REVSTAT – Statistical Journal
Volume 7, Number 1, April 2009, 23–36

THE SVM APPROACH FOR BOX–JENKINS MODELS

Authors: Saeid Amiri
– Dep. of Energy and Technology, Swedish Univ. of Agriculture Sciences, P.O. Box 7032, SE 750 07 Uppsala, Sweden
saeid.amiri@et.slu.se

Dietrich von Rosen
– Dep. of Energy and Technology, Swedish Univ. of Agriculture Sciences, P.O. Box 7032, SE 750 07 Uppsala, Sweden
Dietrich.von.Rosen@et.slu.se

Silvelyn Zwanzig
– Department of Mathematics, Uppsala University, Box 480, SE 751 06 Uppsala, Sweden
zwanzig@math.uu.se

Abstract:

- Support Vector Machine (SVM) is known in classification and regression modeling. It has been receiving attention in the application of nonlinear functions. The aim is to motivate the use of the SVM approach to analyze the time series models. This is an effort to assess the performance of SVM in comparison with ARMA model. The applicability of this approach for a unit root situation is also considered.

Key-Words:

- *Support Vector Machine; time series analysis; unit root.*

AMS Subject Classification:

- 49A05, 78B26.

1. INTRODUCTION

Time series analysis is the study of observations made sequentially in time. It is a complicated field in statistics because of direct and indirect effects of time on the variables in the model. The essential difference between the modeling via time series and ordinary method is that data points taken over time may have an internal relation that should be accounted for. It can be a correlation structure, a trend, seasonality and so on.

Time series can be studied in the time domain and in the time frequency domain. The time domain is more known among researchers in sciences whereas the frequency domain has many applications in engineering. Time domain is modeled by two main approaches. The traditional approach has been given in Box and Jenkins (1970) in their influential book, includes a systematic class of models called autoregressive integrated moving average (ARIMA) (see, for example, Shumway and Stoffer (2000) and Pourahmadi (2001)). A defining feature of these models is that they are multiplicative models, meaning that observed data are assumed to result from the products of factors involving differential or difference equation operators responding to a white noise input.

Other approaches use additive models or structural models. In this approach, it is assumed that the observations include sum of components, each of which deals with a specified time series structure. None of them have inferential tools such as the Box–Jenkins model, for example model selection, parameter estimation and model validation. ARIMA model can therefore be considered as a benchmark model in evaluating the performance of new method. Support Vector Machine is one of the new methods in modeling that has good performance in classification and regression analysis. A few papers have tried to use it for time series, see Müller (1997) and Murkharejee (1997). They have considered dynamic models e.g., the Mackey class equation was used to show the efficiency of SVM.

We are motivated to use SVM because of its ability in dealing with stationary as well as non-stationary series. Moreover, contrary to the traditional methods of time series analysis (autoregressive or structural models that assume normality and stationarity of the series), SVM makes no prior assumptions about the data.

The paper contains five sections and is organized as follows. In Section 2, the necessary theoretical background is provided and the SVM modeling is concisely described. In Section 3, it is shown that the approach of time series modeling can be written as a SVM model. Section 4 includes the discussion of the data and also present the results. Finally some conclusions are given in Section 5.

2. SUPPORT VECTOR MACHINE

During the last decades many researchers have been working on SVM in a variety of fields and it has in fact been a very active field. SVM has impacted on improving the statistical learning method and has been used to solve problems in classification. The SVM approach has improved the modeling, especially for nonlinear models. The review of Burges (1998), Cristianini and Shaw-Taylor (2000) and Bishop (2006) help to understand the concept of SVM. For more details see Vapnik (1995) and Vapnik (1998). Let us briefly consider the SVM regression approach.

