Statistica Sinica **20** (2010), 747-770

DIMENSION REDUCTION IN TIME SERIES

Jin-Hong Park[1], T. N. Sriram[2] and Xiangrong Yin[2]

[1] *College of Charleston and* [2] *University of Georgia*

Abstract: In this article, we develop a sufficient dimension reduction theory for time series. This does not require specification of a model but seeks to find a $p \times d$ matrix $\Phi_{\rm d}$ with the smallest possible number d ($\leq p$) such that the conditional distribution of $x_t|X_{t-1}$ is the same as that of $x_t|\Phi_{\rm d}^T X_{t-1}$, where $X_{t-1} = (x_{t-1}, \ldots, x_{t-p})^T$, resulting in no loss of information about the conditional distribution of the series given its past p values. We define the subspace spanned by the columns of $\Phi_{\rm d}$ as the time series central subspace and estimate it by maximizing Kullback-Leibler distance. We show that the estimator is consistent when p and d are known. In addition, for unknown d and p, we propose a consistent estimator of d and a graphical method to determine p. Finally, we present examples and a data analysis to illustrate a theory that may open new research avenues in time series analysis.

Key words and phrases: Density estimator, Kullback-Leibler distance, nonlinear time series, threshold, time series central subspace.

1. Introduction

Time series analysis has been an active area of research for decades, its intrinsic nature is that of correlated observations. This severely restricts the direct applicability of many conventional statistical methodologies that are primarily suited for analyzing independent and identically distributed data. Unique challenges posed by time series data sets have given rise to two broad approaches: *the time domain approach* and *the frequency domain approach*. While there are many useful parametric and nonparametric methods for analyzing time series data, there is a never-ending quest to build new methodologies to analyze the time series data that arise in a variety of fields such as economics, meteorology, engineering, geophysics, social, and environmental science.

Estimation approaches from classical regression theory are useful in building linear/nonlinear models for time series data $\{x_1, \cdots, x_t; t \geq 1\}$, where there is an obvious dependence between x_t and the past values $\{x_{t-1}, \ldots, x_1\}$; see e.g., Brockwell and Davis (1996), Shummway and Stoffer (2000), Fan and Yao (2003), Tsay (2005), and Wei (2006).

To address the issue of dimension reduction in time series, Xia and Li (1999), and Xia, Tong, and Li (1999, 2002) considered a single-index model that avoids

the curse of dimensionality. Recently, Xia, Tong, Li, and Zhu (2002) proposed a dimension reduction method in regression that is also applicable to time series with known lag; however, their focus is only on estimation of dimensions in the mean function. For an ensemble of time series, Li and Shedden (2002) presented a dimension reduction method that identifies a small number of independent time series components such that each time series in the ensemble is a different linear combination of the components. Their notion of dimension reduction, however, differs from our proposal below. Becker and Fried (2003) used a dynamic version of Sliced Inverse Regression (SIR; Li (1991)) as an exploratory tool for analyzing multivariate time series, where the lag is chosen using preliminary information. Hall and Yao (2005) discussed an estimation method that approximates the conditional distribution function of x_t given the past using a single linear combination of the past. To the best of our knowledge, there is no formal sufficient dimension reduction theory in time series that overcomes the curse of dimensionality without making specific model assumptions, or using specific numbers of dimensions and lag. Development of such a formal theory for time series, and illustration of its use in practice, are the main goals of this article.

The primary goal of time series analysis is forecasting, which requires inference about the conditional distribution of $x_t|X_{t-1}$, for some suitable lag $p \geq 1$, $X_{t-1} = (x_{t-1}, \ldots, x_{t-p})^T$. Typically, the lag p is not known. However, there are diagnostic ways and estimation methods for determining a value of p before proceeding with the inference (Ng and Perron (2005)). It is also important to note that with known p, we may only need a few linear combinations of X_{t-1} in the final model (Xia and Li (1999), and Xia, Tong, and Li (1999, 2002)), determination of which is one of our main focuses.

In Section 2, we develop a theory of sufficient dimension reduction in time series by introducing a notion called time series central subspace. In Section 3, we propose an estimation of the time series central subspace and state the main results. In Section 4, we carry out Monte Carlo simulations, followed by analysis of data sets. In Section 5, we give a brief discussion summarizing our approach and results. All the necessary proofs are given in the Appendix.

2. Central Subspace in Time Series

In time series it is useful to make inference about the conditional distribution of x_t given the past, $x_{t-1}, \ldots, x_1$. However, in many data sets one determines a value of $p \geq 1$ to make inference about the conditional distribution of $x_t|X_{t-1}$, for some $p \geq 1$. Here, we first assume that such a lag value p exists and is known, and then consider the case of unknown p.

Our goal is to find finitely many linear combinations, $\Phi_1^T X_{t-1}, \ldots, \Phi_q^T X_{t-1}$, $q \leq p$, such that the conditional distribution of $x_t|X_{t-1}$ is same as the conditional

distribution of $x_t|(\Phi_1^T X_{t-1}, \ldots, \Phi_q^T X_{t-1})$. This is equivalent to finding a $p \times q$ matrix $\Phi = (\Phi_1, \ldots, \Phi_q)$ such that

$$x_t \perp\!\!\!\perp X_{t-1}|\Phi^T X_{t-1}, \tag{2.1}$$

that is to say, x_t is independent of X_{t-1} given $\Phi^T X_{t-1}$. Then the $p\times 1$ vector X_{t-1} can be replaced by the $q \times 1$ vector $\Phi^T X_{t-1}$ without loss of information. This represents a useful reduction in the dimension of X_{t-1} where all the information in X_{t-1} about x_t is contained in the q-linear combinations.

We define a dimension reduction subspace for x_t on X_{t-1} as any subspace $\mathcal{S}(\Phi)$ of $\mathbb{R}^p$ for which (2.1) holds. Note that (2.1) holds trivially for $\Phi = I_p$ (Identity matrix), which implies that a dimension reduction space always exists. Since our primary aim is to reduce the dimension we seek a minimum dimension reduction space for x_t on X_{t-1}. To this end, we define the intersection of all dimension reduction spaces as a **Time Series Central Subspace**, denoted by $\mathcal{S}_{x_t|X_{t-1}}(\Phi_{\mathrm{d}})$, if the intersection is itself a dimension reduction space, where $\dim(\mathcal{S}_{x_t|X_{t-1}}(\Phi_{\mathrm{d}})) = \mathrm{d}$ and $\Phi_{\mathrm{d}} = (\Phi_1, \ldots, \Phi_d)$. Clearly, a time series central subspace is a minimum dimension reduction subspace; this provides an initial phase when an adequate parsimoniously parameterized time series model is not yet available.

Although our notion of time series central subspace bears similarity to the central subspace in regression (Cook (1994, 1998a)), an important difference is that the former implicitly depends on lag p, usually unknown in practice and requiring estimation, whereas p is usually known in regression. The definition of time series central subspace is general enough to include many linear and nonlinear time series models, as shown below. Nonetheless, we do require Φ to be independent of time t.

Time series central subspaces may not exist. The following Proposition guarantees the existence of a time series central subspace. We omit its proof because it follows from arguments similar to those in Cook (1998a). For a more general result on the existence, see Yin, Li and Cook (2008).

Proposition 1. *Let $\mathcal{S}(\eta)$ and $\mathcal{S}(\gamma)$ be dimension reduction subspaces for x_t on X_{t-1}. If X_{t-1} has a density $f(\mathbf{x}_{t-1}) > 0$ for $\mathbf{x}_{t-1} \in \Omega_{X_{t-1}} \subset \mathbb{R}^p$ and $f(\mathbf{x}_{t-1}) = 0$ otherwise, and if $\Omega_{X_{t-1}}$ is a convex set, then $\mathcal{S}(\eta)\cap\mathcal{S}(\gamma)$ is a dimension reduction subspace.*

As an illustration, let $p = 3$, $X_{t-1} = (x_{t-1}, x_{t-2}, x_{t-3})^T$, and set $x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \phi_3 x_{t-3} + \varepsilon_t$, where x_t's and ε_t are normal random variables and $\{X_{t-1}\}$ is independent of ε_t. Then the vector $(\phi_1, \phi_2, \phi_3)^T$ forms a basis for a time series central subspace. On the other hand, for fixed $p = 3$, when $x_t = \phi_1 x_{t-1}$,

we have that $\mathcal{S}((1,0,0)^T)$, $\mathcal{S}((0,1,0)^T)$, and $\mathcal{S}((0,0,1)^T)$ are all minimum dimension reduction subspaces. In this case, there does not exist a time series central subspace because the intersection of dimension reduction subspaces is empty.

