Martingale Difference Divergence Matrix and Its Application to Dimension Reduction for Stationary Multivariate Time Series

Chung Eun Lee and Xiaofeng Shao

Department of Statistics, University of Illinois at Urbana Champaign, Champaign, IL

ABSTRACT

In this article, 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 the so-called martingale difference divergence matrix (MDDM), which can quantify the conditional mean independence of \( V \in \mathbb{R}^p \) given \( U \in \mathbb{R}^q \) and also encodes the number and form of linear combinations of \( V \) that are conditionally 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. Supplementary materials for this article are available online.

1. Introduction

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 by Box and Tiao (1977), the factor models by Peña and Box (1987), the scalar component analysis by Tiao and Tsay (1989), the independent component analysis by Back and Weigend (1997), the principal component analysis by Stock and Watson (2002), and the dynamic orthogonal component analysis by 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), and Eichler, Motta, and von Sachs (2011), among others.

In this article, 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 \( Y_t \in \mathbb{R}^p \), \( t \in \mathbb{Z} \) be a mean zero \( p \)-variate stationary time series, then the optimal predictor of \( Y_{n+1} \) given the past information set \( \mathcal{F}_n = \sigma(Y_{-n}, \ldots, Y_t) \) 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, \( M \in \mathbb{R}^{p \times p} \), such that \( MY_t = Z_t = [Z_{1t}, Z_{2t}]' \), where \( Z_{1t} \in \mathbb{R}^q \) and \( Z_{2t} \in \mathbb{R}^{p-q} \), 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, and Reichlin (2000, 2005), Bai and Ng (2002), Bai (2003), Stock and Watson (2005), Pan and Yao (2008), Lam, Yao, and Bathia (2011), and Lam and Yao (2012), among others. A distinction from the static factor models in the existing literature is that our error process, that is, \( \varepsilon_t = Y_t - E(Y_t | \mathcal{F}_{t-1}) \) 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...
provides a review of martingale difference divergence and its ear dependence. Since non-Gaussian and nonlinear time series linear analog of cumulative MDDM, which only measures linear dependence of the past, we perform the eigen-decomposition for the MDDM at several lags to account for the underlying time series structure, either jointly or in a pairwise fashion. 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) and Lam and Yao (2012). Note that the inference in the latter work is based on a linear analog of cumulative MDDM, which only measures linear dependence. Since non-Gaussian 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 article is organized as follows. Section 2 provides a review of martingale difference divergence and its sample estimate. In Section 3, we introduce the definition of martingale difference divergence matrix and its properties. Our dimension reduction methodology for conditional mean is presented in Section 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 5. Section 6 presents two real data illustrations and Section 7 concludes. Technical details are relegated to supplementary material.

A word on notation. Let $i = \sqrt{-1}$ be the imaginary unit. For $x \in \mathbb{C}^p$, we use $x^*$ for "x-conjugate-transpose" (conjugate for scalars). The scalar product of vectors $x$ and $y$ is denoted by $\langle x, y \rangle$. 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 $x = (x_1, \ldots, x_p) \in \mathbb{C}^p$ as $|x|^p = |x^p| = x_1^2 + \cdots + x_p^2$, and if $x = (x_1, \ldots, x_p) \in \mathbb{R}^p$, it is sometimes denoted as $|\|x\|_2 = \sqrt{x_1^2 + \cdots + x_p^2}$. For a square matrix $A$, spectral norm of $A$ is defined as $\|A\|_2 = \sqrt{\lambda_{\max}(A^T A)}$ and Frobenius norm of $A$ is denoted by $\|A\|_F$, where $\|A\|_2 = \sqrt{tr(A^T A)}$ and $tr(A) = \sum_{i=1}^p A_{i,i}$. A random vector $X \in \mathbb{L}^q$ if $E|X|^q < \infty$.

2. Review of Martingale Difference Divergence

For $U \in \mathbb{R}^q$ and $V \in \mathbb{R}$, 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 $U$, that is,

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

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

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

where $g_{V,U}(s) = E(V e^{\mu,U \cdot s})$, $g_V = E(V)$, $g_U(s) = E(e^{\mu,U \cdot s})$, and $c_q = \pi^{(1+q)/2}/\Gamma((1+q)/2)$. The definition can be regarded as an extension of distance covariance (Székely, 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 (1) holds. Furthermore, if $E(|V|^2 + |U|^2_q) < \infty$, then

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

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 $V \in \mathbb{R}^p$, $p \geq 1$, then the characteristic function-based definition (2) still applies. Under the assumption that $E(|V|^p_q + |U|^2_q) < \infty$, Park et al. showed

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

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 $\overline{V}_n = \frac{1}{n} \sum_{k=1}^n V_k$, $a_{kl} = V_k V_l$, $\overline{a}_k = \frac{1}{n} \sum_{l=1}^n a_{kl} = V_k \overline{V}_n$, and $\overline{A} = \overline{a}_k - \overline{V}_n\overline{V}_n$. Then

$$\text{MDD}_n(V|U)^2 = -\frac{1}{n^2} \sum_{k,l=1}^n A_{kl} B_{kl},$$

where $g_{V,U}(s) = \frac{1}{n} \sum_{k=1}^n V_k e^{\mu,U \cdot s}$, $g_V = \frac{1}{n} \sum_{k=1}^n V_k$, and $g_U(s) = \frac{1}{n} \sum_{k=1}^n e^{\mu,U \cdot s}$. 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

$$\text{MDD}_n(V|U)^2 = -\frac{1}{n^2} \sum_{k,l=1}^n (V_k - \overline{V}_n)^T (V_l - \overline{V}_n) B_{kl}.$$
It turns out that the above definition can be further simplified as

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

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 (Székely and Rizzo 2014; Park, Shao, and Yao 2015) to define an unbiased estimator, but it unfortunately complicates the asymptotic analysis in Section 4.

