A Single Index Model procedure for interpolation intervals in Time Series

Andrés M. Alonso [a,*]

[a] *Department of Statistics*

Universidad Carlos III de Madrid, Spain

Ana E. Sipols [b]

[b] *Department of Statistics and Operational Research*

Universidad Rey Juan Carlos, Spain

Silvia Quintas [c]

[c] *Department of Sciences*

Universidad Europea de Madrid, Spain

Abstract

In this paper we propose a procedure that uses single index model to construct interpolation intervals for a general class of linear processes. We present an extensive Monte Carlo experiment studying the finite sample properties of this procedure. Finally, we illustrate the performance of the proposed method with a real data example.

Key words: single index model, interpolation intervals, sieve bootstrap.

1 Introduction

In Alonso and Sipols (2008), a sieve bootstrap procedure for constructing interpolation intervals for a general class of linear processes is proposed. This procedure treats interpolation as a prediction with restrictions on future values, and it uses nonparametric estimators of the conditional quantiles. As usual, the behavior of these nonparametric estimators depends on the number of restrictions; it is known that nonparametric estimators suffer from the curse of dimensionality, i.e., the estimators become worse as the dimension increase. In this paper we propose to use a procedure based on a single index model, in an attempt to minimize the effect of the number of variables or restrictions.

The single index models (SIM) are widely used in applied quantitative science. For example censored Tobit models, binary choice models or errors-in-variables models (for details and more examples see Ichimura, 1993). The SIM consists of resuming the effect of the $k-$covariate variable $\boldsymbol{X}$ in a unique variable, called index, in order to capture most information regarding the relation between a response variable Y and the set of explanatory variables $\boldsymbol{X}$, thereby avoiding the "curse of dimensionality". Specifically, given a response variable Y and a set of explanatory variables $\boldsymbol{X}$, the model can be written as

$$Y = \eta(\boldsymbol{\lambda}'\mathbf{X}) + \epsilon, \tag{1}$$

where $\boldsymbol{\lambda}$ is a vector of $k \times 1$ parameters, with $||\boldsymbol{\lambda}|| = 1$ and $E(\epsilon|\boldsymbol{X}) = 0$. Both the link function η and the parameter vector $\boldsymbol{\lambda}$ are unknown.

* Corresponding author. Department of Statistics

Universidad Carlos III de Madrid, 28903 Getafe (Madrid), Spain

Tel: +34-916249591; fax: +34-916249849.

Email address: andres.alonso@uc3m.es (Andrés M. Alonso).

Using the results of Guerrero and Peña (2000), the following expression was obtained by Alonso (2001) for a general class of time series linear model with Gaussian innovations:

$$\boldsymbol{X}_\tau \,|\boldsymbol{X}_O \sim \mathcal{N}(-(\boldsymbol{H}'\boldsymbol{\Sigma}^{-1}\boldsymbol{H})^{-1}\boldsymbol{H}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X}, \boldsymbol{H}'\boldsymbol{\Sigma}^{-1}\boldsymbol{H}), \tag{2}$$

where $\boldsymbol{X}_\tau = (X_{\tau_1}, X_{\tau_2}..., X_{\tau_m})$ is the missing observations vector, $\boldsymbol{X}$ is the complete observed series with zeros in positions $(\tau_1, \tau_2, ..., \tau_m)$, $\boldsymbol{X}_O = \boldsymbol{X} \setminus \boldsymbol{X}_\tau$. $\boldsymbol{H}$ is a $T \times m$ matrix such that $H_{\tau_i,i} = 1$ and $H_{i,j} = 0$ in another case, and $\boldsymbol{\Sigma}$ is the $T \times T$ autocovariance matrix.

Notice that the above expression implies that the conditional distribution of $\boldsymbol{X}_\tau \,|\boldsymbol{X}_O$ depends on $\boldsymbol{X}_O$ only through the index $-(\boldsymbol{H}'\boldsymbol{\Sigma}^{-1}\boldsymbol{H})^{-1}\boldsymbol{H}'\boldsymbol{\Sigma}^{-1}\mathbf{X}$. This result gave rise to using the SIM methodology which takes into account past and future observations, $\boldsymbol{X}_P = (X_1, X_2, ..., X_{\tau_1-1})$, $\boldsymbol{X}_F = (X_{\tau_m+1}, X_{\tau_m+2}, ..., X_T)$ respectively, as explanatory variables, and missing observations, $\boldsymbol{X}_\tau$, as the response variable.

In this paper we will assume that the conditional distribution of $\boldsymbol{X}_\tau$ depends on the sample only through an index, i.e.

$$f\left(\boldsymbol{X}_\tau \,|\boldsymbol{X}_P, \boldsymbol{X}_F\right) = f\left(\boldsymbol{X}_\tau \,|\boldsymbol{\lambda}'(\boldsymbol{X}'_P, \boldsymbol{X}'_F)'\right). \tag{3}$$

The major advantage of the above assumption is that the function f can be estimated by nonparametric procedures, that are not affected by the dimension of the vector since $\boldsymbol{\lambda}'(\mathbf{X}'_P, \mathbf{X}'_F)$ has dimension equal to one.

The rest of the paper is organized into the following sections. Section 2 briefly describes the SIM estimation method proposed by Yu and Ruppert (2002). Section 3 is based on Alonso and Sipols (2008), where the bootstrap procedure for constructing interpolation intervals is described by substituting

the standard nonparametric techniques with the SIM procedure. Section 4 presents the results of an extensive Monte Carlo study and Section 5 illustrates the behavior of the proposed procedure with a real data example. A brief summary of conclusions is given in Section 6.

