Electric Power Demand Forecasting Based on Cointegration Analysis and a Support Vector Machine

ZHANG XING-PING, GU RUI
School of Business Administration,
North China Electric Power University
2 Beinong Road, Zhuxinzhuang, Deshengmenwai, Beijing 102206
CHINA

Abstract: In the process of cointegration analysis, electricity consumption is chosen as the explained variable, and GDP per capita, heavy industry share, and efficiency improvement are chosen as the explanatory variables; then a cointegration model is put forward, which shows that there is a cointegration relationship between the explained variable and explanatory variables. The explained and explanatory variables are input into a support vector machine (SVM), and a Gaussian radial basis function is taken as the kernel function. So an electricity demand forecasting model based on multivariate SVM is established. The example provides evidence for the validity of the forecasting model.

Key words: Support vector machine; Multivariate time series; Unit root test; Cointegration analysis; Embedded dimension; Electricity demand forecasting

1 Introduction

Electricity demand forecasting is the basis for electricity planning. Many scholars [1~5] have applied econometrics to study electricity demand and its main determining factors is usually analyzed correctly in theory, but it is greatly affected by fluctuations in the sample data. A lot of non-linear programming and combinational forecasting methods such as fuzzy logic methods are applied widely in electric load forecasting. But results produced by fuzzy logic methods are quite difficult to express and set up, and the parameters are not easy to modulate[6,7]. A new machine learning technique called support vector machines (SVM) is not only helpful for solving problems involving small sample, devilish learning, high dimension and local minima, but also enables strong generalizability. So SVM can be widely applied in electric load forecasting. some research [8~12] indicates that SVM has distinct advantages in electric load forecasting. SVM is seldom used in forecasting the electricity demand, and when it is, actual electricity consumption is taken as the only input variable of the SVM, while the major factors which impact electric power demand are not considered [13].

In this paper, a cointegration model from econometrics is applied to prove that GDP per capita, heavy industry share, and efficiency improvement are the key factors determining electricity demand. Taking these determining factors and actual electricity consumption as the input variables of the SVM, and selecting the rational kernel function of the SVM, the output variable of future electricity demand is obtained.

2 Multivariate Cointegration Analysis

Project supported by Education Ministry of China (06JC790014)
The authors are grateful to Robert Blohm rb112@columbia.edu for revising the English and making a few mathematical corrections.

of Electricity Consumption

2.1 Cointegration Theory

Cointegration theory seeks to determine whether there is a stationary relationship among nonstationary economic variables, and whether there is a long-term equilibrium relationship among them. It avoids the disadvantages of unreliable regression results generated by spurious regression, and it can differentiate long-term stationary relationships from short-term dynamic relationships among variables. Before cointegration analysis came along, the combination of variables had to be stationary. The variable autoregression model, which includes *g* variables and *k* lags, is expressed as:

$$y_t = \beta_1 y_{t-1} + \beta_2 y_{t-2} + \cdots + \beta_k y_{t-k} + \mu_t \qquad (1)$$

Supposed all y_t are I(1); then a suitable transformation of equation (1) is made, and the error correction model is obtained as:

$$\Delta y_t = \mathbf{\Pi} y_{t-k} + \sum_{i=1}^{k-1} \mathbf{\Gamma}_i \Delta y_{t-i} + \mu_t \qquad (2)$$

where $\mathbf{\Pi} = \sum_{j=1}^{k} \beta_j - \mathbf{I}_g$, $\mathbf{I}_g$ is the *g*-step unit matrix, and $\mathbf{\Gamma}_i = \sum_{j=1}^{i} \beta_j - \mathbf{I}_g$.

Matrix $\mathbf{\Pi}$ is the coefficient matrix which reflects the long-term relationships of the variables. When the variables are in a long-term equilibrium state, the difference in the first variables of equation (2) is the zero vector, and E(μ_t)=0; so, $\mathbf{\Pi} y_{t-k} = 0$ when the variables are in a long-term equilibrium state, and this can be judged by calculating the rank and the eigenvalues of matrix $\mathbf{\Pi}$.