**3. Martingale Difference Divergence Matrix**

For $U \in \mathbb{R}^p$ and $V \in \mathbb{R}^q$, it is possible that there exists a linear combination of $V$, say $\alpha \in \mathbb{R}^p$, such that $E(\alpha^T V|U) = E(\alpha^T V)$ even though $V$ 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 1** Martingale Difference Divergence Matrix.

Given $V = (V_1, \ldots, V_p)^T \in \mathbb{R}^p$, $U \in \mathbb{R}^q$,

$$\text{MDDM}(V|U) = \frac{1}{c_q} \int_{\mathbb{R}^q} G(s)G(s)^T ds,$$

where $G(s) = \text{cov}(V, e^{s^T U}) = (G_1(s), \ldots, G_p(s))^T$ for $s \in \mathbb{R}^q$, $G_j(s) = \text{cov}(V_j, e^{s^T U})$.

Note that the $(i, j)$th entry of the $p \times p$ matrix MDDM equals to MDDM(V|U)$^2$, whereas the $(i, i)$th entry is

$$\text{MDDM}(V|U)_{ij} = \frac{1}{c_q} \int_{\mathbb{R}^q} G_i(s)G_j(s)^T ds = \text{MDDM}(V|U)^2_{ij},$$

that is, 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 1.** If $E(|V_j|^2 + |U|^2) < \infty$, then

$$\text{MDDM}(V|U) = -E[(V - E(V))(V - E(V))^T (U - U |_q)],$$

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

**Lemma 1** implies that MDDM(V|U) is a $p \times p$ matrix with the $(i, j)$th entry equal to MDDM$_n(V|U)_{ij} = -E[(V_i - E(V_i))(V_j - E(V_j))^T (U_i - U_j |_q)]$, provided that $E(|V_j|^2 + |U|^2) < \infty$. Since $G_i(s) = 0, \forall s \Leftrightarrow E(V_i|U) = E(V_i)$, we have that MDDM$_n(V|U)_{ij} = 0$, provided that $E(V_i|U) = E(V_i)$ or $E(V_i|U) = E(V_i)$. It is also worth noting that $tr(\text{MDDM}(V|U)) = \text{MDDM}(V|U)^2$ in (3).

By elementary matrix algebra, it is not difficult to show that Lemma 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 1.** For $V \in \mathbb{R}^p$ and $U \in \mathbb{R}^q$, if $E(|V_j|^2 + |U|^2) < \infty$, 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 1.** We shall provide a discussion on a possible analog of MDDM to measure the linear dependence between two vectors $V \in \mathbb{R}^p$ and $U \in \mathbb{R}^q$. Define

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

where $\text{cov}(V, U)$ is a $p \times q$ matrix with the $(i, j)$th entry being $\text{cov}(V_i, U_j)$. It is easy to show that $L(V|U)$ is a real, symmetric, and positive semidefinite matrix. Then there exists a nonzero $\alpha \in \mathbb{R}^p$, such that $\text{cov}(\alpha^T V, U) = 0$ (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 rank(MDDM(V|U)) $\geq$ rank(L(V|U)).

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

$$\text{MDDM}_n(V|U) = -\frac{1}{n^2} \sum_{k,l=1}^{n} (V_k - \overline{V}_n)(V_l - \overline{V}_n)^T |U_k - U_l|_q.$$
Definition 2. The cumulative MDDM is defined as
\[ \Gamma_{k_0} = \Gamma_{k_0}^{(1)} := \sum_{j=1}^{k_0} \text{MDDM}(Y_t | Y_{t-j}). \]  

Since MDDM is a positive semidefinite matrix, \( \Gamma_{k_0} \) 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 \( \Gamma_{k_0} \) is given by \( \Gamma_{k_0} = \sum_{j=1}^{k_0} \text{MDDM}_n(Y_t | Y_{t-j}) \).

Remark 2. Our definition follows the common practice in time series analysis, where the cumulative contribution from various lags is added up in a pairwise fashion. Alternatively, we could adopt a joint approach, that is, we can define
\[ \Gamma_{k_0}^{(2)} = \text{MDDM}(Y_t^1(Y_t^{t-1} \cdots Y_t^{t-k_0})^T) \]
and its sample estimate as \( \Gamma_{k_0}^{(2)} = \text{MDDM}_n(Y_t^1(Y_t^{t-1} \cdots Y_t^{t-k_0})^T) \). For a given \( k_0 \), it seems difficult to know whether inference based on \( \Gamma_{k_0}^{(2)} \) is preferred for the given series at hand. We shall compare the finite sample performance for inferences based on \( \Gamma_{k_0}^{(j)}, j = 1, 2 \) in simulation studies.

4.1. Principal Conditional Mean Components (PCMC)

As outlined in Section 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 \( \Gamma_{k_0} \) (or \( \Gamma_{k_0}^{(2)} \)); see Theorem 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 \( \Gamma_{k_0} \) 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 \( \gamma_j \) be an orthonormal eigenvector of \( \Gamma_{k_0} \) corresponding to the eigenvalue \( \lambda_j \). Then we have
\[ \gamma_j^T \Gamma_{k_0} \gamma_j = \lambda_j, \quad j = 1, \ldots, p. \]

If we let \( M = (\gamma_1, \ldots, \gamma_p) \) and \( \Lambda = \text{diag}(\lambda_1 \geq \cdots \geq \lambda_p) \), then \( \Gamma_{k_0} M = \Lambda M \) by spectral decomposition of \( \Gamma_{k_0} \). Therefore, the rank of \( \Gamma_{k_0} \) is \( s \), which means that there exist \( p - s \) linearly independent combinations \( \{\gamma_{s+1}, \ldots, \gamma_p\} \). Make MDDM(\( Y_t Y_{t-s}^\top \)) \( = 2 \), \( j = 1, \ldots, k_0 \), \( i = s+1, \ldots, p \). Since all these linear combinations live in the null space of \( \Gamma_{k_0} \), they can be readily estimated based on eigen-decomposition of \( \Gamma_{k_0} \).

