Journal of Data Science **5**(2007), 85-101

Nonparametric Modelling of Quarterly Unemployment Rates

Lijian Yang
Michigan State University

Abstract: A seasonal additive nonlinear vector autoregression (SANVAR) model is proposed for multivariate seasonal time series to explore the possible interaction among the various univariate series. Significant lagged variables are selected and additive autoregression functions estimated based on the selected variables using spline smoothing method. Conservative confidence bands are constructed for the additive autoregression function. The model is fitted to two sets of bivariate quarterly unemployment rate data with comparisons made to the linear periodic vector autoregression model. It is found that when the data does not significantly deviate from linearity, the periodic model is preferred. In cases of strong nonlinearity, however, the additive model is more parsimonious and has much higher out-of-sample prediction power. In addition, interactions among various univariate series are automatically detected.

Key words: Autoregression, BIC, confidence bands, prediction error, seasonality, splines, unemployment rate.

1. Introduction

It has been well known that nonlinearity exists widely in macroeconomic time series, see, for example, Huang and Yang (2004) for empirical evidence of nonlinearity in the US unemployment rates. Generally speaking, when deviation from linear time series model is significant, nonparametric autoregression is more appropriate for the identification and forecasting of time series unless there is convincing evidence of a simpler parametric nonlinear structure that generates the data series. Hence, nonparametric smoothing of nonlinear autoregressive time series can be extremely useful for time series analysis, not only for exploratory study, but also for robust model selection and prediction.

Non- and semi- parametric smoothing estimation of unknown functions have found many applications in the last two decades. In the time series literature, Robinson (1983) first applied kernel (Nadaraya-Watson) method to estimate autoregression function of unknown form. Other significant contributions include: Györfi, Härdle, Sarda and Vieu (1989), Auestad and Tjøstheim (1990), Tjøstheim

and Auestad (1994), Yao and Tong (1994), Fan and Yao (1998), Härdle, Tsybakov and Yang (1998), Yang and Tschernig (2002), to name a few from the great number of published articles in this area. One common feature of these cited works is that they are all based on the local least squares method of kernel/local polynomial regression, and thus are all computationally intensive. Also unaddressed is the issue of the "curse of dimensionality", which refers to the lack of accuracy in estimating multivariate functions of general nonparametric form. In the context of time series modelling, the high dimensionality can be the result of the time series being multivariate and/or potentially too many lagged variables being significant for forecasting. For the two unemployment rates studied in this paper, 24 potentially significant lagged variables are examined.

The issue of the "curse of dimensionality" can be dealt with via additive modelling, as first proposed in Hastie and Tibshirani (1990). More recent works on the subject of additive models include Chen and Tsay (1993), Tjøstheim and Auestad (1994), Linton and Nielsen (1995), Sperlich, Tjøstheim and Yang (2002), Huang and Yang (2004). In particular, Huang and Yang (2004) had taken a different approach to the estimation of nonparametric additive regression function, by polynomial spline smoothing instead of kernel smoothing. The advantage of polynomial spline is that only one least squares problem needs to be solved to obtain estimates of all function values, rather than solving a least squares problem to estimate each function value. Typically, this means that the spline estimation of additive model can be thousands of times faster than standard kernel based methods, such as in Linton and Nielsen (1995) or Sperlich, Tjøstheim and Yang (2002). Spline smoothing has been used in other dimension reduction models, for example, varying coefficient model as in Huang, Wu and Zhou (2002).

In this paper we extend the additive autoregression model of Huang and Yang (2004) to multivariate seasonal time series. One example of such series is the bivariate series consisting of quarterly unemployment rates of men and women in the US. What makes such multivariate series different from univariate series is that there may exist significant interaction among the various univariate series. This turns out to be the case for the men's and women's unemployment rates, while it turns out differently for another bivariate series. In both cases, however, the same data-driven inference procedures are applied without prior assumptions of interaction or lack thereof. Hence the issue of interaction is decided by "letting the data speak for themselves".

In section 2 we will formulate a seasonal additive nonlinear vector autoregression model (SANVAR), and discuss its use in identifying the functional structure in seasonal vector time series, via spline smoothing. In section 3, we briefly describe an asymptotically conservative confidence band for the nonparametric autoregression function, also based on the polynomial spline method. Section 4

discusses the findings on two bivariate quarterly unemployment rates data, and draws some general conclusions about the benefits and precautions when fitting the SANVAR model.

2. The SANVAR Model

Our aim in this section is to develop a general modelling framework, called the seasonal additive nonlinear vector autoregression model (SANVAR), for a multivariate seasonal process $\{Y_{t,\gamma}\}_{t=1,\gamma=1}^{n,d}$. There are d different series, the γ-th of which is $\{Y_{t,\gamma}\}_{t=1}^{n}$. Each of the series is seasonal with S seasons. The approach we take to incorporate seasonality is distinct from those in Lütkepohl (1993), Wolters (1992). We model the d series and S seasons in the form

$$\begin{aligned} Y_{\tau S+s,\gamma} &= m_{s,\gamma}\left(Y_{\tau S+s-j,\beta}\right)_{(j,\beta)\in\Lambda_\gamma} + \xi_{\tau S+s,\gamma}, \\ m_{s,\gamma}\left(y_{j,\beta}\right)_{(j,\beta)\in\Lambda_\gamma} &= m_{0,s,\gamma} + \sum_{(j,\beta)\in\Lambda_\gamma} m_{s,\gamma,j,\beta}\left(y_{j,\beta}\right), 1 \le \gamma \le d, \tau \ge M/S \end{aligned} \quad (2.1)$$

where $s \in \{1, ..., S\}$ indicates the season, Λ_γ is a subset of $\{1, ..., M\} \times \{1, ..., d\}$ for significant lagged variables, $\{\xi_{\tau S+s,\gamma}\}_{1\le\gamma\le d,\tau\ge M/S}$ are martingale differences with respect to the σ-field generated by variables $\{Y_{\tau S+s-j,\beta}\}_{j>0,1\le\beta\le d}$, and the multivariate autoregression function $m_{s,\gamma}$ is an additive function of variables $Y_{\tau S+s-j,\beta}, (j,\beta) \in \Lambda_\gamma$. The largest lag index allowed M is typically taken to be a multiple of S, and the size of set Λ_γ is limited to be no more than a fixed integer $\lambda_{\max}$. Each component function $m_{s,\gamma,j,\beta}$ satisfies the identifiability condition $Em_{s,\gamma,j,\beta}\left(Y_{t-j,\beta}\right) \equiv 0$, as is common in additive modelling. If all the component functions $m_{s,\gamma,j,\beta}$ are restricted to be linear, the model is a periodic vector autoregressive (PVAR) model as in Lütkepohl (1993).

When one fits a SANVAR model (2.1) to a time series γ, $\{Y_{t,\gamma}\}_{t=1}^{n}$, the lag set Λ_γ is unknown a priori and has to be selected. Thus, every index pair $(j,\beta) \in \{1, ..., M\} \times \{1, ..., d\}$ could potentially be in Λ_γ. For many real time series data, however, most of the functions $m_{s,\gamma,j,\beta}$ turn out to be insignificant, as one will see in section 4.

