

Single-Index Additive Vector Autoregressive Time Series Models

YEHUA LI

Department of Statistics, University of Georgia

MARC G. GENTON

Department of Statistics, Texas A&M University

ABSTRACT. We study a new class of nonlinear autoregressive models for vector time series, where the current vector depends on single-indexes defined on the past lags and the effects of different lags have an additive form. A sufficient condition is provided for stationarity of such models. We also study estimation of the proposed model using P-splines, hypothesis testing, asymptotics, selection of the order of the autoregression and of the smoothing parameters and nonlinear forecasting. We perform simulation experiments to evaluate our model in various settings. We illustrate our methodology on a climate data set and show that our model provides more accurate yearly forecasts of the El Niño phenomenon, the unusual warming of water in the Pacific Ocean.

Key words: autoregressive, climate, multivariate, nonlinear, penalized spline, prediction, time series

1. Introduction

In many time series problems, two or more random variables evolve over time. These variables not only have relationships with each other, but also depend on previous values in time. Although in many situations we are only interested in predicting one variable in the future, we need to consider all of these variables as a vector time series. One of the most important models for prediction of vector time series is the linear vector autoregressive (VAR) model, i.e. the present value in the vector time series has a linear relationship with the past. However, there is a need for nonlinear models (see Fan & Yao, 2003).

There is a vast amount of literature on nonlinear autoregressive models and we only mention the following papers that are particularly relevant. Chen & Tsay (1993a,b) proposed functional coefficient autoregressive models and nonlinear additive ARX models. Huang & Yang (2004) proposed an order selection procedure for nonlinear additive autoregressive models using a Bayesian information criterion (BIC). Huang & Shen (2004) proposed a polynomial spline approach for functional coefficient autoregressive models. All these methods were proposed for univariate time series and there has been much less development in nonlinear VAR models. The most relevant work that we are aware of for multivariate time series is Härdle *et al.* (1998), where the authors modelled the autoregression as a high-dimensional non-parametric function estimated with a local linear estimator. However, if the dimension of the vector time series is high, or the model involves many lags in the past, this method will suffer from the ‘curse of dimensionality’.

To circumvent the curse of dimensionality, many recent ideas have focused on the single-index model (see Carroll *et al.*, 1997; Yu & Ruppert, 2002). Xia & Härdle (2006) studied single-index models in time series using the minimum average variance estimation (MAVE) method (Xia *et al.*, 2002), whereas Wang & Yang (2009) studied estimation of single-index models using B-splines. However, again, most of these methods are designed for univariate times series.

In this paper, we aim at developing a new class of nonlinear VAR models. It has a dimension reduction flavour in that the current vector only depends on single-indexes on the past vectors. Moreover, the nonlinear relationship with the past lags has an additive model structure (Hastie & Tibshirani, 1990) in order to avoid the curse of dimensionality. All of the non-parametric functions in the model are univariate. We provide an overall treatment of estimation, hypothesis testing and nonlinear forecasting under the proposed model.

Our work is motivated by the challenging problem of short-term prediction of climate, i.e. predicting the evolution of climate on timescales of a season to a year. We use historical data to fit a multivariate time series model and perform short-term prediction. Currently, only linear VAR models are used in practice and we want to investigate the use of nonlinear models. More specifically, we consider monthly data over the tropical Pacific Ocean of sea surface temperature, sea surface height and wind-stress during 1960–2004. Our goal is to predict the El Niño phenomenon, the unusual warming of water in the Pacific Ocean. Indeed, El Niño denotes a disruption of the ocean–atmosphere system and is believed to have important consequences for climate around the world.

Our paper is organized in the following way. We introduce our single-index additive VAR model in section 2 and provide a sufficient condition for stationarity of the model. In section 3, we propose a backfitting procedure for estimating the unknown parts of the model. Specifically, we estimate the non-parametric link functions using penalized splines. Other issues in estimation are discussed as well, e.g. selecting the order of the autoregression and the smoothing parameters. A nonlinear forecasting procedure is also described. Some asymptotic results are provided in section 4, as well as an inference procedure based on the asymptotic theory in order to test linearity of the link functions. Various simulation results are provided in section 5 and the analysis of the aforementioned climate data set is presented in section 6. We conclude the paper with some Supporting remarks in section 7. Technical proofs are provided in the Appendix and in the online Supporting Information.

2. Model

2.1. Definition and assumptions

We consider vector time series data described at time t by a d -dimensional vector $Y_t = (Y_{1t}, \dots, Y_{dt})^T$. We propose the following single-index additive VAR model of order p , SIAVAR(p):

$$Y_t = \sum_{j=1}^p g_j(A_j Y_{t-j}) + \epsilon_t, \tag{1}$$

where $A_j = [\alpha_{1j}, \dots, \alpha_{dj}]^T$, $g_j(A_j Y_{t-j}) = [g_{1j}(\alpha_{1j}^T Y_{t-j}), \dots, g_{dj}(\alpha_{dj}^T Y_{t-j})]^T$, g_{ij} are unknown univariate link functions, α_{ij} are d -dimensional index weight vectors and ϵ_t is the error term with $E(\epsilon_t | Y_{t'}, t' < t) = \mathbf{0}$ and $\text{cov}(\epsilon_t | Y_{t'}, t' < t) = V_\epsilon$. Suppose the data we observe are $\{Y_t; t = 1, 2, \dots, n\}$, where n is the sample size. By looking at the i th component in model (1), we can see that Y_{it} is modelled as an additive model on single-indexes defined on the past p lags, i.e.

$$Y_{it} = \sum_{j=1}^p g_{ij}(\alpha_{ij}^T Y_{t-j}) + \epsilon_{it}, \quad i = 1, \dots, d.$$

Model (1) has an identifiability issue between α_{ij} and $g_{ij}(\cdot)$, which is commonly seen in single-index models. As one can see, $\tilde{\alpha}_{ij} = c\alpha_{ij}$ and $\tilde{g}_{ij}(u) = g_{ij}(c^{-1}u)$ will serve the same purpose as α_{ij} and $g_{ij}(\cdot)$ for any $c \neq 0$. To make α_{ij} and g_{ij} identifiable, we set $\|\alpha_{ij}\| = 1$ and let the first component in α_{ij} be positive. Similar constraints were used in, for example,

Yu & Ruppert (2002) and Wang & Yang (2009). Because of the additive structure in (1), the g_j s are identified up to the addition of a constant vector. These identifiability issues are common in single/multiple-index models and additive models.

One can easily see that when all the link functions g_{ij} are linear functions, model (1) reduces to the usual linear VAR model of order p , VAR(p):

$$Y_t = \alpha_0 + \sum_{j=1}^p A_j Y_{t-j} + \epsilon_t,$$

where $\alpha_0 = (I_d - \sum_{j=1}^p A_j)\mu$ with $\mu = E(Y_t)$ and I_d is the identity matrix. Therefore, our SIAVAR(p) model, unlike VAR(p), can detect nonlinear relationships.

This model is very different from directly applying projection pursuit regression (PPR) models (Friedman & Stuetzle, 1981) or MAVE (Xia *et al.*, 2002) to Y_{it} using $(Y_{t-1}^T, \dots, Y_{t-p}^T)^T$ as covariates, because these methods are based on models with the following form

$$Y_{it} = \sum_{k=1}^K g_{ik}(\alpha_{ik,11} Y_{1,t-1} + \dots + \alpha_{ik,1d} Y_{d,t-1} + \dots + \alpha_{ik,p1} Y_{1,t-p} + \dots + \alpha_{ik,pd} Y_{d,t-p}) + \epsilon_{it}.$$

Notice that the weight vectors in such models have length $d \times p$. As a result, the index mixes the effects from the past p lags, making the model difficult to interpret. By contrast, the link function g_{ij} in our model still has the nice interpretation of being the effect of the j th lag in the past to the i th component of the present vector. One benefit of our model is that we can test which lag in the past has a nonlinear relationship with the present vector. In our El Niño data, and many other applications, we find that the current value Y_t is often linearly related to certain lags in the past, while nonlinearly related to other lags (see our results in section 6). By maintaining an additive structure for different lags, our proposed model can capture such features.

By construction, our model also differs from the PPR or MAVE in the following ways: (a) we allow some or all of the link functions to be linear without any identifiability problem, because the indexes are defined on different lags; (b) the index vectors α_{ij} and $\alpha_{i'j'}$ do not have to be orthogonal to each other for $j \neq j'$. This is not the case for PPR or MAVE models.

We would also like to stress the importance of jointly modelling all components of the vector time series. Joint modelling is particularly important to prediction, which is one of the fundamental tasks of time series analysis. When making a k -step ahead prediction, we need all the vector values along the way. More details are given in section 3.4.

2.2. Sufficient conditions for stationarity

Chen & Tsay (1993a) developed sufficient conditions for stationarity of a functional coefficient autoregressive model, which contains the additive autoregressive model as a special case. An & Huang (1996) introduced some sufficient conditions for the existence of a stationary distribution for general nonlinear autoregressive models. However, all these results were developed for univariate time series. In addition, all the sufficient conditions in the literature are not necessary. Therefore, when it comes to a specific model, one can still develop better sufficient conditions by taking into account the special features of the model.

Here, we introduce some sufficient conditions for the existence of stationary distributions for our SIAVAR models. By theorem 2.2 in Fan & Yao (2003), for the existence of a stationary distribution, it suffices to show that the Markov model associated with the time series is *geometrically ergodic* (see Fan & Yao, 2003, definition 2.4).

