y_{n+1}\|}{p^{1/2}} = \frac{\|\widehat{\mathcal{A}}\widehat{X}_{n+1}^{(1)} - Y_{n+1}\|}{p^{1/2}},$$

where $\widehat{Y}_{n+1}^{(1)}$ is the one-step ahead prediction for Y_{n+1} based on $(Y_1, \cdots, Y_n)$ and $\widehat{X}_{n+1}^{(1)}$ is the one-step ahead prediction for X_{n+1} based on a parametric model fitted to the estimated factor series $(\widehat{X}_1, \cdots, \widehat{X}_n)$. FE quantifies the prediction accuracy of a dimension reduction method and smaller value of FE indicates more accurate one-step ahead prediction. Multi-step ahead prediction can be done similarly; see Section 2.6.1.

2.6.1 GDP Data

The first data set we analyze is the quarterly change in seasonally adjusted GDP, in percentage, for five countries, i.e., United States (US), Canada (CA), United Kingdom (GBR), South Korea (KO), and Taiwan (TW), which is available from the Organization for Economic Cooperation and Development Web site. This data set has been analyzed by Matteson and Tsay (2011) using the so-called DOC (Dynamic Orthogonal Component) method. The main idea of DOC is that by a contemporary linear transformation of the original p-dimensional time series, the resulting p-dimensional time series does not have any significant cross-correlations, so univariate time series models can be fitted to each component of transformed time series and dimension reduction can be achieved this way. We use the data from the first quarter of 1981 through the last quarter of 2009. Thus, the length of the series $n = 116$ and the dimension $p = 5$.

According to Figure 1 in Matteson and Tsay (2011), there seems no obvious nonstationarity in all time series. To realistically measure the forecasting error, we divide the data set into training set and testing set. Approximately 80% of the data which contains the first 92 data points is included into the training set and the remaining 24 data points are included into the testing set. Specifically, we use a rolling-window approach, and get $\widehat{Y}^{(h)}_{92+j}, h = 1, 2, 3, 4$ based on $(Y_j, \cdots Y_{91+j})$ for $j = 1, 2, \cdots 24 - h, h = 1, 2, 3, 4$ and then report the average of forecasting errors over $24 - h, h = 1, 2, 3, 4$ periods. The one, two, three, and four-step ahead forecasts are implemented using the following procedure: (1) For $j = 1, \cdots, 24 - h, h = 1, 2, 3, 4$, we apply PCMC to the $(Y_j, \cdots Y_{91+j})$ and estimate the number of factors and the factor series; (2) A vector autoregressive model is fitted to the estimated factor series with the order chosen by the AIC and followed by a refinement of the optimal VAR model, which is achieved by hard thresholding the VAR coefficient matrix based on t-ratio (thresholding value equals to 1.7) using "refVAR" in MTS package. The reason for refining VAR model is the over-parameterization of the optimal VAR model considering the limited time series length. We also tried setting the thresholding value to be 2, which gives similar results; (3) the h-step ahead forecast is then $\widehat{\mathcal{A}}\widehat{X}^{(h)}_{92+j}, h = 1, 2, 3, 4$, where $\widehat{\mathcal{A}}$ is the estimated factor loading matrix, and $\widehat{X}^{(h)}_{92+j}$ is the h-step ahead prediction based on the fitted VAR model to the lower-dimensional factor series. DOC method is also carried out in a similar fashion. Once an uncorrelated p-dimensional time series is obtained, (1) optimal AR models selected by AIC are fitted to each univariate time series; (2) the h-step ahead predictions are computed for each univariate time series and transformed back into the original scale to compute FE.

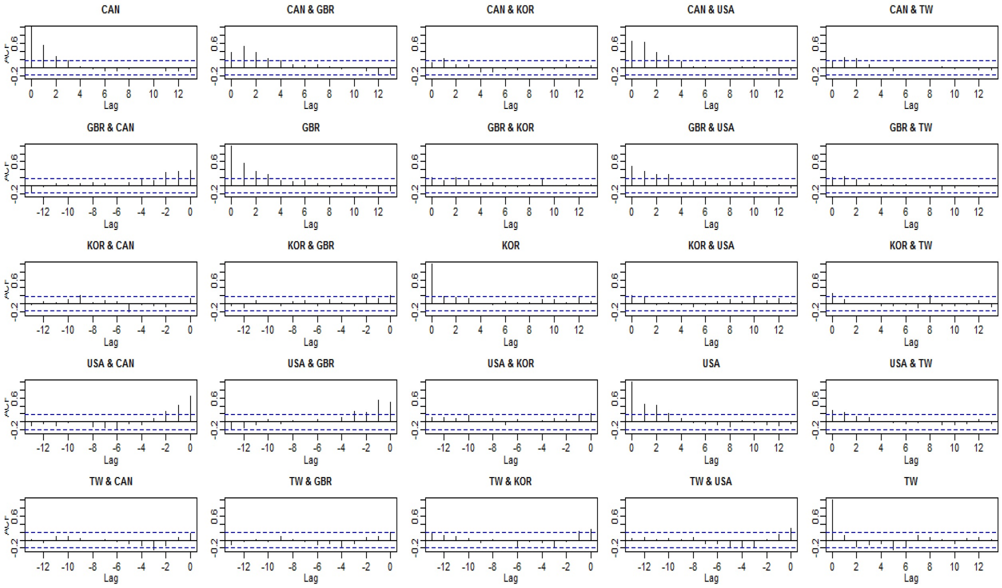

Figure 2.1: Auto and cross-correlations of GDP of five countries

From the autocorrelation plot in Figure 2.1, significant autocorrelations and cross correlations exist up to lag 3. In Table 2.6, we present the average forecasting error and the proportion of estimated number of factors based on $\mathcal{L}_{k_0}$ and $\Gamma^{(1)}_{k_0}$ for $k_0 = 1, 2, 3$. It can be seen that Γ_{k_0}-based approach could noticeably outperform DOC-based one and $\mathcal{L}_{k_0}$-based one in terms of forecasting error and its advantage seems quite uniform across (h, k_0), where h indicates the hth ahead prediction and k_0 is the number of lags involved. The performance of $\mathcal{L}_{k_0}$ and Γ_{k_0} depends on the choice of k_0, and in this case $k_0 = 3$ often delivers the optimal forecasting error (unreported results show that increasing k_0 beyond 3 does not help reduce the forecasting error significantly). Noticeably, Γ_{k_0}-based approach appears less sensitive to the choice of k_0 as compared to $\mathcal{L}_{k_0}$-based ones. It is worth noting that the gain in forecasting accuracy by Γ_{k_0}-based approach (as compared to the $\mathcal{L}_{k_0}$-based one) is completely due to the use of a different matrix to extract the number of factors and factor loading matrix, as we apply the same procedure of fitting VAR model to the estimated factor series.

Table 2.6: Mean of FE and $\widehat{r}, \widehat{s}$ for GDP data

	$\mathcal{L}_{k_0}$				Γ_{k_0}				DOC	
	FE	$\widehat{r}$			FE	$\widehat{s}$			FE	$\widehat{s}$
		$\widehat{r} = 2$	$\widehat{r} = 3$	$\widehat{r} = 4$		$\widehat{s} = 2$	$\widehat{s} = 3$	$\widehat{s} = 4$		$\widehat{s} = 5$
one-step ahead prediction										
$k_0 = 1$	1.377	0	0.25	0.75	1.129	0	1	0	1.345	1
$k_0 = 2$	1.301	0.042	0.75	0.208	1.129	0	1	0	1.345	1
$k_0 = 3$	1.17	0	1	0	1.147	0	1	0	1.345	1
two-step ahead prediction										
$k_0 = 1$	1.597	0	0.26	0.74	1.31	0	1	0	1.326	1
$k_0 = 2$	1.398	0	0.78	0.22	1.281	0	1	0	1.326	1
$k_0 = 3$	1.326	0	1	0	1.255	0	1	0	1.326	1
three-step ahead prediction										
$k_0 = 1$	1.509	0	0.27	0.73	1.366	0	1	0	1.32	1
$k_0 = 2$	1.431	0	0.77	0.23	1.296	0	1	0	1.32	1
$k_0 = 3$	1.318	0	1	0	1.255	0	1	0	1.32	1
four-step ahead prediction										
$k_0 = 1$	1.528	0	0.29	0.71	1.326	0	1	0	1.358	1
$k_0 = 2$	1.304	0	0.76	0.24	1.289	0	1	0	1.358	1
$k_0 = 3$	1.28	0	1	0	1.274	0	1	0	1.358	1

2.6.2 7-city Temperature Series

The second data set we analyze is the monthly temperature series for 7 cities: Nanjing, Dongtai, Huoshan, Hefei, Shanghai, Anqing and Hangzhou in Eastern China. The series run from January 1954 to December 1998, a portion of which have been analyzed by Pan and Yao (2008). The length of the data is $n = 540$ and the dimension $p = 7$. Figure 2.2 suggests that there exist strong seasonal dependence.

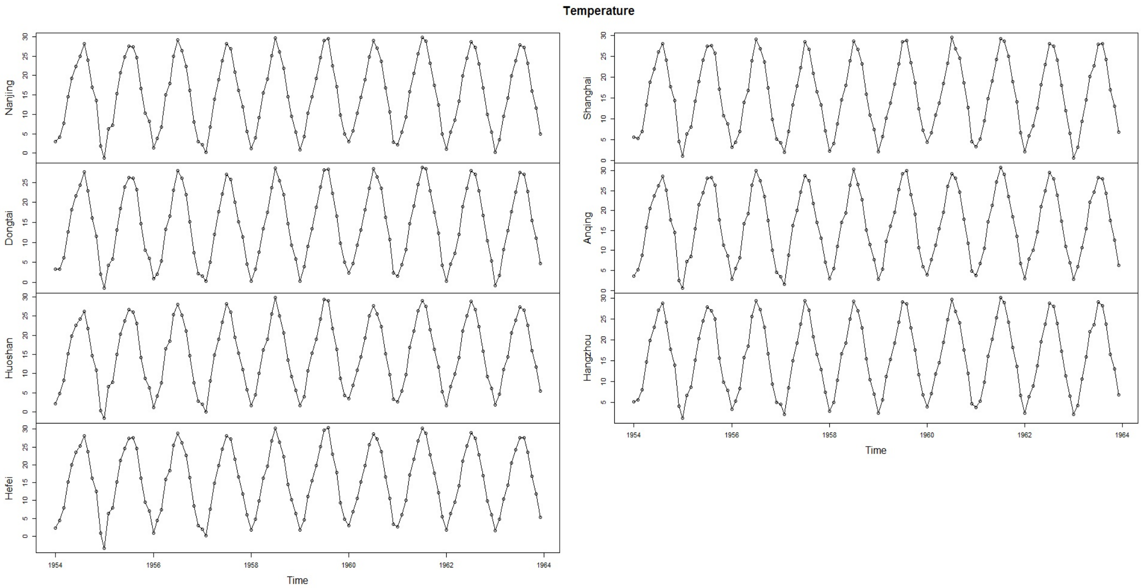

Figure 2.2: Monthly temperatures for seven cities in Eastern China

Following the approach in Section 2.6.1, we use the first 80% of the data which contains the first 433 data points as the training set and the remaining 107 data points are included in the testing set. Using a rolling-window approach, we get $\widehat{Y}^{(1)}_{433+j}$ based on $(Y_j, \cdots, Y_{432+j})$ for $j = 1, \cdots, 107$ and then report the average of forecasting errors over 107 periods. The one-step ahead forecast is implemented using the following procedure: (1) For $j = 1, \cdots, 107$, we apply PCMC to the training data $(Y_j, \cdots, Y_{432+j})$ and estimate the number of factors and factor series; (2) A (possibly multivariate) seasonal ARIMA $(0, 0, 1) \times (0, 1, 1)_{12}$ model is fitted to the estimated factor series. The order of this particular model is determined by checking ACF plots of estimated factor series, as shown in Figure 2.3 for one particular training data; the model fits the estimated factor series quite well, as the residual autocorrelation is fairly small for most training data; see Figure 2.4 for the average of the absolute value of the acf of residual series after fitting the above seasonal model to the estimated factor series from MDDM-based approach. Similar findings apply to the $\mathcal{L}_{k_0}$-based one. (3)

the one-step ahead forecast is then $\widehat{\mathcal{A}}\widehat{X}^{(1)}_{433+j}$, where $\widehat{\mathcal{A}}$ is the estimated factor loading matrix, and $\widehat{X}^{(1)}_{433+j}$ is the one-step ahead prediction based on the fitted seasonal ARIMA model to the lower dimensional factor series. In Table 2.7, we present the average forecasting error and the proportion of estimated number of factors based on $\mathcal{L}_{k_0}$ and $\Gamma^{(1)}_{k_0}$. The choice of $k_0 = 12,\ 24,\ 36$ is based on the consideration that the time series is apparently seasonal with period 12.

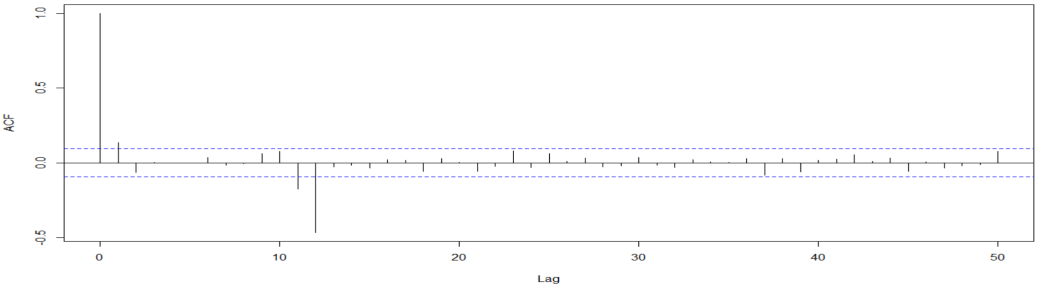

Figure 2.3: ACF plot of estimated factor series selected by $\Gamma^{(1)}_{k_0}$-based method

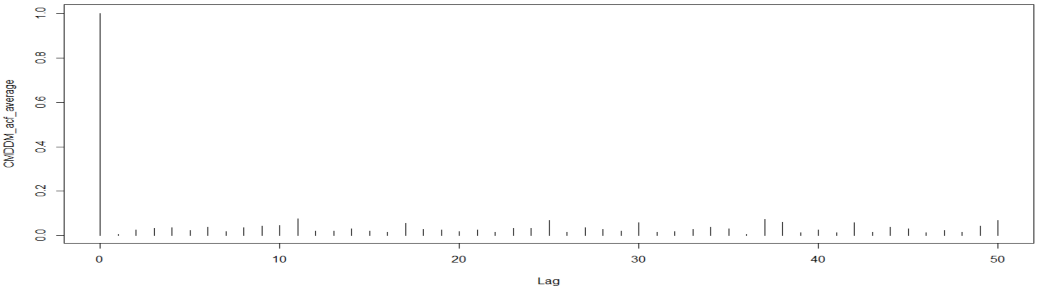

Figure 2.4: Average of absolute value of ACF of residual series after fitting the seasonal model to the estimated factor series determined by the $\Gamma^{(1)}_{k_0}$-based method

Table 2.7: Mean of FE and $\widehat{r}, \widehat{s}$ for 7-city monthly temperature data

	$\mathcal{L}_{k_0}$				Γ_{k_0}			
	FE	$\widehat{r}$			FE	$\widehat{s}$		
		$\widehat{r}=1$	$\widehat{r}=2$	$\widehat{r}>2$		$\widehat{s}=1$	$\widehat{s}=2$	$\widehat{s}>2$
$k_0 = 12$	1.894	0	1	0	1.113	1	0	0
$k_0 = 24$	1.898	0	1	0	1.113	1	0	0
$k_0 = 36$	1.892	0	1	0	1.113	1	0	0

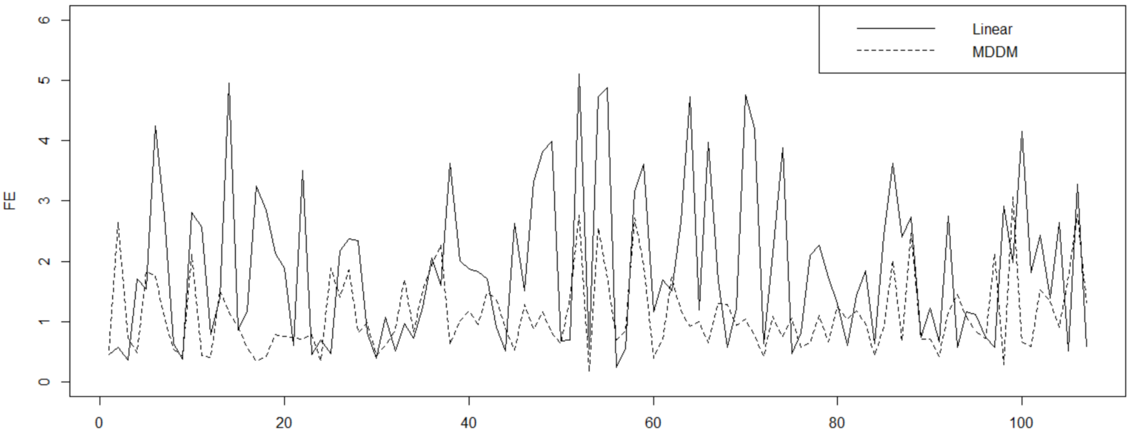

Figure 2.5: Forecast errors computed from 107 training-testing sets with $k_0 = 12$

It can be seen from Table 2.7 that $\mathcal{L}_{k_0}$-based approach and Γ_{k_0}-based one deliver different number of factors and the performance of both methods are stable with respect to the choice of $k_0 = 12, 24, 36$. The Γ_{k_0}-based method has substantially smaller FEs than those of $\mathcal{L}_{k_0}$-based method, as shown in Figure 2.5. It suggests that using one factor model, as estimated by the MDDM-based approach for all training data, can lead to more accurate forecasting. Note that the underlying factor series is non-stationary, which is not covered by our theory. Nevertheless, it shows the potential applicability of our approach to non-stationary time series. For both real data examples, it would be interesting to fit multivariate nonlinear time series models to the estimated factor series after applying Γ_{k_0}-based dimension reduction method. We did not pursue this step since there seems no general guidance on how such modeling can be conducted.

2.7 Summary and Conclusion

In this paper, we propose the so-called martingale difference divergence matrix to quantify the conditional mean (in)dependence of a random vector $V \in R^p$ on $U \in R^q$. The MDDM encodes the number and form of linear combinations of V that are conditional mean independent of U. Building on this property, we generalize MDDM to the time series context and introduce cumulative MDDM, which can approximately quantify the conditional mean independence of Y_t upon the past information $\mathcal{F}_{t-1}$. Dimension reduction for a multivariate time series is then achieved by estimating the number and form of linear combinations that are conditionally mean independent of the past, which is encoded in the cumulative MDDM. Compared to the use of linear dependence metric in Lam, Yao and Bathia (2011) and Lam and Yao (2012), our cumulative MDDM is a natural nonlinear analogue and can capture unknown nonlinear mean dependence. We also present a static factor model representation for our dimension reduction framework and discuss the subtle difference from the static factor model that was studied in previous literature. Since we typically do not know a priori whether nonlinear dependence exists in practice, it might be safe/robust to use our MDDM-based dimension reduction approach. In terms of implementation, since sample MDDM has a V-statistic form, it can be readily calculated. The estimation of the number of factors and factor loading matrix can be conveniently implemented by spectral decomposition of sample cumulative MDDM.

For MDDM-based dimension reduction of conditional mean, we provide some theoretical results to justify the validity of our method. Although our theory is obtained only for the case p is fixed, we investigate the finite sample performance in both small p and large p settings in our simulation studies. In the small-p setting, our limited simulation results show that our MDDM-based approach can be as effective as the linear counterpart in Lam and Yao (2012) for linear Gaussian time series and can outperform the latter for nonlinear time series. The merits of the MDDM-based dimension reduction are further supported by two real data illustrations. Although the simulation suggests our approach may still be useful in the large p setting, there is currently no theoretical support. As seen from Li, Wang and Yao (2016), there can be complications with the ratio-based estimator (see (2.4.2)) in the high-dimensional setting, and one may have to resort to random matrix theory to derive a sensible estimator for the number of factors. In addition, the finite sample simulation results and data illustrations show that in some cases, the results can depend on the choice of k_0, which is the number of lags included in the cumulative MDDM. It would be desirable to develop a data driven rule for k_0 besides the visual inspection of the (partial) autocorrelation plot. Furthermore, strict stationarity is assumed throughout, and it would be interesting to extend the MDDM-based methodology to allow nonstationary series; see Pan and Yao (2008), Motta, Hafner and von Sachs (2011), and Eichler, Motta and von Sachs (2011). Also an extension to dimension reduction for conditional variance-covariance matrix using MDDM and its variant would be interesting. The research along these directions are well underway.

Chapter 3

Dimension Reduction for Multivariate Volatility

3.1 Background

Volatility is a crucial quantity in economics and finance since it represents measurement of risk and often an estimate of volatility is required in order to conduct tasks of economics and finance such as hedging. It has been empirically documented that the volatility of multivariate time series is changing over time and it is essential to model time-varying multivariate volatility [see Engle (1982), Bollerslev (1986)]. A main difficulty in the multivariate volatility modeling is the curse of dimensionality. If the dimension of time series is p, the volatility matrix is of dimension $p(p+1)/2$, and GARCH models without any structral constraints would require $O(p^4)$ number of parameters, thus dimension reduction is often necessary in volatility modeling even for moderate p. There is a large and growing literature on the dimension reduction for volatility modeling. Here we mention several representative lines of research, notably GARCH models with structural constraints (e.g. Bollerslev (1990), Engle, Ng and Rothschild (1990), Engle (2002), Weide (2002), Pelletier (2006)) and the use of principal component analysis (PCA) and variations (e.g. Chen, Härdle and Spokoiny (2007), Fan, Wang and Yao (2008), Matteson and Tsay (2011), Hu and Tsay (2014) and Li, Gao, Li and Yao (2016)). In the application of PCA, some of the articles mentioned above used the covariance matrix to quantify the conditional variance dependence of multivariate time series, which may fail to capture the nonlinear volatility dependence. There are a few exceptions. For example, the generalized kurtosis matrix was recently developed by Hu and Tsay (2014), which can measure certain degree of nonlinear dependence and forms the core of the so-called principal volatility component analysis. In particular, applying eigen decomposition to the so-called cumulative generalized kurtosis matrix, which is the summation of generalized kurtosis matrix at different time lags, can lead to an effective estimation of the number and forms of linear combinations that are conditionally heteroscedastic. More recently, Li, Gao, Li and Yao (2016) further proposed a way of capturing nonlinear dependence to generalize the principal volatility component analysis of Hu and Tsay (2014) by extending an indicator function based approach used in Pan, Polonik and Yao (2010) and Fan, Wang and Yao (2008) for a related problem. However, their nonlinear metric requires the selection of user-chosen quantities and can be computationally costly to implement.

In this paper, we introduce new matrix objects, the so-called volatility martingale difference divergence matrix (VMDDM, hereafter) and vec volatility martingale difference divergence matrix (vecVMDDM, hereafter) to measure both linear and nonlinear conditional variance dependendence. VMDDM and vecVMDDM can be viewed as extensions of martingale difference divergence matrix in Lee and Shao (2016), which measures the conditional mean dependence. We demonstrate the usage of VMDDM and vecVMDDM in dimension reduction of volatility context by applying new matrix objects to two dimension reduction frameworks: One is the principal volatility component analysis (PVCA, hereafter) proposed by Hu and Tsay (2014) and generalized by Li, Gao, Li and Yao (2016). Here, the goal is to estimate the number of linear combinations that exhibit conditional heteroscedasticity and the volatility space [Li, Gao, Li and Yao (2016)] which is the space spanned by these linear combinations. In the other framework, we assume that there exsit conditionally uncorrelated components (CUC, hereafter) [Fan, Wang and Yao (2008)] after a linear transformation and the objective is to estimate the orthogonal transformation matrix. Our proposed metrics are characteristic function-based, and they are conceptually simple, easy to implement and requires less number of user-chosen quantities.

The rest of the paper is organized as follows. We introduce VMDDM, its properties and application to principal volatility component analysis in Section 3.2. In Section 3.3, we propose vecVMDDM and describe its corresponding application to dimension reduction in the context of conditionally uncorrelated components model. To demonstrate the finite sample performance of dimension reduction for volatility with VMDDM and vecVMDDM, simulation results are presented in Section 4.4.1 and data examples are collected in Section 3.5. Section 3.6 concludes.

A word on notation. Let $i = \sqrt{-1}$ be the imaginary unit. The scalar product of vectors x and y is denoted by $< x, y >$. For a complex-valued function $f(\cdot)$, the complex conjugate of f is denoted by $\overline{f}$ and $|f|^2 = f\overline{f}$. Denote the Euclidean norm of $x = (x_1, \cdots, x_p) \in \mathbf{C}^p$ as $|x|_p$, where $|x|_p^2 = x_1\overline{x}_1 + \cdots + x_p\overline{x}_p$, and if $x = (x_1, \cdots, x_p) \in R^p$, it is sometimes denoted as $\|x\|$, where $\|x\|^2 = x_1^2 + \cdots x_p^2$. For a square matrix $A = (A_{i,j})_{i,j=1,\cdots,p}$, spectral norm of A is denoted as $\|A\|_2$, where $\|A\|_2 = \sqrt{\lambda_{max}(A^TA)}$ and Frobenius norm of A is denoted by $\|A\|_F$, where $\|A\|_F = \sqrt{tr(A^TA)}$ and $tr(A) = \sum_{i=1}^p A_{i,i}$. A random vector $x \in \mathcal{L}^s$ if $E|x|_p^s < \infty$.

3.2 Principal Volatility Component Analysis

Modeling and inference for volatility is of primary importance in the analysis of econometric and financial time series. Empirical study of multivariate volatility is difficult, in part because of high dimensionality of the volatility matrix, and also due to the positive semi-definiteness constraints on the volatility matrix, which needs to be satisfied by its sample estimator. Many approaches have been proposed to model and estimate volatility matrix; see Tsay (2010, Chapter 10) for a brief discussion of these methods. Recently, Hu and Tsay (2014) and Li, Gao, Li and Yao (2016) proposed methods to achieve dimension reduction of volatility by assuming a factor model, which is briefly reviewed below.

Let $Y_t = (Y_{1,t}, \cdots, Y_{p,t})^T$ denote a p-dimensional stationary time series. We assume $E(Y_t|\mathcal{F}_{t-1}) = 0$ for simplicity as our focus is on the volatility. Define the volatility matrix of Y_t as $\Sigma_t = \text{cov}(Y_t|\mathcal{F}_{t-1}) = E(Y_tY_t^T|\mathcal{F}_{t-1})$, which is a $p \times p$ matrix. To perform dimension reduction for Σ_t, Hu and Tsay (2014) imposed a linear structure, i.e.,

$$vec(\Sigma_t) = c_0 + \sum_{i=1}^{\infty} C_i vec(Y_{t-i}Y_{t-i}^T), \tag{3.2.1}$$

where $vec(M)$ denotes the column-stacking vector of the matrix M, c_0 is a p^2-dimensional positive constant vector and C_i are $p^2 \times p^2$ constant matrices for $i > 0$. Thus the process Y_t has conditional heteroscedasticity if and only if $C_i \neq 0$ for some $i > 0$, which is equivalent to the fact that $Y_tY_t^T$ is correlated with $Y_{t-i}Y_{t-i}^T$ for some $i > 0$. This observation motivates them to define the lag-l generalized kurtosis matrix

$$\mathcal{G}_l = \sum_{i=1}^{p}\sum_{j=i}^{p} \text{cov}(Y_tY_t^T, x_{ij,t-l})\text{cov}(Y_tY_t^T, x_{ij,t-l})^T,$$

where $x_{ij,t-l}$ is a nonlinear function of $Y_{i,t-l}Y_{j,t-l}$, i.e., $x_{ij,t-l} = \phi(Y_{i,t-l}Y_{j,t-l})$, with $\phi(\cdot)$ chosen to be Huber's function. In particular,

$$\phi(y) = y\mathbf{1}(|y| \le c^2) + \{2c\sqrt{y} - c^2\}\mathbf{1}(y > c^2) + \{c^2 - 2c\sqrt{|y|}\}\mathbf{1}(y < -c^2), \quad (3.2.2)$$

where c is a pre-specified constant. The cumulative generalized kurtosis matrix is then defined as

$$\Omega_{k_0} = \sum_{l=1}^{k_0} \mathcal{G}_l$$

to measure the ARCH(k_0), $k_0 < \infty$ effects in Y_t and $k_0 = \infty$ for general GARCH-type models. From its definition, we can see that cumulative generalized kurtosis matrix measures the cumulative linear dependence of $Y_tY_t^T$ on $\{x_{ij,t-l}\}$.

Motivated by Hu and Tsay's proposal, we seek linear combinations of Y_t, say m^TY_t, such that m^TY_t has no conditional heteroscedasticity, i.e., $E((m^TY_t)^2|\mathcal{F}_{t-1}) = E(m^TY_t)^2$, which is equivalent to $m^T\Sigma_t m = m^T\Sigma m$, where $\Sigma = E(Y_tY_t^T)$ is the unconditional covariance matrix of Y_t. This is further equivalent to the fact that

$$m^TE(\{Y_tY_t^T - \Sigma\}|\mathcal{F}_{t-1})m = 0, \quad (3.2.3)$$

which implies that $m^T\Omega_{k_0}m = 0$. However, since they use the linear metric to measure the uncorrelatedness of $Y_tY_t^T$ with $x_{ij,t-l}$ for all $i \le j$ and $l = 1, 2, \cdots$, their procedure, which is based on Ω_{k_0} and its sample estimate, may not be able to fully capture nonlinear dependence. To this end, Li, Gao, Li and Yao (2016) adopted an indicator function-based approach and formulated the PVCA as an equivalent factor model to acheive dimension reduction for Σ_t. Specifically, let

$$Y_t = AX_t + \epsilon_t,$$

where $A \in R^{p\times s}$ is a factor loading matrix, $X_t \in R^s$ is a latent factor which generates the conditional heteroscedasticity of Y_t and ϵ_t is an error process which exhibits conditional homoscedasticity. They defined a volatility space $\mathcal{M}$ as the space spanned by colums of the matrix A which is assumed to satisfy $A^TA = I_s$, where s is the number of factors. It is important to note that matrix A is not unique, but volatility space $\mathcal{M}$ is unique. Under the above factor model,

$$\Sigma_t = A\Sigma_{x,t}A^T + \Sigma_\epsilon,$$

where $\Sigma_{x,t} = cov(X_t|\mathcal{F}_{t-1})$. Let $B \in R^{p\times(p-s)}$ be a matrix such that (A, B) forms a $p \times p$ orthogonal matrix. Then

$$E[(Y_tY_t^T - \Sigma)I(Y_{t-k} \in W)]B = 0, \quad \forall W \in \mathcal{B}_t,$$

where $\mathcal{B}_t$ is any π-class such that the σ-algebra generated by $\mathcal{B}_t$ is $\mathcal{F}_{t-1}$ and $I(\cdot)$ is an indicator function. By using the above fact, they built the following matrix,

$$\Psi_{k_0} = \sum_{k=1}^{k_0}\sum_{W\in\mathcal{B}} w(W)\{E[(Y_tY_t^T - \Sigma)I(Y_{t-k} \in W)]\}^2$$

where $w(\cdot)$ is a nonnegative weight function, $\mathcal{B}$ is a countable sequence of subsets and one example of $(w(\cdot), \mathcal{B})$ adopted in Fan, Wang and Yao (2008) and Li, Gao, Li and Yao (2016) is $w(\cdot) = \frac{1}{n}$, $\mathcal{B} = \{u \in R^p : |u| \le |Y_t|, t = 1, 2, \cdots, n\}$. According to the definition of Ψ_{k_0}, it is easily seen that Ψ_{k_0} is a positive semidefinite matrix. Similar to Hu and Tsay (2014), they seek $m \in R^p$ such that m^TY_t has no conditional heteroscedasticity which corresponds to those in the orthogonal complement of the volatiliy space. Since

$$E[(Y_tY_t^T - \Sigma)I(Y_{t-k} \in W)]m = 0,$$

implies $m^T\Psi_{k_0}m = 0$.