2 Single Index Model estimation procedure

In general, SIM estimation takes place in two stages. First the vector of coefficients, $\boldsymbol{\lambda}$, is estimated. Then, using the index values as single explanatory variable, the link function η is estimated using standard non-parametric techniques. There are a variety of methods to estimate the single-index models parameters. Geenens and Delecroix (2006) give a review of many them; comparing M-estimators, like semiparametric least squares (Ichimura, 1993) and semiparametric maximum likelihood (Klein and Spady, 1993; Delecroix, Härdle and Hristache, 2003), and direct estimators like average derivative estimator (ADE): unweighted average derivatives (UADE)(Härdle and Stoker, 1989) and density-weighted average derivatives (DWADE)(Powell, Stock and Stoker, 1989). Their simulation study points out that once the sample size grows, the M-estimators outperform the direct estimators. In our case we will always have large sample sizes because the SIM estimates are calculated on the bootstrap samples.

Some of the problems and weaknesses presented in the above mentioned methods are dealt with by Yu and Ruppert (2002), who propose the use of penalized spline (which may be classified as an M-estimation approach). Some of the advantages of this method are that it is rapid and computationally stable, and it uses standard nonlinear least squares software.

The Yu and Ruppert's method involves the simultaneous estimation of all parameters in the following partially linear single index model:

$$y_i = \eta(\boldsymbol{\lambda}'\mathbf{x}_i) + \boldsymbol{\beta}'\mathbf{z}_i + \epsilon_i, \tag{4}$$

where $\mathbf{x}_i \in \mathbb{R}^d$, $\mathbf{z}_i \in \mathbb{R}^{d_z}$, $y_i \in \mathbb{R}$, $\boldsymbol{\lambda} \in \mathbb{R}^d$ (unknown single index parameter), $\boldsymbol{\beta} \in \mathbb{R}^{d_z}$ (unknown linear parameter)and $\eta : \mathbb{R} \to \mathbb{R}$ is an unknown univariate function; $\{\epsilon_i\}$ is an independent error process with zero mean and variance σ_0^2. $\{\epsilon_i\}$ is assumed independent of $\{(\mathbf{x}_i, \mathbf{z}_i)\}$. In order to have an identifiable model, the following conditions are imposed: $||\boldsymbol{\lambda}|| = 1$ and the first nonzero element of $\boldsymbol{\lambda}$ is positive. In our case, we have a simpler model since we can assume that $\boldsymbol{\beta} \equiv 0$.

Their method assume that the link function can be written by

$$\eta(u) = \delta + \delta_1 u + \ldots + \delta_p u^p + \sum_{k=1}^{K} \delta_{p+k}(u - \kappa_k)_+^p, \tag{5}$$

where $\mathbf{B}(u) = (1 \; u \ldots u^p \; (u - \kappa_1)_+^p \ldots \; (u - \kappa_K)_+^p)'$ is a spline basis, $\boldsymbol{\delta} = (\delta, \delta_1, \ldots \delta_{p+K})'$ is the spline coefficient vector and $\{\kappa_k\}_{k=1}^K$ are the spline knots. The knots are chosen at equally spaced sample quantiles of the estimated index values $\boldsymbol{\lambda}'\mathbf{x}$.

If we denote $\mathbf{v}_i = (\mathbf{x}_i'\mathbf{z}_i')'$ and $\boldsymbol{\theta} = (\boldsymbol{\lambda}'\boldsymbol{\beta}'\boldsymbol{\delta}')'$, then the spline model can be written as:

$$\eta(u) = \boldsymbol{\delta}'\mathbf{B}(u), \tag{6}$$

and the mean function $E[y_i|\mathbf{v}_i; \boldsymbol{\theta}]$ is given by:

$$m(\mathbf{v}_i; \boldsymbol{\theta}) = \boldsymbol{\delta}'\mathbf{B}(\boldsymbol{\lambda}'\mathbf{x}_i) + \boldsymbol{\beta}'\mathbf{z}_i. \tag{7}$$

The model (4)-(7) can be estimated by the penalized least squares method

which minimizes the following expression:

$$Q_{n,\alpha}(\boldsymbol{\theta}) = n^{-1} \sum_{i=1}^{n} \{y_i - m(\mathbf{v}_i; \boldsymbol{\theta})\}^2 + \alpha \boldsymbol{\delta}' \mathbf{D} \boldsymbol{\delta}, \tag{8}$$

where $\mathbf{D}$ is an appropriate positive semidefinite symmetric matrix and $\alpha \geq 0$ is a penalty parameter. Once we have an estimate of $\boldsymbol{\theta}$, we can use a standard nonparametric estimator for the unknown function η (see details in Yu and Ruppert, 2002).

3 A time series bootstrap procedure for interpolation intervals

Alonso and Sipols (2008) provide a procedure for calculating the interpolation intervals, using a nonparametric estimator of the conditional distribution function. This section briefly describes their procedure, which is a modification of that proposed by Alonso, Peña and Romo (2002) as applied to time series with missing observations.

Let $\{X_t\}_{t \in Z}$ be a stationary process with $E[X_t] = \mu_X$ which admits a $MA(\infty)$ representation

$$X_t - \mu_X = \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j}, \quad \psi_0 = 1, \quad t \in Z \tag{9}$$

where $\{\varepsilon_t\}_{t \in \mathbb{Z}}$ is a sequence of independent and identically distributed random variables with $E(\varepsilon_t) = 0$ and $V(\varepsilon_t) = \sigma^2$. The coefficients $\{\psi_j\}_{j=0}^{+\infty}$ have, at most, a polynomial decay, and the polynomial $\Psi(z) = \sum_{j=0}^{\infty} \psi_j z^j$ is bounded and different from zero in $|z| \leq 1$. Then $\{X_t\}_{t \in Z}$ admits an autoregressive representation:

$$\sum_{j=0}^{\infty} \phi_j (X_{t-j} - \mu_X) = \varepsilon_t, \quad \phi_0 = 1 \tag{10}$$

and the polynomial $\Phi(z) = \sum_{j=0}^{\infty} \phi_j z^j$ is bounded and different from zero in $|z| \leq 1$. This $AR(\infty)$ representation motivates the sieve bootstrap proposed

by Kreiss (1988) and Bühlmann (1997).

The procedure proposed by Alonso and Sipols (2008) consists of the following steps:

1. Given a sample $X = (X_1, X_2, ..., X_n)$, select the autoregressive approximation of order $p = p(n)$ using the Bayesian information criteria (BIC, see, Schwarz 1978).