Theorem 1

Assume ϵ_t are independent and identically distributed (i.i.d.) with a positive density function, and the link functions in (1) satisfy the following conditions:

$$\sup_{|x| \leq M} |g_{ij}(x)| < \infty, \text{ for any } M, \tag{2}$$

$$\lim_{|x| \rightarrow \infty} \frac{|g_{ij}(x) - c_{ij}x|}{|x|} = 0, \tag{3}$$

for some constants c_{ij} , and for all $i = 1, \dots, d, j = 1, \dots, p$. Let $C_j = \text{diag}(c_{1j}, \dots, c_{dj})$, $\tilde{A}_j = C_j A_j$. If the roots of

$$|\zeta^p I_d - \zeta^{p-1} \tilde{A}_1 - \dots - \zeta \tilde{A}_{p-1} - \tilde{A}_p| = 0 \tag{4}$$

are all inside the unit complex disk, then model (1) is geometrically ergodic.

Theorem 1 generalizes theorem 3.1 in An & Huang (1996) to the vector time series setting. To appreciate the condition given in theorem 1, we consider $g_{ij}(x) = c_{ij}x$, i.e. model (1) reduces to a linear VAR(p) model. In this case, the condition given in theorem 1 is the sufficient condition for a linear VAR model to be causal. Theorem 1 implies that if the link functions are bounded within a compact set and behave like causal linear link functions when the values of $\{Y_{t-1}, \dots, Y_{t-p}\}$ become large, then the time series given in model (1) is geometrically ergodic, and therefore stationary. All the SIAVAR models discussed in our simulation studies in section 5 are examples of processes satisfying the conditions in theorem 1.

3. Estimation via P-splines

3.1. Penalized criterion

We shall consider a penalized spline (Eilers & Marx, 1996) method in estimating the non-parametric functions g_{ij} . There are many spline bases available. For convenience, we adopt the truncated power series basis in Ruppert *et al.* (2003):

$$g_{ij}(u) = \sum_{k=1}^m \delta_{ij,k} u^{k-1} + \sum_{k=1}^K \delta_{ij,k+m} (u - \kappa_{ij,k})_+^{m-1},$$

where m and K are the order and number of knots for the spline functions, the $\delta_{ij,k}$ s are the spline coefficients, $\kappa_{ij} = \{\kappa_{ij,1}, \kappa_{ij,2}, \dots, \kappa_{ij,K}\}$ is the set of knots for g_{ij} , and $(\cdot)_+$ denotes the positive part. If we put the spline basis functions into a vector

$$\mathbf{b}_{ij}(u) = \{1, u, u^2, \dots, u^{m-1}, (u - \kappa_{ij,1})_+^{m-1}, \dots, (u - \kappa_{ij,K})_+^{m-1}\}^T, \tag{5}$$

then $g_{ij}(u) = \delta_{ij}^T \mathbf{b}_{ij}(u)$ with $\delta_{ij} = (\delta_{ij,1}, \dots, \delta_{ij,K+m})^T$. Notice that modelling g_{ij} as spline functions is an approximation commonly used in semiparametric regression (see, e.g. Yu & Ruppert, 2002; Qu & Li, 2006; Apanasovich *et al.*, 2008).

The parameters in the model can be estimated by minimizing the following penalized least squares criterion:

$$Q_{n,\lambda}(\boldsymbol{\theta}) = n^{-1} \sum_{t=p+1}^n \left\| \mathbf{Y}_t - \sum_{j=1}^p \mathbf{g}_j(A_j \mathbf{Y}_{t-j}) \right\|^2 + \sum_{i,j} \lambda_{ij} \boldsymbol{\delta}_{ij}^T D_{ij} \boldsymbol{\delta}_{ij}, \tag{6}$$

where $\boldsymbol{\lambda} = \{\lambda_{ij}\}$ are the smoothing parameters, and $\boldsymbol{\theta} = [\boldsymbol{\theta}_i = \{\boldsymbol{\alpha}_{ij}, \boldsymbol{\delta}_{ij}; j = 1, \dots, p\}, i = 1, \dots, d]$. Here, the D_{ij} are positive semi-definite matrices representing roughness penalties on g_{ij} . A

typical choice of D_{ij} is $D_{ij} = \text{diag}\{\mathbf{0}_m, I_K\}$, i.e. a diagonal matrix with the last K diagonal entries equal to 1 and the rest equal to 0 (Ruppert *et al.*, 2003).

We can decompose the penalized least squares criterion (6) into the sum of penalized least squares for each component of the vector time series \mathbf{Y}_t . Let

$$Q_{n,\lambda_i,i}(\boldsymbol{\theta}_i) = n^{-1} \sum_{t=p+1}^n \left\{ Y_{it} - \sum_{j=1}^p g_{ij}(\boldsymbol{\alpha}_{ij}^T \mathbf{Y}_{t-j}) \right\}^2 + \sum_{j=1}^p \lambda_{ij} \boldsymbol{\delta}_{ij}^T D_{ij} \boldsymbol{\delta}_{ij},$$

then $Q_{n,\lambda}(\boldsymbol{\theta}) = \sum_{i=1}^d Q_{n,\lambda_i,i}(\boldsymbol{\theta}_i)$. In other words, we can fit a multiple-index model to each component of the vector time series.

3.2. Fitting algorithm

As the index vector $\boldsymbol{\alpha}_{ij}$ must have norm 1, we need to reparameterize it to get rid of this constraint. There are two commonly used reparameterizations:

$$\boldsymbol{\alpha}_{ij} = \left(\sqrt{1 - \|\boldsymbol{\gamma}_{ij}\|^2}, \boldsymbol{\gamma}_{ij}^T \right)^T, \tag{7}$$

where $\boldsymbol{\gamma}_{ij}$ is a $(d - 1)$ -dimensional vector in the unit disk, i.e. $\|\boldsymbol{\gamma}_{ij}\| \leq 1$; and

$$\boldsymbol{\alpha}_{ij} = (1 + \|\boldsymbol{\gamma}_{ij}\|^2)^{-1/2} (1, \boldsymbol{\gamma}_{ij}^T)^T, \tag{8}$$

where $\boldsymbol{\gamma}_{ij}$ is any $(d - 1)$ -dimensional vector. The only difference between the two reparameterizations is that (7) allows the first component of $\boldsymbol{\alpha}_{ij}$ to be 0. However, (8) is often used in practice, because it does not require constrained minimization.

For fixed order p and smoothing parameters λ , the estimating algorithm is described as follows:

- 1 To start, we let g_{ij} be identity functions and fit a linear VAR model to get an initial value for $\boldsymbol{\alpha}_{ij}$, then standardize them by $\boldsymbol{\alpha}_{ij}^{(0)} = \boldsymbol{\alpha}_{ij} / \|\boldsymbol{\alpha}_{ij}\|$, and set $g_{ij}^{(0)}(x) = \|\boldsymbol{\alpha}_{ij}\| x$.
- 2 Each penalized least squares criterion $Q_{n,\lambda_i,i}(\boldsymbol{\theta}_i)$ is minimized separately. For a fixed component i , $(\boldsymbol{\alpha}_{ij}, g_{ij})_{j=1}^p$ are updated iteratively following the backfitting idea. More details are provided below.
- 3 Repeat step 2 until the values of $(\boldsymbol{\alpha}_{ij}, g_{ij})_{j=1}^p$ converge. Use the same procedure for every component of the vector time series.

A detailed algorithm for the backfitting procedure for step 2 is described as follows. To update $(\boldsymbol{\alpha}_{ij}, g_{ij})$ for a specific lag j , we minimize $Q_{n,\lambda_i,i}$ with respect to $(\boldsymbol{\alpha}_{ij}, \boldsymbol{\delta}_{ij})$ while taking the rest of the parameters as fixed. If we define

$$Y_{it}^* = Y_{it} - \sum_{j' \neq j} g_{ij'}(\boldsymbol{\alpha}_{ij'}^T \mathbf{Y}_{t-j'}),$$

then this updating step is equivalent to fitting a single-index model for Y_{it}^* against \mathbf{Y}_{t-j} . We then iterate this procedure for all j .

Following a standard argument for backfitting (see the arguments for explaining the BRUTO algorithm in Chen & Tsay, 1993b, p. 957), the backfitting procedure described above is guaranteed to converge. The objective function $Q_{n,\lambda_i,i}$ is minimized with respect to one block of the parameters holding the rest fixed, then alternates between different blocks of the parameter vector. The objective function is decreased at each step; therefore, the algorithm will stop at a (local) minimum after a sufficient number of iterations.

We consider two ways to fit the single-index model in each iteration. The first algorithm is given by Yu & Ruppert (2002). We define the knots of g_{ij} to be on the quantiles of

$\{\alpha_{ij}^T Y_{t-j}\}_{t=p+1}^n$. As recommended in Ruppert *et al.* (2003), for a monotone function, 10–15 knots are usually adequate. We then augment the data and minimize the following nonlinear least squares criterion with respect to $(\alpha_{ij}, \delta_{ij})$:

$$\left\| \begin{pmatrix} Y_i^* \\ \mathbf{0} \end{pmatrix} - \begin{pmatrix} \mathbf{m}_{ij}(\alpha_{ij}, \delta_{ij}) \\ \sqrt{\lambda_{ij}} D_{ij}^{1/2} \delta_{ij} \end{pmatrix} \right\|^2 = Q_{n, \lambda_i, i}(\theta_i),$$

where $Y_i^* = (Y_{it}^*)_{t=p+1, \dots, n}^T$, $\mathbf{m}_{ij}(\alpha_{ij}, \delta_{ij}) = \{g_{ij}(\alpha_{ij}^T Y_{t-j}); t=p+1, \dots, n\}^T$. To get rid of the constraints on α_{ij} , we use the reparameterization (8) and minimize over $(\gamma_{ij}, \delta_{ij})$ instead. Minimization of this nonlinear least squares criterion can be carried out using some standard nonlinear minimization package. The function `nlm` in R (R Development Core Team, 2008) is a possible choice.