In statistics, the aim of modeling is often to find a function $f(x)$ which predicts y in a model $y = f(x) + error$. It is not easy to find $f(x)$. It can be interpolated by using mathematical methods and approximated by using statistical methods. Via some statistical criteria like sum of squares or maximum likelihood, ML, the model can be exploited. To evaluate the procedure, one needs a criterion or loss function. It is defined as "ignoring observation which error is less than ϵ",

$$L(x, y, f) \;=\; \big|y - f(x)\big|_\epsilon = \max\big(0, \big|y - f(x)\big| - \epsilon\big)\,.$$

It is called "ϵ-insensitive error function". Another loss function is Huber's loss function which is the squared distance between the observations and the function, see Cristianini and Shaw-Taylor (2000) and Hasti *et al.* (2001). In Figure 1, the points outside the tube around the function are called slack variables which is shown by ξ_{1i} and ξ_{2j} for above and below the tube, respectively. The value of the points inside the tube is zero and outside is nonzero. To find ξ_{1i} and ξ_{2j}, one should estimate parameters by the error function as below,

$$\begin{aligned} \text{minimize} \quad & \sum_{i=1}^{N}(\xi_{1i} + \xi_{2i}) + \frac{\lambda}{2}\,\|W\|^2\,, \\ \text{subject to} \quad & y_i \le f + \epsilon + \xi_{1i}\,, \\ & y_i \ge f - \epsilon - \xi_{2i}\,, \\ & \xi_{1i}, \xi_{2j} \ge 0\,. \end{aligned}$$

By using the Lagrange multiplier to find parameters and optimize by the Karush–Kuhn–Tucker condition, $f(x)$ can be shown to equal

$$f(x) = \sum_{i=1}^{N} \alpha_i\, k(x, x_i)\;, \tag{2.1}$$

where α_i are support vectors, i.e. those points that contribute in the prediction. All points within the tube have $\alpha_i = 0$ and a few of α_i are nonzero. In (2.1), $k(x, x_i)$ is the kernel function, which is an inner product of variables, i.e.,

$$k(x, x_i) = \big\langle \phi(x), \phi(x_i) \big\rangle\;. \tag{2.2}$$

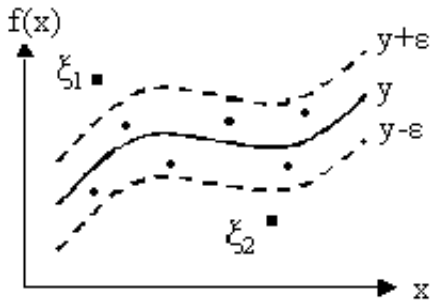

Figure 1: SVM regression with insensitive tube, slack variables ξ_1, ξ_2 and observations.

The following are some kernels:

Linear kernel	$k(x, x') = \langle x, x' \rangle$,
Polynomial kernel	$k(x, x') = \big(a\langle x, x' \rangle + k\big)^d$,
Radial Basis Function kernel (RBF)	$k(x, x') = \exp\big(-\sigma \|x - x'\|^2\big)$,
Laplacian kernel	$k(x, x') = \langle x, x' \rangle \exp\big(-\sigma \|x - x'\|\big)$.

Other kernels are the hyberbolic tangent kernel, the spline kernel, the Bessel and the ANOVA RBF kernel. The number of kernels is unlimited and new kernels can be found by combining existing ones (for more information see Burges (1998), Shaw-Taylor (2000) and Karatzoglou *et al.* (2007)). There are several advantages and disadvantages; SVM is based on the kernel, hence the suitable kernel selection is most important step. However, in practice one needs to study only a few kernel functions (Burges (1998)). The key in SVM is the transformation of a nonlinear problem to a higher dimensional linear space using the kernel function. SVM is not based on any assumptions about the distribution.

3. TIME SERIES ANALYSIS

The Box–Jenkins approach involves identifying an appropriate ARMA process by a mathematical model for forecasting. This model is a combination of AR and MA models. AR(p) is defined as bellow,

$$x_{t+1} = \sum_{j=1}^{p} \phi_j \, x_{t+1-j} + \epsilon_{t+1} \ . \tag{3.1}$$

If one considers the series to be deterministic as linear dynamic systems, a method based on the linear measure such as ARMA model can be used for analysis of the series. However, observed real data are rarely normally distributed and