Remark **3.2.1**. It is worth mentioning that conditional variance dependence has been taken into account in Pan, Polonik and Yao (2010) who proposed the so-called innovation expansion approach. More specifically, Ψ_{k_0} in Li, Gao, Li and Yao (2016) appeared in Pan, Polonik and Yao (2010) to measure the conditional variance dependence but the two papers differ in the way they estimate the number of factors and the volatility space. In particular, Li, Gao, Li and Yao's approach is based on spectral decomposition of Ψ_{k_0} and they adopt the ratio-based estimator (see (3.2.7)) to estimate the number of factors.

Note that Li, Gao, Li and Yao (2016) do not assume linear structure (see (3.2.1)) in Hu and Tsay (2014) and their Ψ_{k_0} incorporates nonlinear conditional variance dependence of Y_t upon $\mathcal{F}_{t-1}$, and can be regarded as a nonlinear analogue of Ω_{k_0}. However, a practical drawback associated with this approach is that it requires selections of several user-chosen parameters when computing Ψ_{k_0} i.e., k_0, $w(\cdot)$, $\mathcal{B}$. Especially, for $w(\cdot)$, $\mathcal{B}$, there seems no clear guidelines to follow for selecting these user-chosen parameters and little is known about the impact of two user-chosen parameters in practice.

3.2.1 Volatility Martingale Difference Divergence Matrix

In a recent article, Lee and Shao (2016) constructed a matrix called MDDM which measures conditional mean dependence between two random vectors and applied it to achieve dimension reduction for conditional mean of stationary multivariate time series. Below we propose an extension of MDDM to measure the conditional variance dependence of two random vectors and use it to do dimension reduction for volatility. In this section, we provide the definition of volatility MDDM and present its key properties.

For $Y \in R^p$, $X \in R^q$ and $E(Y|X) = E(Y) = 0$, suppose that there is a linear combination of Y, say $\alpha \in R^p$, such that $E[(\alpha^T Y)^2|X] = E[(\alpha^T Y)^2]$ although Y is not necessarily conditionally variance independent of X, i.e. $E[YY^T|X] \neq E[YY^T]$ a.s.. Our goal is to find a matrix that encodes the number and the form of linear combinations of Y that are conditionally variance independent of X.

For $q = 1$, the generalized kurtosis matrix is defined as
$KM = \text{cov}(YY^T, X)\text{cov}(YY^T, X)^T$ which is a real and symmetric $p \times p$ matrix. We can rewrite it as

$$KM = E((YY^T - \Sigma)(Y'(Y')^T - \Sigma)^T X^T X')$$

where (X', Y') is an iid copy of (X, Y) and $\Sigma = E(YY^T)$.

To measure the conditional variance independence of Y on X, i.e., $\text{var}(Y|X) = \text{var}(Y)$, which is equivalent to $E(YY^T|X) = \Sigma$ under the assumption that $E(Y|X) = 0$, we first note an analogy between the $\text{MDDM}(\cdot|\cdot)$ and $L(\cdot|\cdot)$ in Definition 2.3.1 and Remark 2.3.1. Write

$$L(Y|X) = \text{cov}(Y, X)\text{cov}(Y, X)^T = E((Y - E(Y))(Y' - E(Y'))^T X^T (X')),$$

$$MDDM(Y|X) = -E[(Y - E(Y))(Y' - E(Y'))^T |X - X'|_q],$$

which shows that $MDDM(Y|X)$ replaces $X^T X'$ in $L(Y|X)$ by $-|X - X'|_q$ in its definition. Based on this intuition and the definition of KM, we define the volatility martingale difference divergence matrix below.

Definition **3.2.1**. *Volatility Martingale Difference Divergence Matrix*

Given $Y = (Y_1, \cdots, Y_p) \in R^p, X \in R^q$ *and assume that* $Y \in \mathcal{L}^4$, $X \in \mathcal{L}^2$ *and* $E(Y|X) = 0$*. Define that*

$$VMDDM(Y|X) = -E[(YY^T - \Sigma)(Y'(Y')^T - \Sigma)^T|X - X'|_q]$$

Note that $VMDDM(Y|X)$ is a real and symmetric $p \times p$ matrix.

Proposition **3.2.1**. *Assume that* $Y \in \mathcal{L}^4$, $X \in \mathcal{L}^2$ *and* $E(Y|X) = 0$*. Then we have that*

1. $VMDDM(Y|X)$ *is positive semidefinite.*
2. *The rank of* $VMDDM(Y|X)$ *is equal to* $p - h$*, where* h *is the number of linearly independent combinations* $\alpha_1, \cdots, \alpha_h$*, such that* $E((\alpha_j^T Y)^2|X) = E((\alpha_j^T Y)^2)$ *for* $j = 1, \cdots, h$*.*

It can be seen from the proof of Proposition 3.2.1 that the conditional variance independence of $\alpha^T Y$ on X is equivalent to conditional mean independence of $YY^T\alpha$ on X. Given a random sample $(X_t, Y_t)_{t=1}^n$ from the joint distribution of (X, Y), let $\overline{Y}_n = n^{-1}\sum_{t=1}^n Y_t$. Then we define the sample VMDDM as

$$VMDDM_n(Y|X) = -\frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n (Y_iY_i^T - \Sigma_n)(Y_jY_j^T - \Sigma_n)^T|X_i - X_j|_q,$$

where $\Sigma_n = n^{-1}\sum_{t=1}^n (Y_t - \overline{Y}_n)(Y_t - \overline{Y}_n)^T$ is the sample estimator of Σ.

Recall that Ψ_{k_0} defined in Li, Gao, Li and Yao (2016) is also a nonlinear analogue of Ω_{k_0} in Hu and Tsay (2014). The matrix that corresponds to Ψ_{k_0} is $LG(Y|X) = \sum_{W \in \mathcal{B}} w(W)\{E[(YY^T - \Sigma)I(X \in W)]\}^2$ which can be expressed as

$$LG(Y|X) = E[(YY^T - \Sigma)(Y'(Y')^T - \Sigma)^T \sum_{W \in \mathcal{B}} w(W)I(X \in W)I(X' \in W)]$$

The main difference between $VMDDM(Y|X)$ and $LG(Y|X)$ is the kernel function of X where $VMDDM(Y|X)$ and $LG(Y|X)$ have $-|X - X'|_q$ and $\sum_{W \in \mathcal{B}} w(W)I(X \in W)I(X' \in W)$, resepectively. It would be interesting to see which inference method (i.e., the one based on $LG(Y|X)$ versus the one based on $VMDDM(Y|X)$) delivers a better estimate of the number of factors and volatility space and we shall address this question in our simulations.

3.2.2 Cumulative Volatility Martingale Difference Divergence Matrix

As we discussed at the beginning of Section 3.2, our goal is to quantify the conditional variance dependence of Y_t on $\mathcal{F}_{t-1} = \sigma(Y_{t-1}, Y_{t-2}, \cdots)$. In practice, since we only have a finite stretch of observations from the process Y_t, we approximate the conditional variance dependence of Y_t on $\mathcal{F}_{t-1}$ by Y_t on $\mathcal{F}_{t-1,t-k_0} = \sigma(Y_{t-1}, \cdots, Y_{t-k_0})$, where k_0 is a pre-specified fixed integer. This approximation is quite common in time series literature and considered reasonable for certain time series models. For instance, if the time series model is $ARCH(k_0)$, then dependence of volatility on past information is captured in $\mathcal{F}_{t-1,t-k_0}$. Below we define so-called cumulative volatility martingale difference divergence matrix.

Definition **3.2.2**. *Cumulative Volatility Martingale Difference Diveregence Matrix*

Let $Y_t \in R^p$ be a time series process with $E[Y_t|\mathcal{F}_{t-1}] = 0$. The cumulative volatility martingale difference diveregence matrix is defined as

$$V_{k_0} = \sum_{l=1}^{k_0} VMDDM(Y_t|Y_{t-l}). \tag{3.2.4}$$

Since VMDDM is a positive semidefinite $p \times p$ matrix, V_{k_0} is also positive semidefinite. Note that $VMDDM(Y_t|Y_{t-l})$ depends on the time lag l but not on t due to stationarity. The sample estimate of V_{k_0} is given by $\widehat{V}_{k_0} = \sum_{l=1}^{k_0} VMDDM_n(Y_t|Y_{t-l})$. Following Hu and Tsay (2014), we can explore the PVCA based on spectral decomposition of $\widehat{V}_{k_0}$ and the details are presented in Section 3.2.3.

Remark **3.2.2**. Our definition of V_{k_0} differs from that in Hu and Tsay (2014) and Li, Gao, Li and Yao (2016) in several aspects. For Hu and Tsay (2014), we use a fixed k_0 whereas Hu and Tsay (2014) used ∞ at the population level, and used a growing sequence of truncation lags $k_0(n)$ in their sample estimator. While the number of lags included is always finite for a given sample size n, the asymptotic analysis seems quite different for fixed k_0 or growing $k_0(n)$. Furthermore, we note that Ω_{k_0} is cumulating the dependence from various lags in an entrywise and pairwise fashion, and V_{k_0} collects dependence only in a pairwise fashion since our VMDDM$(Y|X)$ is well defined for $X \in R^q$, $q > 1$, whereas the generalized kurtosis matrix is only defined for $X \in R^1$. The key difference between Li, Gao, Li and Yao (2016) and our approach can be explained as follows. If α is a linear combination of Y_t that are conditionally variance independent of $Y_{t-k}, k = 1, \cdots, k_0$ then

$$E[(Y_tY_t^T - \Sigma)I(Y_{t-k} \in W)]\alpha = 0, \quad \forall W \in \mathcal{B}_t, \quad \forall k = 1, \cdots, k_0, \tag{3.2.5}$$

and this is equivalent to $\Psi_{k_0}\alpha = \sum_{k=1}^{k_0} \sum_{W \in \mathcal{B}_t} w(W)\{E[(Y_tY_t^T - \Sigma)I(Y_{t-k} \in W)]\}^2\alpha =$

$0 \Leftrightarrow \alpha^T \Psi_{k_0} \alpha = 0$. By contrast, our approach hinges on the observation that

$$Cov(Y_t Y_t^T, e^{i<s,Y_{t-k}>})\alpha = 0, \quad \forall s \in R^q, \quad \forall k = 1, \cdots, k_0. \tag{3.2.6}$$

This is equivalent to

$$\begin{aligned} |Cov(Y_t Y_t^T, e^{i<s,Y_{t-k}>})\alpha|^2 = 0, \forall k = 1, \cdots, k_0 \quad &\Leftrightarrow \quad VMDDM(Y_t|Y_{t-k})\alpha = 0, \forall k = 1, \cdots, k_0 \\ &\Leftrightarrow \quad V_{k_0}\alpha = 0 \\ &\Leftrightarrow \quad \alpha^T V_{k_0} \alpha = 0. \end{aligned}$$

Thus Li, Gao, Li and Yao (2016) used an indicator function-based approach whereas we adopt a characteristic function-based approach. Moreover, Ψ_{k_0} contains three user-chosen parameters such as k_0, $w(\cdot)$, $\mathcal{B}$ whereas V_{k_0} has one user-chosen parameter k_0. Thus V_{k_0} is more convenient and straightforward to implement.

Neither method assumed structural assumptions as (3.2.1), which is imposed in Hu and Tsay (2014). Generally speaking, the two approaches: ours and that in Li, Gao, Li and Yao (2016) have roots from the two approaches to measure conditional mean dependence: indicator function-based and characteristic function-based, which have long existed in econometrics and statistics. See Bierens (1982, 1990), Escanciano (2006) for some representative works on characteristic function-based approaches to specification testing in econometrics, and Stute (1997), Koul and Stute (1999) and Zhu (2003), among others for the use of indicator function approach to model checking in statistics. In general, it seems that neither one dominates the other. We shall examine the finite sample performance and see which metric is more effective in terms of estimating the volatility space and factor number.

3.2.3 Principal Volatility Component Analysis with V_{k_0}

In the context of PVCA, we have two specific goals in order to achieve dimension reduction for conditional variance matrix. One is to identify the number of linear combinations of Y_t that are conditionally variance independent of the past. The other refers to estimating the form of linear combinations of Y_t that exhibit conditional homoscedasticity. It turns out that these two goals can be achieved by doing spectral decomposition of V_{k_0}. Let $(\lambda_j, \gamma_j)_{j=1}^p$ be eigenvalues and eigenvectors of V_{k_0}. Then

$$\gamma_j^T V_{k_0} \gamma_j = \lambda_j, \quad j = 1, \cdots, p.$$

Assume that the rank of V_{k_0} is s, then due to the fact that V_{k_0} is positive semidefinite, $\lambda_j > \lambda_{s+1} = \lambda_{s+2} = \cdots = \lambda_p = 0,\ j = 1, \cdots, s$. This implies that

$$\begin{aligned} & \gamma_j^T V_{k_0} \gamma_j = \sum_{k=1}^{k_0} MDD(Y_t Y_t^T \gamma_j | Y_{t-k})^2 = 0, \quad j = s+1, \cdots, p \\ \Leftrightarrow \quad & E[(\gamma_j^T Y_t)^2 | Y_{t-k}] = E[(\gamma_j^T Y_t)^2] \ \ a.s., \quad j = s+1, \cdots, p, \quad k = 1, \cdots, k_0. \end{aligned}$$

Therefore, the eigenvectors corresponding to zero eigenvalues of V_{k_0} are the linear combinations of Y_t that have conditional homoscedasticity.

To estimate s which is the rank of V_{k_0}, we adopt the ratio-based estimator used in Lam, Yao and Bathia (2011), Lam and Yao (2012) and Li, Gao, Li and Yao (2016). Let $(\widehat{\lambda}_j, \widehat{\gamma}_j)_{j=1}^p$ be the estimates of eigenvalues and eigenvectors of $\widehat{V}_{k_0}$. Let

$$\widehat{s} = argmin_{1 \leq j \leq p-1} \frac{\widehat{\lambda}_{j+1}}{\widehat{\lambda}_j}. \tag{3.2.7}$$

The reason for using ratio-based estimator is that our method has a close connection to Li, Gao, Li and Yao (2016) and that it is fast and easy to implement. Next we present several assumptions and asymptotic results for our method.

Assumption **3.2.1**. *1.* $\lambda_1 > \lambda_2 > \cdots > \lambda_s > 0 = \lambda_{s+1} = \cdots = \lambda_p$.

2. $(Y_t)_{t\in N}$ *is a strictly stationary and* β*-mixing process. There exist* $\delta > 0$ *such that* $E[|Y_t|^{10+5\delta}] < \infty$ *and for* $\delta' \in (0,\delta)$, $\beta(k) = (k^{-(2+\delta')/\delta'})$.

3. $(Y_t)_{t\in N}$ *is a strictly stationary and* m*-dependent process and* $E[|Y_t|^{10}] < \infty$.

Theorem **3.2.1**. *Let conditions 1, 2 in Assumption 3.2.1 hold. Then as* $n \to \infty$*, it holds that*

1. $\widehat{\lambda}_i - \lambda_i = O_p(n^{-1/2})$ *for* $i = 1, \cdots, s$.

2. $\widehat{\gamma}_i - \gamma_i = O_p(n^{-1/2})$ *for* $i = 1, \cdots, s$.

Let conditions 1, 3 in Assumption 3.2.1 hold. Then as $n \to \infty$*, it holds that*

3. $\widehat{\lambda}_i = O_p(n^{-1})$ *for* $i = s+1, \cdots, p$.

Remark **3.2.3**. Theorem 3.2.1 is an analogue of Theorem 4.1 in Lee and Shao (2016), where the same result is obtained for cumulative MDDM. It is worth noting that Theorem 3.2.1 is developed for the fixed p case and it is different from Theorem 1 in Li, Gao, Li and Yao (2016). The latter aurthors have shown that the estimator of volatilty space based on Ψ_{k_0} is consistent based on the metric $d(\widehat{\mathcal{M}}, \mathcal{M})$ in (3.4.1) under the assumption that the number of factors is known and no theoretical results for the estimator of the number of factors are presented in their paper. Here, we derive the convergence rates for estimated eigenvalues and eigenvectors of V_{k_0}, which easily lead to the fact that $P(\widehat{s} \geq s) \to 1$ i.e., the probability of underestimating the true number of factors s goes to zero. However, we are unable to show the consistency of $\widehat{s}$ due to some techinical difficulties.

3.3 Conditionally Uncorrelated Components

To overcome the difficulty which comes from overparameterization in GARCH type models, Fan, Wang and Yao (2008) proposed the so-called conditionally uncorrelated components (CUC) model by assuming that the observed data Y_t is a linear combination of CUCs. Specifically, the CUC model can be formulated as follows.

ASSUMPTION **3.3.1**. *Assume that*

$$Y_t = A_0 Z_t, \quad E[Z_t|\mathcal{F}_{t-1}] = 0, \quad E[Z_{i,t}Z_{j,t}|\mathcal{F}_{t-1}] = 0, \ \forall i \neq j, \tag{3.3.1}$$

where $var(Y_t) = I_p$, $Z_t = (Z_{1,t}, Z_{2,t}, \cdots, Z_{p,t})^T$ are CUCs such that $var(Z_t) = I_p$, A_0 is an orthogonal matrix by construction i.e., $var(Y_t) = A_0 var(Z_t) A_0^T = A_0 A_0^T = I_p$.

By (3.3.1), the following relationship is implied,

$$\Sigma_t = A_0 \Sigma_{z,t} A_0^T, \quad \Sigma_{z,t} = var(Z_t|\mathcal{F}_{t-1}) = diag(\sigma_{1,t}^2, \cdots, \sigma_{p,t}^2),$$

where $\sigma_{i,t}^2 = var(Z_{i,t}|\mathcal{F}_{t-1}), i = 1, \cdots, p$. Therefore, if A_0 is accurately estimated by $\widehat{A}_0$, estimated CUCs can be obtained by $\widehat{Z}_t = \widehat{A}_0^T Y_t$. Due to the fact that CUCs are conditionally uncorrelated upon the past, univariate volatility models are fitted to each estimated component. From the aspect of multivariate volatility modeling, this approach reduces the number of parameters substantially and guarantees positive semidefiniteness of estimated volatility of Y_t. Here our main interest is the orthogonal matrix $A_0 = (a_{01}, \cdots, a_{0p}), a_{0j} \in R^p, j = 1, \cdots, p$ which is not identifiable in terms of the order of $(a_{01}, \cdots, a_{0p})$ and the sign. To measure the closeness of the truth A_0 and its estimator $\widehat{A}_0 = (\widehat{a}_{01}, \cdots, \widehat{a}_{0p})$, we use the D-distance which is invariant of the change of the order and sign, i.e.,

$$D(A_0, \widehat{A}_0) = 1 - \frac{1}{p}\sum_{i=1}^{p} max_{1\leq j\leq p}|a_{0i}^T \widehat{a}_{0j}|. \qquad (3.3.2)$$

Here we are only interested in the first step of CUC analysis which is the estimation of A_0 and we propose an alternative method of estimating A_0 by employing our MDD-based metric. Before introducing our method, we first provide a brief review of the estimation method used in Fan, Wang and Yao (2008). Their approach is based on the fact that Condition (3.3.1) is equivalent to

$$\sum_{W\in\mathcal{B}_t} |E[Z_{i,t}Z_{j,t}I(Y_{t-k}\in W)]| = 0, \qquad (3.3.3)$$

for any π-class $\mathcal{B}_t \subset \mathcal{F}_{t-1}$ such that the σ-algebra generated by $\mathcal{B}_t$ is equal to $\mathcal{F}_{t-1}$. They defined an objective function $\Phi_{k_0}(\cdot)$ as

$$\begin{aligned}\Phi_{k_0}(M) &= \sum_{k=1}^{k_0}\sum_{1\leq i<j\leq p}\sum_{W\in\mathcal{B}} w(W)|E[m_i^T Y_t Y_t^T m_j I(Y_{t-k}\in W)]| \\ &= \sum_{k=1}^{k_0}\sum_{1\leq i<j\leq p}\sum_{W\in\mathcal{B}} w(W)|m_j^T \otimes m_i^T E[vec(Y_t Y_t^T)I(Y_{t-k}\in W)]|,\end{aligned}$$

where $w(\cdot)$ is a nonnegative weight function, $\mathcal{B}$ is a countable sequence of subsets and $M = (m_1, \cdots, m_p), m_i \in R^p, i = 1, \cdots, p$. Correspondingly, estimator of A_0 is $\widehat{A}_0 = argmin_M \widehat{\Phi}_{k_0}(M)$ subject to the constraint that M is orthogonal, where $\widehat{\Phi}_{k_0}(\cdot)$ is the sample counterpart of $\Phi_{k_0}(\cdot)$. Note that this approach suffers from the same drawback as mentioned in Remark 3.2.2 for Ψ_{k_0}-based approach, i.e., the selection of $w(\cdot)$ and $\mathcal{B}$. Here we propose an alternative approach which is relatively simple to implement and can be computationally more efficient.

3.3.1 vec Volatility Martingale Difference Divergence Matrix and Cumulative vecVMDDM

For $Y_t \in R^p$ and $E[Y_t|\mathcal{F}_{t-1}] = 0$, our goal in this section is to estimate an orthogonal matrix $A_0 = (a_{01}, \cdots, a_{0p})$ such that the volatilty matrix of $A_0^T Y_t$ is a diagonal matrix. In other words, $a_{0i}^T Y_t$ and $a_{0j}^T Y_t, i \neq j$ are conditionally uncorrelated given the past values of Y_t, i.e., $E[a_{0i}^T Y_t Y_t^T a_{0j}|\mathcal{F}_{t-1}] = 0$. We can view the above relationship as conditional mean independence of $a_{0i}^T Y_t Y_t^T a_{0j}$ on $\mathcal{F}_{t-1}$. Hence, we can define an alternative MDD-based objective function $G_{k_0}(\cdot)$. For $M = (m_1, \cdots, m_p)$, we define

$$\begin{aligned}
G_{k_0}(M) &= \sum_{k=1}^{k_0} \sum_{1 \le i < j \le p} MDD(m_i^T Y_t Y_t^T m_j | Y_{t-k})^2 \\
&= \sum_{k=1}^{k_0} \sum_{1 \le i < j \le p} MDD(m_j^T \otimes m_i^T vec(Y_t Y_t^T) | Y_{t-k})^2 \\
&= \sum_{k=1}^{k_0} \sum_{1 \le i < j \le p} m_j^T \otimes m_i^T MDDM(vec(Y_t Y_t^T) | Y_{t-k}) m_j \otimes m_i \\
&= \sum_{1 \le i < j \le p} m_j^T \otimes m_i^T \{\sum_{k=1}^{k_0} MDDM(vec(Y_t Y_t^T) | Y_{t-k})\} m_j \otimes m_i \\
&= \sum_{1 \le i < j \le p} (m_j \otimes m_i)^T \otimes (m_j^T \otimes m_i^T) \{vec(\sum_{k=1}^{k_0} MDDM(vec(Y_t Y_t^T) | Y_{t-k}))\}
\end{aligned}$$

This expression motivates us to define vecVMDDM and cumulative vecVMDDM displayed below.

Definition **3.3.1**. *vec Volatility Martingale Difference Divergence Matrix*

Given $Y \in R^p, X \in R^q$ and assume that $Y \in \mathcal{L}^4, X \in \mathcal{L}^2$ and $E[Y|X] = 0$. We define

$$\begin{aligned} vecVMDDM(Y|X) &= MDDM(vec(YY^T)|X) \\ &= -E[\{vec(YY^T - \Sigma)\}\{vec(Y'(Y')^T - \Sigma)\}^T|X - X'|_q]. \end{aligned}$$

Observe that $vecVMDDM(Y|X)$ is a real, symmetric $p^2 \times p^2$ matrix. Based on the sample $(X_t, Y_t)_{t=1}^n$ from the joint distribution of (X, Y), the sample $vecVMDDM(Y|X)$ is defined by

$$vecVMDDM_n(Y|X) = -\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} vec(Y_iY_i^T - \Sigma_n)vec(Y_jY_j^T - \Sigma_n)^T|X_i - X_j|_q$$

Proposition **3.3.1**. Let $Y = (Y_1, \cdots, Y_p)^T \in R^p$, $X \in R^q$ and $VMDDM(Y|X) = [VMDDM(Y|X)_{i,j}]_{i,j=1}^p, vecVMDDM(Y|X) = [vecVMDDM(Y|X)_{i,j}]_{i,j=1}^{p^2}$. Then

1. $vecVMDDM(Y|X)$ is positive semidefinite.

2. $tr(VMDDM(Y|X)) = tr(vecVMDDM(Y|X)) = \sum_{i=1}^p \sum_{j=1}^p MDD(Y_iY_j|X)^2$.

3. $VMDDM(Y|X)_{i,j} = \sum_{k=1}^p vecVMDDM(Y|X)_{(i-1)p+k,(j-1)p+k}$.

According to Proposition 3.3.1, we can retrieve the $p \times p$ matrix $VMDDM(Y|X)$ from the $p^2 \times p^2$ matrix $vecVMDDM(Y|X)$. Moreover, it is easy to show that $vecVMDDM(Y|X)$ is a positive semidefinite matrix due to the nonnegative definiteness of $MDDM$.

Definition **3.3.2**. *Cumulative vec Volatility Martingale Difference Divergence Matrix Given $Y_t \in R^p$, the cumulative vec volatility martingale difference divergence matrix is defined by*

$$vecV_{k_0} = \sum_{l=1}^{k_0} vecVMDDM(Y_t|Y_{t-l})$$

Since vecVMDDM is a positive semidefinite $p^2 \times p^2$ matrix, $vecV_{k_0}$ is also positive semidefinite. Furthermore, the sample estimate of $vecV_{k_0}$ is
$\widehat{vecV}_{k_0} = \sum_{l=1}^{k_0} vecVMDDM_n(Y_t|Y_{t-l})$.

3.3.2 Estimation of CUC Model with $vecV_{k_0}$

In this section, we introduce an alternative approach to estimating A_0 in the CUC model by employing the new matrix object $vecV_{k_0}$ which effectively summarizes conditional variance dependence between two random vectors. Note that

$$G_{k_0}(M) \quad = \quad \sum_{1 \le i < j \le p} (m_j \otimes m_i)^T \otimes (m_j^T \otimes m_i^T) vec(vecV_{k_0})$$

and $G_{k_0}(A_0) = 0$ due to condition (3.3.1). Therefore, our estimator of A_0 is

$$\widehat{A}_0 = argmin_M \widehat{G}_{k_0}(M) \text{ subject to } MM^T = M^T M = I_p,$$

where $\widehat{G}_{k_0}(\cdot)$ is a sample counterpart of $G_{k_0}(\cdot)$ which simply replaces vecVMDDM with $vecVMDDM_n$, i.e.,

$$\widehat{G}_{k_0}(M) \quad = \quad \sum_{1 \le i < j \le p} (m_j \otimes m_i)^T \otimes (m_j^T \otimes m_i^T) vec(\widehat{vecV}_{k_0})$$

In order to remove the constraint that $M^T M = MM^T = I_p$, we present a useful representation of an orthogonal matrix M.

$$M = \Pi_{1 \le i < j \le p} R_{ij}(\theta_{ij}), \quad -\pi \le \theta_{ij} \le \pi,$$

where $R_{ij}(\theta_{ij})$ is an identity matrix I_p with (i,i) and (j,j)th elements being replaced by $cos(\theta_{ij})$ and (i,j), (j,i)th elements being replaced by $sin(\theta_{ij})$, $-sin(\theta_{ij})$, respectively. With this representation of an orthogonal matrix M, the optimization problem with a constraint is transformed into an unconstrainted minimization. This representation is commonly used in the literature of dimension reduction for multivariate time series; see Matteson and Tsay (2011), Fan, Wang and Yao (2008) and Weide (2002), among others.

In our experience, $\widehat{G}_{k_0}(\cdot)$ is simpler and faster to compute than the objective function $\widehat{\Phi}_{k_0}(\cdot)$ used in Fan, Wang and Yao (2008) and estimation of A_0 based on $\widehat{G}_{k_0}(\cdot)$ is computationally cheaper. The computational advantage comes from the separation of $(m_j \otimes m_i)^T \otimes m_j^T \otimes m_i^T$ and $vec(\widehat{vecV}_{k_0})$ in $\widehat{G}_{k_0}$. In other words, $vec(\widehat{vecV}_{k_0})$ only needs to be computed once whereas the original procedure used in Fan, Wang and Yao (2008) needs to calculate $|m_j^T \otimes m_i^T E[vec(Y_tY_t^T)I(Y_{t-k} \in W)]|, W \in \mathcal{B}$ many times due to the fact that $|m_j^T \otimes m_i^T E[vec(Y_tY_t^T)I(Y_{t-k} \in W)]|$ cannot be separated as in our case.

Below we present a theoretical result under suitable moments and dependence conditions on Y_t.

Assumption **3.3.2**. *1. $(Y_t)_{t\in N}$ is a strictly stationary and β-mixing process. There exist $\delta > 0$ such that $E[|Y_t|^{10+5\delta}] < \infty$ and for $\delta' \in (0,\delta)$, $\beta(k) = (k^{-(2+\delta')/\delta'})$.*

2. *There exist a $p \times p$ orthogonal matrix A_0 such that minimizes $G_{k_0}(\cdot)$. Furthermore, the minimum value of $G_{k_0}(\cdot)$ is obtained at an orthogonal matrix A if and only if $D(A_0, A) = 0$. (unique minimizer)*

3. *$G_{k_0}(A_0) - G_{k_0}(A) \leq -aD(A_0, A)$ for any orthogonal matrix A such that $D(A_0, A)$ is smaller than a small but fixed constant and $a > 0$ is a constant.*

THEOREM **3.3.1**. *Let $k_0 \geq 1, p \geq 1$ be fixed integers. Under conditions 1, 2 in Assumption 3.3.2,*

1. $sup_A|\widehat{G}_{k_0}(A) - G_{k_0}(A)| = O_p(n^{-1/2})$ *and* $D(\widehat{A}_0, A_0) \to^p 0$ *as* $n \to \infty$.

   If additionally condition 3 holds, then as $n \to \infty$*, it holds that*

2. $D(\widehat{A}_0, A_0) = O_p(n^{-1/2})$.

Theorem 3.3.1 and Assumption 3.3.2 condition 2 and 3 are analogous to Theorem 1 and Assumptions (b)-(e) in Fan, Wang and Yao (2008). Here we require stronger moment assumption, which seems hard to relax based on our current technical argument.

3.4 Numerical Simulations

In this section, we study the finite sample performance of our VMDDM-based and vecVMDDM-based dimension reduction approaches of a volatility matrix via simulations in Sections 3.4.1 and 3.4.2, respectively. In particular, we focus on the dimension reduction of a volatility matrix by PVCA in Section 3.4.1 and compare with the methods in Hu and Tsay (2014) and Li, Gao, Li and Yao (2016), which are based on Ω_{k_0} and Ψ_{k_0}. In Section 3.4.2, we compare our G_{k_0}-based approach with the Φ_{k_0}-based counterpart by Fan, Wang and Yao (2008). In our simulations, we tried several different values of k_0 to assess the sensitivity of our dimension reduction method with respect to the choice of k_0. For each example, we replicate the simulation 1000 times.

3.4.1 PVCA

In this subsection, our main focus is on estimating the volatility space and number of factors. Four volatility models have been examined and finite sample performance for our V_{k_0}-based approach, Ω_{k_0}-based approach proposed by Hu and Tsay (2014) and Ψ_{k_0}-based approach suggested by Li, Gao, Li and Yao (2016) have been compared. In order to compare the performance of estimating volatility space, we treat the number of factors as known for Example 3.4.1 and Example 3.4.2 following Li, Gao, Li and Yao (2016). For Example 3.4.3 and Example 3.4.4, we consider both cases, i.e., the number of factors as known and unknown. We let $c = 2.5$ (see (3.2.2)) for Ω_{k_0} and $w(\cdot) = \frac{1}{n}$, $\mathcal{B} = \{u \in R^p : |u| \leq |Y_t|, t = 1, \cdots, n\}$ for Φ_{k_0}. Below we report the results for $n = 250, 500, 1000$ and $k_0 = 1, 5$. The following criteria are adopted to measure the estimation accuracy.