The second algorithm consists in profiling on the value of α_{ij} . Given the value of α_{ij} , define the knots of the spline basis to be on the quantiles of $u_{ij,t}(\alpha_{ij}) = \alpha_{ij}^T Y_{t-j}$. Then we can equivalently minimize

$$n^{-1} \sum_t [Y_{it}^* - \delta_{ij}^T \mathbf{b}_{ij}\{u_{ij,t}(\alpha_{ij})\}]^2 + \lambda_{ij} \delta_{ij}^T D_{ij} \delta_{ij}. \tag{9}$$

One can easily see that, with a fixed α_{ij} , the value of δ that minimizes the penalized least squares criterion (9) has a simple closed form

$$\hat{\delta}_{ij}(\alpha_{ij}) = [B_{ij}^T\{u_{ij,t}(\alpha_{ij})\} B_{ij}\{u_{ij,t}(\alpha_{ij})\} + n\lambda_{ij} D_{ij}]^{-1} B_{ij}^T\{u_{ij,t}(\alpha_{ij})\} Y_i^*,$$

where $B_{ij}\{u_{ij,t}(\alpha_{ij})\} = [\mathbf{b}_{ij}\{u_{ij,p+1}(\alpha_{ij})\}, \dots, \mathbf{b}_{ij}\{u_{ij,n}(\alpha_{ij})\}]^T$.

We then plug $\hat{\delta}_{ij}(\alpha_{ij})$ into (9), and minimize this profile penalized least squares criterion with respect to α_{ij} . Again, the reparameterization defined in (8) is used to get rid of the constraint on α_{ij} . This becomes a minimization problem in $(d - 1)$ -dimensional space, which can be done by any standard optimization package. In the vector time series setting, when the dimension is not high, this algorithm works very well. This second algorithm is used in all of our simulations and in our data analysis.

Remark 1. The spline basis in (5) includes an intercept term for all lags j . Notice that, when $p > 1$, the intercepts of g_{ij} as in definition (1) are unidentifiable, as one can add constants to the g_{ij} s without changing the model as long as these constants add up to 0. This is a well-known phenomenon in additive models (see discussions in Hastie & Tibshirani, 1990). The backfitting procedure described above, however, still works. The non-parametric link functions that we estimate are the versions with the constraint, $\sum_t \hat{g}_{ij}(\hat{\alpha}_{ij}^T Y_{t-j}) = 0$ for $i = 1, \dots, d$ and $j = 2, \dots, p$. However, this non-identifiability issue does affect our inference procedure, see our discussion below.

Remark 2. In our algorithm, the knots of \hat{g}_{ij} are placed at fixed quantiles of $\hat{\alpha}_{ij}^T Y_{t-j}$, which would vary for different iterations. In fact, as long as the knots are reasonably placed, the performance of the P-spline estimator is mainly controlled by the smoothing parameter λ_{ij} . To get rid of the variation caused by the changing of knots, we suggest to fix the knots after a sufficient number of iterations when the values of the α_{ij} s become stable.

Another possible way to place the knots is the following. After centring the time series, the scatter plot of Y_t falls in a ball in the \mathbb{R}^d space around the origin. We now have an idea of the range of $\alpha^T Y$ no matter what α is (because α is a unit vector), and we put equally spaced knots in this range. This will result in a real fixed knots scenario, but it does not work as well as choosing the knots adaptively.

3.3. Order and smoothing parameter selection

Huang & Yang (2004) proposed using BIC to select the order of additive autoregressive models. They showed that the BIC criterion can consistently identify the correct model as the sample size goes to ∞ . We adapt the same idea to our SIAVAR models.

In our semiparametric setting, we assume the smoothing parameters $\lambda_{ij} = o(n^{-1/2})$, so that the influence of the penalty can be ignored for large sample size. The degrees of freedom in the model are

$$\text{d.f.} = \sum_i \text{d.f.}_i, \quad \text{where} \quad \text{d.f.}_i = p(d - 1 + m + K).$$

Here, d.f._i are the degrees of freedom for the penalized least squares criterion $Q_{n,\lambda_i,i}(\theta)$, where $d - 1$ is the number of free parameters in α_{ij} and $m + K$ is the number of parameters defining g_{ij} . Then the BIC criterion is defined as

$$\text{BIC}(p) = \sum_{i=1}^d \text{BIC}_i(p), \quad \text{where} \quad \text{BIC}_i(p) = n \log(\text{RSS}_{p,i,n}/n) + \text{d.f.}_i \log(n),$$

and where $\text{RSS}_{p,i,n}$ is the residual sum of squares obtained by minimizing $Q_{n,\lambda_i,i}(\theta)$ with order p .

When the sample size is moderate, we need the penalty to regularize the estimator of the non-parametric link functions and we can incorporate the smoothing parameter selection into the BIC criterion above. When $\lambda_{ij} \neq 0$, with the estimated index weight vector $\hat{\alpha}_{ij}$, the hat matrix for \hat{g}_{ij} is

$$H_{ij}(\lambda_{ij}) = B_{ij}\{u_{ij,t}(\hat{\alpha}_{ij})\} [B_{ij}^T\{u_{ij,t}(\hat{\alpha}_{ij})\} B_{ij}\{u_{ij,t}(\hat{\alpha}_{ij})\} + n\lambda_{ij}D_{ij}]^{-1} B_{ij}^T\{u_{ij,t}(\hat{\alpha}_{ij})\}.$$

Then the degrees of freedom for fitting \hat{g}_{ij} are $\text{tr}\{H_{ij}(\lambda_{ij})\}$. Notice that these degrees of freedom will be $m + K$ if $\lambda_{ij} = 0$. Therefore, we should correct the degrees of freedom in the BIC criterion by

$$\text{d.f.}_i(\lambda_i) = p(d - 1) + \sum_{j=1}^p \text{tr}\{H_{ij}(\lambda_{ij})\}.$$

Minimizing the BIC with respect to both p and λ helps to choose the order of the model and the smoothing parameters.

Minimizing $\text{BIC}(p, \lambda)$ can be computationally expensive. Especially when d and p are large, we have many smoothing parameters to choose. We use an algorithm which combines backfitting and smoothing parameter selection. For a fixed p , we update λ_{ij} iteratively. Each time, we minimize $\text{BIC}(p, \lambda)$ with respect to one λ_{ij} using a grid search while holding the other λ s fixed. We continue until the BIC value converges to a minimum. The algorithm is guaranteed to converge, because the BIC value decreases at each step. This idea is similar in spirit with the BRUTO method in additive models (see Hastie & Tibshirani, 1990, section 9.4.3). Chen & Tsay (1993b, p. 957) also described the idea of updating one smoothing parameter at a time. The only difference is that we use BIC instead of the approximated generalized cross-validation (GCV) criterion.

Our unified procedure for choosing the order and smoothing parameters, combined with the adaptive backfitting algorithm mentioned above, worked very well in both our simulation studies and data analysis.

3.4. Prediction

One important task for time series models is to make prediction in the future. However, multi-step prediction in nonlinear time series models is not trivial. As stated in Fan & Yao

(2003), to do k -step ahead prediction, simply repeating one-step plug-in k times usually does not predict well. We use a nonlinear prediction method proposed by Huang & Shen (2004) based on the bootstrap (Efron & Tibshirani, 1993). The algorithm is as follows.

We generate B artificial series for the future:

$$\hat{Y}_{n+k}^{(b)} = \sum_{j=1}^p \hat{g}_j(\hat{A}_j \hat{Y}_{n+k-j}^{(b)}) + \epsilon_{n+k}^{(b)}, \quad k = 1, 2, \dots, \quad b = 1, \dots, B,$$

where the \hat{g}_j s and \hat{A}_j s are estimates, and the $\epsilon_{n+k}^{(b)}$ are sampled with replacement from the residual vectors. We can obtain a point prediction of Y_{n+k} by using either the mean or median of the bootstrap samples, i.e.

$$\hat{Y}_{n+k} = \frac{1}{B} \sum_{b=1}^B \hat{Y}_{n+k}^{(b)} \quad \text{or} \quad \tilde{Y}_{n+k} = \text{median}(\hat{Y}_{n+k}^{(b)}; b = 1, \dots, B).$$

The choice depends on the optimality criterion: if we want to choose a predictor that minimizes the mean squared prediction error, we should use the bootstrap mean; if we want to minimize the absolute prediction error, we should use the bootstrap median. With the bootstrap samples, we can also obtain interval predictors and density predictors for \hat{Y}_{n+k} . In our numerical examples, we typically use $B = 2000$.

When implementing this procedure, we also followed the truncation rule in Huang & Shen (2004). To generate $\hat{Y}_{n+k}^{(b)}$, if any $\hat{A}_j \hat{Y}_{n+k-j}^{(b)}$ falls outside the range of its historical values, the spline estimator \hat{g}_j becomes unreliable. If this happens in an early step of this bootstrap series, we will truncate the value of $\hat{A}_j \hat{Y}_{n+k-j}^{(b)}$ to the historical limit; if this scenario happens in later steps of the bootstrap series, we discard the entire series and begin a new one.

4. Asymptotic theory and inference

4.1. Preliminaries

In this paper, we adopt a semiparametric setting where we model the link functions g_{ij} as linear combinations of spline bases with fixed knots. Similar settings were adopted in Yu & Ruppert (2002), Qu & Li (2006) and Apanasovich *et al.* (2008). There are three reasons for the fixed knots assumption. First, spline functions are highly flexible; hence, the bias incurred by restricting g_{ij} to the spline space is rather small compared to the variation of the estimator. By theorem 20.3 in Powell (1981), for any smooth function $g \in C^l[a, b]$, the L^∞ approximation error by a spline is $O\{h^{\min(m,l)}\}$, where m is the order of the spline and h is the maximum distance between the neighbouring knots. Therefore, given a reasonable number of knots, the spline models are flexible enough for all the inference and prediction purposes in this paper. Secondly, by fixing the number of knots, the algorithms are numerically stable. Finally, theoretical results for such semiparametric methods are much easier to derive and understand.

