QUANTITATIVE FINANCE VOLUME 3 (2003) 1–10 RESEARCH PAPER
INSTITUTE OF PHYSICS PUBLISHING quant.iop.org

Estimating GARCH models using support vector machines*

Fernando Pérez-Cruz[1], Julio A Afonso-Rodríguez[2] and Javier Giner[3]

[1] Department of Signal Theory and Communications, University Carlos III, Leganés, 28911 Madrid, Spain
[2] Department of Institutional Economics, Economic Statistics and Econometrics, University of La Laguna, 38071 Tenerife, Canary Islands, Spain
[3] Department of Financial Economy and Accounting, University of La Laguna, 38071 Tenerife, Canary Islands, Spain

E-mail: fernando@tsc.uc3m.es, jafonsor@ull.es and jginer@ull.es

Received 17 February 2002, in final form 20 February 2003
Published
Online at stacks.iop.org/Quant/3

Abstract

Support vector machines (SVMs) are a new nonparametric tool for regression estimation. We will use this tool to estimate the parameters of a GARCH model for predicting the conditional volatility of stock market returns. GARCH models are usually estimated using maximum likelihood (ML) procedures, assuming that the data are normally distributed. In this paper, we will show that GARCH models can be estimated using SVMs and that such estimates have a higher predicting ability than those obtained via common ML methods.

1. Introduction

Financial returns series are mainly characterized by having a zero mean, exhibiting high kurtosis and little, if any, correlation. The squares of these returns often present high correlation and persistence, which makes ARCH-type models suitable for estimating the conditional volatility of such processes; see Engle (1982) for the seminal work, Bollerslev *et al* (1994) for a survey on volatility models and Engle and Patton (2001) for several extensions. The ARCH parameters are usually estimated using maximum likelihood (ML) procedures that are optimal when the data is drawn from a Gaussian distribution.

Support vector machines (SVMs) are state-of-the-art tools for linear and nonlinear input–output knowledge discovery (Vapnik 1998, Schölkopf and Smola 2001). SVMs can be employed for solving pattern recognition and regression estimation problems. SVMs have been developed in the machine learning community and resemble, in some ways, a neural network (NN). But SVMs are superior to most common NNs (such as multi-layered perceptron or radial basis function networks) due to the SVM optimization procedure giving not only the weights of the network but also its architecture. Furthermore, one of the most desirable properties when using a SVM is that its optimizing functional is quadratic and linearly restricted, meaning that it only presents a single minimum without any local undesirable solutions.

In this paper we propose using a SVM instead of a ML method to estimate GARCH parameters. The benefits of the SVM in regression (also known as a support vector regressor; SVR) lies in not assuming that there is a probability density function (pdf) over the return series and it adjusts the parameters relying on the empirical risk minimization inductive principle (Vapnik 1998). The SVR defines an insensitivity zone (detailed in section 2) that means it can deal with any pdf. Therefore, if the variable to be estimated is not

* Paper presented at Applications of Physics in Financial Analysis (APFA) 3, 5–7 December 2001, Museum of London, UK.

  PII: S1469-7688(03)33983-1

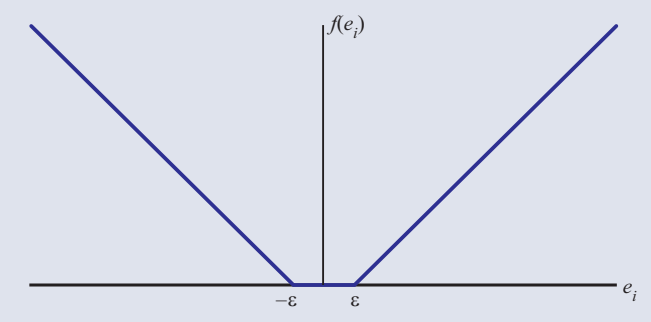

Figure 1. The cost function associated with SVR errors.

sampled from a Gaussian distribution, the SVR can actually lead to better predictions than those obtained using a least squares (ML) approach.

The rest of the paper is outlined as follows. We will introduce the SVM for regression estimation in section 2. Section 3 is devoted to the GARCH model and its estimation via the SVR and ML procedures, including a simulation experiment. In section 4 we consider the estimation of a GARCH (1, 1) model with six different financial data sets, along with the usual statistics employed in time-series analysis to assess the adequacy of the model when using both estimation methods. We compare the ML and SVR forecasts within the GARCH model in section 5 using the same financial series as in section 4. The paper ends in section 6 with a discussion and suggestion for further work.

2. The SVMs

The SVR needs to work with a training set to adjust its parameters and afterwards the machine can be used to predict any possible outcome. The prediction over the samples used for training purposes is known as *in-sample* prediction. For the samples the algorithm did not use out-sample prediction, also known, respectively, as the training and generalization (test) sets in the machine-learning community. The SVR needs a labelled training data set ($\boldsymbol{x}_i \in \Re^d$ and $y_i \in \Re$, for $i = 1, \ldots, n$, where $\boldsymbol{x}_i$ is the input vector and y_i is its corresponding label)[4], to solve the regression estimation problem ($y_i = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b$, where $\boldsymbol{w}$ and b define the linear regressor), i.e. to find the values of $\boldsymbol{w}$ and b. The SVR uses the penalty function shown in figure 1, in which the samples with a prediction error ($e_i = y_i - \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i - b$) lower than ε in absolute terms are not penalized and those samples with a prediction error greater than ε are linearly penalized.

The regression SVM is an extension of the SVM for classification proposed in 1992 (Boser *et al* 1992). The SVR cost function has been proposed to follow the structural risk minimization (SRM) principle (Vapnik 1998). The SRM principle is an upper bound in the prediction error. It depends on the empirical error and on a factor that measures the complexity of the used regressor. The SRM defines a nested set of possible functions that approximates the given series and chooses the best one according to the one that provides the minimum SRM; more details can be found in Vapnik (1998), Schölkopf and Smola (2001). The SRM principle needs a nonzero insensitivity zone in order to provide an upper bound of the error less than infinity, but this ε can be made arbitrarily small. The samples that present a prediction error greater than ε are linearly penalized. This decision is based on a result by Huber (1964), in which he demonstrated that the best cost function over the worst model over any probability density function of y given $\boldsymbol{x}(p(y/\boldsymbol{x}))$ is the linear cost function (absolute value penalization of the error). Therefore, if the pdf $p(y/\boldsymbol{x})$ is unknown the best cost function is the linear penalization over the errors. The mixture of these two results gives rise to the SVM cost function, which warrantees that the SRM is finite and the best solution when $p(y/\boldsymbol{x})$ is unknown. Most regression estimation problems use a quadratic loss function, $f(e_i) = e_i^2$ (least squares), but for those problems in which y_i has not been drawn from a Gaussian distribution, the least square techniques are suboptimal (Vapnik 1982), and can lead to severely mismatched solutions for some densities. Furthermore, the value of ε can be optimally set if the probability density function over e_i is known, as shown in Smola *et al* (1998).