For the fitting of SANVAR, we use the adaptive spline approach, which is described here in detail. For all seasons s and every index pair $(j,\beta) \in \{1, ..., M\} \times \{1, ..., d\}$, one denote by the interval $[a_{j,\beta}, b_{j,\beta}]$ the range of variable $\{Y_{\tau S+s-j,\beta}\}_{\tau\ge M/S}$, which is divided into $N+1$ equally-spaced subintervals. Here $N = N_n = \left[k\left(n/S\right)^{1/(2p+3)}\right]$ in which k is a tuning constant (default set to 1), and p is an integer no more than the degree of smoothness of the component functions (default is set to 1). The N interior endpoints of these subintervals are labelled as $a_{j,\beta}^{(1)}, ..., a_{j,\beta}^{(N)}$, which form the knot sequence for the explanatory

variable $Y_{t-j,\beta}$. Next we define the spline basis as the set of the following functions

$$\begin{aligned} B_{j,\beta}^{(1)}(y) &= y, ..., B_{j,\beta}^{(p)}(y) = y^p \\ B_{j,\beta}^{(p+1)}(y) &= \left(y - a_{j,\beta}^{(1)}\right)_+^p, ..., B_{j,\beta}^{(p+N)}(y) = \left(y - a_{j,\beta}^{(N)}\right)_+^p \end{aligned}$$

where $x_+ = x$ if $x > 0$, 0 otherwise. Linear combinations of these spline basis are piecewise smooth up to order p called spline functions. Since $N_n \to \infty$ as $n \to \infty$, all functions of smoothness order p can be approximated on interval $[a_{j,\beta}, b_{j,\beta}]$ by such linear combinations and so for every index pairs (s,γ) and (j,β) the function $m_{s,\gamma,j,\beta}(y)$ is approximated by spline functions.

To estimate the component functions, we have to solve an ordinary least squares problem of the form

$$\sum_{M<\tau S+s\le n} \left\{ Y_{\tau S+s,\gamma} - c_{s,\gamma} - \sum_{j=1}^{M}\sum_{\beta=1}^{d}\sum_{l=1}^{p+N_n} c_{s,\gamma,j,\beta}^{(l)} B_{j,\beta}^{(l)}\left(Y_{\tau S+s-j,\beta}\right) \right\}^2 \tag{2.2}$$

and the solution $\left\{\hat{c}_{s,\gamma}, \hat{c}_{s,\gamma,j,\beta}^{(l)}\right\}$ will then provide estimators

$$\begin{aligned} \hat{m}_{s,\gamma,j,\beta}(y) &= \sum_{l=1}^{p+N} \hat{c}_{s,\gamma,j,\beta}^{(l)} B_{j,\beta}^{(l)}(y) - \sum_{M<\tau S+s\le n}\sum_{l=1}^{p+N} \hat{c}_{s,\gamma,j,\beta}^{(l)} B_{j,\beta}^{(l)}\left(Y_{\tau S+s-j,\beta}\right) A^{-1} \\ \hat{m}_{0,s,\gamma} &= \hat{c}_{s,\gamma} + \sum_{M<\tau S+s\le n}\sum_{j=1}^{M}\sum_{\beta=1}^{d}\sum_{l=1}^{p+N} \hat{c}_{s,\gamma,j,\beta}^{(l)} B_{j,\beta}^{(l)}\left(Y_{\tau S+s-j,\beta}\right) A^{-1}, \end{aligned} \tag{2.3}$$

where $A = \sum_{M<\tau S+s\le n} 1$. Fortunately, one typically will need only a small number of these estimators for the SANVAR modelling. To identify which of these are significant, a BIC criterion is defined for each subset $\Lambda = \{(j_1,\beta_1), ..., (j_\lambda,\beta_\lambda)\} \subset \{1, ..., M\} \times \{1, ..., d\}$ where $j_1 \le \cdots \le j_\lambda$. Let $\left\{\hat{c}_{s,\gamma-\tau}, \hat{c}_{s,\gamma,j,\beta-\tau}^{(l)}\right\}$ be the solution of the least squares problem (2.2), but the sum is over $(j,\beta) \in \Lambda$ and with the squared error term at time τ removed, for every integer τ that satisfies $j_\lambda < \tau S + s \le n$. Then for every $1 \le \gamma \le d$, the BIC criterion for the γ-th series is defined as

$$\begin{aligned} \mathrm{BIC}_\gamma(\Lambda) = \{1 + \lambda(p+N_n)\}\frac{\ln n_S}{n_S} + \ln\Bigg[\frac{1}{S}\sum_{s=1}^{S}\frac{1}{n_{s,j_\lambda,S}} \sum_{j_\lambda<\tau S+s\le n} \{Y_{\tau S+s,\gamma} \\ - \hat{c}_{s,\gamma-\tau} - \sum_{(j,\beta)\in\Lambda}\sum_{l=1}^{p+N_n} \hat{c}_{s,\gamma,j,\beta-\tau}^{(l)} B_{j,\beta}^{(l)}\left(Y_{\tau S+s-j,\beta}\right)\}^2 \Bigg] \end{aligned} \tag{2.4}$$

where $n_S = n/S, n_{s,j_\lambda,S}$ = the number of integers τ that satisfy $j_\lambda < \tau S + s \leq n$.

The set of significant variables selected by the BIC is then defined as

$$\hat{\Lambda}_\gamma = \underset{\Lambda \subset \{1,\ldots,M\}\times\{1,\ldots,d\}}{\operatorname{argmin}} \mathrm{BIC}_\gamma(\Lambda), \qquad 1 \leq \gamma \leq d$$

and under reasonable assumptions (Huang and Yang 2004), it can be shown that

$$\hat{\Lambda}_\gamma \to \Lambda_\gamma \text{ in probability, as } n \to \infty, .1 \leq \gamma \leq d \tag{2.5}$$

Notice that in the above steps, if all basis $B_{j,\beta}^{(l)}, l \geq 2$ are removed, the result would be the PVAR model. Once this set $\hat{\Lambda}_\gamma$ is obtained, the ordinary least squares problem (2.2) is solved only once based on this set and the resulting function estimates $\hat{m}_{s,\gamma,j,\beta}$ are used to build the estimated SANVAR model. This intelligent identification of a parsimonious model can be used for improving forecasting. In the next section, we discuss forecasting based on SANVAR model.

3. Confidence Bands

Suppose that by using the BIC criterion (2.4), a set $\hat{\Lambda}_\gamma$ of lags has been determined for series γ of the multivariate time series, $1 \leq \gamma \leq d$. The consistency property in (2.5) allows one to take the estimated set $\hat{\Lambda}_\gamma$ for the true set Λ_γ, for the sake of simpler notation. The estimated SANVAR model is of the form

$$Y_{\tau S+s,\gamma} = \hat{Y}_{\tau S+s,\gamma} + \hat{\xi}_{\tau S+s,\gamma}, \hat{Y}_{\tau S+s,\gamma} = \hat{m}_{s,\gamma}\left(Y_{\tau S+s-j,\beta}\right)_{(j,\beta)\in\Lambda_\gamma}, \tag{3.1}$$

$1 \leq \gamma \leq d, \tau \geq M/S$, with predicted values $\hat{Y}_{\tau S+s,\gamma}$, residuals $\hat{\xi}_{\tau S+s,\gamma}$, and where the multivariate additive function

$$\hat{m}_{s,\gamma}\left(y_{j,\beta}\right)_{(j,\beta)\in\Lambda_\gamma} = \hat{m}_{0,s,\gamma} + \sum_{(j,\beta)\in\Lambda_\gamma} \hat{m}_{s,\gamma,j,\beta}\left(y_{j,\beta}\right) \tag{3.2}$$

with univariate functions $\hat{m}_{s,\gamma,j,\beta}$ and constants $\hat{m}_{0,s,\gamma}$ as defined in (2.3). In this section, a procedure is described for the construction of simultaneous confidence intervals, or, confidence bands, for functions $m_{s,\gamma}$ based on the estimated SANVAR model (3.1).