The BIC criteria is used in order to obtain a parsimonious model. Of course, other model selection criteria can be used, e.g., the AICC criterion proposed by Hurvich and Tsai (1989).

2. Obtain the parameter's estimates: $\widehat{\phi}_p = \left(\widehat{\phi}_1, \widehat{\phi}_2, ..., \widehat{\phi}_p\right)'$ by means of the resolution of the following Yule-Walker equations:

$$\widehat{\mathbf{\Gamma}}_p \widehat{\phi}_p = -\widehat{\gamma}_p, \tag{11}$$

where $\widehat{\mathbf{\Gamma}}_p = \left[\widehat{R}\left(i-j\right)\right]_{i,j=1}^{p}$, $\widehat{\gamma}_p = \left(\widehat{R}(1), ..., \widehat{R}(p)\right)'$, and

$$\widehat{R}(j) = \sum_{t=1}^{n-|j|} a_t a_{t+|j|} \left(Y_t - \overline{Y}\right)\left(Y_{t+|j|} - \overline{Y}\right) \Big/ \sum_{t=1}^{n-|j|} a_t a_{t+|j|},$$

where $Y_t = a_t X_t$, a_t represents the state of the observation at time t, i.e., $a_t = 1$ if X_t is observed and $a_t = 0$ if X_t is a missing observation.

The estimators (11) were proposed by Parzen (1963) and its asymptotical properties have been studied by Dunsmuir and Robinson (1981). If the number of missing observations remains constant with the sample size n, then the effect of substituting $\sum_{t=1}^{n-|j|} a_t a_{t+|j|}$ by n in (11) is asymptotically zero.

3. Calculate the residuals:

$$\widehat{\varepsilon}_t = \sum_{j=0}^{p} \widehat{\phi}_j \left(X_{t-j} - \overline{X}\right) \text{ with } \widehat{\phi}_0 = 1,\ t \in \Upsilon_p, \tag{12}$$

where $\Upsilon_p = \{t : \prod_{i=0}^{p} a_{t-i} = 1\}$.

4. Obtain the empirical distribution function of the centered residuals:

$$\widetilde{F}_{\widetilde{\varepsilon}}(x) = (\#\Upsilon_p - p)^{-1} \sum_{t \in \Upsilon_p} I\left(\widetilde{\varepsilon}_t \leq x\right), \tag{13}$$

where $\widetilde{\varepsilon}_t = \widehat{\varepsilon}_t - \widehat{\varepsilon}^{(\cdot)}$ and $\widehat{\varepsilon}^{(\cdot)} = (\#\Upsilon_p - p)^{-1} \sum\limits_{t \in \Upsilon_p} \widehat{\varepsilon}_t$.

5. Obtain a resample of *i.i.d.* observations: $\varepsilon_t^* \sim \widetilde{F}_{\widetilde{\varepsilon}}$.

6. Define X_t^* by the relation:

$$\sum_{t=0}^{p} \widehat{\phi}_j \left(X_{t-j}^* - \overline{X}\right) = \varepsilon_t^*, \tag{14}$$

where the first p observations are equal to $\overline{X}$.

7. Given $(X_1^*, \ldots, X_n^*)$, obtain $Y_t^* = a_t X_t^*$, and calculate the autoregressive parameter estimates, $\widehat{\boldsymbol{\phi}}_p^*$, as in step 2.

8. Calculate the future observations by the relation:

$$X_{T'+h}^* - \bar{X} = -\sum_{j=1}^{p} \widetilde{\phi}_j^* (X_{T'+h-j}^* - \bar{X}) + \varepsilon_{T'+h}^*, \tag{15}$$

where $h = 1, 2, \ldots, T - T'$, $X_t^* = X_t$ for $T' - p \leq t \leq T'$, and $T' = \max\{t \in \Upsilon_p : t < \tau\}$.

From step 8, the bootstrap distribution function for the $(T - T') \times 1$ vector $(X_{T'+1}^*, X_{T'+2}^*, \ldots, X_T^*)$ conditional to $(X_{T'-p}^* = X_{T'-p}, X_{T'-p+1}^* = X_{T'-p+1}, \ldots, X_{T'}^* = X_{T'})$ is obtained, i.e., the distribution of the "future" observations from time $T'+1$ up to time T conditional to its past. From this multidimensional distribution one could obtain $F_{X_\tau^*}^*$, the distribution of X_τ^* conditional to $\boldsymbol{Y}_{T'+1}^{*T} \setminus Y_\tau^* = \boldsymbol{Y}_{T'+1}^{T} \setminus Y_\tau$. Notice that $[T'+1, T]$ includes all the missing observations. In the particular case of m consecutive missing observations $T' = \max\{t \in \Upsilon_p : t < \tau\} = \tau - 1$, where τ is the index of the first missing observation in the block.

The steps from 5 to 8 are repeated B times in order to obtain an estimation of $F^*_{X^*_\tau}$ using one of the following nonparametric estimators (e.g., see Cao et al, 1997):

$$F^*_{X^*_\tau}(x|\boldsymbol{y}) = \frac{\sum_{b=1}^{B} I\left(X^{*(b)}_\tau \leq x\right) \boldsymbol{K}\left((\boldsymbol{y} - \boldsymbol{Y}^{*(b)\,T}_{T'+1})/h\right)}{\sum_{b=1}^{B} \boldsymbol{K}\left((\boldsymbol{y} - \boldsymbol{Y}^{*(b)\,T}_{T'+1})/h\right)} \tag{16}$$

or

$$F^*_{X^*_\tau}(x|\boldsymbol{y}) = \frac{\sum_{b=1}^{B} \mathbb{K}\left((x - X^{*(b)}_\tau)/h\right) \boldsymbol{K}\left((\boldsymbol{y} - \boldsymbol{Y}^{*(b)\,T}_{T'+1})/h\right)}{\sum_{b=1}^{B} \boldsymbol{K}\left((\boldsymbol{y} - \boldsymbol{Y}^{*(b)\,T}_{T'+1})/h\right)}, \tag{17}$$

