ARTICLE IN PRESS

Available online at www.sciencedirect.com

ScienceDirect

Expert Systems with Applications xxx (2008) xxx–xxx

Expert Systems with Applications

www.elsevier.com/locate/eswa

Forecasting volatility based on wavelet support vector machine

Ling-Bing Tang[a,b], Ling-Xiao Tang[c], Huan-Ye Sheng[a,*]

[a] *Computational Finance Laboratory, Department of Computer Science and Engineering, Shanghai Jiao Tong University, 800 Dongchuan Road, Minhang District, Shanghai 200240, China*

[b] *Department of Computer and Electronic Engineering, Hunan Business College, Yuelu Road, Yuelu District, Changsha 410205, China*

[c] *School of Economics, Changsha University of Science and Technology, 45 Chiling Road, Tianxin District, Changsha 410076, China*

Abstract

One of the challenging problems in forecasting the conditional volatility of stock market returns is that general kernel functions in support vector machine (SVM) cannot capture the cluster feature of volatility accurately. While wavelet function yields features that describe of the volatility time series both at various locations and at varying time granularities, so this paper construct a multidimensional wavelet kernel function and prove it meeting the mercer condition to address this problem. The applicability and validity of wavelet support vector machine (WSVM) for volatility forecasting are confirmed through computer simulations and experiments on real-world stock data.

Keywords: Volatility forecasting; Wavelet support vector machine (WSVM); Mercer condition

1. Introduction

Volatility often plays a crucial role in measuring total risk of financial assets, evaluating option prices and conducting hedging strategies (Day & Lewis, 1988; Harvey & Whaley, 1991; Hull & White, 1987; Poterba & Summers, 1986). Since the seminal works of Engle (1982) and Bollerslev (1986) about heteroskedastic return series models, it has become widely recognized in the academic finance literature that ARCH family are effective methods for volatility forecasting (Franses & van Dijk, 1996, Li & Mak, 1994). To enhance the forecasting performance of GARCH farther, neural network was introduced into this field, which has the functional flexibility to capture the non-linear relationship between past return innovations and future volatility. Donaldson and Kamstra (1997) have proposed the use of neural network-GARCH model to capture volatility effects in stock returns. Meissner and Kawano (2001) use a combined GARCH-neural network approach to capture the volatility smile of options on high-tech stocks. Neural networks confirm their usefulness in modeling the conditional volatility of stock returns due to their data-driven and nonparametric weak properties but one of the important weaknesses of neural networks is that they cannot avoid to get trapped in local minima (Bishop, 1996, Haykin, 1994). However, SVM, a novel type of neural network developed by Vapnik and his coworkers in 1995, can perfectly deal with this problem (Vapnik, 1995, 1998). Compared with most common neural network, SVM based on the structural risk minimization principle and linearly constrained quadratic programming theory can obtain better generation performance. In addition, the solution of SVM is unique and globally optimal. Consequently, Perez-Cruz proposed GARCH–SVM model and proved that forecasting volatility using SVM is not only feasible but also effective (Perez-Cruz, Afonso-Rodriguez, & Giner, 2003).

The prediction performance of SVM is greatly dependent upon the selection of kernel functions. There are many kinds

[*] Corresponding author. Tel.: +86 21 54748314.
E-mail addresses: lingbingtang@gmail.com (L.-B. Tang), lingxiaotang@gmail.com (L.-X. Tang), hysheng@sjtu.edu.cn (H.-Y. Sheng).

0957-4174/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2008.01.047

of existent support vector kernels such as the Gaussian and polynomial kernels used to map the data in input space to a high-dimensional feature space in which the problem becomes linearly separable (Schölkopf, Burges, & Smola, 1999). Since wavelet function can describe time series both at various locations and at varying time granularities (Daubechies, 1990, 1992, Mallat, 1989, 1998), it should describe the cluster feature of volatility well. Therefore, it is valuable for us to research the problem of whether a desirable performance could be achieved if we combine SVM with wavelet theory to construct a multidimensional wavelet kernel function to predict the conditional volatility of stock market returns based on GARCH model.

The objective of this paper is to evaluate the performance of wavelet kernel for volatility prediction according to GARCH model by comparing it with the Gaussian kernel in SVM. This paper is organized as follows: Section 2 provides a brief introduction to the theory of SVM for regression estimation. Section 3 describes how to construct wavelet kernel and prove that it is admissible support vector kernel. Section 4 discusses about the experimental results on both simulated and real data sets, followed by conclusions in the last section.

2. Theory of SVM for regression estimation

In the ε-insensitive support vector regression, our goal is to find a function *f*(*x*) that has ε deviation from the actually obtained target y_i or all training data and at the same time is as flat as possible. Suppose *f*(*x*) takes the following form:

$$f(x) = w \cdot x + b, \quad w \in X, \quad b \in R. \tag{1}$$

So, if we have a w with small norm, then we can say that f is flat. One way to do this is to minimize $\|w\|^2$ using the euclidean norm (Smola & Scholkopf, 1998) subject to the constraints of the so called ε-insensitive band

$$\begin{aligned} \min \quad & \frac{1}{2}\|w\|^2 \\ \text{Subject to} \quad & y_i - w \cdot x_i - b \leqslant \varepsilon \\ & w \cdot x_i + b - y_i \leqslant \varepsilon \\ & i = 1, 2, \ldots, l. \end{aligned} \tag{2}$$

One has to solve this problem in order to obtain an ε-insensitive SVR solution. Usually, we need to allow for some errors. We introduce slack variables ξ_i, ξ_i^* to cope with this situation. This case is called soft margin formulation. Specifically, we solve the following problem:

$$\begin{aligned} \min \quad & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}(\xi_i + \xi_i^*) \\ \text{Subject to} \quad & y_i - w \cdot x_i - b \leqslant \varepsilon + \xi_i \\ & w \cdot x_i + b - y_i \leqslant \varepsilon + \xi_i^* \\ & \xi_i, \xi_i^* \geqslant 0, C > 0 \\ & i = 1, 2, \ldots, l. \end{aligned} \tag{3}$$

where C determines the trade-off between the flatness of the *f*(*x*) and the amount up to which deviations larger than ε are tolerated.