When all the endogenous variables are I(1), and when all the variables of $\mathbf{\Pi} y_{t-k}$ are I(0), the stochastic error term is a stationary process. If $0 < Rank(\mathbf{\Pi}) = m < g$, there are matrices $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ such that $\mathbf{\Pi} = \boldsymbol{\alpha}\boldsymbol{\beta}^{\mathrm{T}}$, So equation (2) transforms into equation (3).

$$\Delta y_t = \boldsymbol{\alpha}\boldsymbol{\beta}^{\mathrm{T}} y_{t-k} + \sum_{i=1}^{k-1} \mathbf{\Gamma}_i \Delta y_{t-i} + \mu_t \qquad (3)$$

Each row of the matrix $\boldsymbol{\beta}^{\mathrm{T}} y_{t-k}$ is a stationary combined variable, that is, each row is a linear combined form which enables the variables $y_{1,t-1}, y_{2,t-1}, \cdots, y_{g,t-1}$ to be cointegrated.

2.2 Explained and Explanatory Variables

Lots of documents show that GDP plays the most important role in determining electricity consumption in China. Thus there is a positive correlation between electricity consumption and GDP.

In China, the share of industrial electricity consumption is rising, from 71.75% in 2000 to 74.89 ％ in 2006. Most of the electricity volume is consumed by heavy industry: in 2006 for example, electricity consumption by heavy industrial took up 60.26% of all electricity consumption, and 79.71% of industrial electricity consumption. The breakdown of electricity consumption has been changing in China; electricity consumption by light industry increased 1.87% and by heavy industry decreased 0.14% in 2006. So the heavy industry share, or the ratio of heavy industry production value to gross industry production reflects changing industrial structure.

As the science and technology level has steadily increased since 1997, the comprehensive social and technology level index rose 1.5% in 2006 to 47.11%, increased in comparison with that of last year. Consequently, efficiency improvement plays an important role in electricity consumption; so the ratio of increase in industrial value to industrial electricity consumption is used to reflect efficiency improvement.

So electricity consumption (*Q*) is chosen as the explained variable, and GDP per capita (*PCGDP*), heavy industry share (*HIS*), and efficiency improvement (*EI*) are chosen as the explanatory variables. The sample space is from 1985 to 2005， and the impact of inflation is removed.

2.3 The Cointegration Model

Because the economic variables in a time series are usually nonstationary, and there is neither randomness nor a definite tendency, the sample data should be transformed by taking the natural log so as to reduce vibration, and by taking the difference so as to eliminate instability and heteroscedasticity. Before cointegration analysis, the Augment Dikey-Fuller (ADF) test was applied to test whether a data series is stationary. The null hypothesis is that the data series is nonstationary. The test results are shown in Table1. (△ expresses the first order difference)

Table 1 ADF unit root test results on variables

Variables	ADF Test Statistic	5% Critical Value	Conclusion
LNQ	-3.423	-3.710	non-stationary
ΔLNQ	-3.168*	-3.066	stationary
LNPCGDP	-1.217	-3.691	non-stationary
ΔLNPCGDP	-4.107*	-3.733	stationary
LNHIS	-1.093	-3.658	non-stationary
ΔLNHIS	-4.413*	-3.674	stationary
LNEI	-1.316	-3.691	non-stationary
ΔLNEI	-3.802*	-3.733	stationary

Note:"*" expresses MacKinnon critical values for rejection of hypothesis of a unit root under the 5% significance level

In table 1 all the original values of the variables are less in absolute value than the ADF test statistic's critical value at the 5% significance level; so we fail to reject the null hypothesis at the 5% significance level. But all the computed ADF test statistic values of the first difference of the variables are greater in absolute value than the ADF test statistic's critical value at the 5% significance level, and so the null hypothesis is rejected at the 5% significance level, and so all the variables are I(1), and this meets the conditions for cointegration analysis. In other words, from 1985 to 2005, there may be a cointegration relationship between electricity consumption and the explanatory variables.

The cointegration test needs to be run to find whether there is a cointegration relationship. The null hypothesis is that there is no cointegration relationship between electricity consumption and the explanatory variables. All the observed series contain a time trend; so, the cointegration test model contains the intercept and time trend. The results of the Johansen cointegration test are shown in table 2.