where $\boldsymbol{K}$ is a Kernel with dimension $T - T' - 1$, and $\mathbb{K}(x) = \int_{-\infty}^{x} \boldsymbol{K}(s)ds$. For simplicity in the notation, in (16) and (17) Y_τ is omitted in $\boldsymbol{Y}^T_{T'+1} \setminus Y_\tau$. From these estimators of the conditional distribution, the conditional quantiles can be derived. Alternatively, since our interest is in obtaining interpolation intervals, the following nonparametric quantile estimator can be used:

$$Q^*(\theta)(\boldsymbol{y}) = \frac{Q^{*L}(\theta)(\boldsymbol{y}) + Q^{*U}(\theta)(\boldsymbol{y})}{2}, \tag{18}$$

where

$$Q^{*L}(\theta)(\boldsymbol{y}) = \min_{1 \leq b \leq B} \left\{ X^{*(b)}_\tau / \sum_{\ell=1}^{B} \mathbf{1}(X^{*(\ell)}_\tau \leq X^{*(b)}_\tau) W_\ell(\boldsymbol{y}) \leq \theta \right\}, \tag{19}$$

$$Q^{*U}(\theta)(\boldsymbol{y}) = \min_{1 \leq b \leq B} \left\{ X^{*(b)}_\tau / \sum_{\ell=1}^{B} \mathbf{1}(X^{*(\ell)}_\tau \leq X^{*(b)}_\tau) W_\ell(\boldsymbol{y}) > \theta \right\} \tag{20}$$

and $W_\ell(y)$ are the nonparametric weights (e.g., see Cao et al, 1997).

3.1 SIM methodology

We propose a modification of the Alonso and Sipols (2008) procedure based on SIM methodology assuming that the conditional distribution $F_{X_\tau}(x|\boldsymbol{y})$ can be approximated by $F_{X_\tau}(x|\boldsymbol{\lambda'y})$.

The steps of the previous algorithm are modified as follows: (i) It is not

necessary to modify steps 1–6 of the procedure proposed in Section 3, and (*ii*) Steps 7 and 8 are not carried out since our procedure works with the bootstrap series generated in step 6.

The steps 5 and 6 are repeated B times in order to obtain estimations of the indexes, $\boldsymbol{\lambda}_\tau$ for $\tau = \tau_1, \tau_2, \ldots, \tau_m$ using the estimation procedure described in Section 2, i.e., we obtain $\widehat{\boldsymbol{\lambda}}_\tau$ by estimating the model (4)-(8) with response variable $y = X^*_\tau$, explanatory variable $\mathbf{x} = [\mathbf{X}^{*\prime}_P, \mathbf{X}^{*\prime}_F]'$ and $\beta = 0$.

Then, an estimation of $F^*_{X^*_\tau}$ is obtained using the following nonparametric estimator:

$$F^*_{X^*_\tau}(x|\boldsymbol{y}) = \frac{\sum_{b=1}^{B} I\left(X^{*(b)}_\tau \leq x\right) K\left((\widehat{\boldsymbol{\lambda}}'_\tau \boldsymbol{y}' - \widehat{\boldsymbol{\lambda}}'_\tau [\mathbf{X}^{*\prime}_P, \mathbf{X}^{*\prime}_F]')/h\right)}{\sum_{b=1}^{B} K\left((\widehat{\boldsymbol{\lambda}}'_\tau \boldsymbol{y}' - \widehat{\boldsymbol{\lambda}}'_\tau [\mathbf{X}^{*\prime}_P, \mathbf{X}^{*\prime}_F]')/h\right)} \tag{21}$$

where $\mathbf{X}^*_P$ and $\mathbf{X}^*_F$ are the previously defined series of past and future observations, respectively; and $\widehat{\boldsymbol{\lambda}}_\tau$ is the estimated index for the missing value at position τ. Notice that the estimator (21) entails that the distribution $F_{X_Y}(x|\boldsymbol{y})$ can be approximated by $F_{X_Y}(x|\boldsymbol{\lambda}'\boldsymbol{y})$, i.e., similar to a single index model.

It is noteworthy that, in this proposal, the kernel is evaluated in a one-dimensional variable, thus avoiding the problem of dimensionality of estimators (16) and (17).

4 Simulation Results

Alonso and Sipols (2008) considered three different alternatives, since the behavior of the nonparametric estimator (16) and (17) depends on the dimension of the conditioning vector $\boldsymbol{Y}^T_{T'+1} \setminus Y_\tau$: (CT) conditioning up to observation T; (CP) conditioning up to observation $T' + m + \widehat{p}$ where m is the number

of missing observations and $\widehat{p}$ is the selected autoregressive order; and (CO) conditioning up to observation $T' + m + p_{opt}$ where p_{opt} is the order that minimize the mean squared error (MSE) of the nonparametric conditional mean as interpolator of X_τ.

Considering that the best results were obtained when using CP and CO, in this section we will compare the CP and CO approaches with the SIM methodology. For the SIM procedure, we will consider two cases: (SIM) conditioned on all past ($\boldsymbol{X}_P$) and future ($\boldsymbol{X}_F$) observations, and (SIMP) conditioned up to future observation $\tau_m + \widehat{p}$ and from past observation $\tau_1 - \widehat{p}$, where τ_1 and τ_m are the index of the first and last missing observation, respectively; and $\widehat{p}$ is the selected autoregressive order.

A simulation study is therefore carried out in order to evaluate how our methods (SIM, SIMP) perform compared to (CP, CO) and the additive outlier approach (AO) proposed by Gómez et al (1997, 1999). For this last approach, we use the program TRAMO (**T**ime series **R**egression with **A**RIMA noise, **M**issing observations and **O**utliers) developed by Gómez and Maravall (1996). This program was used in Gómez et al (1997, 1999) and it provides the missing values estimation and its standard error. Then, assuming normality, a Gaussian interpolation interval can be derived.