The lagrangian function will help us to formulate the dual problem, which will give us a quadratic programming problem formulation. Next we construct the dual problem. The reason is that solving the primal problem is difficult due to many variables. If we use the dual problem formulation, we can decrease the variables and the size of the problem becomes smaller. Specifically,

$$\begin{aligned} \max \quad & -\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}(a_i - a_i^*) \cdot (a_j - a_j^*) \cdot x_i \cdot x_j \\ & - \varepsilon \cdot \sum_{i=1}^{l}(a_i + a_i^*) + \sum_{i=1}^{l} y_i \cdot (a_i - a_i^*) \\ \text{Subject to} \quad & \sum_{i=1}^{l}(a_i - a_i^*) = 0 \\ & a_i, a_i^* \in (0, C). \end{aligned} \tag{4}$$

Next we consider the non-linear case. First of all, we need to map the input space into the feature space and try to find a regression hyperplane in the feature space (Schölkopf et al., 1999). We can accomplish that by using the kernel function $k(x, y)$. In other words we replace k as follows:

$$k(x, y) = \varphi(x) \cdot \varphi(y) \tag{5}$$

Therefore, we can replace dot product of vectors in the feature space by using kernel functions. So, the problem becomes

$$\begin{aligned} \max \quad & -\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}(a_i - a_i^*) \cdot (a_j - a_j^*) \cdot k(x_i, x_j) \\ & \varepsilon \cdot \sum_{i=1}^{l}(a_i + a_i^*) + \sum_{i=1}^{l} y_i \cdot (a_i - a_i^*) \\ \text{Subject to} \quad & \sum_{i=1}^{l}(a_i - a_i^*) = 0 \\ & a_i, a_i^* \in (0, C). \end{aligned} \tag{6}$$

At the optimal solution, we obtain

$$\begin{aligned} w^* &= \sum_{i=1}^{l}(a_i - a_i^*) \cdot k(x_i, x) = 0 \\ f(x) &= \sum_{i=1}^{l}(a_i - a_i^*) \cdot k(x_i, x) + b^*. \end{aligned} \tag{7}$$

The main difference between the linear and the non-linear case is that in the non-linear case ω is not given explicitly anymore. On the other hand, by using dot products, it is defined uniquely (Smola & Scholkopf, 1998). Also, we work in the feature space not in the input space.

Please cite this article in press as: Tang, L. -B. et al., Forecasting volatility based on wavelet support vector machine, *Expert Systems with Applications* (2008), doi:10.1016/j.eswa.2008.01.047

3. Constructing and proving wavelet kernel

Let $x = (x_1, \dots, x_N) \in R^N$, Φ be a scaling function and Ψ a wavelet that yields an orthogonal basis of $L^2(R)$. We denote $\theta^0 = \Phi$ and $\theta^1 = \psi$. To any integer $0 \leqslant \lambda < 2^N$ written in binary form $\lambda = \lambda_1 \dots \lambda_N$, we define N-dimensional function

$$\psi^\lambda(x) = \theta^{\lambda_1}(x_1)\dots\theta^{\lambda_N}(x_N)$$
$$\theta^0 = \phi; \theta^1 = \psi$$

obviously,

$$\psi^0(x) = \phi(x_1)\dots\phi(x_N)$$

then, the family obtained by dilating and translating the $2^N - 1$ wavelets for $\lambda \neq 0$

$$\left\{\psi^\lambda_{j,k}(x) = 2^{-N*j/2}\psi^\lambda\left(\frac{x_1 - 2^j k}{2^j}, \dots, \frac{x_N - 2^j k}{2^j}\right)\right\}_{1\leqslant\lambda<2^N, j,k\in Z} \tag{8}$$

is an orthonormal basis of $L^2(R^N)$ (Mallat, 1998).

Theorem 1. *Let $\psi^\lambda(x)$ be a mother wavelet, let j and k denote the dilation and translation, respectively. If s, $t \in R^N$, then dot-product wavelet kernels are*

$$k(s,t) = \sum_{1\leqslant\lambda<2^N, j,k\in Z} \psi^\lambda_{j,k}(s)\psi^\lambda_{j,k}(t) \tag{9}$$

Proof. We prove the dot-product wavelet kernel is admissible support vector kernels.

We denote another orthonormal basis of L^2 (R^N)

$$\left\{\psi^{\lambda'}_{j',k'}(x) = 2^{-N*j'/2}\psi^{\lambda'} \times \left(\frac{x_1 - 2^{j'}k'}{2^{j'}}, \dots, \frac{x_N - 2^{j'}k'}{2^{j'}}\right)\right\}_{1\leqslant\lambda'<2^N, j',k'\in Z}$$

According to dual frame theory,

$$\forall 1 \leqslant \lambda' < 2^N, \quad j', k' \in Z,$$
$$\psi^{\lambda'}_{j',k'} = \sum_{1\leqslant\lambda<2^N,\ j,k\in Z} \langle \psi^{\lambda'}_{j',k'}, \overline{\psi^\lambda_{j,k}} \rangle_{L^2(R^N)} \psi^\lambda_{j,k} \tag{10}$$

where $\overline{\psi^\lambda_{j,k}} = \psi^\lambda_{j,k}$, because $\{\psi^\lambda_{j,k}\}_{1\leqslant\lambda<2^N, j,k\in Z}$ is orthonormal basis.