- $d(\widehat{\mathcal{M}}, \mathcal{M})$ [Li, Gao, Li and Yao (2016)]

$$d(\widehat{\mathcal{M}}, \mathcal{M}) = \sqrt{1 - \frac{tr(\widehat{A}\widehat{A}^T A A^T)}{s}}, \tag{3.4.1}$$

where s is the number of factors. $d(\widehat{\mathcal{M}}, \mathcal{M})$ is used to measure the discrepancy between $\mathcal{M}(A)$ and $\mathcal{M}(\widehat{A})$. AA^T is a projection matrix onto the linear space $\mathcal{M}(A)$ since $A^T A = I_s$ and $d(\widehat{\mathcal{M}}, \mathcal{M}) \in [0, 1]$. $d(\widehat{\mathcal{M}}, \mathcal{M}) = 1$ if and only if the two spaces are orthogonal with each other and $d(\widehat{\mathcal{M}}, \mathcal{M}) = 0$ if and only if two spaces are identical. Thus, smaller value of $d(\widehat{\mathcal{M}}, \mathcal{M})$ indicates more accurate estimation of volatility space.

- $d(\widehat{A}, A)$ [Li, Gao, Li and Yao (2016)]

$$d(\widehat{A}, A) = 1 - \frac{\{\sum_{t=1}^n (Y_t - \overline{Y}_n)^T \widehat{A} A^T (Y_t - \overline{Y}_n)\}^2}{\{\sum_{t=1}^n (Y_t - \overline{Y}_n)^T \widehat{A}\widehat{A}^T (Y_t - \overline{Y}_n)\}\{\sum_{t=1}^n (Y_t - \overline{Y}_n)^T A A^T (Y_t - \overline{Y}_n)\}}$$

$d(\widehat{A}, A)$ measures the linear dependence of $\widehat{A}^T Y_t$ and $A^T Y_t$ when A is a vector. Here $d(\widehat{A}, A) \in [0,1]$, $d(\widehat{A}, A) = 1$ if $\widehat{A}^T Y_t$ and $A^T Y_t$ are uncorrelated and $d(\widehat{A}, A) = 0$ if $\widehat{A}^T Y_t$ and $A^T Y_t$ are perfectly correlated. Therefore smaller value of $d(\widehat{A}, A)$ corresponds to better estimate of underlying factor series.

Example **3.4.1**. This example is from Li, Gao, Li and Yao (2016). One ARCH(1) time series are generated for X_t, i.e., $X_t = \sigma_t e_t$, $\sigma_t^2 = 1 + 0.9X_{t-1}^2$, where e_t is an iid standard normal sequence. The factor loading matrix $A = [1.0, 0.7, -0.1, -0.7]^T$. Then the data is generated by $Y_t = AX_t + \epsilon_t$, where ϵ_t are iid from $N(0, I_p/p)$. For this example, it can be derived that $var(Y_t|\mathcal{F}_{t-1}) = A\sigma_t^2 A^T + I_p/p$ so the dependence on $Y_{t-1}Y_{t-1}^T$ is quite linear.

Table 3.1: Mean, standard error (*in the bracket*) of d-distance of Example 3.4.1

		Ω_{k_0}	Ψ_{k_0}	V_{k_0}
$d(\widehat{\mathcal{M}}, \mathcal{M})$	$n = 250, k_0 = 1$	0.0452 (0.0296)	0.0373 (0.0241)	0.0410 (0.0247)
	$n = 250, k_0 = 5$	0.0447 (0.0278)	0.0373 (0.0226)	0.0400 (0.0234)
$d(\widehat{A}, A)$	$n = 250, k_0 = 1$	0.0002 (0.0005)	0.0001 (0.0003)	0.0002 (0.0003)
	$n = 250, k_0 = 5$	0.0002 (0.0004)	0.0001 (0.0002)	0.0001 (0.0002)
$d(\widehat{\mathcal{M}}, \mathcal{M})$	$n = 500, k_0 = 1$	0.0276 (0.0158)	0.0229 (0.0134)	0.0259 (0.0145)
	$n = 500, k_0 = 5$	0.0280 (0.01595)	0.0235 (0.0134)	0.0260 (0.0144)
$d(\widehat{A}, A)$	$n = 500, k_0 = 1$	5.966e-05 (9.833e-05)	4.226e-05 (6.912e-05)	5.155e-05 (7.781e-05)
	$n = 500, k_0 = 5$	6.088e-05 (9.521e-05)	4.288e-05 (6.392e-05)	5.127e-05 (7.277e-05)
$d(\widehat{\mathcal{M}}, \mathcal{M})$	$n = 1000, k_0 = 1$	0.0184 (0.0100)	0.0143 (0.0075)	0.0170 (0.0089)
	$n = 1000, k_0 = 5$	0.0187 (0.0102)	0.0152 (0.0080)	0.0173 (0.0092)
$d(\widehat{A}, A)$	$n = 1000, k_0 = 1$	2.260e-05 (3.043e-05)	1.334e-05 (1.801e-05)	1.873e-05 (2.377e-05)
	$n = 1000, k_0 = 5$	2.332e-05 (3.178e-05)	1.502e-05 (2.039e-05)	1.953e-05 (2.684e-05)

As seen from Table 3.1, Ψ_{k_0}-based approach slightly outperforms the other two approaches and our method slightly outperforms Ω_{k_0}-based approach, although all three methods are very comparable. It is interesting that Ψ_{k_0} and V_{k_0}-based approaches are slightly superior to Ω_{k_0}-based counterpart in terms of $d(\widehat{\mathcal{M}}, \mathcal{M})$ and $d(\widehat{A}, A)$, since the dependence of volatility process over the past is fairly linear and therefore Ω_{k_0}-based approach is expected to perform well. Overall, when n increases, $d(\widehat{\mathcal{M}}, \mathcal{M})$ and $d(\widehat{A}, A)$ get smaller for all methods. Furthermore, it shows that all three methods seem to have consistent performances with respect to the choice of k_0.

Example **3.4.2**. In this example, the linear ARCH(1) model for X_t in Example 3.4.1 is replaced by a nonlinear model, i.e., a stochastic volatility model.

$$X_t = e^{h_t/2}, \quad h_t = 0.3 + 0.6(h_{t-1} - 0.3) + 0.3\eta_t,$$

where e_t, η_t are all iid standard normal sequences and independent from each other. An examination of the sufficient conditions to ensure the stationarity for stochastic volatility model (see Equation (3.40) in Chapter 3.12 of Tsay (2010)) shows that the above model admits a stationary solution. The data is generated by $Y_t = AX_t + \epsilon_t$, where ϵ_t are iid from $N(0, I_p/p)$. Like Example 3.4.1, we consider $p = 4$ and $var(Y_t|\mathcal{F}_{t-1}) = Ae^{h_t}A^T + I_p/p$ which depends on $(Y_{t-j}Y_{t-j}^T)_{j=1}^{\infty}$ in a very nonlinear fashion.

Table 3.2: Mean, standard error (*in the bracket*) of d-distance of Example 3.4.2

		Ω_{k_0}	Ψ_{k_0}	V_{k_0}
$d(\widehat{\mathcal{M}}, \mathcal{M})$	$n = 250, k_0 = 1$	0.2601 (0.1235)	0.2220 (0.1005)	0.1367 (0.0639)
	$n = 250, k_0 = 5$	0.1384 (0.0612)	0.1215 (0.0529)	0.0899 (0.0499)
$d(\widehat{A}, A)$	$n = 250, k_0 = 1$	0.0163 (0.0475)	0.0100 (0.0102)	0.0036 (0.0036)
	$n = 250, k_0 = 5$	0.0036 (0.0035)	0.0028 (0.0025)	0.0024 (0.0291)
$d(\widehat{\mathcal{M}}, \mathcal{M})$	$n = 500, k_0 = 1$	0.2487 (0.1201)	0.2198 (0.1011)	0.1244 (0.0577)
	$n = 500, k_0 = 5$	0.1284 (0.0565)	0.1137 (0.0495)	0.0733 (0.0307)
$d(\widehat{A}, A)$	$n = 500, k_0 = 1$	0.0140 (0.0347)	0.0098 (0.0103)	0.0029 (0.0030)
	$n = 500, k_0 = 5$	0.0030 (0.0028)	0.0024 (0.0022)	0.0010 (0.0008)
$d(\widehat{\mathcal{M}}, \mathcal{M})$	$n = 1000, k_0 = 1$	0.2443 (0.1175)	0.2109 (0.0967)	0.1157 (0.0536)
	$n = 1000, k_0 = 5$	0.1236 (0.0528)	0.1067 (0.0464)	0.0662 (0.0280)
$d(\widehat{A}, A)$	$n = 1000, k_0 = 1$	0.0133 (0.0333)	0.0087 (0.0086)	0.0025 (0.0023)
	$n = 1000, k_0 = 5$	0.0028 (0.0024)	0.0021 (0.0018)	0.0008 (0.0006)

From Table 3.2, our V_{k_0}-based approach outperforms the other two methods in all cases with substantially smaller $d(\widehat{\mathcal{M}}, \mathcal{M})$ and $d(\widehat{A}, A)$. By comparison, the Ω_{k_0}-based approach appears inferior to V_{k_0}-based and Ψ_{k_0}-based counterparts, which is presumably due to its inability to capture strong nonlinear dependence of the volatility process. Notice that both Ψ_{k_0}-based approach and V_{k_0}-based one aim to capture not only linear but also nonlinear dependence of volatility process and this example is the case where volatility appears to have nonlinear dependence. It is interesting that our V_{k_0}-based approach performs significantly better than Ψ_{k_0}-based counterpart, suggesting that V_{k_0} summarizes dependence of volatility more efficiently than Ψ_{k_0} for this case. When k_0 increases, the ability to estimate the true volatility space improves for all methods, showing sensitivity with respect to the choice of k_0.

Example **3.4.3**. This example is also from Li, Gao, Li and Yao (2016). The factor $X_t = (x_{1,t}, x_{2,t})^T$ is generated by two ARCH(1) processes.

$$\begin{aligned} x_{1,t} &= \sigma_{1,t} e_{1,t}, \quad \sigma_{1,t}^2 = 1 + 0.8 x_{1,t-1}^2 \\ x_{2,t} &= \sigma_{2,t} e_{2,t}, \quad \sigma_{2,t}^2 = 2 + 0.9 x_{2,t-1}^2 \end{aligned}$$

where $e_{i,t}, i = 1, 2$ are all iid standard normal variables. For the factor loading matrix,

$$A = \begin{pmatrix} 0 & 0.7 \\ \sqrt{2}/2 & -0.1 \\ 0 & -0.7 \\ \sqrt{2}/2 & 0.1 \end{pmatrix}$$

The data is defined by $Y_t = AX_t + \epsilon_t$, where $\epsilon_{i,t}, i = 1, 2$ are iid from $N(0, I_p/p)$ and independent from $e_{i,t}, i = 1, 2$.

Table 3.3: Mean, standard error (*in the bracket*) of d-distance of Example 3.4.3 when the number of factors is known and unknown

		Ω_{k_0}	Ψ_{k_0}	V_{k_0}
$d(\widehat{\mathcal{M}}, \mathcal{M})$	$n = 250, k_0 = 1$	0.0433 (0.0371)	0.0382 (0.0246)	0.0537 (0.0593)
	$n = 250, k_0 = 5$	0.0505 (0.0491)	0.0344 (0.0223)	0.0475 (0.0564)
$d(\widehat{\mathcal{M}}, \mathcal{M})$	$n = 500, k_0 = 1$	0.0264 (0.0258)	0.0229 (0.0144)	0.0360 (0.0454)
	$n = 500, k_0 = 5$	0.0366 (0.0515)	0.0237 (0.0139)	0.0333 (0.0413)
$d(\widehat{\mathcal{M}}, \mathcal{M})$	$n = 1000, k_0 = 1$	0.0169 (0.0105)	0.0141 (0.0074)	0.0221 (0.0332)
	$n = 1000, k_0 = 5$	0.0232 (0.0311)	0.0151 (0.0078)	0.0220 (0.0264)

	Ω_{k_0}				Ψ_{k_0}				V_{k_0}			
	$d(\widehat{\mathcal{M}}, \mathcal{M})$	$\widehat{s}$			$d(\widehat{\mathcal{M}}, \mathcal{M})$	$\widehat{s}$			$d(\widehat{\mathcal{M}}, \mathcal{M})$	$\widehat{s}$		
		$\widehat{s} = 1$	$\widehat{s} = 2$	$\widehat{s} = 3$		$\widehat{s} = 1$	$\widehat{s} = 2$	$\widehat{s} = 3$		$\widehat{s} = 1$	$\widehat{s} = 2$	$\widehat{s} = 3$
$n = 250, k_0 = 1$	0.2939 (0.3282)	0.386	0.614	0	0.1834 (0.2805)	0.222	0.778	0	0.2956 (0.3281)	0.386	0.609	0.005
$n = 250, k_0 = 5$	0.3405 (0.3355)	0.455	0.545	0	0.2551 (0.3184)	0.331	0.669	0	0.2944 (0.3286)	0.384	0.607	0.009
$n = 500, k_0 = 1$	0.2440 (0.3209)	0.324	0.676	0	0.0954 (0.2133)	0.108	0.892	0	0.2684 (0.3287)	0.359	0.64	0.001
$n = 500, k_0 = 5$	0.3085 (0.3382)	0.418	0.582	0	0.2041 (0.3039)	0.267	0.733	0	0.2766 (0.3287)	0.37	0.626	0.004
$n = 1000, k_0 = 1$	0.1825 (0.2966)	0.242	0.758	0	0.0355 (0.1204)	0.031	0.969	0	0.2213 (0.3168)	0.298	0.701	0.001
$n = 1000, k_0 = 5$	0.2644 (0.3322)	0.36	0.64	0	0.1134 (0.2428)	0.143	0.857	0	0.2579 (0.3306)	0.351	0.648	0.001

According to Table 3.3, when the number of factors is known, the performances of Ω_{k_0}-based, Ψ_{k_0}-based and V_{k_0}-based methods are comparable for $k_0 = 1$ and 5 with Ψ_{k_0}-based approach slightly outperforming the other two. For this example, the true number of factors s is 2 and if we treat the number of factors as unknown, Ψ_{k_0}-based approach outperforms the other two approaches in terms of smaller $d(\widehat{\mathcal{M}}, \mathcal{M})$ and higher proportion of correctly identifying the number of factors. However, it seems that the sensitivity of Ψ_{k_0}-based approach with respect to the choice of k_0 is quite high as compared to the other two methods.

EXAMPLE **3.4.4**. In this example, the two linear ARCH(1) models for $X_t = (x_{1,t}, x_{2,t})^T$ in Example 3.4.3 are replaced by nonlinear models, i.e., stochastic volatility models

$$\begin{aligned} x_{1,t} &= e^{h_{1,t}/2}, \quad h_{1,t} = 0.3 + 0.6(h_{1,t-1} - 0.3) + 0.3\eta_{1,t} \\ x_{2,t} &= e^{h_{2,t}/2}, \quad h_{2,t} = 0.5 + 0.6(h_{2,t-1} - 0.5) + 0.5\eta_{2,t} \end{aligned}$$

where $\eta_{i,t}, i = 1, 2$ are all iid standard normal sequences and independent from each other. Still the data is generated by $Y_t = AX_t + \epsilon_t$, where ϵ_t are iid from $N(0, I_p/p)$ and independent from $\eta_{i,t}, i = 1, 2$. Like Example 3.4.3, we consider $p = 4$ and $var(Y_t|\mathcal{F}_{t-1}) = A \begin{pmatrix} e^{h_{1,t}} & 0 \\ 0 & e^{h_{2,t}} \end{pmatrix} A^T + I_p/p$ which depends on $(Y_{t-j}Y_{t-j}^T)_{j=1}^{\infty}$ in a very nonlinear fashion.

From Table 3.4, if the number of factors is known, we see that V_{k_0}-based approach outperforms Ω_{k_0}-based and Ψ_{k_0}-based approach, presumably due to its capability of capturing strong nonlinear dependence of volatility. When k_0 increases, performances of all methods enhance significantly. When the number of factors is unknown, V_{k_0}-based method is still superior to Ω_{k_0}-based and Ψ_{k_0}-based methods in terms of $d(\widehat{\mathcal{M}}, \mathcal{M})$ and the proportion of correctly identifying the number of factors. Recall that this example has strong nonlinear dependence of volatility. Thus limited simulation evidence seems to suggest that V_{k_0}-based approach is more efficiently dealing with nonlinear dependence of volatility than Ψ_{k_0}-based and Ω_{k_0}-based counterparts.

Table 3.4: Mean, standard error (*in the bracket*) of d-distance of Example 3.4.4 when the number of factors is known and unknown

		Ω_{k_0}	Ψ_{k_0}	V_{k_0}
$d(\widehat{\mathcal{M}}, \mathcal{M})$	$n = 250, k_0 = 1$	0.1589(0.0694)	0.1599 (0.0682)	0.0957 (0.0461)
	$n = 250, k_0 = 5$	0.0888 (0.0355)	0.0853 (0.0338)	0.0654 (0.0376)
$d(\widehat{\mathcal{M}}, \mathcal{M})$	$n = 500, k_0 = 1$	0.1467 (0.0674)	0.1527 (0.0673)	0.0859 (0.0394)
	$n = 500, k_0 = 5$	0.0785 (0.0295)	0.0794 (0.0295)	0.0526 (0.0285)
$d(\widehat{\mathcal{M}}, \mathcal{M})$	$n = 1000, k_0 = 1$	0.1402 (0.0623)	0.1457 (0.0638)	0.0794 (0.0317)
	$n = 1000, k_0 = 5$	0.0734 (0.0288)	0.0766 (0.0297)	0.0453 (0.0257)

	Ω_{k_0}				Ψ_{k_0}				V_{k_0}			
	$d(\widehat{\mathcal{M}}, \mathcal{M})$	$\widehat{s}$			$d(\widehat{\mathcal{M}}, \mathcal{M})$	$\widehat{s}$			$d(\widehat{\mathcal{M}}, \mathcal{M})$	$\widehat{s}$		
		$\widehat{s} = 1$	$\widehat{s} = 2$	$\widehat{s} = 3$		$\widehat{s} = 1$	$\widehat{s} = 2$	$\widehat{s} = 3$		$\widehat{s} = 1$	$\widehat{s} = 2$	$\widehat{s} = 3$
$n = 250, k_0 = 1$	0.2692 (0.2416)	0.219	0.709	0.072	0.2681 (0.2397)	0.213	0.716	0.071	0.130 (0.1505)	0.058	0.934	0.008
$n = 250, k_0 = 5$	0.1313 (0.1621)	0.07	0.928	0.002	0.1120 (0.1322)	0.044	0.955	0.001	0.0715 (0.0738)	0.01	0.988	0.002
$n = 500, k_0 = 1$	0.2791 (0.2554)	0.25	0.701	0.049	0.2783 (0.2507)	0.24	0.692	0.068	0.1171 (0.1412)	0.05	0.947	0.003
$n = 500, k_0 = 5$	0.1114 (0.1442)	0.053	0.947	0	0.1002 (0.1178)	0.034	0.963	0.003	0.0532 (0.0351)	0.001	0.998	0.001
$n = 1000, k_0 = 1$	0.3139 (0.2746)	0.317	0.653	0.03	0.3170 (0.2720)	0.314	0.634	0.052	0.1037 (0.1258)	0.039	0.96	0.001
$n = 1000, k_0 = 5$	0.1198 (0.1689)	0.074	0.926	0	0.0971 (0.1167)	0.033	0.966	0.001	0.0453 (0.0256)	0	0.999	0.001

3.4.2 CUC

In this subsection, our goal is to estimate an orthogonal matrix $A_0 = (a_{01}, \cdots, a_{0p})$ which transforms the multivariate time series into conditionally uncorrelated components. We consider two different methods of estimating a constant matrix A_0, our G_{k_0}-based approach and Φ_{k_0}-based approach used by Fan, Wang and Yao (2008). Two different volatility processes are generated with $n = 500$ or 1000 and $k_0 = 1$ or 5. For each example, mean and standard error of D-distances [see (3.3.2)] are computed in order to measure the precision of $\widehat{A}_0$. Observe that $D(\widehat{A}_0, A_0) \in [0,1]$ for any orthogonal matrices A_0 and $\widehat{A}_0$. Moreover, if A_0 is obtained by permuting or reflecting the columns of $\widehat{A}_0$, then $D(\widehat{A}_0, A_0) = 0$. Similary, $D(\widehat{A}_0, A_0) = 1$ if and only if the two matrices A_0 and $\widehat{A}_0$ are orthogonal with each other. Hence, smaller value of $D(\widehat{A}_0, A_0)$ refers to a better estimate of an orthogonal matrix A_0.

Example **3.4.5**. This example is from simulation section of Fan, Wang and Yao (2008). Three GARCH(1,1) processes are generated for CUCs $Z_t = (z_{1,t}, z_{2,t}, z_{3,t})^T$.

$$\begin{aligned} z_{1,t} &= \epsilon_{1,t}\sigma_{1,t}, \quad \sigma_{1,t}^2 = 0.02 + 0.9\sigma_{1,t-1}^2 + 0.04z_{1,t-1}^2 + 0.04z_{3,t-1}^2 \\ z_{2,t} &= \epsilon_{2,t}\sigma_{2,t}, \quad \sigma_{2,t}^2 = 0.1 + 0.8\sigma_{2,t-1}^2 + 0.1z_{2,t-1}^2 \\ z_{3,t} &= \epsilon_{3,t}\sigma_{3,t}, \quad \sigma_{3,t}^2 = 0.28 + 0.6\sigma_{3,t-1}^2 + 0.12z_{3,t-1}^2 \end{aligned}$$

where $\epsilon_{i,t}, i = 1,2,3$ are iid standard normal. Furthermore, the transformation matrix A_0 is set to be

$$A_0 = \begin{pmatrix} 0 & 0.5 & 0.866 \\ 0 & 0.866 & -0.5 \\ -1 & 0 & 0 \end{pmatrix},$$

which is orthogonal. As $var(\epsilon_{i,t}) = 1, i = 1,2,3$ and the sum of coefficients of $\sigma_{i,t}^2$ and $z_{i,t}^2$ is smaller than 1 for $i = 1,2,3$, Z_t admits a stationary solution. Recall that the data Y_t is generated by $Y_t = A_0 Z_t$ and the conditional distribution $Z_t|\mathcal{F}_{t-1}$ follows $N(0, diag(\sigma_{1,t}^2, \sigma_{2,t}^2, \sigma_{3,t}^2))$.

From Table 3.5, means and standard errors of both Φ_{k_0} and G_{k_0}-based approaches are small for all cases which indicates that the estimation of A_0 is reasonably accurate. Overall, if n increases, then both methods produce better estimates of A_0 as D-distances decreases. For most cases, both methods are comparable in terms of estimating the transformation matrix A_0 and sometimes Φ_{k_0}-based approach is better than G_{k_0}-based counterpart.

EXAMPLE **3.4.6**. In this example, we replace $Z_t = (z_{1,t}, z_{2,t}, z_{3,t})^T$ with the following nonlinear volatility process TGARCH(1,1). Then the data $Y_t = A_0 Z_t$, where A_0 is defined in Example 3.4.5. Here $Z_t|\mathcal{F}_{t-1}$ follows $N(0, diag(\sigma^2_{1,t}, \sigma^2_{2,t}, \sigma^2_{3,t}))$, where

$$\begin{aligned}
z_{1,t} &= \epsilon_{1,t}\sigma_{1,t}, \quad \sigma^2_{1,t} = 0.02 + 0.4\sigma^2_{1,t-1} + 0.36z^2_{1,t-1}S_{1,t-1} + 0.04z^2_{1,t-1} \\
z_{2,t} &= \epsilon_{2,t}\sigma_{2,t}, \quad \sigma^2_{2,t} = 0.1 + 0.8\sigma^2_{2,t-1} + 0.09z^2_{2,t-1}S_{2,t-1} + 0.01z^2_{2,t-1} \\
z_{3,t} &= \epsilon_{3,t}\sigma_{3,t}, \quad \sigma^2_{3,t} = 0.28 + 0.6\sigma^2_{3,t-1} + 0.27z^2_{3,t-1}S_{3,t-1} + 0.03z^2_{3,t-1}
\end{aligned}$$

with

$$S_{i,t-1} = \begin{cases} 1 & \text{if } z_{i,t-1} < 0 \\ 0 & \text{if } z_{i,t-1} \geq 0 \end{cases}, \quad i = 1, 2, 3.$$

Table 3.5: Mean, standard error (*in the bracket*) of D-distance of Example 3.4.5 and Example 3.4.6

Example 3.4.5	Φ_{k_0}	G_{k_0}
$n = 500, k_0 = 1$	0.1017 (0.0769)	0.1363 (0.0774)
$n = 500, k_0 = 5$	0.0982 (0.0750)	0.1297 (0.0792)
$n = 1000, k_0 = 1$	0.0877 (0.0735)	0.1219 (0.0780)
$n = 1000, k_0 = 5$	0.0778 (0.0692)	0.1075 (0.0740)
Example 3.4.6	Φ_{k_0}	G_{k_0}
$n = 500, k_0 = 1$	0.1793 (0.2933)	0.1484 (0.3002)
$n = 500, k_0 = 5$	0.0835 (0.1415)	0.0636 (0.1405)
$n = 1000, k_0 = 1$	0.0645 (0.0617)	0.0180 (0.0219)
$n = 1000, k_0 = 5$	0.0483 (0.0508)	0.0152 (0.0214)

According to Table 3.5, when n increases, D-distance gets smaller for both methods which demonstrates that the ability to estimate A_0 improves. If $n = 500$, the finite performance of both methods are comparable and both provide fairly good estimator since D-distances are small. We can see that when $n = 1000$, our method noticeably outperforms the Φ_{k_0}-based approach with smaller D-distance.

We shall summarize the findings based on limited simulations. (1) For a dimension reduction of volatility by the PVCA method, V_{k_0}-based approach can be superior to the existing methods (i.e., Ω_{k_0}, Ψ_{k_0}-based counterparts) if the volatility exhibits strong nonlinear dependence when the number of factors are known or estimated. (2) Even when the volatility seems to be quite linearly dependent, V_{k_0}-based approach performs slightly better than Ω_{k_0}-based counterparts but is slightly inferior to Ψ_{k_0}-based one. (3) For the estimation of CUC model, if the volatility dependence is fairly linear, Φ_{k_0}-based approach produces better estimate of the transformation matrix A_0 than our G_{k_0}-based counterpart. But G_{k_0}-based method can outperform Φ_{k_0}-based one in certain nonlinear dependence cases.

3.5 Data Illustrations

In this section, we further compare our approach with the existing counterparts via two real stocks data sets. These two real data sets have been analyzed by Li, Gao, Li and Yao (2016) and Fan, Wang and Yao (2008), respectively.

3.5.1 6 Stocks Data

The first data set is the daily log returns of six stocks from January 2nd, 2002 to July 10th, 2008: Bank of America Corporation, Dell Inc., JPMorgan Chase & Co., FedEx Corporation, McDonald's Corporation, American International Group. The length of the daily log returns is $n = 1642$ and the dimension is $p = 6$. We apply Ω_{k_0}-based, Ψ_{k_0}-based and V_{k_0}-based approaches to this data set with $k_0 = 5$ as Li, Gao, Li and Yao (2016) did. All three methods estimate the number of factors $\widehat{s} = 1$ which means that there is one factor series describing the volatility behavior of six different daily log returns. Table 3.6 displays the ratio of eigenvalues of each approach which convinces us that there is one underlying factor series from this data set. Table 3.7 reports estimated factor loading matrices for three methods which appear quite similar.

Table 3.6: Ratios of eigenvalues for 6 Stocks Data

	Ω_{k_0}	Ψ_{k_0}	V_{k_0}
λ_2/λ_1	0.0538	0.0271	0.0469
λ_3/λ_2	0.8584	0.8185	0.9649
λ_4/λ_3	0.3057	0.7097	0.6156
λ_5/λ_4	0.6653	0.6559	0.8170
λ_6/λ_5	0.5005	0.3200	0.4926

Table 3.7: Estimates of factor loading matrix for 6 Stocks Data

	Ω_{k_0}	Ψ_{k_0}	V_{k_0}
Bank of America Corporation	0.3184	0.3922	0.3681
Dell Inc.	0.3408	0.3138	0.3173
JPMorgan Chase & Co.	0.6834	0.6492	0.6752
FedEx Corporation	0.2033	0.2224	0.2155
McDonald's Corporation	0.1898	0.1263	0.1289
American International Group	0.4880	0.5107	0.4949

3.5.2 4 Stocks Data

Fan, Wang and Yao (2008) have examined this data set which is the daily log returns of four stocks from January 2nd, 1991 to December 31st, 2000: Standard and Poors 500 index, Cisco System, Intel Corporation and Sprint. Therefore, the length of time series is $n = 2527$ and the dimension of the data is $p = 4$. In order to remove the conditional mean of dailiy log returns, VAR(2) is fitted to the log return series and the normalized residual series (in terms of having an identity variance matrix) are considered as Y_t. We applied Φ_{k_0}-based and G_{k_0}-based approach with $k_0 = 1, 5$ to estimate the transformation matrix A_0. The order of VAR model was chosen by Fan, Wang and Yao (2008) using AIC and $M(i)$ in Tiao and Box (1981) which is a test statistic testing whether the data is a stationary VAR(i) model.

Estimates of A_0 with Φ_{k_0}-based and G_{k_0}-based approachs are shown in Table 3.8 and in order to measure the dissimilarity of two estimates, $D(\widehat{A}_0^{\Phi}, \widehat{A}_0^{G})$ is computed which is 0.0110 when $k_0 = 1$ and 0.0047 when $k_0 = 5$, where $\widehat{A}_0^{\Phi}$ is an estimate of A_0 with Φ_{k_0}-based approach and $\widehat{A}_0^{G}$ is an estimate of A_0 with G_{k_0}-based counterparts. It seems that $\widehat{A}_0^{\Phi}$ and $\widehat{A}_0^{G}$ are similar as seen from the small D-distance.

Table 3.8: Estimates of $A_0 = (a_{01}, a_{02}, a_{03}, a_{04})$ for 4 Stocks Data

$\Phi_{k_0}, k_0 = 1$				$\Phi_{k_0}, k_0 = 5$			
a_{01}	a_{02}	a_{03}	a_{04}	a_{01}	a_{02}	a_{03}	a_{04}
-0.33457	-0.24585	0.26645	0.86984	-0.33259	-0.26747	0.28013	0.85987
0.93745	0.0040591	0.063769	0.34219	0.93907	-0.011209	0.075579	0.33512
-0.088254	0.9692	0.080919	0.2152	-0.081316	0.9635	0.081392	0.24173
-0.038095	0.013752	-0.95833	0.28279	-0.030218	0.0027784	-0.95352	0.29981
$G_{k_0}, k_0 = 1$				$G_{k_0}, k_0 = 5$			
a_{01}	a_{02}	a_{03}	a_{04}	a_{01}	a_{02}	a_{03}	a_{04}
-0.4286	0.77203	0.24497	0.40032	0.35647	0.83266	0.24057	0.3489
0.88559	0.45384	0.097962	0.012946	0.065864	0.35139	0.039431	-0.93308
-0.1751	0.34634	0.099429	-0.91624	-0.92908	0.32207	0.17065	0.062921
-0.037157	0.27936	-0.95943	0.0085813	-0.07353	0.2819	-0.9547	0.060626

3.6 Discussions and Conclusions

In this paper, we proposed two new matrix objects, the so-called volatility martingale difference divergence matrix and vec volatility martingale difference divergence matrix which measure the conditional variance dependence of random vectors $Y \in R^p$ on $X \in R^q$ under the assumption that $E[Y|X] = 0$. The VMDDM and vecVMDDM can be viewed as extensions of MDDM proposed by Lee and Shao (2016) which measures the conditional mean dependence. We apply the VMDDM and its cumulative version to PVCA following the work by Hu and Tsay (2014) and Li, Gao, Li and Yao (2015), and the vecVMDDM to the estimation of CUC model proposed by Fan, Wang and Yao (2008). Simulation results suggest that our MDD-based approach performs comparably well and it can outperform the existing counterparts when the volatility dependence is strongly nonlinear. Further our new MDD-based matrix objects are simple to calculate, and have advantages in terms of computational time and convenience of implementation. Theoretical results are also obtained under suitable moment and weak dependence conditions and they provide good justification for the large sample behavior of our estimators.