tend to have marginal distributions with heavier tails. It has been shown that most of the financial time series are nonlinear (see, for example, Soofi and Cao (2002)). Based on the second scenario, we should use the method which has the capability to capture both the linearities and the nonlinearities of the series (see, for example, Hassani *et al.* (2009a) and Hassani *et al.* (2009b)). Here the nonlinear model can be written as

$$x_{t+1} = \sum_{j=1}^{p} \phi_j \, h_j(x_{t+1-j}) + \epsilon_{t+1} \ , \qquad \epsilon_{t+1} \sim N(0, \sigma^2) \ , \tag{3.2}$$

$$x_{t+1} = \big(h_1(x_t), ..., h_p(x_{t+1-p})\big) \begin{pmatrix} \phi_1 \\ \vdots \\ \phi_p \end{pmatrix} , \tag{3.3}$$

$$x = H\,\phi \ , \tag{3.4}$$

where $H = \big(h_1(\cdot), ..., h_p(\cdot)\big)$ and $\phi = (\phi_1, ..., \phi_p)^{\mathrm{T}}$. If H is known, the parameters can be estimated. To simplify assume $\underline{x}_t = (x_t, x_{t-1}, ..., x_{t+1-p})$, $p < t$. The parameters of the model can be estimated by the conditional ML:

$$\begin{aligned} L(\phi, \sigma | \underline{x}_p) &= f(x_{p+1}|\underline{x}_p)\, f(x_{p+2}|\underline{x}_{p+1}) \cdots f(x_t|\underline{x}_{t-1}) \\ &= \prod_{i=p}^{t-1} f(x_{i+1}|\underline{x}_i) \\ &= \prod_{i=p}^{t-1} \frac{1}{\sqrt{2\pi}\,\sigma} \, \exp - \frac{\big(x_{i+1} - \sum_{j=1}^{p} \phi_j \, h(x_{i+1-j})\big)^2}{2\,\sigma^2} \\ &= \left(\frac{1}{2\pi\sigma^2}\right)^{(t-p)/2} \exp - \sum_{i=p}^{t-1} \frac{\big(x_{i+1} - \sum_{j=1}^{p} \phi_j \, h(x_{i+1-j})\big)^2}{2\,\sigma^2} \ . \end{aligned} \tag{3.5}$$

Thus, one needs to minimize,

$$SS = \sum_{i=p}^{t-1} \left(x_{i+1} - \sum_{j=1}^{p} \phi_j \, h_j(x_{i+1-j}) \right)^2 = \sum_{i=p}^{t-1} (x_{i+1} - H_i\,\phi)^2 \ . \tag{3.6}$$

To improve the accuracy of the estimation procedure, one can use a penalty function,

$$SS2 = \sum_{i=p}^{t-1} (x_{i+1} - H_i\,\phi)^2 + \lambda \|\phi\| \ = \ (x - H\phi)^{\mathrm{T}} (x - H\phi) + \lambda \|\phi\| \ , \tag{3.7}$$

$$\frac{\partial SS2}{\partial \phi} = 0 \implies -H^{\mathrm{T}}(x - H\phi) + \lambda\,\phi = 0 \ ,$$

which implies that

$$H\phi = (HH^{\mathrm{T}} + \lambda\, I)^{-1} HH^{\mathrm{T}} x \ , \tag{3.8}$$

where HH^{T} is a matrix of inner product of the observations. It is quite straightforward to show that (3.8) can be written as an inner product. Therefore, the nonlinear equation can be written as a kernel function,

$$x_{t+1} = f(\underline{x}_t) + e_{t+1} = \sum_{i=1}^{p} \phi_i \, h_i(x_{t+1-i}) + e_{t+1} = \sum_{i=1}^{t} \alpha_i \, k(\underline{x}_t, \underline{x}_i) + e_{t+1} \, . \tag{3.9}$$

Another formula that can be considered is the use of time index, as independent, in the model. This is a reasonable variable as the time series data are collected during time,

$$x_t = \sum_{i=1}^{t} \alpha_i \, k(\underline{x}_t, i) \, . \tag{3.10}$$

Let us now consider the moving average model of order q, MA(q),

$$x_t = \sum_{j=0}^{q} \theta_j \, w_{t-j} \, , \qquad w_t \sim N(0, \sigma^2) \, . \tag{3.11}$$