At this point, we would like to comment on the increasing knots asymptotics for P-splines, which is more rigorous in theory. It is well-known that the theoretical difficulty of increasing knots asymptotics has hindered the popularization of P-splines. Recent developments by Li & Ruppert (2008) provided insights into this problem. However, their increasing knots asymptotic results were limited to constant and linear splines, which do not have continuous derivatives. How these theories can be applied to single-index models or additive models are interesting open questions that still need to be investigated. Nevertheless, as shown in our simulation studies, the fixed knots asymptotic theory we develop provides inference procedures with good accuracy.

4.2. Asymptotic results

For any index set \mathcal{T} , denote by $\mathcal{F}(\mathbf{Y}_t, t \in \mathcal{T})$ the σ -field generated by the variables $\{\mathbf{Y}_t, t \in \mathcal{T}\}$. The α -mixing coefficient of the process \mathbf{Y}_t is defined as

$$\alpha(k) = \sup_t \{P(A \cap B) - P(A)P(B), A \in \mathcal{F}(\mathbf{Y}_{t'}, t' \leq t), B \in \mathcal{F}(\mathbf{Y}_{t'}, t' \geq t+k)\}.$$

For convenience, denote $\boldsymbol{\mu}(\mathbf{Y}_{t-j}, j=1, \dots, p; \boldsymbol{\theta}) = \sum_j \mathbf{g}_j(A_j \mathbf{Y}_{t-j})$. Here, $\boldsymbol{\mu}$ is a d -dimensional vector function on $\boldsymbol{\theta}$ with the i th entry $\mu_i(\mathbf{Y}_{t-j}, j=1, \dots, p; \boldsymbol{\theta}_i), i=1, \dots, d$, and $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^T, \dots, \boldsymbol{\theta}_d^T)^T$ is the parameter vector after reparameterization (7) or (8), $\boldsymbol{\theta}_i = \{\gamma_{ij}, \delta_{ij}; j=1, \dots, p\}$. We make the following assumptions:

- (A1) The parameter space Θ is compact and the true parameter $\boldsymbol{\theta}^*$ is an interior point of Θ .
- (A2) The process \mathbf{Y}_t is strictly stationary. There exists a constant $\eta > 2$, such that

$$E|\epsilon_{i,t}|^\eta < \infty \quad \text{and}$$

$$E\left\{\left|\frac{\partial}{\partial \theta_{i,k}} \mu_i(\mathbf{Y}_{t-j}, j=1, \dots, p; \boldsymbol{\theta})\right|^\eta \middle| \boldsymbol{\theta}^*\right\} < \infty, \quad \text{for } i=1, \dots, d, k \leq \dim(\boldsymbol{\theta}_i)$$

and for all $\boldsymbol{\theta} \in \Theta$.

- (A3) \mathbf{Y}_t is α -mixing, with $\alpha(k) = O(k^{-\nu})$, as $k \rightarrow \infty$, for some $\nu > \eta/(\eta - 2)$.
- (A4) The function

$$Q(\boldsymbol{\theta}) = E\{\|\boldsymbol{\mu}(\mathbf{Y}_{t-j}, j=1, \dots, p; \boldsymbol{\theta}) - \boldsymbol{\mu}(\mathbf{Y}_{t-j}, j=1, \dots, p; \boldsymbol{\theta}^*)\|^2 \mid \boldsymbol{\theta}^*\}$$

has a unique minimum at $\boldsymbol{\theta} = \boldsymbol{\theta}^*$.

- (A5) The function $\boldsymbol{\mu}$ is twice continuously differentiable in a neighbourhood of $\boldsymbol{\theta}^*$. The quantity

$$\Omega_{i\ell}(\boldsymbol{\theta}^*) = V_{\epsilon, i\ell} E\left\{\frac{\partial}{\partial \boldsymbol{\theta}_i} \mu_i(\mathbf{Y}_{t-j}, j=1, \dots, p; \boldsymbol{\theta}_i^*) \frac{\partial}{\partial \boldsymbol{\theta}_\ell} \mu_\ell(\mathbf{Y}_{t-j}, j=1, \dots, p; \boldsymbol{\theta}_\ell^*)^T \middle| \boldsymbol{\theta}^*\right\}$$

exists for all $i, \ell = 1, \dots, d$. Let $\Omega(\boldsymbol{\theta})$ be the matrix with the (i, ℓ) th sub-matrix equal to $\Omega_{i\ell}(\boldsymbol{\theta}^*)$. The matrix $\Omega(\boldsymbol{\theta}^*)$ is non-singular.

We have the following asymptotic results.

Theorem 2

Under assumptions A1–A5, and if the smoothing parameters $\lambda_{ij} = o(1)$ as the sample size $n \rightarrow \infty$, for all i and j , then the sequence of penalized least squares estimator $\hat{\boldsymbol{\theta}}$, which minimize $Q_{n,\lambda}(\boldsymbol{\theta})$ in (6), converge to $\boldsymbol{\theta}^*$ with probability 1.

Theorem 3

Under assumptions A1–A5, and if the smoothing parameters $\lambda_{ij} = o(n^{-1/2})$, then the penalized least squares estimator has the following asymptotic distribution:

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \Rightarrow \text{Normal}\{\mathbf{0}, \Psi^{-1}(\boldsymbol{\theta}^*)\Omega(\boldsymbol{\theta}^*)\Psi^{-1}(\boldsymbol{\theta}^*)\},$$

where $\boldsymbol{\theta}^*$ is the true parameter and $\Psi(\boldsymbol{\theta}^*) = \text{diag}\{\Omega_{ii}(\boldsymbol{\theta}^*)\}_{i=1}^d$ is a block diagonal matrix.

Proofs of the theorems are somewhat similar to those of Yu & Ruppert (2002) but adapted to the time series setting, therefore they are omitted here. They can be found in our online Supporting Information.

Remark 3. Assumption 1 for our asymptotic theory is not satisfied for our model set-up in section 3, because the intercept terms of the g_{ij} s are not identifiable, see remark 1. Because of

the non-identifiability issue, the parameter space is degenerated. This problem can be easily fixed: we redefine the basis for the link functions in (5) as, $\mathbf{b}_{i1}^* = \mathbf{b}_{i1}$,

$$\mathbf{b}_{ij}^* = \{u, u^2, \dots, u^{m-1}, (u - \kappa_{ij,1})_+^{m-1}, \dots, (u - \kappa_{ij,k})_+^{m-1}\}^T, \quad \text{for } j = 2, \dots, p.$$

In order to make inference based on the asymptotic theory, we need to combine the intercept terms of the g_{ij} s after the backfitting procedure in section 3, such that only g_{i1} has an intercept term for all i . Without further complication, we will also refer to the new spline basis as \mathbf{b}_{ij} .

4.3. Covariance estimation and inference

The results in theorem 3 are given under the condition that the smoothing parameters will diminish to 0 with a rate $o(n^{-1/2})$ so that the roughness penalty does not have any effect on the asymptotic results. However, in reality we need to take into account the effect of the penalty for the following two reasons. First, with a finite sample size, the λ_{ij} are not exactly 0; therefore, the effect of the roughness penalty cannot be ignored. Secondly, if some of the link functions g_{ij} are truly linear, our smoothing parameter selection procedure will choose large values for λ_{ij} , and these values may not converge to 0 even if the sample size becomes large.

We next discuss estimation of the asymptotic covariance matrix of $\hat{\theta}$, taking into account non-trivial smoothing parameters. The following arguments are similar to those in section 3.2 of Yu & Ruppert (2002). To emphasize the relationship between the estimator with the smoothing parameters, we denote the penalized least squares estimator as $\hat{\theta}(\lambda)$. We notice that $\hat{\theta}(\lambda)$ is the solution of the estimating equation

$$\sum_t \psi_t(\theta, \lambda) = \mathbf{0},$$

where

$$\psi_t(\theta, \lambda) = -\frac{\partial}{\partial \theta} \mathbf{Y}_t^\top (\mathbf{Y}_{t-j}, j = 1, \dots, p; \theta) \{ \mathbf{Y}_t - \mu(\mathbf{Y}_{t-j}, j = 1, \dots, p; \theta) \} + D(\lambda)\theta,$$

and $D(\lambda)$ is a block-diagonal matrix with $d \times p$ blocks on the diagonal, representing the roughness penalty for the $d \times p$ link functions. Each of these blocks is controlled by the smoothing parameter λ_{ij} .

From the theory of estimating equations (Carroll *et al.*, 2006, appendix 3; Gray, 1994), the sandwich formula for the variance of $\hat{\theta}(\lambda)$ is given by

$$\hat{V}_{sw} = \hat{\Psi}\{\hat{\theta}(\lambda)\}^{-1} \hat{\Omega}\{\hat{\theta}(\lambda)\} \hat{\Psi}\{\hat{\theta}(\lambda)\}^{-T}, \tag{10}$$

where

$$\hat{\Psi}(\theta) = \sum_t \frac{\partial}{\partial \theta^\top} \psi_t(\theta, \lambda) \quad \text{and} \quad \hat{\Omega}(\theta) = \sum_t \psi_t(\theta, \lambda) \psi_t(\theta, \lambda)^\top.$$

Notice that we can obtain the covariance estimator of the reparameterized index vector $\hat{\gamma}_{ij}$ from the corresponding sub-block in \hat{V}_{sw} , denoted as $\hat{V}_{\gamma_{ij}}$. To get the covariance estimator of α_{ij} , $\hat{V}_{\alpha_{ij}}$, we can use the delta method. Define the Jacobian matrix for the reparameterization (8), $J(\gamma) = \partial \alpha / \partial \gamma^\top$, which is a $d \times (d - 1)$ matrix. Then $\hat{V}_{\alpha_{ij}} = J(\hat{\gamma}_{ij}) \hat{V}_{\gamma_{ij}} J(\hat{\gamma}_{ij})^\top$.