Table 2 Results of Johansen Cointegration test

Eigenvalue	Likelihood	5 Percent Critical Value	Hypothesized No. of CE(s)
0.7561	60.950	47.856	None*
0.7184	34.141	29.797	At most 1*
0.3170	10.645	15.495	At most 2
0.1380	2.821	3.841	At most 3

Note: "*" expresses it is significant under 5% confidence level

The results in table 2 show that the Likelihood ratio of the first two eigenvalues is bigger than the critical value at the 5% significance level; therefore there is a long-term equilibrium relationship between electricity consumption and the explanatory variables. The normalized cointegration coefficients are shown in table3.

Table 3 Normalized Cointegration Coefficients

LNQ	LNPCGDP	LNHIS	LNEI	C
1.0000	1.01 (0.028)	0.13 (0.065)	-0.86 (0.040)	1.27

Note: the number in parenthesis in the table is the asymptotic standard error.

So the cointegration function is stated as:

$$\overset{\wedge}{LNQ} = 1.27 + 1.01LNPCGDP + 0.13LNHI\mathrm{S} - 0.86LNEI \quad (4)$$

If the residual series of equation (4) is stationary, there is a cointegration relationship between electricity and the explanatory variables; otherwise, there is no cointegration relationship. So the Johansen cointegration test is run to test whether the residual series is stationary, and the test results are shown in table 4.

Table 4 ADF Unit Toot Test Results on Residual Series

ADF Test Statistic	1 Percent Critical Value	5 Percent Critical Value	10 Percent Critical Value
-4.01	-4.62	-3.71	-3.30

The 5% critical value of the ADF test statistic is

-3.71; so, the computed ADF test statistic value of -4.01indicates that there are no unit roots in the residual series; that is, the residual series is stationary. So there is a cointegratoin relationship between electricity consumption and the explanatory variables.

In equation (4) the coefficients of the explanatory variables are the elasticity of *Q* with respect to the explanatory variables. That is, a 1% increase in *PCGDP* leads to, on average, a 1.01% increase in *Q*, a 1% increase in *HIS* increases *Q* by 0.13%, and a 1% improvement in *EI* decreases *Q* by 0.86% on average.

3 Multivariate SVM Model

3.1 Regression Arithmetic of SVM

Suppose $T=\{(x_1,y_1),\cdots,(x_i,y_i),\cdots,(x_p,y_p)\}$, where $x_i\in R^m$ is the input variable, $y_i\in R$ is the corresponding output value and p is the total number of the data points. Then the SVM regression function is:

$$f(x)=(\boldsymbol{\omega}\cdot\Phi(x))+b \tag{5}$$

where $\Phi(\cdot)$ is a non-linear mapping function, $\boldsymbol{\omega}$ is a weight vector, and b is the error term. $\boldsymbol{\omega}$ and b are estimated by:

$$\min R(\boldsymbol{\omega})=\frac{1}{2}\|\boldsymbol{\omega}\|^2+C\sum_{i=1}^{p}L_\varepsilon(y_i,f(x_i)) \tag{6}$$

where C is the punishment parameter, which is considered to specifies the trade-off between empirical risk and the model's flatness. $\frac{1}{2}\|\boldsymbol{\omega}\|^2$ is the normalization term. $L_\varepsilon(y_i,f(x_i))$ is called the ε-insensitive loss function, which is defined as:

$$L_\varepsilon(y_i,f(x_i))=\max(|y_i-f(x_i)|-\varepsilon,0) \tag{7}$$

In equation (7) the loss equals zero if the forecasting error is less than ε; otherwise the loss not less than ε. In order to represent the distance from actual values to the corresponding boundary values of the ε-band, two positive slack variables ξ and ξ^* are introduced. Then, equation (6) is transformed into the following constrained form:

$$J=\frac{1}{2}\|\boldsymbol{\omega}\|^2+C\sum_{i=1}^{p}(\xi_i^*+\xi_i) \tag{8}$$

$$s.t.\begin{cases} y_i-[\boldsymbol{\omega},\Phi(x)]-b\le\varepsilon+\xi_i^* & \xi_i^*\ge 0\\ [\boldsymbol{\omega},\Phi(x)]+b-y_i\le\varepsilon+\xi_i & \xi_i\ge 0\end{cases}$$