And then, according to reproducing kernel theory, we have

$$k(s,t) = \sum_{1\leqslant\lambda<2^N, j,k\in Z} \psi^\lambda_{j,k}(s) \times \sum_{1\leqslant\lambda'<2^N, j',k'\in Z} \psi^{\lambda'}_{j',k'}(t) \left\langle \psi^\lambda_{j,k}, \psi^{\lambda'}_{j',k'} \right\rangle_{L^2(R^N)}$$

Let $x^1, \dots, x^l \in R^N$ and $a_1, \dots, a_l \in R$,

$$\begin{aligned}
\sum_{p,q}^{l} a_p a_q K(x^p, x^q) &= \sum_{p,q}^{l} a_p a_q \sum_{1\leqslant\lambda<2^N, j,k\in Z} \psi^\lambda_{j,k}(x^p) \\
&\quad \times \sum_{1\leqslant\lambda'<2^N, j',k'\in Z} \psi^{\lambda'}_{j',k'}(x^q) \left\langle \psi^\lambda_{j,k}, \psi^{\lambda'}_{j',k'} \right\rangle_{L^2(R^N)} \\
&= \sum_{p,q}^{l} a_p a_q \sum_{\substack{1\leqslant\lambda<2^N, j,k\in Z \\ 1\leqslant\lambda'<2^N, j',k'\in Z}} \psi^\lambda_{j,k}(x^p)\psi^{\lambda'}_{j',k'}(x^q) \\
&\quad \times \left\langle \psi^\lambda_{j,k}, \psi^{\lambda'}_{j',k'} \right\rangle_{L^2(R^N)} \\
&= \left\langle \sum_{p}^{l} \sum_{1\leqslant\lambda<2^N, j,k\in Z} a_p \psi^\lambda_{j,k}(x^p)\psi^\lambda_{j,k}(.), \right. \\
&\quad \left. \times \sum_{q}^{l} \sum_{1\leqslant\lambda'<2^N, j',k'\in Z} a_q \psi^{\lambda'}_{j',k'}(x^q)\psi^{\lambda'}_{j',k'}(.) \right\rangle_{L^2(R^N)} \\
&= \left\| \sum_{p}^{l} \sum_{1\leqslant\lambda<2^N, j,k\in Z} a_p \psi^\lambda_{j,k}(x^p)\psi^\lambda_{j,k}(.) \right\|^2_{L^2(R^N)} \\
&\geqslant 0
\end{aligned}$$

Hence, dot-product kernels satisfy Mercer's condition. Therefore, it is admissible support vector kernel.

4. Experiment results

4.1. Simulated data sets

Two simulated data sets are examined in the first series of experiment. Each data set is generated by a GARCH(1,1) model

$$y_t = \mu + \sigma_t \varepsilon_t \tag{11}$$
$$\sigma_t^2 = \omega + \alpha y_{t-1}^2 + \beta\sigma_{t-1}^2 \tag{12}$$

where y_t is daily return and ε_t is innovation, an uncorrelated process with zero mean and unit variance. For the sake of simplicity, the mean of a financial return series μ is often neglected. In addition, the parameters ω, α and β must satisfy $\omega > 0$, α, $\beta \geqslant 0$ to ensure that the conditional variance σ_t^2 is positive. The experimental setup is as follows: $\mu = 0$, $\omega = 0.1$, $\alpha = 0.4$ and $\beta = 0.5$and a disturbance term ε_t distributed first as Gaussian and then as a Student's t with four degrees of freedom (kurtosis = 4). This second distribution tries to model the excess of kurtosis that appears in real financial series. Every time series consists of 1040 samples. The first 520 samples are used for training and the remaining 520 samples for test.

The SVM is used to forecast the volatility according to GARCH(1,1) model which is used to eliminate arch effects. It means that the better the forecasting performance of SVM, the better the standardized observation ($\hat{\varepsilon}_t = y_t/\hat{\sigma}_t$) is fitting the normal distribution. Therefore, the purpose of this experiment is to primarily validate the performance of wavelet kernel based on fit testing of standard observation, compared with Gaussian kernel which is one of the

Please cite this article in press as: Tang, L. -B. et al., Forecasting volatility based on wavelet support vector machine, *Expert Systems with Applications* (2008), doi:10.1016/j.eswa.2008.01.047

first support vector kernels for most of learning problem. Its expression is $k(s,t) = \exp(-\gamma\|s-t\|^2)$, where $\gamma > 0$ is a free parameter. The input variables are the lagged conditional variance σ^2_{t-1} and squared return y^2_{t-1}. The output variable is current conditional variance σ^2_t. The values of a width size γ, the penalty parameter C and the tube size ε are, respectively, chosen as 0.2666, 12.789 and 0.00008 by using cross validation. The same values of C and ε are used in WSVM for compare. The wavelet that has been used is a Daubechies wavelet with 4 vanishing moments. The dilation parameter j in wavelet kernel is set to $[-2,0]$. The Kolmogorov–Simirnov distance (KS) and Anderson–Darling distance (AD) are used as the criterion for the goodness of fit testing. They are defined as follows:

$$\mathrm{KS} = \sup_{x\in\Re}|F_s(x) - \widetilde{F}(x)| \tag{13}$$

$$\mathrm{AD} = \sup_{x\in\Re}\frac{|F_s(x) - \widetilde{F}(x)|}{\sqrt{\widetilde{F}(x)(1-\widetilde{F}(x))}} \tag{14}$$

where $F_s(x)$ is the empirical sample distribution and $\widetilde{F}(x)$ is the cumulative distribution function of the estimated parametric density and emphasizes the deviations around the median of the distribution. The AD statistic more accentuates the discrepancies in the tails than KS statistic.

We have run 10 independent trails with the same setup and reported the best value of the corresponding statistic over all trials. The superiority of the wavelet kernel over the Gaussian kernel in SVM is seen by examining Table 1. The table reports the computed KS and AD statistics for two simulated stocks under both Gaussian and Student's t distributional assumptions. They are referred to as Data-1 and Data-2. It shows that whether for in training set or test set, the value of KS statistic for the $\hat{\varepsilon}_t$ forecasted by wavelet kernel is below that of KS statistic for the $\hat{\varepsilon}_t$ forecasted by Gaussian kernel. The same is true for the AD statistic. It shows that the $\hat{\varepsilon}_t$ forecasted by wavelet kernel is better fit for Gaussian distribution than $\hat{\varepsilon}_t$ forecasted by Gaussian kernel. As for Student's t distributional assumption, we can draw the same conclusion. The wavelet kernel can give better predictions thanks to its good time–frequency property which can describe any kind of time series both at various locations and at varying time granularities.