We have run a simulation experiment with series of length 100 generated from the following models:

$$\text{Model 1: } (1 - 0.8B)X_t = \varepsilon_t \tag{22}$$

and

$$\text{Model 2: } X_t = (1 - 0.7B)\varepsilon_t, \tag{23}$$

where the $\{\varepsilon_t\}$ are i.i.d. F_ε. The error distributions F_ε considered are the

standard normal $\mathcal{N}(0,1)$, a shifted exponential distribution with zero mean and scale parameter equal to one, and a contaminated distribution 0.9 F_1 + 0.1 F_2 with $F_1 \sim \mathcal{N}(-1,1)$ and $F_2 \sim \mathcal{N}(9,1)$.

We have considered different patterns of missing data: (i) one missing observation at a fixed position $t = 10$, 50, 90; and (ii) five consecutive missing observations at positions $t = 45$ to 49. These models and patterns of missing data were considered by Gómez et al (1997).

In order to approximate the distribution $F^*_{X^*_\tau}$ conditional to $Y^{*T}_{T'+1} \setminus Y^*_\tau = Y^T_{T'+1} \setminus Y_\tau$, for CP and CO, Alonso and Sipols (2008) used the following bandwidth:

$$h = \left(\frac{4}{2+d}\right)^{1/(d+4)} B^{-1/(d+4)} s, \tag{24}$$

where $s^2 = d^{-1} \sum_i s_{ii}$, s_{ii} are the diagonal elements of the covariance matrix, S, of vector $(X^{*(b)}_{T'+m+1}, X^{*(b)}_{T'+m+2}, \ldots, X^{*(b)}_{T'+m+d})$, d is the number of conditioning observations, and B is the number of bootstrap replications.

For SIM and SIMP, we use the previous expression with $d = 1$ since we are considering a one-dimensional variable, and s is the standard deviation of $\boldsymbol{\lambda}_\tau[\mathbf{X}^{*(b)}_P, \mathbf{X}^{*(b)}_F]$. In this case, the bandwidth formula coincides with Silverman's rule of thumb (Härdle et al(2004)):

$$h = \left(\frac{4}{3}\right)^{1/5} B^{-1/5} s \tag{25}$$

In all methods we have considered $B = 1000$ bootstrap replications. To compare the different interpolation intervals, we use their mean coverage and the proportions of observations lying out to the left and to the right of the interval. Tables 1-3 present the coverage results for Model 1 using the three distribution error, a nominal level of 90% and the above mentioned patterns

of missing observations. Tables 4-6 present the results for Model 2.

The main conclusions are the following:

- SIM does not always outperform CP and CO in terms of mean coverage.
- SIMP outperforms SIM, CP and CO methods in terms of mean coverage.
- SIM, SIMP, CP and CO methods outperform the AO approach in terms of mean coverage.

From Tables 1-6 we can derive the following results: (*i*) In some cases the mean coverage obtained with SIM is outperformed by the other methods. That is due to the fact that least squares methods for estimating index $\boldsymbol{\lambda}$ are less accurate if there are a large number of explanatory variables; (*ii*) SIMP outperforms the SIM in terms of mean coverage in almost all the cases, i.e., SIMP attains a coverage closer to the nominal value. Notice that SIMP takes into account a smaller number of variables ; (*iii*) SIMP outperforms CO and CP as it is known that nonparametric estimators suffer from the curse of dimensionality, i.e., the estimators becomes worse as the dimension increases; (*iv*) Regarding the different positions of an isolated missing observation, we observe that SIM and SIMP perform similarly in the considered positions, the exception being the MA model with contaminated innovations; (*v*) Regarding the pattern of five consecutive missing observations, in general, we observe a better coverage at the beginning and at the end of the missing block; (*vi*) The SIM, SIMP, CP and CO approaches outperform the AO approach in almost all the considered cases, which causes a lower coverage of their interpolation intervals. Moreover, the AO approach fails in the case of non-Gaussian error distribution having, in some cases, coverage lower than 80%.

5 Real data example

Here we study the monthly sales (in kiloliters) of red wine by Australian winemakers (1980-1992), considered in Brockwell and Davis (2001). In order to illustrate the proposed procedure, we consider a pattern of twelve consecutive missing observations.

In Figure 1, we present the interpolation relative errors using the proposed procedure, SIMP, the CO procedure with $p_{opt} = 4$, and using a fully parametric method selected by using the TRAMO program developed by Gómez and Maravall (1996). We observe that SIMP and AO methods have a similar performance in terms of relative errors. The SIMP has a MSE, 0.0017, comparable to the obtained MSE, 0.0013, using TRAMO. However, if we compare them in terms of coverage, the SIMP approach obtains 94% and the AO approach obtains 69.2%. We observe that SIMP and CO methods have a similar performance in terms of relative errors. The MSE obtained by SIMP is comparable to the obtained MSE, 0.0016, using CO and if we compare them in terms of coverage, the SIMP approach obtains 94% and the nonparametric conditional mean approach obtains 93%. These results are consistent with the simulation results reported in Section 4. In this real application the results by SIMP and CO are quite similar, but in the Monte Carlo simulations SIMP outperforms CO in all the cases.

In Figure 2 we illustrate the results of the proposed procedure for the pre-

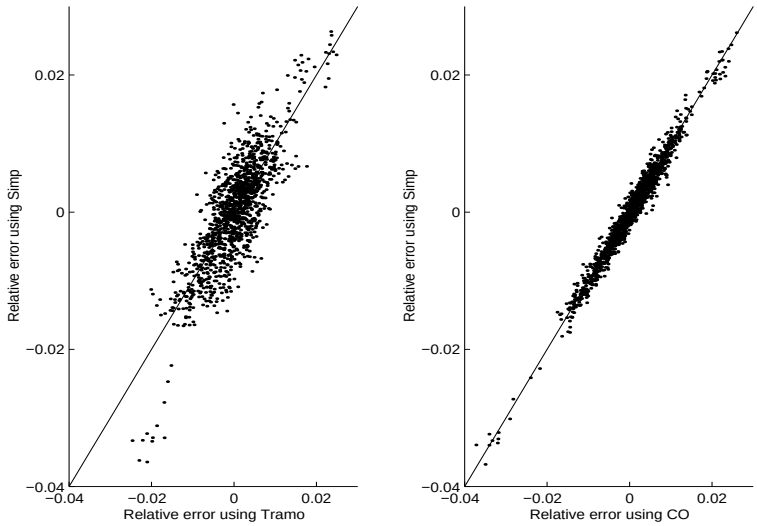

Figure 1. Interpolation relative errors using the CO, AO and SIMP approaches.

vious pattern of twelve consecutive missing observations at positions 67,...,78. We can see that interpolation intervals capture the underlying dynamics in the data.