Recently, confidence bands for univariate regression functions have been developed by Xia (1998), Claeskens and Van Keilegom (2003). The basic idea of constructing asymptotic confidence bands from polynomial spline estimation is proposed in Wang and Yang (2005), which is limited to univariate regression (this means $d = M = 1$) and piecewise constant (i.e., $p = 0$) and piecewise linear splines (i.e., $p = 1$). Yang (2004) extended the procedure of Wang and Yang

(2005) to additive model, using piecewise linear spline ($p = 1$) and wild bootstrap. The method is adopted to SANVAR model, and the steps are described here. The confidence level is taken to be $1 - \alpha$, where α has a default value of 0.05.

Step 1 Let $\{\delta_{\tau,b}\}_{M<\tau S+s\le n}, b = 1, ..., 400$ be i.i.d. samples of the following discrete distribution

$$\delta_{\tau,b} = \begin{cases} 2^{-1}\left(1-\sqrt{5}\right) & \text{with probability } 10^{-1}\left(5+\sqrt{5}\right) \\ 2^{-1}\left(1+\sqrt{5}\right) & \text{with probability } 10^{-1}\left(5-\sqrt{5}\right) \end{cases}, M < \tau S+s \le n.$$

Step 2 For any $1 \le b \le 400$, define the b-th wild bootstrap sample as

$$Y_{\tau S+s,\gamma,b} = \hat{m}_{s,\gamma}\left(Y_{\tau S+s-j,\beta}\right)_{(j,\beta)\in\Lambda_\gamma} + \delta_{\tau,b}\hat{\xi}_{\tau S+s,\gamma}, 1 \le \gamma \le d, \tau \ge M/S \tag{3.3}$$

with residuals $\hat{\xi}_{\tau S+s,\gamma}$, and where the multivariate additive functions $\hat{m}_{s,\gamma}$ are as defined in (3.2). Next, solve the least squares problem for the wild bootstrap sample

$$\sum_{M<\tau S+s\le n}\left\{Y_{\tau S+s,\gamma,b} - c_{s,\gamma} - \sum_{j=1}^{M}\sum_{\beta=1}^{d}\sum_{l=1}^{p+N_n} c^{(l)}_{s,\gamma,j,\beta}B^{(l)}_{j,\beta}\left(Y_{\tau S+s-j,\beta}\right)\right\}^2$$

and use the solution $\left\{\hat{c}_{s,\gamma,b}, \hat{c}^{(l)}_{s,\gamma,j,\beta,b}\right\}$ to obtain the b-th bootstrap estimator

$$\hat{m}_{s,\gamma,b}\left(y_{j,\beta}\right)_{(j,\beta)\in\Lambda_\gamma} = \hat{c}_{s,\gamma,b} + \sum_{j=1}^{M}\sum_{\beta=1}^{d}\sum_{l=1}^{p+N_n}\hat{c}^{(l)}_{s,\gamma,j,\beta,b}B^{(l)}_{j,\beta}\left(y_{j,\beta}\right).$$

Step 3 Define an inflation factor $r_{d_\gamma,N_n,\alpha} = z^{-1}_{1-\alpha/2}\sqrt{A^2}$, where A^2 is the $100(1 - \alpha/\left(N_n+1\right)^{d_\gamma})$-th quantile of the chi-square distribution with degrees of freedom $2d_\gamma$, where one denotes the number of variables in Λ_γ as d_γ. Denote by $\hat{m}^{L,\alpha/2}_{s,\gamma}\left(y_{j,\beta}\right)_{(j,\beta)\in\Lambda_\gamma}$ and $\hat{m}^{U,\alpha/2}_{s,\gamma}\left(y_{j,\beta}\right)_{(j,\beta)\in\Lambda_\gamma}$ respectively the lower and upper $100\left(1-\alpha/2\right)\%$ quantiles of the set$\{\hat{m}_{s,\gamma,b}\left(y_{j,\beta}\right)_{(j,\beta)\in\Lambda_\gamma}\}_{1\le b\le 400}$, and

$$\begin{aligned}\widehat{m}^{L,\alpha/2}_{s,\gamma}\left(y_{j,\beta}\right)_{(j,\beta)\in\Lambda_\gamma} &= \hat{m}_{s,\gamma}\left(y_{j,\beta}\right)_{(j,\beta)\in\Lambda_\gamma} + \\ &\quad \{\hat{m}^{L,\alpha/2}_{s,\gamma}\left(y_{j,\beta}\right)_{(j,\beta)\in\Lambda_\gamma} - \hat{m}_{s,\gamma}\left(y_{j,\beta}\right)_{(j,\beta)\in\Lambda_\gamma}\} \\ &\quad \times r_{d_\gamma,N_n,\alpha} \qquad (3.4)\\ \widehat{m}^{U,\alpha/2}_{s,\gamma}\left(y_{j,\beta}\right)_{(j,\beta)\in\Lambda_\gamma} &= \hat{m}_{s,\gamma}\left(y_{j,\beta}\right)_{(j,\beta)\in\Lambda_\gamma} + \\ &\quad \{\hat{m}^{U,\alpha/2}_{s,\gamma}\left(y_{j,\beta}\right)_{(j,\beta)\in\Lambda_\gamma} - \hat{m}_{s,\gamma}\left(y_{j,\beta}\right)_{(j,\beta)\in\Lambda_\gamma}\} \\ &\quad \times r_{d_\gamma,N_n,\alpha}. \qquad (3.5)\end{aligned}$$

Then $\left[\widehat{m}_{s,\gamma}^{L,\alpha/2}\left(y_{j,\beta}\right)_{\psi}, \widehat{m}_{s,\gamma}^{U,\alpha/2}\left(y_{j,\beta}\right)_{\psi}\right]$ is the $100\left(1-\alpha\right)\%$ confidence band of the function $m_{s,\gamma}\left(y_{j,\beta}\right)_{\psi}$ of the d_{γ} variables $\left(Y_{\tau S+s-j,\underline{e}ta}\right)_{\psi}$, where the subscript ψ denotes $(j,\beta) \in \Lambda_{\gamma}$.