In our SIAVAR model, an important inference problem is to assess whether the link functions are truly nonlinear. We solve this problem by testing the hypothesis, $H_0: g_{ij}$ is linear, directly via a Wald test. In our setting, let $\delta_{ij,(h)}$ be the spline coefficients of g_{ij} with degree higher than 1. Then the hypothesis above is equivalent to $H_0: \delta_{ij,(h)} = \mathbf{0}$. The Wald test statistic is defined as

$$W = \hat{\delta}_{ij,(h)}^T \hat{V}^{-1}(\hat{\delta}_{ij,(h)}) \hat{\delta}_{ij,(h)},$$

where $\hat{V}(\hat{\delta}_{ij,(h)})$ is the estimated asymptotic covariance matrix of $\hat{\delta}_{ij,(h)}$. Under the null hypothesis, the asymptotic distribution of the Wald statistic is a Chi-squared distribution with degrees of freedom equal to $\dim(\delta_{ij,(h)})$.

5. Simulation study

5.1. Simulation 1: estimation, inference and predictions

In the first simulation study, we consider a ($d=3$)-dimensional nonlinear vector time series, following the SIAVAR(2) model given in (1) with

$$A_1 = \frac{1}{\sqrt{6}} \begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix}, \quad A_2 = \frac{1}{\sqrt{3}} \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix},$$

$$g_{11}(x) = -0.4(3 - x^2)/(1 + x^2), \quad g_{12}(x) = 0.6\{3 - (x - 0.5)^3\}/\{1 + (x - 0.5)^4\},$$

$$g_{21}(x) = \{0.4 - 2 \exp(-x^2/2)\}x, \quad g_{22}(x) = 0.3x,$$

$$g_{31}(x) = \{0.4 - 2 \cos(4x) \exp(-x^2)\}x, \quad g_{32}(x) = 0.25x.$$

We also let the error terms ϵ_{it} be i.i.d. random variables with a Uniform $[-1, 1]$ distribution. We simulate a series from the SIAVAR(2) model with length $n=500$ plus 10 values as validation for prediction, and then repeat this 200 times.

First, we want to verify if the model of the simulation satisfies the conditions in theorem 1. We find that conditions (2) and (3) are satisfied for $C_1 = \text{diag}(0, 0.4, 0.4)$ and $C_2 = \text{diag}(0, 0.3, 0.25)$. Then the roots for eigenequation (4) are $\{0, 0, 0, -0.369, 0.163, 0.859\}$, which are all inside the unit complex disk. By theorem 1, the process generated by this model is stationary.

We fit the SIAVAR model to the first 500 data points for each simulated data set. We use cubic splines with 10 knots to model each non-parametric link function. The BIC criterion proposed in section 3.3 chose the right order $p=2$ for all the 200 repetitions.

Estimation of the index weight vectors α_{ij} is summarized in Table 1. We can see that the bias of the estimator is very small compared with the Monte Carlo standard error (MC SE). The mean values of the standard error estimators (SW SE) based on the sandwich formula

Table 1. *Simulation 1: true value, bias, Monte Carlo standard error (MC SE) of the estimated index weights and mean of the estimated standard error using the sandwich formula (SW SE), over 200 simulations*

	$\alpha_{11,1}$	$\alpha_{11,2}$	$\alpha_{11,3}$	$\alpha_{12,1}$	$\alpha_{12,2}$	$\alpha_{12,3}$
True value	0.816	0.408	0.408	0.577	0.577	0.577
Bias	-0.002	-0.004	0.001	0.000	0.002	-0.004
MC SE	0.027	0.046	0.046	0.023	0.022	0.023
SW SE	0.026	0.044	0.043	0.021	0.022	0.024
	$\alpha_{21,1}$	$\alpha_{21,2}$	$\alpha_{21,3}$	$\alpha_{22,1}$	$\alpha_{22,2}$	$\alpha_{22,3}$
True value	0.408	0.816	0.408	0.577	0.577	0.577
Bias	-0.001	0.001	-0.003	0.000	-0.020	-0.002
MC SE	0.026	0.019	0.029	0.084	0.092	0.097
SW SE	0.025	0.016	0.028	0.076	0.085	0.086
	$\alpha_{31,1}$	$\alpha_{31,2}$	$\alpha_{31,3}$	$\alpha_{32,1}$	$\alpha_{32,2}$	$\alpha_{32,3}$
True value	0.408	0.408	0.816	0.577	0.577	0.577
Bias	-0.001	-0.006	0.003	-0.002	-0.008	-0.026
MC SE	0.018	0.020	0.012	0.101	0.116	0.131
SW SE	0.021	0.023	0.015	0.091	0.099	0.106

Table 2. *Simulation 1: empirical frequency of rejecting the hypothesis $H_0: g_{ij}$ is linear, over 200 simulations. Only g_{22} and g_{32} are truly linear. Nominal size of the test is 0.05*

g_{11}	g_{12}	g_{21}	g_{22}	g_{31}	g_{32}
1.000	1.000	1.000	0.045	1.000	0.040

(10) are also provided in Table 1, and they are very close to the Monte Carlo standard errors. We also checked the Q–Q plots of the empirical distribution of $\hat{\boldsymbol{x}}_{ij,k}$ against the normal for all i, j and k , and these empirical distributions were very close to the normal distribution.

We also perform the Wald test proposed in section 4.3 to test linearity for each g_{ij} . We let the nominal size of the test be 0.05. The empirical rejection rates for each hypothesis over 200 simulations are reported in Table 2. Notice that all true g_{ij} are nonlinear except for g_{22} and g_{32} . Therefore, the results in Table 2 show that, when the sample size is large, the test has size close to the nominal value, and it has very high power.

Next, we compare the prediction power of our method with those of linear VAR models. In environmental sciences, the absolute prediction error (APE) is often of great interest. For any predictor \hat{Y} , we define the k -step ahead APE as

$$\text{APE}_i(k) = E|\hat{Y}_{i,n+k} - Y_{i,n+k}|, \quad i = 1, \dots, d.$$

For all of our numerical examples (simulations and the analysis of the El Niño data in section 6), when predicting from our SIAVAR model, we use the bootstrap median as the predictor, as described in section 3.4. We did try using a bootstrap mean as the predictor in all our examples, and we obtained essentially the same results.

We use standard R package (function `ar` in the `stats` package) to fit linear VAR models to the simulated data sets. The BIC chose the order $p=2$ for all 200 data sets. By contrast, the AIC criterion would pick different orders for different simulated data sets, but the chosen orders all ranged between $p=2$ and 5. We therefore fit linear VAR(p) models for each of these orders using the first 500 values in each simulated data set. The fitted models are then used to predict the last 10 values, and the APEs are calculated. For comparison, we also calculate the APE of a persistence forecast, which means we use the last observed value in the time series, \mathbf{Y}_n , as the predictor for all future values.

Define the k -step ahead relative APE of the SIAVAR(2) model versus the linear VAR(p) model as the ratio of the k -step ahead APE of SIAVAR(2) over that of the linear VAR(p) model, denoted as $\text{RELAPe}_i(k; p, 2)$, where the index i indicates the i th component of the vector. We define $\text{RELAPe}_i(k; 0, 2)$ to be the relative APE of the persistence forecast.

In Fig. 1, we show the plots of $\text{RELAPe}_i(k; p, 2)$ against k for all of the three components of the vectors. As we can see, all of the linear VAR models behave similarly: the RELAPE curves are lower than 1 when k is small, close to 1 when k becomes large. This indicates that the SIAVAR(2) model is better than any of the linear VAR models when predicting for the near future but is about the same for long-term prediction. Indeed, the correlation of the process defined in this simulation study decays quickly. Therefore, when predicting for the far future none of these models can perform a lot better than the mean value of the process. We can also see that the persistence forecast is far worse than any of these models.