Let \( \hat{\gamma}_j, \hat{\gamma}_{j+1}, \ldots, \hat{\gamma}_p \) be the \( p \) pairs of eigenvalues and eigenvectors of \( \hat{\Gamma}_{k_0} \), where \( \hat{\lambda}_1 \hat{\geq} \hat{\lambda}_2 \hat{\geq} \cdots \hat{\geq} \hat{\lambda}_p \) and the eigenvectors \( \{\hat{\gamma}_j\}_{j=1}^p \) are orthonormal. To estimate \( s \), the rank of \( \hat{\Gamma}_{k_0} \), we adopt a ratio-based estimator following the practice of Lam, Yao, and Bathia (2011), Lam and Yao (2012), that is,
\[ \hat{s} = \arg \min_{1 \leq i \leq R} \frac{\hat{\lambda}_{i+1}}{\hat{\lambda}_i}, \]
where \( R \) is an integer satisfying \( s \leq R < p \). There are other methods developed for the estimation of \( s \), see, for example, Bai and Ng (2002, 2007) and Hallin and Liska (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 by 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 by Lam, Yao, and Bathia (2011) and Lam and Yao (2012); see Section 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 \( j = 1, \ldots, p \), we can then estimate \( \gamma_j \) by \( \hat{\gamma}_j \). Since \( \gamma_j \) might be replaced by \( -\hat{\gamma}_j \), 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 \( Y_t \in \mathbb{N} \) be a strictly stationary and \( \beta \)-mixing process. Assume that there exist \( \delta > 0 \) such that \( E[|Y_t|^6 + \delta] < \infty \).

For a \( \delta' \in (0, \delta) \) : \( \beta(n) = O(n^{-\delta'/2}) \).

(C2') Let \( Y_t \in \mathbb{N} \) be a strictly stationary and \( m \)-dependent process. Assume that \( E[|Y_t|^6] < \infty \).

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

Theorem 2 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.

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 an \( r \times 1 \) latent factor time series, which is 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
Once an estimator of the analysis of sample estimate of the positive semidefinite matrix implies that certain linear combinations of responding to basically a linear counterpart of our cumulative MDDM. Following the practice in Pan and Yao (2008), Lam, Yao, and Bathia (2011), and Lam and Yao (2012), we assume that

**Assumption 1.** \( A^T A = I_r, A^T B = 0, \) and \( B^T B = I_{p-r} \).

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 \( \mathcal{M}(A) \) 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 \( \hat{A} \) is obtained, it is natural to estimate the factor process \( X_t \) by \( \hat{X}_t = \hat{A}^T Y_t \) and the residuals are \( \hat{e}_t = (I_p - \hat{A}^T) Y_t \). The next step is to fit a parsimonious model to \( \hat{X}_t \), which may be achieved by rotating \( \hat{X}_t \) appropriately, or equivalently modeling \( H^T \hat{X}_t \), where \( H \) is an \( r \times r \) orthogonal matrix.

Based on the above model assumptions, we derive that

\[
B^T Y_t = B^T A X_t + B^T \epsilon_t = B^T \epsilon_t \tag{7}
\]

\[
A^T Y_t = A^T A X_t + A^T \epsilon_t = X_t + A^T \epsilon_t, \tag{8}
\]

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 \( \epsilon_t \) 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 \( k = 1, 2, \ldots \), thus the columns of \( B \) are orthonormal eigenvectors of \( \text{cov}(Y_{t+k}, Y_t) \) corresponding to zero eigenvalues. This observation led them to focus on the following matrix:

\[
\mathcal{L}_{k_0} = \sum_{j=1}^{k_0} \text{cov}(Y_{t+j}, Y_t) \text{cov}(Y_{t+j}, Y_t)^T = \sum_{j=1}^{k_0} L(Y_t|Y_{t-j}) \tag{9}
\]

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 \( \Gamma_{k_0} \); see Remark 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 \( Y_{t-1}, \ldots, Y_{t-k_0} \) in a pairwise fashion, whereas \( \Gamma_{k_0} \) encodes the number and form of linear combinations of \( Y_t \) that are conditionally mean independent of \( Y_{t-1}, \ldots, Y_{t-k_0} \) in a pairwise fashion.

In time series analysis, \( \text{cov}(Y_{t+j}, Y_t) \) is used to measure the linear dependence at lag \( j \), whereas MDDDM(\( Y_{t+j}|Y_{t} \)) 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)

\[
Y_t = E(Y_t|\mathcal{F}_{t-1}) + \eta_t, \tag{10}
\]

where \( \eta_t = Y_t - E(Y_t|\mathcal{F}_{t-1}) \) is a martingale difference sequence by construction. Assume that \( E(Y_t|\mathcal{F}_{t-1}) = A Z_t \), where \( A \) 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 \( \mathcal{M}(A) \) is uniquely defined. Note that \( \mathcal{M}(A) \) is the column space of \( M = (\gamma_1, \ldots, \gamma_s) \). For the convenience of discussion, we assume that there is a \( p \times (p-s) \) matrix \( B \), such that

**Assumption 2.** \( A^T A = I_r, A^T B = 0, \) and \( B^T B = I_{p-r} \).

Following the argument in the derivation of (7), we have that

\[
B^T Y_t = B^T E(Y_t|\mathcal{F}_{t-1}) + B^T \eta_t
= B^T A Z_t + B^T \eta_t = B^T \eta_t, \tag{11}
\]

\[
A^T Y_t = A^T E(Y_t|\mathcal{F}_{t-1}) + A^T \eta_t
= A^T A Z_t + A^T \eta_t = Z_t + A^T \eta_t, \tag{12}
\]

where \( \{B^T \eta_t\} \) and \( \{A^T \eta_t\} \) are martingale difference sequences. Based on (12), it is easy to see that \( Z_t = E[A^T Y_t|\mathcal{F}_{t-1}] \). Suppose we can obtain good estimates of \( A \) and \( B \) (or the corresponding column spaces), say \( \hat{A} \) and \( \hat{B} \), then we can estimate \( \hat{Z}_t \) by \( \hat{Z}_t = \hat{A}^T Y_t \) and the residuals are \( \hat{\eta}_t = (I_p - \hat{A}^T \hat{A}) Y_t \). A lower dimensional model can be fitted to \( \{\hat{Z}_t\} \) 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 \( A Z_t \) 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 \( A X_t \) 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., \( B \)) are sought to make \( B^T Y_t \) a martingale difference sequence; compare (7) and (11). 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 \( (X_t, \epsilon_t) \), and the inference is based on the cumulative linear matrix \( \Gamma_{k_0} \), which encodes the linear dependence, whereas under the FM-MDE, we naturally focus on cumulative MDDM \( \Gamma_{k_0} \), which characterizes conditional mean independence. To shed some light on the difference, we consider the following state space model.