Using the wild bootstrap sample (3.3) is justified by the same reason as in Sperlich, Tjøstheim and Yang (2002), i.e., in terms of conditional moments up to order two, for any $1 \leq b \leq 400$, the b-th bootstrap sample

$$\left\{\left(Y_{\tau S+s-j,\beta}\right)_{(j,\beta)\in\Lambda_{\gamma}}, \delta_{\tau,b}\hat{\xi}_{\tau S+s,\gamma}\right\}_{1\leq\gamma\leq d,\tau\geq M/S}$$

always mimicks the original sample $\left\{\left(Y_{\tau S+s-j,\beta}\right)_{(j,\beta)\in\Lambda_{\gamma}}, \xi_{\tau S+s,\gamma}\right\}_{1\leq\gamma\leq d,\tau\geq M/S}$, due to the fact that $E\left(\delta_{\tau,b}\right) \equiv 0, \mathrm{var}\left(\delta_{\tau,b}\right) \equiv 1$. The performance of the above wild bootstrap procedure has also been examined via Monte-Carlo study in Yang (2004). In particular, simulation experiments have shown that the procedure is extremely robust in regard to the number of bootstrap samples as long as it is higher than 400.

While the above construction is easily implemented in any software, in this paper we have done all computing in the environment of XploRe [see Härdle, Hlavka and Klinke (2000)]. The above wild bootstrap confidence band, according to the Monte-Carlo results of Yang (2004), is asymptotically conservative. In other words, for any $1 \leq \gamma \leq d$

$$\mathrm{liminf}_{n\to\infty} P[\ A_n\ ] \geq 1-\alpha \tag{3.6}$$

where A_n denotes the event

$$\left[\widehat{m}_{s,\gamma}^{L,\alpha/2}\left(y_{j,\beta}\right)_{(j,\beta)\in\Lambda_{\gamma}} \leq m_{s,\gamma}\left(y_{j,\beta}\right)_{(j,\beta)\in\Lambda_{\gamma}} \leq \widehat{m}_{s,\gamma}^{U,\alpha/2}\left(y_{j,\beta}\right)_{(j,\beta)\in\Lambda_{\gamma}} \text{ for all } y_{j,\beta}\right].$$

In addition, Yang (2004) had also provided some Monte Carlo evidence that the confidence band narrows at the rate of $n^{-2/5}\log^{1/2}(n)$ as $n \to \infty$.

In the next section, we will apply the BIC criterion and the wild bootstrap confidence band to some unemployment series and discover some nontrivial dependence structures in these series.

4. Unemployment Rates

In this section we will closely examine four sets of quarterly unemployment rate data collected from the Current Population Survey (SIC) at the US Bureau of Labor Statistics. The first two series are the quarterly unemployment rates of all men 20 years & over, and all women 20 years & over, regardless of ethnic

origins, family status, occupation, profession and race, from 1948 to 2002. The other two series consist of the quarterly unemployment rates of all whites 16 years & over, and all African Americans 16 years & over, regardless of ethnic origins, family status, occupation, profession and sexes, from 1972 to 2003.

The approach is to model respectively the first two jointly and the last two jointly as bivariate time series, of $S = 4$ seasons. For the first data, since there are a total of 220 quarters, the combined time series is $\{R_{t,\gamma}\}_{t=1,\gamma=1}^{220,2}$ where

$$\begin{aligned} R_{t,1} &= \text{unemployment rate of men 20 years and over in quarter } t \\ R_{t,2} &= \text{unemployment rate of women 20 years and over in quarter } t, \end{aligned}$$

while for the second, there are a total of 124 quarters, and the combined series is $\{R_{t,\gamma}\}_{t=1,\gamma=1}^{124,2}$ where

$$\begin{aligned} R_{t,1} &= \text{unemployment rate of whites 16 years and over in quarter } t \\ R_{t,2} &= \text{unemployment rate of African Americans 16 years and over in quarter } t. \end{aligned}$$

It is more convenient to write the time series in the rescaled time τ as $R_{\tau S+s,\gamma}, s \in \{1, ..., S\}, \tau = 0, 1, ...$ Preliminary examination suggests fourth differencing, thus we actually analyze $Y_{\tau S+s,\gamma} = \nabla_S R_{\tau S+s,\gamma}$.

For the two bivariate data sets, we use the beginning 90% of the data to estimate the model and then calculate the out-of-sample prediction error for the last 10% of the data. Both SANVAR and PVAR models are used for comparison. The definition of $Y_{\tau S+s-S,\gamma}$ as $Y_{\tau S+s,\gamma} = R_{\tau S+s,\gamma} - R_{\tau S+s-S,\gamma}$ leads one to define the forecasts of $R_{\tau S+s,\gamma}$ in terms of the forecasts of $Y_{\tau S+s,\gamma}$, i.e., $\hat{R}_{\tau S+s,\gamma} = \hat{Y}_{\tau S+s,\gamma} + R_{\tau S+s-S,\gamma}$.

For the men/women data, the fitted PVAR models give the following forecasting equations

$$\begin{aligned} \hat{Y}_{4\tau+1,1} &= -0.030 - 1.039Y_{4\tau-1,1} + 2.006Y_{4\tau,1} \\ \hat{Y}_{4\tau+2,1} &= -0.002 - 0.749Y_{4\tau,1} + 1.309Y_{4\tau+1,1} \\ \hat{Y}_{4\tau+3,1} &= -0.011 - 0.406Y_{4\tau+1,1} + 1.191Y_{4\tau+2,1} \\ \hat{Y}_{4\tau+4,1} &= 0.005 - 0.702Y_{4\tau+2,1} + 1.475Y_{4\tau+3,1} \qquad (4.1) \\ \hat{Y}_{4\tau+1,2} &= -0.019 + 0.421Y_{4\tau-4,1} - 0.471Y_{4\tau-6,2} - 0.981Y_{4\tau-2,2} + 1.549Y_{4\tau,2} \\ \hat{Y}_{4\tau+2,2} &= 0.013 + 0.006Y_{4\tau-3,1} + 0.067Y_{4\tau-5,2} - 0.357Y_{4\tau-1,2} + 1.073Y_{4\tau+1,2} \\ \hat{Y}_{4\tau+3,2} &= 0.012 + 0.094Y_{4\tau-2,1} + 0.009Y_{4\tau-4,2} - 0.207Y_{4\tau,2} + 0.878Y_{4\tau+2,2} \\ \hat{Y}_{4\tau+4,2} &= -0.019 + 0.134Y_{4\tau-1,1} - 0.069Y_{4\tau-3,2} - 0.432Y_{4\tau+1,2} \\ &\quad + 1.029Y_{4\tau+3,2} \qquad (4.2) \end{aligned}$$