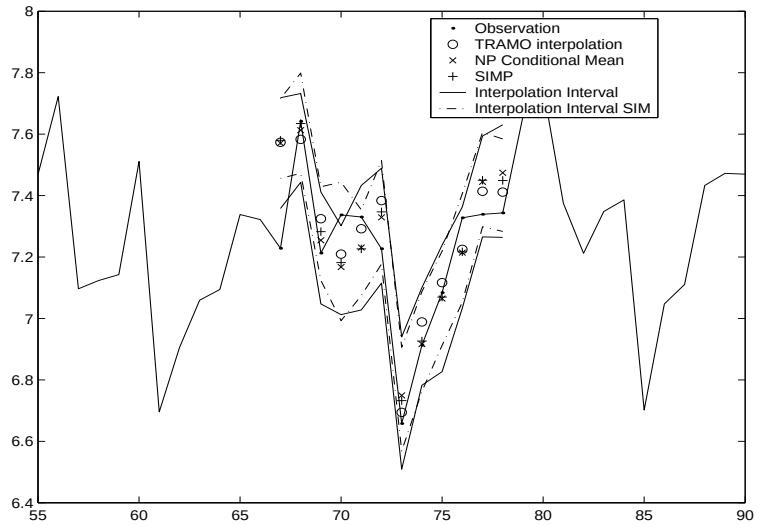

Figure 2. Interpolation results for series Australian red wine sale data (1980-1992).

6 Conclusions and Extensions

In this article we proposed a single-index model methodology in order to construct interpolation intervals. The goal achieved by our approach is that it avoids the "curse of dimensionality" in nonparametric estimations, outperforming the results of Alonso and Sipols (2008). The result is supported by our simulations and real data studies.

References

Alonso, A.M (2001). Resampling techniques and missing values in time series. *PhD thesis*, Universidad Carlos III de Madrid.

Alonso, A. M., Peña, D. and Romo, J. (2002). Forecasting time series with sieve bootstrap, *Journal of Statistical Planning and Inference*, **100**, 1-11.

Alonso, A.M. and Sipols, A.E. (2008). A time series bootstrap procedure for interpolation intervals. *Computational Statistics and Data Analysis*, **52**, 1792-1805.

Beveridge, S. (1992). Least squares estimation of missing values in time series. *Communications in Statistics Theory and Methods*, **21**, 3479-3496.

Brockwell, P.J. and Davis, R.A. (2001). Introduction to time series and forecasting. *Springer-Verlag, New York.*

Bühlmann, P. (1997) Sieve bootstrap for time series, *Bernoulli*, **3**, 123-148.

Cao, R., Delgado, M. A. and González-Manteiga, W. (1997). Nonparametric curve estimation: An overview, *Investigaciones Económicas*, **21**, 209-252.

Delecroix, M., Härdle, W. and Hristache, M. (2003). Efficient estimation in conditional single-index regression. *Journal of Multivariate Analysis*, **86**(2), 213-226.

Dunsmuir, W. and Robinson, P. (1981). Asymptotic theory for time series containing missing and amplitude modulated observations, *Sankhyā. The Indian Journal of Statistics Series A*, **43**, 260-281.

Geenens, G. and Delecroix, M. (2006). A survey about Single-Index Models

theory. *International Journal of Statistics and Systems*, **1**, 203-230.

Gómez, V. and Maravall, A. (1996). Programs TRAMO (Time Series Regression with ARIMA noise, Missing observations and Outliers) and SEATS (Signal Extraction in ARIMA Time Series). Instruction for the user, Working Paper 9628, Bank of Spain, Madrid.

Gómez, V., Maravall, A. and Peña, D. (1997). Missing observations in ARIMA models. Working paper 9701, Banco de España, Madrid.

Gómez, V., Maravall, A. and Peña, D. (1999). Missing observations in ARIMA models: Skipping approach versus, additive outlier approach, *Journal of Econometrics*, **88**, 341-363.

Guerrero, V. M. and Peña, D. (2000). Linear combination of restrictions and forecasts in time series analysis. *Journal of Forecasting*, **19**, 103-122.

Härdle, W. and Stoker, T. M. (1989). Investigating smooth multiple regression by the method of average derivatives. *Journal of American Statistical Association*, **84**, 986-995.

Härdle, W., Müller, M., Sperlich, S. and Werwatz, A. (2004). Nonparametric and Semiparametric Models. *Springer, Berlin.*

Hurvich, C.M. and Tsai, C.-L. (1989). Regression and time series model selection in small samples, *Biometrika*, **76**, 297-307.

Ichimura, H. (1993). Semiparametrix least square (SLS) and weighted SLS estimation of single-index models, *Journal of Econometrics*, **58**, 71-120.

Klein, R. and Spady, R. (1993). An efficient semiparametric estimator for binary response models, *Econometrica*, **61**, 387-421.

Kreiss, J.-P. (1988). Asymptotic Statistical Inference for a class of Stochastic Processes, Habilitationsschrift, University of Hamburg.

Ljung, G.M. (1989). A note on the estimation of missing values in time series. *Communications in Statistics. Simulation and Computation*, **18**, 459-465.

Ljung, G. M. (1993). On outliers detection in time series. *Journal of the Royal Statistical Society. Series B*, **55**, 559-567.