We conclude this section with identification of the bases for time series central subspaces corresponding to three well-known time series models. For an autoregressive model of order p, $x_t = \phi_1 x_{t-1} + \cdots + \phi_p x_{t-p} + \varepsilon_t$, where $\{\varepsilon_t\}$ is a white noise sequence independent of $X_{t-1} = (x_{t-1}, \ldots, x_{t-p})^T$, the vector $(\phi_1, \ldots, \phi_p)^T$ forms a basis for a time series central subspace with $d = 1$. A general Threshold Autoregressive model of order p is defined in Tong and Lim (1980); also see Chan, Petrucelli, Tong and Woolford (1985), and Xia, Li and Tong (2007). More specifically, for an integer l, let $-\infty = r_0 < r_1 < \cdots < r_l = \infty$ denote the thresholds and define $x_t = \sum_{k=1}^{l} \{\sum_{i=1}^{p} \phi_i^{(k)} x_{t-i} + \varepsilon_t^{(k)}\}\, I(x_{t-p^*} \in (r_{k-1}, r_k])$, where $I(A)$ is the indicator function of a set A, x_{t-p^*} is the threshold variable for some fixed $1 \le p^* \le p$, for each $1 \le k \le l$, $\{\varepsilon_t^{(k)}\}$ is a white noise sequence independent of $X_{t-1} = (x_{t-1}, \ldots, x_{t-p})^T$, and $\{\varepsilon_t^{(j)}\}$ is independent of $\{\varepsilon_t^{(j')}\}$ for $1 \le j \ne j' \le l$. Let $\Phi_k = (\phi_1^{(k)}, \ldots, \phi_p^{(k)})^T$ for $1 \le k \le l$. If for example, $l \le p$ and $\Phi_i \not\propto \Phi_j$ for $i \ne j$, where $\not\propto$ is 'not proportional to', then the matrix $\Phi = (\Phi_1, \ldots, \Phi_l, \ell_{p^*})$ forms a basis for a time series central subspace, where ℓ_{p^*} is a vector with its p^*th element equal to 1 and is 0 elsewhere. Finally, consider an autoregressive conditionally heteroscedastic (ARCH) model (Engle (1982)), $x_t = \sigma_t \varepsilon_t$, where ε_t are independent and identically distributed random variables with mean 0 and variance 1, and for $q < p$, $\sigma_t^2 = \theta_0 + \theta_1 x_{t-1}^2 + \theta_2 x_{t-2}^2 + \cdots + \theta_q x_{t-q}^2$. Then, Φ is composed of vectors e_i (a vector with ith element equal to 1 and 0 elsewhere), for $i = 1, \ldots, q$. In addition, if we know that the model is ARCH, then we can set $X_{t-1} = (x_{t-1}^2, \ldots, x_{t-p}^2)^T$, which implies $d = 1$; hence our estimation procedure may be more efficient.

For the rest of the article, we assume the existence of the time series central subspace. Examples above show that our definition of time series central subspace is general. However, it does not include moving average, autoregressive-moving average, or generalized autoregressive conditionally heteroscedastic models, which may be viewed as infinite order autoregression.

3. Estimation

To estimate $\mathcal{S}_{x_t|X_{t-1}}(\Phi_{\mathrm{d}})$, we begin by assuming that $d(\le p)$ and p in X_{t-1} are known. It is natural to ask if the estimation of Φ_{d} may be carried out using well-known sufficient dimension reduction methods in regression such as SIR (Li (1991)) and Sliced Average Variance Estimation (SAVE; Cook and Weisberg (1991)). These are powerful methods in regression, but they impose a stringent requirement on the distribution of X_{t-1}, that severely limits their use in time

series analysis; see Xia, Tong, Li, and Zhu (2002, p.365). Li (1991) points out that removal of outliers upon closer examination of the distribution of regressors may help the analysis using SIR; however, the removal may create problems in time series analysis. The minimum average variance estimation method of Xia, Tong, Li, and Zhu (2002) is another useful way of addressing the problem of dimension reduction in time series. However, their approach focuses only on estimation of dimensions in the mean function.

All this leads us to suggest a sufficient dimension reduction method for time series analysis, that stems from the expected conditional log-likelihood approach of Yin and Cook (2005) for single-index regression in the independent and identically distributed case. We believe that our sufficient dimension reduction method is particularly suitable for time series analysis because it does not constrain the distribution of X_{t-1}, nor does it limit consideration to the mean or variance functions.

3.1. Expected conditional log-likelihood

Let $p(\cdot,\cdot)$, $p(\cdot|\cdot)$, and $p(\cdot)$ denote joint, conditional, and marginal densities, respectively. For $p \times q$ matrices h with $q \le p$, we consider an objective function $\Psi(\mathrm{h})$ that measures the mutual information (Cover and Thomas (1991)) between $\mathrm{h}^T X_{t-1}$ and x_t, defined by

$$\Psi(\mathrm{h}) = \mathrm{E}\left\{ \log \frac{p(\mathrm{h}^T X_{t-1}, x_t)}{p(x_t)p(\mathrm{h}^T X_{t-1})} \right\} = \mathrm{E}\left\{ \log \frac{p(x_t|\mathrm{h}^T X_{t-1})}{p(x_t)} \right\}. \tag{3.1}$$

We want to maximize this objective function over all $p \times d$ matrices h with $\mathrm{h}^T\mathrm{h} = I_d$. Since $p(x_t)$ does not involve h, maximizing $\Psi(\mathrm{h})$ is the same as maximizing the expected conditional log-likelihood. This mutual information is the Kullback-Leibler divergence between the joint density, $p(\mathrm{h}^T X_{t-1}, x_t)$, and the product of the marginal densities, $p(x_t)p(\mathrm{h}^T X_{t-1})$, quantifying the dependence of x_t on $\mathrm{h}^T X_{t-1}$. The following proposition shows that this is a reasonable method for identifying the time series central subspace.

Proposition 2. *Let h_1, h_2 and h_d be $p\times q_1$, $p\times q_2$, and $p\times d$ matrices, respectively, where $q_1, q_2, d \le p$.*

(i) *If $\mathcal{S}(h_1) = \mathcal{S}(h_2)$, then $\Psi(h_1) = \Psi(h_2)$.*

(ii) *$\Psi(I_p) \ge \Psi(h_1)$, and equality holds if and only if $x_t \perp\!\!\!\perp X_t | h_1^T X_{t-1}$. Consequently, $\Psi(I_p) = \Psi(\Phi_d) \ge \Psi(h_d)$, and equality holds if and only if $\mathcal{S}_{x_t|X_{t-1}}(\Phi_d) = \mathcal{S}(h_d)$.*

(iii) *$\Psi(h_1) \ge 0$. Moreover, if $d > q_2 > q_1 \ge 1$, then $\Psi(I_p) = \Psi(\Phi_d) = \max_{h_d} \Psi(h_d) > \max_{h_2} \Psi(h_2) > \max_{h_1} \Psi(h_1)$.*

Part (i) of Proposition 2 says that only $\mathcal{S}(\mathrm{h})$ matters when maximizing $\Psi(\mathrm{h})$ and not the particular basis of the subspace. Hence, we may use the constraint $\mathrm{h}^T\mathrm{h} = I_d$ for identifiability. Part (ii) helps us confirm whether $\mathcal{S}(\mathrm{h})$ is a dimension reduction subspace or not by comparing $\Psi(\mathrm{h})$ with $\Psi(I_p)$, if $\Psi(I_p)$ is known. More importantly, Part (ii) says that $\arg\max_{\mathrm{h}_d} \Psi(\mathrm{h}_d)$ is always a basis for the time series central subspace. Part (iii) provides a theoretical justification for sequential search of the time series central subspace by showing that the information content increases with the dimension until dimension d is achieved. This result will also be useful in Section 3.3.

3.2. Computational algorithm

If all the densities were known, then we could use the first equality in (3.1) as the basis of a sample version

$$\Psi_n(\mathrm{h}) = \frac{1}{n}\sum_{t=1}^{n} \log \frac{p(\mathrm{h}^T X_{t-1}, x_t)}{p(x_t)p(\mathrm{h}^T X_{t-1})},$$

and maximize it over all $p \times d$ matrices h subject to the constraint. In practice, however, the densities in $\Psi_n(\mathrm{h})$ are not known and we have to estimate them nonparametrically. For this, we need a one-dimensional density estimate of $p(x_t)$ and, for fixed h, multi-dimensional density estimates of $p(\mathrm{h}^T X_{t-1}, x_t)$ and $p(\mathrm{h}^T X_{t-1})$. General guidelines for choice of kernels and selection of bandwidths can be found in Silverman (1986) and Scott (1992).

In our computations, we use a Gaussian kernel for one-dimensional density estimation, and product Gaussian kernels for multi-dimensional density estimation. More specifically, let G denote the univariate Gaussian kernel, and $(u_1, \ldots, u_k)^T$ be the $k \times 1$ random vector, for $k \geq 1$. Denote the ith observation by $(u_{1i}, \ldots, u_{ki})^T$, then the k-dimensional density estimate is:

$$p_n(u_1, \ldots, u_k) = \left(n \prod_{j=1}^{k} a_{nj}\right)^{-1} \sum_{i=1}^{n} \prod_{j=1}^{k} G\left(\frac{u_j - u_{ji}}{a_{nj}}\right), \tag{3.2}$$

where $a_{nj} = b_k s_j n^{-1/(4+k)}$ for $j = 1, \ldots, k$, $b_k = \{4/(k+2)\}^{1/k+4}$, and s_j is the corresponding sample standard deviation of u_j that must be updated during iteration. Inclusion of s_j in the bandwidth term is not necessary; however, doing so usually improves the estimation. The constant b_k is the optimal bandwidth in the sense of minimizing mean integrated square error (Silverman (1986, p.87), and Scott (1992, p.152)). This choice performs well in our simulations and data analysis.

Finally, we replace the densities in $\Psi_n(\mathrm{h})$ by their corresponding estimates defined in (3.2) and maximize

$$\hat{\Psi}_n(\mathrm{h}) = \frac{1}{n}\sum_{t=1}^{n} \log \frac{p_n(\mathrm{h}^T X_{t-1}, x_t)}{p_n(x_t)p_n(\mathrm{h}^T X_{t-1})}$$

over all $p \times d$ matrices h satisfying the constraint $\mathrm{h}^T\mathrm{h} = I_d$. A method that naturally incorporates this constraint is the Sequential Quadratic Programming procedure (Gill, Murray and Wright (1981, Chap. 6)). The code for our algorithm is available in *MATLAB* from the authors.

Note that it is also possible to construct an alternative sample version of $\Psi(\mathrm{h})$ in (3.1). One may construct this using, for example, the local linear (or polynomial) smoother of conditional density proposed by Fan, Yao and Tong (1996) with smoothing parameters chosen based on the Residual Squares Criterion of Fan and Gijbels (1995) (also see Fan, Heckman and Wand (1995)), and thus carry out maximization of the resulting sample version. Incidentally, note that our approach is same as local constant approximation. There are also a host of other nonparametric smoothers of conditional density we can use, such as those proposed by De Gooijer and Zerom (2003) and Hyndman and Yao (2002). We plan to pursue this alternative approach in future. Here, we are encouraged by the large sample properties of our estimator seen in Section 3.4, and by the performance of our method in simulations and data analysis in Section 4.