We shall conclude by mentioning several future directions. It would be interesting to investigate the choice of k_0 in PVCA as we see it can have an impact on the finite sample performance. A data-driven choice of k_0 that works well in the case of strong linear/nonlinear dependence is needed. It would be important to understand the behavior of the proposed approaches when the dimension p is high from both theoretical and numerical angles. High dimensional stock return time series are nowadays very common, so an extension to allow high dimension would be practically relevant but seems challenging. Another related issue is that we assume stationarity throughout the paper. Given the nonstationarity of many real time series, it would be useful to come up with a dimension reduction approach that accommodate nonstationarity. We leave these topics for future work.

Chapter 4

Testing the Conditional Mean Independence for Functional Data

4.1 Background

Functional data analysis (FDA) has emerged as an important area of statistics which provides convenient and informative tools for the analysis of data objects of high or infinite dimension. It is generally applicable to problems which are difficult to cast into a framework of scalar or vector observations. In many situations, even if standard scalar or vector based approaches are applicable, the functional data based approach can often provide a more natural and parsimonious description of the data, and lead to more accurate inference and prediction. The area of FDA has been growing rapidly in the recent decade since Ramsay and Silverman's (2005) excellent monograph, which provides a systematic account of the existing methodologies and tools to deal with data of functional nature. See Ferraty and Vieu (2010), Horváth and Kokoszka (2012), and Kokoszka and Reimherr (2017) for recent book-length treatments of FDA.

In the literature, functional linear model with scalar or functional response Y and functional or vector covariates X have been extensively studied; see e.g. Cuevas, Febrero, and Fraiman (2002), Cardot et al. (2003), Chiou, Müller, and Wang (2004), Müller and Stadtmüller (2005), Yao, Müller, and Wang (2005a, 2005b), Cai and Hall (2006), Chiou and Müller (2007), among others. There has also been extensions of nonparametric regression models and inference to functional data; see e.g. Ferraty et al. (2011), Lian (2011), and Ferraty, Van Keilegom, and Vieu (2012). Most of the above-mentioned papers focus on modeling the conditional mean of the response variable Y given the covariates X using either linear model or nonparametric models. An important problem in conditional mean modeling is to assess whether X contributes to the conditional mean of Y, i.e., whether we have enough evidence to reject the following null hypothesis

$$H_0 : E[Y|X] = E[Y], \text{ almost surely}$$

based on a random sample $(X_i, Y_i)_{i=1}^n$. If H_0 is supported by the data, then there is no need to pursue a regression model for the mean of Y given X. In this paper, we shall address this testing problem when both Y and X can be either function-valued or vector-valued. It is worth noting that our test can be extended to do diagnostic checking for functional linear models but we shall leave that to future work.

To the best of our knowledge, the above testing problem has been first investigated by Kokoszka et al. (2008) for functional response and functional covariates. Specifically, they assumed a functional linear model, i.e., $Y(t) = \int_0^1 \varphi(t,s)X(s)ds + \epsilon(t)$, $t \in [0,1]$, where $\epsilon(\cdot)$ is an error process that is independent of the covariates and $\varphi(\cdot,\cdot)$ is a square integrable function on $[0,1] \times [0,1]$. They proposed a χ^2-based test for the nullity of the φ, i.e., $H_0 : \varphi(s,t) = 0, \ \forall s,t$, which implies conditional mean independence of Y given X under the linear model assumption. Their procedure relies on the use of functional principal component analysis (FPCA) for both X and Y, and their test statistic measures the correlation of the finite-dimensional scores of X and Y. More recently, Patilea et al. (2016) introduced a nonparametric test for the predictor effect on a functional response allowing covariates to be either function-valued or vector-valued. Their test is nonparametric in the sense that no linear model assumption is imposed, but it requires the choice of 5 user-chosen quantities when X is function-valued and its implementation seems quite complex. Similar to Kokoszka et al. (2008), their test also projects the functional data to a finite dimensional space and constructs test statistics via the finite dimensional projections. Thus these two existing tests may have low power when the dependence of Y on X is along the directions that are orthogonal to the ones used. In the related diagnostic checking problem for functional linear models, Chiou and Müller (2007) proposed a randomization test and recommended to use residual plots based on functional principal component scores of residual processes for diagnostic purposes; Gabrys et al. (2010) proposed goodness-of-fit test statistics that aim to detect serial correlation in the error.

In this article, we shall introduce a new nonparametric test to test H_0 versus

$$H_1 : P(E(Y|X) = E(Y)) < 1,$$

where both the response Y and the covariate X can be either function-valued or vector-valued. The main contribution of our work lies in the following aspects: (1) we first generalize the martingale difference divergence (MDD, hereafter) [Shao and Zhang (2014); Park, Shao and Yao (2015)], which characterizes the conditional mean independence of Y given X when both X and Y are vector-valued, to the functional setting. Note that MDD can be viewed as an analogue of distance covariance [Székely, Rizzo, and Bakirov 2007], which measures the (in)dependence of two random vectors. The so-called functional martingale difference divergence (FMDD) is shown to fully characterize the conditional mean independence based on certain results developed by Lyons (2013), who extended the distance covariance from Euclidean space to metric space. (2) We then define the $\mathcal{U}$-centering [Székely and Rizzo (2014)] based sample estimate of FMDD, which is shown to be unbiased, and its limiting null distribution is shown to be nonpivotal; (3) We propose a wild bootstrap approach to approximate the limiting null distribution, and asymptotic behavior of bootstrap test statistic is carefully studied under both the null and alternatives. In particular, bootstrap consistency under the null and limiting power under the local alternative that is in the n^{-a}, $a > 0$ neighborhood of the null hypothesis is derived. An appealing feature of our test is that there is no tuning parameter or user-chosen number involved, and the test does not impose any linear or parametric model assumption so it is model-free. Through numerical simulations, we show that our test has accurate size and fairly high power relative to the tests developed by Kokoszka et al. (2008) and Patilea et al. (2016).

The rest of this paper is organized as follows. Section 4.2 introduces functional martingale difference divergence (FMDD) as an analog of MDD and its sample version to construct the test statistic. In Section 4.3, we describe the testing procedure including the use of wild bootstrap to obtain the critical values and establish asymptotic validity of the test. Simulation results are presented in Section 4.4 to examine the finite sample performance of the new test in comparison with the tests developed by Kokoszka et al. (2008) and Patilea et al. (2016). Section 4.5 concludes and technical details are included in Appendix.

We introduce some notation. Let $i = \sqrt{-1}$ be the imaginary unit and $\mathcal{L}_2(\mathcal{I})$ be the separable Hilbert space consisting of all the square intergrable curves defined on $\mathcal{I} = [0,1]$ with the inner product,

$$< f, g >= \int_{\mathcal{I}} f(u)g(u)du, \ f, g \in \mathcal{L}_2(\mathcal{I}).$$

Also the vector product of vectors x and y is denoted by $< x, y >= x^T y$. For a complex-valued function $f(\cdot)$, the complex conjugate of f is denoted by $\overline{f}$ and $|f|^2 =< f, \overline{f} >$. Denote the Euclidean norm of $x = (x_1, \cdots, x_p) \in \mathbf{C}^p$ as $|x|$, where $|x|^2 =< x, \overline{x} >= x_1\overline{x}_1 + \cdots + x_p\overline{x}_p$, and if $x \in R^p(\mathcal{L}_2(\mathcal{I}))$, it is denoted as $|x|$, where $|x|^2 =< x, x >$.

4.2 Functional Martingale Difference Divergence

To introduce the new metric FMDD for functional data, we shall provide a brief review of the MDD. For U $\in R^q$ and V$\in R^p$, where q and p are fixed positive integers, Shao and Zhang (2014), Park, Shao and Yao (2015) proposed the so-called martingale difference divergence (MDD) to measure the conditional mean (in)dependence of V on U, i.e.,

$$E(V|U) = E(V), \text{ almost surely.} \tag{4.2.1}$$

Specifically $MDD(V|U)$ is defined as the nonnegative number that satisfies

$$MDD(V|U)^2 \ = \ \frac{1}{c_q}\int_{R^q} \frac{|g_{V,U}(s) - g_V g_U(s)|^2}{|s|^{1+q}} ds, \tag{4.2.2}$$

where $g_{V,U}(s) = E(Ve^{i<s,U>})$, $g_V = E(V)$, $g_U(s) = E(e^{i<s,U>})$, and $c_q = \pi^{(1+q)/2}/\Gamma((1+q)/2)$. A key property of MDD is that $MDD(V|U)^2 = 0$ if and only if (4.2.1) holds, thus MDD completely characterizes the conditional mean independence of V on U. Furthermore, if $E(|V|^2 + |U|^2) < \infty$, then

$$MDD(V|U)^2 = -E[(V - E(V))^T (V^{'} - E(V^{'}))|U - U^{'}|], \tag{4.2.3}$$

where $(V^{'}, U^{'})$ is an independent copy of (V, U).

Considering the definition of MDD in (4.2.3), we naturally define an analogue of MDD that is well defined for functional response Y or functional covariate X by replacing the vector product with the inner product associated with the separable Hilbert space, e.g., $\mathcal{L}_2(\mathcal{I})$. Note that Y and X are in metric spaces $(\mathcal{L}_y, |\cdot|_y)$ and $(\mathcal{L}_x, |\cdot|_x)$, respectively, i.e., $Y \in \mathcal{L}_y,\ X \in \mathcal{L}_x$. Throughout the paper, $(\mathcal{L}_y, \mathcal{L}_x)$ can be $(\mathcal{L}_2(\mathcal{I}), \mathcal{L}_2(\mathcal{I}))$ or (R^p, R^q) or $(\mathcal{L}_2(\mathcal{I}), R^q)$ or $(R^p, \mathcal{L}_2(\mathcal{I}))$. For the convenience of presentation, we do not distinguish between $|\cdot|_y$ and $|\cdot|_x$ but use $|\cdot|$ for both cases.

Definition **4.2.1**. *Functional Martingale Difference Divergence*
For $Y \in \mathcal{L}_y$ and $X \in \mathcal{L}_x$, we define

$$FMDD(Y|X) = -E[< Y - \mu_Y, Y^{'} - \mu_Y > |X - X^{'}|],$$

where μ_Y is the mean function of Y and $(X^{'}, Y^{'})$ is an iid copy of (X, Y).

To show that FMDD fully characterizes the conditional mean independence, we provide the following proposition, which is shown by using several results in Lyons (2013).

Proposition **4.2.1**. *For $Y \in \mathcal{L}_y$, $X \in \mathcal{L}_x$ with $E[|X| + |Y|] < \infty$ and $E[|X - \mu_X||Y - \mu_Y|] < \infty$, we have*

1. *$FMDD(Y|X) \geq 0$.*

2. $FMDD(Y|X) = 0$ if and only if H_0 is true.

Inspired by unbiased estimation of MDD in Park, Shao, and Yao (2015), we construct an unbiased estimator of FMDD by adopting the $\mathcal{U}$-centering approach [Székely and Rizzo (2014), Park, Shao, and Yao (2015), and Zhang, Yao, and Shao (2017)].

Definition **4.2.2**. *Given the iid observations $(X_i, Y_i)_{i=1}^n$ from the joint distribution of (X, Y) where X and Y can be either function-valued or vector-valued, an unbaised estimator of $FMDD(Y|X)$ is defined as*

$$FMDD_n(Y|X) = \frac{1}{n(n-3)} \sum_{i \neq j} \widetilde{A}_{ij} \widetilde{B}_{ij}.$$

Here, $\widetilde{A}_{ij}, \widetilde{B}_{ij}$ are the $\mathcal{U}$-centered (i, j)th element of the matrices defined as

$$\widetilde{A}_{ij} = \begin{cases} a_{ij} - a_{i\cdot} - a_{\cdot j} + a_{\cdot\cdot} & i \neq j \\ 0 & i = j \end{cases}, \quad \widetilde{B}_{ij} = \begin{cases} b_{ij} - b_{i\cdot} - b_{\cdot j} + b_{\cdot\cdot} & i \neq j \\ 0 & i = j, \end{cases}$$

where $a_{ij} = |X_i - X_j|$,

$$a_{i\cdot} = \frac{1}{n-2} \sum_{l=1}^n a_{il}, \;\; a_{\cdot j} = \frac{1}{n-2} \sum_{k=1}^n a_{kj}, \;\; a_{\cdot\cdot} = \frac{1}{(n-1)(n-2)} \sum_{k,l=1}^n a_{kl}.$$

In addition, $b_{ij} = \frac{1}{2}|Y_i - Y_j|^2$ and $b_{i\cdot}$, $b_{\cdot j}$, $b_{\cdot\cdot}$ are defined similarly as $a_{i\cdot}$, $a_{\cdot j}$, $a_{\cdot\cdot}$.

Using the same argument shown in Appendix A.1 of Székely and Rizzo (2014) and (3.4) in Park, Shao, and Yao (2015), it is not difficult to show that $FMDD_n(Y|X)$ is an unbiased estimator of $FMDD(Y|X)$ and it has the expression below.

$$FMDD_n(Y|X) = \frac{1}{\binom{n}{4}} \sum_{i<j<q<r} h(Z_i, Z_j, Z_q, Z_r),$$

where

$$h(Z_i, Z_j, Z_q, Z_r) = \frac{1}{4!} \sum_{(s,t,u,v)}^{(i,j,q,r)} (a_{st}b_{uv} + a_{st}b_{st} - a_{st}b_{su} - a_{st}b_{tv}), \tag{4.2.4}$$

with $Z_i = (X_i, Y_i)$, $\sum_{(s,t,u,v)}^{(i,j,q,r)}$ is the summation over all permutations of the 4-tuple of indices (i, j, q, r). For example, if $(i, j, q, r) = (1, 2, 3, 4)$, then there exist 24 permutations including $(1, 2, 3, 4), \cdots, (4, 3, 2, 1)$. Then (s, t, u, v) can be any permutation of $(1, 2, 3, 4)$ and $\sum_{(s,t,u,v)}^{(1,2,3,4)}$ is the sum of all possible permutations of $(1, 2, 3, 4)$.

In the following, we state the consistency and weak convergence of $FMDD_n(Y|X)$ as an estimator of $FMDD(Y|X)$, which are analogous to Theorems 3 and 4 in Shao and Zhang (2014).

Proposition **4.2.2**. *Under $E[|X| + |Y|] < \infty,\ E[|X - \mu_X||Y - \mu_Y|] < \infty$, we have*

$$FMDD_n(Y|X) \to^{a.s.} FMDD(Y|X).$$

Theorem **4.2.1**. *Assume that $E[|X|^2+|Y|^2] < \infty,\ E[|X-\mu_X|^2|Y-\mu_Y|^2] < \infty$. Under the null H_0, we have*

$$nFMDD_n(Y|X) \to^{D} \sum_{k=1}^{\infty} \lambda_k (G_k^2 - 1),$$

where (G_k) is a sequence of zero mean, unit variance Gaussian random variables which are mutually independent and (λ_k) is a sequence of eigenvalues corresponding to eigenfunctions $(\psi_k(\cdot))$ such that

$$J(z, z') = \sum_{k=1}^{\infty} \lambda_k \psi_k(z) \psi_k(z')$$

where $z = (x, y)$, $J(z, z^{'}) = U(x, x')V(y, y')$, $U(x, x') = |x - x'| + E[|X - X'|] - E[|x - X^{'}|] - E[|X - x'|]$, $V(y, y') = - < y - \mu_Y, y' - \mu_Y >$, *and* (ψ_k) *is an orthonormal sequence i.e.,*

$$E[\psi_j(Z)\psi_k(Z)] = \begin{cases} 1, & \textit{if } j = k \\ 0, & \textit{if } j \neq k \end{cases}$$

Recall that our goal is to test $H_0 : E[Y|X] = E[Y]$ $a.s.$ which is equivalent to $FMDD(Y|X) = 0$. According to Theorem 4.2.1, it is appropriate for us to define our test statistic as

$$T_n = nFMDD_n(Y|X).$$

To understand the behavior of T_n when the null does not hold, we shall study the limiting distribution of T_n under (1) local alternative $H_{1,n} : Y = \mu_Y + \frac{g(X)}{n^a} + \epsilon,\ a > 0$, where $g : \mathcal{L}_x \to \mathcal{L}_y$ satisfies $E[g(X)] = 0$, $FMDD(g(X)|X) > 0$ and $\epsilon \in \mathcal{L}_y$ is nondegenerate and satisfies $E[\epsilon|X] = 0$ $a.s.$, $P(< g(X), \epsilon > \neq 0) > 0$. (2) fixed alternative $H_1 : FMDD(Y|X) > 0$.

Theorem **4.2.2**. *Assume that* $E[|X|^2 + |g(X)|^2 + |\epsilon|^2] < \infty$, $E[|X - \mu_X|^2(|g(X)|^2 + |\epsilon|^2)] < \infty$. *Under the local alternative* $H_{1,n}$, *and*

(i) if $0 < a < 1/2$,

$$T_n \to^p \infty.$$

(ii) if $a = 1/2$,

$$T_n \to^D c + G + \sum_{k=1}^{\infty} \lambda_k (G_k^2 - 1).$$

Here $c = FMDD(g(X)|X) > 0$ *and* G *is a normal random variable with zero mean and variance equal to* $4var(K_1(\mathcal{Z}))$ *which is possibly correlated with* (G_k), *where* $\mathcal{Z} = (X, \epsilon)$ *and* $K_1(z_1) = E[U(x_1, X)V(\epsilon_1, g(X))]$.

(iii) if $a > 1/2$,

$$T_n \to^D \sum_{k=1}^{\infty} \lambda_k (G_k^2 - 1).$$

THEOREM **4.2.3**. *Assume that* $E[|X|^2 + |Y|^2] < \infty$, $E[|X - \mu_X|^2 |Y - \mu_Y|^2] < \infty$. *Under the alternative* H_1, *we have*

$$\sqrt{n}(FMDD_n(Y|X) - FMDD(Y|X)) \to^D N(0, 4\sigma_1^2),$$

where $\sigma_1^2 = var(K(Z))$, $Z = (X, Y)$, *and* $K(z) = E[U(x, X)V(y, Y)]$.

Note that the limiting null distribution of our test statistic is nonpivotal in Theorem 4.2.1. Hence we use the wild bootstrap method to approximate the limiting null distribution of the test statistic and details are given in the next section.

4.3 Bootstrap-based Test

Since the limiting null distribution of our test statistic T_n is nonpivotal, we propose a wild bootstrap procedure to approximate the null distribution and show its asymptotic validity. Note that $FMDD_n(Y|X) = \frac{1}{n(n-3)} \sum_{i \neq j} \widetilde{A}_{ij} \widetilde{B}_{ij}$ is a U-statistic [see (4.2.4) in Section 4.2] and its mean is zero under the null hypothesis. Therefore, we follow the approach of Dehling and Mikorsch (1994) who proposed weighted bootstrap for U-statistics with external random variables $(\eta_j)_{j=1}^n$. Below is the wild bootstrap procedure.

1. Generate the bootstrap statistic.

$$FMDD_n^*(Y|X)^b = \frac{1}{n(n-3)} \sum_{i \neq j} \eta_i \widetilde{A}_{ij} \widetilde{B}_{ij} \eta_j \tag{4.3.1}$$

   where $\eta_i, i = 1, \cdots, n$ are iid with zero mean and unit variance, e.g., standard normal random variables.

2. Repeat 1 for B times and collect $(T_{n,b}^*)_{b=1}^B$, where $T_{n,b}^* = nFMDD_n^*(Y|X)^b$.

3. Obtain the $(1-\alpha)$th quantile of $(T_{n,b}^*)_{b=1}^B$, $Q_{(1-\alpha),n}^*$ and set it as the critcal value for the test with significance level α.

4. Reject the null hypothesis if T_n is greater than the critical value $Q_{(1-\alpha),n}^*$ and accept H_0 otherwise.

Remark **4.3.1**. Patilea et al. (2016) also proposed a wild bootstrap procedure to improve the finite sample performance. It is worth pointing out the difference between the two wild bootstrap procedures. In particular, Patilea et al. (2016) perturbed the response Y directly, i.e., $Y_i^* := \eta_i Y_i, \ \forall i$. In other words, they computed their bootstrap test statistic based on a new bootstrap sample $(X_i, Y_i^*)_{i=1}^n$ and they need to compute their bootstrapped test statistic starting from the very first step which includes dimension reduction procedure through FPCA and finding the least favorable direction toward the null hypothesis, so their test can be computationally costly to implement. By contrast, for our wild bootstrap procedure, $(\widetilde{A}_{ij}, \widetilde{B}_{ij})$ only needs to be computed once and our test is simpler and faster to implement than that in Patilea et al. (2016).

In order to examine the asymptotic behavior of bootstrap test statistic, we first introduce notations of the bootstrap order [see Remark 1 in Chang and Park (2003)] and bootstrap consistency [see Definition 2 in Li, Hsiao, and Zinn (2003)].

Definition **4.3.1**. *Let T_n^* be a bootstrap statistic that depends on the random sample $\{Z_i\}_{i=1}^n$. We define $T_n^* = o_p^*(1)$ a.s. if*

$$P^*(|T_n^*| > \epsilon) \to 0 \ a.s.,$$

for any $\epsilon > 0$, where P^* is conditional probability given $\{Z_i\}_{i=1}^n$. Moreover, we define $T_n^* = O_p^*(1)$ a.s. if, for every $\epsilon > 0$, there exists a constant $M > 0$ such that for large n,

$$P^*(|T_n^*| > M) < \epsilon.$$

Notice that $O_p^*(1)$ and $o_p^*(1)$ are for bootstrap sample asymptotics which have similar definition with $O_p(1)$ and $o_p(1)$. It is straightforward to extend those to $O_p^*(c_n)$ and $o_p^*(c_n)$ based on the similarity to $O_p(1)$ and $o_p(1)$, where c_n is a nonconstant deterministic sequence.

Definition **4.3.2**. *Let T_n^* be a bootstrap statistic that depends on the random sample $\{Z_i\}_{i=1}^n$. We say that $(T_n^*|Z_1, Z_2, \cdots)$ converges to $(T|Z_1, Z_2, \cdots)$ in distribution almost surely if for any sequence T_n^*, such that $(T_n^*|Z_1, Z_2, \cdots)$ converges to $(T|Z_1, Z_2, \cdots)$ almost every sequence $(Z_1, Z_2, \cdots)$ and the following notation is used to denote convergence in distribution almost surely.*

$$T_n^* \to^{D^*} T \ a.s.$$

We introduce the following theorem that is useful for deriving the asymptotic distribution of bootstrap test statistic T_n^*.

Theorem **4.3.1**. *Suppose $\mathcal{H}$ is a symmetric kernel satisfying $E[\mathcal{H}(Z, Z')^4] < \infty$ and $U_n = \frac{1}{n(n-1)} \sum_{i \neq j} \mathcal{H}(Z_i, Z_j)$. Further assume that $\{W_i\}$ is an iid sequence of random* *variables with* $E[W_1] = 0$, $E[W_1^2] = 1$, $E[W_1^4] < \infty$. *Then the bootstrap statistic* $nU_n^* = \frac{1}{n-1}\sum_{i\neq j}\mathcal{H}(Z_i, Z_j)W_iW_j$, *has the following asymptotic distribution.*

$$nU_n^* \to^{D^*} \sum_{k=1}^{\infty}\nu_k(N_k^2 - 1) \ a.s.,$$

where (N_k) *is a sequence of zero mean, unit variance Gaussian random variables which are mutually independent.*

Note that the result of Theorem 4.3.1 can be viewed as an extension of Theorem 3.1 in Dehling and Mikorsch (1994) to functional data although our theoretical argument is considerably different from that in Dehling and Mikorsch (1994). Based on Theorem 4.3.1, we are ready to examine the asymptotic distribution of our bootstrap statistic T_n^* under the null, local and fixed alternatives.

Theorem **4.3.2**. *Assume that* $E[|X|^4+|Y|^8] < \infty$, $E[|X-\mu_X|^4|Y-\mu_Y|^4] < \infty$, $E[\eta^4] < \infty$. *Under the null* H_0, *we have*

$$T_n^* \to^{D^*} \sum_{k=1}^{\infty}\lambda_k(G_k^2 - 1) \ a.s.,$$

where (λ_k, G_k) *are defined in Theorem 4.2.1.*

Theorem **4.3.3**. *Assume that* $E[|X|^4 + |g(X)|^8 + |\epsilon|^8] < \infty$, $E[|X - \mu_X|^4|\epsilon|^4] < \infty$, $E[\eta^4] < \infty$. *Under the local alternative* $H_{1,n}$, *and*

(i) if $0 < a < 1/2$,

$$P(T_n \geq Q^*_{(1-\alpha),n}|H_{1,n}) \to 1,$$

where $Q^*_{(1-\alpha),n}$ *is the* $(1-\alpha)$*th quantile of the bootstrap test statistic.*

(ii) if $a = 1/2$,

$$P(T_n \geq Q^*_{(1-\alpha),n}|H_{1,n}) \to P(\mathcal{G}_1 \geq Q_{(1-\alpha),\mathcal{G}_0} - c),$$

where $\mathcal{G}_1 = G + \sum_{k=1}^{\infty} \lambda_k (G_k^2 - 1)$ follows the asymptotic distribution of $T_n - c$ under $H_{1,n}$ when $a = 1/2$, $Q_{(1-\alpha),\mathcal{G}_0}$ is the $(1-\alpha)$th quantile of the limiting null distribution.

(iii) if $a > 1/2$,

$$P(T_n \geq Q^*_{(1-\alpha),n} | H_{1,n}) \to \alpha.$$

Under the fixed alternative H_1 with the same assumptions in Theorem 4.3.2, we have

$$P(T_n \geq Q^*_{(1-\alpha),n} | H_1) \to 1.$$

Remark **4.3.2**. Patilea et al. (2016) considered the following local alternatives $H_{1,n}$: $E[Y] = \mu_Y + r_n \delta(X)$, where r_n satisfies certain constraints which implies $r_n n^{1/2} \to \infty$ and showed the consistency in Theorem 3.8 of their paper. By comparison, we show that our test has nontrivial power under the local alternative that approaches the null hypothesis at the rate of $1/\sqrt{n} < r_n$ in Theorem 4.3.3, where Patilea et al.'s (2016) smoothing-based test is unable to detect. Therefore, we can conclude that our test is more powerful than the one in Patilea et al. (2016) in terms of capability of detecting the local alternative that approaches the null at a faster rate.

4.4 Numerical Simulations

In this section, we study the finite sample performance of our FMDD-based conditional mean independence test. For convenience, we denote our test, Patilea et al.'s (2016) test, and Kokoszka et al.'s (2008) test in the tables as FMDD, PSS, and KMSZ, respectively. In particular, Example 4.4.1 considers functional response Y and univariate covariate X and compares with the test in Patilea et al. (2016). For other examples with functional response Y and functional covariates X, we compare our FMDD-based test with both PSS and KMSZ which use FPCA when constructing their test statistics. In our simulations, we tried several different values of nominal level α, 10%, 5%, 1% to assess the sensitivity of our test with respect to the choice of nominal levels. For each example, bootstrap sample size is equal to 499 and $\{\eta_i\}_{i=1}^n$ are from the following distribution [see Mammen (1993)] which is same as the one used in Section 4 in Patilea et al. (2016).

$$\eta_i = \begin{cases} \frac{-(\sqrt{5}-1)}{2} & w.p.\ \frac{\sqrt{5}+1}{2\sqrt{5}} \\ \frac{(\sqrt{5}+1)}{2} & w.p.\ 1-\frac{\sqrt{5}+1}{2\sqrt{5}} \end{cases}$$

In order to compute the size and power of tests, 5000 replicates are generated for every example.

4.4.1 Simulations

Example **4.4.1**.

Example 4.4.1 is adopted from Patilea et al. (2016) where the data $(X_i, Y_i)_{i=1}^n$ is generated by

$$\begin{aligned} Y_i(t) &= \mu(t) + \epsilon_i(t), \quad 1 \le i \le n, \\ \mu(t) &= 0.01e^{-4(t-0.3)^2}, \quad t \in [0,1], \end{aligned}$$

where $\epsilon_i(t)$ are independent Brownian Bridges and is independent of X_i, and X_i follows log-normal distribution with mean 3 and standard deviation 0.5. Therefore, under this data generating process, X_i is independent of Y_i. In order to evaluate the power of a test, we consider the following data generating process,

$$Y_i(t) \;=\; \mu(t)X_i + \epsilon_i(t), \;\; 1 \leq i \leq n,$$

where ϵ_i and X_i are generated in the same fashion as described above. In this example, we consider $n = 100, \; 200$. Recall that the test proposed by Patilea et al. (2016) involves several user-chosen parameters. Specifically, when function Y and variable X are considered, Patilea et al. (2016) requires one user-chosen parameter, bandwidth h and we let $h = c_h n^{-2/9}, \; c_h = 0.75, \; 1.00, \; 1.25$ following the recommendation in their Section 4.1 in Patilea et al. (2016).

From Table 4.1, the empirical sizes of both tests are reasonably close to the nominal levels. Comparing empirical sizes of PSS tests with different values of the bandwidth parameter h, there is no uniformly best h. In other words, different combinations of (n, α) have different values of h which produce the most accurate size. For the empirical powers, our test outperforms PSS test noticeably, which is consistent with our theory. Overall, when n increases, the empirical power increases for both tests.

Table 4.1: Size and Power of the two tests for Example 4.4.1

		$\alpha = 10\%$		$\alpha = 5\%$		$\alpha = 1\%$	
Size		$n = 100$	$n = 200$	$n = 100$	$n = 200$	$n = 100$	$n = 200$
	FMDD	0.104	0.0992	0.0562	0.0528	0.0156	0.0114
$c_h = 0.75$	PSS	0.11	0.1104	0.0586	0.0592	0.0118	0.0144
$c_h = 1.00$	PSS	0.1112	0.1112	0.0554	0.0578	0.0114	0.0146
$c_h = 1.25$	PSS	0.1064	0.1122	0.0524	0.0576	0.013	0.0182
Power		$n = 100$	$n = 200$	$n = 100$	$n = 200$	$n = 100$	$n = 200$
	FMDD	0.8308	0.9878	0.7344	0.9668	0.4782	0.8762
$c_h = 0.75$	PSS	0.4	0.793	0.2834	0.6984	0.1302	0.4612
$c_h = 1.00$	PSS	0.3802	0.7868	0.2634	0.6818	0.1088	0.4358
$c_h = 1.25$	PSS	0.3492	0.7732	0.2268	0.6614	0.088	0.406

Example **4.4.2**.

This example is also from Patilea et al. (2016) where both Y and X are functional data. The data is generated by the following functional linear model,

$$Y_i(t) = \int_0^1 \xi(s,t)X_i(s)ds + \epsilon_i(t), t \in [0,1],$$

where $X_i(t),\ \epsilon_i(t)$ are independent Brownian Bridges and $\xi(s,t) = c\cdot\exp(t^2/2 + s^2/2)$, $c = 0,\ 0.75$ and we let $n = 40,\ 100$. Note that PSS and KMSZ tests require several user-chosen parameters. For PSS test, the bandwidth parameter $h = n^{-2/9}$, the penalty value $\alpha_n = 2$, the initial guess for the direction $\gamma_0^{(q)} = (1, 1, \cdots, 1)/\sqrt{q} \in R^q$, q is chosen as the minimum integer that explains 95% of the variance of X, and we use the sequential algorithm described in Section 3.5 in their paper with a grid size equal to 50 and these settings are the same as those used in their simulation study. For KMSZ tests, p and q are chosen by the minimum values which explain at least 95% of variances of Y and X, respectively.

According to Table 4.2, our FMDD-based test is superior to the other two tests with respect to the empirical size and power. In particular, size performances of all three tests are comparable with FMDD-based test and our test slightly outperforms the other two tests. For all αs, KMSZ test shows slight conservative size compared to the other two tests. Under the alternatives, all three tests produce fairly high empirical powers for all cases where our test always has the highest power, especially for $n = 40$. Notice that Y follows the functional linear model for this example and therefore KMSZ test is expected to perform well since KMSZ test is tailored for the functional linear model. It is interesting that FMDD-based test performs better than KMSZ test indicating that projecting the functional data to a finite dimensional space could lead to some loss of power, especially when the sample size is small.