**Example 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 \( W_t = h(\epsilon_{2t}, \epsilon_{2(t-1)}, \ldots) \), \( t \in \mathbb{Z} \) be an \( r \)-dimensional nonlinear stationary causal process, where \( \{\epsilon_{2t}\}_{t \in \mathbb{Z}} \) is the \( r \)-dimensional mean zero iid innovation process that is independent of the \( p \)-dimensional error process \( (\epsilon_{1t})_{t \in \mathbb{Z}} \). 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 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 \( \mathcal{M}(A) = \mathcal{M}(A) \), that is, the two factor loading spaces are identical. It would be interesting to see which inference method (i.e., the one based on \( \mathcal{L}_k \) vs. the one based on \( \Gamma_{k_0} \)) 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 \( Y_t \), which admits both representations (i.e., FM-WNE and FM-MDE), the two could coincide as demonstrated in the following proposition.

**Proposition 1.** Suppose that Assumptions (1) and (2) hold. If \( \{\epsilon_t, \mathcal{F}_t\} \) is a martingale difference sequence and \( \mathcal{M}(A) = \mathcal{M}(A) \), then we have \( \mathcal{M}(A) = \mathcal{M}(A) \).

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

**Example 2.** Let \( Y_t = A_1W_{1t} + [A_1, A_2]W_{2t} + \epsilon_{1t} \), 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 an \( r \)-dimensional stationary causal process as defined in Example 1 and is not a white-noise sequence, \( W_{2t} \) is a \( q \)-dimensional vector white-noise sequence but not martingale difference sequence, and \( \epsilon_{1t} \) and \( \epsilon_{2t} \) are iid mean zero. (iii), The three processes \( \{W_{1t}\}, \{W_{2t}\} \), and \( \{\epsilon_{1t}\} \) are mutually independent.

Let \( W_{2t} = (W_{2t}^1, W_{2t}^2, \ldots, W_{2t}^q)^T \), where \( W_{2t}^j \) is of dimension \( r \) and \( W_{2t}^q \) is of dimension \( q-r \). Then the model can be reformulated as

\[
Y_t = [A_1, A_2] \left( \begin{array}{c} W_{2t}^1 \quad W_{2t}^2 \quad \ldots \quad W_{2t}^q \end{array} \right) + \epsilon_{1t}.
\]

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 \( \mathcal{M}(A) = \mathcal{M}(A) \), whereas under the framework of FM-MDE, \( E(Y_t | \mathcal{F}_{t-1}) = A_1E(W_{1t} | \mathcal{F}_{t-1}) + [A_1, A_2]E(W_{2t} | \mathcal{F}_{t-1}) \). Then \( s > q > r \) and \( \mathcal{M}(A) = \mathcal{M}([A_1, A_2]) \); see Table 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 6.4.

### 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, Sriram, and Yin 2010) and central mean subspace (Park, Sriram, and Yin 2009) of the conditional distribution of \( Y_t \) given \( (Y_{t-1}, \ldots, Y_{t-d}) \), 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, Sriram, and Yin (2009, 2010) mainly focuses on the dimension reduction of covariates, which are naturally defined as the lagged observations \( (Y_{t-1}, \ldots, Y_{t-d}) \) 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. target 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 to spirit to the recently developed envelop models by Cook, Li, and Chiaromonte (2010), 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, Sriram, and Yin (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

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

that is, given \( (Y_{t-1}, \ldots, Y_{t-k_0}) \), \( Y_t \) is conditionally mean independent of \( (Y_{t-k_0-1}, Y_{t-k_0-2}, \ldots) \). Thus, the determination of \( k_0 \) itself is a nontrivial task. If the series follows a vector autoregressive model, then partial autocorrelation function (ACF) provides an indication about the magnitude of \( k_0 \). Alternatively, we can first look at the lag \( j \) sample MDD, that is, \( \text{MDD}_j(Y_t | Y_{t-j})^2 \), 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 \( k_0 \) be an integer multiple of the period (see Section 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.

---

**Table 1.** Factor model representations for the state space model.

| Representation | FM-WNE | FM-MDE |
|---------------|--------|--------|
| \( A = D_A \) | \( X_t = W_t \) | \( Z_t = E(W_t | \mathcal{F}_{t-1}) \) |
| \( \epsilon_t = \epsilon_{1t} \) | \( \eta_t = Y_t - D_tE(W_t | \mathcal{F}_{t-1}) \) |
| Example 1     |        |        |
| \( A = A_1 \) | \( X_t = W_t \) | \( \epsilon_t = [A_1; A_2]W_{2t} + \epsilon_{1t} \) |
| \( \eta_t = Y_t - A_tE(W_t | \mathcal{F}_{t-1}) \) | \( [A_1, A_2]E(W_{2t} | \mathcal{F}_{t-1}) \) |
| Example 2     |        |        |

---

[Note: The table content reflects the text, but for the sake of this demonstration, the actual table content is not transcribed here.]
An additional practical and methodological issue is that after we obtain the rank of $\Gamma_{k}$, say by $\hat{s}$, and the null space of $\Gamma_{k}$, by $M_{1} = [\hat{\gamma}_{T+1}, \ldots, \hat{\gamma}_{p}]$, it would be useful to verify that the transformed series $M_{1}Y_{t}$ are indeed conditionally mean independent of $F_{t-1}$ (or $F_{t-1}, t \neq k_{0}$ in a pairwise fashion). To this end, one can look at the magnitude of $H_{n} = \sum_{j=1}^{k_{0}} \text{MDDM}^{2}_{j}(M_{1}Y_{t} | Y_{t-j})$ 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.