3.3. Estimation of dimension d and lag p

In practice, we need to estimate d and p since they are usually unknown. Reiterating a comment made in Section 2, an important difference between dimension reduction in regression and that in time series is that, in the former context, we need only estimate d since p is usually known (Li (1992), Schott (1994), Cook (1998b), and Xia, Tong, Li, and Zhu (2002)), whereas, the lag p also requires estimation in our context. Motivated by Proposition 2 (iii) and by a recent work of Woo and Sriram (2006) for finite mixtures, we propose an estimator of d using the estimating function $\hat{\Psi}_n$ above. In addition, we propose a graphical method for determining p that differs from traditional approaches in time series analysis (Ng and Perron (2005)).

The estimators of d and p, respectively, are defined as follows.

Step 1: Note that if $p = 1$, then $d = 1$, and there is no need for dimension reduction. The procedure starts by fixing a value of lag $p(\geq 2)$ and determines

$$\hat{d}_p = \min\Big\{k(\leq (p-1)) : \hat{c}_k \leq \tau_{p,n}\Big\}, \tag{3.3}$$

where $\hat{c}_k = \hat{\Psi}_n(\hat{\mathrm{h}}_{p,(k+1)}) - \hat{\Psi}_n(\hat{\mathrm{h}}_{p,k})$ with $\hat{\mathrm{h}}_{p,k} = \arg\max_{\mathrm{h}_k} \hat{\Psi}_n(\mathrm{h}_k)$, and the maximization is over all $p \times k$ matrices h_k, $\{\tau_{p,n}; n \geq 1\}$ a sequence of non-negative threshold values chosen in such a way that it converges to zero as $n \to \infty$. For our simulations and data analysis in Section 4, we set the threshold value $\tau_{p,n} = 0$ and take $\chi^2_p(\alpha)/(2n)$, where $\chi^2_p(\alpha)$ is the $100(1-\alpha)$ percentile of Chi-square distribution with p degrees of freedom.

The procedure defined in (3.3) successively compares $\hat{c}_k$ (> 0 because of Proposition 2 (iii)) with the threshold value, and stops at the first value of k for which $\hat{c}_k$ is at or below the threshold. This yields an estimate $\hat{d}_p$ of d for a given value of p, which in turn yields an estimate $\hat{\mathrm{h}}_{p,\hat{d}_p}$ of the time series central subspace with the maximum value $\hat{\Psi}_n(\hat{\mathrm{h}}_{p,\hat{d}_p})$. Obviously, if $\hat{c}_k$ never falls below the threshold, then $\hat{d}_p = p$.

Step 2: Repeat Step 1 for each $p = 2, 3, \ldots$. This process yields a sequence of estimates $\{\hat{d}_p\}$ and a corresponding sequence of maximum values $\{\hat{\Psi}_n(\hat{\mathrm{h}}_{p,\hat{d}_p})\}$. Two procedures, graphical or information criteria, may be used to determine p. Graphically, we can plot $\{\hat{\Psi}_n(\hat{\mathrm{h}}_{p,\hat{d}_p})\}$ versus p and look for the value of $\hat{p}$ at which $\hat{\Psi}_n(\hat{\mathrm{h}}_{\hat{p},\hat{d}_{\hat{p}}})$ is essentially the largest; that is, the subsequent values of $\hat{\Psi}_n(\hat{\mathrm{h}}_{p,\hat{d}_p})$ are about the same or less than $\hat{\Psi}_n(\hat{\mathrm{h}}_{\hat{p},\hat{d}_{\hat{p}}})$, creating a shoulder-like situation at $p = \hat{p}$. Hence the name *Shoulder Plot*. This gives us an estimate of lag p.

We may also use the Akaike Information Criterion (AIC) or the Bayesian information criterion (BIC) to determine p;

$$BIC : \hat{p} = \arg\min_p \{-2n\hat{\Psi}_n(\hat{\mathrm{h}}_{p,\hat{d}_p}) + p\hat{d}_p \ln(n)\}$$
$$AIC : \hat{p} = \arg\min_p \{-2n\hat{\Psi}_n(\hat{\mathrm{h}}_{p,\hat{d}_p}) + 2p\hat{d}_p\}.$$

For instance, in Model 2 (see Section 4), given $\hat{d}_p = 1$, BIC and AIC detect the correct lag ($p = 6$) 79% and 76% of times, respectively. These are slightly less accurate than our answers using *Shoulder plot* in Section 4, but are still reasonable. We use the *Shoulder Plot* in all our examples because it has visual appeal and simplicity.

Note that Steps 1 and 2 yield estimates $\hat{p}$ and $\hat{d}_{\hat{p}}$. The idea behind a *Shoulder Plot* is similar to an *Elbow Plot*, which plots the (decreasing) eigenvalues against the serial numbers in a principal component analysis. In data analysis, our process begins with estimation of d and p, followed by estimation of the time series central subspace. Our calculations do require use of multi-dimensional density estimation. However, even if the value of lag p is large, the dimensionality of densities used in our iterative algorithm to determine d is at most $(d+1)$, or close to it. In practice, d is usually small, say, $d \leq 3$, and our multi-dimensional density estimation does not suffer from the curse of dimensionality.

It is known that sufficient dimension reduction methods such as SIR and SAVE have a naturally nested structure for the extracted linear combinations. In our context, our simulations do suggest that $\mathcal{S}(\hat{\mathrm{h}}_{p,i}) \subseteq \mathcal{S}(\hat{\mathrm{h}}_{p,i+1})$ for $i = 1, \ldots, d$. However, we do not yet have a theoretical proof of this observation. Nonetheless, observe that this nested structure is a stronger result than the one in Proposition 2 (iii).

Note that p is fixed in the definition of time series central subspace. In practice, p is usually unknown and there are different methods to estimate it. Our *Shoulder Plot* approach is one viable way of estimating p. For instance, we recognize that there are two possible ways of using our approach: One adopt a 'dual' approach where we fix d and search for the best p and then find the best pair of d and p; search for the best d and p using the matrix of dimension by lag. It is possible that these two approaches lead to different values of p. Nevertheless, different values of p may eventually lead us to satisfactory models with possibly different structures, providing a good fit of the data at hand. Despite having different structures, all these models are sufficient, hence their respective dimensions are also sufficient.

3.4. Consistency of the estimators

In this section, we establish the consistency of the estimate of the time series central subspace and of d. Unlike in Section 3.2, we do not restrict to Gaussian kernels. Suppose M_i is a sequence of k-dimensional random vectors with density $p(\cdot)$ and distribution function F. Consider a density estimator of $p(\cdot)$ as

$$f_n(M) = \frac{1}{na_n^k} \sum_{i=1}^{n} K\left(\frac{M - M_i}{a_n}\right)$$

for $M \in \mathbb{R}^k$, where $K : \mathbb{R}^k \to \mathbb{R}_+$ is a probability density, $\lim K(M) = 0$ uniformly for $\|M\| \to \infty$, $a_n > 0$, and $\lim_{n\to\infty} a_n = 0$. Let $\kappa_\iota = \{t : f_n(x_t) > \iota, f_n(h^T X_{t-1}) > \iota, f_n(h^T X_{t-1}, x_t) > \iota\}$ for any fixed $p \times d$ matrix h such that $\mathrm{h}^T\mathrm{h} = I_d$. Here ι is chosen in the following way: let $\epsilon \to 0$, and $\iota \to 0$, but $(\epsilon/\iota) \to 0$ as $n \to \infty$ for $\epsilon > 0$ and $\iota > 0$. Let n_ι be the number of observations whose indices are not in κ_ι.

Theorems 1 and 2 stated below are proved in the Appendix utilizing Lemmas 1 and 2 found there. In Theorem 1, the distance for convergence is $\|(I_p - \hat{\Phi}_{\mathrm{q}}\hat{\Phi}_{\mathrm{q}}^T)\Phi_{\mathrm{d}}\|^2$ (Xia, Tong, Li, and Zhu (2002)).

Theorem 1. *Assume the conditions of Lemma* 1 *and that* $(n_\iota/n) \to 0$ *in probability as* $n \to \infty$. *Let* $\hat{\Phi}_n = \arg\max_h \hat{\Psi}_n^\iota(h)$ *and* $\Phi_d = \arg\max_h \Psi(h)$, *where* $\hat{\Psi}_n^\iota(h) = (1/n)\sum_{t=1}^n J(t \in \kappa_\iota) \log(f_n(h^T X_{t-1}, x_t)/[f_n(x_t) f_n(h^T X_{t-1})])$, $J(t \in \kappa_\iota)$ *is the indicator function for* κ_ι, $\Psi(h)$ *is as in* (3.1) *and the maximization is over*

all $p \times d$ matrices h such that $h^T h = I_d$. Then $\hat{\Phi}_n$ converges to Φ_d with probability one as $n \to \infty$.

Theorem 2. *Assume the conditions of Lemma* 1*, Lemma* 2*, and Theorem* 1*. Let $\hat{d}_p^{\iota} = \min\{k(\leq (p-1)) : \hat{c}_k^{\iota} \leq \tau_{p,n}\}$, where $\hat{c}_k^{\iota}$ is the same as $\hat{c}_k$ defined at* (3.3) *with $\hat{\Psi}_n$ replaced by $\hat{\Psi}_n^{\iota}$ defined in Theorem* 1*. If for each fixed p, $\tau_{p,n} \to 0$ as $n \to \infty$, then $\hat{d}_p^{\iota}$ converges to d with probability one as $n \to \infty$, where d is the dimension of time series central subspace.*

Corollary. *Under the conditions of Theorems* 1 *and* 2*, we have that $\hat{\Phi}_{n,\hat{d}_p^{\iota}}$ converges to Φ_d. Here, $\hat{\Phi}_{n,\hat{d}_p^{\iota}}$ is the same as $\hat{\Phi}_n$ with d replaced by $\hat{d}_p^{\iota}$.*

The proof of the Corollary follows from same arguments as in Theorem 1; hence, it is omitted. Restriction to the set κ_{ι} in Theorems 1 and 2 above is a common truncation tool for proving theoretical results (See, e.g., Härdle and Stoker (1989)). In practice, we do not use such a restriction.