The previous procedure follows by using a nonlinear function,

$$x_t = \sum_{j=0}^{q} \theta_j \, h(w_{t-j}) \, .$$

It is difficult to decide about the distribution of $h(\cdot)$ beforehand. With the assumption $h(w_{t-j}) \sim N(\mu_n, \sigma_n^2)$, there is no improvement for modeling. However, if the model is invertible, we can write MA as AR and follow the previous model. Hence, there are two problems: the distribution of $h(\cdot)$ and the invertibility of the model which make the behavior of MA a bit unclear for using kernel. The similar problem exists for ARMA(p, q). There are two viewpoints: first, ignorance of MA in the model and considering ARMA(p, q) as AR, and second, if ARMA(p, q) is invertible, then ARMA can be written as AR directly. At any rate, the procedure of AR process can be used.

Let us now consider a unit root process:

$$x_t = \mu + x_{t-1} + w_t = \mu + \mu + x_{t-2} + w_{t-1} + w_t = \cdots = t\mu + x_0 + \sum_{i=0}^{t} w_i \, . \tag{3.12}$$

This is a problem for the Box–Jenkins approach as it violates the stationarity condition, and therefore one can not formulate the Box–Jenkins model (see, for example, Brockwell and Davis (1991)). The modeling of the unit root has been discussed extensively in the literature. There exist some statistical tests for diagnosis and also modeling in the special conditions. Equation (3.12) tells us that the unit root has a regression form of time but because of dependency between

observations, the common regression can not be used for it. In this case, one can use SVM, using the previous discussion and rewriting it as kernel formula. It is not based on the distribution and hence the dependency does not affect on it. It should be noted that, if $\mu = 0$ then this model has major drawback and behaves randomly.

4. APPLICATIONS

In this section, the applicability of SVM for time series analysis is considered. In order to performs the comparison, two different criteria are used: sum of squared residuals (SSR) and Akaike Information Criterion (AIC). AIC is calculated based on $\ln \widehat{\sigma}_k^2 + \frac{2k}{n}$, where $\widehat{\sigma}_k^2 = \frac{SSR}{n}$, k and n are the number of parameters and observations, respectively. In the following, the SVM approach is used in the modeling of AR(2), MA(1) and ARMA$(2,1)$ process.

4.1. AR

Here we use the series that has been used in Brockwell and Davis (1991), Example 9.2.1. The series includes 200 observations. Table 1 shows SSR and AIC of AR(2) and SVM with different kernels. SVM has been calculated using equation (3.9). In the table, the results of a few kernels are presented as SSR of other kernels were larger than AR(2). The results show the efficiency of the Laplacian kernel in comparison with the Box–Jenkins modeling. It should be noted that RBF with $\sigma = 50$ fitted fairly well.

Table 1: SSR and AIC of AR(2) and SVM with different kernels.

Model	SSR	AIC
AR(2)	176.99	−0.102
RBF[1]	171.73	−0.136
RBF[2]	144.33	−0.368
Bessel[1]	161.16	−0.176
Bessel[2]	194.46	0.009
Laplacian[1]	100.83	−0.664
Laplacian[2]	202.68	0.330
linear	177.75	−0.102
poly[3]	176.43	−0.085

[1] Fitted by $\sigma = 10$.
[2] Fitted by $\sigma = 50$.
[3] With 2 degrees.

The calculations in Table 2 are based on equation (3.10). This model uses the time as an independent variable. The table shows how much fitting has been improved. The Laplacian kernel and Bessel kernel have smaller SSR than AR, but other kernels have greater SSR than AR. These values show the Bessel kernel has been fitted well, but its variation is very large. The variation of Laplacian kernel is small in comparison with the Bessel kernel, and hence it seems to be more reliable to use. The Laplacian kernel, for this model, is better than the previous models.

Table 2: Modeling directly based on time for AR(2) with different kernels.