2.1. SVR optimization

The SVR is stated as a constrained optimization problem

$$\min_{\boldsymbol{w},b,\xi_i,\xi_i^*} \left\{ \tfrac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \right\} \tag{2.1}$$

subject to

$$y_i - \boldsymbol{w}^{\mathrm{T}}\boldsymbol{\phi}(\boldsymbol{x}_i) - b \leqslant \varepsilon + \xi_i \tag{2.2}$$

$$\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\phi}(\boldsymbol{x}_i) + b - y_i \leqslant \varepsilon + \xi_i^* \tag{2.3}$$

$$\xi_i, \xi_i^* \geqslant 0 \tag{2.4}$$

where $\boldsymbol{\phi}(\cdot)$ is a nonlinear transformation to a higher dimensional space ($\boldsymbol{x}_i \in \Re^d \rightarrow \boldsymbol{\phi}(\boldsymbol{x}_i) \in \Re^H, d \leqslant H$). The SVR defines a linear regressor in the transformed space ($\Re^H$), also known as feature space, which is nonlinear in the input space, unless $\boldsymbol{\phi}(\boldsymbol{x}_i) = \boldsymbol{x}_i$ (linear regression). ξ_i and ξ_i^* are positive slack variables, introduced to deal with the samples with prediction errors greater than ε. The parameter C is the penalization applied over the samples with prediction error greater than ε. This problem is usually solved introducing constraints (2.2)–(2.4) using Lagrange multipliers, leading to the minimization of

$$\begin{aligned} L_P = {} & \tfrac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{n}\xi_i + \xi_i^* \\ & - \sum_{i=1}^{n}\alpha_i(\varepsilon + \xi_i - y_i + \boldsymbol{w}^{\mathrm{T}}\boldsymbol{\phi}(\boldsymbol{x}_i) + b) - \sum_{i=1}^{n}\mu_i\xi_i \\ & - \sum_{i=1}^{n}\alpha_i^*(\varepsilon + \xi_i^* + y_i - \boldsymbol{w}^{\mathrm{T}}\boldsymbol{\phi}(\boldsymbol{x}_i) - b) - \sum_{i=1}^{n}\mu_i^*\xi_i^* \end{aligned} \tag{2.5}$$

[4] For compactness we will use matrix notation. The vectors will be column vectors denoted by bold lower case, the matrices will be bold upper case and the scalars will be italic (lower case and occasionally upper case). The dot product between the columns vectors will be shown as matrix multiplication ($\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}$), where $^{\mathrm{T}}$ indicates the matrix transpose operation.

with respect to $\boldsymbol{w}$, b, ξ_i and ξ_i^* and its maximization with respect to the Lagrange multipliers, α_i, α_i^*, μ_i and μ_i^*. To solve this problem we need to compute the Karush–Kunh–Tucker (KKT) conditions (Fletcher 1987), that states some conditions over the variables in (2.5) to be an admissible solution set, being (2.2)–(2.4),

$$\frac{\partial L_P}{\partial \boldsymbol{w}} = \boldsymbol{w} - \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)\boldsymbol{\phi}(\boldsymbol{x}_i) = 0 \tag{2.6}$$

$$\frac{\partial L_P}{\partial b} = \sum_{i=1}^{n}\alpha_i - \alpha_i^* = 0 \tag{2.7}$$

$$\frac{\partial L_P}{\partial \xi_i} = C - \alpha_i - \mu_i = 0 \tag{2.8}$$

$$\frac{\partial L_P}{\partial \xi_i^*} = C - \alpha_i^* - \mu_i^* = 0 \tag{2.9}$$

$$\alpha_i, \alpha_i^*, \mu_i, \mu_i^* \geqslant 0 \tag{2.10}$$

$$\alpha_i\{\varepsilon + \xi_i - y_i + \boldsymbol{w}^{\mathrm{T}}\boldsymbol{\phi}(\boldsymbol{x}_i) + b\} = 0 \tag{2.11}$$

$$\alpha_i^*\{\varepsilon + \xi_i^* - \boldsymbol{w}^{\mathrm{T}}\boldsymbol{\phi}(\boldsymbol{x}_i) - b + y_i\} = 0 \tag{2.12}$$

$$\mu_i\xi_i = 0 \qquad \text{and} \qquad \mu_i^*\xi_i^* = 0. \tag{2.13}$$

The usual procedure to optimize the SVR introduces the KKT conditions (2.6)–(2.9) into (2.5), leading to the maximization of

$$L_d = \varepsilon\sum_{i=1}^{n}(\alpha_i + \alpha_i^*) - \sum_{i=1}^{n}\sum_{j=1}^{n}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)\boldsymbol{\phi}^{\mathrm{T}}(\boldsymbol{x}_i)\boldsymbol{\phi}(\boldsymbol{x}_j) \tag{2.14}$$

subject to (2.7) and $0 \leqslant \alpha_i, \alpha_i^* \leqslant C$. This procedure can be solved using quadratic programming (QP) schemes (Schölkopf and Smola 2001). To solve (2.14) we do not need to know the nonlinear mapping $\boldsymbol{\phi}(\cdot)$, only its reproducing kernel in Hilbert Space (RKHS) $\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \boldsymbol{\phi}^{\mathrm{T}}(\boldsymbol{x}_i)\boldsymbol{\phi}(\boldsymbol{x}_j)$ (Schölkopf and Smola 2001).

One of the most difficult tasks when solving the SVR is choosing the appropriate ε, because it dramatically depends with the given data, making very hard to choose *a priori* a good value. There is an alternative formulation, known as ν-SVR (Schölkopf *et al* 2000), which yields the same solution as the SVR, also known as ε-SVR, in which the ε is replaced by ν. The value of ν roughly determines the fraction of support vectors, therefore it gives the complexity of the machine and it is bounded between 0 and 1. Furthermore, the solution is not very sensitive to this ν and for a wide range of them the obtained ε is close to the optimal one. Details about the ν-SVR formulation, in addition to the proof that ν roughly represents the fraction of support vectors and the fact that the solution is not very sensitive to the values ν can be found in Schölkopf *et al* (2000).

The SVR can also be solved by relying on an iterative re-weighted least squares (IRWLS) procedure, that is easier to implement and it is much faster than the usual QP schemes, as shown in Pérez-Cruz *et al* (2000) for the regular SVR and in Pérez-Cruz and Artés-Rodr ıguez (2001) for the ν-SVR, which is what we will use throughout this paper.