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 (9)). In our simulations, we tried $\Gamma(1), j = 1, 2$ for several $k_{0}$’s 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 finite methods, two steps are involved in the estimation. The first step corresponds to the estimation of the true number of factors, that is, $r$ or $s$ using ratio-based estimator (see (6)), 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, that is, $\mathcal{M}(A)$ (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 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** ($D(A, \cdot, \cdot)$) (Pan and Yao 2008)
  \[
  D_{1}(A, \hat{A}) = \frac{[\text{tr}(\hat{A}^{T}(I_p - AA^{T})\hat{A}) + \text{tr}(\hat{B}^{T}AA^{T}\hat{B})]}{p^{1/2}},
  \]
  where $\hat{B}$ is a basis of an orthogonal complement of the column space spanned by $\hat{A}$. $D_{1}(A, \hat{A})$ is used to measure the distance between $\mathcal{M}(A)$ and $\mathcal{M}(\hat{A})$. Note that under Assumption 2, $AA^{T}$ is a projection matrix onto the linear space $\mathcal{M}(A)$ and $D_{1}(A, \hat{A}) \in [0, 1]$. $D_{1}(A, \hat{A}) = 0$ if and only if $\mathcal{M}(A) = \mathcal{M}(\hat{A})$, and $D_{1}(A, \hat{A}) = 1$ if and only if $\mathcal{M}(A) = \mathcal{M}(\hat{B})$.

- **Root mean squared error (RMSE)** (Lam, Yao, and Bathia 2011)
  \[
  \text{RMSE} = \left(\frac{1}{n} \sum_{t=1}^{n} \frac{||\hat{X}_{t} - AX_{t}||^{2}}{pn}\right)^{1/2},
  \]
  which measures the overall closeness of the estimated signal $\hat{AX}_{t}$ to the true signal $AX_{t}$. Smaller value of RMSE indicates more accurate estimation of underlying factor series.

We shall investigate the following examples.

**Example 3.** Example 3 is adopted from the simulation study of Pan and Yao (2008) with slight modification so the model falls into the framework of Example 1. We define the factor series $X_{t} = (x_{1,t}, x_{2,t}, x_{3,t})^{T}$ as
\[
\begin{align*}
x_{1,t} & = 0.8x_{1,t-1} + e_{1,t}, \\
x_{2,t} & = e_{2,t} + 0.9e_{2,t-1} + 0.3e_{2,t-2}, \\
x_{3,t} & = -0.5x_{3,t-1} - e_{3,t} + 0.8e_{3,t-1},
\end{align*}
\]
where $e_{1,t}, i = 1, 2, 3$ are all iid standard normal variables. The observed data $Y_{t} = (Y_{1,t}, \ldots, Y_{p,t})^{T}$ are defined by
\[
Y_{i,t} = \begin{cases} 
  x_{i,t} + e_{i,t} & \text{for } i = 1, 2, 3 \\
  e_{i,t} & \text{for } i = 4, \ldots, p,
\end{cases}
\]
where $e_{i,t}, i = 1, 2, \ldots, p$ are iid standard normal variables and independent from $\{e_{j,t}\}, j = 1, 2, 3$. We consider several different combinations of $(p, n, k_{0})$, that is, $p = 5, 10, 20$, $n = 300, 600, 1000$, 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, $A = A = (I_{3}, 0_{p-3})^{T}$. Note that when $k_{0} = 1$, $\Gamma(k_{0})$ and $\Gamma^{(2)}(k_{0})$ become the same so some results are identical in Table 2.

It appears from Table 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 $\Gamma(k_{0})$ or $\Gamma^{(2)}(k_{0})$, the performance is quite stable with respect to $k_{0}$.

**Example 4.** In this example, the linear ARMA model for $X_{t}$ in Example 3 is replaced by a nonlinear model, where $X_{t} = (x_{1,t}, x_{2,t}, x_{3,t})^{T}$ is defined as
\[
\begin{align*}
x_{1,t} & = -(0.9e^{-0.2x_{1,t-1}^{2}})x_{1,t-1} + e_{1,t}, \\
x_{2,t} & = (0.5e^{-0.4x_{2,t-1}^{2}} + 0.4)x_{2,t-1} + e_{2,t}, \\
x_{3,t} & = (0.1e^{-x_{3,t-1}^{2}} + 0.7)x_{3,t-1} + e_{3,t},
\end{align*}
\]
Then the data $Y_{t} = AX_{t} + e_{t}$, where $A = (I_{3}, 0_{p-3})^{T}$, $e_{1,t}, e_{2,t}$ are iid standard normal and independent from each other. Like Example 3, 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 Wei (2007), $X_{t}$ admits a stationary solution and the model falls into the framework in Example 1. For this example, the true number of factors $r$ and $s$ are still 3 and $\mathcal{M}(A) = \mathcal{M}(A)$. From Table 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 3.
Table 2. Mean, standard error (in the bracket) of $D$-distance and $\hat{r}$, $x^*$ with 1000 replicates for Example 3.

| $i$ | $\hat{r}$ | $\hat{r} < 3$ | $\hat{r} = 3$ | $\hat{r} > 3$ | $\hat{r}$ | $\hat{r} < 3$ | $\hat{r} = 3$ | $\hat{r} > 3$ |
|-----|------------|---------------|---------------|---------------|------------|---------------|---------------|---------------|
| $p = 5$, $n = 300$ | | 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 |