Table 1
KS and AD statistic for simulated data: Gaussian distribution (Data-1) and Student's $t(4)$ distribution (Data-2)

Data set	Training set				Test set			
	Gaussian kernel		Wavelet kernel		Gaussian kernel		Wavelet kernel	
	KS	AD	KS	AD	KS	AD	KS	AD
Data-1	2.3886	13.3249	1.2378	2.9104	1.4572	3.2079	1.2618	1.9009
Data-2	1.9352	7.5821	1.8985	6.9929	1.9905	7.9406	1.9284	7.7945

Table 2
Descriptive statistics of the daily returns

	DAXINDX	FRCAC40	FTSE100	JAPDOWA	SPCOMP
Mean	0.00439	0.00212	0.00719	0.00374	0.01517
S.D.	0.10893	0.17430	0.13595	0.16883	0.18270
Skewness	−0.15341	−0.02455	0.04861	0.39928	−0.35353
Kurtosis	4.54437	3.77320	5.44952	6.74197	5.15058
$r(1)$	0.0370	0.0307	0.0573	0.0025	0.0229
$Q(20)$	34.0088	19.8654	35.0099	13.7866	43.4745
$r2(1)$	0.0108	0.0547	0.1755	0.1088	0.0277
$Q2(20)$	62.4672	82.2997	299.6639	183.4667	41.8994

$r(1)$, autocorrelation of order 1 of the original observations y_t; $r2(1)$,autocorrelation of order 1 of the squared observations y_t^2; $Q(20)$, Box–Ljung statistic for y_t (31.4 is the 5% critical value); $Q2(20)$, Box–Ljung statistic for y_t^2 (31.4 is the 5% critical value).

4.2. Financial data sets

The data examined in the experiment are composed of the following daily indices: DAXINDX, FRCAC40, FTSE100, JAPDOWA and SPCOMP. These stock market indices p_t are then transformed into daily returns y_t by 100 times their log differences

$$y_t = 100\ln(p_t/p_{t-1}) \tag{15}$$

All the index data encompass the period from January 1, 1992 to December 31, 1997. There are 1560 observations for each time series of daily return. Each whole data set is divided into five overlapping training and testing sets according to the walk-forward testing routine (Kaastra & Boyd, 1996). Each training and test set is moved forward through the time series by 130 observations, in which there are a total of 520 observations in the training set, 520 observations in the test set. The optimal values of C, ε and γ are chosen based on cross validation. The same values of C and ε are used in the WSVM. The same method is also used in the WSVM to choose the dilation j. The results are collated and the best results are recorded as follow, which are gained from the second sets (July 1992–July 1996).

Table 2 gives the descriptive statistics of the daily returns. It indicates that the return series show little correlation while its squares show high correlation coefficients. The values of standard Box Ljung statistics reinforce this evidence. Table 2 also depicts that all the series illustrate zero means and excess kurtosis for the normal distribution value. The FTSE100 returns series is plotted in Fig. 1. Fig. 2 depicted nonparametric estimates of the pdf of returns together with the corresponding normal density. And the autocorrelation coefficients are showed in Fig. 3. These figures can confirm the results reported in Table 2 about clustering, heavy tailedness and long-range dependence of this returns. Thus, it is

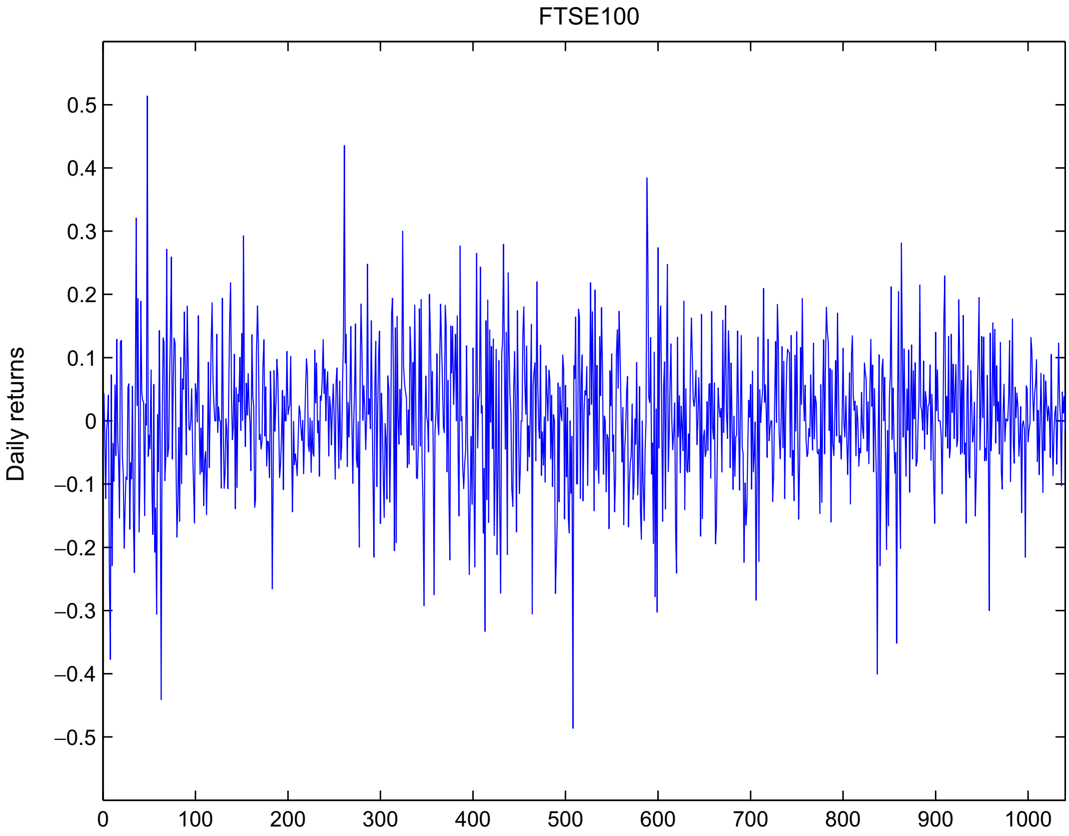

Fig. 1. Daily returns of y_t.