Table 4.2: Size and Power of the three tests for Example 4.4.2

	$\alpha = 10\%$		$\alpha = 5\%$		$\alpha = 1\%$	
Size	$n = 40$	$n = 100$	$n = 40$	$n = 100$	$n = 40$	$n = 100$
FMDD	0.1068	0.1056	0.0602	0.0556	0.015	0.011
PSS	0.133	0.1156	0.0698	0.0632	0.0146	0.0164
KMSZ	0.0898	0.0902	0.0378	0.0394	0.004	0.0054
Power	$n = 40$	$n = 100$	$n = 40$	$n = 100$	$n = 40$	$n = 100$
FMDD	1	1	0.9998	1	0.9956	1
PSS	0.898	1	0.8274	1	0.6246	0.9996
KMSZ	0.9892	1	0.9572	1	0.6986	1

Example **4.4.3**.

In this example, we generate the functional response Y by quadratic form of the covariate X which is also considered in Patilea et al. (2016).

$$Y_i(t) = c \cdot (X_i(t)^2 - 1) + \epsilon_i(t), t \in [0, 1],$$

where $X_i(t)$ and $\epsilon_i(t)$ are independent Brownian Motion and Brownian Bridge and $c = 0,\ 0.5$. Furthermore, other settings including user-chosen parameters for the existing two tests are the same as Example 4.4.2.

Table 4.3 reports the empirical sizes and powers for three tests. By comparison, our FMDD-based test appears to outperform the other two tests for most of the cases in terms of more accurate empirical size and higher empirical power. Moreover, KMSZ test appears inferior to PSS and FMDD-based counterparts with respect to the size and power, presumably due to its inability of capturing nonlinear dependence between Y and X. Under the null hypothesis, it seems that FMDD-based test produces more accuarate sizes than the other two tests, especially when $n = 40$. Except for $n = 100,\ \alpha = 1\%$, FMDD-based test is the most powerful one among the three. When $n = 40$, our test has noticeably higher power than the other two tests. Note that FMDD-based and PSS tests aim to detect not only linear but also nonlinear depedence between Y and X and this example has strong nonlinear dependence. Limited simulation evidence seems to suggest that our FMDD-based test is more powerful than PSS test against the alternative hypothesis where there exist strong nonlinear depedence between functional data. This could be due to the fact that PSS test uses FPCA when constructing their test statistic while our test statistic is constructed by preseving the functional form of the data. Hence, it seems that some loss of power might occur due to the use of a dimension reduction device when computing a test statistic.

Table 4.3: Size and Power of the three tests for Example 4.4.3

	$\alpha = 10\%$		$\alpha = 5\%$		$\alpha = 1\%$	
Size	$n = 40$	$n = 100$	$n = 40$	$n = 100$	$n = 40$	$n = 100$
FMDD	0.1078	0.104	0.0562	0.0552	0.0118	0.0116
PSS	0.113	0.1076	0.0614	0.054	0.0146	0.014
KMSZ	0.0918	0.0934	0.038	0.0422	0.0048	0.0062
Power	$n = 40$	$n = 100$	$n = 40$	$n = 100$	$n = 40$	$n = 100$
FMDD	0.6954	0.9988	0.3904	0.9904	0.07	0.691
PSS	0.288	0.9868	0.1618	0.9728	0.0394	0.9148
KMSZ	0.333	0.3862	0.2066	0.2674	0.0534	0.1052

4.5 Discussion and Conclusions

In this paper, we propose a novel metric, namely the functional martingale difference divergence, to measure the conditional mean dependence of Y given X, where Y and X can be elements in a separable Hilbert space, e.g., $\mathcal{L}_2(\mathcal{I})$. The FMDD is a natural extension of the MDD proposed by Shao and Zhang (2014), and is shown to fully characterize the conditional mean independence. We further propose to use the $\mathcal{U}$-centering based sample estimate of FMDD as our test statistic (up to a normalizing constant) and study its limiting behavior under both the null and alternative hypothesis. Since the limiting null distribution of our test statistic is not pivotal, we use a wild bootstrap method to approximate the limiting null distribution and show its consistency under the null. The limiting distributions of the bootstrap test statistic are further derived under the local and fixed alternatives, which show that our test have nontrivial power to detect the local alternatives that lie within $1/\sqrt{n}$-neighborhood of the null. Compared to the two existing tests developed by Kokoszka et al. (2008) and Patilea et al. (2016), our test does not require linear model assumption and a choice of user-chosen numbers, and is thus model free and tuning parameter free. Additionally, our test does not involve dimension reduction using functional PCA, and treats function-valued and vector-valued responses and covariates in a unifed fashion. Through numerical simulations, we show that our test exhibits fairly accurate size in small sample and the power is noticeably higher than the two above-mentioned tests in most cases, consistent with our theoretical result on approximate power. From the computational and practical viewpoint, our test is much more convenient to implement and is less costly in computation, compared to the other nonparametric test by Patilea et al. (2016).

To conclude, we mention two related future research topics. On one hand, diagnostic checking for functional linear model is worth investigating given the prevalence of functional linear model in practical applications. Given Y and X that are both function-valued, we want to test

$$H_0 : E(Y|X) = \Phi X,$$

where Φ is a square integrable operator. A natural extension seems to consist of the following three steps: (1), estimate Φ by $\widehat{\Phi}_n$, which typically involves regularization [see Ramsay and Silverman (2005)]; (2) obtain the residuals $\widehat{\epsilon}_j = Y_j - \widehat{\Phi}_n X_j$ for $j = 1, \cdots, n$; (3) Apply the FMDD-based test to $(X_j, \widehat{\epsilon}_j)_{j=1}^n$, as under the null, we have $E(\epsilon|X) = 0$, where $\epsilon = Y - E(Y|X)$ is the population counterpart of $\hat{\epsilon}$. One complication is that the estimation effect from replacing ϵ by $\hat{\epsilon}$ may show up in the limiting null distribution, and it is unclear whether the wild bootstrap is capable of capturing that effect. A careful theoretical investigation is needed. On the other hand, it would be interesting to extend the idea to test for the conditional quantile independence owing to a natural connection between conditional quantile independence and conditional mean independence when the response Y is a scalar-valued variable; see Shao and Zhang (2014). Also see Kato (2012) for estimation in functional linear quantile regression when the response Y is a scalar random variable. When Y is function-valued, Chowdhury and Chaudhuri (2016) recently advanced nonparametric quantile regression to functional data based on spatial depth and quantiles. It would be intriguing to see how FMDD can play a role in the model checking and testing for nonparametric quantile regression models.

References

Alexander, C. O. (2001) *Market Models*, New York: Wiley.

Back, A. D. and Weigend, A. S. (1997) First application of independent component analysis to extracting structure from stock returns. *International Journal of Neural Systems*, **8**, 473-484.

Bai, J. (2003) Inferential theory for factor models of large dimensions. *Econometrica*, **71**, 135-171.

Bai, J. and Ng, S. (2002) Determining the number of factors in approximate factor models. *Econometrica*, **70**, 191-221.

Bai, J. and Ng, S. (2007) Determining the number of primitive shocks in factor models. *Journal of Business and Economic Statistics*, **25**, 52-60.

Bathia, N., Yao, Q. and Ziegelmann, F. (2010) Identifying the finite dimensionality of curve time series. *The Annals of Statistics*, **38**, 3352-3386.

Bierens, H. (1982). Consistent model specification test. *Journal of Econometrics*, **20**, 105-134.

Bierens, H. (1990). A consistenct conditional moment test of functional form. *Econometrica*, **58**, 1443-1458.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, **31**, 307-327.

Bollerslev, T. (1990). Modelling the coherence in short-run nominal exchange rates: a multivariate generalized arch model. *Review of Economics and Statistics*, **72**, 498-505.

Box, G. E. P. and Tiao, G. C. (1977) A canonical analysis of multiple time series. *Biometrika*, **64**, 335-365.

Brillinger, D. R. (1981) *Time Series, Data Analysis and Theory (expanded ed.)*, San Francisco: Holden-Day.

Cai, T., and Hall, P. (2006) Prediction in functional linear regression. *The Annals of Statistics*, **34**, 2159-2179.

Cardot, H., Ferraty, F., Mas, A., and Sarda, P. (2003) Testing hypothesis in the functional linear model. *Scandinavian Journal of Statistics*, **30**, 241-255.

Cattell, R.B. (1966) The scree test for the number of factors. *Multivariate Behavioral Research*, **1**, 245-276.

Chang, Y. and Park, J. (2003) A sieve bootstrap for the test of a unit root. *Journal of Time Series Analysis*, **24**, 379-400.

Chen, Y., Härdle, W. and Spokoniy, V. (2007) Portfolio value at risk based on independent component analysis. *Journal of Computational and Applied Mathematics*, **205**, 594-607.

Chiou, J., and Müller, H. (2007) Diagnostics for functional regression via residual processes. *Computational Statistics and Data Analysis.* **15**, 4849-4863.

Chiou, J., Müller, H., and Wang, J. (2004) Functional response models. *Statistica Sinica*, **14**, 675-693.

Chowdhury, J. and Chaudhuri, P. (2016) Nonparametric depth and quantile regression for functional data. *Preprint*, at https://arxiv.org/pdf/1607.03752v1.pdf.

Cook, R. D., Li, B., and Chiaromonte, F. (2010) Envelope models for parsimonious and efficient multivariate linear regression. *Statistica Sinica*, 20, 927-1010.

Cuevas, A., Febrero, M., and Fraiman, R. (2002) Linear functional regression: the case of fixed design and functional response. *The Canadian Journal of Statistics*, **30**, 285-300.

Dehling, H. and Mikorsch T. (1994) Random quadratic forms and the bootstrap for U-statistics. *Journal of Multivatiate Analysis*, **51**, 392-413.

Dehling, H., Durieu, O., and Volny, D. (2009) New techniques for empirical processes of dependent data. *Stochastic Processes and their Applications*, **119**, 3699-3718.

Dunford, N. and Schwartz, J. T. (1963) *Linear Operators, Spectral Theory, Self Adjoint Operators in Hilbert Space, Part 2.* Wiley-Interscience.

Eichler, M., Motta, G. and von Sachs, R. (2011) Fitting dynamic factor models to non-stationary time series. *Journal of Econometrics*, **163**, 51-70.

Engle, R. (1982). Autoregressive conditional heteroskedasticity with estimates of the variance of united kingdom inflation. *Econometrica*, **50**, 987-1007.

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

Engle, R. and Kroner, K. (1995). Multivariate simultaneous generalized arch. *Econometric Theory*, **11**, 122-150.

Engle, R., Ng, V. and Rothschild, M. (1990). Asset pricing with a factor-arch covariance structure, empirical estimates for treasury bills. *Journal of Econometrics*, **45**, 213-237.

Engle, R. and Sheppard, K. (2001). Theoretical and empirical properties of dynamic conditional correlation multivariate garch. UCSD Discussion Paper 2001-15.

Escanciano, J. (2006). Goodness-of-fit tests for linear and nonlinear time series models. *Journal of the American Statistical Association*, **101**, 531-541.

Fan, J., Wang, M. and Yao, Q. (2008). Modelling multivariate volatilities via conditionally uncorrelated components. *Journal of the Royal Statistical Society, Series B*, **70**, 676-702.

Ferraty, F., Laksaci, A., Tadj, A., and Vieu, P. (2011) Kernel regression with functional response. *Electronic Journal of Statistics*, **5**, 159-171.

Ferraty, F., Van Keilegom, I., and Vieu, P. (2012) Regression when both response and predictor are functions. *Journal of Multivariate Analysis*, **109**, 10-28.

Ferraty, F. and Vieu, P. (2010) *Nonparametric Functional Data Analysis: Theory and Practice.* Springer, New York.

Forni, M., Hallin, M. and Reichlin, L. (2000) The generalized dynamic-factor model: identification and estimation. *Review of Economics and Statistics*, **82**, 540-554.

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

Gabrys, R., Horváth, L., and Kokoszka, P. (2010) Tests for error correlation in the functional linear model. *Journal of the American Statistical Association*, **105**, 1113-1125.

Geweke, J. (1977) The dynamic factor analysis of economic time series models. *Latent Variables in Socio-Economic Models*, eds. D. J. Aigner and A. S. Goldberger, Amsterdam: North-Holland, pp, 365-383.

Granger, C. W., Maasoumi, E. and Racine, J. (2004) A dependence metric for possibly nonlinear processes. *Journal of Time Series Analysis*, **25**, 649-669.

Greene, W. (2007) *Econometric Analysis.* Pearson Education India.

Hallin, M. and Liška, R. (2007) Determining the number of factors in the general dynamic factor model. *Journal of the American Statistical Association*, **102**, 603-617.

Hoeffding, W. (1961) The strong law of large numbers for U-statistics. Inst. Statist.

Univ. of North Carolina, Mimeo Report, No. 302.

Hong, Y. M. (1999) Hypothesis testing in time series via the empirical characteristic function: a generalized spectral density approach. *Journal of the American Statistical Association*, **94**, 1201-1220.

Horváth, L., and Kokoszka, P. (2012) *Inference for Functional Data with Applications.* Springer, New York.

Hu, Y. P. and Tsay, R. S. (2014) Principal volatility component analysis. *Journal of Business and Economics Statistics*, **32**, 153-164.

Kato, K. (2012) Estimation in functional linear quantile regression. *Annals of Statistics*, **40**, 3108-3136.

Kokoszka, P., Maslova, I., Sojka, J., and Zhu, L. (2008) Testing for lack of dependence in the functional linear model. *The Canadian Journal of Statistics*, **36**, 1-16.

Kokoszka, P. and Reimherr, M. (2017) *Introduction to Functional Data Analysis.* Chapman and Hall/CRC.

Koul, H. and Stute, W. (1999). Nonparametric model checks for time series. *Annals of Statistics*, **27**, 204-236.

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

Lam, C., Yao, Q. and Bathia, N. (2011) Estimation of latent factors for high-dimensional time series. *Biometrika*, **98**, 901-918.

Lee, A. J. (1990) *U-Statistics: Theory and Practice.* Marcel Dekker, Inc.

Lee, C. and Shao, X. (2016). Martingale difference divergence matrix and its application to dimension reduction for stationary multivariate time series. *Journal of the American Statistical Association*, in press.

Lehmann, E. L. and Romano, J. P. (2005) *Testing Statistical Hypotheses.* 3rd edition, Springer, New York.

Li, Q., Hsiao, C. and Zinn, J. (2003) Consistent specication tests for semiparametric/nonparametric models based on series estimation methods. *Journal of Econometrics*, **112**, 295-325.

Li, W., Gao, J., Li, K. and Yao, Q. (2016). Modelling multivariate volatilites via latent common factors. *Journal of Business & Economiv Statistics*, **34**, 564-573.

Li, Z., Wang, Q. and Yao, J. (2016) Identifying the number of factors from singular values of a large sample auto-covariance matrix. *Annals of Statistics*, forthcoming.

Lian, H. (2011) Convergence of functional k-nearest neighbor regression estimate with functional responses. *Electronic Journal of Statistics*, **5**, 31-40.

Lyons, R. (2013) Distance covariance in metric spaces. *The Annals of Probability*, **41**, 3284-3305.

Mammen, E. (1993) Bootstrap and wild bootstrap for high dimensional linear models. *The Annals of Statistics*, **21**, 255-285.

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

Motta, G., Hafner, C. M. and von Sachs, R. (2011) Locally stationary factor models: identification and nonparametric estimation. *Econometric Theory*, **27**, 1279-1319.

Müller, H., and Stadtmüller, U. (2005) Generalized functional linear models. *The Annals of Statistics*, **33**, 774-805.

Ombao, H., von Sachs, R. and Guo, W. (2005) SLEX analysis of multivariate nonstationary time series. *Journal of the American Statistical Association*, **100**, 519-531.

Pan, J., Polonik, W. and Yao, Q. (2010). Estimating factor models for multivariate volatilities: an innovation expansion method. *Proceedings of COMPSTAT'2010* Physica-Verlag HD, 2010, 305-314.

Pan, J. and Yao, Q. (2008) Modeling multiple time series via common factors. *Biometr -ika*, **95**, 365-379.

Park, J. H., Sriram, T. N., and Yin, X. (2009) Central mean subspace in time series. *Journal of Computational and Graphical Statistics*, 18, 717-730.

Park, J. H., Sriram, T. N., and Yin, X. (2010) Dimension reduction in time series. *Statistica Sinica*, 20, 747-770.

Park, T., Shao, X. and Yao, S. (2015) Partial martingale difference correlation. *Electronic Journal of Statistics*, **9**, 1492-1517.

Patilea, V., Sánchez-Sellero, C. and Saumard, M. (2016) Testing the predictor effect on a functional response. *Journal of the American Statistical Association*, **111**, 1684-1695.

Pelletier, D. (2006). Regime switching for dynamic correlations. *Journal of Econometrics*, **131**, 445-473.

Peña, D. and Box, G. E. P. (1987) Identifying a simplifying structure in time series. *Journal of the American Statistical Association*, **82**, 836-843.

Ramsay, J. O. and Silverman, B. W. (2005) *Functional Data Analysis.* 2nd edition, Springer, New York.

Resnick, S. I. (2005) *A Probability Path*, Springer Science & Business Media.

Serfling, R. J. (1980) *Approximation Theorems of Mathematical Statistics*, John Wiley & Sons.

Shao, X. (2011) Testing for white noise under unknown dependence and its applications to goodness-of-fit for time series models. *Econometric Theory*, **27**, 312-343.

Shao, X. and Zhang, J. (2014) Martingale difference correlation and its use in high dimensional variable screening. *Journal of the American Statistical Association*,

109, 1302-1318.

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

Stock, J. H. and Watson, M. W. (2005) Implications of dynamic factor models for VAR analysis. Working Paper 11467, National Bureau of Economic Research.

Stoffer, D. (1999) Detecting common signals in multiple time series using the spectral envelope. *Journal of the American Statistical Association*, **94**, 1341-1356.

Stute, W. (1997). Nonparametric model checks for regression. *Annals of Statistics*, **25**, 613-641.

Székely, G. J. and Rizzo, M. L. (2014) Partial distance correlation with methods for dissimilarities. *Annals of Statistics*, 42, 2382-2412.

Székely, G. J., Rizzo, M. L. and Bakirov, N. K. (2007) Measuring and testing independence by correlation of distances. *Annals of Statistics*, **35**, 2769-2794.

Tiao, G. and Box, G. (1981). Modeling multiple times series with applications. *Journal of the American Statistical Association*, **76** 802-816.

Tiao, G. C. and Tsay, R. S. (1989) Model specification in multivariate time series. *Journal of the Royal Statistical Society, Series B*, **51**, 157-213.

Tsay, R. S. (2010) *Analysis of Financial Time Series*, 3rd edition. Hoboken, NJ:Wiley.

Weide, R. (2002) Go-garch: a multivariate generalized orthogonal garch model. *Journal of Applied Econometrics*, **17**, 549-564.

Xia, Y., Tong, H., Li, W. K., and Zhu, L. X. (2002) An adaptive estimation of dimension reduction. *Journal of the Royal Statistical Society, Ser. B*, 64, 363-410.

Yao, F., Müller, H., and Wang, J. (2005a) Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association*, **100**, 577-590.

Yao, F., Müller, H., and Wang, J. (2005b) Functional linear regression analysis for longitudinal data. *The Annals of Statistics*, **33**, 2873-2903.

Zhang, X., Yao, S. and Shao, X. (2017) Conditional mean and quantile dependence testing in high dimension. *The Annals of Statistics*, in press.

Zhou, Z. (2012) Measuring nonlinear dependence in time-series, a distance correlation approach. *Journal of Time Series Analysis*, **33**, 438-457.

Zhu, L. (2003). Model checking of dimension-reduction type for regression. *Statistica Sinica*, **13**, 283-296.

Appendix A

Supplementary Materials Including Proofs

Proof of Lemma 2.3.1: For $j, k = 1, \cdots, p$,

$$\begin{aligned} G_j(s)G_k(s)^* &= E[(V_j - E(V_j))e^{i<s,U>}]E[(V_k^{'} - E(V_k^{'}))e^{-i<s,U^{'}>}] \\ &= E[(V_j - E(V_j))(V_k^{'} - E(V_k^{'}))e^{i<s,U-U^{'}>}] \\ &= -E[(V_j - E(V_j))(V_k^{'} - E(V_k^{'}))(1 - cos(s < U - U^{'} >))] + A \end{aligned}$$

with A representing the term that vanishes when the integral is evaluated. Integrating the above term and using Lemma 1 in Székeley et al. (2007), we can derive that

$$MDDM_{jk}(V|U) = -E[(V_j - E(V_j))(V_k^{'} - E(V_k^{'}))|U - U^{'}|_q]$$

Therefore, $MDDM(V|U) = -E[(V - E(V))(V^{'} - E(V^{'}))^T|U - U^{'}|_q]$. ♢

Proof of Theorem 2.3.1: The first assertion is a direct consequence of Lemma 2.3.1. Regarding the second one, let $m = (m_1, \cdots, m_p)^T \in R^p$, $m \neq 0$, and $Z = m^T V$ be a linear combination of V that satisfies $E(Z|U) = E(Z)$, then $MDDM(Z|U) = 0$ and

$m^T MDDM(V|U)m = MDDM(Z|U) = 0$, implying that $MDDM(V|U)$ is singular. On the other hand, assume that $MDDM(V|U)$ is singular and m is in its null space, i.e., $MDDM(V|U)m = 0$. Since $MDDM(V|U)$ is positive semidefinite, we have $m^T MDDM(V|U)m = MDDM(m^T V|U) = 0$, which implies that $E(m^T V|U) = E(m^T V)$, i.e., a linear combination of V is conditionally mean independent of U. The conclusion follows. $\diamondsuit$

Proof of Theorem 2.4.1: We shall treat the case $k_0 = 1$ only as the more general case can be handled in a similar fashion but at the expense of lengthy details. The main idea of the proof is to use Lemma A.1 in Kneip and Utikal (2001), which quantifies the changes of eigenvalues and eigenvectors when passing from a matrix C to a perturbed matrix $C + E$. In our setting, we let $C = \Gamma_1 = MDDM(Y_t|Y_{t-1})$, and $E = \widehat{\Gamma}_1 - \Gamma_1$. Then for $j = 1, \cdots, s$, we get by applying part (a) of that lemma that

$$\widehat{\lambda}_j - \lambda_j \quad = \quad tr(\gamma_j \gamma_j^T \{\widehat{\Gamma}_1 - \Gamma_1\}) + R_1, \tag{A.0.1}$$

where $|R_1| \leq \frac{6\|\widehat{\Gamma}_1 - \Gamma_1\|_2^2}{\min_{\lambda \in EG(\Gamma_1), \lambda \neq \lambda_j} |\lambda - \lambda_j|}$. Here $EG(C) = (\lambda_1(C), \cdots, \lambda_p(C))$ denotes the set of eigenvalues of the $p \times p$ matrix C. To obtain the order of $\widehat{\lambda}_j - \lambda_j$, we shall show that

$$\|\widehat{\Gamma}_1 - \Gamma_1\|_2^2 = O_p(n^{-1}) \tag{A.0.2}$$

Note that

$$\begin{aligned} \|\widehat{\Gamma}_1 - \Gamma_1\|_2^2 \;&\leq\; \|\widehat{\Gamma}_1 - \Gamma_1\|_F^2 \\ &\leq\; \sum_{i=1}^{p} \sum_{j=1}^{p} |MDDM_n(Y_t|Y_{t-1})_{ij} - MDDM(Y_t|Y_{t-1})_{ij}|^2 \end{aligned}$$

Let $Y_t = (Y_{1,t}, Y_{2,t}, \cdots, Y_{p,t})^T$. Write

$$\begin{aligned} MDDM_n(Y_t|Y_{t-1}) &= -\frac{1}{(n-1)^2}\sum_{t_1=2}^{n}\sum_{t_2=2}^{n}(Y_{t_1}-\overline{Y}_{n-1})(Y_{t_2}-\overline{Y}_{n-1})^T|Y_{t_1-1}-Y_{t_2-1}|_p \\ MDDM_n(Y_t|Y_{t-1})_{i,j} &= \frac{(n-2)}{(n-1)}\{(U_{1,n})_{i,j}+(U_{2,n})_{i,j}+(U_{3,n})_{i,j}+(U_{4,n})_{i,j}\}, \end{aligned}$$

where $\overline{Y}_{n-1} = (n-1)^{-1}\sum_{t=2}^{n} Y_t$, and $MDDM_n(Y_t|Y_{t-1})_{i,j}$ is the (i,j)th entry of $MDDM_n(Y_t|Y_{t-1})$. Furthermore,

$$(U_{1,n})_{ij} = -\frac{1}{(n-1)(n-2)}\sum_{t_1=2}^{n}\sum_{t_2\neq t_1}(Y_{i,t_1}-E(Y_{i,t_1}))(Y_{j,t_2}-E(Y_{j,t_2}))|Y_{t_1-1}-Y_{t_2-1}|_p$$

$$(U_{2,n})_{ij} = -\frac{1}{(n-1)(n-2)}\sum_{t_1=2}^{n}\sum_{t_2\neq t_1}(Y_{i,t_1}-E(Y_{i,t_1}))(E(Y_{j,t_2})-(\overline{Y}_{n-1})_j)|Y_{t_1-1}-Y_{t_2-1}|_p$$

$$(U_{3,n})_{ij} = -\frac{1}{(n-1)(n-2)}\sum_{t_1=2}^{n}\sum_{t_2\neq t_1}(E(Y_{i,t_1})-(\overline{Y}_{n-1})_i)(Y_{j,t_2}-E(Y_{j,t_2}))|Y_{t_1-1}-Y_{t_2-1}|_p$$

$$(U_{4,n})_{ij} = -\frac{1}{(n-1)(n-2)}\sum_{t_1=2}^{n}\sum_{t_2\neq t_1}(E(Y_{i,t_1})-(\overline{Y}_{n-1})_i)(E(Y_{j,t_2})-(\overline{Y}_{n-1})_j)|Y_{t_1-1}-Y_{t_2-1}|_p$$

and $(U_{1,n})_{i,j}$ is a U-statistic of order 2 for the stationary time series $\{Z_t = (Y_t^T, Y_{t-1}^T)^T\}$. The kernel function for $(U_{1,n})_{ij}$ is

$$g(Z_2, Z_2') = -\{Y_{i,2}-E(Y_{i,2})\}\{Y_{j,2}'-E(Y_{j,2}')\}|Y_1-Y_1'|_p,$$

where $Z_2 = (Y_2^T, Y_1^T)^T$ and $Z_2' = (Y_2^{'T}, Y_1^{'T})^T$. Under our condition (C2), we have that $E(|g(Z_2,Z_2')|^{2+\delta}) < \infty$ and $E(|g(Z_2,Z_{2+k})|^{2+\delta}) < \infty, k = 1,\cdots,n$ by using Cauchy-Swartz inequality. It then follows from Theorem 1 in Yoshihara (1976) that $|(U_{1,n})_{ij} - MDDM(Y_t|Y_{t-1})_{ij}|^2 = O_p(n^{-1})$ for $i,j = 1,\cdots,p$. Since $|E(Y_{i,t_1}) - (\overline{Y}_{n-1})_i| = O_p(n^{-1/2})$ for $i = 1,\cdots,p$, $(U_{2,n})_{i,j}, (U_{3,n})_{i,j}, (U_{4,n})_{i,j}$ are $O_p(n^{1/2})$. Thus, these facts yield (A.0.2). Then the conclusion that $\widehat{\lambda}_j - \lambda_j = O_p(n^{-1/2})$, $j = 1, 2, \cdots, s$ follows since $|tr(\gamma_j \gamma_j^T \{\widehat{\Gamma}_1 - \Gamma_1\})| \leq \|\gamma_j\|^2 \|\widehat{\Gamma}_1 - \Gamma_1\|_2 = O_p(n^{-1/2})$ and $|R_1| = O_p(n^{-1})$ under condition (C1).

Regarding the eigenvector, we apply part (b) of that lemma and get that for $j = 1, \cdots, s$,

$$\widehat{\gamma}_j - \gamma_j = -S_j(\Gamma_1)(\widehat{\Gamma}_1 - \Gamma_1)\gamma_j + R_3,$$

where $S_j(\Gamma_1) = \sum_{h \neq j} \frac{1}{\lambda_h - \lambda_j} \gamma_h \gamma_h^T$ and $\|R_3\|_2 \leq \frac{6\|\widehat{\Gamma}_1 - \Gamma_1\|_2^2}{\min_{\lambda \in EG(\Gamma_1), \lambda \neq \lambda_j} |\lambda - \lambda_j|^2}$. Then $\sqrt{n}\|R_3\|_2 = O_p(n^{-1/2})$ and $\| - S_j(\Gamma_1)(\widehat{\Gamma}_1 - \Gamma_1)\gamma_j\|_2^2 = \sum_{h \neq j} \frac{\{\gamma_h^T (\widehat{\Gamma}_1 - \Gamma_1)\gamma_j\}^2}{(\lambda_h - \lambda_j)^2} = O_p(n^{-1})$ which yields that $\|\widehat{\gamma}_j - \gamma_j\|_2 = O_p(n^{-1/2})$.