Example 5. This example is from Lam, Yao, and Batia (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 $A$, 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 $Y_t$ are defined as $Y_t = AX_t + \epsilon_t$, where $\epsilon_t$ is a random sample of $N(0, \Sigma)$ and is independent of $X_t$, where $\Sigma = (\sigma_{ij})^{p \times p}$ is defined as

\[
\sigma_{ij} = \begin{cases} 
\frac{1}{2} & \text{if } i \neq j \\
\frac{1}{2} + (i - j - 1)^2 & \text{if } i = j, H = 0.9, \\
1 & \text{if } i = j, H = 1.0.
\end{cases}
\]

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 1. When $k_0 = 3.$

Table 3. Mean, standard error (in the bracket) of $D$-distance and $\hat{r}$, $\hat{s}$ with 1000 replicates for Example 4.

| $i$ | $\hat{r}$ | $\hat{r} < 3$ | $\hat{r} = 3$ | $\hat{r} > 3$ | $\hat{r}$ | $\hat{r} < 3$ | $\hat{r} = 3$ | $\hat{r} > 3$ |
|-----|------------|---------------|---------------|---------------|------------|---------------|---------------|---------------|
| $p = 5$, $n = 300$ | | 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 | 0.15 (0.05) | 0.018 | 0.982 | 0 | 0.15 (0.05) | 0.018 | 0.982 | 0 |
| $p = 10$, $n = 1000$ | 0.391 (0.32) | 0 | 0.586 | 0.414 | 0.17 (0.02) | 0 | 1 | 0 | 0.17 (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 |

$k_0 = 1$
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 (6)). According to Table 4, the performance of \( \mathcal{L}_{k_0} \)-based and \( \Gamma_{k_0} \)-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 6. In this example, we replace the data-generating process of \( w_t \) in Example 5 by a nonlinear one as follows:

\[
\begin{align*}
\nu_t &= \begin{cases} 
0.5 + 0.05e^{-0.01w_{t-1}} & \text{if } w_{t-1} < 5, z_1 \sim N(0, 1), \\
+0.9w_{t-1} + z_1 & \text{if } w_{t-1} < 5, \\
0.9e^{-0.01w_{t-1}} + w_{t-1} + z_1 & \text{if } w_{t-1} \geq 5, 
\end{cases} 
\end{align*}
\]

Furthermore, \( Y_i \) is defined as \( Y_i = AX_i + \epsilon_i \), where error \( \epsilon_i \) is generated from \( N(0, 0.25 \Sigma) \) and \( \Sigma \) is defined in Example 5. Other parts of the model, such as \( A \) are exactly the same as Example 5, along with the combinations of \( (p, n, k_0) \). From Table 4, our \( \Gamma_{k_0} \)-based approach outperforms \( \mathcal{L}_{k_0} \)-based counterpart in all cases under both criteria with the advantage more pronounced when \( k_0 = 1 \). When \( k_0 = 5 \), \( \Gamma_{k_0} \)-based approach is slightly inferior to \( \Gamma_{k_0} \)-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 5), the inability of \( \mathcal{L}_{k_0} \)-based method to accommodate nonlinear dependence is amplified. It is also worth noting that for \( \Gamma_{k_0}, \Gamma_{k_0}^{(2)} \), and \( \mathcal{L}_{k_0} \), the performance can depend on \( k_0 \) to some extent.

Example 7. Example 7 is constructed by following Example 2 where \( r \) and \( s \) are different. The factor loading matrix \( A = ([A_1]_{10 \times 2}, [A_2]_{10 \times 1}) \) is a \( 10 \times 3 \) matrix. For each column of \( A \), the first five elements are iid \( U(-2, 2) \) and the rest five elements are set to 0. The time series \( X_t = ((W_{1t} + W_{2t})^T, W_{3t}^T)^T \), where \( W_{1t} = (W_{1t,1}, W_{1t,2})^T \) and \( W_{2t} = (W_{3t,1}, W_{3t,2})^T \). They are \( W_{1t,1} = u_t, W_{1t,2} = u_{t-1}, u_t = 0.5c_{t-1} + c_t \), where \( c_t = N(0, 8^2) \) and \( c_{t-1} \) follows \( N(0, 1.5^2) \). Then the data are generated by \( Y_t = AX_t + \epsilon_{1t} \), \( \epsilon_{1t} \sim N(0, 0.5 \times I_{10}) \) and independent from \( \{e_{i,j} \}, i = 1, 3 \). Note that \( W_{1t} \) is from a stationary MA(1) model and \( W_{3t} \), \( i = 3, 4 \) 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 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 that is, \( \mathcal{M}(A) = \mathcal{M}(A_1), \mathcal{M}(A_1) = \mathcal{M}(\{A_1, A_2\}) \). This is because \( W_{3t} \) and \( W_{4t} \) are white-noise sequence but not margingale difference sequence; see Example 2.3 in Shao (2011). According to Table 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 \( \Gamma_{k_0} \) 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 seems 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 7. Overall, the finite sample performance of MDDM-based method is quite encouraging.

### 6. Data Illustrations

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

- Forecasting error (FE) (Lam, Yao, and Bathia 2011)

\[
FE = \frac{\| \hat{Y}_{n+1}^{(1)} - Y_{n+1} \|}{p^{1/2}} = \frac{\| \hat{X}_{n+1}^{(1)} - Y_{n+1} \|}{p^{1/2}},
\]

where \( \hat{Y}_{n+1}^{(1)} \) is the one-step-ahead prediction for \( Y_{n+1} \) based on \( (Y_1, \ldots, Y_n) \) and \( \hat{X}_{n+1}^{(1)} \) is the one-step-ahead prediction for \( X_{n+1} \) based on a parametric model fitted to the estimated factor series \( (X_1, \ldots, X_n) \). 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 6.1.