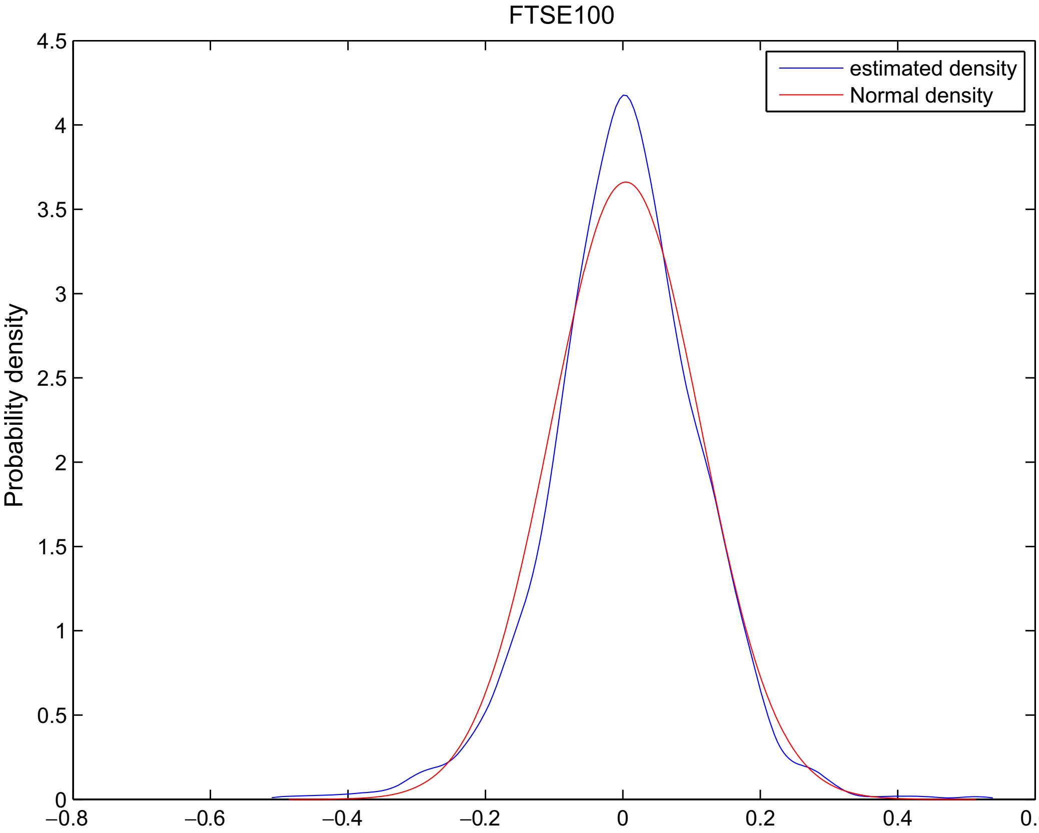

Fig. 2. Probability densities of y_t.

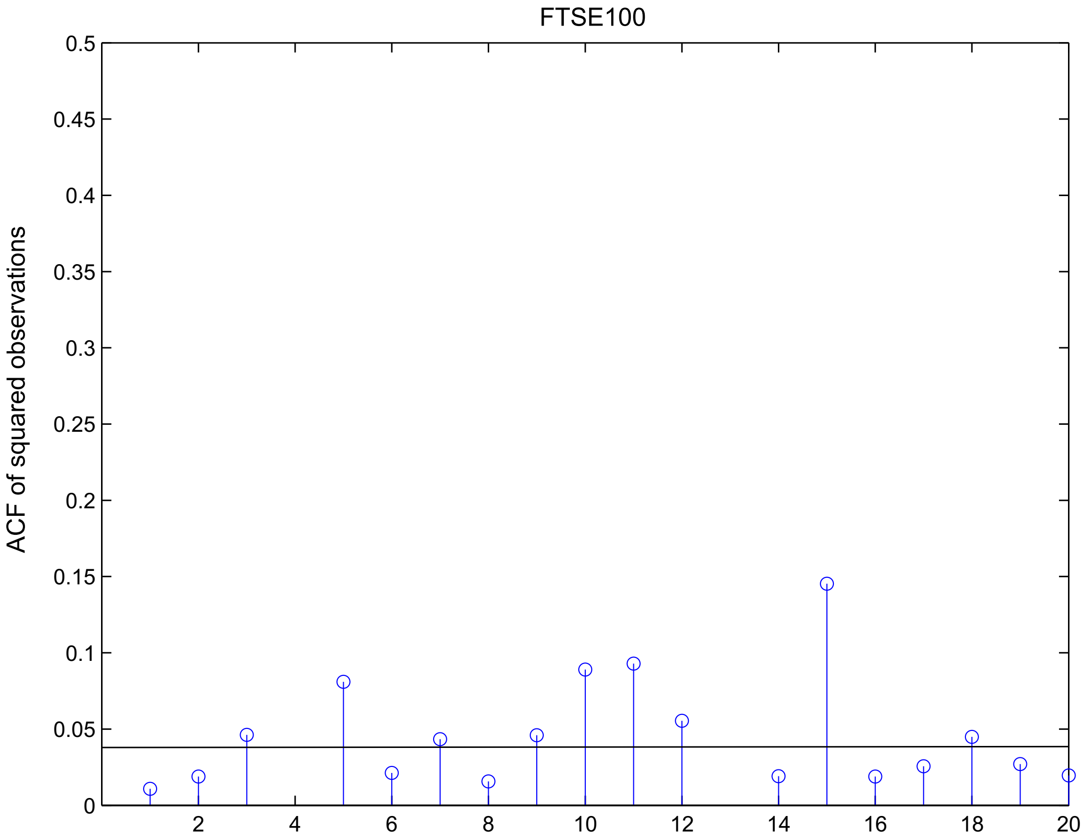

Fig. 3. Autocorrelation function of squared observations y_t^2.

reasonable that modeling their time varying conditional variance using GARCH.