Parzen, E. (1963). On spectral analysis with missing observations and amplitude modulation, *Sankhyā. The Indian Journal of Statistics Series A*, **25**, 383-392.

Peña, D. and Maravall, A. (1991). Interpolation, outliers, and inverse autocorrelations. *Communications in Statistics. Theory and Methods*, **20**, 3175-3186.

Pourahmadi, M. (1989). Estimation and interpolation of missing values of a stationary time series. *Journal of Time Series Analysis*, **10**, 149-169.

Powel, J. L., Stock, J. H. and Stoker, T. M. (1989). Semiparametric estimation of index coefficients. *Econometrica*, **57**(6), 1403-1430.

Schwarz, G. (1978). Estimating the dimension of a model, *Annals of Statistics*, **6**, 461-464.

Wincek, M. A. and Reinsel, G. C. (1986). An exact maximum likelihood estimation procedure for regression-ARMA time series models with possibly nonconsecutive data. *Journal of the Royal Statistical Society. Series B*, **48**, 303-313.

Yu, Y. and Ruppert, D. (2002). Penalized Spline Estimation for Partially Lin-

ear Single-Index Models. *Journal of the American Statistical Association*, **97**, 1042-1054.

Table 1
Simulation results for Model AR using Gaussian innovations.

Missing obs.	Method	Coverage	Coverage (left/right)
10	CP	93	4.2/2.8
	CO	93.8	3.4/2.8
	SIM	90.4	3.2/6.4
	SIMP	91.6	3.6/4.8
	AO	94	4/2
50	CP	90.8	6.3/2.9
	CO	90.4	5.8/3.8
	SIM	90.6	4/5.4
	SIMP	90.4	5.4/4.2
	AO	88	7/5
90	CP	91.8	2.8/5.4
	CO	91.5	4/4.5
	SIM	91.8	2.4/5.8
	SIMP	90.6	2.8/6.6
	AO	92	3/5
RP	CP	87.5	7.2/5.8
	CO	87.8	7.4/5.6
	SIM	91.2	4.6/4.2
	SIMP	90	5.8/4.2
	AO	86	7/7
45	CP	86.8	4/9.2
	CO	87	9/4
	SIM	87.2	9/3.8
	SIMP	87.4	8.6/4
	AO	88	9/3
46	CP	90.2	5.6/4.2
	CO	90.4	6.2/3.4
	SIM	90.4	5.5/4.1
	SIMP	90.2	6.4/3.4
	AO	89	7/4
47	CP	85	5/10
	CO	84.4	5.6/9.4
	SIM	87.6	7.8/4.6
	SIMP	86.2	5.8/8
	AO	82	7/11
48	CP	84	6.8/9.2
	CO	85	6/9
	SIM	89.2	8/2.8
	SIMP	90.4	5.8/3.8
	AO	80	8/12
49	CP	93.4	3/3.6
	CO	94.2	2.4/3.4
	SIM	87.8	4.4/7.8
	SIMP	92.6	2.2/5.2
	AO	93	1/6

Table 2
Simulation results for Model AR using exponential innovations.

Missing obs.	Method	Coverage	Coverage (left/right)
10	CP	92.2	7.2/0.6
	CO	91.0	8.0/1.0
	SIM	92.2	5/2.8
	SIMP	89.1	9.5/1.4
	AO	82	10/8
50	CP	95.2	3.2/1.6
	CO	95.6	2.8/1.6
	SIM	91.6	5.6/2.8
	SIMP	90	5/5
	AO	81	9/10
90	CP	91.2	6/2.8
	CO	91.4	6/2.6
	SIM	88.8	7.4/3.8
	SIMP	90.8	6/3.2
	AO	83	9/8
RP	CP	92	5/3
	CO	92	6/2
	SIM	91	6/3
	SIMP	90.6	4.6/4.8
	AO	81	7.2/11.8
45	CP	91	5.5/3.5
	CO	91	5.6/3.4
	SIM	90	6.7/3.3
	SIMP	89.5	6.3/4.2
	AO	77	9/14
46	CP	90.6	5.4/4
	CO	91	6/3
	SIM	91.2	5/3.8
	SIMP	90.6	5.4/4
	AO	79	9/12
47	CP	87	6/7
	CO	87	7.2/5.8
	SIM	88	7/5
	SIMP	88.8	7/4.2
	AO	68	8/24
48	CP	93	3/4
	CO	93	4/3
	SIM	91.8	3/5.2
	SIMP	89.2	4/6.8
	AO	73	7/20
49	CP	93.4	5.6/1.0
	CO	93.6	4.6/1.8
	SIM	92.8	5.6/1.6
	SIMP	92.4	5/2.6
	AO	81	8/11

Table 3
Simulation results for Model AR using contaminated innovations.

Missing obs.	Method	Coverage	Coverage (left/right)
10	CP	91	2/7
	CO	91.6	2.2/6.2
	SIM	91.4	4/4.6
	SIMP	90.6	3/6.4
	AO	77.6	7/15.4
50	CP	91.8	4.6/3.6
	CO	91	6.2/2.8
	SIM	91.2	4.6/4.2
	SIMP	90.2	5.2/4.6
	AO	79	8/13
90	CP	93	2/5
	CO	92	3/5
	SIM	90.8	1.6/7.6
	SIMP	90.2	4.6/5.2
	AO	75.8	6.2/18
RP	CP	91	4.6/4.4
	CO	92	4.6/3.4
	SIM	91.4	4.6/4
	SIMP	91.2	4/4.8
	AO	73	8/19
45	CP	90	6/4
	CO	90.2	5/4.3
	SIM	90.2	5.4/4.4
	SIMP	90.2	5/4.3
	AO	74.6	12.8/12.6
46	CP	91	6/3
	CO	91	5.4/3.6
	SIM	89.8	6/4.2
	SIMP	89.4	5.6/5
	AO	65.2	16.6/18.2
47	CP	91	4.7/4.3
	CO	90.4	5.6/4.0
	SIM	91.2	5.6/3.2
	SIMP	90.8	6/3.2
	AO	64.6	15.8/19.6
48	CP	93	5/2
	CO	91.4	4.8/3.8
	SIM	91.8	4.4/3.8
	SIMP	90.6	5.6/3.8
	AO	61	16/23
49	CP	96.8	2.2/1
	CO	95.4	3.2/1.4
	SIM	88	10.0/2
	SIMP	92.4	4.2/3.4
	AO	62.2	10.2/27.6

Table 4
Simulation results for Model MA using Gaussian innovations.