Model	SSR	AIC
Laplacian[1]	56.60	−1.252
Laplacian[2]	21.55	−2.217
Bessel[1]	29.50	−1.830
Bessel[2]	980.17	1.619

[1] Fitted by $\sigma = 10$.
[2] Fitted by $\sigma = 50$.

Moreover, consider AR(2) with $x_t = x_{t-1} - 0.9\,x_{t-2} + \omega_t$. This model is stationary and hence the Box–Jenkins model fits very well. To compare the Box–Jenkins model with SVM, the simulation of this model is performed 1000 times with 100 observations. The results for the Box–Jenkins model and different kernels are shown in Table 3. The first two columns include the results of using (3.9)

Table 3: Percent and order of model in simulation of AR.

Model	model based on x_t		model based on t	
	percent	order	percent	order
AR(2)	0.020	6.93	0.006	2.93
RBF[1]	0.283	3.67	0.00	9.18
RBF[2]	0.000	4.43	0.00	6.00
Bessel[1]	0.023	3.77	0.00	7.90
Bessel[2]	0.000	5.85	0.994	1.00
tangent[1]	0.000	12.51	0.000	12.63
tangent[2]	0.000	12.49	0.000	12.53
splinedot	0.000	14.51	0.000	14.42
spline1	0.000	14.48	0.000	14.36
Laplacian[1]	0.540	2.17	0.000	3.92
Laplacian[2]	0.003	6.27	0.000	2.14
linear	0.020	6.36	0.000	10.21
poly[3]	0.110	5.52	0.000	10.22
ANOVA[1]	0.000	10.98	0.000	7.52
ANOVA[2]	0.000	10.01	0.000	4.99

[1] Fitted by $\sigma = 10$.
[2] Fitted by $\sigma = 50$.
[3] With 2 degrees.

and the second two columns include the results of using (3.10). The order column is the mean of orders of models in all of the simulations and the percent shows how many times the model has the smallest SSR in the simulations. As it appears from Table 3, the Laplacian kernel in 54% time has minimum SSR using x_t, but Bessel kernel has minimum SSR using time as explanatory variable. The results of Table 3 is similar to those obtained in Table 1. Therefore, the Bessel and Laplacian kernel are suitable for AR. Table 2 also shows that the fitted model based on the time index as an explanatory variable has better performance than a model based on x_t.

4.2. MA

The Example 10.4.2 of Brockwell and Davis (1991) is a MA(1) process with 160 observations. Here we use the same series to examine the performance of the SVM modeling. The results are presented in Table 4.

Table 4: SSR and AIC of MA(1) and SVM with different kernel.

Model	SSR	AIC
MA(1)	147	−0.072
Bessel[1]	227.373	0.388
Bessel[2]	198.415	0.252
Laplacian[1]	178.720	0.123
Laplacian[2]	79.282	−0.689

[1] Fitted by $\sigma = 10$.
[2] Fitted by $\sigma = 50$.

The results show that the Laplacian kernel with large σ has been fitted very well to MA(1) and also SSR of using Bessel kernel is close to MA(1), but other kernels have not good performance. As it is mentioned above, SVM has a better performance for a AR(p) model than a MA model. For a AR model, the Laplacian kernel with small σ has smallest SSR, but for MA, the Laplacian kernel with larger σ has smallest SSR. For more clarification, see Table 5 which shows the result of the simulation $y_t = \omega_t + 0.5\,\omega_{t-1}$ with 100 observations. This includes the order and the percent of different models in comparison with the Box–Jenkins model. The results confirm the previous results that indicate the Laplacian kernel with large σ has fitted better, almost 88%, than other methods.

Table 5: Percent and order of model in simulation of MA.

Model	percent	order
MA(1)	0.000	8.08
RBF[1]	0.000	8.54
RBF[2]	0.000	5.00
Bessel[1]	0.000	6.512
Bessel[2]	0.112	2.90
tangent[1]	0.000	12.59
tangent[2]	0.000	12.44
Spline[1]	0.000	14.50
Spline[2]	0.000	14.46
Laplacian[1]	0.000	3.00
Laplacian[2]	0.888	1.11
linear	0.000	10.68
poly[3]	0.000	13.68
ANOVA[1]	0.000	6.89
ANOVA[2]	0.000	4.00

[1] Fitted by $\sigma = 10$.
[2] Fitted by $\sigma = 50$.
[3] With 2 degrees.