Table 1. Some of the kernels used in the SVR.

Linear	$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_j$
Polynomial	$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_j + 1)^k$
RBF	$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp(-\|\|\boldsymbol{x}_i - \boldsymbol{x}_j\|\|^2/(2\sigma^2))$

Some of the most widely used kernels are shown in table 1, where k is a natural number and σ is a real positive number. We must recall that the Mercer theorem (Schölkopf and Smola 2001) states the necessary and sufficient conditions for any function $\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j)$ to be a kernel in a Hilbert space.

3. Estimating GARCH models using SVR

3.1. The GARCH (1, 1) model

The GARCH (1, 1) model provides a simple representation of the main statistical characteristics of a return series for a wide range of assets and, consequently, it is extensively used to model real financial time series. It serves as a natural benchmark for the forecast performance of heterocedastic models based on ARCH. In the simplest set up, if y_t follows a GARCH (1, 1) model, then

$$\begin{aligned} y_t &= \mu + \sigma_t\varepsilon_t \\ \sigma_t^2 &= \omega + \alpha y_{t-1}^2 + \beta\sigma_{t-1}^2 \end{aligned} \tag{3.1}$$

where ε_t is an uncorrelated process with zero mean and unit variance. Although the mean of a financial return series could not be zero it can be usually neglected ($\mu \approx 0$), without significantly degrading the performance of the proposed model. Following the definition in (3.1), the conditional variance σ_t^2 is a stochastic process assumed to be a constant plus a weighted average of the last period's forecast, σ_{t-1}^2, and the last period's squared observation, y_{t-1}^2. The parameters ω, α and β must satisfy $\omega > 0$, $\alpha, \beta \geqslant 0$ to ensure that the conditional variance is positive. The parameter ω has to be strictly positive for the process y_t not to degenerate. The process y_t is stationary iff $\alpha + \beta < 1$.

This model is usually estimated using the (conditionally) Gaussian log-likelihood function and maximizing it through an iterative algorithm such as BHHH (Berndt *et al* 1974), because the function to be maximized is nonlinear in its arguments. The estimates are called ML when the Gaussian distribution is the underlying pdf that the data has been sampled from, if this is not the case it is called quasi-ML. Bollerslev and Wooldridge (1992) showed the consistency of these estimates in this case, which does not ensure that for a finite sample set it is the best estimate.

The SVR can also be used to estimate this model. The selection of ν by cross validation will help to adjust the GARCH parameters to the optimal pdf described by the return series without knowing it beforehand. If the pdf is not Gaussian, the SVR will probably improve the results attained using ML procedures.

The SVR needs to deal with observable variables in the model to be estimated. In equation (3.1) both σ_t (as the

dependent variable) and $\boldsymbol{x}_t = [y_{t-1}\sigma_{t-1}]^{\mathrm{T}}$ (as the regressor vector) must be known in order to set the weights (α_i) in equation (2.14). With the time series being persistent, we decided to measure σ_t as a moving average of the contemporaneous and four lagged squared returns around each time point in the in-sample set, that is

Q.2

$$\hat{\sigma}_t^2 = \tfrac{1}{5}\sum_{k=0}^{4} y_{t-k}^2 \tag{3.2}$$

and we can use now the series of y_t and σ_t to train the SVM. The value of C has been set to 10 as a compromise value between the regularization of $\boldsymbol{w}$ and the weight of the errors (the solution is not very sensitive to this parameter). The value of ν has been set using eightfold cross validation (eightfold CV; Bishop 1995). We have divided the in-sample set in eight disjoint sets and have used seven of them to train several machines with different values of ν. The eighth set has been used to compute the validation error of each machine for each specific value of ν. This process is repeated seven times, leaving aside in each repetition one of the sets for computing the validation error. Finally, all the validation errors for the same ν value are added together and the best ν value is the one pointed out by the minimum over the validation error. This ν value is the one used to train the machine with the full in-sample set. Further details on how to compute hyper-parameters in machine learning can be found in Bishop (1995). The software that has been used for solving the SVR has been placed on the http://www.tsc.uc3m.es/~fernando web page.

3.2. A simulation experiment

As an illustration of how the SVM works with GARCH-type models we have performed a simulation study. We have generated time series that follow the model in (3.1) setting $\mu = 0$, $\omega = 0.1$, $\alpha = 0.1$ and $\beta = 0.8$ and a disturbance term ε_t distributed first as Gaussian and then as a Student's t with six degrees of freedom (kurtosis = 6). This second distribution tries to model the excess of kurtosis that appears in real financial series.

We have drawn 2000 samples for each model and have used the first 1000 samples to train the model, i.e. to obtain the parameter estimates using ML and SVR methods, and have used the remaining 1000 samples to assess the quality of the obtained model. To measure the predicting ability of both estimates, we have used R^2 statistics, as described in detail in section 5. Basically, the larger the R^2 value is, the better the model predicts real data. We have reported the obtained results for the Gaussian and Student's t error distributions, respectively, in tables 2 and 3, with ν taking all possible values and $C = 10$. The reported results are the mean values of 50 independent trials.

In table 2, we report the R^2 statistics for the in-sample period and the out-sample period. The best result over the out-sample period, in which the predicting ability can be best measured (the in-sample estimate can be over-fitted), is obtained for the ML estimate because for Gaussian data the ML procedure gives the best estimate. But one notices that the best result for the SVM is not far from the ML estimate and the solution is very stable for a wide range of possible ν values; recall that the values of ν are limited to between 0 and 1.

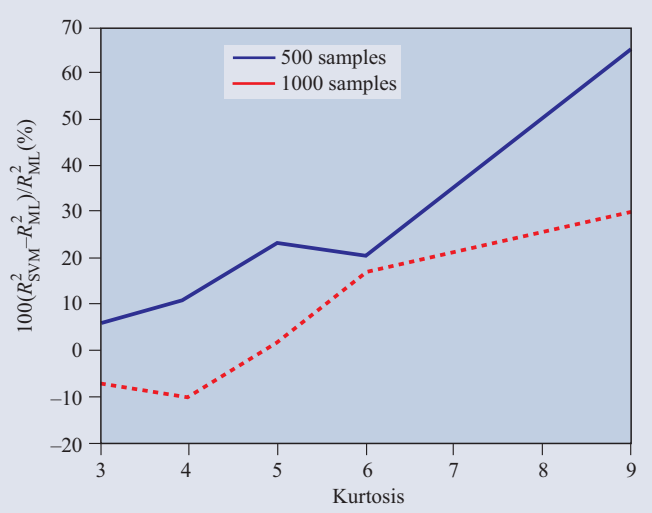

Figure 2. A comparison of the average R^2 statistic for simulated data with a different number of samples and kurtosis levels for ε_t.