5.2. Simulation 2: long-term prediction

We now use a second simulation study to illustrate the long-term prediction power of our methods under the situation that the correlation in the time series is strong. We simulate from

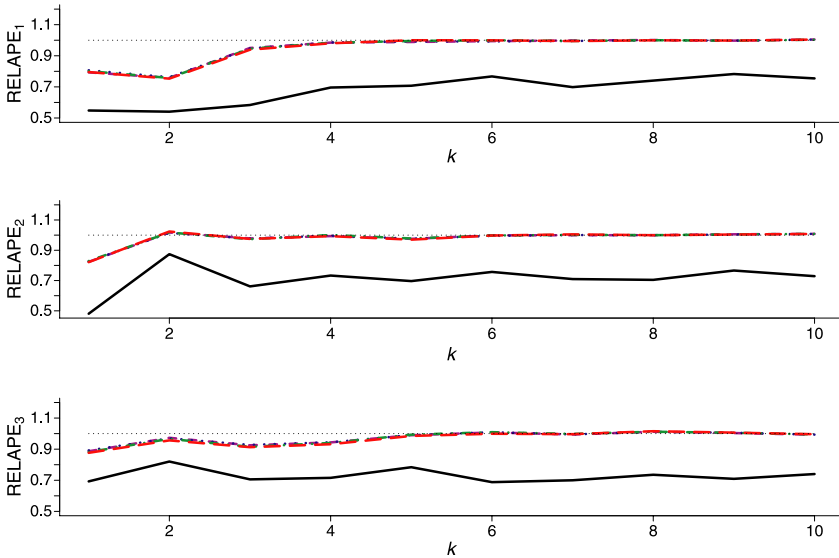


Fig. 1. $RELAPE_i(k; p, 2)$ for simulation 1. The three panels correspond to the three components of the vectors. In each plot, the solid curve is the relative APE of SIAVAR(2) versus persistence forecast; the dashed curve is the relative APE of SIAVAR(2) versus linear VAR(2); the dotted curve is the relative APE of SIAVAR(2) versus linear VAR(3); the dot-dashed curve is the relative APE of SIAVAR(2) versus linear VAR(4); the long dashed curve is the relative APE of SIAVAR(2) versus linear VAR(5).

another SIAVAR(2) model with $d = 3$, and

$$A_1 = \begin{pmatrix} 0.95 & 0.18 & 0.00 \\ 0.00 & 0.95 & 0.06 \\ 0.00 & -0.10 & 0.95 \end{pmatrix}, \quad A_2 = \begin{pmatrix} 0.86 & 0.50 & -0.13 \\ 0.78 & 0.55 & 0.29 \\ 0.77 & -0.31 & 0.55 \end{pmatrix},$$

$$g_{11}(x) = g_{21}(x) = g_{31}(x) = x, \quad g_{12}(x) = -0.5(x + 1)I(x < -1) - 0.5(x - 1)I(x > 1),$$

$$g_{22}(x) = -1.5x/(1 + 0.5x^2), \quad g_{32}(x) = -(0.4 - \exp(-x^2/2))x.$$

We set $\epsilon_t \sim$ i.i.d. $\text{Normal}\{\mathbf{0}, \text{diag}(0.03, 0.04, 0.05)\}$.

Part of the goal of this simulation is to generate a process that resembles the El Niño data. Under this model, Y_t has very strong linear relationship with Y_{t-1} , with diagonal elements of A_1 equal to 0.95. The second lag relationships in the time series are nonlinear. These features are similar to those we observed in the El Niño data, see our analysis results in section 6. The standard deviations of the errors for the three components of the time series are about 0.2, which might seem somewhat small, but they are roughly what we observed in the El Niño data. On the other hand, they are not extremely small compared with the standard deviations of the three components of Y_t , which are 1.02, 0.76 and 1.66.

Again, we first verify the conditions in theorem 1. We find that the model above satisfies conditions (2) and (3) for $C_1 = I_3$ and $C_2 = \text{diag}(-0.5, 0, -0.4)$, and the roots for the eigenequation (4) are $\{0, 0.967, 0.638 \pm 0.375i, 0.304 \pm 0.358i\}$ which are all inside the unit complex disk.

We simulate 200 independent series from the model above. As before, each data set is split into a training set with sample size $n = 500$ and a test set with size 12. For each data set, we fit a SIAVAR(2) model and linear VAR models with order $p = 2, 3, 4, 5$ to the training set, then use the test set to compute the APE. In Fig. 2, we plot the APE against the number of steps ahead for each component of the vector and for each fitted model. For comparison, the APE curves of the persistence forecast are also provided in these plots.

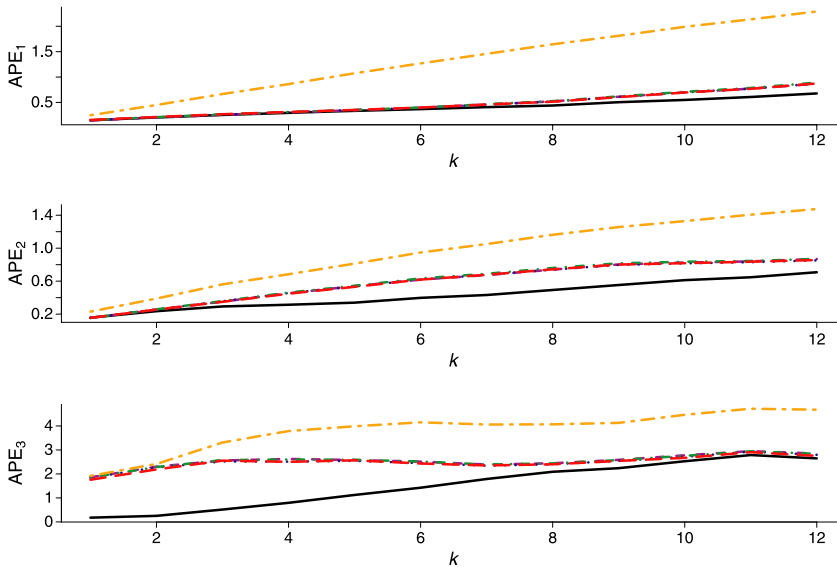


Fig. 2. Absolute prediction errors for simulation 2. The three panels correspond to the three variables in the vector time series. In each plot, the curves are $APE(k)$ against k , where k is the number of steps ahead. The lower solid curve is the APE for the SIAVAR(2) model, and the upper two-dashed curve is the APE for the persistence forecast. The middle curves are almost undistinguishable and here the dashed curve is for the VAR(2), the dotted curve is for the VAR(3), the dot-dashed curve is for the VAR(4) and the long dashed curve is for the VAR(5).

As we can see, the SIAVAR model has significantly smaller prediction errors than the linear VAR models. For the first component of the vector, the autocorrelation is very high and decays slowly. Hence, the nonlinear model can perform much better than linear models for long-term predictions, such as 12 steps ahead. For the third component of the vector, the autocorrelation decays faster and we observe a similar pattern as in Simulation 1: the nonlinear model is better for short-term prediction, and when it approaches 12 steps ahead, the nonlinear model gives similar results as the linear VAR models. Again, the persistence forecast is much worse than any of these models.

6. Climate data analysis

We consider monthly data of sea surface temperature, sea surface height and wind stress over the tropical Pacific Ocean during the period 1960–2004. The spatial information of these three variables (20°S–20°N, 125°E–70°W, at a $3^\circ \times 2^\circ$ resolution) is typically condensed by atmospheric scientists by means of empirical orthogonal functions, i.e. principal components; see, for example, Xue *et al.* (2000). Seasonal effects were also removed. The result is a trivariate time series, which is particularly important for studying the El Niño phenomenon.

6.1. SIAVAR model versus linear VAR models

We split this time series into a training set for the period 1960–96 and a testing set for the period 1997–2004. Our goal is to fit linear and nonlinear VAR models to the training data and compare the APE for the next 12 months using the test set.

We first fit linear VAR models to the training data, where the AIC criterion picked the order $p=5$ and the BIC picked $p=2$. We thus fitted all linear VAR models from $p=2$ to

$p=5$ and use these models to predict the values in the test set. We calculate the $APE(k)$, where k is the number of steps ahead, as follows. For one-step ahead prediction, given all values before 1997, we do a one-step ahead prediction for January 1997 and get one realization of one-step prediction error; we do a one-step prediction for February 1997 given all values before to get the second realization of the one-step prediction error. We repeat this procedure until December 2004. We then average the 96 one-step prediction errors to get $APE(1)$. Similarly, $APE(2)$ is the average of 95 two-step ahead prediction errors, and so on. We obtain $APE(k)$ for $k=1, \dots, 12$ and plot them as curves. The APE curves are given in Fig. 3 for each of the three variables in the vector time series, and for each of the linear VAR models for $p=2-5$.

We then fit our SIAVAR model to the data. The BIC criterion picked the order $p=2$. There are six non-parametric link functions in the model, the estimates of which are shown in Fig. 4. To check how well these curves are fitted to the data, we give the scatter plot of $(\hat{\alpha}_{ij}^T Y_{t-j}, Y_{it} - \sum_{j' \neq j} \hat{g}_{ij'}(\hat{\alpha}_{ij'}^T Y_{t-j'}))$ in each of the six panels in Fig. 4. As we can see, these functions fit well into the data and some of them, g_{22} and g_{32} , clearly exhibit some nonlinearity. Our Wald test also confirmed that g_{22} and g_{32} are statistically significantly nonlinear, with p -value $= 1.04 \times 10^{-5}$ and 6.04×10^{-4} respectively. All the other functions are not significantly nonlinear.

We also apply the nonlinear prediction procedure described in section 3.4. We obtain the APE curves for the SIAVAR(2) model in the same way as we did for linear VAR models, and plot them as solid curves in Fig. 3. By comparing these APE curves with those of the linear VAR models, we can see that the SIAVAR model clearly improved the APE for the second and third components of the time series. For the first component, the SIAVAR model has APE comparable with that of the linear models for $k=1, \dots, 5$, and has lower APE for $k \geq 6$. We also plotted the APE curves for persistence forecasts as reference curves. We found that, over all, persistence forecast is much worse than any of these models.