This constrained optimization problem is solved by using the following Lagrangian form:

$$\max H(\partial,\partial^*)=-\frac{1}{2}\sum_{i=1}^{p}\sum_{j=1}^{p}(\partial_i,\partial_i^*)(\partial_i,\partial_i^*)K(x_i,x_j)$$
$$+\sum_{i=1}^{p}\partial_i^*(y_i-\varepsilon)-\varepsilon\sum_{i=1}^{p}(y_i+\varepsilon)$$

$$s.t.\begin{cases}\sum_{i=1}^{p}(\partial_i-\partial_i^*)=0\\ \partial_i,\partial_i^*\in[0,C]\end{cases} \tag{9}$$

where ∂_i,∂_i^* are Lagrangian multipliers, and $\partial_i-\partial_i^*\ne 0$ i.e. corresponding data points are a support vector. By the Lagrange multipliers ∂_i and ∂_i^* calculated, an optimal desired weight vector of the regression hyperplane is obtained:

$$\omega^*=\sum_{i=1}^{p}(\partial_i-\partial_i^*)K(x_i,x) \tag{10}$$

Hence, the regression function is:

$$f(x)=\sum_{i=1}^{p}(\partial_i-\partial_i^*)K(x_i,x)+b \tag{11}$$

where $K(x_i,x)$ is called the kernel function. The value of the kernel function equals the inner product of $\Phi(x_i)$ and $\Phi(x)$, which are produced by mapping x_i and x into a higher dimensional feature space; that is:

$$K(x_i, x) = \Phi(x_i, x) \tag{12}$$

3.2 Multivariate SVM Model

For a univariate time series $\{x_1, x_2, \cdots, x_n\}$, training sample sets, $\{x_1, x_2, \cdots, x_m\} \to \{x_{m+1}\}$, $\{x_2, x_3, \cdots, x_{m+1}\} \to \{x_{m+2}\}$, $\cdots$ are established.

$$\mathbf{X} = \begin{bmatrix} x_1 & x_2 & \cdots & x_m \\ x_2 & x_3 & \cdots & x_{m+1} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n-m} & x_{\mathrm{n}-m+1} & \cdots & x_{n-1} \end{bmatrix} \tag{13}$$

$$\mathbf{Y} = [x_{m+1} \quad x_{m+2} \quad \cdots \quad x_n]^{\mathrm{T}} \tag{14}$$

$\{x_i, x_{i+1}, \cdots, x_{i+m-1}\}$ is the input vector, $\{x_{i+m}\}$ is the output value. m is the embedded dimension.

Supposed that we have observed an l-dimensional multivariate time series:

$$\{X_n\}_{n=1}^{N} = \{(x_{1,n}, x_{2,n}, \cdots, x_{l,n})\} \tag{15}$$

As in the case of a univariate time series, we make a state space reconstruction:

$$\begin{aligned} \mathbf{V}_n = \{ & x_{1,n}, x_{1,n-1}, \cdots, x_{1,n-m_1+1}; \\ & x_{2,n}, x_{2,n-1}, \cdots, x_{2,n-m_2+1}; \cdots; \\ & x_{l,n}, x_{l,n-1}, \cdots, x_{l,n-m_l+1}\}^{\mathrm{T}} \end{aligned} \tag{16}$$

m_i is the embedded dimension of i-th variable, $i = 1,2,\cdots,l$. The node quantity is the sum of the embedded dimensions in the multivariate time series, namely:

$$m = m_1 + m_2 + \cdots + m_l \tag{17}$$

4 Example

4.1 Input and Output Variables

The model (4) indicates that the *PCGDP*,*HIS* and *EI* are the determining factors of *Q*, and there is a long term equilibrium relationship among them. So *PCGDP*, *HIS*, *EI* and actual *Q* are input into the SVM. In order to eliminate dimensional diversity in the variations in each time series, data is normalized into the interval [0, 1].

4.2 Computing Results

Comparing the results calculated by 4 kinds of kernel function, the following Gaussian radial basis function was applied in the SVM.

$$K(x_i, x) = \exp(-|x_i - x|/2\sigma^2) \tag{18}$$

According to the principle of minimum error [14], let ε =0.0008, σ =3.5, and *C*=10000. The forecasted values of electricity demand are shown in table 5.