whereas the SANVAR forecasting equations are

$$\hat{Y}_{4\tau+1,1} = -0.503 - 0.333Y_{4\tau-2,1} + 1.015Y_{4\tau,1}$$

$$
\begin{aligned}
& -0.365(Y_{4\tau-2,1}+0.667)_+ - 0.456(Y_{4\tau-2,1}-1.467)_+ \\
& +1.137(Y_{4\tau,1}+0.800)_+ - 1.210(Y_{4\tau,1}-1.000)_+ \\
\hat{Y}_{4\tau+2,1} &= -0.050 - 0.251Y_{4\tau-1,1} + 0.850Y_{4\tau+1,1} \\
& -0.048(Y_{4\tau-1,1}+0.43)_+ - 0.449(Y_{4\tau-1,1}-1.233)_+ \\
& +0.072(Y_{4\tau+1,1}+1.333)_+ + 0.196(Y_{4\tau+1,1}-1.033)_+ \\
\hat{Y}_{4\tau+3,1} &= -0.155 + 0.019Y_{4\tau,1} + 0.569Y_{4\tau+2,1} \\
& -0.110(Y_{4\tau,1}+0.800)_+ - 0.595(Y_{4\tau,1}-1.000)_+ \\
& +0.392(Y_{4\tau+2,1}+0.667)_+ - 0.048(Y_{4\tau+2,1}-1.467)_+ \\
\hat{Y}_{4\tau+4,1} &= -0.190 - 0.155Y_{4\tau+1,1} + 0.756Y_{4\tau+3,1} \\
& +0.122(Y_{4\tau+1,1}+1.333)_+ - 0.757(Y_{4\tau+1,1}-1.033)_+ \\
& +0.339(Y_{4\tau+3,1}+0.433)_+ - 0.090(Y_{4\tau+3,1}-1.233)_+ \qquad (4.3) \\
\hat{Y}_{4\tau+1,2} &= -0.611 - 0.116Y_{4\tau-2,1} + 0.344Y_{4\tau,1} \\
& -0.408(Y_{4\tau-2,1}+0.667)_+ - 0.533(Y_{4\tau-2,1}-1.467)_+ \\
& +1.237(Y_{4\tau,1}+0.800)_+ - 0.595(Y_{4\tau,1}-1.000)_+ \\
\hat{Y}_{4\tau+2,2} &= -0.495 - 0.304Y_{4\tau-1,1} + 0.394Y_{4\tau+1,1} \\
& +0.096(Y_{4\tau-1,1}+0.433)_+ - 0.408(Y_{4\tau-1,1}-1.233)_+ \\
& +0.263(Y_{4\tau+1,1}+1.333)_+ + 0.511(Y_{4\tau+1,1}-1.033)_+ \\
\hat{Y}_{4\tau+3,2} &= -0.663 - 0.532Y_{4\tau,1} + 0.567Y_{4\tau+2,1} \\
& +0.735(Y_{4\tau,1}+0.800)_+ - 0.840(Y_{4\tau,1}-1.000)_+ \\
& +0.113(Y_{4\tau+2,1}+0.667)_+ + 0.001(Y_{4\tau+2,1}-1.467)_+ \\
\hat{Y}_{4\tau+4,2} &= -0.484 - 0.125Y_{4\tau+1,1} + +0.265Y_{4\tau+3,1} \\
& +0.179(Y_{4\tau+1,1}+1.333)_+ - 0.507(Y_{4\tau+1,1}-1.033)_+ \\
& +0.604(Y_{4\tau+3,1}+0.433)_+ - 0.299(Y_{4\tau+3,1}-1.233)_+ \qquad (4.4)
\end{aligned}
$$

In Figures 1 and 2, the forecasts $\hat{R}_{t,1}, \hat{R}_{t,2}, t = 201, ..., 220$ are plotted according to computation from the SANVAR equations (4.3), (4.4) and the PVAR equations (4.1) and (4.2). The Mean Squared Prediction Error (MSPE) is evaluated for each model as $\frac{1}{20}\sum_{t=201}^{220}\left(\hat{R}_{t,\gamma}-R_{t,\gamma}\right)^2, \gamma = 1, 2$. The SANVAR forecasts come with confidence bands as given in (3.4) and (3.5) of section 3. From these plots, one can see clearly that the SANVAR model is superior to the PVAR model. For the series of men, the SANVAR model is only slightly better in prediction power, while for the series for women, the SANVAR model is twice as powerful as the PVAR model in prediction. For both men's and women's series, the confidence bands appear rather narrow and follow the trends well. Notice that these confidence bands are simultaneous confidence intervals, not simultaneous prediction intervals (which need to account for extra noise), hence the excellent

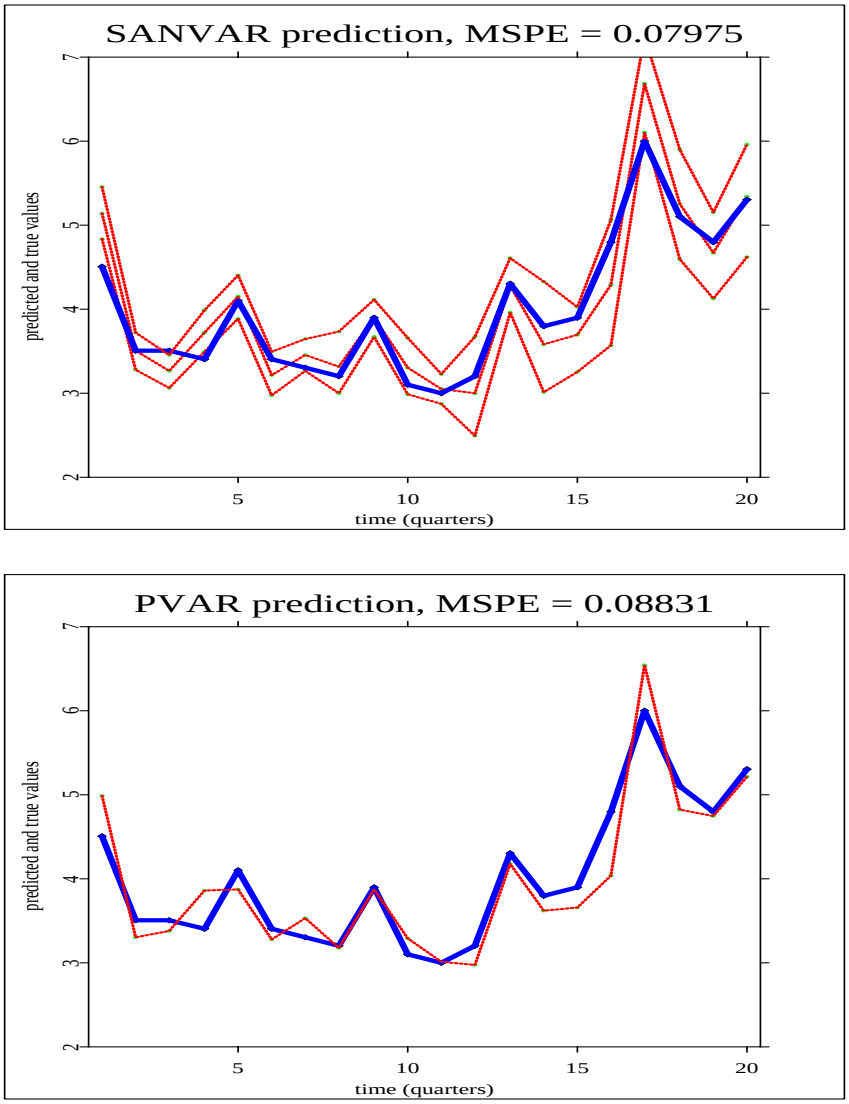

Figure 1: Forecasting the men's quarterly unemployment rates of 1998-2002, based on men's and women's unemployment rates of 1948-1997. The solid thick line represents the actual unemployment rates during 1998-2002, the thin dashed line represents the forecasts. Both the parametric PVAR and the nonparametric SANVAR models are used. In the plot for SANVAR model, nonparametric confidence band for the predicted means are also plotted. The MSPE is calculated as the mean squared prediction error between the predicted and true unemployment rates.