Table 3 gives the descriptive statistics of standardized observations $\hat{\varepsilon}_t$ forecasted by Gaussian kernel. It is not able to reinforce the conclusion of simulation experiment that there is still the excess kurtosis present in the standardized observations. This may be explained by the fact that simulation experiment is simple and explicit to drive out interference of some unknown factors which may affect the results of experiment on real world data sets. But the squared returns show a reduction of the standard Box Ljung statistics compared to the values given in Table 2. This means that its autocorrelations are not significant

Table 3
Descriptive statistics of the standardized observations $\hat{\varepsilon}_t$ forecasted by Gaussian kernel for the test set

	DAXINDX	FRCAC40	FTSE100	JAPDOWA	SPCOMP
Skewness	−0.39592	0.01965	−0.28084	0.24417	−0.28832
Kurtosis	5.33291	4.15950	3.92138	9.81959	6.25969
r(1)	0.0036	−0.0146	0.0234	0.0618	0.0491
Q(20)	16.9932	14.3174	20.1930	20.6503	27.1804
*r*2(1)	−0.0619	−0.0743	−0.0868	−0.0003	−0.0864
*Q*2(20)	16.3267	28.3718	43.5108	7.9827	26.0096

r(1), autocorrelation of order 1 of the original observations $\hat{\varepsilon}_t$; *r*2(1),autocorrelation of order 1 of the squared observations $\hat{\varepsilon}_t^2$; *Q*(20), Box–Ljung statistic for $\hat{\varepsilon}_t$ (31.4 is the 5% critical value); *Q*2(20), Box–Ljung statistic for $\hat{\varepsilon}_t^2$ (31.4 is the 5% critical value).

Table 4
Descriptive statistics of the standardized observations $\hat{\varepsilon}_t$ forecasted by wavelet kernel for the test set

	DAXINDX	FRCAC40	FTSE100	JAPDOWA	SPCOMP
Skewness	−0.38127	0.01891	−0.32129	0.38146	−0.23904
Kurtosis	5.30953	4.17155	3.90672	9.82767	5.95657
r(1)	0.0026	−0.0138	0.0207	0.0547	0.0420
Q(20)	17.1629	14.2383	19.1633	19.2300	27.0344
*r*2(1)	−0.0599	−0.0716	−0.0711	−0.0072	−0.0865
*Q*2(20)	16.7309	27.8910	49.2659	7.8744	27.5811

r(1), autocorrelation of order 1 of the original observations $\hat{\varepsilon}_t$; *r*2(1),autocorrelation of order 1 of the squared observations $\hat{\varepsilon}_t^2$; *Q*(20), Box–Ljung statistic for $\hat{\varepsilon}_t$ (31.4 is the 5% critical value);Q2(20), Box–Ljung statistic for $\hat{\varepsilon}_t^2$ (31.4 is the 5% critical value).

Table 5
Results on the training set

Stock indices	Gaussian kernel			Wavelet kernel		
	NMSE	NMAE	HR	NMSE	NMAE	HR
DAXINDX	0.7589	0.8493	0.6784	0.7529	0.8458	0.6807
FRCAC40	0.7524	0.8080	0.7133	0.7318	0.7942	0.7133
FTSE100	0.7412	0.7678	0.7044	0.7106	0.7467	0.7152
JAPDOWA	0.7067	0.8203	0.7081	0.6919	0.8055	0.7057
SPCOMP	0.7628	0.8819	0.6853	0.7369	0.8535	0.6922
t-value	$4.5542 > t_{0.1,4} = 1.5332$					

Table 6
Results on the test set

Stock indices	Gaussian kernel			Wavelet kernel		
	NMSE	NMAE	HR	NMSE	NMAE	HR
DAXINDX	0.7627	0.8284	0.7179	0.7614	0.8290	0.7298
FRCAC40	0.7292	0.7958	0.7239	0.7268	0.7939	0.7356
FTSE100	0.6690	0.7284	0.7211	0.6703	0.7243	0.7509
JAPDOWA	0.7949	0.8891	0.6936	0.7603	0.8740	0.6842
SPCOMP	0.7476	0.8436	0.7077	0.7014	0.8114	0.7077
t-value	$1.6825 > t_{0.1,4} = 1.5332$					

any more. Table 4 reports the descriptive statistics of standardized observations $\hat{\varepsilon}_t$ forecasted by wavelet kernel. The same conclusion can also be drawn by analyzing it. Furthermore, by comparing the results of the two tables, it can be seen that in terms of descriptive statistics of standardized observations, the results are comparable between the two kernels. We must point out that both Gaussian kernel and wavelet kernel in SVM combined with GARCH seem able to describe enough the dynamics of the squared returns series.

The prediction performance is evaluated using the following statistical metrics: normalized mean squared error (NMSE), normalized mean absolute error (NMAE) and the hit rate (HR). These metrics are calculated as follows:

$$\mathrm{NMSE} = \sqrt{\sum_{t=1}^{N}(\hat{\sigma}_t^2 - y_t^2)^2 \Big/ \sum_{t=1}^{N}(y_{t-1}^2 - y_t^2)^2} \tag{16}$$

$$\mathrm{NMAE} = \sum_{t=1}^{N}|\hat{\sigma}_t^2 - y_t^2| \Big/ \sum_{t=1}^{N}|y_{t-1}^2 - y_t^2| \tag{17}$$

$$\mathrm{HR} = \frac{1}{N}\sum_{t=1}^{N} q_t, \quad q_t = \begin{cases} 1 & : (\hat{\sigma}_t^2 - y_{t-1}^2)(y_t^2 - y_{t-1}^2) \geqslant 0 \\ 0 & : \text{else} \end{cases} \tag{18}$$

where N represents the total number of data points in the test set. $\hat{\sigma}^2$ denotes the predicted conditional variance. y represents the predicted return. y denotes the actual return. The NMSE relates the mean square error of the predicted volatility $\hat{\sigma}_t^2$ by SVM to the mean square error of the naive model $\hat{\sigma}_t^2 = y_{t-1}^2$. The NMAE is more robust against outliers in comparison with NMSE. They are the measures of the deviation between the actual and predicted values. The smaller the values of them, the closer are the predicted time series values to the actual values. On the contrary, the larger the value of HR as a measure of how often the model gives the correct direction of change of volatility, the better is the performance of prediction.