To show (ii), we note that part (a) of Lemma A1 of Kneip and Utikal (2001) implies that

$$\sum_{j=s+1}^{p} (\widehat{\lambda}_j - \lambda_j) = tr(\gamma_{s+1}\gamma_{s+1}^T \{\widehat{\Gamma}_1 - \Gamma_1\}) + R_2,$$

where $|R_2| \leq \min(p - s, s) \frac{6\|\widehat{\Gamma}_1 - \Gamma_1\|_2^2}{\min_{\lambda \in EG(\Gamma_1), \lambda \neq \lambda_{s+1}} |\lambda - \lambda_{s+1}|} = O_p(n^{-1})$. Furthermore, we write $tr(\gamma_{s+1}\gamma_{s+1}^T \{\widehat{\Gamma}_1 - \Gamma_1\}) = \gamma_{s+1}^T \widehat{\Gamma}_1 \gamma_{s+1} - \gamma_{s+1}^T \Gamma_1 \gamma_{s+1} = \frac{n-2}{(n-1)}(V_{1,n} + V_{2,n} + V_{3,n} + V_{4,n})$,

where

$$\begin{aligned}
V_{1,n} &= \frac{-1}{(n-1)(n-2)}\sum_{t_1=2}^{n}\sum_{t_2\neq t_1}\{\gamma_{s+1}^T Y_{t_1} - E(\gamma_{s+1}^T Y_{t_1})\}\{\gamma_{s+1}^T Y_{t_2} - E(\gamma_{s+1}^T Y_{t_2})\} \\
&\quad \times |Y_{t_1-1} - Y_{t_2-1}|_p \\
V_{2,n} &= \frac{-1}{(n-1)(n-2)}\sum_{t_1=2}^{n}\sum_{t_2\neq t_1}\{\gamma_{s+1}^T Y_{t_1} - E(\gamma_{s+1}^T Y_{t_1})\}\{E(\gamma_{s+1}^T Y_{t_2}) - \gamma_{s+1}^T \overline{Y}_{n-1}\} \\
&\quad \times |Y_{t_1-1} - Y_{t_2-1}|_p \\
V_{3,n} &= \frac{-1}{(n-1)(n-2)}\sum_{t_1=2}^{n}\sum_{t_2\neq t_1}\{E(\gamma_{s+1}^T Y_{t_1}) - \gamma_{s+1}^T \overline{Y}_{n-1}\}\{\gamma_{s+1}^T Y_{t_2} - E(\gamma_{s+1}^T Y_{t_2})\} \\
&\quad \times |Y_{t_1-1} - Y_{t_2-1}|_p \\
V_{4,n} &= \frac{-1}{(n-1)(n-2)}\sum_{t_1=2}^{n}\sum_{t_2\neq t_1}\{E(\gamma_{s+1}^T Y_{t_1}) - \gamma_{s+1}^T \overline{Y}_{n-1}\}\{E(\gamma_{s+1}^T Y_{t_2}) - \gamma_{s+1}^T \overline{Y}_{n-1}\} \\
&\quad \times |Y_{t_1-1} - Y_{t_2-1}|_p
\end{aligned}$$

Since $\lambda_{s+1} = 0$, $MDD(\gamma_{s+1}^T Y_2|Y_1) = 0$, i.e., $E(\gamma_{s+1}^T Y_2|Y_1) = E(\gamma_{s+1}^T Y_2)$ almost surely. This implies that $V_{1,n}$ is a degenerate U-statistic of order 1. Thus

$$\begin{aligned}
E(V_{1,n}^2) &= O(n^{-4})\sum_{t_1=2}^{n}\sum_{t_3=2}^{n}\sum_{t_2=2}^{t_1-1}\sum_{t_4=2}^{t_3-1} E\{\{\gamma_{s+1}^T Y_{t_1} - E(\gamma_{s+1}^T Y_{t_1})\} \\
&\quad \times\{\gamma_{s+1}^T Y_{t_2} - E(\gamma_{s+1}^T Y_{t_2})\}\{\gamma_{s+1}^T Y_{t_3} - E(\gamma_{s+1}^T Y_{t_3})\} \\
&\quad \times\{\gamma_{s+1}^T Y_{t_4} - E(\gamma_{s+1}^T Y_{t_4})\}|Y_{t_1-1} - Y_{t_2-1}|_p |Y_{t_3-1} - Y_{t_4-1}|_p\} \\
&= O(n^{-4})\sum_{t_1=2}^{n}\sum_{t_2=2}^{t_1-1}\sum_{t_4=2}^{t_1-1} E\{\{\gamma_{s+1}^T Y_{t_1} - E(\gamma_{s+1}^T Y_{t_1})\}^2\{\gamma_{s+1}^T Y_{t_2} - E(\gamma_{s+1}^T Y_{t_2})\} \\
&\quad \{\gamma_{s+1}^T Y_{t_4} - E(\gamma_{s+1}^T Y_{t_4})\}|Y_{t_1-1} - Y_{t_2-1}|_p |Y_{t_1-1} - Y_{t_4-1}|_p\} \\
&= O(n^{-4})\sum_{t_1=2}^{n}\sum_{t_2=2}^{t_1-1}\sum_{t_4=2}^{t_2} E\{\{\gamma_{s+1}^T Y_{t_1} - E(\gamma_{s+1}^T Y_{t_1})\}^2\{\gamma_{s+1}^T Y_{t_2} - E(\gamma_{s+1}^T Y_{t_2})\} \\
&\quad \{\gamma_{s+1}^T Y_{t_4} - E(\gamma_{s+1}^T Y_{t_4})\}|Y_{t_1-1} - Y_{t_2-1}|_p |Y_{t_1-1} - Y_{t_4-1}|_p\}
\end{aligned}$$

To simplify the notation, we denote

$H(t_1, t_1-1, t_2, t_2-1, t_4, t_4-1) = E(\xi(t_1, t_1-1, t_2, t_2-1)\xi(t_1, t_1-1, t_4, t_4-1))$, where

$$\begin{aligned}
\xi(t_1, t_1-1, t_2, t_2-1) &= \{\gamma_{s+1}^T Y_{t_1} - E(\gamma_{s+1}^T Y_{t_1})\}\{\gamma_{s+1}^T Y_{t_2} - E(\gamma_{s+1}^T Y_{t_2})\} \\
&\quad \times |Y_{t_1-1} - Y_{t_2-1}|_p \\
\xi(t_1, t_1-1, t_4, t_4-1) &= \{\gamma_{s+1}^T Y_{t_1} - E(\gamma_{s+1}^T Y_{t_1})\}\{\gamma_{s+1}^T Y_{t_4} - E(\gamma_{s+1}^T Y_{t_4})\} \\
&\quad \times |Y_{t_1-1} - Y_{t_4-1}|_p
\end{aligned}$$

Then $E(V_{1,n}^2) = O(n^{-4}) \sum_{t_1=2}^{n} \sum_{t_2=2}^{t_1-1} \sum_{t_4=2}^{t_2} H(t_1, t_1-1, t_2, t_2-1, t_4, t_4-1)$. Write

$$\begin{aligned}
& H(t_1, t_1-1, t_2, t_2-1, t_4, t_4-1) \\
= \;& E(\xi(t_1, t_1-1, t_2, t_2-1)\xi(t_1, t_1-1, t_4, t_4-1)) \\
= \;& E[E\{\xi(t_1, t_1-1, t_2, t_2-1)\xi(t_1, t_1-1, t_4, t_4-1)|\mathcal{F}_{t_4}, Y_{t_1}, Y_{t_1-1}\}] \\
= \;& E[\xi(t_1, t_1-1, t_4, t_4-1)E\{\xi(t_1, t_1-1, t_2, t_2-1)|\mathcal{F}_{t_4}, Y_{t_1}, Y_{t_1-1}\}]
\end{aligned}$$

Under the m-dependence assumption for $\{Y_t\}$, we shall show that $H(t_1, t_1-1, t_2, t_2-1, t_4, t_4-1) = 0$ whenever $|(t_2-1) - t_4| > m$ and $|(t_1-1) - t_2| > m$. To see this, we note that

$$\begin{aligned}
& E\{\xi(t_1, t_1-1, t_2, t_2-1)|\mathcal{F}_{t_4}, Y_{t_1}, Y_{t_1-1}\} \\
= \;& E\{\{\gamma_{s+1}^T Y_{t_1} - E(\gamma_{s+1}^T Y_{t_1})\}\{\gamma_{s+1}^T Y_{t_2} - E(\gamma_{s+1}^T Y_{t_2})\}|Y_{t_1-1} - Y_{t_2-1}|_p|\mathcal{F}_{t_4}, Y_{t_1}, Y_{t_1-1}\} \\
= \;& \{\gamma_{s+1}^T Y_{t_1} - E(\gamma_{s+1}^T Y_{t_1})\}E\{\{\gamma_{s+1}^T Y_{t_2} - E(\gamma_{s+1}^T Y_{t_2})\}|Y_{t_1-1} - Y_{t_2-1}|_p|\mathcal{F}_{t_4}, Y_{t_1}, Y_{t_1-1}\} \\
= \;& \{\gamma_{s+1}^T Y_{t_1} - E(\gamma_{s+1}^T Y_{t_1})\} \times \\
& E[E\{\{\gamma_{s+1}^T Y_{t_2} - E(\gamma_{s+1}^T Y_{t_2})\}|Y_{t_1-1} - Y_{t_2-1}|_p|Y_{t_2-1}, \mathcal{F}_{t_4}, Y_{t_1}, Y_{t_1-1}\}|\mathcal{F}_{t_4}, Y_{t_1}, Y_{t_1-1}] \\
= \;& \{\gamma_{s+1}^T Y_{t_1} - E(\gamma_{s+1}^T Y_{t_1})\} \times \\
& E[|Y_{t_1-1} - Y_{t_2-1}|_p E\{\{\gamma_{s+1}^T Y_{t_2} - E(\gamma_{s+1}^T Y_{t_2})\}|Y_{t_2-1}, \mathcal{F}_{t_4}, Y_{t_1}, Y_{t_1-1}\}|\mathcal{F}_{t_4}, Y_{t_1}, Y_{t_1-1}]
\end{aligned}$$

Since $\lambda_{s+1} = 0$, which implies that $E(\gamma_{s+1}^T Y_{t_2} | Y_{t_2-1}) = E(\gamma_{s+1}^T Y_{t_2})$ almost surely. Due to the independence between Y_{t_2} and $(\mathcal{F}_{t_4}, Y_{t_1}, Y_{t_1-1})$ under the m-dependence assumption, we can derive that

$$E\{\{\gamma_{s+1}^T Y_{t_2} - E(\gamma_{s+1}^T Y_{t_2})\} | Y_{t_2-1}, \mathcal{F}_{t_4}, Y_{t_1}, Y_{t_1-1}\} = E\{\{\gamma_{s+1}^T Y_{t_2} - E(\gamma_{s+1}^T Y_{t_2})\} | Y_{t_2-1}\} = 0,$$

which implies that $E\{\xi(t_1, t_1-1, t_2, t_2-1) | \mathcal{F}_{t_4}, Y_{t_1}, Y_{t_1-1}\} = 0$ and $H(t_1, t_1-1, t_2, t_2-1, t_4, t_4-1) = 0$.

Under the finite 6th moment assumption for Y_t, it follows from Cauchy-Swartz inequality that $|H(t_1, t_1-1, t_2, t_2-1, t_4, t_4-1)| \le C$ for any (t_1, t_2, t_4). Thus we have

$$\begin{aligned} E(V_{1,n}^2) &= O(n^{-4}) \sum_{t_1=2}^{n} \sum_{|(t_1-1)-t_2| \le m} \sum_{t_4=2}^{t_2} H(t_1, t_1-1, t_2, t_2-1, t_4, t_4-1) \\ &+ O(n^{-4}) \sum_{t_1=2}^{n} \sum_{|(t_1-1)-t_2| > m} \sum_{|(t_2-1)-t_4| \le m} H(t_1, t_1-1, t_2, t_2-1, t_4, t_4-1) \\ &= O(n^{-2}), \end{aligned}$$

which yields $V_{1,n} = O_p(n^{-1})$. By a similar but simpler argument, we can show that $V_{j,n} = O_p(n^{-1})$ for $j = 2, 3, 4$. Hence $\sum_{j=s+1}^{p} (\widehat{\lambda}_j - \lambda_j) = O_p(n^{-1})$ which implies that $\widehat{\lambda}_j - \lambda_j = O_p(n^{-1})$ for $j = s+1, \cdots, p$. The proof is now complete.

$\diamondsuit$

Proof of Proposition 2.4.1: If $(\epsilon_t, \mathcal{F}_t)$ is a martingale difference sequence, it implies that $E(\epsilon_t | \mathcal{F}_{t-1}) = 0$, which leads to

$$E(Y_t | \mathcal{F}_{t-1}) = E(AX_t + \epsilon_t | \mathcal{F}_{t-1}) = AE(X_t | \mathcal{F}_{t-1})$$

The conclusion follows by letting $Z_t = E(X_t | \mathcal{F}_{t-1})$ and $\mathcal{A} = A$. $\diamondsuit$

Proof of Proposition 3.2.1: To show the first assertion, we define $V_\alpha = YY^T\alpha$ for any $\alpha \in R^p$, which is a $p \times 1$ random vector. Let $V'_\alpha = Y'(Y')^T\alpha$. Then by definition

$$\begin{aligned} MDD(V_\alpha|X)^2 &= -E[(V_\alpha - E(V_\alpha))^T(V'_\alpha - E(V'_\alpha))|X - X'|_q] \\ &= -\alpha^T E[(YY^T - \Sigma)(Y'(Y')^T - \Sigma)|X - X'|_q]\alpha \\ &= \alpha^T VMDDM(Y|X)\alpha \geq 0. \end{aligned}$$

So $VMDDM(Y|X)$ is positive semidefinite.

To show the second assertion, we note that for any $\alpha \neq 0$,

$$\begin{aligned} VMMDM(Y|X)\alpha = 0 &\iff \alpha^T VMDDM(Y|X)\alpha = 0 \\ &\iff MDD(V_\alpha|X)^2 = 0 \\ &\iff E(V_\alpha|X) = E(V_\alpha) \text{ a.s.} \\ &\iff E((YY^T - \Sigma)|X)\alpha = 0 \text{ a.s.} \\ &\iff \alpha^T E((YY^T - \Sigma)|X)\alpha = 0 \text{ a.s.} \\ &\iff E((\alpha^T Y)^2|X) = E((\alpha^T Y)^2) \text{ a.s.} \end{aligned}$$

Thus if $VMDDM(Y|X)$ is singular and α is in its null space, then the conditional variance of $\alpha^T Y$ given X is a constant. On the other hand, if for $\alpha \neq 0$, $E((\alpha^T Y)^2|X) = E((\alpha^T Y)^2)$, then $VMDDM(Y|X)\alpha = 0$ according to the equivalence relations stated above. Thus $VMDDM(Y|X)$ is singular. The conclusion follows. ♢

Proof of Theorem 3.2.1: For simplicity, we prove the above theorem assuming $k_0 = 1$ and the proof for general k_0 can be extended in a similar fashion. The main arguments follow from that used in the proof of Theorem 4.1 in Lee and Shao (2016).

According to Lemma A.1 in Kneip and Utikal (2001),

$$\widehat{\lambda}_i - \lambda_i = tr(\gamma_i \gamma_i^T (\widehat{V}_1 - V_1)) + R_1, \ \ |R_1| \leq \frac{6\|\widehat{V}_1 - V_1\|_2^2}{min_{\lambda \in EG(V_1), \lambda \neq \lambda_i} |\lambda - \lambda_i|}, \ \ i = 1, \cdots, s$$

We claim that $\|\widehat{V}_1 - V_1\|_2^2 = O_p(n^{-1})$.

Note that

$$\begin{aligned} \|\widehat{V}_1 - V_1\|_2^2 &\leq \|\widehat{V}_1 - V_1\|_F^2 \\ &\leq \sum_{i=1}^{p} \sum_{j=1}^{p} |(\widehat{V}_1)_{ij} - (V_1)_{ij}|^2 \\ &\leq \sum_{i_1=1}^{p} \sum_{j_1=1}^{p} \sum_{i_2=1}^{p} \sum_{j_2=1}^{p} |(\widehat{V}_1)_{i_1 j_1 i_2 j_2} - (V_1)_{i_1 j_1 i_2 j_2}|^2 \end{aligned}$$

where $(\widehat{V}_1)_{i_1 j_1 i_2 j_2} = -\frac{1}{(n-1)^2} \sum_{t_1=2}^{n} \sum_{t_2=2}^{n} (Y_{i_1,t_1} Y_{j_1,t_1} - \Sigma_{n,i_1,j_1})(Y_{i_2,t_2} Y_{j_2,t_2} - \Sigma_{n,i_2,j_2}) |Y_{t_1-1} - Y_{t_2-1}|_p$ and $\Sigma_n = \frac{1}{n} \sum_{t=1}^{n} (Y_t - \overline{Y}_n)(Y_t - \overline{Y}_n)^T$ where $\overline{Y}_n = \frac{1}{n} \sum_{t=1}^{n} Y_t$.

Note that

$$(\widehat{V}_1)_{i_1 j_1 i_2 j_2} = \frac{(n-2)}{(n-1)} \{(U_{1,n})_{i_1 j_1 i_2 j_2} + (U_{2,n})_{i_1 j_1 i_2 j_2} + (U_{3,n})_{i_1 j_1 i_2 j_2} + (U_{4,n})_{i_1 j_1 i_2 j_2}\}$$

where

$$
\begin{aligned}
(U_{1,n})_{i_1j_1i_2j_2} &= -\frac{1}{(n-1)(n-2)}\sum_{t_1}\sum_{t_1\neq t_2}(Y_{i_1,t_1}Y_{j_1,t_1}-\Sigma_{i_1,j_1})(Y_{i_2,t_2}Y_{j_2,t_2}-\Sigma_{i_2,j_2}) \\
&\quad \times|Y_{t_1-1}-Y_{t_2-1}|_p \\
(U_{2,n})_{i_1j_1i_2j_2} &= -\frac{1}{(n-1)(n-2)}\sum_{t_1}\sum_{t_1\neq t_2}(Y_{i_1,t_1}Y_{j_1,t_1}-\Sigma_{i_1,j_1})(\Sigma_{i_2,j_2}-\Sigma_{n,i_2,j_2}) \\
&\quad \times|Y_{t_1-1}-Y_{t_2-1}|_p \\
(U_{3,n})_{i_1j_1i_2j_2} &= -\frac{1}{(n-1)(n-2)}\sum_{t_1}\sum_{t_1\neq t_2}(\Sigma_{i_1,j_1}-\Sigma_{n,i_1,j_1})(Y_{i_2,t_2}Y_{j_2,t_2}-\Sigma_{i_2,j_2}) \\
&\quad \times|Y_{t_1-1}-Y_{t_2-1}|_p \\
(U_{4,n})_{i_1j_1i_2j_2} &= -\frac{1}{(n-1)(n-2)}\sum_{t_1}\sum_{t_1\neq t_2}(\Sigma_{i_1,j_1}-\Sigma_{n,i_1,j_1})(\Sigma_{i_2,j_2}-\Sigma_{n,i_2,j_2}) \\
&\quad \times|Y_{t_1-1}-Y_{t_2-1}|_p
\end{aligned}
$$

and $(U_{1,n})_{i_1,j_1,j_2,j_2}$ is a U-statistic of order 2 with the following kernel.

$$
g(Z_2,Z_2^{'}) = -(Y_{i_1,2}Y_{j_1,2}-\Sigma_{i_1,j_1})(Y_{i_2,2}^{'}Y_{j_2,2}^{'}-\Sigma_{i_2,j_2})|Y_1-Y_1^{'}|_p
$$

where $Z_t=(Y_t^T,Y_{t-1}^T)^T$. By applying Theorem in Yoshihara (1976), $|(U_{1,n})_{i_1,j_1,i_2,j_2}-(V_1)_{i_1,j_1,i_2,j_2}|_2^2=O_p(n^{-1})$ for $i_1,j_1,i_2,j_2=1,\cdots,p$. Since $|\Sigma_{i_1,j_1}-\Sigma_{n,i_1,j_1}|=O_p(n^{-1/2})$ and $|\Sigma_{i_2,j_2}-\Sigma_{n,i_2,j_2}|=O_p(n^{-1/2})$, $(U_{i,n})_{i_1,j_1,i_2,j_2}=O_p(n^{-1/2})$ for $i=2,3,4$. Therefore, $|tr(\gamma_i\gamma_i^T(\widehat{V}_1-V_1)|\leq\|\gamma_j\|^2\|\widehat{V}_1-V_1\|_2=O_p(n^{-1/2})$ and $|R_1|=O_p(n^{-1})$. Finally, $\widehat{\lambda}_i-\lambda_i=O_p(n^{-1/2}),\ \ i=1,\cdots,s.$

Based on part (b) of Lemma A.1 in Kneip and Utikal (2001),

$$
\widehat{\gamma}_i-\gamma_i=-S_i(V_1)(\widehat{V}_1-V_1)\gamma_i+R_3,\ \ i=1,\cdots,s
$$

where $S_i(V_1)=\sum_{h\neq i}\frac{1}{\lambda_h-\lambda_i}\gamma_h\gamma_h^T$, $\|R_3\|\leq\frac{6\|\widehat{V}_1-V_1\|_2^2}{min_{\lambda\in EG(V_1),\lambda\neq\lambda_i}|\lambda-\lambda_i|^2}=O_p(n^{-1})$.

$$
\begin{aligned}
\| - S_i(V_1)(\widehat{V}_1 - V_1)\gamma_i\|_2^2 &= \sum_{h \neq i} \frac{(\gamma_h^T(\widehat{V}_1 - V_1)\gamma_i)^2}{(\lambda_h - \lambda_i)^2} \\
&\leq \sum_{h \neq i} \frac{1}{(\lambda_h - \lambda_i)^2} \|\gamma_h\|^2 \|\widehat{V}_1 - V_1\|_2^2 \|\gamma_i\|^2 = O_p(n^{-1})
\end{aligned}
$$

Therefore, $\widehat{\gamma}_i - \gamma_i = O_p(n^{-1/2}), \quad i = 1, \cdots, s.$

In order to show the third assertion in Theorem 3.2.1, we start from part (a) of the lemma.

$$
\sum_{i=s+1}^{p} (\widehat{\lambda}_i - \lambda_i) = \sum_{i=s+1}^{p} \widehat{\lambda}_i = tr(\gamma_{s+1}\gamma_{s+1}^T(\widehat{V}_1 - V_1)) + R_2
$$

Where $|R_2| \leq min(p-s, s)\frac{6\|\widehat{V}_1 - V_1\|_2^2}{min_{\lambda \in EG(V_1), \lambda \neq \lambda_i}|\lambda - \lambda_{s+1}|} = O_p(n^{-1})$.

$$
\begin{aligned}
tr(\gamma_{s+1}\gamma_{s+1}^T(\widehat{V}_1 - V_1)) &= \gamma_{s+1}^T \widehat{V}_1 \gamma_{s+1} - \gamma_{s+1}^T V_1 \gamma_{s+1} \\
&= \frac{(n-2)}{(n-1)}(\mathcal{V}_{1,n} + \mathcal{V}_{2,n} + \mathcal{V}_{3,n} + \mathcal{V}_{4,n})
\end{aligned}
$$

where

$$\begin{aligned}
\mathcal{V}_{1,n} &= \frac{-1}{(n-1)(n-2)}\sum_{t_1=2}^{n}\sum_{t_2\neq t_1}\gamma_{s+1}^T(Y_{t_1}Y_{t_1}^T-\Sigma)(Y_{t_2}Y_{t_2}^T-\Sigma)^T\gamma_{s+1}|Y_{t_1-1}-Y_{t_2-1}|_p \\
\mathcal{V}_{2,n} &= \frac{-1}{(n-1)(n-2)}\sum_{t_1=2}^{n}\sum_{t_2\neq t_1}\gamma_{s+1}^T(Y_{t_1}Y_{t_1}^T-\Sigma)(\Sigma-\Sigma_n)^T\gamma_{s+1}|Y_{t_1-1}-Y_{t_2-1}|_p \\
\mathcal{V}_{3,n} &= \frac{-1}{(n-1)(n-2)}\sum_{t_1=2}^{n}\sum_{t_2\neq t_1}\gamma_{s+1}^T(\Sigma-\Sigma_n)(Y_{t_2}Y_{t_2}^T-\Sigma)^T\gamma_{s+1}|Y_{t_1-1}-Y_{t_2-1}|_p \\
\mathcal{V}_{4,n} &= \frac{-1}{(n-1)(n-2)}\sum_{t_1=2}^{n}\sum_{t_2\neq t_1}\gamma_{s+1}^T(\Sigma-\Sigma_n)(\Sigma-\Sigma_n)^T\gamma_{s+1}|Y_{t_1-1}-Y_{t_2-1}|_p
\end{aligned}$$

Note that $\lambda_{s+1}=0$ which implies $V_1\gamma_{s+1}=0 \Leftrightarrow MDD(\gamma_{s+1}^T Y_2|Y_1)^2=0$

$\Leftrightarrow E[(\gamma_{s+1}^T Y_2)^2|Y_1]=E[(\gamma_{s+1}^T Y_2)^2]\ \ a.s.$

$$\begin{aligned}
E(\mathcal{V}_{1,n}^2) &= O(n^{-4})\sum_{t_1=2}^{n}\sum_{t_3=2}^{n}\sum_{t_2=2}^{t_1-1}\sum_{t_4=2}^{t_3-1}E[\gamma_{s+1}^T(Y_{t_1}Y_{t_1}^T-\Sigma)(Y_{t_2}Y_{t_2}^T-\Sigma)^T\gamma_{s+1} \\
&\times\ \gamma_{s+1}^T(Y_{t_3}Y_{t_3}^T-\Sigma)(Y_{t_4}Y_{t_4}^T-\Sigma)^T\gamma_{s+1}|Y_{t_1-1}-Y_{t_2-1}|_p|Y_{t_3-1}-Y_{t_4-1}|_p] \\
&= O(n^{-4})\sum_{\substack{t^{(1)},t^{(2)},t^{(3)},t^{(4)}\\ t_1>t_2,t_3>t_4}}E[\gamma_{s+1}^T(Y_{t_1}Y_{t_1}^T-\Sigma)(Y_{t_2}Y_{t_2}^T-\Sigma)^T\gamma_{s+1} \\
&\times\ \gamma_{s+1}^T(Y_{t_3}Y_{t_3}^T-\Sigma)(Y_{t_4}Y_{t_4}^T-\Sigma)^T\gamma_{s+1}|Y_{t_1-1}-Y_{t_2-1}|_p|Y_{t_3-1}-Y_{t_4-1}|_p] \\
&= O(n^{-4})\{\sum_{\substack{t^{(1)},t^{(2)},t^{(3)},t^{(4)}\\ t_1>t_2,t_3>t_4\\ (t^{(1)}-t^{(2)})>m}}E[\gamma_{s+1}^T(Y_{t_1}Y_{t_1}^T-\Sigma)(Y_{t_2}Y_{t_2}^T-\Sigma)^T\gamma_{s+1} \\
&\times\ \gamma_{s+1}^T(Y_{t_3}Y_{t_3}^T-\Sigma)(Y_{t_4}Y_{t_4}^T-\Sigma)^T\gamma_{s+1}|Y_{t_1-1}-Y_{t_2-1}|_p|Y_{t_3-1}-Y_{t_4-1}|_p] \\
&+ \sum_{\substack{t^{(1)},t^{(2)},t^{(3)},t^{(4)}\\ t_1>t_2,t_3>t_4\\ (t^{(1)}-t^{(2)})\leq m}}E[\gamma_{s+1}^T(Y_{t_1}Y_{t_1}^T-\Sigma)(Y_{t_2}Y_{t_2}^T-\Sigma)^T\gamma_{s+1} \\
&\times\ \gamma_{s+1}^T(Y_{t_3}Y_{t_3}^T-\Sigma)(Y_{t_4}Y_{t_4}^T-\Sigma)^T\gamma_{s+1}|Y_{t_1-1}-Y_{t_2-1}|_p|Y_{t_3-1}-Y_{t_4-1}|_p]\}
\end{aligned}$$

where $t^{(i)}$ is the i-th largest integer among (t_1,t_2,t_3,t_4), i.e. If $(t_1,t_2,t_3,t_4)=(5,3,4,2)$, then $t^{(1)} = t_1 = 5, \;\; t^{(2)} = t_3 = 4, \;\; t^{(3)} = t_2 = 3, \;\; t^{(4)} = t_4 = 2.$

Let $H(t_1, t_1-1, t_2, t_2-1, t_3, t_3-1, t_4, t_4-1) = E[\xi(t_1, t_1-1, t_2, t_2-1)\xi(t_3, t_3-1, t_4, t_4-1)]$, where $\xi(t_1, t_1-1, t_2, t_2-1) = \gamma_{s+1}^T(Y_{t_1}Y_{t_1}^T - \Sigma)(Y_{t_2}Y_{t_2}^T - \Sigma)^T\gamma_{s+1}|Y_{t_1-1} - Y_{t_2-1}|_p$. Then

$$E(\mathcal{V}_{1,n}^2) = O(n^{-4})\{\sum_{\substack{t^{(1)},t^{(2)},t^{(3)},t^{(4)} \\ t_1>t_2,t_3>t_4 \\ (t^{(1)}-t^{(2)})>m}} H(t_1, t_1-1, t_2, t_2-1, t_3, t_3-1, t_4, t_4-1)$$

$$+ \sum_{\substack{t^{(1)},t^{(2)},t^{(3)},t^{(4)} \\ t_1>t_2,t_3>t_4 \\ (t^{(1)}-t^{(2)})\le m}} H(t_1, t_1-1, t_2, t_2-1, t_3, t_3-1, t_4, t_4-1)\}.$$

If $t^{(1)} - t^{(2)} > m$ and $t^{(1)} = t_1, \;\; t^{(2)} = t_3$,

$$\begin{aligned}
&H(t_1, t_1-1, t_2, t_2-1, t_3, t_3-1, t_4, t_4-1) = E[\xi(t_1, t_1-1, t_2, t_2-1) \\
&\times\xi(t_3, t_3-1, t_4, t_4-1)] \\
= \;& E[E[\xi(t_1, t_1-1, t_2, t_2-1)\xi(t_3, t_3-1, t_4, t_4-1)|\mathcal{F}_{t_3}]] \\
= \;& E[E[\xi(t_1, t_1-1, t_2, t_2-1)|\mathcal{F}_{t_3}]\xi(t_3, t_3-1, t_4, t_4-1)] \\
= \;& E[E[\gamma_{s+1}^T(Y_{t_1}Y_{t_1}^T - \Sigma)|Y_{t_1-1} - Y_{t_2-1}|_p|\mathcal{F}_{t_3}](Y_{t_2}Y_{t_2}^T - \Sigma)^T\gamma_{s+1}\xi(t_3, t_3-1, t_4, t_4-1)] \\
= \;& E[E[E[\gamma_{s+1}^T(Y_{t_1}Y_{t_1}^T - \Sigma)|\mathcal{F}_{t_3}, Y_{t_1-1}]|Y_{t_1-1} - Y_{t_2-1}|_p|\mathcal{F}_{t_3}](Y_{t_2}Y_{t_2}^T - \Sigma)^T \\
&\times\gamma_{s+1}\xi(t_3, t_3-1, t_4, t_4-1)] \\
= \;& 0.
\end{aligned}$$

Similarly, the other cases such as ($t^{(1)} - t^{(2)} > m$ and $t^{(1)} = t_1, \;\; t^{(2)} = t_2$), ($t^{(1)} - t^{(2)} > m$ and $t^{(1)} = t_3, \;\; t^{(2)} = t_1$), ($t^{(1)} - t^{(2)} > m$ and $t^{(1)} = t_3, \;\; t^{(2)} = t_4$) have $H(t_1, t_1, t_2, t_2-1, t_3, t_3-1, t_4, t_4-1) = 0$.

Therefore, $E(\mathcal{V}_{1,n}^2) = O(n^{-4}) \sum_{\substack{t^{(1)},t^{(2)},t^{(3)},t^{(4)} \\ t_1>t_2,t_3>t_4 \\ (t^{(1)}-t^{(2)})\leq m}} H(t_1, t_1-1, t_2, t_2-1, t_3, t_3-1, t_4, t_4-1)$.

When $t^{(1)} - t^{(2)} \leq m$, $t^{(2)} - t^{(3)} - 1 > m$, $t^{(3)} - t^{(4)} - 1 > m$ and $t^{(1)} = t_1, \quad t^{(2)} = t_3$,

$$\begin{aligned}
& H(t_1, t_1-1, t_2, t_2-1, t_3, t_3-1, t_4, t_4-1) \\
= \quad & E[\xi(t_1, t_1-1, t_2, t_2-1)\xi(t_3, t_3-1, t_4, t_4-1)] \\
= \quad & E[E[\xi(t_1, t_1-1, t_2, t_2-1)\xi(t_3, t_3-1, t_4, t_4-1)|Y_{t_1}, Y_{t_1-1}, Y_{t_3}, Y_{t_3-1}, Y_{t_4}, Y_{t_4-1}]] \\
= \quad & E[E[\xi(t_1, t_1-1, t_2, t_2-1)|Y_{t_1}, Y_{t_1-1}, Y_{t_3}, Y_{t_3-1}, Y_{t_4}, Y_{t_4-1}]\xi(t_3, t_3-1, t_4, t_4-1)] \\
= \quad & E[\gamma_{s+1}^T(Y_{t_1}Y_{t_1}^T - \Sigma)E[(Y_{t_2}Y_{t_2}^T - \Sigma)^T\gamma_{s+1}|Y_{t_1-1} - Y_{t_2-1}|_p \\
\times \quad & |Y_{t_1}, Y_{t_1-1}, Y_{t_3}, Y_{t_3-1}, Y_{t_4}, Y_{t_4-1}]\xi(t_3, t_3-1, t_4, t_4-1)] \\
= \quad & E[\gamma_{s+1}^T(Y_{t_1}Y_{t_1}^T - \Sigma)E[E[(Y_{t_2}Y_{t_2}^T - \Sigma)^T\gamma_{s+1}|Y_{t_1}, Y_{t_1-1}, Y_{t_3}, Y_{t_3-1}, Y_{t_4}, Y_{t_4-1}, Y_{t_2-1}] \\
\times \quad & |Y_{t_1-1} - Y_{t_2-1}|_p|Y_{t_1}, Y_{t_1-1}, Y_{t_3}, Y_{t_3-1}, Y_{t_4}, Y_{t_4-1}]\xi(t_3, t_3-1, t_4, t_4-1)] \\
= \quad & 0.
\end{aligned}$$

Furthermore, it can be shown that the other cases, ($t^{(2)} - t^{(3)} - 1 > m$, $t^{(3)} - t^{(4)} - 1 > m$ and $t^{(1)} = t_1, \quad t^{(2)} = t_2$), ($t^{(2)} - t^{(3)} - 1 > m$, $t^{(3)} - t^{(4)} - 1 > m$ and $t^{(1)} = t_3, \quad t^{(2)} = t_1$), ($t^{(2)} - t^{(3)} - 1 > m$, $t^{(3)} - t^{(4)} - 1 > m$ and $t^{(1)} = t_3, \quad t^{(2)} = t_4$) have $H(t_1, t_1-1, t_2, t_2-1, t_3, t_3-1, t_4, t_4-1) = 0$ through a similar fashion.