4. Simulation and Data Analysis

In this section, we use the measures proposed by Ye and Weiss (2003) and Xia, Tong, Li, and Zhu (2002) to assess the accuracy of our estimates. Ye and Weiss (2003) use *the vector correlation coefficient* (Hotelling (1936)) $\rho = |\hat{\Phi}_{\mathrm{d}}^T \Phi_{\mathrm{d}} \Phi_{\mathrm{d}}^T \hat{\Phi}_{\mathrm{d}}|^{1/2}$, where $|A|$ denotes the determinant of a matrix A. Note that $0 \leq \rho \leq 1$, and when $\rho = 1$, $\mathcal{S}_{x_t|X_{t-1}}(\hat{\Phi}_{\mathrm{d}}) = \mathcal{S}_{x_t|X_{t-1}}(\Phi_{\mathrm{d}})$. Therefore, higher values of ρ imply that the two spaces are closer and, hence, the estimates are more accurate. On the other hand, the method in Xia, Tong, Li, and Zhu (2002) (see also Li, Zha and Chiaromonte (2005)) measures the distance between $\mathcal{S}_{x_t|X_{t-1}}(\hat{\Phi}_{\mathrm{q}})$ and $\mathcal{S}_{x_t|X_{t-1}}(\Phi_{\mathrm{d}})$ using $m^2 = \|(I_p - \Phi_{\mathrm{d}}\Phi_{\mathrm{d}}{}^T)\hat{\Phi}_{\mathrm{q}}\|^2$ if $q < d$, $m^2 = \|(I_p - \hat{\Phi}_{\mathrm{q}}\hat{\Phi}_{\mathrm{q}}^T)\Phi_{\mathrm{d}}\|^2$ if $q \geq d$. Here, smaller values of m^2 yield more accurate estimates. For each simulation, we used sample sizes $n = 100$, 200 and 300, and we performed 200 Monte Carlo replications. The error $\{\varepsilon_t\}$ was taken to be a sequence of independent standard normal random variables.

Model 1. $x_t = -1 - \cos((\pi/2)(x_{t-1} + 2x_{t-4})) + 0.2\varepsilon_t(-2x_{t-1} + 2x_{t-2} - 2x_{t-3} + x_{t-4} - x_{t-5} + x_{t-6})$, where $p = 6$ and $d = 2$. Here, in addition to a nonlinear mean function, the error term ε_t involves a linear function depending on the past of the series. For $p = 6$ and $p = 10$, Table 1 gives the average values of ρ and m^2, respectively for fixed $d = 2$. The two different m^2 values in each cell correspond to estimates $\hat{\Phi}_1$ and $\hat{\Phi}_2$, respectively. The results show that the estimates are accurate. In addition, accuracy is better when correct lag ($p = 6$) is used as compared to using a wrong lag ($p = 10$). More importantly, the results attest to the fact that our estimation procedure can perform reasonably well even when

Table 1. Model 1: Average values of accuracy measures ρ and m^2 based on 200 Monte Carlo replications.

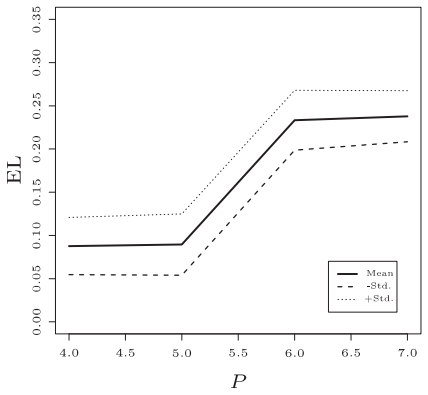
n	lag p	ρ	m^2	
100	10	0.7000	0.3006	0.2720
100	6	0.8711	0.0924	0.1460
200	10	0.8356	0.1519	0.1918
200	6	0.9379	0.0304	0.0843
300	10	0.8954	0.1087	0.1278
300	6	0.9713	0.0127	0.0429

Table 2. Model 2: Average values of accuracy measures ρ and m^2, and frequency of estimated dimension for 0-threshold, 0.05-threshold and 0.01-threshold, all based on 200 Monte Carlo replications. The true dimension is $d = 1$.

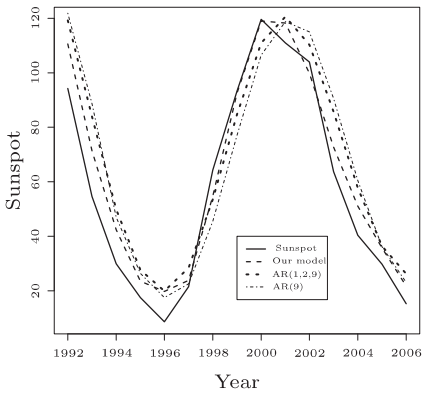
n	lag p	ρ	m^2	0-threshold	0.05-threshold	0.01-threshold
100	10	0.9965	0.0069	f_1=114	f_1=180	f_1=197
				f_{2+}=86	f_{2+}=20	f_{2+}=3
100	6	0.9985	0.0030	$f_1 = 46$	f_1=174	f_1=196
				f_{2+}=154	f_{2+}=26	f_{2+}=4
200	10	0.9989	0.0021	f_1=94	f_1=191	f_1=197
				f_{2+}=106	f_{2+}=9	f_{2+}=3
200	6	0.9995	0.0010	f_1=115	f_1=194	f_1=199
				f_{2+}=85	f_{2+}=6	f_{2+}=1
300	10	0.9994	0.0011	f_1=165	f_1=199	f_1=200
				f_{2+}=35	f_{2+}=1	f_{2+}=0
300	6	0.9997	0.00005	f_1=184	f_1=200	f_1=200
				f_{2+}=16	f_{2+}=0	f_{2+}=0

the conditional mean is nonlinear and the conditional variance is also a function of the past. To infer p using *Shoulder Plot*, we set $d = 2$ and computed $\hat{\Psi}_n(\hat{\mathrm{h}}_{p,2})$ with $p = 4, 5, 6, 7$ for 100 simulated data sets, each with $n = 300$. For 45% of the data sets, the *Shoulder Plot* correctly indicated that $\hat{p} = 6$. Note that the percentage of correct identification is relatively low. However, this is expected because there is a dimension in the variance function.

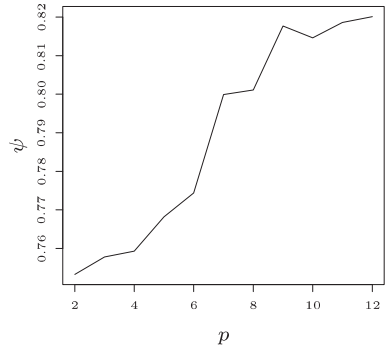
Model 2. $x_t = -1 - \cos((\pi/2)(x_{t-3} + 2x_{t-6})) + 0.2\varepsilon_t$, where $p = 6$ and $d = 1$. Table 2 shows that our estimates of Φ_1 are rather accurate, even with the wrong lag $p = 10$. In Table 2, we report f_i, the frequency of $\hat{d}_p = i$, based on 200 Monte Carlo replications using threshold values $\tau_{p,n} = 0$ (0-threshold), $\chi^2_p(0.05)/(2n)$ (0.05-threshold), and $\chi^2_p(0.01)/(2n)$ (0.01-threshold). Here, f_{i+} denotes the frequency of $\hat{d}_p \geq i$. For sample sizes $n = 100$ and 200, Table 2 indicates that, in terms of correctly identifying d, $\hat{d}_p$ with 0.05-threshold and



a. *Shoulder plot* for Model 2.



b. Forecast for Wolf yearly sunspot data:.



c. *Shoulder plot* for Wolf yearly sunspot data.

Figure 1. a: *Shoulder plot* of average values ("Mean"), average − standard deviation values ("-Std.") and average + standard deviation values ("+Std.") of $\hat{\Psi}_n(\hat{\mathrm{h}}_{p,1})$ versus $p = 4, 5, 6, 7$, based on 100 simulated datasets, each with sample size $n = 300$; b: Overlay plot of observed sunspot numbers (Sunspot) and forecast values from AR(9), AR(1, 2, 9), and our model: Years 1992−2006; c: *Shoulder plot* of $\hat{\Psi}_n(\hat{\mathrm{h}}_{p,2})$ values (second column of Table 5) versus $p = 2, \ldots, 12$.

0.01-threshold perform better (even when $p = 10$) than that with 0-threshold. For $n = 300$, the performance of $\hat{d}_p$ with 0-threshold is much better, but 0.05- and 0.01- thresholds show near perfect performance.