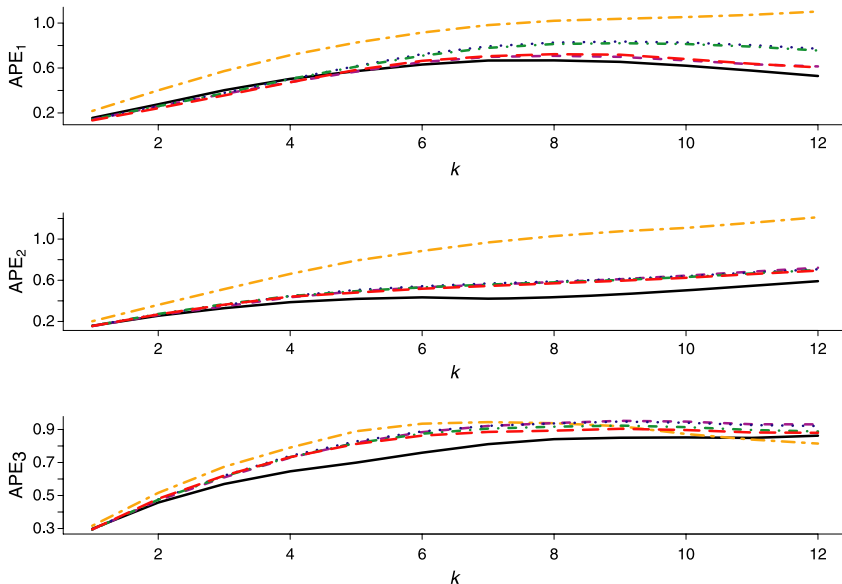


Fig. 3. Absolute prediction errors for the El Niño data. The three panels correspond to the three variables in the vector time series. In each plot, the curves are $APE(k)$ against k , where k is the number of steps ahead in months. The solid curve is the APE for the SIAVAR(2) model, the dashed curve is for the VAR(2), the dotted curve is for the VAR(3), the dot-dashed curve is for the VAR(4), the long dashed curve is for the VAR(5) and the two-dashed curve is the APE for the persistence forecast.

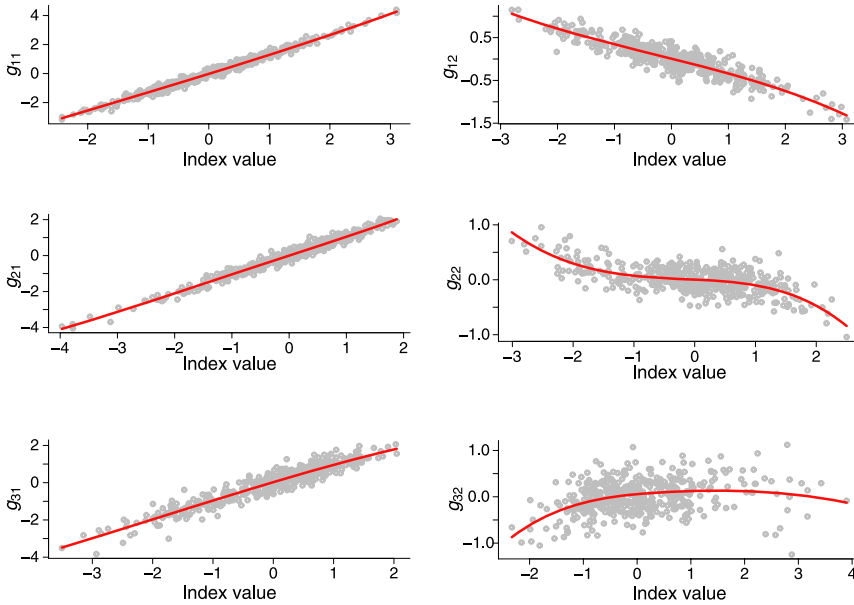


Fig. 4. Estimated link functions for the El Niño data. The scatter plots in each panel are $(\hat{\alpha}_{ij}^T Y_{t-j}, Y_{it} - \sum_{j' \neq j} \hat{g}_{ij'}(\hat{\alpha}_{ij'}^T Y_{t-j'}))$, and the curves are the estimated link function $\hat{g}_{ij}(\cdot)$.

6.2. Comparison with functional autoregressive models

We provide a comparison of our SIAVAR model with the functional autoregressive (FAR) models proposed by Besse *et al.* (2000), who also analysed some climate data related to El Niño. It is worth pointing out that the El Niño data analysed in our paper are different from those in Besse *et al.* (2000): the data were collected for different variables, from a different region of the ocean and over a different time period.

The FAR models are a class of scalar time series models, in which the time courses (or trajectories) in different years are viewed as functional data (Ramsay & Silverman, 2005). The FAR(1) models in Besse *et al.* (2000) predict the entire trajectory of the time series for the next year using the trajectory in the current year as the predictor.

We implement the two best models in their paper, the smooth FAR(1) model and the local FAR(1) model. The smooth FAR(1) models the relationship between the time series in the current year with that of the past year using a linear model. Due to the high dimensionality of functional data, a dimension reduction via principal component analysis is conducted before fitting the linear model. The local FAR(1) is based on a similar idea, except that the linear model is fitted locally using kernel weights. Both models have tuning parameters which determine the performance of the model: for the smooth FAR(1), the number of principal components q in the linear model is the tuning parameter; for the local FAR(1), we need to choose the number of principal components q and the bandwidth h for the kernel weights. We choose these tuning parameters using cross-validation as suggested in Besse *et al.* (2000). For more details on these models, we refer the readers to the original paper.

We fit the two FAR models to the training set, separately for each component of the time series. Then, we use the fitted model to do prediction in the test set. Unlike the prediction procedure for the SIAVAR model, the FAR model gives a prediction for an entire year. In the test set, we use the annual data in 1996 to predict the whole year of 1997, then use the annual data in 1997 to predict 1998, and so on. The APES are calculated and compared with

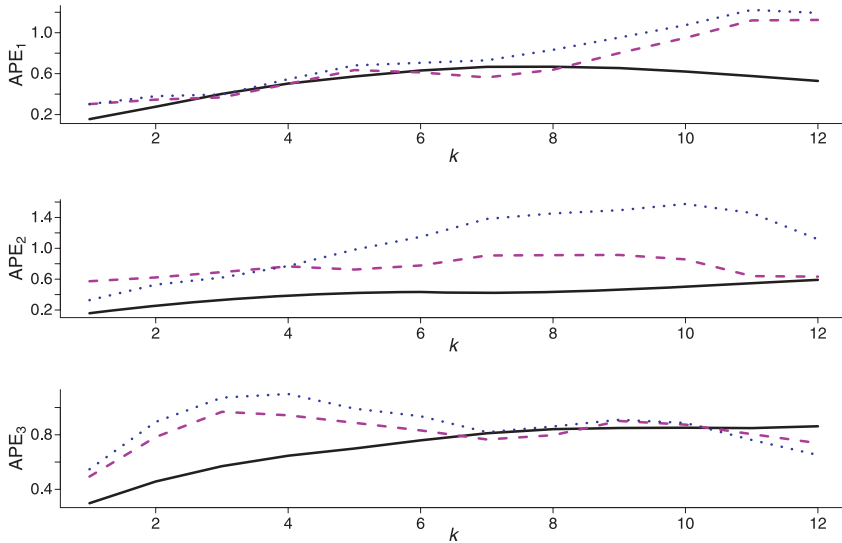


Fig. 5. Comparison of the SIAVAR model with the FAR(1) model for the El Niño data. The three panels correspond to the three variables in the vector time series. In each plot, the curves are $APE(k)$ against k , where k is the number of steps ahead in months. The solid curve is the APE for the SIAVAR(2) model, the dashed curve is for the smooth FAR(1), the dotted curve is for the local FAR(1).

those of the SIAVAR model in Fig. 5. As we can see, overall, the predictions from the FAR model are worse than those from our SIAVAR model for these data.

7. Conclusion

We have proposed a single-index additive VAR time series model, which is parsimonious and flexible. It has the dimension reduction flavour of single-index models, and also inherits the interpretability of additive models. We allow some lag effects to be linear and some to be nonlinear.

We described a backfitting procedure to fit the model using penalized splines. We proposed a BIC criterion to choose both the order of the model and the smoothing parameters. This procedure has been proved to be quite successful in simulations. An asymptotic Wald-type test was described to determine which of the link functions are nonlinear.

We also used a bootstrap method to perform nonlinear prediction for future observations. Our method was successfully applied to a climate data set, with less APE than linear VAR models and FAR models. This result also shows the importance of developing nonlinear models for vector time series.

Acknowledgements

Genton’s research was supported in part by a National Science Foundation CMG grant ATM-0620624 and by Award no. KUS-C1-016-04, made by King Abdullah University of Science and Technology (KAUST). The authors thank the editor, the associate editor and two referees for constructive suggestions that have improved the content and presentation of this article. The authors also thank Salil Mahajan and Ramalingam Saravanan from the Department of Atmospheric Sciences at Texas A&M University for providing the climate data set.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Proofs of theorems 2 and 3, and formulas to calculate the sandwich variance estimator (10).