Under the condition that $E[|Y_t|^{10}] < \infty$, $|H(t_1, t_1-1, t_2, t_2-1, t_3, t_3-1, t_4, t_4-1)| \leq$

C for any t_1, t_2, t_3, t_4. Therefore,

$$\begin{aligned} E(\mathcal{V}_{1,n}^2) &= O(n^{-4})\{ \sum_{\substack{t^{(1)},t^{(2)},t^{(3)},t^{(4)} \\ t_1>t_2,t_3>t_4 \\ (t^{(1)}-t^{(2)})\leq m \\ t^{(2)}-t^{(3)}-1\leq m}} H(t_1, t_1-1, t_2, t_2-1, t_3, t_3-1, t_4, t_4-1) \\ &+ \sum_{\substack{t^{(1)},t^{(2)},t^{(3)},t^{(4)} \\ t_1>t_2,t_3>t_4 \\ (t^{(1)}-t^{(2)})\leq m \\ t^{(2)}-t^{(3)}-1>m \\ t^{(3)}-t^{(4)}-1\leq m}} H(t_1, t_1-1, t_2, t_2-1, t_3, t_3-1, t_4, t_4-1)\} = O(n^{-2}). \end{aligned}$$

Thus, $\mathcal{V}_{1,n} = O_p(n^{-1})$. Similarly, $\mathcal{V}_{i,n} = O_p(n^{-1})$ for $i = 2, 3, 4$. Therefore, $\sum_{i=s+1}^{p}(\widehat{\lambda}_i - \lambda_i) = O_p(n^{-1})$ and this implies $\widehat{\lambda}_i - \lambda_i = O_p(n^{-1})$ for $i = s+1, \cdots, p$. $\diamondsuit$

Proof of Proposition 3.3.1: To show the second assertion,

$$\begin{aligned} VMDDM(Y|X)_{ij} &= \sum_{k=1}^{p} -E[(Y_iY_k - \Sigma_{ik})(Y_j^{'}Y_k^{'} - \Sigma_{jk})|X - X^{'}|_q] \\ &= \sum_{k=1}^{p} vecVMDDM(Y|X)_{(i-1)p+k,(k-1)p+j} \\ &= \sum_{k=1}^{p} vecVMDDM(Y|X)_{(i-1)p+k,(j-1)p+k}. \end{aligned}$$

With the above result, we can further show the first assertion in Proposition 3.3.1 as follows,

$$\begin{aligned} tr(VMDDM(Y|X)) &= \sum_{i=1}^{p} VMDDM(Y|X)_{ii} \\ &= \sum_{i=1}^{p}\sum_{k=1}^{p} -E[(Y_iY_k - \Sigma_{ik})(Y_i^{'}Y_k^{'} - \Sigma_{jk})|X - X^{'}|_q] \\ &= \sum_{i=1}^{p}\sum_{k=1}^{p} MDD(Y_iY_k|X)^2 \\ &= \sum_{i=1}^{p}\sum_{k=1}^{p} vecVMDDM(Y|X)_{(i-1)p+k,(i-1)p+k} \\ &= tr(vecVMDDM(Y|X)^2). \end{aligned}$$

$\diamondsuit$

Before we start the proof of Theorem 3.3.1, we first claim that $G_{k_0}(\cdot), \widehat{G}_{k_0}(\cdot)$ are Lipschitz continuous on $\mathcal{H}$ with D-distance, where $\mathcal{H}$ is a set of all $p \times p$ orthogoanl matrices.

Lemma A.2.1. *For any $U, V \in \mathcal{H}$, it holds that*

$$
\begin{aligned}
|G_{k_0}(U) - G_{k_0}(V)| &\leq c\ tr(vecV_{k_0})D(U,V)^{1/2} && \text{(A.0.3)} \\
|\widehat{G}_{k_0}(U) - \widehat{G}_{k_0}(V)| &\leq c\ tr(\widehat{vecV}_{k_0})D(U,V)^{1/2} && \text{(A.0.4)}
\end{aligned}
$$

where $c > 0$ is a general constant.

Proof of Lemma A.2.1: Let $U = (u_1, \cdots, u_p)^T, V = (v_1, \cdots, v_p)^T$, $(\lambda_i, \gamma_i)_{i=1}^{p^2}$ be eigenvalues and eigenvectors of $vecV_{k_0} \in R^{p^2 \times p^2}$ and

$$
\mathcal{U}_{ij} = \sum_{k=1}^{k_0} MDD(u_i^T Y_t Y_t^T u_j | Y_{t-k})^2,\ \mathcal{V}_{ij} = \sum_{k=1}^{k_0} MDD(v_i^T Y_t Y_t^T v_j | Y_{t-k})^2.
$$

We assume that $u_i^T v_i \in [0,1],\ \forall i = 1, \cdots, p$ and $D(U,V) = 1 - \frac{1}{p}\sum_{i=1}^{p} u_i^T v_i$ (meaning that $max_{1 \leq j \leq p} u_i^T v_j = u_i^T v_i, \forall i = 1, \cdots, p$) by arranging the orders and directions of $\{u_j\}_{j=1}^{p}$ and $\{v_j\}_{j=1}^{p}$.

$$\begin{aligned}
|\mathcal{U}_{ij} - \mathcal{V}_{ij}| &= |u_j^T \otimes u_i^T vecV_{k_0} u_j \otimes u_i - v_j^T \otimes v_i^T vecV_{k_0} v_j \otimes v_i| \\
&= |u_j^T \otimes u_i^T \sum_{l=1}^{p^2} \lambda_l \gamma_l \gamma_l^T u_j \otimes u_i - v_j^T \otimes v_i^T \sum_{l=1}^{p^2} \lambda_l \gamma_l \gamma_l^T v_j \otimes v_i| \\
&= |\sum_{l=1}^{p^2} \lambda_l \{u_j^T \otimes u_i^T \gamma_l \gamma_l^T u_j \otimes u_i - v_j^T \otimes v_i^T \gamma_l \gamma_l^T v_j \otimes v_i\}| \\
&\leq \sum_{l=1}^{p^2} \lambda_l \{|u_j^T \otimes u_i^T \gamma_l \gamma_l^T u_j \otimes u_i - v_j^T \otimes v_i^T \gamma_l \gamma_l^T v_j \otimes v_i|\} \\
&\leq \sum_{l=1}^{p^2} \lambda_l \{|(u_j^T \otimes u_i^T - v_j^T \otimes v_i^T) \gamma_l \gamma_l^T u_j \otimes u_i| \\
&\quad + |v_j^T \otimes v_i^T \gamma_l \gamma_l^T (u_j \otimes u_i - v_j \otimes v_i)|\} \\
&\leq \sum_{l=1}^{p^2} \lambda_l \{|(u_j^T \otimes u_i^T - v_j^T \otimes v_i^T) \gamma_l| \; \|\gamma_l\| \|u_j \otimes u_i\| \\
&\quad + \|v_j \otimes v_i\| \|\gamma_l\| \; |\gamma_l^T (u_j \otimes u_i - v_j \otimes v_i)|\} \\
&\leq \sum_{l=1}^{p^2} \lambda_l \{\|u_j \otimes u_i - v_j \otimes v_i\| \|\gamma_l\| + \|\gamma_l\| \|u_j \otimes u_i - v_j \otimes v_i\|\} \\
&= 2\|u_j \otimes u_i - v_j \otimes v_i\| \sum_{l=1}^{p^2} \lambda_l
\end{aligned}$$

Let $u_j = (u_{j1}, \cdots, u_{jp})^T, v_j = (v_{j1}, \cdots, v_{jp})^T$.

$$\begin{aligned}
\|u_j \otimes u_i - v_j \otimes v_i\|^2 &= \sum_{s=1}^{p}\sum_{t=1}^{p}\{u_{js}u_{it} - v_{js}v_{it}\}^2 \\
&= \sum_{s=1}^{p}\sum_{t=1}^{p}\{u_{js}(u_{it} - v_{it}) + (u_{js} - v_{js})v_{it}\}^2 \\
&= \sum_{s=1}^{p} u_{js}^2 \sum_{t=1}^{p}(u_{it} - v_{it})^2 + \sum_{j=1}^{p}(u_{js} - v_{js})^2 \sum_{t=1}^{p} v_{it}^2 \\
& \quad +2\sum_{s=1}^{p} u_{js}(u_{js} - v_{js}) \sum_{t=1}^{p}(u_{it} - v_{it})v_{it} \\
&= \sum_{t=1}^{p}(u_{it} - v_{it})^2 + \sum_{s=1}^{p}(u_{js} - v_{js})^2 \\
&+ \quad 2\sum_{s=1}^{p}(u_{js}^2 - u_{js}v_{js})\sum_{t=1}^{p}(u_{it}v_{it} - v_{it}^2) \\
&= (2 - 2u_i^T v_i) + (2 - 2u_j^T v_j) - 2(1 - u_j^T v_j)(1 - u_i^T v_i) \\
&\leq (2 - 2u_i^T v_i) + (2 - 2u_j^T v_j)
\end{aligned}$$

Therefore,

$$\begin{aligned}
\|u_j \otimes u_i - v_j \otimes v_i\| &\leq \sqrt{2}\sqrt{(1 - u_i^T v_i) + (1 - u_j^T v_j)} \\
&\leq \sqrt{2}(\sqrt{1 - u_i^T v_i} + \sqrt{1 - u_j^T v_j})
\end{aligned}$$

Finially, we have

$$\begin{aligned}
|G_{k_0}(U) - G_{k_0}(V)| &= |\sum_{i=1}^{p}\sum_{i<j}(\mathcal{U}_{ij} - \mathcal{V}_{ij})| \\
&\leq \sum_{i=1}^{p}\sum_{i<j}|\mathcal{U}_{ij} - \mathcal{V}_{ij}| \\
&\leq \sum_{i=1}^{p}\sum_{i<j}2(\sum_{l=1}^{p^2}\lambda_l)\|u_j \otimes u_i - v_j \otimes v_i\| \\
&\leq 2\sqrt{2}(\sum_{l=1}^{p^2}\lambda_l)\sum_{i=1}^{p}\sum_{i<j}(\sqrt{1 - u_i^T v_i} + \sqrt{1 - u_j^T v_j}) \\
&\leq 2\sqrt{2}(\sum_{l=1}^{p^2}\lambda_l)\{\sum_{i=1}^{p}\sum_{j=1}^{p}\sqrt{1 - u_i^T v_i} + \sum_{i=1}^{p}\sum_{j=1}^{p}\sqrt{1 - u_j^T v_j}\} \\
&= 4p\sqrt{2}(\sum_{l=1}^{p^2}\lambda_l)\sum_{i=1}^{p}\sqrt{1 - u_i^T v_i} \\
&\leq 4p\sqrt{2}(\sum_{l=1}^{p^2}\lambda_l)\sqrt{p}\sqrt{\sum_{i=1}^{p}(1 - u_i^T v_i)} \\
&= 4p^2\sqrt{2}(\sum_{l=1}^{p^2}\lambda_l)D(U,V)^{1/2}
\end{aligned}$$

Note that $(\sum_{l=1}^{p^2}\lambda_l)$ is $tr(vecV_{k_0})$ and this completes the proof of (A.0.3). (A.0.4) can be shown in a similar fashion as (A.0.3).

$\diamondsuit$

Proof of Theorem 3.3.1: From Lee and Shao (2016) and the proof of Theorem 3.2.1 (for V_{k_0}), we have $\|\widehat{vecV}_{k_0} - vecV_{k_0}\|_2 = O_p(n^{-1/2})$ by applying Theorem in Yoshihara (1976). With this fact, we will show $|\widehat{G}_{k_0}(A) - G_{k_0}(A)| = O_p(n^{-1/2})$, for any $A \in \mathcal{H}$.

$$\begin{aligned}|\widehat{G}_{k_0}(A) - G_{k_0}(A)| &= |\sum_{i=1}^{p}\sum_{i<j} a_j^T \otimes a_i^T(\widehat{vecV}_{k_0} - vecV_{k_0})a_j \otimes a_i| \\ &\leq \sum_{i=1}^{p}\sum_{i<j} \|a_j \otimes a_i\| \|\widehat{vecV}_{k_0} - vecV_{k_0}\|_2 \|a_j \otimes a_i\| \\ &= \sum_{i=1}^{p}\sum_{i<j} \|\widehat{vecV}_{k_0} - vecV_{k_0}\|_2 \\ &= \frac{p(p-1)}{2}\|\widehat{vecV}_{k_0} - vecV_{k_0}\|_2 = O_p(n^{-1/2})\end{aligned}$$

Therefore, $sup_{A\in\mathcal{H}}|\widehat{G}_{k_0}(A) - G_{k_0}(A)| = O_p(n^{-1/2})$.

With this result and Lemma A.2.1, $D(\widehat{A}_0, A_0) \rightarrow^p 0$ as $n \rightarrow \infty$ by the argmax mapping theorem (Theorem 3.2.2 and Corollary 3.2.3) in van der Vaart and Wellner (1996) .

Additionally, if we assume the condition 3 of Assumption 3.3.2, then

$$\widehat{G}_{k_0}(A_0) - \widehat{G}_{k_0}(A) = G_{k_0}(A_0) - G_{k_0}(A) + O_p(n^{-1/2}) \leq -aD(A_0, A) + O_p(n^{-1/2})$$

When $A = \widehat{A}_0$, $\widehat{G}_{k_0}(A_0) - \widehat{G}_{k_0}(\widehat{A}_0)$ has to be a non-negative number by the definition of $\widehat{A}_0$. Thus, $D(A_0, \widehat{A}_0,) = O_p(n^{-1/2})$ otherwise $\widehat{G}_{k_0}(A_0) - \widehat{G}_{k_0}(\widehat{A}_0)$ becomes negative.
Proof of Proposition 4.2.1: If $E[Y|X] = \mu_Y$ a.s., it is clear that $FMDD(Y|X) = 0$. We only need to show the other direction i.e., $FMDD(Y|X) = 0$ implies $E[Y|X] = \mu_Y$ a.s. Without loss of generality, we can assume that $\mu_Y = 0$ (otherwise we work with $Y - \mu_Y$). By Theorem 3.16 and Proposition 3.1 in Lyon (2013), there exists an embedding $\phi : \mathcal{L}_2(\mathcal{I})$ (or R^q) $\rightarrow \mathcal{H}$ such that $|x - x^{'}| = |\phi(x) - \phi(x^{'})|^2$ and $\beta_\phi(v) = \int \phi(x)dv(x)$ is injective on the set of measures v on $\mathcal{L}_2(\mathcal{I})$ (or R^q) such that $|v|$ has a finite first moment, i.e., $\int |x - o|d|v|(x) < \infty$, for some $o \in \mathcal{L}_x$. Using this

result and the definition of FMDD, we have

$$\begin{aligned} FMDD(Y|X) &= 2\int <y,y^{'}><\phi(x)-\beta_\phi(\mu),\phi(x^{'})-\beta_\phi(\mu)> d\theta(x,y)\theta(x^{'},y^{'}) \\ &= |E[Y\otimes\phi(X)]|^2 \geq 0, \end{aligned}$$

where μ denotes the distribution of X and θ is the joint distribution of (X,Y). Hence, $FMDD(Y|X)=0$ implies that

$$E[Y\otimes\phi(X)] = \int y\phi(x)d\theta(x,y) = 0 \quad a.s.$$

For any Borel set $B \subseteq \mathcal{L}_2(\mathcal{I})$ (or R^q) and $k \in \mathcal{L}_2(\mathcal{I})$ (or R^p), define the sign measure,

$$v_k(B) = \int <y,k> 1_B(x)d\theta(x,y) = E[<Y,k>1_B(X)],$$

where $|v_k|$ has a finite first moment under the assumptions that $E[|X|+|Y|]<\infty$ and $E[|X-\mu_X||Y-\mu_Y|]<\infty$. Then we have

$$\beta_\phi(v_k) = \int <y,k>\phi(x)d\theta(x,y) =< \int y\phi(x)d\theta(x,y),k >= 0.$$

The injectivity of β_ϕ gives $v_k(B) = E[<Y,k>1_B(X)] = 0$. Thus, by the definition of conditional mean independence, we have

$$E[<Y,k>|X] = 0, \tag{A.0.5}$$

for any $k \in \mathcal{L}_2(\mathcal{I})$ (or R^p). Therefore, (A.0.5) implies that $E[Y|X]=\mu_Y$, which completes the proof of Proposition 4.2.1.

$\diamondsuit$

Proof of Proposition 4.2.2: Following the arguments in Section 1.1 of the supplement of Zhang et al. (2017), we can show that $FMDD_n(Y|X)$ is an unbiased estimator of $FMDD(Y|X)$, and it is a fourth-order U-statistic which has the form of

$$\begin{aligned} FMDD_n(Y|X) &= \frac{1}{\binom{n}{4}} \sum_{i<j<q<r} h(Z_i, Z_j, Z_q, Z_r), \\ h(Z_i, Z_j, Z_q, Z_r) &= \frac{1}{4!} \sum_{(s,t,u,v)}^{(i,j,q,r)} (a_{st}b_{uv} + a_{st}b_{st} - a_{st}b_{su} - a_{st}b_{tv}), \end{aligned}$$

where $\sum_{(s,t,u,v)}^{(i,j,q,r)}$ denotes the summation over all permutations of the 4-tuple of indices (i, j, q, r) and $Z_i = (X_i, Y_i)$. Under the assumption that $E[|X| + |Y|] < \infty$ and $E[|X - \mu_X||Y - \mu_Y|] < \infty$, we have

$$\begin{aligned} E[|h(Z_i, Z_j, Z_q, Z_r)|] \leq & \frac{1}{4!} \sum_{(s,t,u,v)}^{(i,j,q,r)} E|a_{st}b_{uv} + a_{st}b_{st} - a_{st}b_{su} - a_{st}b_{tv}| \\ \leq & E[|X - X'|]E[|Y - \mu_Y|]^2 + E[|X - X'||Y - \mu_Y||Y' - \mu_Y|] \\ & + 2E[|X - X'||Y - \mu_Y|]E[|Y - \mu_Y|] < \infty. \end{aligned}$$

Proposition 4.2.2 follows from the law of large numbers for U-statistics [see e.g. Hoeffding (1961) and Lee (1990)]. ♢

Proof of Theorem 4.2.1: For $c = 1, 2, 3, 4$, define

$$h_c(z_1, \cdots, z_c) = E[h(z_1, \cdots, z_c, Z_{c+1}, \cdots, Z_4)],$$

where $z_i = (x_i, y_i)$ for $1 \leq i \leq 4$. Denote by $Z' = (X', Y')$ and $Z'' = (X'', Y'')$ two independent copies of $Z = (X, Y)$. When $FMDD(Y|X) = 0$, following the calculations in Section 1.2 of the supplement of Zhang et al. (2017), we have $h_1(z) = 0$ and

$h_2(z, z') = U(x, x')V(y, y')/6$ for $z = (x, y)$ and $z' = (x', y')$. Under the assumption $E[|X|^2 + |Y|^2] < \infty$ and $E[|X - \mu_X|^2 |Y - \mu_Y|^2] < \infty$, we have $E[h(Z_i, Z_j, Z_q, Z_r)^2] < \infty$. Applying Theorem 5.5.2 in Serfling (1980), we obtain $nFMDD_n(Y|X) \rightarrow^D \sum_{k=1}^{\infty} \lambda_k (G_k^2 - 1)$. ♢

Proof of Theorem 4.2.2: Under the local alternative $H_{1,n} : Y = \mu_Y + \frac{g(X)}{n^a} + \epsilon$, we have

$$\begin{aligned}
|Y_i - Y_j|^2 &= \frac{1}{n^{2a}} |g(X_i) - g(X_j)|^2 + |\epsilon_i - \epsilon_j|^2 + \frac{2}{n^a} < g(X_i) - g(X_j), \epsilon_i - \epsilon_j > \\
&= \frac{1}{n^{2a}} |g(X_i) - g(X_j)|^2 + |\epsilon_i - \epsilon_j|^2 \\
&+ \frac{1}{n^a} \left\{ |g(X_i) + \epsilon_i - (g(X_j) + \epsilon_j)|^2 - |g(X_i) - g(X_j)|^2 - |\epsilon_i - \epsilon_j|^2 \right\}
\end{aligned}$$

Using the above result, $FMDD_n(\frac{g(X)}{n^a} + \epsilon | X)$ can be decomposed into three terms.

$$\begin{aligned}
FMDD_n(\frac{g(X)}{n^a} + \epsilon | X) &= \frac{1}{n^{2a} n(n-3)} \sum_{i \neq j} \widetilde{A}_{ij} \widetilde{B}_{ij}^g + \frac{1}{n(n-3)} \sum_{i \neq j} \widetilde{A}_{ij} \widetilde{B}_{ij}^{\epsilon} \\
&+ \frac{1}{n^a n(n-3)} \sum_{i \neq j} \widetilde{A}_{ij} (\widetilde{B}_{ij}^{g+\epsilon} - \widetilde{B}_{ij}^g - \widetilde{B}_{ij}^{\epsilon}), \qquad \text{(A.0.6)}
\end{aligned}$$

where $\widetilde{B}_{ij}^{\epsilon} = e_{ij} - e_{i\cdot} - e_{\cdot j} + e_{\cdot\cdot}$, with $e_{ij} = \frac{1}{2}|\epsilon_i - \epsilon_j|^2$, and $e_{ij},\ e_{i\cdot},\ e_{\cdot j},\ e_{\cdot\cdot}$ are defined similarly as $b_{ij},\ b_{i\cdot},\ b_{\cdot j},\ b_{\cdot\cdot}$. Moreover $\widetilde{B}_{ij}^{g+\epsilon},\ \widetilde{B}_{ij}^g$ are defined similarly by replacing (ϵ_i, ϵ_j) in $\widetilde{B}_{ij}^{\epsilon}$ with $(g(X_i) + \epsilon_i, g(X_j) + \epsilon_j)$ and $(g(X_i), g(X_j))$.

Since $\frac{1}{n(n-3)} \sum_{i \neq j} \widetilde{A}_{ij} \widetilde{B}_{ij}^{\epsilon}$ is a degenerate U-statistic whereas $\frac{1}{n(n-3)} \sum_{i \neq j} \widetilde{A}_{ij} \widetilde{B}_{ij}^g$, $\frac{1}{n(n-3)} \sum_{i \neq j} \widetilde{A}_{ij} (\widetilde{B}_{ij}^{g+\epsilon} - \widetilde{B}_{ij}^g - \widetilde{B}_{ij}^{\epsilon})$ are nondegenerate U-statistics (to be shown below), (A.0.6) implies that

$$\begin{aligned}
nFMDD_n(\frac{g(X)}{n^a} + \epsilon | X) &= n^{1-2a} FMDD(g(X)|X) + O_p(n^{1/2-2a}) \\
&+ O_p(1) + O_p(n^{1/2-a}). \qquad \text{(A.0.7)}
\end{aligned}$$

We shall consider three scenarios: *(i)* $0 < a < 1/2$, *(ii)* $a = 1/2$, *(iii)* $a > 1/2$.

(i) $0 < a < 1/2$:

Based on (A.0.7), we can easily show that $nFMDD_n(\frac{g(X)}{n^a}+\epsilon|X) \rightarrow^p \infty$ which implies that our test has consistency under this scenario.

(iii) $a > 1/2$:

Similarly, using (A.0.7) and Theorem 4.2.1, we have

$$\begin{aligned} nFMDD_n(\frac{g(X)}{n^a}+\epsilon|X) &= \frac{1}{(n-3)}\sum_{i\neq j}\widetilde{A}_{ij}\widetilde{B}^{\epsilon}_{ij} \\ &\quad +o_p(1) \rightarrow^D \sum_{k=1}^{\infty}\lambda_k(G_k^2-1), \qquad \text{(A.0.8)} \end{aligned}$$

which is same as the limiting null distribution.

(ii) $a = 1/2$:

When $a = 1/2$, $FMDD_n(\frac{g(X)}{\sqrt{n}}+\epsilon|X)$ can be written as a sum of linear combination of two U-statistics, $U_n^{\epsilon},\ U_n^{g,\epsilon}$ and a sequence of random variables c_n where $nc_n \rightarrow^{a.s.} FMDD(g(X)|X) = c$, i.e.,

$$\begin{aligned} FMDD_n(\frac{g(X)}{\sqrt{n}}+\epsilon|X) &= \frac{1}{n^2(n-3)}\sum_{i\neq j}\widetilde{A}_{ij}\widetilde{B}^{g}_{ij} \\ &+ \frac{1}{n(n-3)}\sum_{i\neq j}\widetilde{A}_{ij}\{\widetilde{B}^{\epsilon}_{ij}+\frac{1}{\sqrt{n}}(\widetilde{B}^{g+\epsilon}_{ij}-\widetilde{B}^{g}_{ij}-\widetilde{B}^{\epsilon}_{ij})\} \\ &= c_n + U_n^{\epsilon} + \frac{1}{\sqrt{n}}U_n^{g,\epsilon}. \qquad \text{(A.0.9)} \end{aligned}$$

Specifically, $U_n^{\epsilon},\ U_n^{g,\epsilon}$ are mean zero U-statistics of degree 4, i.e.,

$$\begin{aligned} U_n^{\epsilon} &= \frac{1}{n(n-3)}\sum_{i\neq j}\widetilde{A}_{ij}\widetilde{B}^{\epsilon}_{ij} = \frac{1}{\binom{n}{4}}\sum_{i<j<q<r}H_1(\mathcal{Z}_i,\mathcal{Z}_j,\mathcal{Z}_q,\mathcal{Z}_r), \\ U_n^{g,\epsilon} &= \frac{1}{n(n-3)}\sum_{i\neq j}\widetilde{A}_{ij}(\widetilde{B}^{g+\epsilon}_{ij}-\widetilde{B}^{g}_{ij}-\widetilde{B}^{\epsilon}_{ij}) = \frac{1}{\binom{n}{4}}\sum_{i<j<q<r}h_1(\mathcal{Z}_i,\mathcal{Z}_j,\mathcal{Z}_q,\mathcal{Z}_r), \end{aligned}$$

where $\mathcal{Z}_i = (X_i, \epsilon_i)$,

$$H_1(\mathcal{Z}_i, \mathcal{Z}_j, \mathcal{Z}_q, \mathcal{Z}_r) = \frac{1}{4!} \sum_{(s,t,u,v)}^{(i,j,q,r)} a_{st}(e_{uv} + e_{st} - e_{su} - e_{tv}),$$

$$h_1(\mathcal{Z}_i, \mathcal{Z}_j, \mathcal{Z}_q, \mathcal{Z}_r) = \frac{1}{4!} \sum_{(s,t,u,v)}^{(i,j,q,r)} a_{st}(m_{uv} + m_{st} - m_{su} - m_{tv})$$

and $m_{ij} =< g(X_i) - g(X_j), \epsilon_i - \epsilon_j >$.

Define

$$\widetilde{h}_{1,i}(z_1, \cdots, z_i) = E[h_1(z_1, z_2, \cdots, z_i, \mathcal{Z}_{i+1}, \cdots, \mathcal{Z}_4)],$$

$$\widetilde{H}_{1,i}(z_1, \cdots, z_i) = E[H_1(z_1, z_2, \cdots, z_i, \mathcal{Z}_{i+1}, \cdots, \mathcal{Z}_4)] \ , i = 1, 2, 3, 4.$$

By the results in the supplement of Zhang et al. (2017) and calculations, we show that $\widetilde{h}_{1,1}(z_1) = \frac{1}{2}E[U(x_1, X)V(\epsilon_1, g(X))]$, $\widetilde{H}_{1,1}(z_1) = 0$,

$$\begin{aligned}
\widetilde{h}_{1,2}(z_1, z_2) &= \frac{1}{6}\{U(x_1, x_2)(V(\epsilon_1, g(x_2)) + V(g(x_1), \epsilon_2)) \\
&+ E[U(x_1, X)V(\epsilon_1, g(X))] + E[U(x_2, X)V(\epsilon_2, g(X))] \\
&+ E[(U(x_1, X) - U(x_2, X))(V(\epsilon_1, g(X)) - V(\epsilon_2, g(X)))]\} \\
\widetilde{H}_{1,2}(z_1, z_2) &= \frac{1}{6}U(x_1, x_2)V(\epsilon_1, \epsilon_2)
\end{aligned}$$

$$
\begin{aligned}
\widetilde{h}_{1,3}(z_1,z_2,z_3) &= \frac{1}{12}\{(2U(x_1,x_2)-U(x_2,x_3)-U(x_1,x_3)) \\
&\times (V(\epsilon_1,g(x_2))+V(g(x_1),\epsilon_2))) \\
&+ (2U(x_1,x_3)-U(x_1,x_2)-U(x_2,x_3))(V(\epsilon_1,g(x_3))+V(g(x_1),\epsilon_3))) \\
&+ (2U(x_2,x_3)-U(x_1,x_2)-U(x_1,x_3))(V(\epsilon_2,g(x_3))+V(g(x_2),\epsilon_3))) \\
&+ E[(2U(x_1,X)-U(x_2,X)-U(x_3,X))V(\epsilon_1,g(X))] \\
&+ E[(2U(x_2,X)-U(x_1,X)-U(x_3,X))V(\epsilon_2,g(X))] \\
&+ E[(2U(x_3,X)-U(x_1,X)-U(x_2,X))V(\epsilon_3,g(X))]\} \\
\widetilde{H}_{1,3}(z_1,z_2,z_3) &= \frac{1}{12}\{(2U(x_1,x_2)-U(x_2,x_3)-U(x_1,x_3))V(\epsilon_1,\epsilon_2) \\
&+ (2U(x_1,x_3)-U(x_1,x_2)-U(x_2,x_3))V(\epsilon_1,\epsilon_3) \\
&+ (2U(x_2,x_3)-U(x_1,x_2)-U(x_1,x_3))V(\epsilon_2,\epsilon_3)\}
\end{aligned}
$$

Note that $\widetilde{h}_{1,4}(z_1,z_2,z_3,z_4) = h_1(z_1,z_2,z_3,z_4)$ and $\widetilde{H}_{1,4}(z_1,z_2,z_3,z_4) = H_1(z_1,z_2,z_3,z_4)$. Under the assumptions that $E[|X|^2+|g(X)|^2+|\epsilon|^2] < \infty$, $E[|X-\mu_X|^2(|g(X)|^2 + |\epsilon|^2)] < \infty$, it is guaranteed that $var(h_1(\mathcal{Z},\mathcal{Z}',\mathcal{Z}'',\mathcal{Z}''')) < \infty$ and $var(H_1(\mathcal{Z},\mathcal{Z}',\mathcal{Z}'',\mathcal{Z}''')) < \infty$. Moreover, by the results in Section 5.2.1 (page 182) and Lemma 5.1.5A in Serfling (1980), we have $0 < var(\widetilde{h}_{1,1}(\mathcal{Z})) < \infty$ and obtain

$$
\sqrt{n}U_n^{g,\epsilon} = \frac{4}{\sqrt{n}}\sum_{i=1}^{n}\widetilde{h}_{1,1}(\mathcal{Z}_i) + \mathcal{R}_{1,n}, \tag{A.0.10}
$$

where $\mathcal{R}_{1,n}$ is asymptotically negligible.

Similarly we have $0 < var(\widetilde{H}_{1,2}(\mathcal{Z},\mathcal{Z}')) < \infty$ and obtain

$$
nU_n^{\epsilon} = \frac{6}{(n-1)}\sum_{i\neq j}\widetilde{H}_{1,2}(\mathcal{Z}_i,\mathcal{Z}_j) + \mathcal{R}_{2,n}, \tag{A.0.11}
$$

where $\mathcal{R}_{2,n}$ is asymptotically negligible.

Based on (A.0.10) and (A.0.11), we deduce that

$$\begin{aligned}\sqrt{n}U_n^{g,\epsilon} + nU_n^{\epsilon} &= \frac{4}{\sqrt{n}}\sum_{i=1}^{n}\widetilde{h}_{1,1}(\mathcal{Z}_i) + \frac{6}{(n-1)}\sum_{i\neq j}\widetilde{H}_{1,2}(\mathcal{Z}_i,\mathcal{Z}_j) + \mathcal{R}_n \\ &= \sqrt{n}\mathcal{U}_{n1} + n\mathcal{U}_{n2} + \mathcal{R}_n, \qquad \text{(A.0.12)}\end{aligned}$$

where $\mathcal{R}_n$ is asymptotically negligible and

$$\mathcal{U}_{n1} = \frac{4}{n}\sum_{i=1}^{n}\widetilde{h}_{1,1}(\mathcal{Z}_i),\ \mathcal{U}_{n2} = \frac{6}{n(n-1)}\sum_{i\neq j}\widetilde{H}_{1,2}(\mathcal{Z}_i,\mathcal{Z}_j).$$

Next we shall find the limiting distribution of $\sqrt{n}U_n^{g,\epsilon} + nU_n^{\epsilon}$. Applying Dunford and Schwartz (1963) to $6\widetilde{H}_{1,2}(\mathcal{Z},\mathcal{Z}^{'})$, we obtain

$$6\widetilde{H}_{1,2}(\mathcal{Z},\mathcal{Z}^{'}) = \sum_{k=1}^{\infty}\lambda_k\psi_k(\mathcal{Z})\psi_k(\mathcal{Z}^{'}),$$

where $(\lambda_k,\psi_k(\cdot))$ is a sequence of eigenvalues and eigenfunctions of $6\widetilde{H}_{1,2}$ and the eigenfunctions are orthogonal, i.e., $E[\psi_i(\mathcal{Z})\psi_j(\mathcal{Z})] = \delta_{ij}$.