In table 3, the best result is obtained for the SVM with $\nu = 0.1$. Although the ML procedure gives consistent estimates, the SVM can give better predictions thanks to its robust cost function, which allows one to obtain good predictions for any distribution of ε_t, and to its regularization term, $\|\boldsymbol{w}\|^2$.

Finally, to assess the behaviour of both schemes with different numbers of samples, we repeated the previous trials with 1000 samples (the first 500 for the in-sample set and the last 500 for the out-sample set) and with 2000 samples (1000 for the in-sample set and 1000 for the out-sample set), using the same values for μ, ω, α and β. We considered ε_t to be distributed as Gaussian and Student's t random variables with ten, seven, six and five degrees of freedom, which, respectively, lead to kurtosis values of 4, 5, 6 and 9. We have set the value of ν for the SVM to be equal to 0.1 and $C = 10$. We have run 50 independent trials for each distribution and number of samples and we have reported the mean values for the out-sample R^2 statistics in figure 2.

In figure 2, one first notices that as the kurtosis gets further from 3 the SVM is a better predictor than the ML procedure for both sample sizes. This occurs because the ML procedure is tuned only for Gaussian distributions whereas the SVM is appropriate for any distribution, which means it gives better results for nonGaussian distributions.

It is also relevant to mention that for the reduced set, the SVM outperforms the ML procedure for every kurtosis value

$$\frac{R^2_{\mathrm{SVM}} - R^2_{\mathrm{ML}}}{R^2_{\mathrm{ML}}} > 0$$

even when ε_t is distributed as a Gaussian random variable. This can be explained by VC theory: when we have very few samples sometimes a simple model is better than the actual model, which allows the SVM with the regularization over $\boldsymbol{w}$ to seek better solutions, even when normality assumptions

Q.3

Table 2. Average R^2 statistic for simulated data: Gaussian distribution.

	SVM									ML
ν	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	
In-sample R^2	4.49	4.31	4.18	3.92	3.94	3.28	3.28	3.80	3.50	5.32
Out-sample R^2	3.25	3.30	3.23	3.11	3.10	2.49	2.56	2.91	2.71	**3.51**

Table 3. Average R^2 statistic for simulated data: Student's $t(6)$ distribution.

	SVM									ML
ν	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	
In-sample R^2	3.34	3.31	2.64	2.58	2.57	2.73	2.60	2.79	2.42	4.13
Out-sample R^2	**2.97**	2.75	2.53	2.42	2.24	2.40	1.96	2.03	2.21	2.54

hold. As the number of samples in the in-sample set increases, the SVM and the ML method will both tend to the same solution because both are consistent, but for reduced sets, the SVM can give a much better result.

4. Empirical modelling

To illustrate the main empirical properties often observed in high-frequency financial time series, table 4 contains descriptive statistics of six financial time series observed daily. Let p_t be the observed daily price at time t and y_t the corresponding daily return defined by

$$y_t = \ln p_t - \ln p_{t-1}.$$

The considered series are returns of four international stock market indexes: the S&P100 index observed from January 1996 until October 2000; the FTSE100 index observed from January 1995 until December 2000; IBEX35 of the Madrid Stock Exchange observed from January 1990 until December 1999; and the NIKKEI index from January 1995 until November 1999. The returns of two stock equities: General Motors and Hewlett Packard were also analysed from January 1996 until October 2000. In table 4, it is possible to observe that all the series show almost zero means and excess kurtosis (always above 3) for the normal distribution value. We must point out that the return series show little or no correlation and its squares show high-correlation coefficients. The analysis of serial correlation levels using standard Box–Ljung statistics, named Q(20), indicates that the S&P100, FTSE100 and IBEX35 are significantly correlated. However, this evidence disappears when the test is corrected for conditional heteroskedasticity as proposed by West and Cho (1995). This corrected test is denoted in table 4 by Q*(20). The difference that can be observed among them seems to reinforce the evidence in favour of modelling their time-varying conditional variance using GARCH schemes.

Furthermore, the analysis of the squared observations shows significant correlation and persistence. Thus, the Box–Ljung statistics corresponding to squared observations (asymptotically equivalent to the test for ARCH effects of Engle (1982)) are greater than their critical values for all series, so we can expect to find more serial dependence on these series when being modelled.

We have plotted in figure 3 the S&P100 and General Motors returns series. It can readily be seen that the volatility concentrates itself in clusters, i.e. periods of high and low volatility can be observed in the data. We have also depicted nonparametric estimates of the pdf of returns together with the corresponding normal density (same mean and variance). The density plots confirm the results reported in table 4 about the returns being heavily tailed. Finally, correlograms of the squared return series, y_t^2, are also reported. The volatility clustering is reflected in the significant correlations of squared returns. The y_t^2 autocorrelation coefficients are larger and last longer (persistent) than those of the y_t series.

Table 5 reports the ML estimates of the parameters of the GARCH (1, 1) model for all the considered returns series. The original series of length N was divided into two sets: the in-sample (training set) with the first $N/2$ observations and the out-sample (test set) with the last $N/2$ observations. In this table it is possible to observe that all the series considered have significant ARCH effects and high persistence measured as $\hat{\alpha} + \hat{\beta}$ (except the HP series that presents $\hat{\alpha} + \hat{\beta} = 0.6712$). We check the model estimation using several statistics based on the standardized observations $\hat{\varepsilon}_t = y_t/\hat{\sigma}_t$, where $\hat{\sigma}_t$ is the estimated volatility from the GARCH model. The standardized observations still have heavy tails, as can be noted when scrutinizing table 5. However, the autocorrelations of squared returns are no longer significant using either the usual Box–Ljung statistic or the corrected test suggested by Li and Mak (1994), in which the correction factors depend on the model specification and the results of the ML estimation. Therefore, the GARCH (1, 1)-ML model seems able to adequately represent the dynamics of the squared returns series considered, although it is not able to explain the excess kurtosis present in the standardized observations.

We show in table 6 the same parameters for the SVR estimation, but we must point out that the parameters are quite different due to the different optimization procedure being used for both schemes. We have also included the value of the parameter ν of the SVR (this parameter was computed using cross validation techniques) as previously discussed. For every series this gave rise to a value that was far from the optimal one if the data came from a Gaussian random variable and we also report the value of ε associated with the optimal ν. The C parameter was set to 10 for all cases. Therefore, the analysis of the standardized observations in table 6 shows that the

Table 4. Descriptive statistics of the daily returns (N, sample size; $r(1)$, autocorrelation of order 1 of the original observations y_t; $r2(k)$, autocorrelation of order k of the squared observations y_t^2; Q(20), Box–Ljung statistic for y_t (31.4 is the 5% critical value); Q*(20), modified Box–Ljung statistic for y_t suggested by West and Cho (1995) (31.4 is the 5% critical value); Q2(20), Box–Ljung statistic for y_t^2 (31.4 is the 5% critical value)).