4.3. ARMA

Next we consider ARMA$(2, 1)$ with 200 observations from Brockwell and Davis (1991), Example 9.2.3. Table 6 shows SSR and AIC of ARMA$(2, 1)$ and different kernels. The first two columns include the results of using (3.9) and the second two columns include the results of using (3.10). It admits the efficiency of Laplacian kernel for the ARMA model. As it appears from the results, the Laplacian kernel has the smallest SSR in both cases.

Table 6: SSR and AIC of ARMA and SVM with different kernels.

Model	model based on x_t		model based on t	
	SSR	AIC	SSR	AIC
ARMA$(2, 1)$	197.16	0.0157		
RBF[1]	244.16	0.209	1536.55	2.048
RBF[2]	176.26	−0.008	1216.10	1.815
Bessel[1]	201.50	0.037	1460.53	2.018
Bessel[2]	195.39	0.006	56.82	−1.228
Laplacian[1]	116.96	−0.526	350.00	0.569
Laplacian[2]	200.14	0.010	46.76	−1.443

[1] Fitted by $\sigma = 10$.
[2] Fitted by $\sigma = 50$.
[3] With 2 degrees.

To simulate ARMA$(2,1)$, consider $x_t = 0.4\,x_{t-1} + 0.5\,x_{t-2} + \omega_t + 0.2\,\omega_{t-1}$. The simulation results are based on 1000 replications of 100 observations. The results of ARMA$(2,1)$ using the Box–Jenkins and SVM, using different kernels, were presented in Table 7. The results are similar to those obtained in Table 6, which is based on a time series data. As it appears from the table, in both models, equation (3.9) and (3.10), the Laplacian kernel has better performance than others. The Laplacian kernel, using x_t and time as explanatory variables, with $\sigma = 10$ has the smallest SSR in 92.3% and 66% of the simulations, respectively.

Table 7: Percent and order of model in simulation of ARMA$(2,1)$.

Model	model based on x_t		model based on t	
	percent	order	percent	order
ARMA$(2,1)$	0.000	8.80	0.000	9.00
RBF[1]	0.020	5.14	0.000	7.99
RBF[2]	0.000	3.33	0.000	4.93
Bessel[1]	0.000	4.23	0.000	6.47
Bessel[2]	0.002	3.86	0.044	2.33
tangent[1]	0.000	12.56	0.000	12.60
tangent[2]	0.000	12.46	0.000	12.39
Spline[1]	0.000	14.56	0.000	14.47
Spline[2]	0.000	14.40	0.000	14.52
Laplacian[1]	0.923	1.14	0.660	1.48
Laplacian[2]	0.045	3.81	0.296	2.39
Linear	0.000	9.51	0.000	10.87
Poly[3]	0.000	8.67	0.000	10.12
ANOVA[1]	0.000	9.64	0.000	6.48
ANOVA[2]	0.000	7.83	0.000	3.90

[1] Fitted by $\sigma = 10$.
[2] Fitted by $\sigma = 50$.
[3] With 2 degrees.

4.4. Unit root

Let us now consider the application of SVM for a unit root process. The model $x_t = x_{t-1} + \omega_t$ with 100 observations is simulated 1000 to study the SVM performance. The results of SVM modeling for the simulated series are presented in Table 8. For a better understanding of the SVM performance in modeling, the order of the model is presented in comparison with the other competitive methods and also the percent. In this case, modeling by ARMA model is impossible because of the non stationarity property of the series. Nonstationarity can often

be associated with different trends in the signal or heterogeneous segments with different local statistical properties. Table 8 indicates that the Laplacian kernel has been fitted very well to the series.

Table 8: Percent and order of the model in simulation of a unit root process.