Missing obs.	Method	Coverage	Coverage (left/right)
10	CP	86.6	9.0/4.4
	CO	90.2	6.4/3.4
	SIM	90.6	5.4/4
	SIMP	90.2	6/3.8
	AO	84.3	10.0/5.7
50	CP	87	7.2/5.8
	CO	89	8.8/2.2
	SIM	90.4	7.2/2.4
	SIMP	89	8.5/2.5
	AO	85.2	12.8/2.0
90	CP	89.2	5.2/5.6
	CO	91	6/3
	SIM	92.6	4.2/3.2
	SIMP	90	5/5
	AO	76	14.0/10.0
RP	CP	89	3.8/7.2
	CO	88.8	3.8/7.4
	SIM	91.2	5/3.8
	SIMP	90.4	5.5/4.1
	AO	78	13/9
45	CP	83.2	9.4/7.4
	CO	85.6	8.6/5.8
	SIM	86	9/5
	SIMP	86.2	8.4/5.4
	AO	86	10/4
46	CP	85.0	4.8/10.2
	CO	88.6	3.4/8
	SIM	90.2	6.4/3.4
	SIMP	89	2/9
	AO	88.8	3.2/8
47	CP	84	6.2/9.8
	CO	88.2	3.2/8.6
	SIM	85.2	6/8.8
	SIMP	88.2	3.6/8.2
	AO	86	5/9
48	CP	80.6	11.6/7.8
	CO	85.2	8.4/6.4
	SIM	82.8	10.6/6.6
	SIMP	86.6	8/5.4
	AO	86	10/4
49	CP	82.0	12.2/5.8
	CO	84.8	10.6/4.6
	SIM	81.4	12.2/6.4
	SIMP	85.6	10/4.4
	AO	84	12/4

Table 5
Simulation results for Model MA using exponential innovations.

Missing obs.	Method	Coverage	Coverage (left/right)
10	CP	88.2	5.4/6.6
	CO	88.2	6.4/5.4
	SIM	92.8	3/4.2
	SIMP	90.4	4.6/5
	AO	83	12/5
50	CP	85.8	6.2/8.0
	CO	87.6	6.6/5.8
	SIM	90	2.8/7.2
	SIMP	89.4	3.6/7
	AO	82	9/9
90	CP	90.2	2.8/7.0
	CO	91.2	2.6/6.2
	SIM	91.2	5/3.8
	SIMP	90.4	4.6/5
	AO	78	13/9
RP	CP	89	3.8/7.2
	CO	88.8	3.8/7.4
	SIM	89.6	6.8/3.6
	SIMP	90.4	5.4/4.2
	AO	77.6	10.6/11.8
45	CP	83	10.8/6.2
	CO	85	9.2/5.8
	SIM	83.2	10.8/6
	SIMP	86.4	9.4/4.2
	AO	89	4/7
46	CP	84.2	7.0/8.8
	CO	83.8	6.6/9.6
	SIM	84	7.2/8.8
	SIMP	85	6/9
	AO	85	6/9
47	CP	82.8	6.4/10.8
	CO	87	4.4/8.6
	SIM	83	7/10
	SIMP	88	3.4/8.6
	AO	85	4/11
48	CP	78.4	11/10.6
	CO	82.6	9.6/7.8
	SIM	77.4	12/10.6
	SIMP	88	7/5
	AO	88	7.0/5.0
49	CP	87.6	7.0/5.4
	CO	90	6.4/3.6
	SIM	88.2	8.4/3.4
	SIMP	89.2	6.2/4.6
	AO	87	7/6

Table 6
Simulation results for Model MA using contaminated innovations.

Missing obs.	Method	Coverage	Coverage (left/right)
10	CP	88	5.3/6.7
	CO	91.8	3.2/5
	SIM	87	6.3/6.7
	SIMP	91.1	3.5/5.4
	AO	69.4	13.2/17.4
50	CP	89.2	6.0/4.8
	CO	89.0	6.2/4.8
	SIM	89.8	6.0/4.2
	SIMP	90.2	7.2/2.6
	AO	72.8	14.4/12.8
90	CP	87.2	7.6/5.2
	CO	87.0	6.8/6.2
	SIM	87.8	6/6.2
	SIMP	90.2	5/4.8
	AO	72.8	11.8/15.4
RP	CP	88.6	5.4/6
	CO	89.8	5.2/5
	SIM	88.8	5/6.2
	SIMP	90.1	7.3/2.6
	AO	75.8	9.4/14.8
45	CP	87.8	4.2/8.0
	CO	88.8	3.8/7.4
	SIM	87.2	3.4/9.4
	SIMP	89	3.4/7.6
	AO	83.2	4.2/12.6
46	CP	84.4	8.0/7.6
	CO	86.0	7.6/6.4
	SIM	84.2	7.8/8
	SIMP	87	6.6/6.4
	AO	82.2	7.6/10.2
47	CP	85.0	7.8/7.2
	CO	86.2	7.4/6.4
	SIM	85.8	7/7.2
	SIMP	86	6.8/7.2
	AO	84.8	7.2/8.0
48	CP	82.6	7.0/10.4
	CO	86.4	5.4/8.2
	SIM	83.4	9/7.6
	SIMP	87	4.4/8.6
	AO	84.6	5.2/10.2
49	CP	86.0	9.8/4.2
	CO	88.4	8.2/3.4
	SIM	87.2	8/4.8
	SIMP	89.4	7.2/3.4
	AO	84.4	7.8/7.8