To infer p using *Shoulder Plot*, we set $d = 1$ and computed $\hat{\Psi}_n(\hat{\mathrm{h}}_{p,1})$ with $p = 4, 5, 6, 7$ for 100 replicated data sets, each with $n = 300$. For 95% of the data sets, the *Shoulder Plot* correctly indicated $\hat{p} = 6$. We also computed the average and the standard deviation of $\hat{\Psi}_n(\hat{\mathrm{h}}_{p,1})$ values for each $p = 4, 5, 6, 7$ based on the 100 simulated data sets. The resulting *Shoulder Plot* in Figure 1a clearly

Table 3. Model 3: Average values of accuracy measures ρ and m^2, and frequency of estimated dimension for 0-threshold, 0.05-threshold and 0.01-threshold, all based on 200 Monte Carlo replications. The true dimension is $d = 2$.

n	lag p	ρ	m^2		0-threshold	0.05-threshold	0.01-threshold
100	10	0.90	0.10	0.10	$f_1 = 49$ f_2=67 f_{3+}=84	f_1=137 f_2=54 f_{3+}=9	f_1=155 f_2=45 f_{3+}=0
100	6	0.97	0.04	0.03	f_1=26 f_2=52 f_{3+}=122	f_1=41 f_2=148 f_{3+}=11	f_1=50 f_2=145 f_{3+}=5
200	10	0.97	0.03	0.02	f_1=60 f_2=80 f_{3+}=60	f_1=85 f_2=83 f_{3+}=32	f_1=113 f_2=71 f_{3+}=16
200	6	0.99	0.01	0.01	f_1=23 f_2=82 f_{3+}=95	f_1=34 f_2=147 f_{3+}=23	f_1=37 f_2=149 f_{3+}=17
300	10	0.98	0.02	0.02	f_1=58 f_2=85 f_{3+}=57	f_1=71 f_2=91 f_{3+}=38	f_1=87 f_2=84 f_{3+}=29
300	6	0.99	0.01	0.01	f_1=14 f_2=93 f_{3+}=93	f_1=21 f_2=156 f_{3+}=23	f_1=21 f_2=159 f_{3+}=20

indicated that $\hat{p} = 6$.

Model 3. $x_t = -1 - \cos((\pi/2)(x_{t-1})) - \cos((\pi/2)(1/\sqrt{5})(x_{t-3} + 2x_{t-6})) + 0.2\varepsilon_t$, where $p = 6$ and $d = 2$. Table 3 shows that the accuracy of estimates of Φ_2 are reasonable. As for estimation of d, Table 3 shows that $\hat{d}_p$ with 0.05-threshold and 0.01-threshold correctly estimated the true dimension, $d = 2$, about 73% to 80% of the times, for all sample sizes and when the lag p is 6. However, it also shows that wrong specification of lag adversely affected the performance of $\hat{d}_p$, resulting in severe underestimation for 0.05- and 0.01-threshold. On the other hand, the 0-threshold did not perform well and, in fact, considerably overestimated the true dimension.

Finally, we set $d = 2$ and computed $\hat{\Psi}_n(\hat{\mathrm{h}}_{p,2})$ with $p = 4, 5, 6, 7$ for 100 simulated data sets, each with $n = 300$. For 94% of the data sets, the *Shoulder Plot* correctly indicated that $\hat{p} = 6$.

Model 4.

$$\begin{aligned} x_t = {} & -1 + (0.4)\Big(\frac{1}{\sqrt{5}}\Big)(x_{t-1} + 2x_{t-4}) - \cos\left(\Big(\frac{\pi}{2}\Big)\Big(\frac{1}{\sqrt{5}}\Big)(x_{t-3} + 2x_{t-6})\right) \\ & + \exp\left(-\Big(\frac{1}{\sqrt{15}}\Big)^2(-2x_{t-1} + 2x_{t-2} - 2x_{t-3} + x_{t-4} - x_{t-5} + x_{t-6})^2\right) \\ & + 0.2\varepsilon_t, \end{aligned}$$

where $p = 6$ and $d = 3$. This is Example 3 of Xia, Tong, Li, and Zhu (2002). Table 4 shows that the accuracy of estimates of Φ_3 are better for large sample

Table 4. Model 4: Average values of accuracy measures ρ and m^2, and frequency of estimated dimension for 0-threshold, 0.05-threshold and 0.01-threshold, all based on 200 Monte Carlo replications. The true dimension is $d = 3$.

n	p	ρ	m^2	0-threshold	0.05-threshold	0.01-threshold
100	10	0.65	0.25 0.18	f_1=20 f_2=105	f_1=41 f_2=155	f_1=62 f_2=138
			0.23	f_3=63 f_{4+}=12	f_3=1 f_{4+}=0	f_3=0 f_{4+}=0
100	6	0.85	0.11 0.06	f_1=2 f_2=67	f_1=10 f_2=187	f_1=28 f_2=172
			0.10	f_3=102 f_{4+}=29	f_3=3 f_{4+}=0	f_3=0 f_{4+}=0
200	10	0.79	0.14 0.10	f_1=11 f_2=125	f_1=14 f_2=183	f_1=17 f_2=183
			0.15	f_3=63 f_{4+}=1	f_3=3 f_{4+}=0	f_3=0 f_{4+}=0
200	6	0.93	0.07 0.02	f_1=32 f_2=33	f_1=45 f_2=114	f_1=49 f_2=134
			0.04	f_3=135 f_{4+}=0	f_3=41 f_{4+}=0	f_3=17 f_{4+}=0
300	10	0.86	0.09 0.08	f_1=9 f_2=110	f_1=11 f_2=179	f_1=12 f_2=184
			0.09	f_3=79 f_{4+}=2	f_3=10 f_{4+}=0	f_3=4 f_{4+}=0
300	6	0.96	0.04 0.01	f_1=1 f_2=67	f_1=1 f_2=159	f_1=1 f_2=180
			0.02	f_3=132 f_{4+}=0	f_3=40 f_{4+}=0	f_3=19 f_{4+}=0

sizes and the true lag. Comparison of our m^2 values in Table 4 with those in Table 2 of Xia, Tong, Li, and Zhu (2002) shows that the refined minimum average variance estimation method of Xia, Tong, Li, and Zhu (2002) performed better than our estimation method. However, this is to be expected because, unlike our method, their method focuses on estimation of dimensions in the mean function. On the other hand, their method may not perform well for our Model 1, where there is a dimension in the error term. Table 4 shows that $\hat{d}_p$ using the 0-threshold correctly estimated the true dimension, $d = 3$, about 51% to 68% of the time for all sample sizes and when $p = 6$. Note that the performance of $\hat{d}_p$ with 0-threshold was slightly better than that of Xia, Tong, Li, and Zhu (2002) for $n = 100$, but the latter method performed better than ours for $n = 200$ and 300, as shown in their Table 2. Table 4 also shows that wrong specification of lag adversely affected the performance of $\hat{d}_p$, resulting in severe underestimation for all the three thresholds. On the other hand, even when the lag was correctly specified, the 0.05- and 0.01-threshold did not perform well for any sample size.

To infer p using *Shoulder Plot*, we set $d = 3$ and computed $\hat{\Psi}_n(\hat{\mathrm{h}}_{p,3})$ with $p = 4, 5, 6, 7$ for 100 simulated data sets, each with $n = 300$. For 96% of the data sets, the *Shoulder Plot* correctly indicated that $\hat{p} = 6$.

Results for models 2, 3 and 4 above seem to suggest that when $d = 1$ or 2, the 0.05-threshold performs better, but when $d = 3$, the 0-threshold performs better. Thus, we suggest use of 0.05-threshold to estimate d when we prefer a smaller dimension; otherwise, use 0-threshold. Our simulations also indicate that the *Shoulder plot* is an effective way of identifying the unknown lag p.

Table 5. Wolf yearly sunspot data: $\hat{\Psi}_n(\hat{\mathrm{h}}_{p,d})$ values for $p = 2, \ldots, 12$, and $d = 1, 2$ and 3. For each p, $\hat{d}_p$ determined by 0.05-threshold is denoted by *.

p	d=1	d=2	d=3
2	0.7229	0.7533*	N/A
3	0.7259	0.7578*	0.7185
4	0.7258	0.7593*	0.7581
5	0.7255	0.7682*	0.7768
6	0.7377	0.7744*	0.7825
7	0.7590	0.7999*	0.8049
8	0.7634	0.8011*	0.8069
9	0.7846	0.8177*	0.8197
10	0.7843*	0.8146	0.8235
11	0.7837	0.8186*	0.8256
12	0.7818	0.8201*	0.8153

Wolf Yearly Sunspot Data: This data set has 307 observations for the years 1700 to 2006 and it has been extensively studied by many authors using various linear and nonlinear models; see, for example, Yule (1927), Bartlett (1950), Whittle (1954), Brillinger and Rosenblatt (1967), Xia, Tong, and Li (1999), and Wei (2006). We compare the results of our data analysis with those in Wei (2006), who fits the following three autoregressive models: (i) AR(2) , (ii) AR(9) , and (iii) AR(1, 2, 9). Based on estimated error variance, Wei (2006) concludes that both models (ii) and (iii) are adequate for the data. For our analysis, we considered the sunspot numbers for the years 1700 to 1991 with $n = 292$. Our process began with estimation of d and p, followed by estimation of time series central subspace. We then built a model, based on which we computed the sunspot number forecasts for the remaining fifteen years: 1992 to 2006.

To estimate d, we set $p = 2, \ldots, 12$ and computed $\hat{\Psi}_n(\hat{\mathrm{h}}_{p,d})$ for $d = 1, 2$ and 3. Table 5 lists the $\hat{\Psi}_n(\hat{\mathrm{h}}_{p,d})$ values for each p and d, except for the trivial case $p = 1$. Using (3.3) with 0.05-threshold, we concluded that $\hat{d}_p = 2$ for all $2 \le p \le 12$, except when $p = 10$, for which $\hat{d}_p = 1$. Note that $\hat{d}_p$ for each p is indicated by an asterisk in Table 5. Since $\hat{d}_p = 2$ uniformly for all p, except for $p = 10$, we decided to use $d = 2$. The *Shoulder Plot* in Figure 1c, based on $\hat{\Psi}_n(\hat{\mathrm{h}}_{p,2})$ values for $p = 2, \ldots, 12$ (2nd column in Table 5), seems to indicate that the largest value is at $p = 12$, but the *Shoulder* is at $p = 9$, where the value is essentially very close to the largest value. That is, the subsequent values are about the same or less than $\hat{\Psi}_n(\hat{\mathrm{h}}_{9,2})$. In fact, our determination of $p = 9$ agrees with other approaches; for example, the *R* software with command '*sunspot.ar <- ar(sunspot.year)*' determines lag $p = 9$.