The results on the training set are listed in Table 5. It can be observed that in all the daily indices, the smaller values

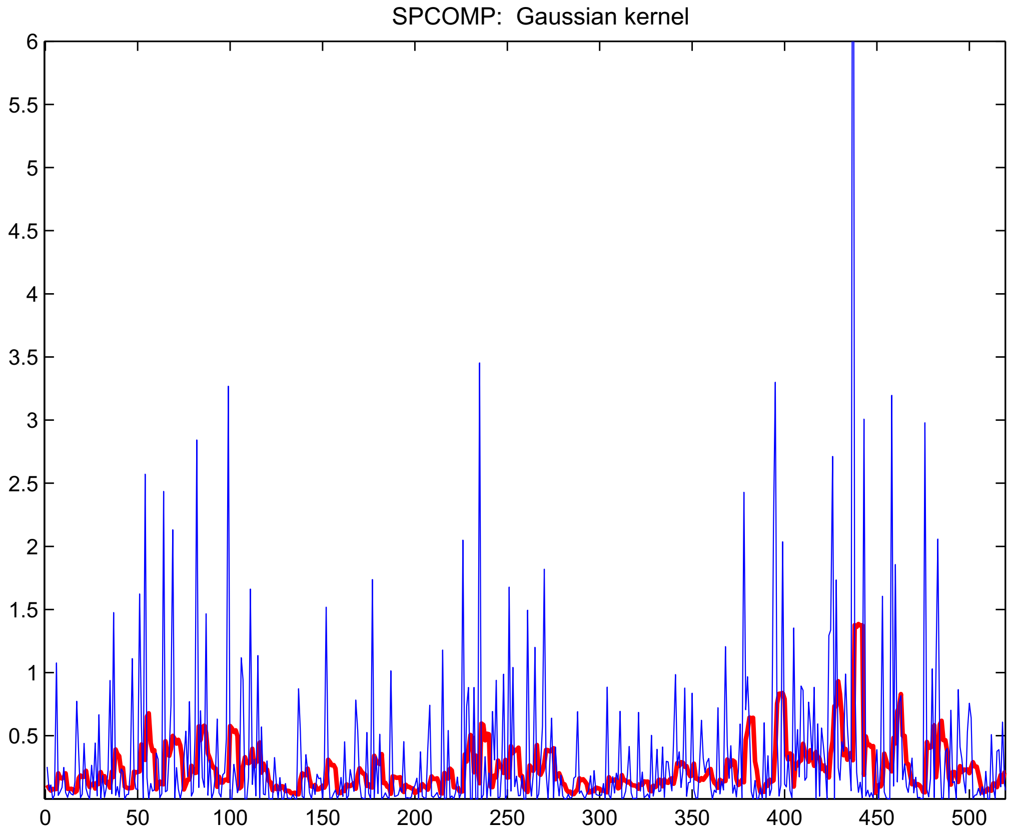

Fig. 4. Squared observations y_t^2 in blue and forecasted volatility $\hat{\sigma}_t^2$ in red by Gaussian kernel. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Please cite this article in press as: Tang, L. -B. et al., Forecasting volatility based on wavelet support vector machine, *Expert Systems with Applications* (2008), doi:10.1016/j.eswa.2008.01.047

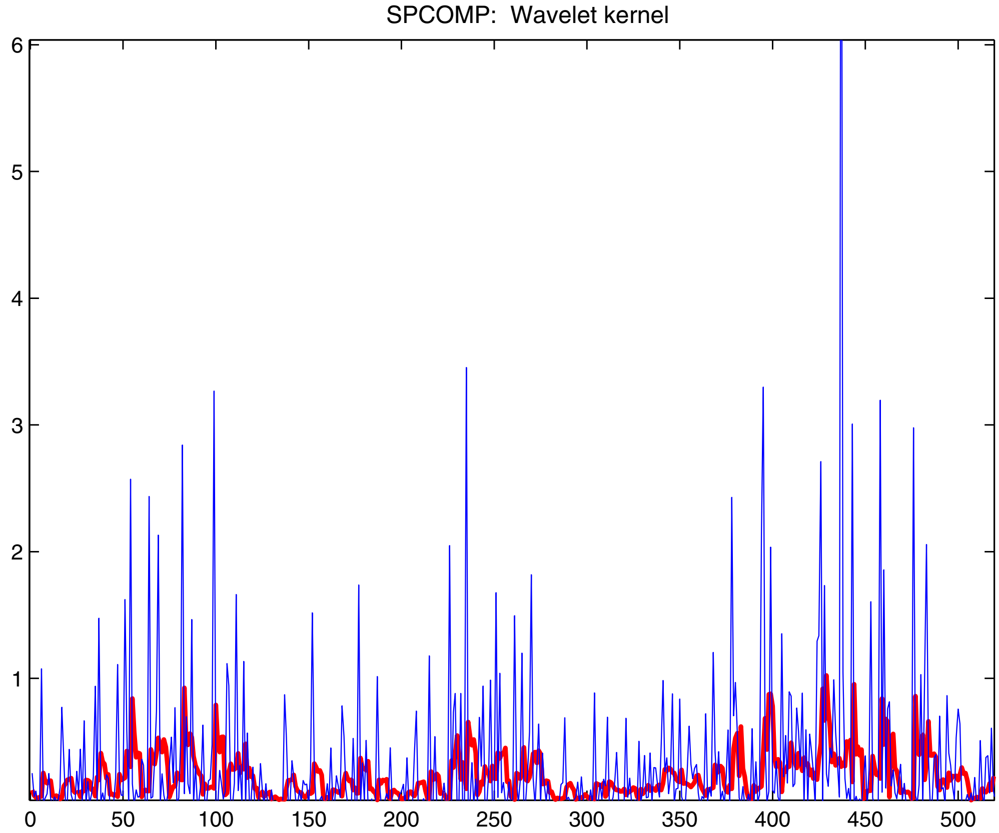

Fig. 5. Squared observations y_t^2 in blue and forecasted volatility $\hat{\sigma}_t^2$ in red by wavelet kernel. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

of NMSE and NMAE are in wavelet kernel. As for HR, with the exclusion of JAPDOWA, in all other daily indices (DAXINDX, FRCAC40, FTSE100, and SPCOMP) the larger values are in wavelet kernel. A paired-test (Montgomery & Runger, 1999) is performed to determine if there is significant difference between the two kernels based on the NMSE of the training set. The calculated t-value indicates that wavelet kernel outperforms Gaussian kernel with 10% significance level for a one-tailed test.