Model	percent	order
RBF[1]	0.00	7.68
RBF[2]	0.019	4.08
Bessel[1]	0.000	6.23
Bessel[2]	0.036	2.77
tangent[1]	0.000	11.63
tangent[2]	0.000	11.36
spline[1]	0.000	13.57
spline[2]	0.000	13.42
Laplacian[1]	0.915	1.14
Laplacian[2]	0.002	5.68
linear	0.000	9.87
poly[3]	0.000	9.08
ANOVA[1]	0.002	5.75
ANOVA[2]	0.024	2.85

[1] Fitted by $\sigma = 10$.
[2] Fitted by $\sigma = 50$.
[3] With 2 degrees.

5. CONCLUSION

Although the Box–Jenkins model is still one of the most applied model in time series analysis, there are several major drawbacks; the Box–Jenkins models are based on the stationarity, but this is often not sufficient, for example modeling unit root process using ARMA approach is impossible.

The results of this study show that the ARMA models can be expressed as SVM. The performance of the SVM modeling is studied in comparison with the Box–Jenkins modeling. Particularly, the Laplacian kernel is superior to others. It is therefore concluded that the use of SVM for the ARMA model is of great interest and should be considered (see Section 3). Moreover, the use of time index, as explanatory variable, in modeling will improve the accuracy of the results (see Tables 3, 6 and 7). To clarify the performance of the SVM for time series analysis, several examples and simulated series are used. The empirical results confirm our theoretical results. Our findings also show that the SVM based on the Laplacian kernel works very well for the unit root process.

ACKNOWLEDGMENTS

The authors gratefully acknowledge Dr. Mats Gustafsson and the referees for the valuable suggestions that led to the improvement of this paper.

REFERENCES

[1] Bishop, C.M. (2006). *Pattern Recognition and Machine Learning*, Springer, New York.

[2] Box, G. and Jenkins, G. (1970). *Time Series Analysis: Forecasting and Control*, Holden-Day, San Francisco.

[3] Brockwell, P.J. and Davis, R.A. (1991). *Time Series: Theory and Methods*, 2nd ed., Springer, New York.

[4] Burges, C.J.C. (1998). A tutorial on support vector machines for pattern recognition, *Data Mining and Knowledge Discovery*, **2**(2), 121–167.

[5] Cristianini, N. and Shaw-Taylor, J. (2000). *An introduction to Support Vector Machine*, Cambridge University Press, New York.

[6] Hassani, H.; Heravi, H. and Zhigljavsky, A. (2009). Forecasting European industrial production with singular spectrum analysis, *International Journal of Forecasting*, doi:10.1016/j.ijforecast.2008.09.007.

[7] Hassani, H.; Dionisio, A. and Ghodsi, M. (2009). The effect of noise reduction in measuring the linear and nonlinear dependency of financial markets, *Nonlinear Analysis: Real World Applications*, doi:10.1016/j.nonrwa.2009.01.004.

[8] Hastie, T.; Tibshirani, R. and Friedman, J. (2000). *Elements of Statistical learning: Data Mining, Inference and Prediction*, Springer, New York.

[9] Karatzoglou, A. *et al.* (2007). *kernel lab package*, `http://cran.r-project.org/src/contrib/Descriptions/kernlab.html`

[10] Müller, K.R. *et al.* (1997). *Predicting time series with support vector machines*, ICANN'97, Berlin, 999–1004.

[11] Murkharejee, S. *et al.* (1997). *Nonlinear Prediction of Chaotic Time Series using Support Vector Machines*, IEEE workshop on Neural Network for Signal Processing.

[12] Pourahmadi, M. (2001). *Foundations of Time Series Analysis and Prediction Theory*, Wiley, New York.

[13] Shumway, R.H. and Stoffer, D.S. (2000). *Time Series Analysis and Its Applications*, Springer, New York.

[14] Soofi, A. and Cao, L. (Eds.) (2002). *Modelling and Forecasting Financial Data: Techniques of Nonlinear Dynamics*, Kluwer Academic Publishers, Boston.

[15] Vapnik, V.N. (1995). *The Nature of Statistical Learning Theory*, Springer, New York.

[16] Vapnik, V.N. (1998). *Statistical Learning Theory*, Wiley, New York.