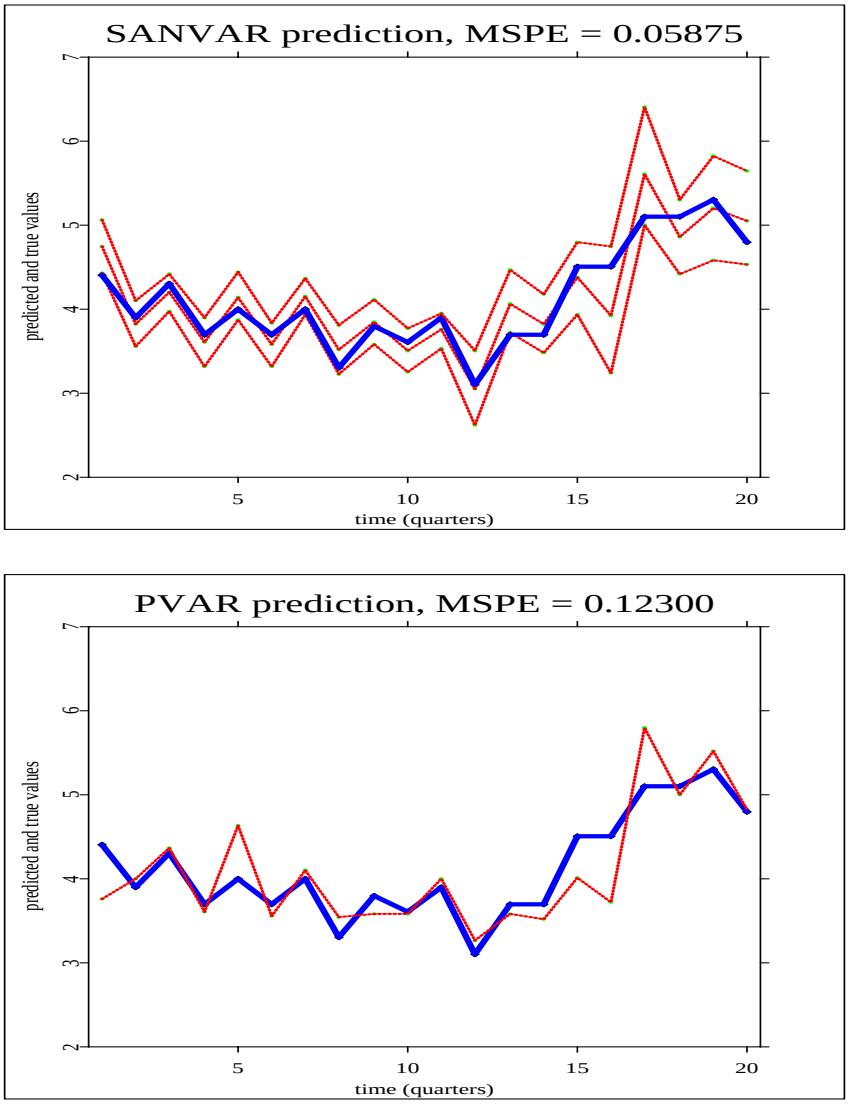

Figure 2: Forecasting the women's quarterly unemployment rates of 1998-2002, based on men's and women's unemployment rates of 1948-1997. The solid thick line represents the actual unemployment rates during 1998-2002, the thin dashed line represents the forecasts. Both the parametric PVAR and the nonparametric SANVAR models are used. In the plot for SANVAR model, nonparametric confidence band for the predicted means are also plotted. The MSPE is calculated as the mean squared prediction error between the predicted and true unemployment rates.

coverage of the actual data path by these bands is all the more remarkable. This is consistent with the conservativeness of the confidence bands as in (3.6). Similar

phenomenon will be observed again for the forecasting of whites and African Americans' unemployment rates, in Figures 3 and 4.

The PVAR model for the men's series is a more parsimonious one than the SANVAR model, as both contain two variables, yet the PVAR equations are three term equations while the SANVAR equations contain seven terms. On the other hand, the SANVAR model for the women's series is more parsimonious than the PVAR model, as each SANVAR equation contains only two variables, versus the four variables of PVAR equations. It is for this reason that the SANVAR is preferred to PVAR for the women's series, but not for the men's series. In addition, the PVAR and SANVAR equations for the men's series are actually quite similar as well. Another interesting phenomenon is that the SANVAR equations for men and women's series are the same in form, both are expressed in terms of the men's series of one and three previous quarters. One explanation is that the women's job condition has been strongly affected by the men's, possibly due to family related factors.

For the white/black data, the PVAR equations are

$$\begin{aligned}
\hat{Y}_{4\tau+1,1} &= -0.025 - 1.098Y_{4\tau-1,1} + 1.956Y_{4\tau,1} \\
\hat{Y}_{4\tau+2,1} &= 0.003 - 0.541Y_{4\tau,1} + 1.232Y_{4\tau+1,1} \\
\hat{Y}_{4\tau+3,1} &= -0.010 - 0.434Y_{4\tau+1,1} + 1.238Y_{4\tau+2,1} \\
\hat{Y}_{4\tau+4,1} &= -0.006 - 0.599Y_{4\tau+2,1} + 1.489Y_{4\tau+3,1} \qquad (4.5) \\
\hat{Y}_{4\tau+1,2} &= -0.027 + 0.527Y_{4\tau,2} - 1.768Y_{4\tau-1,1} + 2.357Y_{4\tau,1} \\
\hat{Y}_{4\tau+2,2} &= -0.041 + 0.655Y_{4\tau+1,2} - 0.563Y_{4\tau,1} + 0.930Y_{4\tau+1,1} \\
\hat{Y}_{4\tau+3,2} &= -0.004 + 0.363Y_{4\tau+2,2} - 0.420Y_{4\tau+1,1} + 1.255Y_{4\tau+2,1} \\
\hat{Y}_{4\tau+4,2} &= 0.011 + 0.834Y_{4\tau+3,2} - 1.242Y_{4\tau+2,1} + 1.411Y_{4\tau+3,1} \qquad (4.6)
\end{aligned}$$

whereas the SANVAR equations are

$$\begin{aligned}
\hat{Y}_{4\tau+1,1} &= -0.164 - 0.702Y_{4\tau-1,1} + 1.491Y_{4\tau,1} \\
&\quad +0.130(Y_{4\tau-1,1} + 0.300)_+ - 2.253(Y_{4\tau-1,1} - 1.100)_+ \\
&\quad +0.253(Y_{4\tau,1} + 0.600)_+ + 1.282(Y_{4\tau,1} - 0.800)_+ \\
\hat{Y}_{4\tau+2,1} &= 0.021 - 0.597Y_{4\tau,1} + 1.322Y_{4\tau+1,1} \\
&\quad +0.211(Y_{4\tau,1} + 0.600)_+ - 0.341(Y_{4\tau,1} - 0.800)_+ \\
&\quad -0.170(Y_{4\tau+1,1} + 0.633)_+ + 0.105(Y_{4\tau+1,1} - 1.333)_+ \\
\hat{Y}_{4\tau+3,1} &= -0.103 + 0.207Y_{4\tau+1,1} + 0.410Y_{4\tau+2,1} \\
&\quad -0.673(Y_{4\tau+1,1} + 0.633)_+ - 0.841(Y_{4\tau+1,1} - 1.333)_+ \\
&\quad +1.063(Y_{4\tau+2,1} + 0.433)_+ + 0.490(Y_{4\tau+2,1} - 1.433)_+ \\
\hat{Y}_{4\tau+4,1} &= 0.073 - 0.627Y_{4\tau+2,1} + 1.681Y_{4\tau+3,1}
\end{aligned}$$