#### 6.1. GDP Data

The first dataset we analyze is the quarterly change in seasonally adjusted GDP, in percentage, for five countries, that is, 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 dataset 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 Matteson and Tsay (2011, fig. 1), there seems no obvious nonstationarity in all time series. To realistically measure the forecasting error, we divide the dataset into training set and testing set. Approximately 80% of the data that contain 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 \( \bar{Y}_{h}^{(k)} \), \( h = 1, 2, 3, 4 \) based on \( (Y_j, \ldots, Y_{91+j}) \) for \( j = 1, 2, 3, \ldots, 24 - h \), \( h = 1, 2, 3, 4 \) 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 \( j = 1, \ldots, 24 - h \), \( h = 1, 2, 3, 4 \), we apply PCMC to the \( (Y_j, \ldots, Y_{91+j}) \) 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 Akaike information criterion (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 equal 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 \( \hat{A}_{22+j}^{(h)} \), \( h = 1, 2, 3, 4 \), where \( \hat{A} \) is the estimated factor loading matrix, and \( \hat{X}_{22+j}^{(h)} \) 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.

From the autocorrelation plot in Figure 1, significant autocorrelations and cross-correlations exist up to lag 3. In Table 6, we present the average forecasting error and the proportion of estimated number of factors based on \( \mathcal{L}_{k_0} \) and \( \Gamma_{k_0}^{(1)} \) for \( k_0 = 1, 2, 3 \). It can be seen that \( \Gamma_{k_0}^{(2)} \)-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 \( h \)-th step-ahead prediction and \( k_0 \)
is the number of lags involved. The performance of $\mathcal{L}_{k_0}$ and $\Gamma_{k_0}$ 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, $\Gamma_{k_0}$-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 $\Gamma_{k_0}$-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.

### 6.2. Seven City Temperature Series

The second dataset we analyze is the monthly temperature series for seven cities: Nanjing, Dongtai, Huoshan, Hefei, Shanghai, Anqing, and Hangzhou in Eastern China. The series run from

| $k_0$ | $\mathcal{L}_{k_0}$ | $\Gamma_{k_0}$ | DOC |
|-------|--------------------|----------------|-----|
|       | $\mathcal{L}_{k_0}$ | $\Gamma_{k_0}$ | DOC |
|       | $\mathcal{L}_{k_0}$ | $\Gamma_{k_0}$ | DOC |
| $k_0 = 1$ | 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 | 1 | 0 | 0.17 | 1.147 | 0 | 1 | 0 | 1.345 | 1 |

#### Table 6. Mean of FE and $\hat{r}, \hat{s}$ for GDP data.

#### Figure 1. Auto and cross-correlations of GDP of five countries.
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 suggests that there exist strong seasonal dependence. Following the approach in Section 6.1, we use the first 80% of the data that contain 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 $\hat{Y}_{433+j}$ based on $(Y_j, \ldots, Y_{432+j})$ for $j = 1, \ldots, 107$ and then report the average of forecasting errors over 107 periods. The one-step-ahead forecast is implemented using the following procedure: (1) for $j = 1, \ldots, 107$, we apply PPMC to the training data $(Y_j, \ldots, Y_{432+j})$ 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 3 for one particular training data; the model fits the estimated factor series quite well, as the residual auto-correlation is fairly small for most training data; see Figure 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 $\hat{X}_{433+j}$, where $\hat{A}$ is the estimated factor loading matrix, and $\hat{\mathcal{X}}_{433+j}$ is the one-step-ahead prediction based on the fitted seasonal ARIMA model to the lower dimensional factor series. In Table 7, we present the average forecasting error and the proportion of estimated number of factors based on $\mathcal{L}_{k_0}$ and $\mathcal{I}_{k_0}^{(1)}$. The choice of $k_0 = 12, 24, 36$ is based on the consideration that the time series is apparently seasonal with period 12.

It can be seen from Table 7 that $\mathcal{L}_{k_0}$-based approach and $\mathcal{I}_{k_0}$-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 $\mathcal{I}_{k_0}$-based method has substantially smaller FEs than those of $\mathcal{L}_{k_0}$-based method, as shown in Figure 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

---

**Figure 2.** Monthly temperatures for seven cities in Eastern China.

**Figure 3.** ACF plot of estimated factor series selected by $\Gamma_{k_0}^{(1)}$-based method.

---

| $k_0$ | $\mathcal{L}_{k_0}$ | $\mathcal{I}_{k_0}$ |
|-------|---------------------|---------------------|
|       | $\hat{r}$           | $\hat{s}$           |
| $\hat{r}$ | $\hat{s}$           | $\hat{r}$           | $\hat{s}$           |
| FE   | $\hat{r} = 1$ | $\hat{r} = 2$ | $\hat{r} > 2$ | $\hat{s} = 1$ | $\hat{s} = 2$ | $\hat{s} > 2$ |
| 12   | 1.894              | 0                   | 1                   | 0                   | 1.113              | 1                   | 0                   |
| 24   | 1.898              | 0                   | 1                   | 0                   | 1.113              | 1                   | 0                   |
| 36   | 1.892              | 0                   | 1                   | 0                   | 1.113              | 1                   | 0                   |
series is nonstationary, which is not covered by our theory. Nevertheless, it shows the potential applicability of our approach to nonstationary 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 the $\Gamma_{k_0}$-based dimension reduction method. We did not pursue this step since there seems no general guidance on how such modeling can be conducted.

7. Summary and Conclusion

In this article, we propose the so-called martingale difference divergence matrix to quantify the conditional mean (in)dependence of a random vector $V \in \mathbb{R}^p$ on $U \in \mathbb{R}^q$. 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 analog 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 (2017), there can be complications with the ratio-based estimator (see (6)) 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.

**Supplementary Materials**

The supplemental materials contain the proofs of Lemma 1, Theorem 1, Theorem 2, and Proposition 1.

**Acknowledgments**

The authors thank Professor Qiwei Yao for providing the temperature data used in this article, and thank Professor David Matteson for providing the GDP dataset and the R code for dynamic orthogonal component analysis. The authors are also grateful to two reviewers and associate editor for constructive comments that led to substantial improvements.

**Funding**

Shao's research was partially supported by NSF under Grant No. DMS14-07037 and DMS16-07489.