The results on the test set in Table 6 provide a better basis for a comparison of the two kernels where over-fitting issues may be neglected. As expected, the results of the test set are worse than those of the training set in terms of NMSE, NMAE and HR. But the similar conclusion can still be achieved. The table illustrates that apart from NMSE of FTSE100, NMAE of DAXINDX and HR of JAPDOWA, as a whole the smaller values of NMSE and NMAE are founded in wavelet kernel and the larger values of HR occurred in wavelet kernel, too. And a paired-test on the NMSE of the test set also shows that wavelet kernel outperforms Gaussian kernel with 10% significance level for a one-tailed test.

The squared observations y_t^2 and the predicted values of $\hat{\sigma}_t^2$ from both two kernels for the test sets are illustrated in Figs. 4 and 5, where only the SPCOMP is drawn, since it has well representative values of NMSE and NMAE in all daily indices. In this investigation, the parameter C and ε are, respectively, chosen to be 3.3464 and 0.0097. The Gaussian kernel parameter γ is fixed at 0.43907. Daubechies wavelet with 4 vanishing moments is applied to WSVM with the dilation j setting to 1. It is clear that both Gaussian and wavelet kernel are enough to grasp the features reflected by the naive model. The predictions made by both kernels are very similar according to Figs. 4 and 5, although, as we can see in Table 6, performance of wavelet kernel is better than that of Gaussian kernel.

5. Conclusion

An effective wavelet kernel which we combine the wavelet theory with SVM to construct for volatility prediction is presented in this paper. The existence of wavelet kernel is proven firstly. And then the forecasting performance of wavelet kernel is evaluated by using two simulated data sets and five real daily indices. As demonstrated in the experiment, though the explanation for the excess kurtosis present in the standardized observation is not satisfied, wavelet kernel forecasts significantly better than Gaussian kernel in all other aspects. The superior performance of wavelet kernel to the Gaussian kernel mostly lies in that wavelet function is a set of bases that can approximate arbitrary functions. Future work will involve a theoretic analysis of the multiscale frame. More sophisticated wavelet kernel which can closely follow the volatility cluster will be explored for further improving the performance of WSVM in volatility prediction.

Acknowledgements

We gratefully acknowledge the many helpful discussions and suggestions we have had with Prof. Dr. Kuno Egle in Department of Econometrics, Statistics and Mathematical Finance, School of Economics and Business Engineering, University of Karlsruhe. We also wish to thank the generous support provided by Baden-Württenberg state government in Germany.

References

Bishop, C. M. (1996). *Neural networks for pattern recognition*. Oxford: Clarendon Press.

Bollerslev, T. (1986). A generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics, 31*, 307–327.

Daubechies, I. (1990). The wavelet transform: Time–frequency localization and signal analysis. *IEEE Transactions on Information Theory, 36*(5), 961–1005.

Daubechies, I. (1992). *Ten lectures on wavelets*. Philadelphia: SIAM.

Day, T. E., & Lewis, C. M. (1988). The behavior of the volatility implicit in the prices of stock in dex options. *Journal of Financial Economics, 22*, 103–122.

Donaldson, R. G., & Kamstra, M. (1997). An artificial neural network-GARCH model for international stock return volatility. *Journal of Empirical Evidence, 4*(1), 17–46.

Engle, R. F. (1982). Autoregressive conditional heteroskedasticity with estimates of the variance of UK inflation. *Econometrica, 50*, 987–1008.

Franses, P. H., & van Dijk, D. (1996). Forecasting stock market volatility using (non-linear) GARCH models. *Journal of Forecasting, 15*, 229–235.

Harvey, C. R., & Whaley, R. E. (1991). S&P100 index option volatility. *Journal of Finance, 46*, 1551–1561.

Haykin, S. (1994). *Neural networks, a comprehensive foundation*. New York: Macmillan College Publishing Company Inc..

Hull, J., & White, A. (1987). The pricing of options on assets with stochastic volatilities. *Journal of Finance, 42*, 281–300.

Kaastra, I., & Boyd, M. (1996). Designing a neural network for forecasting financial and economic time series. *Neurocomputing, 10*, 215–236.

Li, W. K., & Mak, T. K. (1994). On the squared residual autocorrelations in non-linear time series with conditional heteroskedasticity. *Journal of Time Series Analysis, 15*, 627–636.

Mallat, S. (1989). A theory for muliresolution signal decomposition: The wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 11*, 674–693.

Mallat, S. (1998). *A wavelet tour of signal processing*. Boston: Academic Press.

Meissner, G., & Kawano, N. (2001). Capturing the volatility smile of options on high-tech stocks-a combined GARCH-neural network approach. *Journal of Economics and Finance, 25*(3), 276–293 (Fall).

Montgomery, D. C., & Runger, G. C. (1999). *Applied statistics and probability for engineers*. New York: Wiley.

Perez-Cruz, F., Afonso-Rodriguez, J. A., & Giner, J. (2003). Estimating GARCH models using support vector machines. *Quantitative Finance, 3*(3), 163–172.

Poterba, J. M., & Summers, L. H. (1986). The persistence of volatility and stock market fluctuations. *American Economic Review, 76*, 1142–1151.

Schölkopf, B., Burges, C. J. C., & Smola, A. J. (1999). *Advances in kernel methods*. London: The MIT Press.

Smola, A. J. & Schölkopf, B. A. (1998). Tutorial on support vector regression. NeuroCOLT Technical Report NC-TR-98-030, Royal Holloway College, London, UK.

Vapnik, V. N. (1995). *The nature of statistical learning theory*. New York: Springer.

Vapnik, V. N. (1998). *Statistical learning theory*. New York: Wiley.

Please cite this article in press as: Tang, L. -B. et al., Forecasting volatility based on wavelet support vector machine, *Expert Systems with Applications* (2008), doi:10.1016/j.eswa.2008.01.047