$$
\begin{aligned}
& +0.057(Y_{4\tau+2,1}+0.433)_+ + 0.023(Y_{4\tau+2,1}-1.433)_+ \\
& -0.169(Y_{4\tau+3,1}+0.300)_+ - 0.277(Y_{4\tau+3,1}-1.100)_+ \qquad (4.7)\\
\hat{Y}_{4\tau+1,2} &= 0.356 + 1.018Y_{4\tau,2} - 0.613(Y_{4\tau,2}+0.633)_+ + 1.052(Y_{4\tau,2}-1.433)_+ \\
\hat{Y}_{4\tau+2,2} &= -0.121 + 0.874Y_{4\tau+1,2} + 0.031(Y_{4\tau+1,2}+0.867)_+ \\
& \quad + 0.246(Y_{4\tau+1,2}-2.167)_+ \\
\hat{Y}_{4\tau+3,2} &= 0.054 + 0.781Y_{4\tau+2,2} - 0.059(Y_{4\tau+2,2}+1.233)_+ \\
& \quad + 0.237(Y_{4\tau+2,2}-2.033)_+ \\
\hat{Y}_{4\tau+4,2} &= -0.149 + 0.797Y_{4\tau+3,2} + 0.305(Y_{4\tau+3,2}+0.800)_+ \\
& \quad - 0.788(Y_{4\tau+3,2}-1.800)_+ \qquad (4.8)
\end{aligned}
$$

Plots similar to Figures 1 and 2 have also been created for the whites and African Americans, see Figures 3 and 4, based on equations (4.7), (4.5), (4.8) and (4.6) respectively. The PVAR model (4.5) for the unemployment rates of whites actually predicts better than the SANVAR model (4.7). Again, this is due to the fact that the fitted PVAR model for whites has only two explanatory variables and is very similar to the SANVAR model. Therefore, one should always use the linear periodic VAR model for better prediction when the two models produce similar results. For the series of African Americans, the opposite is true, where the SANVAR predicts much better than the PVAR. Also it is worth noticing that for both whites and African Americans, the preferred forecasting model is always a univariate series prediction model. To be precise, the PVAR model (4.5) for the whites and the SANVAR model (4.8) for African Americans both suggest that prediction for different races be best done separately. This strongly suggests that whites and African Americans have been living in parallel economies and there is little interaction of their unemployment rates.

Overall, the SANVAR model is a more robust option than the PVAR model. It nearly always predicts better, except when the series is extremely close to linearity, which is always indicated by the lack of parsimony of the fitted SANVAR model. If the fitted SANVAR model is less parsimonious than the fitted PVAR model (i.e., having more or the same number of variables), one should use the simpler PVAR model for forecasting and inference. In addition, the model is able to detect from the data whether there is any significant interaction among the individual series.

Acknowledgement

The research has been partially supported by National Science Foundation award DMS 0405330, and by award SES 0127722 while being an ASA/NSF/BLS Research Fellow at the U.S. Bureau of Labor Statistics. Helpful comments and

suggestions from Stuart Scott at the U.S. Bureau of Labor Statistics and from an anonymous referee are gratefully acknowledged.

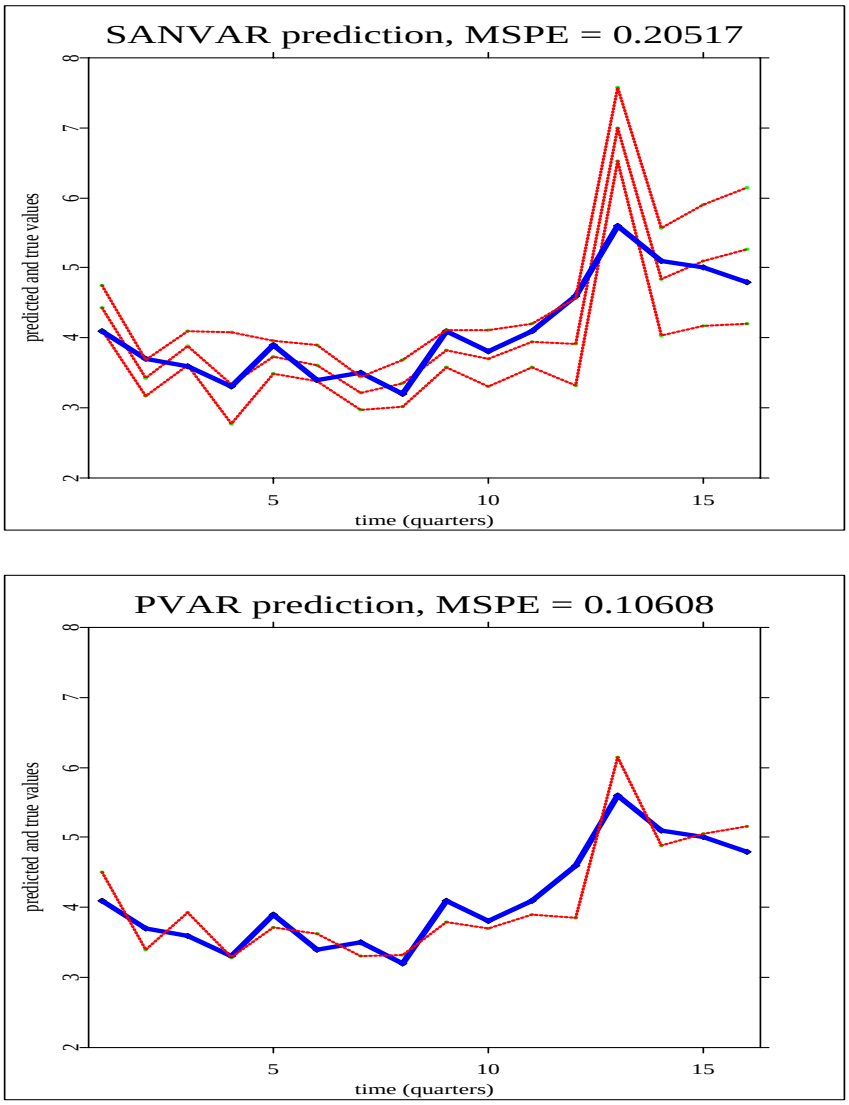

Figure 3: Forecasting the white's quarterly unemployment rates of 1999-2002, based on white's and African American's unemployment rates of 1972-1998. The solid thick line represents the actual unemployment rates during 1999-2002, the thin dashed line represents the forecasts. Both the parametric PVAR and the nonparametric SANVAR models are used. In the plot for SANVAR model, nonparametric confidence band for the predicted means are also plotted. The MSPE is calculated as the mean squared prediction error between the predicted and true unemployment rates.