	S&P100	FTSE100	IBEX35	NIKKEI	GM	HP
N	1220	1480	2009	1070	1220	1220
Mean	7.51E−04	4.78E−04	7.46E−04	−5.11E−05	3.10E−04	8.05E−04
S.D.	0.0120	0.0103	0.0125	0.0154	0.0201	0.0287
Skewness	−0.357	−0.157	−0.360	0.0836	0.0543	−0.00372
Kurtosis	6.51	4.20	6.84	5.63	4.42	6.21
$r(1)$	−0.018	0.057	0.118	−0.042	−0.061	−0.040
Q(20)	32.6	44.9	73.3	22.3	25.0	26.4
Q*(20)	23.9	28.6	29.5	12.7	19.5	22.3
Squared observations y_t^2						
Mean	1.45E−04	1.06E−04	1.57E−04	2.38E−04	4.04E−04	8.25E−04
S.D.	3.38E−04	1.89E−04	3.75E−04	5.13E−04	7.49E−04	1.88E−03
$r2(1)$	0.247	0.130	0.211	0.119	0.112	0.031
$r2(2)$	0.142	0.193	0.197	0.123	0.070	0.032
$r2(5)$	0.131	0.181	0.179	0.107	0.064	−0.015
$r2(10)$	0.042	0.187	0.227	0.016	0.049	0.065
Q2(20)	175.9	770.3	1414.4	161.4	98.9	49.2

Table 5. Estimated GARCH (1, 1) models by ML estimation and diagnostics for the in-sample data set ($r(1)$, autocorrelation of order 1 of the standardized observations $\hat{\varepsilon}_t$; $r2(k)$, autocorrelation of order k of the squared standardized 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); Q2*(20), modified Box–Pierce statistic for $\hat{\varepsilon}_t^2$ suggested by Li and Mak (1994) when the GARCH model is estimated by ML (31.4 is the 5% critical value)).

	S&P100	FTSE100	IBEX35	NIKKEI	GM	HP
ω	6.34E−06	9.90E−08	6.34E−06	1.81E−06	4.07E−05	2.10E−04
α	0.111	0.0220	0.111	0.0318	0.0730	0.132
β	0.828	0.978	0.828	0.958	0.768	0.539
$\alpha+\beta$	0.939	1.000*	0.939	0.990	0.841	0.671
Standardized observations $\hat{\varepsilon}_t = y_t/\hat{\sigma}_t$						
Mean	0.0855	0.0949	0.0327	−0.0015	0.0411	0.0234
S.D.	0.997	0.995	1.00	1.011	0.999	1.000
Skewness	−0.677	−0.251	0.058	0.014	0.258	−0.257
Kurtosis	5.55	4.06	4.29	5.94	3.57	7.27
$r(1)$	0.0671	0.0417	0.1470	−0.0132	−9.81E−04	−0.0449
Q(20)	28.7	19.6	41.3	18.1	31.4	11.7
$r2(1)$	0.0258	−0.0441	0.0106	0.0214	0.0305	−0.0163
$r2(2)$	0.0093	0.0766	−0.0275	0.0248	−0.0193	−0.0107
$r2(5)$	0.0073	−0.0369	−0.0018	0.0415	0.0167	−0.0153
$r2(10)$	−0.0131	−0.0018	0.0426	−0.0072	−0.0414	−0.0181
Q2(20)	8.9	16.8	18.9	6.6	16.0	22.9
Q2*(20)	10.2	18.2	20.4	6.9	15.3	19.2

[a] $1-(\alpha+\beta) = +2.0\text{e}-006$.

remaining errors exhibit relatively high correlation for S&P100 and IBEX35; the squared errors of FTSE100 and IBEX35 also show a high level of correlation when using the standard Box–Ljung test statistic with its asymptotic distribution being based on the normality assumption. We must also point out that the SVM does not assume normality over the return series nor are its errors normally distributed and so the results obtained should not be considered as an erroneous estimate by the SVR, but as an expected consequence of not assuming a given distribution function. The results are almost identical when using the corrected Box–Ljung statistic, Q2*(20), that was also based on the Gaussianity of the conditional distribution, leading to the conclusion that this assumption is far from being valid in our case.

Thus, we encounter a problem when evaluating the model adequacy of the GARCH (1, 1) specification based on the SVR estimates because all the test statistics for model checking are based (even asymptotically) on the Gaussian assumption. Finally, in the next section we will evaluate both methods for estimating the GARCH (1, 1) model in terms of its forecasting ability to replicate the observed series.

5. A comparison of the forecasting ability of the GARCH models

The ability of the GARCH models to provide good estimates of equity and index return volatility is well documented. Many