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

References

- An, H. Z. & Huang, F. C. (1996). The geometrical ergodicity of nonlinear autoregressive models. *Statist. Sinica* **6**, 943–956.
- Apanasovich, T., Ruppert, D., Lupton, J., Popovic, N., Turner, N., Chapkin, R. & Carroll, R. J. (2008). Semiparametric longitudinal-spatial binary regression, with application to colon carcinogenesis. *Biometrics* **64**, 490–500.
- Besse, P. C., Cardot, H. & Stephenson, D. B. (2000). Autoregressive forecasting of some functional climatic variations. *Scand. J. Statist.* **27**, 673–687.
- Carroll, R. J., Fan, J., Gijbels, I. & Wand, M. P. (1997). Generalized partially linear single-index models. *J. Amer. Statist. Assoc.* **92**, 477–489.
- Carroll, R. J., Ruppert, D., Stefanski, L. A. & Crainiceanu, C. (2006). *Measurement error in nonlinear models, a modern perspective*, 2nd edn. Chapman & Hall, London.
- Chan, K. S. & Tong, H. (1985). On the use of the deterministic Lyapunov function for the ergodicity of stochastic difference equations. *Adv. Appl. Probab.* **17**, 666–678.
- Chen, R. & Tsay, R. (1993a). Functional-coefficient autoregressive models. *J. Amer. Statist. Assoc.* **88**, 298–308.
- Chen, R. & Tsay, R. (1993b). Nonlinear additive ARX models. *J. Amer. Statist. Assoc.* **88**, 955–967.
- Efron, B. & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. Chapman & Hall, London.
- Eilers, P. H. C. & Marx, B. D. (1996). Flexible smoothing with B-splines and penalties (with discussion). *Statist. Sci.* **11**, 89–121.
- Fan, J. & Yao, Q. (2003). *Nonlinear time series: nonparametric and parametric methods*. Springer, New York.
- Friedman, J. H. & Stuetzle, W. (1981). Projection pursuit regression. *J. Amer. Statist. Assoc.* **76**, 817–823.
- Gray, R. J. (1994). Spline-based tests in survival analysis. *Biometrics* **50**, 640–652.
- Härdle, W., Tsybakov, A. & Yang, L. (1998). Nonparametric vector autoregression. *J. Statist. Plann. Inference* **68**, 221–245.
- Hastie, T. J. & Tibshirani, R. J. (1990). *Generalized additive models*. Chapman & Hall, New York.
- Huang, J. Z. & Shen, H. (2004). Functional coefficient regression models for nonlinear time series: a polynomial spline approach. *Scand. J. Statist.* **31**, 515–534.
- Huang, J. Z. & Yang, L. (2004). Identification of non-linear additive autoregressive models. *J. Roy. Statist. Soc. Ser. B* **66**, 463–477.
- Li, Y. & Ruppert, D. (2008). On the asymptotics of penalized splines. *Biometrika* **95**, 415–436.
- Powell, M. J. D. (1981). *Approximation theory and methods*. Cambridge University Press, Cambridge, UK.
- Qu, A. & Li, R. (2006). Quadratic inference functions for varying-coefficient models with longitudinal data. *Biometrics* **62**, 379–391.
- R Development Core Team (2008). *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.
- Ramsay, J. O. & Silverman, B. W. (2005). *Functional data analysis*, 2nd edn. Springer, New York.
- Ruppert, D., Wand, M. P. & Carroll, R. J. (2003). *Semiparametric regression*. Cambridge University Press, Cambridge, UK.
- Tweedie, R. L. (1975). Criteria for classifying general Markov chains. *Adv. Appl. Probab.* **8**, 737–771.
- Wang, L. & Yang, L. (2009). Spline estimation of single-index models. *Statist. Sinica* **19**, 765–783.
- Xia, Y. & Härdle, W. (2006). Semi-parametric estimation of partially linear single-index models. *J. Multivariate Anal.* **97**, 1162–1184.
- Xia, Y., Tong, H., Li, W. K. & Zhu, L.-X. (2002). An adaptive estimation of dimension reduction space. *J. Roy. Statist. Soc. Ser. B* **64**, 363–410.

Xue, Y., Leetmaa, A. & Ji, M. (2000). ENSO predictions with Markov models: the impact of sea level. *J. Climate* **13**, 849–871.
 Yu, Y. & Ruppert, D. (2002). Penalized spline estimation for partially linear single-index models. *J. Amer. Statist. Assoc.* **97**, 1042–1054.

Received August 2008, in final form January 2009

Yehua Li, Department of Statistics, University of Georgia, 204 Statistics Building, 101 Cedar Street, Athens, GA 30602, USA.
 E-mail: yehuali@uga.edu

Appendix: Proof of Theorem 1

Lemma 1 (Tweedie, 1975)

Let $\{X_t\}$ be an aperiodic irreducible Markov chain. Suppose that there exists a compact set A , a non-negative measurable function g , positive constants c_1, c_2 and $\rho < 1$ such that

$$E\{g(X_{t+1}) | X_t = x\} \leq \rho g(x) - c_1, \quad \text{for any } x \notin A,$$

$$E\{g(X_{t+1}) | X_t = x\} \leq c_2, \quad \text{for any } x \in A.$$

Then $\{X_t\}$ is geometrically ergodic.

Proof of theorem 1. Let $\tilde{Y}_t = (Y_t^T, Y_{t-1}^T, \dots, Y_{t-p+1}^T)^T$, $\tilde{\epsilon}_t = (\epsilon_t^T, \mathbf{0}^T, \dots, \mathbf{0}^T)^T$ and

$$\mathcal{T}(\tilde{Y}) = \begin{pmatrix} C_1(Y_1)A_1 & C_2(Y_2)A_2 & \cdots & C_{p-1}(Y_{p-1})A_{p-1} & C_p(Y_p)A_p \\ I_d & 0 & \cdots & 0 & 0 \\ 0 & I_d & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & I_d & 0 \end{pmatrix},$$

where $C_j(Y) = \text{diag}\{g_{ij}(\alpha_{ij}^T Y) / (\alpha_{ij}^T Y)\}_{i=1}^d$. Then model (1) can be written as the following Markov model

$$\tilde{Y}_t = \mathcal{T}(\tilde{Y}_{t-1})\tilde{Y}_{t-1} + \tilde{\epsilon}_t. \tag{A.1}$$

One can show that the Markov chain defined by (A.1) is irreducible and aperiodic (Chan & Tong, 1985).

By condition (3), for an arbitrary small $\eta > 0$, there exists a constant M_η such that $|g_{ij}(x) - c_{ij}x| \leq \eta|x|$ for any i, j and x with $|x| > M_\eta$. On the other hand, by condition (2), there exists a bound Δ_η such that $|g_{ij}(x)| \leq \Delta_\eta$ for all i, j and x with $|x| < M_\eta$. We have just shown that

$$|g_{ij}(x) - c_{ij}x| \leq \min(\eta|x|, \Delta_\eta).$$

Define the matrix

$$C = \begin{pmatrix} \tilde{A}_1 & \tilde{A}_2 & \cdots & \tilde{A}_{p-1} & \tilde{A}_p \\ I_d & 0 & \cdots & 0 & 0 \\ 0 & I_d & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & I_d & 0 \end{pmatrix}.$$

Denote $\mathcal{T}(\tilde{\mathbf{Y}})\tilde{\mathbf{Y}}=(\mathbf{v}_1, \dots, \mathbf{v}_p)$, where each \mathbf{v}_j is a d -dimensional vector. We have $\mathbf{v}_j = \mathbf{Y}_{j-1}$ for $j=2, \dots, p$. For $j=1$, denote the i th entry of \mathbf{v}_1 as v_{i1} , then

$$|v_{i1}| = \left| \sum_{j=1}^p g_{ij}(\boldsymbol{\alpha}_{ij}^T \mathbf{Y}_j) \right| \leq \left| \sum_{j=1}^p c_{ij} \boldsymbol{\alpha}_{ij}^T \mathbf{Y}_j \right| + \min \left\{ \eta \sum_{j=1}^p |\boldsymbol{\alpha}_{ij}^T \mathbf{Y}_j|, p\Delta_\eta \right\}$$

Therefore,

$$\|\mathcal{T}(\tilde{\mathbf{Y}})\tilde{\mathbf{Y}}\| \leq \|C\tilde{\mathbf{Y}}\| + \eta \sum_{j=1}^p \|A_j \mathbf{Y}_j\| + p\Delta_\eta \|\mathbf{1}_d\|,$$

where $\mathbf{1}_d$ is a d -dimensional vector of 1s. By standard linear algebra, one can show that the characteristic polynomial of the matrix C , defined as $\det(\xi I_{dp} - C)$, equals to (4). The roots of (4) are all inside the unit circle. Hence, the largest eigenvalue of C , denoted as ξ_1 , has a module less than 1. Therefore,

$$\begin{aligned} E(\|\tilde{\mathbf{Y}}_{t+1}\| \|\tilde{\mathbf{Y}}_t\|) &= E\{\|\mathcal{T}(\tilde{\mathbf{Y}}_t)\tilde{\mathbf{Y}}_t + \tilde{\boldsymbol{\epsilon}}_{t+1}\| \|\tilde{\mathbf{Y}}_t\|\} \\ &\leq \|\mathcal{T}(\tilde{\mathbf{Y}}_t)\tilde{\mathbf{Y}}_t\| + E(\|\tilde{\boldsymbol{\epsilon}}_{t+1}\|) \\ &\leq \|C\tilde{\mathbf{Y}}_t\| + \eta \sum_{j=1}^p \|A_j \mathbf{Y}_j\| + p\Delta_\eta \|\mathbf{1}_d\| + E(\|\boldsymbol{\epsilon}_{t+1}\|) \\ &\leq |\xi_1| \times \|\tilde{\mathbf{Y}}_t\| + \eta \sum_{j=1}^p \|A_j \mathbf{Y}_j\| + p\sqrt{d}\Delta_\eta + E(\|\boldsymbol{\epsilon}_{t+1}\|). \end{aligned}$$

Next, we can choose η to be sufficiently small such that

$$E(\|\tilde{\mathbf{Y}}_{t+1}\| \|\tilde{\mathbf{Y}}_t\|) \leq \xi^* \|\tilde{\mathbf{Y}}_t\| + p\sqrt{d}\Delta_\eta + E(\|\boldsymbol{\epsilon}_{t+1}\|),$$

for some $\xi^* < 1$. As $p\sqrt{d}\Delta_\eta + E(\|\boldsymbol{\epsilon}_{t+1}\|)$ is a finite number, one can find a range \mathcal{M} such that $E(\|\tilde{\mathbf{Y}}_{t+1}\| \|\tilde{\mathbf{Y}}_t\|)$ is bounded by a constant when $\|\tilde{\mathbf{Y}}_t\| \leq \mathcal{M}$, and strictly less than $(1 - \rho)\|\tilde{\mathbf{Y}}_t\|$ when $\|\tilde{\mathbf{Y}}_t\| > \mathcal{M}$. The theorem follows from application of Tweedie’s criterion (lemma 1).