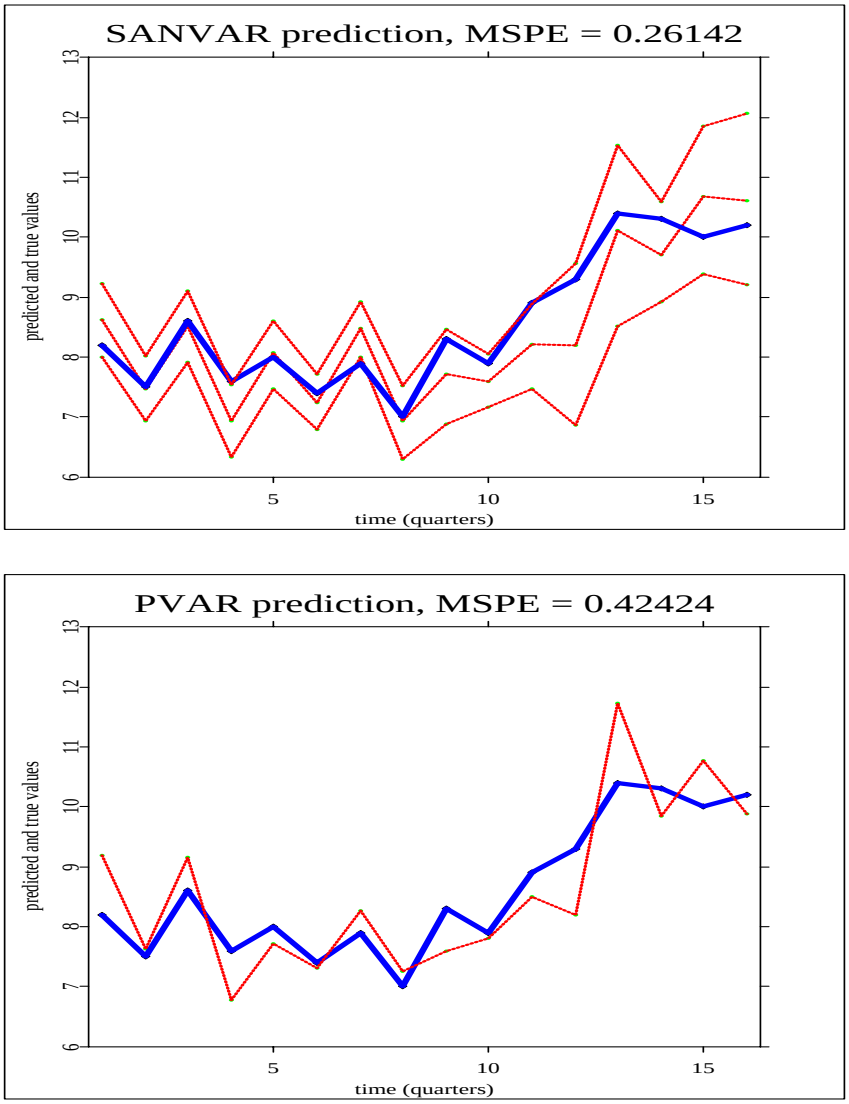

Figure 4: Forecasting the African American's quarterly unemployment rates of 1999-2002, based on white's and African American's unemployment rates of 1948-1998. The solid thick line represents the actual unemployment rates during 1999-2002, the thin dashed line represents the forecasts. Both the parametric PVAR and the nonparametric SANVAR models are used. In the plot for SANVAR model, nonparametric confidence band for the predicted means are also plotted. The MSPE is calculated as the mean squared prediction error between the predicted and true unemployment rates.

References

Auestad, B. and Tjøstheim, D. (1990). Identification of nonlinear time series: first order characterization and order determination. *Biometrika* **77**, 669-687.

Chen, R. and Tsay, R. S. (1993). Nonlinear additive ARX models. *Journal of the American Statistical Association* **88**, 955-967.

Claeskens, G. and Van Keilegom, I. (2003). Bootstrap confidence bands for regression curves and their derivatives. *The Annals of Statistics* **31**, 1852–1884.

Fan, J. and Yao, Q. (1998). Efficient estimation of conditional variance functions in stochastic regression. *Biometrika* **85**, 645-660.

Franses, H. F. (1996). *Periodicity and Stochastic Trends in Economic Time Series*. Oxford University Press.

Györfi, L. , Härdle, W. , Sarda, P. and Vieu, P. (1989). *Nonparametric Curve Estimation From Time Series*. Springer-Verlag.

Härdle, W., Hlavka, Z. and Klinke, S. (2000). *XploRe Application Guide*. Springer-Verlag.

Härdle, W. , Tsybakov, A. and Yang, L. (1998). Nonparametric vector autoregression. *Journal of Statistical Planning and Inference* **68**, 221-245.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models*. Chapman and Hall.

Huang, J. Z. , Wu, C. O. and Zhou, L. (2002). Varying-coefficient models and basis function approximations for the analysis of repeated measurements. *Biometrika*, **89**, 111-128.

Huang, J. and Yang, L. (2004). Identification of nonlinear additive autoregressive models. *Journal of the Royal Statistical Society Series B* **66**, 463-477.

Linton, O. B. and Nielsen, J. P. (1995). A kernel method of estimating structured nonparametric regression based on marginal integration. *Biometrika* **82**, 93-100.

Lütkepohl, H. (1993). *Introduction to Multiple Time Series Analysis*. Springer-Verlag.

Robinson, P. M. (1983). Nonparametric estimators for time series. *Journal of Time Series Analysis* **4**, 185-207.

Sperlich, S., Tjøstheim, D. and Yang, L. (2002). Nonparametric estimation and testing of interaction in additive models. *Econometric Theory* **18**, 197-251.

Tjøstheim, D. and Auestad, B. (1994). Nonparametric identification of nonlinear time series: projections. *Journal of the American Statistical Association* **89**, 1398-1409.

Wang, J. and Yang, L. (2005). Polynomial spline confidence bands for regression curves. *Annals of Statistics*, tentatively accepted.

Wolters, J. (1992). Persistence and seasonality in output and employment of the Federal Republic of Germany. *Rechereches Economiques de Louvain* **58**, 421–439.

Xia, Y. (1998). Bias-corrected confidence bands in nonparametric regression. *Journal of the Royal Statistical Society Series B* **60**, 797-811.

Yang, L. and Tschernig, R. (2002). (2002) Non- and semiparametric identification of seasonal nonlinear autoregression models. *Econometric Theory* **18**, 1408-1448.

Yang, L. (2004). Confidence band for additive regression model, Research Manuscript 635, Michigan State University, Department of Statistics and Probability.

Yao, Q. and Tong, H. (1994). On subset selection in non-parametric stochastic regression. *Statistica Sinica* **4**, 51-70.

Received August 6, 2005; accepted December 6, 2005.

Lijian Yang
Department of Statistics & Probability
Michigan State University
East Lansing, MI 48824
yang@stt.msu.edu