Table 5 Forecasted Electric power demand using SVR

Year	Actual values M (*KW.H*)	Forecasted values M (*KW.H*)	APE (%)
1994	9260.4	9141.1	1.29
1995	10023.4	9833.3	1.90
1996	10764.3	10529.2	2.18
1997	11284.5	11266.3	0.16
1998	11598.5	11652.4	0.47
1999	12305.2	12029.5	2.25
2000	13471.4	13059.2	3.06
2001	14633.5	14404.7	1.56
2002	16331.5	15890.8	2.70
2003	19031.6	18374.2	3.46
2004	21971.4	22794.1	3.74
2005	24940.4	24762.4	0.71
2006	28248.3	28727.5	1.70

$$\mathrm{APE} = \left| \frac{x_i - x_i^*}{x_i} \right| \times 100\%.$$

where x_i^* is the actual value, and x_i is the forecast value.

The mean absolute percentage error of the SVM forecast is 1.94%, and the maximum absolute percentage error is 3.74%. It is proved that a multivariate SVM model might enhance forecast precision effectively.

5 Conclusion

Two conclusions are obtained:

(1) There is a cointegration relationship between electricity consumption and the 3 explanatory variables in China; so, the 3 explanatory variables, GDP per capita, heavy industry share, and efficiency improvement, are determining factors influencing electricity consumption.

(2) Taking GDP per capita, heavy industry share, efficiency improvement, and actual electricity consumption as the input variables, and selecting the Gaussian radial basis function as the kernel function of the SVM, we have shown that the forecast accuracy of the SVM model may be higher than other models'.

Reference:

[1] Yuan Jiahai, Ding Wei, HU Zhaoguang. Cointegration and fluctuation analysis of electric power consumption and economic development[J]. *Power System Technology*, 2006, Vol. 30, No.9.

[2] Chien-Chiang Lee, Chun-Ping Chang. Structural breaks, energy consumption, and economic growth revisited: Evidence from Taiwan [J]. *Energy Economics*, Volume 27, Issue 6, November 2005，pp. 857-872.

[3] Erkan Erdogdu. Electricity demand analysis using cointegration and ARIMA modeling: A case study of Turkey [J]. *Energy Policy*, In Press, Corrected Proof, Posted online 17 April 2006.

[4] Ajith Abraham, Baikunth Nath. A neuro-fuzzy approach for modeling electricity demand in Victoria [J]. *Applied Soft Computing*, Volume 1, Issue 2, August 2001, pp.127-138.

[5] Yemane Wolde-Rufael. Electricity consumption and economic growth: a time series experience for 17 African countries [J]. *Energy Policy*, Volume 34, Issue 10, July 2006, pp.1106-1114.

[6] Taylor J W, Buizza R. Neural network load forecasting with weather ensemble predictions[J]. *IEEE Trans.Power Syst*., 2002, 17（3）, pp.626-632.

[7] Liu Mengliang1, Liu Xiaohua, Gao Rong. Short t6erm load forecasting using wavelet transform and SVM based on similar-days. *Transactions of China Electrotechnical Society*, 2006, 21(11), pp. 59-63.

[8] Xie Hong, Wei Jiangping, Liu Heli. Parameter selection and optimization method of SVM model for short-term load forecasting. *Proceedings of the CSEE*, 2006, 26（22）, pp. 17-22.

[9] Zhang Qian-jin Research on Electric-Power Load Forecasting Based on Support Vector Machine Regression Technique. *Aeronautical Computing Technique.* 2006, 36（4）, pp.105-110.

[10] Xie Hong，Chen Zhiye，Niu Dongxiao, et al．Research on a daily load forecasting model based on wavelet decomposition and climatic influence[J].*Proceedings of the CSEE*，2001，21(5), pp.5-10.

[11] Zhao Dengfu，Pang Wenchen，Zhang Jiangshe，et al．SVM for short term load forecasting based on Bayesian theory and online learning [J]．*Proceedings of the CSEE*，2005，25(13), pp.8-13.

[12] Huang Yuansheng, Zheng Yan, Qi Jianxun.The application of electric power demand forecasting based on LS-SVM. *Chinese Journal of Management Science*, 2005,(10).

[13] Wang Xiaohong, Wu Dehui. An annual electric consumption forecasting model based on least-square support vector machines. *Relay*,2006,(16).

[14]V Cherkassky，Y Ma． Practical selection of SVM parameters and noise estimation for SVM regressions[J]．*Neural Networks*，2004，17(1), pp.113-126