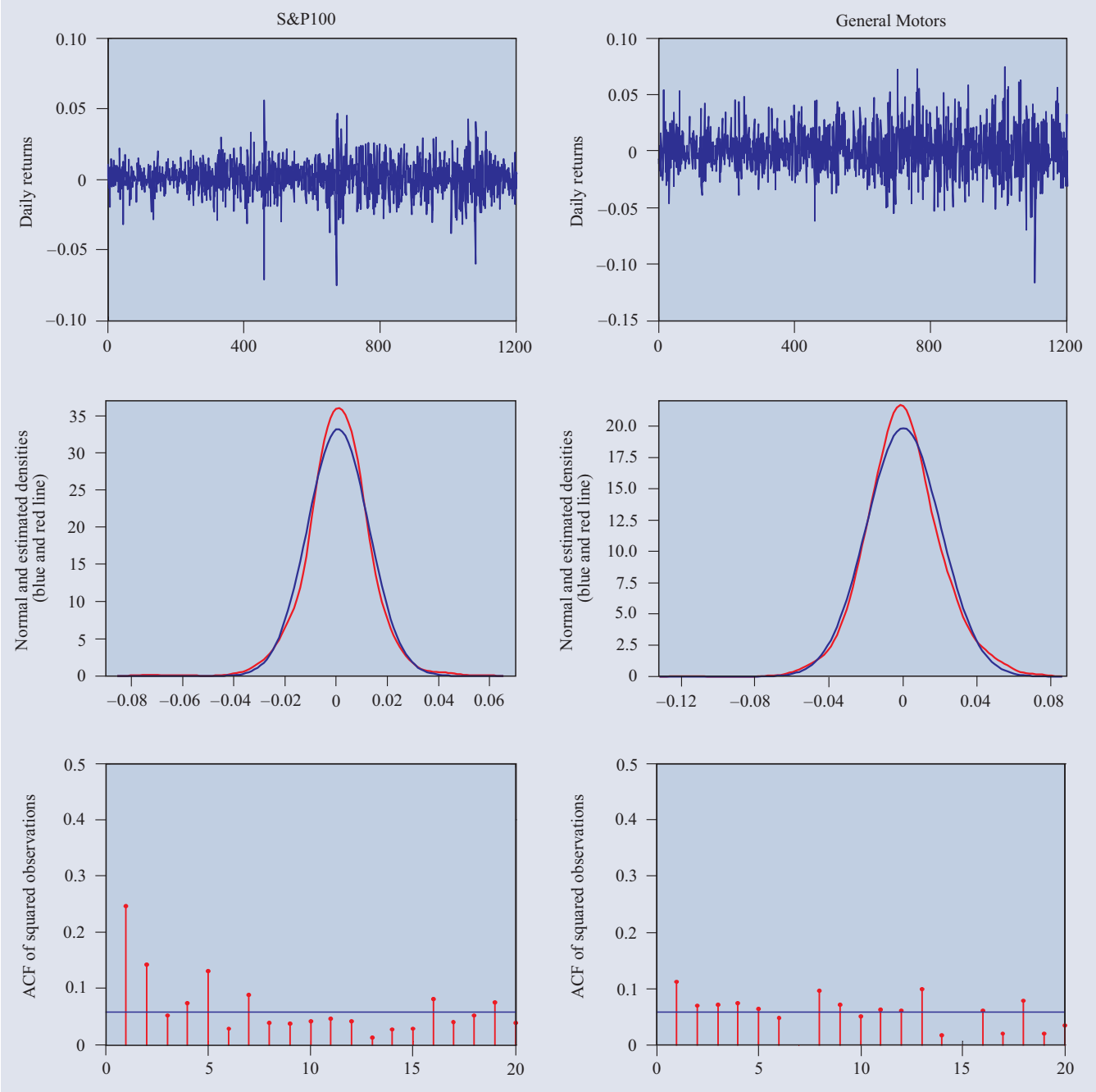

Figure 3. Daily returns and densities of y_t and ACF of y_t^2.

studies show that the parameters of a variety of different GARCH models are highly significant in the sample (see, for example, Bollerslev 1986, 1987, Nelson 1991, Andersen and Bollerslev 1998). There is, however, less evidence that GARCH models provide good forecasts of equity return volatility. Some studies (Franses and Van Dijk 1995, Figlewski 1997) examine the out-sample predictive ability of GARCH models using the daily squared observations as a measure of the realized volatility. All find that a regression of realized volatility on forecast volatility produces a low R^2 statistic, defined below (often less than 10%) and hence the predictive power of the forecasts may be questionable. Recent works are introducing important improvements in the volatility forecasting and testing procedures (Blair *et al* 2001), as we will see in section 5.2, but SVR estimation can be used independently of the initial approach used.

5.1. The daily squared observations as a measure of realized volatility

To forecast the squared daily returns when the underlying process of volatility is GARCH (1, 1) we could employ equation (3.1) so one-period-ahead forecast is simply given by

$$\hat{y}_t^2 = \hat{\omega} + (\hat{\alpha} + \hat{\beta}) y_{t-1}^2.$$

Table 6. Estimated GARCH (1, 1) models by SVR and diagnostics for the in-sample data set ($r(1)$, autocorrelation of order 1 of the standardized observations $\hat{\varepsilon}_t$; $r2(k)$, autocorrelation of order k of the squared standardized 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); Q2*(20), modified Box-Pierce statistic for $\hat{\varepsilon}_t^2$ suggested by Li and Mak (1994) when the GARCH model is estimated using the ML procedure (31.4 is the 5% critical value)).

	S&P100	FTSE100	IBEX35	NIKKEI	GM	HP
ω	1.36E−05	7.43E−06	2.11E−05	$4.03E-05$	3.61E−05	2.47E−05
α	0.102	0.0481	0.0730	0.0478	0.0459	0.0171
β	0.785	0.810	0.774	0.841	0.804	0.924
$\alpha+\beta$	0.886	0.858	0.847	0.888	0.849	0.941
ν	0.150	0.325	0.225	0.100	0.325	0.950
ε	4.61E−05	1.35E−05	3.95E−05	1.19E−04	6.17E−05	2.21E−06
Standardized observations $\hat{\varepsilon}_t = y_t/\hat{\sigma}_t$						
Mean	0.0837	0.0968	0.0274	−9.12E−05	0.0383	0.0214
S.D.	0.939	1.02	0.915	0.777	1.03	1.19
Skewness	−0.668	−0.205	−0.019	0.058	0.241	−0.227
Kurtosis	5.88	4.27	4.01	6.77	3.55	7.15
$r(1)$	0.0604	0.0587	0.1430	−0.0021	0.0006	−0.0399
Q(20)	31.7	21.3	40.2	17.8	30.9	10.2
$r2(1)$	0.0449	−0.0338	0.0390	0.0173	0.0529	0.0261
$r2(2)$	0.0193	0.0690	0.0124	0.0332	−0.0039	0.0018
$r2(5)$	0.0226	−0.0124	0.0399	0.0518	0.0326	−0.0087
$r2(10)$	−0.0006	0.0558	0.0828	0.0124	−0.0398	−0.0259
Q2(20)	9.1	39.8	35.3	10.1	16.5	18.0
Q2*(20)	10.4	63.9	79.7	12.2	23.4	19.2

Given the forecasts, $\hat{y}_t^2$, of the squared returns ,y_t^2, we report the proportion of the sample variation explained by the forecasts with the R^2 statistic (Theil 1971) defined by

$$R^2 = 1 - \frac{\sum_{t=1}^{N}\left(y_t^2 - \hat{y}_t^2\right)^2}{\sum_{t=1}^{N}\left(y_t^2 - \left(\frac{1}{N}\sum_{s=1}^{N} y_t^2\right)\right)^2} = 1 - \frac{\sum_{t=1}^{N}\left(y_t^2 - \hat{y}_t^2\right)^2}{\sum_{t=1}^{N}\left(y_t^2 - \bar{y}_2\right)^2}. \tag{5.1}$$

This relative accuracy statistic indicates that the model accounts for over $100 \times R^2$ per cent of the variability in the observations. For example, $R^2 = 0.11$ means that the model accounts for 11% of the variability in the observations. If $R^2 = 0$, then the model is incapable of extracting the deterministic part of the time series, if there is any. If R^2 is negative this means that the model introduce more variability than the sample mean of the original time series.

Table 7 shows the R^2 values calculated for the six analysed financial returns series using the GARCH model estimated by the SVR and ML schemes.