**References**

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. [216]

Bai, J. (2003), “Inferential Theory for Factor Models of Large Dimensions,” *Econometrica*, 71, 135–171. [216]

Bai, J., and Ng, S. (2002), “Determining the Number of Factors in Approximate Factor Models,” *Econometrica*, 70, 191–221. [216,219]

——— (2007), “Determining the Number of Primitive Shocks in Factor Models,” *Journal of Business and Economic Statistics*, 25, 52–60. [219]

Bathia, N., Yao, Q., and Ziegelmann, F. (2010), “Identifying the Finite Dimensionality of Curve Time Series,” *The Annals of Statistics*, 38, 3352–3386. [219]

Box, G. E. P., and Tiao, G. C. (1977), “A Canonical Analysis of Multiple Time Series,” *Biometrika*, 64, 335–365. [216]

Brillinger, D. R. (1981), *Time Series, Data Analysis and Theory* (expanded ed.), San Francisco, CA: Holden-Day. [216]

Cattell, R. B. (1966), “The Scree Test for the Number of Factors,” *Multivariate Behavioral Research*, 1, 245–276. [219]

Cook, R. D., Li, B., and Chiaromonte, F. (2010), “Envelope Models for Parsimonious and Efficient Multivariate Linear Regression,” *Statistica Sinica*, 20, 927–1010. [221]

Eichler, M., Motta, G., and von Sachs, R. (2011), “Fitting Dynamic Factor Models to Non-Stationary Time Series,” *Journal of Econometrics*, 163, 51–70. [216,229]

Forni, M., Hallin, M., and Reichlin, L. (2000), “The Generalized Dynamic-Factor Model: Identification and Estimation,” *Review of Economics and Statistics*, 82, 540–554. [216]

——— (2005), “The Generalized Dynamic-Factor Model: One-Sided Estimation and Forecasting,” *Journal of the American Statistical Association*, 100, 830–840. [216]

Geweke, J. (1977), “The Dynamic Factor Analysis of Economic Time Series Models,” in *Latent Variables in Socio-Economic Models*, eds. D. J. Aigner, and A. S. Goldberger, Amsterdam, The Netherlands: North-Holland, pp. 365–383. [216]

Granger, C. W., Maasoumi, E., and Racine, J. (2004), “A Dependence Metric for Possibly Nonlinear Processes,” *Journal of Time Series Analysis*, 25, 649–669. [220]

Hallin, M., and iLiska, R. (2007), “Determining the Number of Factors in the General Dynamic Factor Model,” *Journal of the American Statistical Association*, 102, 603–617. [219]

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. [220]

Hu, Y. P., and Tsay, R. S. (2014), “Principal Volatility Component Analysis,” *Journal of Business and Economics Statistics*, 32, 153–164. [219]

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. [216,217,219,221,222,228]

Lam, C., Yao, Q., and Bathia, N. (2011), “Estimation of Latent Factors for High-Dimensional Time Series,” *Biometrika*, 98, 901–918. [216,217,219,221,228]

Li, Z., Wang, Q., and Yao, J. (2017), “Identifying the Number of Factors from Singular Values of a Large Sample Auto-Covariance Matrix,” *Annals of Statistics*, 45, 257–288. [224,228]

Matteson, D. S., and Tsay, R. S. (2011), “Dynamic Orthogonal Components for Multivariate Time Series,” *Journal of the American Statistical Association*, 106, 1450–1463. [216,225]

Motta, G., Hafner, C. M., and von Sachs, R. (2011), “Locally Stationary Factor Models: Identification and Nonparametric Estimation,” *Econometric Theory*, 27, 1279–1319. [229]

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. [216]

Pan, J., and Yao, Q. (2008), “Modeling Multiple Time Series via Common Factors,” *Biometrika*, 95, 365–379. [216,219,227,229]

Park, J.-H., Srinam, T. N., and Yin, X. (2009), “Central Mean Subspace in Time Series,” *Journal of Computational and Graphical Statistics*, 18, 717–730. [221]

——— (2010), “Dimension Reduction in Time Series,” *Statistica Sinica*, 20, 747–770. [221]

Park, T., Shao, X., and Yao, S. (2015), “Partial Martingale Difference Correlation,” *Electronic Journal of Statistics*, 9, 1492–1517. [217,218]

Peña, D., and Box, G. E. P. (1987), “Identifying a Simplifying Structure in Time Series,” *Journal of the American Statistical Association*, 82, 836–843. [216,219]

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. [221,224]

Shao, X., and Wei, B. W. (2007), “Asymptotic Spectral Theory for Nonlinear Time Series,” *Annals of Statistics*, 35, 1773–1801. [222]

Shao, X., and Zhang, I. (2014), “Martingale Difference Correlation and its use in High Dimensional Variable Screening,” *Journal of the American Statistical Association*, 109, 1302–1318. [217]

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. [216]

——— (2005), “Implications of Dynamic Factor Models for VAR Analysis,” Working Paper 11467, National Bureau of Economic Research. [216]

Stoffer, D. (1999), “Detecting Common Signals in Multiple Time Series using the Spectral Envelope,” *Journal of the American Statistical Association*, 94, 1341–1356. [216]

Szekely, G. J., and Rizzo, M. L. (2014), “Partial Distance Correlation with Methods for Dissimilarities,” *Annals of Statistics*, 42, 2382–2412. [218]

Szekely, G. J., Rizzo, M. L., and Bakirov, N. K. (2007), “Measuring and Testing Independence by Correlation of Distances,” *Annals of Statistics*, 35, 2769–2794. [217]

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. [216]

Xia, Y., Tong, H., Li, W. K., and Zhu, L. X. (2002), “An Adaptive Estimation of Dimension Reduction,” *Journal of the Royal Statistical Society*, Series B, 64, 363–410. [221]

Zhou, Z. (2012), “Measuring Nonlinear Dependence in Time-Series, a Distance Correlation Approach,” *Journal of Time Series Analysis*, 33, 438–457. [220]