Based on above determinations, we obtained an estimate of a 9×2 basis matrix $\hat{\Phi}_{\rm d} = (\hat{\Phi}_1, \hat{\Phi}_2)$. Subsequently, we used these estimates and the following trial-and-error approach involving plots to build a time series model. First, we examined 2-dimensional plots of x_t vs $\hat{\Phi}_1^T X_{t-1}$ and x_t vs $\hat{\Phi}_2^T X_{t-1}$, and the 3-dimensional plot of x_t vs $\hat{\Phi}_1^T X_{t-1}$ and $\hat{\Phi}_2^T X_{t-1}$, which revealed a linear pattern. This motivated us to regress x_t on the predictors $\hat{\Phi}_1^T X_{t-1}$ and $\hat{\Phi}_2^T X_{t-1}$ with coefficient estimates significantly different from zero based on t-tests. We then created 2-dimensional and 3-dimensional plots of the resulting residuals vs $\hat{\Phi}_1^T X_{t-1}$ and $\hat{\Phi}_2^T X_{t-1}$. Nonparametric smoothing applied to each of the 2-dimensional residual plots showed cyclical patterns, suggesting that cosine functions might be reasonable approximations. We then fine-tuned the cosine functions so as to match the smoothed curves, determining the following nonlinear terms: $\cos\{(\pi/2)\,\hat{\Phi}_1^T X_{t-1} + \pi\}$ and $\cos\{(\pi/2)\,\hat{\Phi}_2^T X_{t-1} - (\pi/4)\}$. With these linear and nonlinear terms, we fitted a variety of models with and without interactions, and eventually arrived at the following "best" model:

$$\hat{x}_t = 0.48 - 1.25\,\hat{\Phi}_1^T X_{t-1} + 0.75\,\hat{\Phi}_2^T X_{t-1} + 0.50\cos\Big(\frac{\pi}{2}\,\hat{\Phi}_1^T X_{t-1} + \pi\Big) + 0.25\cos\Big(\frac{\pi}{2}\,\hat{\Phi}_2^T X_{t-1} - \frac{\pi}{4}\Big).$$

All the coefficients in the above model were found to be significant with standard error of estimates 0.18, 0.05, 0.03, 0.09, and 0.08, respectively.

To compare our model with models (ii) and (iii) above, we calculated the forecasts for the sunspot numbers for the years 1992 to 2006 and computed the Mean Square Relative Error $= k^{-1}\sum_{t=1}^{k}(z_t - \hat{z}_t)^2/z_t$, where z_t is the observed sunspot number, $\hat{z}_t$ is its forecast value and k is the number of (future) observations. The mean square relative error for our model, models (ii), and (iii) are 2.6714, 6.0688 and 5.8319, respectively. Our model produces an mean square relative error which is less than half those of Wei's models. An overlay plot of forecast values from each of these models and the observed sunspot numbers for the years 1992 to 2006 is given in Figure 1b. This data analysis shows that our approach can lead us to a time series model that outperforms some of the other available models for the sunspot data. Finally, as in Ghaddar and Tong (1981), we used our model to obtain forecast of sunspot numbers for 2007 to 2017, and these are given in Table 6.

5. Discussion

The literature has seen a proliferation of parametric and nonparametric methods for time series analysis. Nonetheless, few use a sufficient dimension reduction approach in time series analysis. Well-known sufficient dimension reduction methods in regression, such as SIR, impose stringent requirements that

Table 6. Wolf yearly sunspot data: Forecasts for the years 2007 to 2017 based on our model.

Year	Sunspot Number
2007	17.1352
2008	24.2818
2009	52.6491
2010	77.5360
2011	93.0935
2012	95.1516
2013	82.9389
2014	64.6360
2015	50.1982
2016	35.8053
2017	21.5584

severely limit their use in time series analysis. In this article, we develop a new theory of sufficient dimension reduction in time series based on the notion of the time series central subspace; this provides an initial phase when an adequate parsimoniously parameterized time series model is not yet available. Although the notion of the time series central subspace bears similarity to that of the central subspace in regression, an important difference is that a time series central subspace implicitly depends on lag p, which is usually unknown in practice and requires estimation. Our definition of time series central subspace is general enough to include many linear and nonlinear time series models.

In our method, we suggest use of either 0-threshold or 0.05-threshold to estimate d. The choice of threshold certainly affects the value of $\hat{d}_p$, which increases as $\tau_{p,n}$ decreases. The use of *Shoulder Plot* to estimate p seems quite informative in our analysis because of its visual appeal and simplicity.

Generally, in nonparametric methods (density estimation as well as nonparametric regression), the choice of kernel is less critical than the choice of bandwidth (Härdle (1990); and Scott (1992, p.133)). Here, the focus is on obtaining directions, and hence the nonparametric estimation using kernels is an intermediate step and not the primary focus per se. Intuitively, even if the estimated densities are not as accurate as required in usual density estimation, but as long as the estimated shapes of densities are similar to the true ones, the estimated directions should not be severely affected. In theoretical studies, multivariate kernels are used to prove theorems. However, in practice, the product kernel is often recommended; see Scott (1992, p.152) and Silverman (1986). In addition, multivariate Gaussian kernel reduces to a product Gaussian kernel if the variables are standardized. We found Gaussian kernels to be adequate because they performed well in the simulations and data analysis reported on in Section 4. Additional

simulations (not reported here) indicate that the choice of bandwidth has little effect in our context.

Overall, the theory of dimension reduction in time series poses many challenges, but a variety of encouraging results presented through simulations seem to suggest that our method has the potential for providing a viable and meaningful alternative to traditional time series analysis. In fact, the performance of our method for Wolf yearly sunspot suggests that it might be quite useful in time series analysis.

In this article, we consider the estimation of a time series central subspace and other related issues. However, new dimension reduction methods such as the central mean subspace (Cook and Li (2002)), which focuses on the mean function of time series, has yet to be developed. Research along these lines is currently underway.

Acknowledgement

We would like to thank the Editors, an associate editor and two referees whose suggestions led to a greatly improved paper. Sriram was supported in part by a NSA grant H982300910027, and Yin was supported in part by a NSF grant DMS-0806120. This study was supported in part by resources provided by the University of Georgia Research Computing Center, a partnership between the Office of the Vice President for Research and the Office of the Chief Information Officer.

Appendix: Assumptions and Proofs

Proof of Proposition 2. (i) If $\mathcal{S}(\mathrm{h}_1) \subseteq \mathcal{S}(\mathrm{h}_2)$, then it is possible to write $\mathrm{h}_1 = \mathrm{h}_2\mathrm{K}$ for some matrix K. Therefore, with simple algebra,

$$\Psi(\mathrm{h}_2) - \Psi(\mathrm{h}_1) = \mathrm{E}\left[\mathrm{E}_{x_t|\mathrm{h}_2^T X_{t-1}}\left\{\log \frac{p(x_t|\mathrm{h}_2^T X_{t-1})}{p(x_t|\mathrm{K}^T\mathrm{h}_2^T X_{t-1})}\right\}\right] \geq 0.$$

The inequality follows from a result on page 14 of Kullback (1959). Suppose $\mathcal{S}(\mathrm{h}_1) = \mathcal{S}(\mathrm{h}_2)$. Then K can be a nonsingular square matrix in the above argument, and the inequality becomes an equality.

(ii) As in (i), we can write

$$\Psi(I_p) - \Psi(\mathrm{h}_1) = E\left[E_{x_t|X_{t-1}}\left\{\log \frac{p(x_t|X_{t-1})}{p(x_t|\mathrm{h}_1^T X_{t-1})}\right\}\right] \geq 0$$

with equality if and only if $p(x_t|X_{t-1}) = p(x_t|\mathrm{h}_1^T X_{t-1})$. But, $x_t \perp\!\!\!\perp X_{t-1}|\mathrm{h}_1^T X_{t-1}$ if and only if $p(x_t|X_{t-1}) = p(x_t|\mathrm{h}_1^T X_{t-1})$. Hence the first assertion in (ii). From

this and the definition of the time series central subspace $\mathcal{S}_{x_t|X_{t-1}}(\Phi_{\mathrm{d}})$, for which the property $x_t \perp\!\!\!\perp X_{t-1}|\Phi_{\mathrm{d}}^T X_{t-1}$ holds, we have that $\Psi(I_p) = \Psi(\Phi_{\mathrm{d}}) \geq \Psi(\mathrm{h}_d)$. Now, if $\Psi(\Phi_{\mathrm{d}}) = \Psi(\mathrm{h}_d)$, then $\mathcal{S}_{x_t|X_{t-1}}(\Phi_{\mathrm{d}}) = \mathcal{S}(\mathrm{h}_d)$ by the uniqueness of the time series central subspace. The other way conclusion follows from (i).

(iii) Since

$$\Psi(\mathrm{h}_1) = \mathrm{E}\left\{ \log \frac{p(x_t, \mathrm{h}_1^T X_{t-1})}{p(x_t)p(\mathrm{h}_1^T X_{t-1})} \right\},$$

$\Psi(\mathrm{h}_1) \geq 0$ by a result of Kullback (1959, p.14).

Let $\alpha = \arg\max_{\mathrm{h}_1} \Psi(\mathrm{h}_1)$, and, for any $p \times (q_2 - q_1)$ matrix β, by using Kullback's result (1959, p.14), we have

$$\begin{aligned} &\max \Psi(\mathrm{h}_2) - \max \Psi(\mathrm{h}_1) \geq \Psi(\alpha, \beta) - \Psi(\alpha) \\ &\quad = \mathrm{E}_{\alpha^T X_{t-1}, \beta^T X_{t-1}} \left[\mathrm{E}_{x_t|\alpha^T X_{t-1}, \beta^T X_{t-1}} \left\{ \log \frac{p(x_t|\alpha^T X_{t-1}, \beta^T X_{t-1})}{p(x_t|\alpha^T X_{t-1})} \right\} \right] \geq 0. \end{aligned}$$

However, equality cannot hold unless $p(x_t|\alpha^T X_{t-1}, \beta^T X_{t-1}) = p(x_t|\alpha^T X_{t-1})$ for any β; that is $x_t \perp\!\!\!\perp \beta^T X_{t-1}|\alpha^T X_{t-1}$, for any β, and hence $x_t \perp\!\!\!\perp X_{t-1}|\alpha^T X_{t-1}$. Thus, $\Psi(I_p) = \Psi(\alpha)$ by the definition of the central subspace. This produces the contradiction that $d \leq (d-1)$.