The SVR is able to explain a higher percentage of all the time series over the out-sample sets except for IBEX35, where the ML method is superior. In addition, the SVR is always able to predict better than the sample mean, which it is not possible for the ML technique, see the HP data set. These results are as expected because the data sets do not resemble a Gaussian distribution and a technique such as the SVR that does not assume normality seems to be able to extract more knowledge from the return series than the usual ML techniques.

We have plotted the predicted values of $\hat{y}_t^2$ by SVR and the squared observations y_t^2 for the S&P100 index for the in-sample and out-sample data sets in figure 4. The prediction made by both methods are very similar, although, as we can see in table 7, the prediction obtained by the SVM explains over a half a percentage more than the explanation made by the ML scheme.

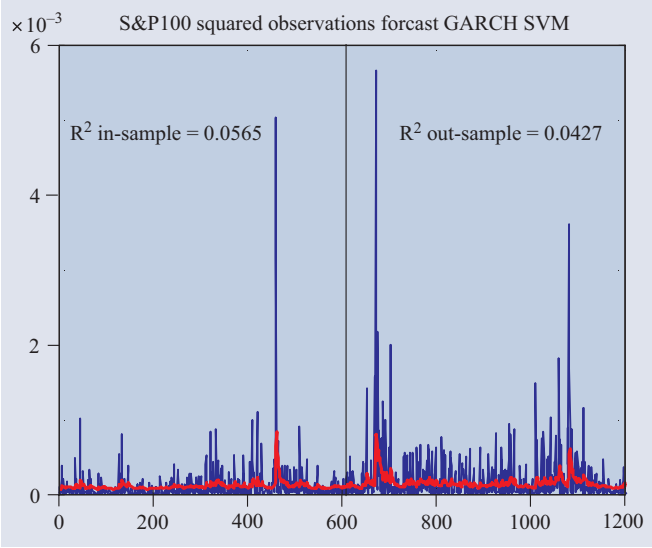

Figure 4. S&P100 (January 1996–October 2000). Squared observations y_t^2 and $\hat{\sigma}_t^2$ for GARCH (1, 1) estimates using SVM procedures.

5.2. The intraday return observations as a measure of realized volatility

The volatility process is not an observable quantity and there have also been different attempts to proposed a computational mechanism for *ex post* estimates of volatility, called realized volatility. The most common method for computing a realized volatility is to square the observed daily returns. If $y_t = \varepsilon_t\sigma_t$, with σ_t being independent of ε_t i.i.d.(0,1), then $E[y_t^2] = E[\varepsilon_t^2\,\sigma_t^2] = E[\sigma_t^2]$. Therefore, accurate forecasting of σ_t^2 will translate in accurate forecasting of y_t^2. This has been the methodology used in the previous sections, independent of the

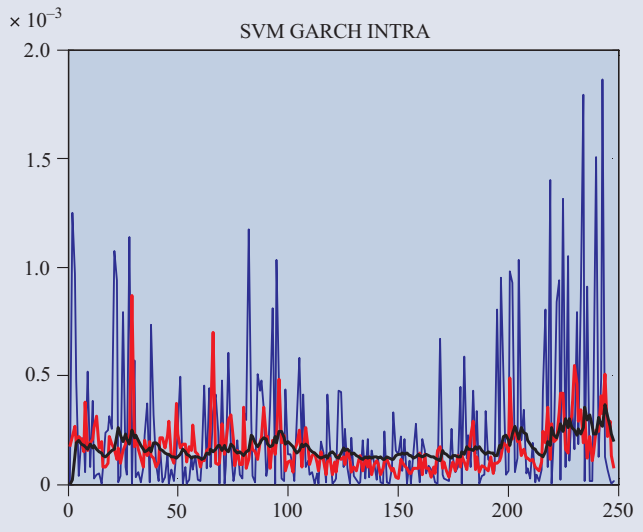

Figure 5. IBEX35 (year 2000). Squared observations y_t^2 in blue (back, thin line), intraday volatility (equation (5.2)) in red (middle) and SVM forecasted volatilities $\hat{\sigma}_t^2$ in black (first).

Table 7. R^2 statistic for GARCH forecasts; the in-sample and out-sample size is $N/2$.

	GARCH ML		GARCH SVR	
	In-sample R^2	Out-sample R^2	In-sample R^2	Out-sample R^2
SP100	0.0466	0.0365	0.0565	**0.0427**
FTSE	0.0911	0.0352	0.0475	**0.0423**
IBEX	0.0590	**0.1341**	0.0502	0.0999
NIKKEI	0.0110	0.0423	0.0108	**0.0479**
GM	0.0175	0.0055	0.0153	**0.0066**
HP	−0.0116	−0.0171	2.16E−04	**0.0048**

estimation of the GARCH parameters, and it can also be seen from figure 4 that the squared volatilities σ_t^2 are quite similar to the squared observations y_t^2.

Andersen and Bollerslev (1998) showed that this method produces inaccurate measures of forecast ability for correctly specified volatility models, such as the former R^2 criterion. Independently of the adequacy of the used model, we will obtain poor predictions for σ_t^2 due to the noisy component ε_t^2. To avoid this problem, one solution is to produce a different volatility measure by sampling intraday data

$$\sigma^2_{\text{intra},t} = \sum_{j=1}^{m} y^2_{(m),t+j/m} \tag{5.2}$$

where m is the sampling frequency. Andersen and Bollerslev (1998) have shown that the noisy component is diminished and that theoretically the realized volatility is then much closer to actual volatility during the day. They found, for foreign exchange data, that the performance of ARCH models improves as the amount of intraday data used to measure the realized volatility increases. Related results for equity and FX data are presented, respectively, in Andersen *et al* (2001a, 2001b).

As an example, we have tested the use of this different measure of realized volatility. We use 15 min intraday data for the IBEX35 index stock market during the year 2000 (248 business days with $m = 34$ quotes each day). We have plotted the squared returns, the intraday volatility (5.2) and the forecasted volatility for the SVR estimation in figure 5. In this plot, it can be seen that the intraday volatility presents a less abrupt behaviour than the squared of the daily observations.

The out-sample R^2 value is improved with this new measure of realized volatility. The out-sample R^2 value for this year, with the parameters estimated using the ML method, increase from 4 to 8%. Using SVM estimation the same measure changes from 3 to 21%. The use of intraday data has increased the explicability of the time series but this increase has been really noticeable for the SVM (over seven times more explicability than with the daily data) than for the ML schemes (about twice as much explicability). Not only do these results show that the intraday volatility measure improves explicability but they also suggest that the SVM is a far more adequate way of estimating the GARCH parameters than the usual ML schemes are.