Note that

$$\begin{aligned}E[6^2\widetilde{H}_{1,2}(\mathcal{Z},\mathcal{Z}^{'})\widetilde{H}_{1,2}(\mathcal{Z},\mathcal{Z}^{''})] &= E[U(X,X^{'})U(X,X^{''})V(\epsilon,\epsilon^{'})V(\epsilon,\epsilon^{''})] = 0 \\ &= E[\sum_k\sum_l\lambda_k\lambda_l\psi_k(\mathcal{Z})\psi_k(\mathcal{Z}^{'})\psi_l(\mathcal{Z})\psi_l(\mathcal{Z}^{''})] \\ &= \sum_k\lambda_k^2E[\psi_k(\mathcal{Z}^{'})]E[\psi_k(\mathcal{Z}^{''})] \\ &= \sum_k\lambda_k^2E[\psi_k(\mathcal{Z}^{'})]^2\end{aligned}$$

which implies that

$$E[\psi_k(\mathcal{Z})] = 0,\ \forall k. \tag{A.0.13}$$

For convenience, we let $\mathcal{U}_{n2}$ be $\frac{6}{n^2}\sum_{i\neq j}\widetilde{H}_{1,2}(\mathcal{Z}_i,\mathcal{Z}_j)$ which will not affect the limiting distribution of $\sqrt{n}U_n^{g,\epsilon}+nU_n^{\epsilon}$. Then the leading term of $\sqrt{n}U_n^{g,\epsilon}+nU_n^{\epsilon}$, which is $\sqrt{n}\mathcal{U}_{n1}+n\mathcal{U}_{n2}$ can be rewritten as

$$\sqrt{n}\mathcal{U}_{n1}+n\mathcal{U}_{n2} \;=\; \frac{4}{\sqrt{n}}\sum_{i=1}^{n}\widetilde{h}_{1,1}(\mathcal{Z}_i)+\sum_{k=1}^{\infty}\lambda_k\left\{(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_k(\mathcal{Z}_i))^2-\frac{1}{n}\sum_{i=1}^{n}\psi_k(\mathcal{Z}_i)^2\right\}.$$

Then we apply multivariate CLT to $\frac{1}{\sqrt{n}}\sum_{i=1}^n\psi_k(\mathcal{Z}_i)$ and $\frac{1}{\sqrt{n}}\sum_i\widetilde{h}_{1,1}(\mathcal{Z}_i)$. Due to (A.0.13) and $E[\widetilde{h}_{1,1}(\mathcal{Z})]=0$, for a fixed positive integer K, we have

$$\begin{pmatrix} \frac{1}{\sqrt{n}}\sum_{i=1}^n\psi_1(\mathcal{Z}_i) \\ \vdots \\ \frac{1}{\sqrt{n}}\sum_{i=1}^n\psi_K(\mathcal{Z}_i) \\ \frac{1}{\sqrt{n}}\sum_{i=1}^n\widetilde{h}_{1,1}(\mathcal{Z}_i)\end{pmatrix} \rightarrow^D Z^*\sim N(\mathbf{0},\Sigma), \tag{A.0.14}$$

where

$$\Sigma=\begin{pmatrix} 1 & \cdots & 0 & E[\psi_1(\mathcal{Z})\widetilde{h}_{1,1}(\mathcal{Z})] \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & 1 & E[\psi_K(\mathcal{Z})\widetilde{h}_{1,1}(\mathcal{Z})] \\ E[\psi_1(\mathcal{Z})\widetilde{h}_{1,1}(\mathcal{Z})] & \cdots & E[\psi_K(\mathcal{Z})\widetilde{h}_{1,1}(\mathcal{Z})] & var(\widetilde{h}_{1,1}(\mathcal{Z}))\end{pmatrix} \tag{A.0.15}$$

We shall use the truncation method to show

$$\sqrt{n}\mathcal{U}_{n1}+n\mathcal{U}_{n2}\rightarrow^D G+\sum_{k=1}^{\infty}\lambda_k(G_k^2-1),$$

where (G_k) are independent standard normal random variables and G is normal random variable with zero mean and variance equal to $16var(\widetilde{h}_{1,1}(\mathcal{Z}))$ which is possibly correlated with (G_k); see (A.0.15).

Define $n\mathcal{U}_{n2}^{(K)} = \frac{1}{n}\sum_{i\neq j}\sum_{k=1}^{K}\lambda_k\psi_k(\mathcal{Z}_i)\psi_k(\mathcal{Z}_j)$ and notice that

$$E[(n\mathcal{U}_{n2} - n\mathcal{U}_{n2}^{(K)})^2] = \frac{1}{n^2}\sum_{i\neq j}\sum_{k=K+1}^{\infty}\lambda_k^2 \to 0, \qquad \text{(A.0.16)}$$

as $K \to +\infty$ due to the fact that $E[6^2\widetilde{H}_{1,2}(\mathcal{Z},\mathcal{Z}')^2] = \sum_{k=1}^{\infty}\lambda_k^2 < \infty$. Then by (A.0.16) and Markov inequality, we can show that for any $\delta > 0$,

$$\lim_{K\to+\infty}\limsup_{n\to\infty} P(|\sqrt{n}\mathcal{U}_{n1} + n\mathcal{U}_{n2} - (\sqrt{n}\mathcal{U}_{n1} + n\mathcal{U}_{n2}^{(K)})| \geq \delta) = 0. \qquad \text{(A.0.17)}$$

Moreover, due to (A.0.14), it is obvious that for any fixed K,

$$\begin{aligned}\sqrt{n}\mathcal{U}_{n1} + n\mathcal{U}_{n2}^{(K)} \quad &= \quad \frac{4}{\sqrt{n}}\sum_{i=1}^{n}\widetilde{h}_1(\mathcal{Z}_i) + \sum_{k=1}^{K}\lambda_k\left\{(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_k(\mathcal{Z}_i))^2 - \frac{1}{n}\sum_{i=1}^{n}\psi_k(\mathcal{Z}_i)^2\right\} \\ &\to^D \quad G + \sum_{k=1}^{K}\lambda_k(G_k^2 - 1). \qquad \text{(A.0.18)}\end{aligned}$$

Since (A.0.17) and (A.0.18) are satisfied, we have the following result by using Theorem 2 in Dehling et al. (2009).

$$\sqrt{n}\mathcal{U}_{n1} + n\mathcal{U}_{n2} \to^D G + \sum_{k=1}^{\infty}\lambda_k(G_k^2 - 1). \qquad \text{(A.0.19)}$$

Therefore, due to (A.0.9), (A.0.12), and (A.0.19), we finally conclude that

$$n\left\{FMDD_n(\frac{g(X)}{\sqrt{n}} + \epsilon|X) - \frac{1}{n}FMDD(g(X)|X)\right\} \to^D G + \sum_{k=1}^{\infty}\lambda_k(G_k^2 - 1).$$

$\diamondsuit$

Proof of Theorem 4.2.3: By Hoeffding decomposition, we have

$$FMDD_n(Y|X) - FMDD(Y|X) = \frac{2}{n}\sum_{i=1}^{n}\{K(Z_i) - FMDD(Y|X)\} + R_n,$$

where R_n is asymptotically negligible. Hence, using Theorem 5.5.1 in Serfling (1980), we obtain

$$\sqrt{n}(FMDD_n(Y|X) - FMDD(Y|X)) \to^D N(0, 4\sigma_1^2),$$

where $\sigma_1^2 = \mathrm{var}(K(Z))$. $\diamondsuit$

Lemma A.4.1. *Let $\{X_i\}_{i\geq 1}$ be a sequence of identically distributed random elements defined on the same probability space $(\Omega, \mathcal{B}, P)$ with $E[|X_1|] < \infty$. Let $Y_n = n^{-1}\max_{1\leq i\leq n}|X_i|$. Then $Y_n \to 0$ almost surely.*

Proof of Lemma A.4.1: For any $\epsilon > 0$, we have

$$\sum_{n=1}^{+\infty} P(|X_n| > \epsilon n) = \sum_{n=1}^{+\infty} P(|X_1| > \epsilon n) < \infty,$$

as $E[|X_1|] < \infty$ [see Lemma 7.5.1 of Resnick (2005)]. By the Borel-Cantelli Lemma, we have $P(\liminf_n[|X_n| \leq \epsilon n]) = 1$. Let $A = \cap_{m=1}^{\infty} \liminf_n[|X_n| \leq n/m]$. Then $P(A) = 1$. For any $w \in A$, there exists $n_0 = n_0(w; m)$ such that for $n \geq n_0(w; m)$, $|X_n| \leq n/m$. Thus we have

$$\begin{aligned} Y_n(w) \leq & n^{-1}\max_{1\leq i\leq n_0-1}|X_i(w)| + n^{-1}\max_{n_0\leq i\leq n}|X_i(w)| \\ \leq & n^{-1}\max_{1\leq i\leq n_0-1}|X_i(w)| + 1/m \to 1/m. \end{aligned}$$

Since m can be arbitrarily large, $\lim_{n\to+\infty} Y_n(w) = 0$, which implies that $Y_n \to 0$ almost surely. $\diamondsuit$

Lemma A.4.2. *If $E[\mathcal{H}(Z,Z')^4] < \infty$ and $\nu_k \neq 0$, then $E[\phi_k(Z)^4] < \infty$, where ν_k is an eigenvalue which corresponds to the kth eigenfunction of $\mathcal{H}$, $\phi_k(\cdot)$.*

Proof of Lemma A.4.2: Note that $\nu_k\phi_k(Z) = E[\mathcal{H}(Z,Z')\phi_k(Z')|Z]$. By the Cauchy-Schwarz inequality and the fact that $E[\phi_k(Z')^2|Z] = E[\phi_k(Z')^2] = 1$, we have

$$\begin{aligned}\nu_k^4 E[\phi_k(Z)^4] =& E[E[\mathcal{H}(Z,Z')\phi_k(Z')|Z]^4]\\ \leq& E[E[\mathcal{H}(Z,Z')^2|Z]^2 E[\phi_k(Z')^2|Z]^2]\\ \leq& E[\mathcal{H}(Z,Z')^4] < \infty.\end{aligned}$$

Thus $E[\phi_k(Z)^4] < \infty$ as $\nu_k \neq 0$. $\diamondsuit$

Proof of Theorem 4.3.1: Let $\mathcal{L}_2(\mu)$ be the space consisting of all square integrable functions with respect to the measure induced by Z (say μ). Let $\mathcal{H}(\cdot,\cdot)$ be a symmetric bivariate function with $E[\mathcal{H}(Z,Z')^2] < \infty$, where Z' is an independent copy of Z. Define the linear operator $(Hf)(s) = \int \mathcal{H}(s,t)f(t)\mu(dt)$ for $f \in \mathcal{L}_2(\mu)$. According to Dunford and Schwartz (1963, p108, Exercise 56), $\mathcal{H}(z,z')$ admits the series decomposition,

$$\mathcal{H}(z,z') = \sum_{k=1}^{\infty} \nu_k\phi_k(z)\phi_k(z'),$$

where $\{\nu_k\}$ and $\{\phi_k\}$ are the eigenvalues and eigenfunctions of H (with respect to μ) respectively, i.e., $H\phi_k = \nu_k\phi_k$ and $E[\phi_i(Z)\phi_j(Z)] = \delta_{ij}$.

Define $\mathcal{H}^{(K)}(Z,Z') = \sum_{k=1}^{K} \nu_k\phi_k(Z)\phi_k(Z^{'})$. As $E[\mathcal{H}(Z,Z')^2] = \sum_{k=1}^{\infty} \nu_k^2 < \infty$, we have

$$\lim_{K\to\infty} E[\{\mathcal{H}(Z,Z^{'})-\mathcal{H}^{(K)}(Z,Z^{'})\}^2]=\lim_{K\to\infty}\sum_{k=K+1}^{\infty}\nu_k^2=0, \tag{A.0.20}$$

which indicates that $\mathcal{H}^{(K)}(Z,Z')$ approximates $\mathcal{H}(Z,Z')$ as $K\to+\infty$. We define $nU_n^*=\frac{1}{n-1}\sum_{i\neq j}\mathcal{H}(Z_i,Z_j)W_i^*W_j^*$ and $nU_n^{(K)*}=\frac{1}{n-1}\sum_{i\neq j}\mathcal{H}^{(K)}(Z_i,Z_j)W_i^*W_j^*$. We first prove that for any $\epsilon>0$,

$$\lim_{K\to+\infty}\limsup_{n\to\infty} P^*(|nU_n^*-nU_n^{(K)*}|>\epsilon)=0 \tag{A.0.21}$$

almost surely. Consider the U-statistic

$$\frac{1}{n(n-1)}\sum_{i\neq j}\{\mathcal{H}(Z_i,Z_j)-\mathcal{H}^{(K)}(Z_i,Z_j)\}^2=\frac{1}{n(n-1)}\sum_{i\neq j}\left\{\sum_{k=K+1}^{\infty}\nu_k\phi_k(Z_i)\phi_k(Z_j)\right\}^2$$

with the kernel $\mathcal{H}^*(Z,Z')=(\mathcal{H}(Z,Z')-\mathcal{H}^{(K)}(Z,Z'))^2$. As $E[\mathcal{H}^*(Z,Z')]=E[(\mathcal{H}(Z,Z')-\mathcal{H}^{(K)}(Z,Z'))^2]=\sum_{k=K+1}^{\infty}\nu_k^2<\infty$, by the strong law of large numbers for U-statistic [see Hoeffding (1961) and Lee (1990)], we obtain

$$\begin{aligned}E^*[(nU_n^*-nU_n^{(K)*})^2]&=E^*\left[\frac{1}{(n-1)^2}\left(\sum_{i\neq j}\sum_{k=K+1}^{\infty}\nu_k\phi_k(Z_i)\phi_k(Z_j)W_i^*W_j^*\right)^2\right]\\&=\frac{1}{(n-1)^2}\sum_{i\neq j}\left(\sum_{k=K+1}^{\infty}\nu_k\phi_k(Z_i)\phi_k(Z_j)\right)^2\\&\to^{a.s.}E\left[\left(\sum_{k=K+1}^{\infty}\nu_k\phi_k(Z)\phi_k(Z^{'})\right)^2\right],\end{aligned}$$

as $n\to+\infty$. Thus (A.0.21) follows from the Markov inequality and (A.0.20).

Next we show that for any fixed K,

$$nU_n^{(K)*} \to^{D^*} \sum_{k=1}^{K} \nu_k (N_k^2 - 1) \; a.s., \tag{A.0.22}$$

where $(N_k) \sim^{i.i.d} N(0,1)$. First, we can rewrite $nU_n^{(K)*}$ as

$$nU_n^{(K)*} = \frac{1}{n} \sum_i \sum_j \left(\sum_{k=1}^{K} \nu_k \phi_k(Z_i) \phi_k(Z_j) W_i^* W_j^* \right) - \frac{1}{n} \sum_i \sum_{k=1}^{K} \nu_k (\phi_k(Z_i) W_i^*)^2 \tag{A.0.23}$$

and for convenience, we let the denominator of $nU_n^{(K)*}$ be n instead of $n-1$. By Lemma A.4.2, $E[\phi_k(Z_i)^4] < \infty$ which implies that $E[\sum_{i=1}^{+\infty} \phi_k(Z_i)^4 / i^2] < \infty$, where $\phi_k(\cdot)$ corresponds to $\nu_k \neq 0$. Define the set

$$\mathcal{A}_k := \left\{ w \in \Omega : \sum_{i=1}^{+\infty} \frac{\phi_k(Z_i(w))^4}{i^2} < \infty \text{ and } \frac{1}{n} \sum_{i=1}^{n} \phi_k(Z_i(w))^b \to E[\phi_k(Z_i)^b] \text{ for } b = 2, 4 \right\}.$$

Then $P(\cap_{k=1}^{(K)} \mathcal{A}_k) = 1$, where $\cap_{k=1}^{(K)}$ is the intersection of indices where eigenvalues $(\nu_k)_{k=1}^{K}$ are nonzero. Conditional on $\{Z_i(w)\}$ with $w \in \cap_{k=1}^{(K)} \mathcal{A}_k$, by Corollary 7.4.1 of Resnick (2005), we have

$$\frac{1}{n} \sum_{i=1}^{n} (W_i^{*2} - 1) \phi_k(Z_i)^2 \to^{a.s.} 0,$$

where $\phi_k(\cdot)$ corresponds to $\nu_k \neq 0$. As $\sum_{i=1}^{n} \phi_k(Z_i)^2 / n \to 1$, we have $\frac{1}{n} \sum_{i=1}^{n} W_i^{*2} \phi_k(Z_i)^2 \to^{a.s.}$ 1, which implies

$$\frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} \nu_k (\phi_k(Z_i) W_i^*)^2 \to^{a.s.} \sum_{k=1}^{K} \nu_k.$$

On the other hand, note that the first term in (A.0.23) can be rewritten as

$$\sum_{k=1}^{K} \nu_k \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} W_i^* \phi_k(Z_i) \right)^2$$

and

$$cov^* \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} W_i^* \phi_s(Z_i), \frac{1}{\sqrt{n}} \sum_{j=1}^{n} W_j^* \phi_t(Z_j) \right) = \frac{1}{n} \sum_{i=1}^{n} \phi_s(Z_i) \phi_t(Z_i)$$

$$\to^{a.s.} E[\phi_s(Z)\phi_t(Z)] = \delta_{st}. \quad \text{(A.0.24)}$$

Similarly, define the set

$$\mathcal{B}_k := \left\{ w \in \Omega : \frac{1}{n} max_{1 \le i \le n} \phi_k(Z_i(w))^2 \to 0 \right\}.$$

By Lemma A.4.1 and $E[\phi_k(Z)^2] < \infty$ for $k = 1, 2, \cdots, K$, we have $P(\cap_{k=1}^{K} \mathcal{B}_k) = 1$ which implies that $P(\cap_{k=1}^{(K)} (\mathcal{A}_k \cap \mathcal{B}_k)) = 1$. Conditional on $\{Z_i(w)\}$ with $w \in \cap_{k=1}^{(K)} (\mathcal{A}_k \cap \mathcal{B}_k)$, we have

$$\frac{\max_{1 \le i \le n} \text{var}^*(W_i^* \phi_k(Z_i))}{\sum_{j=1}^{n} \text{var}^*(W_j^* \phi_k(Z_j))} = \frac{\frac{1}{n} \max_{1 \le i \le n} \phi_k(Z_i)^2}{\frac{1}{n} \sum_{j=1}^{n} \phi_k(Z_j)^2} \to 0. \quad \text{(A.0.25)}$$

By Theorem D.19 in Greene (2007) and the Cramer-Wold device,

$$\left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} W_i^* \phi_{(1)}(Z_i), \ldots, \frac{1}{\sqrt{n}} \sum_{i=1}^{n} W_i^* \phi_{(K)}(Z_i) \right) \to^{D} N(0, I_{(K)}),$$

for almost every realization of $\{Z_i\}$, where $((1), \cdots, (K))$ are indices that correspond to nonzero eigenvalues $(\nu_k)_{k=1}^{K}$, $I_{(K)}$ is the $(K) \times (K)$ identity matrix. Hence, $nU_n^{(K)*} \to^{D^*} \sum_{k=1}^{K} \nu_k (N_k^2 - 1)$ a.s.

Finally, since (A.0.21) and (A.0.22) are both satisfied, we can apply Theorem 2 in Dehling et al. (2009) to conclude that

$$nU_n^* \to^{D^*} \sum_{k=1}^{\infty} \nu_k (N_k^2 - 1) \ a.s.$$

$\diamondsuit$

Proof of Theorem 4.3.2: Recall that $J(Z_i, Z_j) = U(X_i, X_j)V(Y_i, Y_j)$ for $Z_i = (X_i, Y_i)$; see Theorem 4.2.1. We first show that

$$\begin{aligned}
&\operatorname{var}^* \left(\frac{1}{(n-3)} \sum_{i \neq j} (\widetilde{A}_{ij} \widetilde{B}_{ij} - J(Z_i, Z_j)) \eta_i \eta_j \right) \\
=& \frac{1}{(n-3)^2} \sum_{i \neq j} (\widetilde{A}_{ij} \widetilde{B}_{ij} - J(Z_i, Z_j))^2 \to^{a.s.} 0.
\end{aligned}$$

For the ease of notation, write $U_{ij} = U(X_i, X_j)$ and $V_{ij} = V(Y_i, Y_j)$. Notice that

$$\begin{aligned}
\sum_{i \neq j} (\widetilde{A}_{ij} \widetilde{B}_{ij} - U_{ij} V_{ij})^2 =& \sum_{i \neq j} (\widetilde{A}_{ij} \widetilde{B}_{ij} - U_{ij} \widetilde{B}_{ij} + U_{ij} \widetilde{B}_{ij} - U_{ij} V_{ij})^2 \\
\leq& 2 \sum_{i \neq j} (\widetilde{A}_{ij} - U_{ij})^2 \widetilde{B}_{ij}^2 + 2 \sum_{i \neq j} U_{ij}^2 (\widetilde{B}_{ij} - V_{ij})^2 \\
\leq& 4 \sum_{i \neq j} (\widetilde{A}_{ij} - U_{ij})^2 (\widetilde{B}_{ij} - V_{ij})^2 + 2 \sum_{i \neq j} U_{ij}^2 (\widetilde{B}_{ij} - V_{ij})^2 \\
&+ 4 \sum_{i \neq j} (\widetilde{A}_{ij} - U_{ij})^2 V_{ij}^2 \\
\leq& 4 \left(\sum_{i \neq j} (\widetilde{A}_{ij} - U_{ij})^4 \right)^{1/2} \left(\sum_{i \neq j} (\widetilde{B}_{ij} - V_{ij})^4 \right)^{1/2} \\
&+ 2 \left(\sum_{i \neq j} U_{ij}^4 \right)^{1/2} \left(\sum_{i \neq j} (\widetilde{B}_{ij} - V_{ij})^4 \right)^{1/2} \\
&+ 4 \left(\sum_{i \neq j} V_{ij}^4 \right)^{1/2} \left(\sum_{i \neq j} (\widetilde{A}_{ij} - U_{ij})^4 \right)^{1/2}.
\end{aligned}$$

Under the assumption $E[|Y|^8 + |X|^4] < \infty$, we have

$$\frac{1}{n^2}\sum_{i\neq j} U_{ij}^4 \to^{a.s.} EU_{12}^4, \qquad \frac{1}{n^2}\sum_{i\neq j} V_{ij}^4 \to^{a.s.} EV_{12}^4.$$

Thus we only need to show that

$$\frac{1}{n^2}\sum_{i\neq j}(\widetilde{A}_{ij} - U_{ij})^4 \to^{a.s.} 0, \tag{A.0.26}$$

$$\frac{1}{n^2}\sum_{i\neq j}(\widetilde{B}_{ij} - V_{ij})^4 \to^{a.s.} 0. \tag{A.0.27}$$

We only prove (A.0.27) as the proof for the other one is similar. Some algebra shows that

$$\begin{aligned}&\frac{1}{n^2}\sum_{i\neq j}(\widetilde{B}_{ij} - V_{ij})^4 \\ \leq &\frac{C}{n^2}\sum_{i\neq j}\left\{\left(\frac{1}{n}\sum_{l=1}^{n}(b_{il} - E[b_{il}|Y_i])\right)^4 + \left(\frac{1}{n^2}\sum_{k,l=1}^{n}(b_{kl} - E[b_{kl}])\right)^4\right\} + o_{a.s.}(1),\end{aligned}$$

for some constant C. Under the assumption $E[|Y|^8] < \infty$, by the strong law of large numbers for V-statistics, we have

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{n}\sum_{l=1}^{n}(b_{il} - E[b_{il}|Y_i])\right)^4 \\ =&\frac{1}{n^5}\sum_{i=1}^{n}\sum_{l_1,l_2,l_3,l_4=1}^{n}\prod_{k=1}^{4}(b_{il_k} - E[b_{il_k}|Y_i]) \to^{a.s.} 0\end{aligned}$$

due to the fact that $E[\prod_{k=1}^4(b_{il_k} - E[b_{il_k}|Y_i])] = 0$, when (l_1, l_2, l_3, l_4) are distinct indices. Similarly,

$$\frac{1}{n^2}\sum_{i\neq j}\left(\frac{1}{n^2}\sum_{k,l=1}^{n}(b_{kl}-E[b_{kl}])\right)^4 \leq \left(\frac{1}{n^2}\sum_{k,l=1}^{n}(b_{kl}-E[b_{kl}])\right)^4 \to^{a.s.} 0.$$

Therefore, we have

$$\frac{1}{n-3}\sum_{i\neq j}\widetilde{A}_{ij}\widetilde{B}_{ij}\eta_i\eta_j = \frac{1}{n-3}\sum_{i\neq j}J(Z_i,Z_j)\eta_i\eta_j + o_p^*(1)\ a.s., \tag{A.0.28}$$

and the conclusion follows from Theorem 4.3.1. ♢

Remark **A.0.1**. Since (A.0.28) is shown only with the assumption $E[|Y|^8+|X|^4]<\infty$, (A.0.28) is valid under the local alternative with the assumption $E[|\epsilon|^8+|g(X)|^8+|X|^4]<\infty$ and under the fixed alternative with the assumption $E[|Y|^8+|X|^4]<\infty$.

Proof of Theorem 4.3.3: Under the local alternative and the assumption $E[|\epsilon|^8+|g(X)|^8+|X|^4]<\infty$, (A.0.28) remains valid and we further show that

$$\frac{1}{n-3}\sum_{i\neq j}\widetilde{A}_{ij}\widetilde{B}_{ij}\eta_i\eta_j = \frac{1}{n-3}\sum_{i\neq j}U(X_i,X_j)V(\epsilon_i,\epsilon_j)\eta_i\eta_j + o_p^*(1)\ a.s.$$

Then we are left to show

$$\frac{1}{n-3}\sum_{i\neq j}J(Z_i,Z_j)\eta_i\eta_j = \frac{1}{n-3}\sum_{i\neq j}U(X_i,X_j)V(\epsilon_i,\epsilon_j)\eta_i\eta_j + o_p^*(1)\ a.s.$$

Similar to the proof of Theorem 4.3.2, let's consider

$$\begin{aligned}
& var^*\left(\frac{1}{n-3}\sum_{i\neq j}\{J(Z_i,Z_j)-U(X_i,X_j)V(\epsilon_i,\epsilon_j)\}\eta_i\eta_j\right) \\
= \quad & \frac{1}{(n-3)^2}\sum_{i\neq j}U(X_i,X_j)^2 \\
\times \quad & \{\frac{1}{n^{2a}}V(g(X_i),g(X_j))+\frac{1}{n^a}(V(g(X_i)+\epsilon_i,g(X_j)+\epsilon_j) \\
& -V(g(X_i),g(X_j))-V(\epsilon_i,\epsilon_j))\}^2 \\
\leq \quad & O(n^{-2(1+2a)})\left(\sum_{i\neq j}U(X_i,X_j)^4\right)^{1/2}\left(\sum_{i\neq j}V(g(X_i),g(X_j))^4\right)^{1/2} \\
+ \quad & O(n^{-2(1+a)})\left(\sum_{i\neq j}U(X_i,X_j)^4\right)^{1/2}\left(\sum_{i\neq j}V(g(X_i)+\epsilon_i,g(X_j)+\epsilon_j)^4\right)^{1/2} \\
+ \quad & O(n^{-2(1+a)})\left(\sum_{i\neq j}U(X_i,X_j)^4\right)^{1/2}\left(\sum_{i\neq j}V(g(X_i),g(X_j))^4\right)^{1/2} \\
+ \quad & O(n^{-2(1+a)})\left(\sum_{i\neq j}U(X_i,X_j)^4\right)^{1/2}\left(\sum_{i\neq j}V(\epsilon_i,\epsilon_j)^4\right)^{1/2} \\
\longrightarrow^{a.s.} \quad & 0 \qquad \text{(A.0.29)}
\end{aligned}$$

Here (A.0.29) is due to the fact that

$$\begin{aligned}
\frac{1}{n^2}\sum_{i\neq j}U(X_i,X_j)^4 \quad \rightarrow^{a.s.} \quad & E[U(X,X^{'})^4],\ \frac{1}{n^2}\sum_{i\neq j}V(g(X_i)+\epsilon_i,g(X_j)+\epsilon_j)^4 \\
\rightarrow^{a.s.} \quad & E[V(g(X)+\epsilon,g(X^{'})+\epsilon^{'})^4]
\end{aligned}$$

$$\frac{1}{n^2}\sum_{i\neq j}V(g(X_i),g(X_j))^4 \rightarrow^{a.s.} E[V(g(X),g(X^{'}))^4],\ \frac{1}{n^2}\sum_{i\neq j}V(\epsilon_i,\epsilon_j)^4 \rightarrow^{a.s.} E[V(\epsilon,\epsilon^{'})^4],$$

since $E[|g(X)|^4+|\epsilon|^4+|X|^4]<\infty$.

Thus, under the local alternative, we have

$$\frac{1}{n-3}\sum_{i\neq j}\widetilde{A}_{ij}\widetilde{B}_{ij}\eta_i\eta_j = \frac{1}{n-3}\sum_{i\neq j}U(X_i,X_j)V(\epsilon_i,\epsilon_j)\eta_i\eta_j + o_p^*(1)\ a.s.$$

and applying Theorem 4.3.1 to $\frac{1}{n-3}\sum_{i\neq j}U(X_i,X_j)V(\epsilon_i,\epsilon_j)\eta_i\eta_j$, we have

$$T_n^* \to^{D^*} \mathcal{G}_0\ a.s.$$

Similarly, by (A.0.28) and applying Theorem 4.3.1 to $\frac{1}{n-3}\sum_{i\neq j}U(X_i,X_j)V(Y_i,Y_j)\eta_i\eta_j$, under the fixed alternative and the same assumptions in Theorem 4.3.2, we have

$$T_n^* \to^{D^*} \widetilde{\mathcal{G}}_0 := \sum_{k=1}^{\infty}\widetilde{\lambda}_k(\widetilde{G}_k^2-1)\ a.s.,$$

where $(\widetilde{\lambda}_k)$ is a sequence of eigenvalues corresponding to orthonormal eigenfunctions of J under the fixed alternative and $(\widetilde{G}_k)$ is a sequence of zero mean, unit variance Gaussian random variables which are mutually independent.
Note that under the fixed alternative, $FMDD(Y|X)$ is a positive integer which implies that $nFMDD_n(Y|X) \to^{a.s.} +\infty$.

Furthermore, under the local and fixed alternatives, we can show that

$$Q^*_{(1-\alpha),n} \to^p Q_{(1-\alpha),\mathcal{G}_0} \text{ and } Q^*_{(1-\alpha),n} \to^p Q_{(1-\alpha),\widetilde{\mathcal{G}}_0}, \tag{A.0.30}$$

respectively, where $Q_{(1-\alpha),\widetilde{\mathcal{G}}_0}$ is the $(1-\alpha)$th quantile of $\widetilde{\mathcal{G}}_0$. Here (A.0.30) is shown by using (ii) of Lemma 11.2.1 in Lehmann and Romano (2005) and the fact that $\mathcal{G}_0$, $\widetilde{\mathcal{G}}_0$ are continuous random variables which can be shown under $\sum_{k=1}\lambda_k^2 \neq 0$,

$\sum_{k=1} \widetilde{\lambda}_k^2 \neq 0$ and these are implied by the assumptions in $H_{1,n}$, H_1. Finally, the conclusions follow from Theorems 4.2.2 and 4.2.3.

$\diamondsuit$