The following assumptions apply to Lemma 1 stated below and the Theorem 1 stated in Section 3.4.

Bounded variation. Define the $\varnothing_y^{\mathrm{x}}$-operator for functions $g : \mathbb{R}^k \to \mathbb{R}^1$ and $\mathrm{x}, y \in \mathbb{R}^k$ by

$$\varnothing_y^{\mathrm{x}} g = \sum_{(\epsilon_1, \ldots, \epsilon_k) \in \{0,1\}^k} (-1)^{\sum_{j=1}^k \epsilon_j} g\Big\{ \epsilon_1 x_1 + (1-\epsilon_1) y_1 + \cdots + \epsilon_k x_k + (1-\epsilon_k) y_k \Big\}$$

(Reyni (1962)). Suppose $\mathcal{P}$ is the set of all finite partitions of $\mathbb{R}^k$ into rectangles and $\varnothing_y^{\mathrm{x}} g$ is expressed by the corresponding limit if some components of x and y are positive or negative infinity. Then g is said to be of bounded variation if $\sup\{g(s), s \in \mathcal{P}\} < \infty$, where $g(s)$ is defined for $s = \sum_{i=1}^l [x_i, y_i)$ by $\sum_{i=1}^l |\varnothing_{y_i}^{x_i} g|$.

Assumptions A1. As in Sen (1974), $\{\mathrm{h}^T X_t, -\infty < t < \infty\}$ is a stationary ϕ-mixing sequence for any $p \times d$ matrix h, defined on a probability space $(\Omega, \mathscr{A}, P)$ with each $\mathrm{h}^T X_t$ having a continuous distribution F. That is, if $\mathscr{M}_{-\infty}^k$ and $\mathscr{M}_{k+n}^{\infty}$ are σ-fields generated by $\{\mathrm{h}^T X_t, t \leq k\}$ and $\{\mathrm{h}^T X_t, t \geq k+n\}$, respectively, and if $A \in \mathscr{M}_{-\infty}^k$ and $B \in \mathscr{M}_{k+n}^{\infty}$, then for all k, $n \geq 0$, and $\phi_n \geq 0$, $|P(A \cap B) - P(A)P(B)| \leq \phi_n P(A)$, where $\{\phi_n\}$ is independent of h, $\{\phi_n\} \downarrow$ in n and $\lim_{n\to\infty} \phi_n = 0$.

Assumptions A2. For each $m \geq 0$, $A_m(\phi) = \sum_{n=1}^{\infty}(n+1)^m \phi_n^{1/2}$, for $\{\phi_n\}$ in Assumption A1, satisfies $A_m(\phi) < \infty$ for some $m \geq 1$.

Now, under assumptions A1 and A2, the conclusion of Theorem 3.2 of Sen (1974) holds for $\mathrm{h}^T X_t$ with the upper bound $C_\phi \lambda^{-2(m+1)}$ (for $\lambda \geq 1$), where C_ϕ ($< \infty$) only depends on $\{\phi_n\}$. Using this upper bound and the arguments in the proof of Theorem 1(b) of Rüschendorf (1977), one can prove the following lemma. (A similar result as Lemma 1 was established by Masry (1996) over compact subsets of $\mathcal{R}^d$ for the regression function and its partial derivatives for strongly mixing processes).

Lemma 1. *Assume* A1 *and* A2*, and that* $\sum_{n=1}^{\infty}\{\gamma/(\sqrt{n}a_n^k)\}^{2(m+1)} < \infty$ *for all* $\gamma \in \mathbb{R}_+$*, where* $\{a_n\}$ *is a sequence of bandwidths of the kernel density estimator* f_n *defined in Section* 3.4*. Suppose the kernel* K *in* f_n *is of bounded variation. Also, let the density functions satisfy the following conditions:* $p(x_t)$ *is uniformly continuous,* $p(h^T X_{t-1})$ *is uniformly continuous in* h *and* X_{t-1}*, and* $p(h^T X_{t-1}, x_t)$ *is uniformly continuous in* h*,* X_{t-1}*, and* x_t*, where* $h^T h = I_d$*. Then the following results hold with probability one as* $n \to \infty$*:*

$$\sup_{x_t \in \mathbb{R}^1} |f_n(x_t) - p(x_t)| \to 0,$$

$$\sup_{h \in \mathbb{R}^{p\times d}, X_{t-1} \in \mathbb{R}^p} |f_n(h^T X_{t-1}) - p(h^T X_{t-1})| \to 0,$$

$$\sup_{h \in \mathbb{R}^{p\times d}, X_{t-1} \in \mathbb{R}^p, x_t \in \mathbb{R}^1} |f_n(h^T X_{t-1}, x_t) - p(h^T X_{t-1}, x_t)| \to 0.$$

Proof of Theorem 1. Note that the constraint $\mathrm{h}^T\mathrm{h} = I_d$ does not guarantee that a matrix maximizing the objective function $\Psi(\mathrm{h})$ is unique, but the subspace corresponding to it is unique. Hence, for identifiability, we may replace any basis matrix that maximizes the objective function by its orthogonal projection matrix, which is unique. Thus, without loss of generality and for the simplicity of our proof, we assume that the matrix solution is unique.

If $\hat{\Phi}_\mathrm{n}$ does not converge to Φ_d with probability 1, there is a subsequence which is still indexed by n, and a $p \times d$ matrix Φ_0 satisfying $\Phi_0^T \Phi_0 = I_d$ and $\Phi_0 \neq \Phi_\mathrm{d}$, such that $\hat{\Phi}_\mathrm{n} \to \Phi_0$. Thus, for any $\epsilon > 0$ and large enough n, we have

$$f_n(x_t) = p(x_t) + \delta_{1,t}, \tag{A.1}$$

$$f_n(\hat{\Phi}_\mathrm{n}^T X_{t-1}) = p(\hat{\Phi}_\mathrm{n}^T X_{t-1}) + \eta_{2,t} = p(\Phi_0^T X_{t-1}) + \delta_{2,t}, \tag{A.2}$$

$$f_n(\hat{\Phi}_\mathrm{n}^T X_{t-1}, x_t) = p(\hat{\Phi}_\mathrm{n}^T X_{t-1}, x_t) + \eta_{3,t} = p(\Phi_0^T X_{t-1}, x_t) + \delta_{3,t}, \tag{A.3}$$

such that $|\delta_{k,t}| < \epsilon$ for all t and $k = 1, 2, 3$. Note that (A.1) and the first equalities in (A.2) and (A.3) follow from the conclusions of Lemma 1, whereas the uniform

continuity conditions in Lemma 1 lead to the second equalities in equations (A.2) and (A.3). From these, we have

$$\log f_n(x_t) = \log p(x_t) + \log\left\{1 + \frac{\delta_{1,t}}{p(x_t)}\right\},$$
$$\log f_n(\hat{\Phi}_{\mathrm{n}}^T X_{t-1}) = \log p(\Phi_0^T X_{t-1}) + \log\left\{1 + \frac{\delta_{2,t}}{p(\Phi_0^T X_{t-1})}\right\},$$
$$\log f_n(\hat{\Phi}_{\mathrm{n}}^T X_{t-1}, x_t) = \log p(\Phi_0^T X_{t-1}, x_t) + \log\left\{1 + \frac{\delta_{3,t}}{p(\Phi_0^T X_{t-1}, x_t)}\right\},$$

Therefore, by the definition of κ_ι in Section 3.4, for large n, on this set we have $p(x_t) > \iota/2, p(h^T X_{t-1}) > \iota/2, p(h^T X_{t-1}, x_t) > \iota/2$. Since $(\epsilon/\iota) \to 0$, for $\hat{\Psi}_n^\iota$ defined in the Theorem 1 we have that

$$\hat{\Psi}_n^\iota(\hat{\Phi}_n) = \frac{1}{n}\sum_{t=1}^{n} J(t \in \kappa_\iota) \log \frac{p(\Phi_0^T X_{t-1}, x_t)}{p(x_t)p(\Phi_0^T X_{t-1})} + o(1) = \bar{\Psi}_n^\iota(\Phi_0) + o(1).$$

But,

$$\begin{aligned}\bar{\Psi}_n^\iota(\Phi_0) - \Psi(\Phi_0) &= \left\{\frac{1}{n}\sum_{t=1}^{n} \log \frac{p(\Phi_0^T X_{t-1}, x_t)}{p(x_t)p(\Phi_0^T X_{t-1})} - \Psi(\Phi_0)\right\} \\ &\quad - \frac{1}{n}\sum_{t=1}^{n} J(t \in \kappa_\iota^c) \log \frac{p(\Phi_0^T X_{t-1}, x_t)}{p(x_t)p(\Phi_0^T X_{t-1})} \\ &= \tau_1 - \tau_2.\end{aligned}$$

From $(n_\iota/n) \to 0$ and the Ergodic Theorem, both τ_1 and τ_2 tend to 0 with probability one as $n \to \infty$. Hence, $\lim_{n\to\infty} \hat{\Psi}_n^\iota(\hat{\Phi}_{\mathrm{n}}) = \Psi(\Phi_0)$ with probability one. Since $\hat{\Psi}_n^\iota(\hat{\Phi}_{\mathrm{n}}) \geq \hat{\Psi}_n^\iota(\Phi_{\mathrm{d}})$ by definition, taking limit on both sides we get $\Psi(\Phi_0) \geq \Psi(\Phi_{\mathrm{d}})$. On the other hand, by the definition of Φ_{d}, $\Psi(\Phi_0) \leq \Psi(\Phi_{\mathrm{d}})$, and therefore $\Psi(\Phi_0) = \Psi(\Phi_{\mathrm{d}})$. Due to uniqueness $\Phi_0 = \Phi_{\mathrm{d}}$, a contradiction. Therefore, $\hat{\Phi}_{\mathrm{n}} \to \Phi_{\mathrm{d}}$ with probability 1.