These results highlight the expected relative superiority of SVM estimation against the ML procedure, developed under more restrictive distributional assumptions (at least under the assumption of Gaussianity in our case), because the former can be interpreted as a robust estimation procedure. Furthermore, Acosta *et al* (2002) show that, even using simulations of a conditionally Gaussian GARCH (1, 1) process, the ML estimation of the conditional variance results in a measure that usually overestimates the magnitude of volatility in the time series. These authors also suggested that the bias in the estimation of the persistence $(\alpha + \beta)$, increases when the variance of the process is highly integrated. This is a characteristic displayed by all series studied in this document; therefore we would expect the same results for all of them, if we could compute the intraday volatility as in the IBEX35 time series.

6. Discussion

In this paper we have used the SVM to estimate the parameters of a GARCH model instead of using the standard ML procedure and have shown that the SVM is able to give more accurate predictions than regular ML estimation. This can be explained by the different natures of the estimates produced by the ML and SVM methods. The former tries to fit the residuals to a Gaussian distribution that if correct will provide the best fit, but if this is not the case, will give an extra error term caused by forcing the residuals to be Gaussian. The SVM tries to get the best fit with the data, not relying on any prior knowledge, and it only concentrates on minimizing the prediction error with a given machine complexity. If the residuals were Gaussian, the SVM will not give such a good solution as the ML estimate, because this method is based on this pdf model, but it will give a set of residuals that are Gaussian. As one can see from table 6, the residuals given by the SVM are not Gaussian, explaining why the predictive ability of the SVM is greater than ML estimation.

We have used the SVM as a linear machine to replace the ML estimation process and we have not tried to obtain a nonlinear estimation using kernels, in which the comparison with other NNs can be readily applied. We have left for future work the development of nonlinear machines in which the estimate of future volatilities will depend nonlinearly on the past volatilities and observations. We have also left for future work the integration of the estimation similar to the BHHH algorithm, in which one will not need to use an intermediate estimation of σ_t. These future studies will allow us to make a full comparison between NNs and SVMs and explain their advantages compared to linear estimation techniques.

Before ending, we would like to discuss the differences between the SVM and most NNs. The SVM has several properties that make it suitable for solving problem in which a linear and nonlinear dependency has to be estimated from the data. The most relevant property is that the SVM is based on a well established learning theory that is able to give bounds on the expected errors and convergence rate given the number of samples and the machine complexity (measure as the VC dimension) (Vapnik 1998). This is a nonasymptotic theory that holds for any number of samples. In addition, the SVM has two extra desirable properties: the functional to be minimized is quadratic and linearly restricted and the machine architecture is given by the learning procedure. The former property ensures that the solution cannot get trapped in local minima (there are none) and the latter property precludes the need to look for the best connections and hidden layers because the SVM solution provides it. The practitioner only needs to find the hyper-parameters of the SVM, which can be found either by using the SRM bounds or by cross validation.

References

Acosta E, Fernández F and Pérez J 2002 Volatility bias in the Garch model: a simulation study *Working paper 20026-02*(http://www.fcee.ulpgc.es/hemeroteca/) University of Las Palmas de Gran Canaria

Andersen T G and Bollerslev T 1998 Answering the skeptics: yes standard volatility models do provide accurate forecasts *Int. Econ. Rev.* **39** 885–905

Andersen T G, Bollerslev T, Diebold F X and Ebens H 2001a The distribution of realized stock return volatility *J. Financial Economics* **61** 43–76

Andersen T G, Bollerslev T, Diebold F X and Labys P 2001b The distribution of exchange rate volatility *J. Am. Stat. Assoc.* **96** 42–55

Berndt E K, Hall B H, Hall R E and Hausman J A 1974 Estimation inference in nonlinear structural models *Ann. Economy Social Meas.* **4** 653–65

Bishop C M 1995 *Neural Networks for Pattern Recognition* (Oxford: Clarendon)

Blair B J, Poon S H and Taylor S J 2001 Forecasting S&P100 volatility: the incremental information content of implied volatilities and high-frequency index returns *J. Econometrics* **105** 5–26

Bollerslev T 1986 Generalized autoregressive conditional heteroskedasticity *J. Econometrics* **31** 307–27

Bollerslev T 1987 A conditional heteroskedasticity time series model for speculative prices and rates of returns *Rev. Economics Statistics* **69** 542–7

Bollerslev T, Engle R F and Nelson D B 1994 ARCH models *The Handbook of Econometrics* vol 4 (Amsterdam: Elsevier) pp 2959–3038

Bollerslev T and Wooldridge J M 1992 Quasi maximum likelihood estimation and inference in dynamic models with time varying covariances *Econometric Rev.* **11** 143–72

Boser B E, Guyon I M and Vapnik V N 1992 A training algorithm for optimal margin classifiers *5th Annual ACM Workshop on COLT (Pittsburgh, PA)* pp 144–52 **Q.5**

Engle R F 1982 Autoregressive conditional heteroskedasticity with estimates of the variance of UK inflation *Econometrica* **50** 957–1008

Engle R F and Patton A J 2001 What good is a volatility model? *Quant. Finance* **1** 237–45

Figlewski S 1997 Forecasting volatility *Financial Markets. Inst. Instrum.* **6** 1–88

Fletcher R 1987 *Practical Methods of Optimization* (New York: Wiley)

Franses P H and Van Dijk D 1995 Forecasting stock market volatility using (nonlinear) GARCH models *J. Forecasting* **15** 229–35 **Q.6**

Huber P H 1964 Robust estimation of location parameter *Ann. Inst. Stat. Math.* **35**

Li W K and Mak T K 1994 On the squared residual autocorrelations in nonlinear time series with conditional heteroskedasticity *J. Time Series Anal.* **15** 627–36

Nelson D B 1991 Conditional heteroskedasticity in asset returns: a new approach *Econometrica* **59** 347–70

Pérez-Cruz F and Artés-Rodríguez A 2000 An IRWLS procedure for nu-SVR *Proc. ICASSP'01* **Q.7**

Pérez-Cruz F, Navia-Vázquez A, Alarcón-Diana P L and Artés-Rodríguez A 2000 An IRWLS procedure for SVR *Proc. EUSIPCO'00*

Schölkopf B and Smola A 2001 *Learning with Kernels* (Cambridge, MA: MIT Press)

Schölkopf B, Smola A, Williamson R and Bartlett P L 2000 New support vector algorithms *Neural Comput.* **12** 1207–45

Smola A, Murata N, Schölkopf B and Müller K R 1998 Asymptotically optimal choice of ε-loss for support vector machines *Proc. ICANN'98*

Theil H 1971 *Principles of Econometrics* (New York: Wiley)

Vapnik V 1982 *Estimation of Dependences Based on Empirical Data* (Berlin: Springer)

Vapnik V 1998 *Statistical Learning Theory* (New York: Wiley)

West K D and Cho D 1995 The predicitive ability of several models of exchange rate volatility *J. Econometrics* **69** 367–91