Lemma 2. *Assume the conditions of Lemma* 1 *and Theorem* 1*. For each fixed* p *and* k, $\max_{h_{p,k}} \hat{\Psi}_n^\iota(h_{p,k}) \to \max_{h_{p,k}} \Psi(h_{p,k})$, *as* $n \to \infty$.

Proof of Lemma 2. For simplicity, we assume that $\arg\max_{\mathrm{h}_{p,k}} \Psi(\mathrm{h}_{p,k})$ is unique. By the arguments in the proof of Theorem 1 and its conclusion, we have that $\lim_{n\to\infty} \max_{\mathrm{h}_{p,k}} \hat{\Psi}_n^\iota(\mathrm{h}_{p,k}) = \max_{\mathrm{h}_{p,k}} \Psi(\mathrm{h}_{p,k})$, with probability one.

Proof of Theorem 2. Let $c_k = \max_{\mathrm{h}_{p,(k+1)}} \Psi(\mathrm{h}_{p,(k+1)}) - \max_{\mathrm{h}_{p,k}} \Psi(\mathrm{h}_{p,k})$. By Proposition 2 (iii), it follows that $c_k > 0$ if $k < d$, and $c_k = 0$ if $k \geq d$. Hence, $d = \min\{k(\leq (p-1)) : c_k = 0\}$. Moreover, for each k, by Lemma 2 we have $\hat{c}_k^\iota \to c_k$ as $n \to \infty$. Recall that $\hat{d}_p^\iota = \min\{k(\leq (p-1)) : \hat{c}_k^\iota \leq \tau_{p,n}\}$. Since $\tau_{p,n} \to 0$ as $n \to \infty$, it follows that $\hat{d}_p^\iota \to d$ as $n \to \infty$, with probability one.

References

Barlett, M. S. (1950). Periodogram analysis and continuous spectra. *Biometrika* **37**, 1-16.

Becker, C. and Fried, R. (2003). Sliced inverse regression for high-dimensional time series. *Exploratory Data Analysis in Empirical Research: Proceedings of the* 25*th Annual Conference of the Gesellschaft für Klassifickation*, University of Munich, 3-11.

Brillinger, D. R. and Rosenblatt, M. (1967). Asymptotic theory of kth order spectra. *Spectral Analysis of Time Series* (Ed. B. Harris), 153-188. John Wiley, New York.

Brockwell, P. J. and Davis, R. A. (1996). *Introduction to Time Series and Forecasting*. Springer-Verlag, New York.

Chan, K. S., Petrucelli, J. D., Tong, H. and Woolford, S. W. (1985). A multiple-threshold AR(1) model. *J. Appl. Probab.* **22**, 267-279.

Cook, R. D. (1994). On the interpretation of regression plots. *J. Amer. Statist. Assoc.* **89**, 177-190.

Cook, R. D. (1998a). *Regression Graphics: Ideas for Studying Regressions through Graphics*. Wiley, New York.

Cook, R. D. (1998b). Principal Hessian directions revisited (with discussion). *J. Amer. Statist. Assoc.* **93**, 84-100.

Cook, R. D. and Li, B. (2002). Dimension reduction for the conditional mean in regression. *Ann. Statist.* **30**, 455-474.

Cook, R. D. and Weisberg, S. (1991). Discussion of 'sliced inverse regression' by K. C. Li. *J. Amer. Statist. Assoc.* **86**, 328-332.

Cover, T. M. and Thomas, J. A. (1991). *Elements of Information Theory*. John Wiley, New York.

De Gooijer, J. and Zerom, D. (2003). On conditional density estimation. *Statist. Neerlandica* **57**, 159-176.

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica* **50**, 987-1007.

Fan, J. and Gijbels, I. (1995). Data-driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adaptation. *J. Roy. Statist. Soc. Ser. B* **57**, 371-394.

Fan, J., Heckman, M. E. and Wand, M. P. (1995). Local polynomial kernel regression for generalized linear models and quasi-likelihood functions. *J. Amer. Statist. Assoc.* **90**, 141-150.

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

Fan, J., Yao, Q. and Tong, H. (1996). Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems. *Biometrika* **83**, 189-206.

Ghaddar, D. K. and Tong, H. (1981). Data transformation and self-exciting threshold autoregression. *Appl. Statist.* **30**, 238-248.

Gill, P., Murray, W. and Wright, M. H. (1981). *Practical Optimization*. Academic Press, New York.

Hall, P. and Yao, Q. (2005). Approximating conditional distribution functions using dimension reduction. *Ann. Statist.* **33**, 1404-1421.

Härdle, W. (1990). *Applied Nonparametric Regression*. Cambridge University Press.

Härdle, W. and Stoker, T. M. (1989). Investigating smooth multiple regression by the method of average derivatives. *J. Amer. Statist. Assoc.* **84**, 986-995.

Hotelling, H. (1936). Relations between two sets of variates. *Biometrika* **28**, 321-377.

Hyndman, R. J. and Yao, Q. (2002). Nonparametric estimation and symmetry tests for conditional density functions. *Nonparam. Statist.* **14**, 259-278.

Kullback, S. (1959). *Information Theory and Statistics*. Wiley, New York.

Li, K. C. (1991). Sliced inverse regression for dimension reduction (with discussion). *J. Amer. Statist. Assoc.* **86**, 316-342.

Li, K. C. (1992). On principal Hessian directions for data visualization and dimension reduction: another application of Stein's Lemma. *Ann. Statist.* **87**, 1025-1039.

Li, K. and Shedden, K. (2002). Identification of shared components in large ensembles of time series using dimension reduction. *J. Amer. Statist. Assoc.* **97**, 759-765.

Li, B., Zha, H. and Chiaromonte, C. (2005). Contour regression: a general approach to dimension reduction. *Ann. Statist.* **33**, 1580-1616.

Masry, E. (1996). Multivariate local polynomial regression for time series: uniform strong consistency and rates. *J. Time Ser. Anal.* **17**, 571-599.

Ng, S. and Perron, P. (2005). A note on selection of time series models. *Oxford Bulletin of Economics and Statistics* **67**, 115-134.

Reyni, A. (1962). *Wahrscheinlichkeitsrechnung*. VEB, Deutscher Verlag der Wisschenschaften, Berlin.

Rüschendorf, L. (1977). Consistency of estimators for multivariate density functions and for the mode. *Sankhyā* **39**, 243-250.

Schott, J. R. (1994). Determining the dimensionality in sliced inverse regression. *J. Amer. Statist. Assoc.* **89**, 141-148.

Scott, D. W. (1992). *Multivariate Density Estimation: Theory, Practice, and Visualization*. John Wiley, New York.

Sen, P. K. (1974). Weak convergence of multidimensional empirical processes for stationary ϕ-mixing processes. *Ann. Probab.* **2**, 147-154.

Shummway, R. H. and Stoffer, D. S. (2000). *Time Series Analysis and Its Applications*. Springer-Verlag, New York.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London and New York:

Tong, H. and Lim, K. S. (1980). Threshold autoregression, limit cycles and cyclical data. *J. Roy. Statist. Soc. Ser. B* **42**, 245-292.

Tsay, R. S. (2005). *Analysis of Financial Time Series*. John Wiley, New York.

Wei, W. W. S. (2006). *Time Series Analysis: Univariate and Multivariate Methods*. Pearson & Addison Wesley, Boston.

Whittle, P. (1954). A statistical investigation of sunspot observations with special reference to H.Alfven's sunspot model. *Astrophys. J.* **120**, 251-260.

Woo, Mi-Ja and Sriram, T. N. (2006). Robust estimation of mixture complexity. *J. Amer. Statist. Assoc.* **101**, 1475-1486.

Xia, Y. and Li, W. K. (1999). On the estimation and testing of functional coefficient linear models. *Statist. Sinica* **9**, 735-757.

Xia, Y., Li, W. K. and Tong, H. (2007). Threshold variable selection using nonparametric methods. *Statist. Sinica* **17**, 265-287.

Xia, Y., Tong, H. and Li, W. K. (1999). On extended partially linear single-index models. *Biometrika* **86**, 831-842.

Xia, Y., Tong, H. and Li, W. K. (2002). Single-index volatility models and estimation. *Statist. Sinica* **12**, 785-799.

Xia, Y., Tong, H., Li, W. K. and Zhu, L. X. (2002). An adaptive estimation of dimension reduction. *J. Roy. Statist. Soc. Ser. B* **64**, 363-410.

Ye, Z. and Weiss, R. E. (2003). Using the bootstrap to select one of a new class of dimension reduction methods. *J. Amer. Statist. Assoc.* **98**, 968-979.

Yin, X. and Cook, R. D. (2005). Direction estimation in single-index regressions. *Biometrika* **92**, 371-384.

Yin, X., Li, B. and Cook, R. D. (2008). Successive direction extraction from estimating the central subspace in a multiple-index regression. *J. Multivariate Anal.* **99**, 1733-1757.

Yule, G. U. (1927). On a method of investigating periodicities in disturbed series with special reference to Wolfer's sunspot numbers. *Phil. Trans. R. Soc. Lond. A* **226**, 267-298.

Department of Mathematics, College of Charleston, Charleston, SC 29424, U.S.A.

E-mail: parkj@cofc.edu

Department of Statistics, University of Georgia, Athens, GA 30602, U.S.A.

E-mail: tn@stat.uga.edu

Department of Statistics, University of Georgia, Athens, GA 30602, U.S.A.

E-mail: xryin@stat.uga.edu

(Received May 2008; accepted January 2009)